Matrix perturbations: bounding and computing eigenvalues
Reis da Silva, R.J.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Despite the somewhat negative connotation of the word, not every perturbation is a bad perturbation. In fact, while disturbing the matrix entries, many perturbations still preserve useful properties such as the orthonormality of the basis of eigenvectors or the Hermicity of the original matrix.

In the first part of this thesis, some of these property preserving perturbations are analyzed, with a focus on the consequences to the eigenvalues. For Hermitian rank-k perturbations of Hermitian matrices, this resulted in improved Weyl-type bounds for the perturbed eigenvalues. With respect to normal matrices, normality preserving normal perturbations were also considered and, for 2 x 2 and for rank-one matrices, the analysis is now complete. For higher rank, all essentially Hermitian normality perturbations are described. Moreover, the normality preserving augmentation of normal matrices is revisited and complemented with an analysis of the consequences to the eigenvalues. All augmentations resulting in normal matrices with eigenvalues on a quadratic curve in the complex plane are also constructed.

In the second part, the Subspace Projected Approximate Matrix (SPAM) method, an iterative method for the Hermitian eigenvalue problem, is investigated. For certain choices of the approximation matrix, SPAM turns out to be mathematically equivalent to the Lanczos method. While more sophisticated approximations turn SPAM into a boosted version of the Lanczos method, it can also be interpreted as an attempt to enhance a certain instance of the preconditioned Jacobi-Davidson method.
Matrix Perturbations

bounding and computing eigenvalues

Ricardo Reis da Silva

2011
Matrix Perturbations
bounding and computing eigenvalues

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 10 juni 2011, te 10:00 uur

door

Ricardo Jorge Reis da Silva

geboren te Maia, Portugal
Promotiecommissie

Promotor: Prof. dr. R.P. Stevenson
Co-promotor: dr. J.H. Brandts

Overige leden: Prof. dr. B. Fischer
Prof. dr. ir J.E. Frank
dr. G.L.G. Sleijpen
Prof. dr. ir. C. Vuik
Prof. dr. J.J.O.O. Wiegerinck

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Anicca.
This thesis is not only the result of the four years of my doctorate studies at the Korteweg-de Vries Institute for Mathematics but is also the product of all the people who have supported me as student, researcher, teacher, friend, brother and (grand)son. I do not want to miss this opportunity to publicly thank them for all I have received.

First and foremost I would like to thank my co-promotor, Jan Brandts, for the opportunity to have worked with him. Your dedication to your students is outstanding and I was privileged to have witness it first hand. Your help has been precious and your comments have always challenged me to improve. During these years we were able to go over the supervisor-student paradigm and share also our personal lives. I have learned a lot from you at both levels. Thank you for everything.

I want to thank also to my promoter Rob Stevenson for the remarks on the early versions of this thesis. They were always very precise and allowed me to greatly improve this text. I also greatly appreciated the independence you gave me while assisting you in the Numerical Algorithms course.

Jan Wiegerinck has always made me feel welcome at the department. Despite your busy schedule you always found time to answer my questions and requests or to listen to my remarks. Not only did I deeply enjoyed assisting you in Analyse 3 but I have learned a lot and had great fun doing so.

I am in debt also to Gerard Sleijpen. I would not have started my PhD studies if you haven’t made me interested in Numerical Linear Algebra. I have always valued and respected your knowledge, your advices and your sharp comments and the remarks you made on the manuscript have been very helpful in improving this thesis.

Enno was the first other PhD candidate I met when I arrived at the KdV. We shared an office for a while at the beginning and later when we moved to the Science Park. You were always a big help especially in the first times explaining me who was who at the department and helping me with translations of all kind of forms I needed to fill. I regret not having had the chance to watch one of your concerts but maybe there will still be time for that. With Jevgenijs I have had quite some sporty time. Mud walking in Friesland, climbing and slacklining in Amsterdam and even some invitation for ”badminton” which I was unable to
attend. Your good mood and friendliness were contagious and your adventures were always fun to listen to. Thank you both for everything and I wish you the best of lucks with your plans for the future.

I really enjoyed moving to the new building as there was, finally, a group of PhD candidates with which to share some time outside work. Loek, Piotr, Benjamin, Nabi, Paul, Naser, Abdel, Michel I had really a great time with the drinks and the dinners we had and the football games we watched together.

I am grateful to Andries for all the nice lunches. I have learned a lot from our discussions and despite our divergences in teaching methods your dedication to your students is something I set as an example for myself. To Evelien and Hanneke I want to thank the patience with all my questions and help requests for all kind of issues. You were always very kind and always ready to help make everything so much simpler. To my students throughout these years. I can now tell you that you have, unknowingly, been subjects of teaching experiments I have conducted. You gave me the time to learn how to better communicate, and your questions, feedback and critics have made me a better teacher, a better mathematician and a better researcher. Above all, it has been a great pleasure assisting you with your learning processes.

During the last four years many people have come and gone from the KdV institute. To some of them I want to leave a special thanks and wish them a lot of success in their new endeavors. To Kamyar for all the truly fun Fridays we have spent working on our RKFEM. It is unfortunate that we were unable to finish the 3D version but I have really enjoyed the work. To Phyllis, for our talks and discussions. They were always refreshing and have allowed my mind to rest as it wondered away from mathematics. To Niels for your precise observations on the experiences of expats in the Netherlands. Many times I end up in situations where I remember our conversations and just start laughing! I really enjoyed sharing the office with you and your advices on how to handle my time as a PhD candidate have been very useful. I know I still have an offer for a guided tour to Berlin and I hope I will take it soon.

It was super to co-organize the PhDays 2008 and 2009 with such enthusiastic group of people. Albert-Jan, Joris, Jok Tang, Katrijn, Liesbeth and Peter Lucas, not only am I really proud of what we did but thanks to all of you I had such a great time in our meetings.

I owe a great deal also to many people outside the mathematics community. First, to Roxanna, to whom and above everything I am grateful for keeping me
sane during the last years. Whatever the context you have always prevented me from loosing track of what is important. I will do my best to do the same for you during next four years ;) I have always admired the causes you support and the passion with which you do it and I will never forget how you, your family and your friends have welcomed me so wonderfully in Peru in a trip from which I have only good memories. To Ivelina for the company over the last years to visit museums, going to theaters, watching movies, going to the beach and much more. I always had a wonderful time and it was great to share all those things with someone with such a contagious enthusiasm. To Raquel and Arnt for all the delicious and funny dinners and parties. From sportsman to prince, I do not remember ever having to think so much on how to dress for any other place. Above all I always had a lot of fun when we met and it was great to communicate in Portuguese for a change. To Elif for the concerts, the bike tours, the dinners and parties and especially for all the interesting people I have met through you. I wish you a lot of success for your new life.

Obrigado tambm a todos os meus amigos em Portugal em especial Anita, Nuno, Inês, André, Juliana, Inês, António e Joana por me receberem maravilhosamente sempre que estou por Portugal, pelos jantares, pelas tardes em cafés, por me visitarem e por me continuarem a deixar fazer parte das vossas vidas.

À minha irmã, Patricia, também tenho imenso por que agradecer. O facto de eu não estar em Portugal significa que tens o dobro do fardo em organizar coisas, comprar presentes e tomar conta de toda a gente. Tens tido uma paciência fantástica para tudo isso e tens-no feito de uma maneira soberba. Estou mesmo profundamente agradecido.

Finalmente, aos meus pais e avós. Eu sei que vos tem custado especialmente o facto de eu não viver tão perto como gostariam. Infelizmente ainda não vos posso dizer que acabou. Eu sei que não entendem as minhas opções completamente e que, sem duvidar que estejam contentes comigo, gostariam que a minha vida fosse diferente. Eu sinto-me um privilégiado por poder viver da maneira que quero. Viver a minha vida o melhor possível para mim e para os outros é honrar a educação e as oportunidades que me deram. Obrigado por tudo isso e por todo o apoio que me continuam a dar.

Thank you all,

Ricardo

10 June 2011
## Contents

<table>
<thead>
<tr>
<th>List of Figures</th>
<th>xi</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>xv</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Perturbation theory</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Iterative methods for Eigenvalue Problems</td>
<td>13</td>
</tr>
<tr>
<td>2 Perturbation theory</td>
<td>23</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>26</td>
</tr>
<tr>
<td>2.2 Hermitian rank-k perturbations</td>
<td>29</td>
</tr>
<tr>
<td>2.3 Bordered diagonal matrices</td>
<td>39</td>
</tr>
<tr>
<td>2.4 Singular Values</td>
<td>42</td>
</tr>
<tr>
<td>2.5 Numerical illustrations</td>
<td>47</td>
</tr>
<tr>
<td>2.6 Final notes</td>
<td>52</td>
</tr>
<tr>
<td>3 Normal matrices</td>
<td>59</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>60</td>
</tr>
<tr>
<td>3.2 Normality preserving augmentation</td>
<td>66</td>
</tr>
<tr>
<td>3.3 Normality preserving normal perturbations</td>
<td>70</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.4 Further augmentations</td>
<td>78</td>
</tr>
<tr>
<td>3.5 Illustrations</td>
<td>84</td>
</tr>
<tr>
<td>II Iterative methods</td>
<td>93</td>
</tr>
<tr>
<td>4 The Subspace Projected Approximate Matrix method</td>
<td>95</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>96</td>
</tr>
<tr>
<td>4.2 SPAM and other subspace methods</td>
<td>102</td>
</tr>
<tr>
<td>4.3 Selecting the matrix $A_0$ in the SPAM method</td>
<td>109</td>
</tr>
<tr>
<td>4.4 Numerical illustrations</td>
<td>119</td>
</tr>
<tr>
<td>4.5 Conclusions</td>
<td>129</td>
</tr>
<tr>
<td>4.6 Comments on a SPAM linear system solver</td>
<td>130</td>
</tr>
<tr>
<td>5 Summary</td>
<td>133</td>
</tr>
<tr>
<td>5.1 The mathematical model</td>
<td>134</td>
</tr>
<tr>
<td>5.2 From systems to matrices</td>
<td>136</td>
</tr>
<tr>
<td>5.3 This thesis</td>
<td>138</td>
</tr>
<tr>
<td>Notation</td>
<td>139</td>
</tr>
<tr>
<td>Bibliography</td>
<td>141</td>
</tr>
<tr>
<td>Index</td>
<td>149</td>
</tr>
<tr>
<td>Samenvatting</td>
<td>151</td>
</tr>
<tr>
<td>About the Author</td>
<td>153</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Error in the lower bound (and detail for $q \leq 10$) of the smallest eigenvalue of the perturbed matrix $A$ resulting from the Finite Difference discretization ($n = 32$, $\xi = 1/33$) of Problem $(2.17)$ with $E$ a rank-one diagonal matrix resulting from Equation $(2.18)$. . . . . 49

2.2 Error in the lower bound (and detail for $q \leq 10$) of the smallest eigenvalue of a random rank-one perturbation $E$ of the matrix $A$ obtained from the discretization of Poisson equation on a $40 \times 40$ grid. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51

2.3 Error in the lower bound (and detail for $q \leq 10$) of the smallest eigenvalue of the perturbed matrix Poisson with $E$ a rank-2 random matrix. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52

2.4 Error in the lower bound of the smallest eigenvalue of the perturbed matrix $A$ for the matrix Sherman1 when a multiple eigenvalue is clustered (below) and comparison with the naive approach (above). 53

2.5 Error in the lower bound (and detail for $q \leq 20$) of the smallest singular value of $B_r$ where $B$ is the matrix ILLC1850 of dimensions $1850 \times 712$ from the Matrix Market and $c$ is a random $n$-dimensional vector. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53

3.1 Interpolating polynomials of different degree for different values of $\theta$. The seven asterisks represent the eigenvalues. . . . . . . . . . . . . 64
3.2 Eigenvalue perturbation by a rank-one matrix (left) and a rank-$k$ matrix (right). ........................................ 77

3.3 Illustration of the construction in the proof of Theorem 3.4.7. Three of the seven eigenvalues of $A$, indicated by the circles, are already on the quadratic curve $C$, and a rank-4 matrix $Z$ is needed to push the remaining four eigenvalues onto $C$, after which the augmented matrix can be formed. ........................................ 83

3.4 Normality preserving augmentations of a $3 \times 3$ matrix for two different values of $\gamma$. ........................................ 86

3.5 Eigenvalue trajectories of a normality preserving perturbation, and of the same perturbation written as the sum of non-normality preserving normal perturbations. ........................................ 89

4.1 Arrowhead updates from $\hat{A}_{k-1}$ to $\hat{A}_k$ for consecutive values of $k$. ........................................ 99

4.2 Approximating a matrix $A$ from structural engineering from below by a sparser matrix $A_0$ by subtracting from $A$ a definite matrix $H$; the sparsity plot of $A$ (with 3648 nonzero entries) is on the left and of $A_0$ (with 968 nonzero entries) on the right. See the experiments with this pair of matrices in Section 4.4.4] ........................................ 116

4.3 Lanczos method (left) with a random start vector, versus SPAM (right) with the discretized reaction term as approximating matrix $A_0$, and the largest eigenvalue of $A$ as target. ........................................ 123

4.4 From left to right: Lanczos and Full SPAM approximating the largest, the second, the fifth, and the smallest eigenvalue of $A$, using algebraic rank-12 approximation from below (for the smallest eigenvalue, we applied both the methods to $6I - A$). Lanczos and Full SPAM used in each experiment the same start vector.] ........................................ 124

4.5 Lanczos versus Full SPAM and SPAM(1) with diagonal (left), tridiagonal (middle left), and with algebraic rank-6 approximation from below (both middle right and right). In the three leftmost pictures, the target was the largest eigenvalue, at the right it was the smallest eigenvalue (i.e., the largest of $\alpha I - A$). ........................................ 125
4.6 Lanczos versus Full SPAM and SPAM(1) for bcsstk04, bcsstk07 and bcsstk10, with approximating matrices from below. All three methods used the same start vector. 126

4.7 Lanczos and SPAM(1) compared with SPAM(1,ℓ) for small values of ℓ. Left: reaction-diffusion problem, smallest eigenvalue. Other pictures: largest eigenvalue. Middle left: banded matrix; middle right: bcsstk07; right: bcsstk10. The graphs for Lanczos and SPAM(1) can also be found in Figures 4.4, 4.5, and 4.6. 127

4.8 Comparing SPAM(1,ℓ) with JD(1,ℓ) and JD(ℓ). The eigenvalue problems are exactly the same as the corresponding ones in Figure 4.7, and the curves for SPAM(1,ℓ) can be found there as well. 128

5.1 Two stories building. The elasticity of the walls is denoted by the \( a_j \)s, their displacement by the \( \tilde{x}_j \)s and the mass of the ceilings are denoted by the \( m_j \)s for \( j \in \{1,2,3\} \). 134
List of Tables

2.1 Eigenvalues of $X_q + E$ for $q \in \{1, \ldots, 5\}$ ........................................... 28
2.2 Eigenvalues of $X_q + E$ for $q \in \{1, \ldots, 5\}$ from Example 2.2.2 computed using the approach of Theorem 2.2.2 .................................................. 35
2.3 Different lower bounds for the eigenvalues of $A+E$ given by the eigenvalues of $X_3 + E$ and $X'_3 + E$ ................................................................. 37
2.4 Error in $\mu_- := \lambda_{\text{min}}(Y_r + E)$ for $1 \leq r \leq n$ for a first few lower bounds for the matrix obtained from the discretization of Poisson equation on a $40 \times 40$ grid (left) and Sherman1 from Matrix Market (right). The perturbations are random rank-one matrices. ..................... 50
2.5 Eigenvalues of $X_3 + E$ and $A + E$ where $E$ is a non-Hermitian perturbation. .......................................................... 58
Introduction

*Perturbation*,
a disturbance of motion, course, arrangement, or state of equilibrium.
- Merriam-Webster dictionary

Despite the somewhat negative connotation of the word, not every perturbation is a bad perturbation. In fact, while disturbing the entries of a matrix, many perturbations still preserve useful properties such as the orthonormality of the basis of eigenvectors or the Hermicity of the original matrix.

In the coming chapters we will encounter several property-preserving perturbations. Some of these are as ordinary as Hermitian perturbations of Hermitian matrices. Some others, such as normality preserving perturbations, or those yielding approximations from below are less studied. Nevertheless, also these turn out to hide some charming results. For instance, the reader might be surprised to discover that the knowledge of a single eigenpair of a normal matrix is all that is needed to augment it with a row and a column while retaining its normality. Or, that among all the linear combinations of normal matrices, $tA + (1 - t)B$ with $t \in \mathbb{R}$, there are either two or infinitely many normal matrices.
We shall also encounter some pleasant perturbations. Those that affect entries in such a way that the properties of the resulting matrix are more gracious than the ones we started from. We find them, for instance, in the context of iterative methods for eigenvalue problems. These are perturbations that make the computation of the eigenvalues more efficient by introducing extra sparsity and/or a better structure. Determining general suitable perturbations for that purpose is still a difficult task. Nevertheless, in the framework of the Subspace Projected Approximate Matrix method some property-preserving perturbations turn out to be both suitable and to simplify the theoretical analysis.

Broadly speaking, however, this thesis is less about the perturbations themselves and more about the eigenvalues of the resulting matrix. Therefore, among the main results, the reader will find improved bounds for the eigenvalues of Hermitian perturbations of Hermitian matrices and new estimates for the eigenvalues of certain normality preserving perturbations of normal matrices.

1.1 Perturbation theory

Research on eigenvalue problems is one of the most prolific topics in Numerical Linear Algebra [31]. Perturbation analysis, in particular, studies the consequences of small changes in a matrix $A \in \mathbb{C}^{n \times n}$ to its eigenvalues and/or eigenvectors. The perturbation may result from the inherent inexact nature of numerical computations but also from research questions in the context of discretized differential equations, or stability of dynamical systems, to name a few. Independently of their origin, and assuming that the eigendata of $A$ are known, the problem is to determine the eigenvalues (and sometimes also eigenvectors) of the matrix

$$A + E$$

where $E \in \mathbb{C}^{n \times n}$ is some perturbation matrix. In many cases, however, rather than the exact eigenvalues of $A + E$, lower or upper bounds suffice and the deeper our knowledge of the characteristics of $A$ and $E$, the better can it be exploited to improve those bounds.
We may collect the results in perturbation theory for eigenvalues into two major groups. The first includes bounds for the error (absolute or relative) of the perturbation and examples of important results are Hoffman-Wielandt’s ([38], [39, Thm. 6.3.5]) and Bauer-Fike’s theorems. Because we shall refer to Bauer-Fike’s theorem often in the next chapters, we chose to present here also its proof.

**Theorem 1.1.1** (Bauer-Fike [5]). For $n \in \mathbb{N}$, let $\tilde{A}, A$ and $E$ be $n \times n$ matrices satisfying $\tilde{A} = A + E$. Assume, furthermore, that $V \in \mathbb{C}^{n \times n}$ is nonsingular and such that $\Lambda = V^{-1}AV$ is diagonal. Finally, for $j \in \{1, \ldots, n\}$ denote by $\lambda_j$ the eigenvalues of $A$ and let $\mu$ be an eigenvalue of $\tilde{A}$. Then, for at least one eigenvalue, $\lambda_j$, of $A$

$$|\lambda_j - \mu| \leq \kappa_p(V)\|E\|_p,$$

where $\kappa_p(V) = \|V^{-1}\|_p\|V\|_p$ is the condition number of $V$ with respect to the matrix $p$-norm.

**Proof.** For notation simplicity, we define in this proof $\| \cdot \| := \| \cdot \|_p$ for the matrix $p$-norm. If $\mu$ is an eigenvalue of $A$, it is enough to take $\lambda_j = \mu$. We assume, therefore, that $\mu$ is not an eigenvalue of $A$. Owing to the fact that $\mu \in \sigma(\tilde{A})$, the matrix $\tilde{A} - \mu I$ is singular and since $\text{det}(\Lambda - \mu I) \neq 0$ we may write

$$0 = \text{det}(\tilde{A} - \mu I) = \text{det}(\Lambda + V^{-1}EV - \mu I) = \text{det}(\Lambda - \mu I)\text{det}[(\Lambda - \mu I)^{-1}V^{-1}EV + I].$$

Thus, $\text{det}[(\Lambda - \mu I)^{-1}V^{-1}EV - (-1)I] = 0$. This points to $-1$ as an eigenvalue of $(\Lambda - \mu I)^{-1}V^{-1}EV$. Therefore,

$$1 \leq \|(\Lambda - \mu I)^{-1}V^{-1}EV\| \leq \|((\Lambda - \mu I)^{-1}V^{-1})\|\|E\|\|V\|$$

$$= \kappa_p(V)\|(\Lambda - \mu I)^{-1}\|\|E\|.$$

Here, $\kappa_p(V) = \|V^{-1}\|\|V\|$ and the first inequality follows from the fact that, for consistent norms, $\|A\|$ is never smaller than the spectral radius of $A$. This gives

$$1 \leq \max |(\lambda_j - \mu)^{-1}\|E\|\kappa_p(V) \text{ or equivalently}$$

$$\min |\lambda_j - \mu| \leq \kappa_p(V)\|E\|.$$
Then, for at least one $j$, we have $|\lambda_j - \mu| \leq \kappa_p(VE)$, proving the statement.

The other group of perturbation results focuses on the relative position of the eigenvalues of $A$ and $A + E$. Important results include Weyl’s inequalities on Hermitian perturbations of Hermitian matrices but also, in some sense, Cauchy’s Interlace Theorem. The two types of results are not mutually exclusive and they often appear together. Our interest, however, lies with this second group which we now develop in more detail.

1.1.1 Classical theory

Determining all values $\lambda$ for which $Av = \lambda v$ has a nonzero solution $v$ is called the standard algebraic eigenvalue problem. Given such an eigenvector $v$ of unit norm, we may use the same equation to determine the associated eigenvalue by writing $\lambda = v^*Av$. For a matrix $A$ (usually Hermitian), the quadratic form that relates an arbitrary vector in $\mathbb{C}^n$ with a complex scalar is known as the Rayleigh quotient or Rayleigh-Ritz ratio.

**Definition 1.1.2** (Rayleigh quotient). For a matrix $A \in \mathbb{C}^{n \times n}$ the Rayleigh quotient is the map

$$\rho_A : \mathbb{C}^n \setminus \{0\} \to \mathbb{C} : x \mapsto x^*Ax / x^*x.$$ 

For normal matrices, the Rayleigh quotient possesses several agreeable properties that include homogeneity, boundedness and translation invariance (see [62] or [63, §1.5]) and of which we emphasize one:

**Lemma 1.1.3** (Minimal Residual [62]). Given $x \in \mathbb{C}^n \setminus \{0\}$, the scalar $\rho_A(x)$ is the quantity that minimizes $\|Ax - \mu x\|_2$. Otherwise stated,

$$\|Ax - \rho_A(x)x\|_2 \leq \|Ax - \mu x\|_2, \quad \text{for all } \mu \in \mathbb{C}.$$ 

**Proof.** Set $\rho := \rho_A(x)$. Then,

$$\|Ax - \mu x\|_2^2 = x^*x|\mu|^2 - \bar{\mu}x^*Ax - \mu x^*A^*x + x^*A^*Ax$$

$$= x^*x(x^*A^*Ax / x^*x + (\mu - \rho)(\bar{\mu} - \bar{\rho}) - |\rho|^2)$$

$$\geq \|Ax\|_2^2 - |\rho|^2\|x\|_2^2.$$
The statement is proved noting that \( \|Ax - \rho x\|^2_2 = \|Ax\|^2_2 - |\rho|^2 \|x\|^2_2 \).

Such properties make Rayleigh’s quotient an important tool not only in perturbation theory but also for the actual computation of eigenvalues. In fact, assuming that \( \|x\|_2 = 1 \), the norm of the spectral residual, \( \|Ax - \rho_A(x)x\|_2 \), is a measure of how close a given pair, \((x, \rho_A(x))\), formed by a vector and its Rayleigh quotient is to an exact eigenpair of \(A\). It is, therefore, often used as stopping criterion for iterative methods.

For a matrix \(A\), the subset of the complex plane resulting from considering the Rayleigh quotient of every nonzero vector \(x \in \mathbb{C}^n\) constrained to the unit \((n-1)\)-sphere is called the field of values of the matrix \(A\).

**Definition 1.1.4 (Field of values).** Given \(A \in \mathbb{C}^{n \times n}\), the field of values of \(A\) is the set

\[
F(A) = \{x^*Ax : \|x\|_2 = 1, x \in \mathbb{C}^n\}.
\]

The set \(F(A)\) is the image of the unit \((n-1)\)-sphere under the quadratic form \(x^*Ax\) and is closed, bounded and convex in \(\mathbb{C}\). During the last sixty years, several generalizations of the field of values were proposed [40, §1.8]. Examples include Givens field of values generalizing the inner-product [29]; Bauer field of values with different right and left vectors [4, 58, 85]; the \(q\)-field of values [53, 79]; and more recently the generalized two-sided field of values [69] by the author of this thesis. Each generalization attempts to replicate one or more properties of the standard field of values.

If \(A\) is a normal matrix, \(F(A)\) is the convex hull of the eigenvalues of \(A\). In particular, if \(A\) is Hermitian, the field of values is a real interval. In this case, we are able to express the largest and smallest eigenvalues of \(A\) as solutions of constrained optimization problems involving \(F(A)\).

**Theorem 1.1.5 (Rayleigh-Ritz [39]).** If \(A \in \mathbb{C}^{n \times n}\) is a Hermitian matrix with eigenvalues \(\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n\) then

\[
\lambda_1 = \min_{x \in \mathbb{C}^n \setminus \{0\}} \frac{x^*Ax}{x^*x} = \min_{x \in \mathbb{C}^n, \|x\|_2 = 1} x^*Ax,
\]

\[
\lambda_n = \max_{x \in \mathbb{C}^n \setminus \{0\}} \frac{x^*Ax}{x^*x} = \max_{x \in \mathbb{C}^n, \|x\|_2 = 1} x^*Ax.
\]
\[ \lambda_n = \max_{x \in \mathbb{C}^n \setminus \{0\}} \frac{x^*Ax}{x^*x} = \max_{x \in \mathbb{C}^n, \|x\|_2 = 1} x^*Ax \]

and consequently,

\[ \lambda_1 \|x\|_2^2 \leq x^*Ax \leq \lambda_n \|x\|_2^2, \quad \forall x \in \mathbb{C}^n. \]

**Proof.** Let \( n \in \mathbb{N} \) and let \( V \in \mathbb{C}^{n \times n} \) be unitary and such that \( \Lambda = V^*AV \) is diagonal. Recall, that for any vector \( x \in \mathbb{C}^n \) we have \( \|V^*x\|_2^2 = x^*VV^*x = x^*Ix = \|x\|_2^2 \). Consequently,

\[ x^*Ax = (V^*x)^*\Lambda(V^*x) = \sum_{j=1}^n \lambda_j |(V^*x)_j|^2. \]

Now, replacing every \( \lambda_j \) by either \( \lambda_{\text{min}} \) or \( \lambda_{\text{max}} \) we obtain the inequalities

\[ \lambda_{\text{min}} \sum_{j=1}^n |(V^*x)_j|^2 \leq \lambda_j \sum_{j=1}^n |(V^*x)_j|^2 \leq \lambda_{\text{max}} \sum_{j=1}^n |(V^*x)_j|^2 \]

or differently stated

\[ \lambda_1 \|x\|_2^2 \leq x^*Ax \leq \lambda_n \|x\|_2^2. \]

For the remaining statements, note that if \( Ax = \lambda_j x \) then \( x^*Ax = \lambda_j \|x\|_2^2 \). \( \Box \)

If \( A \) is not normal, however, \( F(A) \) is, generally, no longer the convex hull of \( \sigma(A) \) and consequently it is not maximized or minimized in the eigenvectors associated with the largest and smallest eigenvalues of \( A \), even if these are real.

In fact, not only the largest and smallest but every eigenvalue of a Hermitian matrix \( A \) can be expressed in terms of a constrained optimization problem. This result is known as the Courant-Fischer minimax theorem \([20, 25]\) whose proof we present here following the approach from Parlett \([63\text{ Thm. 10.2.1}]\).

**Theorem 1.1.6** (Courant-Fischer \([63]\)). Let \( A \in \mathbb{C}^{n \times n} \) be Hermitian and let \( \mathcal{W}^k \) denote a subspace of \( \mathbb{C}^n \) with dimension indicated by the superscript. Then, the \( j \)th eigenvalue of \( A \) in non-decreasing order of magnitude satisfies

\[ \lambda_j(A) = \min_{\mathcal{W}^j} \max_{0 \neq x \in \mathcal{W}^j} \rho_A(x) = \max_{\mathcal{W}^{n-j+1}} \min_{0 \neq u \in \mathcal{W}^{n-j+1}} \rho_A(u) \]
Proof. The dimensions of $W^j$ and $W^{n-j+1}$ and the basis theorem for finite dimensional spaces ensure that $W^j \cap W^{n-j+1} \neq \{\vec{0}\}$. Denote by $z$ any nonzero vector in $W^j \cap W^{n-j+1}$. Then,

$$\min_{\vec{u} \neq \vec{0} \in W^{n-j+1}} \rho_A(\vec{u}) \leq \rho_A(z, z) \leq \max_{\vec{u} \neq \vec{0} \in W^j} \rho_A(\vec{x}).$$

Because the inequalities are valid for all choices of $W^j$ and $W^{n-j+1}$, we may take the maximum over all of them on the left-hand side and the minimum on the right-hand side leading to

$$\max_{W^{n-j+1}} \min_{\vec{u} \neq \vec{0} \in W^{n-j+1}} \rho_A(\vec{u}) \leq \min_{W^j} \max_{\vec{u} \neq \vec{0} \in W^j} \rho_A(\vec{x}).$$

Recall that, if $x = u = v_k$ is an eigenvector of $A$ associated with eigenvalue $\lambda_k$, its Rayleigh quotient satisfies $\rho_A(v_k) = \lambda_k$. Therefore, equality follows from choosing $W^j = \text{span}\{v_1, \ldots, v_j\}$ and $W^{n-j+1} = \text{span}\{v_j, \ldots, v_n\}$ where for each $k \in \{1, \ldots, n\}$ each $v_k$ is an eigenvector of $A$ associated with eigenvalue $\lambda_k$. \qed

1.1.2 Perturbation results

Because when $A$ and $E$ are Hermitian so is $A + E$, Courant-Fischer’s theorem may be used to estimate its eigenvalues in terms of those of $A$ and $E$. This is the statement of the next theorem, originally presented by Weyl.

**Theorem 1.1.7 (Weyl’s inequalities [81]).** Let $A$ and $E$ be $n \times n$ Hermitian matrices. For $j \in \{1, \ldots, n\}$ denote by $\lambda_j, \mu_j$ and $\delta_j$ the $j$th eigenvalue of $A + E$, $A$ and $E$ respectively. Assume these to be in non-decreasing order of magnitude. Then,

$$\mu_j + \delta_1 \leq \lambda_j \leq \mu_j + \delta_n.$$

**Proof.** For any vector $x \in \mathbb{C}^n$, Theorem 1.1.5 guarantees that

$$\delta_1 x^* x \leq x^* E x \leq \delta_n x^* x.$$

Therefore, using Courant-Fischer’s theorem (Theorem 1.1.6), we have for each
\( j \in \{1, \ldots, n\} \)

\[
\lambda_j = \min_{W_j \subseteq \mathbb{C}^n, \, \delta \neq x \in W_j} \max_{x} \frac{x^*(A + E)x}{x^*x} = \min_{W_j \subseteq \mathbb{C}^n, \, \delta \neq x \in W_j} \max_{x} \left( \frac{x^*Ax}{x^*x} + \frac{x^*Ex}{x^*x} \right)
\]

\[
\geq \min_{W_j \subseteq \mathbb{C}^n, \, \delta \neq x \in W_j} \max_{x} \left( \frac{x^*Ax}{x^*x} + \delta_1 \right) = \mu_j + \delta_1.
\]

The remaining inequality is obtained using similar arguments.

Theorem 1.1.7 provides us with bounds for the eigenvalues of \(A + E\) in terms of the eigenvalues of \(A\) and \(E\). Because the eigenvalues of \(A\) are assumed to be known, the usefulness of Theorem 1.1.7 depends on how much is known of \(E\). For instance, if \(E\) is a rank-one Hermitian positive semidefinite matrix, Theorem 1.1.7 simplifies into the following Corollary.

**Corollary 1.1.8.** Let \(E = ww^*\) be a Hermitian rank-one positive semidefinite matrix. Then,

\[
\mu_j \leq \lambda_j \leq \mu_j + \|w\|^2_2.
\]

Note that Corollary 1.1.8 may, equivalently, be stated in terms of a rank-one negative semidefinite matrix \(E\). In either situation, if the eigenvalues of \(A\) are assumed to be known, this results in easily computable bounds for the eigenvalues of \(A + E\). Corollary 1.1.8 is of interest on its own, but also because perturbing a Hermitian matrix \(A\) with a Hermitian matrix \(E\) can, at least in theory, be studied from a succession of perturbations of \(A\) by Hermitian rank-one matrices.

**Recent contributions**

The following more recent contributions on perturbation theory will prove useful for the contextualization of the results from Chapter 2.

A renewed interest in the analysis of rank-one perturbations of diagonal matrices grew with the development of the divide-and-conquer algorithm [21] for the computation of the eigenvalues of Hermitian tridiagonal matrices. The algorithm divides a tridiagonal matrix into a sum of a block diagonal matrix with tridiagonal blocks and a Hermitian rank-one matrix. Given the diagonalization of the submatrices, the remaining problem is that of determining the eigenvalues of a
Hermitian rank-one perturbation $E = w\sigma w^*$ of a diagonal matrix $D$. If the eigenvalues, $\delta_1, \ldots, \delta_n$, of $D$ are distinct and $w_j \neq 0$, then solving $(D + E)v = \lambda v$ with $w^*v \neq 0$ is equivalent to solving $1 + \sigma w^*(D - \lambda I)^{-1}w = 0$ or, in scalar form, to determining the roots of the so-called secular equation:

$$f(\lambda) = 1 + \sigma \sum_{j=1}^{n} \frac{w_j^2}{\delta_j - \lambda}. \quad (1.1)$$

The eigenvalues of a diagonal plus a rank-one matrix are usually computed using methods based on the secular equation [15, 51, 54, 59, 83]. In Section 5 of [30], Golub describes alternative methods for the eigenvalue problem of a rank-one perturbation of a diagonal matrix. His approach is to transform the original problem into one for which efficient algorithms already exist. Methods based on the secular equation (among others) are mentioned and an alternative $O(n^2)$ algorithm is devised. Bunch et al. [15] analyze the computation of the eigenvalues of rank-one perturbations of general Hermitian matrices. Golub’s results (from [30]) are extended to include the eigenvector computation as well as the possibility of the matrix having multiple eigenvalues.

Weyl’s inequalities, used to bound the eigenvalues of $A + E$ from below by the eigenvalues of $A$, are among the best known and easily computable results. More recently, however, Ipsen and Nadler [46] show how Weyl’s bounds can be improved at a small cost. These authors bound the smallest eigenvalues of $A + E$ from below by the smallest eigenvalue of a $2 \times 2$ eigenproblem resulting from clever manipulations of the matrices $A$ and $E$. Making use of Cauchy’s Interlace Theorem (see Theorem 1.1.10) they propose, also, lower bounds for the interior eigenvalues of $A$. Additional bounds are given for (well-separated) eigenvalues of Hermitian matrices when the perturbation is Hermitian or non-Hermitian with sufficiently small norm.

Perturbation problems arise not only for the eigenvalue but also for the singular value problem. Given a matrix $B$ with known singular values, one wishes to determine the singular values of some perturbed matrix $\tilde{B}$. A survey in perturbation theory for the Singular Value Decomposition is given by Stewart in [76, 77]. It covers results ranging from singular value bounds to low rank approximations.
and singular vectors perturbation. Applications in the field of image analysis and information retrieval can be found in [16] and [78] respectively. Another frequent problem occurs when $\tilde{B}$ is the matrix $B$ after columns or rows were added or removed. Bunch and Nielsen [14] studied this situation and developed algorithms for the efficient computation of the updated singular value decomposition as well as for the updated least squares problem.

1.1.3 Normal matrices

As soon as we leave the safe haven of Hermitian matrices, perturbation results and, in particular, perturbation results in terms of the relative position of the eigenvalues, become scarce. One reason is that for general normal matrices, those whose basis of eigenvectors is still orthonormal, many fundamental questions are still unanswered. We contextualize the results from Chapter 3 with some background on perturbation results for normal matrices.

The location of the eigenvalues of sums of normal matrices has been studied as early as 1955 by Wielandt [82]. He did not, however, require their sum to be normal. Another commonly studied subclass of normal matrices is that of essentially Hermitian matrices [7, 23]. Essentially Hermitian matrices are matrices $E$ that can be written as

$$E = \beta H + \alpha I,$$

where $\alpha, \beta \in \mathbb{C}$ and where $H$ is Hermitian.

Faber and Manteuffel show in [24] that for linear systems with essentially Hermitian system matrix, there exists a variant of the Conjugate Gradient method that still relies on a three term recursion. In the context of the Arnoldi method applied to normal matrices, Huckle [41] proves that irreducible normal tridiagonal matrices are essentially Hermitian. In fact, the upper Hessenberg matrix generated by the Arnoldi method is tridiagonal only for essentially Hermitian matrices. The Arnoldi’s method can, in that case, be simplified to the Lanczos method.

In [42], Huhtanen shows that for every normal matrix $A$ and almost all unimodular $z \in \mathbb{C}$, the skew-Hermitian part $S(zA)$ of $zA$ is a polynomial $p$ of degree at most $n - 1$ in the Hermitian part $H(zA)$. The essentially Hermitian matrices
are exactly the ones for which the degree of \( p \) is at most one, resulting in collinear eigenvalues. Since \( p \) can be retrieved in a modest number of arithmetic operations as a by-product of the Lanczos algorithm applied to \( H(zA) \), this led to the development \cite{12, 43} of efficient structure preserving algorithms for eigenvalue problems and linear systems involving normal matrices.

In \cite{44}, Ikramov investigates which matrices \( R \) can be the upper triangular part of a normal matrix \( A \). Although the full problem remains open, a characterization exists if all diagonal entries of \( R \) lie on the same line in the complex plane, say

\[ \ell : \{ \gamma + \theta \rho \mid \rho \in \mathbb{R} \}, \quad \gamma, \theta \in \mathbb{C} \text{ with } |\theta| = 1. \]

Writing \( R = D + U \) where \( D \) is the diagonal of \( R \), we see that \( \bar{\theta}(D+U - \gamma I) \) is upper triangular with real diagonal and, consequently, that \( H = \bar{\theta}(D+U - \gamma I) + \theta U^* \) is Hermitian. Therefore,

\[ A = \theta H + \gamma I = R + \theta^2 U^* \]

is essentially Hermitian, hence normal, with upper triangular part equal to \( R \). Ikramov and Elsner also studied the normal augmentation problem which is a particular case of the topic of the next section.

### 1.1.4 Augmentation of matrices

A problem related to perturbation of matrices is the one of augmenting a matrix with one or more rows and columns. Formally, we may state it as follows. Given \( A \in \mathbb{C}^{n \times n} \), characterize all \( n \times m \) matrices \( V, W \in \mathbb{C}^{n \times m} \), and all \( \Gamma \in \mathbb{C}^{m \times m} \) such that

\[ A_+ = \begin{bmatrix} A & V \\ W^* & \Gamma \end{bmatrix}, \tag{1.2} \]

satisfies certain properties.

The matrix \( A \) is called a principal submatrix of \( A_+ \).

**Definition 1.1.9** (Principal submatrix). A \( k \times k \) matrix \( A \) is called a principal submatrix of \( A_+ \) if \( A \) is obtained from \( A_+ \) by removing \( n - k \) columns and their corresponding rows.
Note that we may obtain principal submatrices by removing consecutive rows and columns from the top left, as well as from the bottom right corners of $A_+$. The resulting principal submatrices are called leading and trailing respectively.

If $A$ is Hermitian then $A_+$ is also Hermitian if and only if $W = V$ and $\Gamma = \Gamma^*$. In this case, Cauchy Interlace Theorem relates the eigenvalues of the Hermitian matrix $A_+$ with those of $A$.

**Theorem 1.1.10** (Cauchy Interlace [63]). Let $A \in \mathbb{C}^{n \times n}$ be Hermitian and let $V$ and $\Gamma$ be such that

$$
A_+ = \begin{bmatrix}
A & V \\
V^* & \Gamma
\end{bmatrix}
$$

is Hermitian. Denote by

$$
\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n \quad \text{and} \quad \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{n+m}
$$

the eigenvalues of $A$ and $A_+$ respectively. Then

$$
\lambda_j \leq \mu_j \leq \lambda_{j+m} \quad \text{for all} \quad j \in \{1, \ldots, n\}.
$$

In short, Cauchy’s Interlace Theorem states that the eigenvalues of $A$ are interlaced by those of any of its submatrices. The proof does not add relevant information. We therefore only present the result, and direct the reader to [63 §10.1] or [39 Thm 4.3.8] for two possible proofs.

If $A$ is normal and we want $A_+$ to be normal too the situation is more difficult. Ikramov and Elsner [45] dealt with this normality preserving augmentation for $m = 1$ and $m = 2$. In particular for $m = 1$, with the matrices $V, W$ and $\Gamma$ reduced to vectors $v$ and $w$ and a scalar $\gamma$, they established that for an arbitrary complex number $\xi$ of modulus one, $v$ must be an eigenvector of the matrix $A - \xi A^*$ while the vector $w$ equals $w = \xi v$. The scalar $\gamma$ is then dependent on the eigenvalue of $A - \xi A^*$ associated with $v$. The problem in its full generality, however, remains unsolved.

As a final remark on this subject, note that for the case that $m = 1$ and the matrix $A$ is diagonal and $v$ and $w$ are nonzero vectors, then the matrix $A_+$ is a so-
1.2 Iterative methods for Eigenvalue Problems

Although every method to obtain the eigenvalues of a matrix is, in essence, iterative, the characterization as iterative is usually restricted to those that determine only a few eigenvalues (and eigenvectors) of large and usually sparse matrices [80]. These methods restrict the matrix $A$ to a suitable subspace and search for approximations to the eigenvalues of $A$ in that subspace. For this reason, they are also called subspace methods. Important concepts in this context are those of Ritz values and Ritz vectors.

1.2.1 Ritz values and vectors

Let $V$ be an $n \times k$ matrix with mutually orthonormal columns and define a $k$-dimensional subspace of $\mathbb{C}^n$ as

$$\mathcal{V} = \{Vy \mid y \in \mathbb{C}^k\}.$$  

In the context of iterative methods, we shall call the subspace $\mathcal{V}$ as the search subspace. Consider, in addition, a matrix $V_\perp$ whose columns form an orthonormal
basis for the orthogonal complement of $\mathcal{V}$. Then, $(V|V_{\perp})$ is unitary and an arbitrary Hermitian matrix $A$ may be represented with respect to the basis given by the columns of $V$ as

$$
\hat{A} = (V|V_{\perp})^* A (V|V_{\perp}) = \begin{bmatrix} M & R^* \\ R & S \end{bmatrix}.
$$

(1.3)

In Equation (1.3), $M = V^* AV$ is $k \times k$, $R$ is $(n - k) \times k$ and $S = V_{\perp}^* A V_{\perp}$ is $(n - k) \times (n - k)$. Since $A$ is Hermitian so is $M$ and its $k$ eigenvalues,

$$
\mu_1 \leq \mu_2 \leq \ldots \leq \mu_{k-1} \leq \mu_k,
$$

(1.4)

are called the Ritz values of $A$ with respect to $\mathcal{V}$. Any vector $u = Vz$, where $z$ is an eigenvector of $M$ belonging to $u_j$ is called a Ritz vector of $A$ in $\mathcal{V}$, and $(\mu_j, u)$ a Ritz pair. Each Ritz pair can be interpreted as an eigenpair of the $n \times n$ rank-$k$ matrix $VMV^*$ and is seen as an approximation of an eigenpair of $A$. See [47, 63, 71].

An additional remark is in order. If we were to add an extra column, $v$, to $V$, the new matrix $M_{k+1} = V_{k+1}^* A V_{k+1}$ would be

$$
M_{k+1} = \begin{bmatrix} M & V^* Av \\ v^* AV & v^* A v \end{bmatrix}.
$$

(1.5)

The matrix $M$ is a principal submatrix of $M_{k+1}$. When $A$ is Hermitian, both $M$ and $M_{k+1}$ are also Hermitian and from Cauchy Interlace Theorem (Theorem 1.1.10) we know that the eigenvalues of $M$ interlace those of $M_{k+1}$. Moreover, since $A$ and $\hat{A}$ have the same eigenvalues, the eigenvalues of $M$ also interlace the ones of $A$.

Related to Ritz values and vectors is another type of important approximate eigenpairs: harmonic Ritz values and vectors. In the context of iterative methods, and as the number of iterations increases, the Ritz values tend to approximate the eigenvalues located at the extrema of the spectrum of $A$ better than those located in the interior [63, §11]. To approximate interior eigenvalues, Morgan [56] suggested the use of Ritz pairs associated with the inverse of $A$. 
**Definition 1.2.1** (Harmonic Ritz value). The harmonic Ritz values of $A$ with respect to the linear subspace $\mathcal{V}$ are the reciprocals of the Ritz values of $A^{-1}$ with respect to $A\mathcal{V}$.

Paige, Parlett and van der Vorst have (formally) introduced harmonic Ritz values for Hermitian matrices in [60]. In that publication, they make the connection between harmonic Ritz values and Lehmann’s optimal error bounds [50] for the eigenvalues of $A$ and show how, for Krylov subspaces, harmonic Ritz values can be cheaply obtained as by-products of the standard computation.

Other work on harmonic Ritz values and vectors include [73] from Sleijpen and van der Vorst introducing harmonic Ritz values for general matrices, [6] with an extensive study on their relation with Lehmann’s optimal bounds; [17] and [71] looking into a priori and a posteriori error analysis; [33] in the context of non-symmetric linear systems; and [35, 37] with generalizations to different eigenvalue problems.

### 1.2.2 Rayleigh-Ritz procedure

Traditionally, iterative methods are seen as working in two different stages (together forming one *iteration*) which are repeated until a criterion is met. The first includes the process of projecting the matrix in the search subspace and retrieving the Ritz pairs. At this stage, the matrix $S$ from (1.3) is not required. On the other hand, since the product $AV$ was already evaluated to compute $M$, a cheaply available by-product of the process is the matrix $\hat{R} = V_{\perp}R = AV - VM$. The columns $\hat{r}_j$ of $\hat{R}$ are the *residuals* for the respective Ritz pairs $(\mu_j, u_j)$ and satisfy

$$Au_j - \mu_j u_j = \hat{r}_j \perp \mathcal{V}.$$  

At each iteration, one or more Ritz pairs are chosen as approximations for the target eigenpair(s) of the original matrix. The norm of their residuals gives a measure of how close those pairs are to eigenpairs of $A$. Indeed, $(A - \hat{r}_j u_j^*)u_j = \mu_j u_j$ and thus, with $E = -\hat{r}_j u_j^*$, the Bauer-Fike Theorem (Theorem 1.1.1) shows that

$$|\mu_j - \lambda| \leq \|\hat{r}_j\|_2 \quad (1.6)$$
for at least one eigenvalue $\lambda$ of $A$. Note, moreover, that line 11 in Algorithm 1 ensures that a sufficiently close approximation is found in the sense that $|\mu_j - \lambda| \leq \epsilon$ for some $\lambda \in \sigma(A)$ and some $\epsilon \in \mathbb{R}$.

If the Ritz values and vectors selected do not approximate the target eigenpair well enough, it is necessary to expand the search subspace. This leads to the second stage where an attempt is made to obtain a suitable vector with which to expand $\mathcal{V}$. Different approaches for that selection distinguish the different iterative methods. This is indicated in line 14 of Algorithm 1 for general iterative methods for eigenvalue problems. In the coming sections we review three of these expansion strategies each leading to an iterative method for eigenproblems. These will be the Lanczos [49], Jacobi-Davidson [72] and Riccati [9] methods.

**Algorithm 1** General subspace method

1: Input $A$, $v_0$, $\epsilon$;
2: for $q = 1, 2 \ldots$ do
3: \hspace{1em} $v_q = v_{q-1}/\|v_{q-1}\|_2$;
4: \hspace{1em} $w_q = Av_q$;
5: \hspace{1em} for $p = 1, \ldots, q - 1$ do
6: \hspace{2em} $M_{p,q} = v_p^* w_q$; $M_{q,p} = M_{p,q}^*$;
7: \hspace{1em} end for
8: \hspace{1em} $M_{q,q} = v_q^* w_q$;
9: Compute the target eigenpair $(\mu, z)$, with $\|z\|_2 = 1$, of $M$;
10: \hspace{1em} $u = Vz$; \{Ritz vector\}
11: \hspace{1em} $\hat{r} = Au - \mu u$; \{residual\}
12: \hspace{1em} if $\|\hat{r}\|_2 \leq \epsilon$ then
13: \hspace{2em} stop; \{(\mu, u) is the desired approximate eigenpair\}
14: \hspace{1em} else
15: \hspace{2em} Obtain $t$; \{Specific to the different methods\}
16: \hspace{1em} end if
17: \hspace{1em} Orthogonalize $t$ to the columns of $V$
18: end for

### 1.2.3 Lanczos method

Given a Hermitian matrix $A \in \mathbb{C}^{n \times n}$ and an initial search space $\mathcal{V}_1 = \text{span}\{v_1\}$ with $\|v_1\|_2 = 1$, the Lanczos method [49] constructs an orthonormal basis for the
\textit{Krylov subspace}

\[ V_j = \mathcal{K}^j(A, v_1) = \text{span}\{v_1, Av_1, \ldots, A^{j-1}v_1\}. \]

If we orthogonalize the second basis vector \( Av_1 \) with respect to \( v_1 \), the result is
\[ Av_1 - (v_1^* Av_1)v_1 = \hat{r}, \] i.e., the residual associated with the Ritz pair of the current iteration. Thus, assuming that \( \hat{r} \neq 0 \), \( \{v_1, \hat{r}\} \) is an orthogonal basis of \( \mathcal{V}_2 \). For higher dimensions, the expansion is made with the residual, \( \hat{r} \), associated with any of the Ritz pairs of \( A \) in \( \mathcal{V}_j \) and \( \mathcal{V}_{j+1} = \mathcal{V}_j \oplus \text{span}\{\hat{r}\} \). Since all residuals are multiples of one another, this uniquely defines the method.

In each step, a single matrix-vector product with the matrix \( A \) is required and each new expansion vector is normalized and appended to a matrix \( V_j \), whose columns span \( \mathcal{V}_j \). A consequence of the use of orthonormal bases for the Krylov subspaces is that \( M \) is tridiagonal. Formally, we may express this by the \textit{Lanczos relation}

\[ AV_j = V_{j+1}M_{j+1,j}, \quad \text{with} \quad M_{j+1,j} = \begin{bmatrix} M_j & \tau e_j^* \end{bmatrix}. \]

Here, \( M_j \) is a \( j \times j \) tridiagonal matrix, \( e_j \) is the \( j \)th canonical basis vector of \( \mathbb{C}^j \) and \( \tau \) is a scalar.

A method mathematically equivalent to the Lanczos method can be implemented in terms of Algorithm 1 by replacing line 14 with \( t = \hat{r} \). However, due to the special structure of the subspaces it generates, it is possible to obtain a simpler implementation which we present in Algorithm 2.

The simplicity of the Lanczos method makes it a widely studied method. A selection of the most important references as well as a clear and relatively concise discussion of its mathematical properties, algorithmic implementations and variants is given in [3, 32]. The use of Lanczos methods for problems resulting from industrial applications is discussed in [48].

### 1.2.4 Jacobi-Davidson and Riccati methods

An alternative to Krylov subspaces was proposed by Davidson [22] in the method with his name. Rather than expanding the subspace \( \mathcal{V} \) with the residual, Davidson
Algorithm 2  Lanczos method

1: Input: $A, v_0, \epsilon$
2: $\beta_0 = \|v_0\|_2$
3: for $j = 1, 2 \ldots$ do
4: $v_j = v_{j-1}/\beta_{j-1}$
5: $w = Av_j$
6: $w = w - v_{j-1}\beta_{j-1}$
7: $\alpha_j = v_j^*w$
8: $v_j = w - v_j\alpha_j$
9: $\beta_j = \|v_j\|_2$
10: Compute the target eigenpairs of $M_j; \{\text{Ritz pairs}\}$
11: Compute the residual, $\hat{r}$
12: if $\|\hat{r}\|_2 \leq \epsilon$ then
13: stop
14: end if
15: end for

suggested to use the solution of the linear system

$$(D - \lambda I)t = -\hat{r}$$

where $D$ stands for the diagonal of the matrix $A$ to obtain a suitable expansion vector. This idea was improved in the Jacobi-Davidson method by Sleijpen and van der Vorst [72] and in the Riccati method from Brandts [9]. Just like Davidson’s method, Jacobi-Davidson and Riccati methods make use of subspaces without Krylov structure. Unlike in Davidson’s method, however, the new expansion vectors, are sought in the orthogonal complement of the current Ritz vector.

Given a Ritz pair $(\mu, u)$ with residual $\hat{r} \perp V$ and an $n \times (n-1)$ matrix $U_\perp$ such that $(u|U_\perp)$ is unitary, we may transform $A$ as

$$(u|U_\perp)^*A(u|U_\perp) = \begin{bmatrix} \mu & r^* \\ r & S \end{bmatrix}, \tag{1.7}$$

where $r = U_\perp^*\hat{r}$ and $S = U_\perp^*AU_\perp$. Associated with the vector $u \in V$ there is, generally, a unique vector $t \in V_\perp$ such that $u + t$ is an eigenvector of $A$ corresponding to the target eigenvalue. For that reason and the fact that $t \perp u$, the vector $t$ is a so-called orthogonal correction to $u$. Moreover, $t = U_\perp y$ where $y$ can be verified
to satisfy the generalized algebraic Riccati equation

\[(S - \mu I)y = -r + yr^*y,\]

or transformed back to the original basis

\[(I - uu^*)(A - \mu I)(I - uu^*)t = -\hat{r} + t\hat{r}^*t \quad \text{with} \quad t \perp u.\] (1.9)

Solving Equation (1.9), however, is as difficult as solving the original eigenvalue problem. The idea in Jacobi-Davidson is, therefore, to first linearize Equation (1.9) and only then to solve it. The result is Jacobi-Davidson’s correction equation

\[(I - uu^*)(A - \mu I)(I - uu^*)t = -\hat{r} \quad \text{with} \quad t \perp u \quad (1.10)\]

which has a unique solution and is, in practice, only solved approximately. The Riccati method, on the other hand, approximates the solutions of Equation (1.9) by means of a new Rayleigh-Ritz procedure in a smaller, \(\ell\)-dimensional subspace \(X\) of \(U^\perp\). A suitable solution is then chosen as expansion vector for \(V\). The Riccati method can, in this sense, be seen as an enhanced version of Jacobi-Davidson.

Both Jacobi-Davidson and Riccati methods can be implemented using Algorithm 1 with line (14) replaced by

**Solve (approximately)** \((I - uu^*)(A - \mu I)(I - uu^*)t = -\hat{r} \quad \text{for} \quad t \perp u;\)

in the Jacobi-Davidson method and with

**Solve (approximately)** \((I - uu^*)(A - \mu I)(I - uu^*)t = -\hat{r} + t\hat{r}^*t \quad \text{for} \quad t \perp u;\)

**Choose the minimum norm solution;**

in the Riccati method.

Both methods can be adapted to exploit extra available information such as, for instance, approximations for the matrix \(A - \mu I\). This may be seen as a form of preconditioning. This is only useful, however, if the approximation reduces the computational costs while preserving or even improving the rate of convergence towards the solution. Therefore, also for eigenvalue problems, determining good preconditioners is, in the words of Saad [68], “a combination of art and science”.

20 1.2. ITERATIVE METHODS FOR EIGENVALUE PROBLEMS

Recent contributions

A study of inner-outer iteration methods in the context of eigenvalues problems is given in [20]. Jacobi-Davidson and Riccati are compared with more detail in [10] and the analysis is complemented with some numerical experiments highlighting the differences between the two. The use of alternative correction equations for Jacobi-Davidson was studied in [28] and included, in particular, a projection onto the orthogonal complement of the entire current search subspace rather than only the current Ritz vector. Other variants of Jacobi-Davidson are suggested, for instance, in [36] for non-normal matrices and [70] for different types of eigenproblems.

To conclude this introduction, and to improve the overall readability of the remainder of this thesis, we now provide an outline of the different chapter contents and a few notational remarks.

Outline

Hermicity preserving perturbations are a standard choice in the study of eigenvalues bounds and we shall investigate them in Chapter 2. Starting from the study of a simple Hermitian rank-one perturbation $E$ of a Hermitian matrix $A$, we construct a sequence of increasingly better computable bounds for the eigenvalues of $A + E$. The advantage of this approach is that the accuracy of the bounds is dependent on the available eigendata of $A$ and, therefore, is potentially better than that of the classical solutions and the recent work in [46]. Moreover, the techniques suggested extend easily to other eigenvalue and singular value perturbation problems. The contents of this chapter, with the exception of Section 2.6, have been published as [11]:


Unlike the preservation of Hermicity, preserving normality is, in every sense, a complex problem. Chapter 3 makes headways on normality preserving pertur-
bations and augmentations of normal matrices. The latter situation has been addressed earlier by Ikramov and Elsner [45]. However, by recognizing the role played by the eigendata of the original matrix, we not only clarify their solution but also bound the eigenvalues of the augmented matrix. We also study normality preserving perturbations. For $2 \times 2$ and for rank-one matrices the analysis is now complete while for higher rank all essentially Hermitian normality perturbations are described. With this new information and using the stratification of normal matrices suggested by Huhtanen [42] in 2001 we construct all augmentations that result in normal matrices with eigenvalues on a quadratic curve in the complex plane. The contents of this chapter may be found in [67]:


With Chapter 4 the focus of the thesis shifts from perturbation theory to iterative methods for eigenvalue problems. The chapter is devoted to the analysis of the fairly recent and unobserved Subspace Projected Approximate Matrix (SPAM) [69] method. We analyze it from a numerical linear algebra perspective with particular interest for the choice of the approximating matrix. Some property preserving perturbations introduced in Chapter 2 will reappear here as desirable preconditioners not only theoretically but also numerically. At each point we illustrate the performance of SPAM in comparison with well-known methods such as Lanczos [49] or Jacobi-Davidson [72]. Sections 4.1 to 4.5 of this chapter can be found in [12]:


while an extended version of Section 4.6 will be found in [13]

Notation

Although most of the notation used is considered standard, it is useful to clarify a few assumptions. A more complete list of notational symbols and their meanings can be found on page 139.

When appropriate, we choose to label eigenvalues, but also singular values, in non-decreasing order of magnitude

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{n-1} \leq \lambda_n.$$ 

The spectrum of a matrix $A$ is denoted by $\sigma(A)$. Letters at the end of the alphabet represent vectors and these are always column vectors. Lower case Latin letters $j,k$ and occasionally $p$ and $q$ are mostly used as indices. In this thesis, the letter $i$ is used solely to indicate the imaginary unit.

The conjugate transpose of a matrix $A$ is denoted by $A^*$, also for $A$ real. With $I_n$ we mean the $n \times n$ identity matrix and by $e_j$ its $j$th column. Similarly, $O_n$ stands for the zero matrix of size $n \times n$ while $\vec{0}$ stands for any of its columns. The commutator of $A$ and $B$ is denoted by $[A,B]$ and defined as $[A,B] = AB - BA$. Finally, the Hermitian and skew-Hermitian part of a matrix $A$ are represented by $\mathcal{H}(A)$ and $\mathcal{S}(A)$ and defined as $\mathcal{H}(A) = (1/2)(A + A^*)$ and $\mathcal{S}(A) = (1/2)(A - A^*)$, respectively.
Part I

Perturbation theory
Chapter 2

Hermitian matrices

When $A$ is Hermitian and $E$ is Hermitian positive semi-definite, Weyl’s inequalities (Theorem 1.1.7) show that the eigenvalues of $A$ are lower bounds for the eigenvalues of $A + E$. For a long time these were the best available bounds. More recently, however, a result by Ipsen and Nadler [46] improves Weyl’s inequalities. Both (Weyl’s and the result from [46]) offer cheaply computable lower bounds for the eigenvalues of perturbed Hermitian matrices.

In this chapter we improve the bounds from Ipsen and Nadler [46] for the eigenvalues of $A + E$. The starting point is a Hermitian, positive semi-definite, rank-one perturbation, $E$, of a Hermitian matrix, $A$. The bounds proposed rely on the construction of a sequence of matrices approximating $A$ in a certain sense. Each element of this sequence determines an element of a sequence of lower bounds for each eigenvalue of $A + E$. The number of available eigenvalues and eigenvectors of the original matrix determines how close those approximations can be. If the perturbation is of low rank, the new bounds are relatively inexpensive to compute. For the smallest eigenvalue, the first two elements of the sequence correspond to Weyl’s and Ipsen and Nadler’s bounds. For the other eigenvalues, however, our approach is different from that of [46]. We then use the same procedures to bound eigenvalues of bordered diagonal matrices as well as singular values of rank-$k$ perturbations and other modifications of $n \times m$ matrices.
2.1 Introduction

The need to bound the eigenvalues of perturbed matrices arises, for instance, in problems related to the discretization of differential equations. Consider the following reaction-diffusion equation

\[- u'' + su = \lambda u \quad \text{in } I = [0, 1], \quad \text{with } u(0) = u(1) = 0. \quad (2.1)\]

In Equation (2.1), let the function \( s : [0, 1] \rightarrow \mathbb{R} \) be a continuous nonnegative local perturbation, which is zero everywhere except in the interval \( (\frac{1}{2} - \xi, \frac{1}{2} + \xi) \) for some \( \xi \in [0, \frac{1}{2}] \) where it is defined by

\[s(x) = e^{-g(x)} > 0 \quad \text{with} \quad g(x) = \frac{\xi^2}{(x - (1/2 - \xi))(x - (1/2 + \xi))}.\]

Take \( n \in \mathbb{N} \) and define the mesh width \( h = 1/n \) and the grid points \( x_q = qh \) for \( q \in \{0, \ldots, n\} \). The Finite Difference discretization of Problem (2.1) on the grid points \( x_0, \ldots, x_n \) yields a positive definite Hermitian tridiagonal matrix \( A \), resulting from the diffusion part and a Hermitian diagonal matrix \( E \), resulting from \( s \), whose rank is dependent on \( h \) and the variable \( \xi \).

The problem

In this chapter we are interested in lower bounds for the eigenvalues of discrete problems like those of Equation (2.1), i.e., lower bounds for the eigenvalues of \( A + E \). We may state it as follows.

Let \( A, E \in \mathbb{C}^{n \times n} \) be Hermitian matrices with the additional property that \( E \) is of rank-\( k \). Assuming that the eigendata of the matrices \( A \) and \( E \) are known, determine (lower) bounds for the eigenvalues of \( A + E \).

Because, at least in theory, we can always apply a unitary similarity transformation that diagonalizes \( A \) to both \( A \) and \( E \) we may assume, without loss of generality, that \( A \) is diagonal. The general case will be discussed later, in Section 2.2.3.
2.1. INTRODUCTION

For $q \in \{1, \ldots, n\}$ consider the $n \times n$ matrix

$$X_q = \begin{bmatrix}
\Lambda_q \\
\lambda_q I_{n-q}
\end{bmatrix}$$

where

$$\Lambda_q = \begin{bmatrix}
\lambda_1 \\
\ddots \\
\lambda_q
\end{bmatrix},$$

and $\lambda_q$ stands for the $q$th eigenvalue (in non-decreasing order) of $A$. Observe that when $q = 1$ and $q = n$ we have $X_1 = \lambda_1 I_n$ and $X_n = A$, respectively. Moreover,

$$x^*(X_p - X_q)x \geq 0, \quad \text{for all } x \in \mathbb{C}^n, \quad \text{and } p, q \in \{1, \ldots, n\} \quad \text{with } p \geq q$$

and we say that $X_q$ is an approximation from below to $X_p$. In particular, each $X_q$ approximates $A$ from below. Consequently, it follows from Weyl’s theorem (Theorem 1.1.7) that for $j \in \{1, \ldots, n\}$

$$\lambda_j(X_1 + E) \leq \lambda_j(X_2 + E) \leq \ldots \leq \lambda_j(X_{n-1} + E) \leq \lambda_j(A + E).$$

(2.3)

Remark 2.1.1. We can obtain upper bounds for the eigenvalues by using approximations $X_q$ from above, i.e, such that $X_q - A$ is positive semi-definite, or by applying the theory that follows to $-A$ instead of $A$.

The following example illustrates, for two small matrices $A$ and $E$, what we have just discussed.

2.1.1 Example

Let $w := [1, 1, 1, 1]^*$ and consider the symmetric rank-one matrix $E := ww^*$ and the real diagonal matrix

$$A = \begin{bmatrix}
2 & 3 \\
3.1 & 4 \\
4.1
\end{bmatrix} = X_5$$
as well as

\[
X_1 = \begin{bmatrix}
2 & 2 & 2 \\
2 & 2 & 2 \\
2 & 2 & 2
\end{bmatrix},
\quad X_2 = \begin{bmatrix}
2 & 3 & 3 \\
3 & 3 & 3 \\
3 & 3 & 3
\end{bmatrix},
\quad X_3 = \begin{bmatrix}
2 & 3 & 3.1 \\
3 & 3.1 & 3.1 \\
3.1 & 3.1 & 3.1
\end{bmatrix}
\quad \text{and} \quad X_4 = \begin{bmatrix}
2 & 3 & 3.1 \\
3 & 3.1 & 4 \\
3.1 & 4 & 4
\end{bmatrix}.
\]

The eigenvalues of \(X_q + E\) for each \(q \in \{1, \ldots, 5\}\) are displayed in Table 2.1.

<table>
<thead>
<tr>
<th>(X_q + E)</th>
<th>(\lambda_1)</th>
<th>(\lambda_2)</th>
<th>(\lambda_3)</th>
<th>(\lambda_4)</th>
<th>(\lambda_5)</th>
<th>(|X_q - X_{q-1}|)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1 + E)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>(X_2 + E)</td>
<td>2.1716</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>7.8284</td>
<td>1</td>
</tr>
<tr>
<td>(X_3 + E)</td>
<td>2.1821</td>
<td>3.0250</td>
<td>3.1</td>
<td>3.1</td>
<td>7.8929</td>
<td>0.1</td>
</tr>
<tr>
<td>(X_4 + E)</td>
<td>2.2201</td>
<td>3.0473</td>
<td>3.5089</td>
<td>4</td>
<td>8.3236</td>
<td>0.9</td>
</tr>
<tr>
<td>(A + E)</td>
<td>2.2214</td>
<td>3.0475</td>
<td>3.5291</td>
<td>4.0519</td>
<td>8.3503</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 2.1: Eigenvalues of \(X_q + E\) for \(q \in \{1, \ldots, 5\}\)

Observe that each \(X_q\) is defined using the \(q\) smallest eigenvalues of \(A\) and that as \(q\) increases, all eigenvalues of \(X_q + E\) move, indeed, closer to those of \(A + E\) as showed by (2.3). Unfortunately, the effort involved in the computation of the eigenvalues of \(X_q + E\) seems the same as that involved in the computation of those of \(A + E\). If that obstacle could be removed, however, then the eigenvalues of \(X_q + E\) could be used also in practice as lower bounds for the eigenvalues of \(A + E\).

Remark 2.1.2. In the rank-one situation and for given \(j \in \{1, \ldots, n - 1\}\), only if \(q > j\) is the bound for \(\lambda_j\) due to \(X_q\) an improvement over Weyl’s bound.
2.2 Hermitian rank-$k$ perturbations

Example 2.1.1 shows that increasingly accurate lower bounds for the eigenvalues of $A + E$ are easily defined. The error matrix resulting from approximating $A + E$ by $X_q + E$, is

$$S_q = A + E - (X_q + E) = \begin{bmatrix} \Lambda_q & \\ \Lambda_{n-q} & \lambda_q \end{bmatrix} - \begin{bmatrix} \Lambda_q & \\ \lambda_q I_{n-q} & \end{bmatrix} = \begin{bmatrix} O_q & \\ \mathbf{L}_q & \end{bmatrix}$$

where $L_q = \Lambda_{n-q} - \lambda_q I_{n-q}$. More specifically, when considering two succeeding matrices $X_q$ and $X_{q+1}$ we obtain

$$\|S_{q+1}\|_2 = \|S_q\|_2 - (\lambda_{q+1} - \lambda_q).$$

Equation (2.4) estimates the difference expected from using each next element of the sequence. In other words, if $A$ has a multiple eigenvalue, $\lambda_q$, of multiplicity $m$, then all the matrices between $X_q$ and $X_{q+m}$ are equal and, thus, have the same eigenvalues. In this case, the bounds do not change. On the other hand, if the gap between two consecutive eigenvalues is large, we may hope to do better by using the next matrix.

Although the eigenvalues of $X_q + E$ provide lower bounds for the eigenvalues of $A + E$, the reader might be inquiring the usefulness of Example 2.1.1. For any $q \in \{1, \ldots, n\}$, the dimension of the matrix $X_q + E$ is the same as that of $A + E$ and the computation of the eigenvalues of $X_q + E$ seem as costly as the computation of the eigenvalues of $A + E$. Fortunately, as Proposition 2.2.1 will show, that need not be the case.

Proposition 2.2.1 focuses on the perturbation of multiple eigenvalues and, in short, states that a Hermitian rank-one perturbation of a Hermitian matrix, $X$ affects at most single copies of multiple eigenvalues. The remaining copies are not perturbed. The reasoning is that if $X$ has an eigenvalue $\lambda$ of multiplicity $m$ we can find $m - 1$ eigenvectors associated with $\lambda$ orthogonal to the perturbation (eigen)vector $w$. Proposition 2.2.1 suggests, thus, that the computation effort should be directed to those eigenvalues of $X_q + E$ that are not eigenvalues of $X_q$. 
Proposition 2.2.1. Let $\sigma \in \mathbb{R}$, $w \in \mathbb{C}^n$ with $\|w\|_2 = 1$ and define $E := w\sigma w^*$. Let $X \in \mathbb{R}^{n \times n}$ be diagonal with exactly $t$ distinct eigenvalues denoted by $\lambda_p$ ($1 \leq p \leq t$). Write $P_p$ for the orthogonal projection onto the eigenspace $X_p$ corresponding to $\lambda_p$ and define $\theta_p$ by

$$\|P_p w\|_2 = \cos \theta_p \quad \text{where} \quad \theta_p = \angle (w, X_p) \quad \text{is chosen in} \quad \left[-\frac{\pi}{2}, \frac{\pi}{2}\right].$$

Define yet the $t \times t$ matrix

$$Y = \begin{bmatrix} \lambda_1 & \vdots & \lambda_t \end{bmatrix} + y\sigma y^*, \quad \text{where} \quad y = \begin{bmatrix} \cos \theta_1 \\ \vdots \\ \cos \theta_t \end{bmatrix}. $$

Then the eigenvalues of $X + E$ are the eigenvalues of $X$, of which single copies of multiple eigenvalues are replaced by the eigenvalues of $Y$.

Proof. We give the details of the proof for $t = 2$ after which the proof for $t > 2$ should be clear. Let

$$X = \begin{bmatrix} \lambda_1 I_{m_1} \\ \lambda_2 I_{m_2} \end{bmatrix}$$

where $m_1 > 1$ is the multiplicity of $\lambda_1$, $m_2 > 1$ is the multiplicity of $\lambda_2$ and $m_1 + m_2 = n$. For unitary matrices $Q_1 \in \mathbb{C}^{m_1 \times m_1}$ and $Q_2 \in \mathbb{C}^{m_2 \times m_2}$ define

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \quad \text{and thus,} \quad Q^* X Q = X.$$ 

Now, since each $P_p w \in X_p$ we can choose each $Q_p$ such that $Q^* w$ has at most two nonzero entries:

$$Q^* w = \cos \theta_1 e_1 + \cos \theta_2 e_{m_1+1}. $$

Thus,

$$Q^* (X + E) Q = X + Q^* w \sigma w^* Q.$$ 

Note that the second term of the right-hand side has nonzero entries only at two rows and corresponding two columns. By appropriate permutations, those columns
and rows can be placed at the top left. Both \( \lambda_1 \) and \( \lambda_2 \) continue to be eigenvalues of \( X+E \) now with multiplicity \( m_1 - 1 \) and \( m_2 - 1 \) respectively. As for the remaining two eigenvalues they have been replaced by those of \( Y \):

\[
Y = \begin{bmatrix} \lambda_1 & \cos \theta_1 \\ \lambda_2 & \cos \theta_2 \end{bmatrix} + \sigma [\cos \theta_1, \cos \theta_2].
\]

This proves the statement for \( t = 2 \). For \( t > 2 \) the proof is similar.

If the rank of \( E \) is larger than one, a similar situation occurs. This is expressed in Corollary 2.2.2.

**Corollary 2.2.2.** If, in the conditions of Proposition 2.2.1, the matrix \( E \) is Hermitian and of rank-\( k \) (with \( 1 \leq k \leq n \)), then at most \( k \) copies of each distinct eigenvalue of \( X \) are not eigenvalues of \( X+E \). Therefore, at most \( kt \) of the total number of eigenvalues of \( X \) are perturbed.

We conclude from Example 2.1.1 and from the two previous results that only the eigenvalues of \( X_q + E \) that are not eigenvalues of \( X_q \) need to be computed. It remains unclear, however, how that can be achieved. We address this problem in the next section.

### 2.2.1 Computing the bounds

Let \( W \in \mathbb{C}^{n \times k} \) and \( \Sigma \in \mathbb{R}^{k \times k} \) be such that

\[
E = W \Sigma W^* \tag{2.5}
\]

is a rank-\( k \) Hermitian matrix. Consider, now, the following procedure.
Procedure 2.2.3. For any \( j \in \{0, \ldots, n - k\} \) partition the \( n \times k \) matrix \( W \) as \( W_j \), consisting of the first \( j \) rows and \( W_{n-j} \), consisting of the last \( n - j \) rows, i.e.,

\[
W = \begin{bmatrix} W_j \\ W_{n-j} \end{bmatrix}.
\]

Therefore, if \( Q \in \mathbb{C}^{(n-j) \times (n-j)} \) and \( R \in \mathbb{C}^{k \times k} \) are such that

\[
Q \begin{bmatrix} R \\ O \end{bmatrix} = W_{n-j}
\]

is a QR-decomposition of \( W_{n-j} \), we have for the unitary \( G = I_j \oplus Q \) that

\[
G^* E G = \begin{bmatrix} E_{j+k} \\ O \end{bmatrix}
\]

where \( E_{j+k} = \begin{bmatrix} W_j \\ R \end{bmatrix} \Sigma \begin{bmatrix} W_j^* & R^* \end{bmatrix} \)

is a \((j + k) \times (j + k)\) matrix.

The goal of this procedure is to introduce a large zero trailing principal submatrix using a unitary transformation \( G \) that leaves \( \Lambda_q \) invariant, i.e., such that \( G^* \Lambda_q G = \Lambda_q \). We can now use Procedure 2.2.3 to prove the following result.

Theorem 2.2.4. Let \( E \in \mathbb{C}^{n \times n} \) be the Hermitian rank-\( k \) matrix defined in (2.5) and for given \( q \in \{1, \ldots, n\} \), set \( r = q + k - 1 \) if \( q \leq n - k \) and \( r = n \) otherwise. Let \( E_r \in \mathbb{C}^{r \times r} \) be the matrix obtained from \( E \) by Procedure 2.2.3 with \( j = r - k \). Let, moreover, \( X_q \) be the matrix from (2.2) partitioned as

\[
X_q = \begin{bmatrix} Y_q \\ \lambda_q I_{n-r} \end{bmatrix}
\]

where \( Y_q = \begin{bmatrix} \Lambda_q \\ \lambda_q I_{r-q} \end{bmatrix} \)

is an \( r \times r \) matrix. Then,

\[
\sigma(X_q + E) = \{\sigma(Y_q + E_r), \lambda_q\}.
\]

Proof. The particular case of the smallest eigenvalue when \( k = 1 \) was treated in [46]. We look now into the general proof for all eigenvalues and for general \( k \).
Note that $Y_q$ is of size $r \times r$ containing the $r$ smallest eigenvalues of $X_q$ (counting multiplicities). By partitioning $W$ appropriately with $j = r - k$ according to Procedure 2.2.3 we get a unitary matrix $G$ such that

$$G^* E G = \begin{bmatrix} E_r & \cr & O_{n-r} \end{bmatrix}. \tag{2.8}$$

However, Procedure 2.2.3 also ensures that $G^* X_q G = X_q$ and, therefore,

$$G^* (X_q + E) G = \begin{bmatrix} Y_q + E_r & \cr & \lambda_q I_{n-r} \end{bmatrix},$$

proving the statement.

\[\square\]

**Remark 2.2.5.** The matrix $R$ in (2.6) is of size $k \times k$. Consequently, if $q > n - k$ the complete matrix $X_q$ is required and, thus, $r = n$.

**Remark 2.2.6.** If any eigenvalue of $A$ has multiplicity larger than $k$ then by Proposition 2.2.1 the bounds are sharp.

We now provide an example detailing the computation of the lower bounds for a rank-2 perturbation of the matrix $A$ from Example 2.1.1.

### 2.2.2 Example

Let $A \in \mathbb{C}^{5 \times 5}$ and $X_q \in \mathbb{C}^{5 \times 5}$ for $q \in \{1, \ldots, 5\}$ be the matrices from Example 2.1.1 and for $\Sigma = I_2$ and $W = [w|v] \in \mathbb{C}^{5 \times 2}$ let $E = W \Sigma W^*$ be a Hermitian matrix of rank-2. For this example we take the vectors $w = \left[ \begin{array}{cc} 1 & 1 & 1 & 1 & 1 \end{array} \right]^*$ and $v = \left[ \begin{array}{cc} 1 & -1 & 1 & -1 & 1 \end{array} \right]^*$. The $r \times r$ matrices $Y_q$ from (2.7) are displayed next

\[
Y_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad Y_2 = \begin{bmatrix} 2 & 3 \\ 3 & \end{bmatrix}, \quad Y_3 = \begin{bmatrix} 2 & & \\ & 3 & & \\ & & 3.1 & \end{bmatrix},
\]
We shall compute bounds for the eigenvalues of $A + E$ using only $Y_2$ as the procedure for the remaining matrices is similar. In this case, $r = 3$ and following Procedure (2.2.3), we partition $W$ as

$$W = \begin{bmatrix} w_1 & v_1 \\ \tilde{w} & \tilde{v} \end{bmatrix}$$

and compute a QR-decomposition of $\begin{bmatrix} \tilde{w} & \tilde{v} \end{bmatrix}$

$$\begin{bmatrix} \tilde{w} & \tilde{v} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix}.$$

In this example, $\tilde{w}$ and $\tilde{v}$ have length 4 and are already orthogonal. Therefore, with the unitary matrix $G = 1 \oplus Q$ we obtain

$$G^*EG = \begin{bmatrix} 2 & 2 & 2 \\ 2 & 4 & 0 \\ 2 & 0 & 4 \end{bmatrix} = \begin{bmatrix} E_3 \\ O_2 \end{bmatrix}.$$

Finally, computing the eigenvalues of $Y_2 + E_3$ gives us lower bounds for the eigenvalues of $A + E$. In Table 2.2 we display, for $q \in \{1, \ldots, 5\}$ the eigenvalue of $X_q + E$ from Example 2.2.2 computed using the approach of Theorem 2.2.2.
2.2. HERMITIAN RANK-$K$ PERTURBATIONS

Table 2.2: Eigenvalues of $X_q + E$ for $q \in \{1, \ldots, 5\}$ from Example 2.2.2 computed using the approach of Theorem 2.2.2.

<table>
<thead>
<tr>
<th>$X_q + E$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>$\lambda_4$</th>
<th>$\lambda_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 + E$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>$X_2 + E$</td>
<td>2.2984</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>8.7016</td>
</tr>
<tr>
<td>$X_3 + E$</td>
<td>2.3247</td>
<td>3.0494</td>
<td>3.1</td>
<td>7.0506</td>
<td>8.7753</td>
</tr>
<tr>
<td>$X_4 + E$</td>
<td>2.3939</td>
<td>3.4384</td>
<td>3.5641</td>
<td>7.5616</td>
<td>9.1420</td>
</tr>
<tr>
<td>$A + E$</td>
<td>2.3981</td>
<td>3.4384</td>
<td>3.6149</td>
<td>7.5616</td>
<td>9.1870</td>
</tr>
</tbody>
</table>

2.2.3 Perturbations of non-diagonal matrices

For the problem we are considering, we could, without loss of generality, study diagonal matrices instead of arbitrary Hermitian matrices. In practice, however, since diagonalizations of large matrices are expensive it is wise to avoid them whenever possible. The bounds developed earlier can be computed without an explicit diagonalization. In fact, because for these perturbation, the knowledge of a few eigenvalues and eigenvectors suffices, the computational costs to determine the bounds are low.

Let $A$ be Hermitian with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$ and corresponding orthonormal eigenvectors $v_1, \ldots, v_n$. For given $j \in \{0, \ldots, n\}$, write $V_j$ for the matrix with columns $v_1, \ldots, v_j$ and $V_{n-j}$ for the matrix with columns $v_{j+1}, \ldots, v_n$ and let $V = (V_j V_{n-j})$. Now, for each matrix $X_q$ from (2.2), $V^* AV - X_q$ is positive semidefinite. Thus, $T_q = V X_q V^*$ is an approximation of $A$ from below, and for any Hermitian perturbation $E$ of $A$, the eigenvalues of $T_q + E$ are lower bounds for the corresponding eigenvalues of $A + E$. Of course, the eigenvalues of $T_q + E$ are the eigenvalues of $X_q + V^* EV$. Thus, with $E = W \Sigma W^*$ where $W$ is an $n \times k$ matrix with orthonormal columns, the problem that results is a rank-$k$ perturbation of a diagonal matrix.

Assume now that only the eigenpairs belonging to the eigenvalues $\lambda_1, \ldots, \lambda_\ell$ with $1 \leq \ell < n$ are available. Then $V^* EV$ cannot be explicitly computed. But because $W$ has orthonormal columns, so has $V^* W$, and for each $j \in \{0, \ldots, n\}$ we find that

$$I = (V^* W)^* (V^* W) = W^* V_j V_j^* W + W^* V_{n-j} V_{n-j}^* W.$$ 

Note that $Y^* Y = Z^* Z$ for matrices $Y$ and $Z$ of the same size, if and only if
$Y = QZ$ for some unitary transformation $Q$. Therefore, let $Y$ be any solution of the $k \times k$ matrix equation

$$Y^*Y = I - W^*V_j V_j^*W$$

for instance, the $k \times k$ Cholesky factor of the right-hand side, padded with $n - j - k$ zero rows, then $Y = Q_{n-j} V_{n-j}^*W$ for some unitary $(n - j) \times (n - j)$ matrix $Q_{n-j}$. This yields

$$V^*EV = V^*W\Sigma W^*V = \begin{bmatrix} V_j^*W \\ V_{n-j}^*W \end{bmatrix} \Sigma \begin{bmatrix} W^*V_j & W^*V_{n-j} \end{bmatrix}$$

$$= \begin{bmatrix} I_j & Q_{n-j} \end{bmatrix}^* \begin{bmatrix} V_j^*W \\ Y \end{bmatrix} \Sigma \begin{bmatrix} W^*V_j & Y^* \end{bmatrix} \begin{bmatrix} I_j & Q_{n-j} \end{bmatrix}.$$  

If $j \leq \ell$, then the only unknown in the above expression is $Q_{n-j}$. But if $q \leq j$ then the eigenvalues of $X_q + V^*EV$ can still be computed, since $X_q$ commutes with the $n \times n$ unitary transformation, i.e., $X_q + V^*EV$ equals

$$\begin{bmatrix} I_j & Q_{n-j} \end{bmatrix}^* \begin{bmatrix} X_q + \begin{bmatrix} V_j^*W \\ Y \end{bmatrix} \Sigma \begin{bmatrix} W^*V_j & Y^* \end{bmatrix} \begin{bmatrix} I_j & Q_{n-j} \end{bmatrix}.$$  

What remains is a rank-$k$ perturbation of a diagonal matrix with known perturbation, which can be treated as in Theorem 2.2.4. In fact, if $Y$ is indeed upper triangular due to the choice of the Cholesky factor in and below (2.9), it is clear that the eigenvalues of $X_q + V^*EV$ that differ from the ones from $X_q$ are the eigenvalues of the leading principal submatrix of $X_q + V^*EV$ of size $(j + k) \times (j + k)$.

### 2.2.4 Alternative approximations

The matrices chosen in (2.2) are not the only approximations obtained from the eigenvalues of $A$ that approximate $A$ from below. In fact, we may construct other approximations satisfying the same property and, with that, still obtain lower bounds for the same eigenvalue of $A + E$.

In Example 2.1.1 improving the bounds was implicitly associated with a)
Knowing a larger number of eigenvalues of $A$; and b) Increasing the size of the small eigenvalue problem. Obtaining improved bounds without the knowledge of more eigenvalues and without increasing the size of the small eigenvalue problem may prove difficult. However, the construction of matrices leading to bounds of higher quality seems achievable without increasing $r$.

We turn to Example 2.1.1 and consider the following two matrices

$$X_3 := \begin{bmatrix} 2 & \ 3 \\ & 3.1 I_3 \end{bmatrix} \quad \text{and} \quad X'_3 := \begin{bmatrix} 2 & \ 3 I_2 \\ & 4 I_2 \end{bmatrix} \quad (2.10)$$

together with Table 2.3 containing the respective eigenvalues after perturbation with $E$.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>$\lambda_4$</th>
<th>$\lambda_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_3 + E$</td>
<td>2.1821</td>
<td>3.0250</td>
<td>3.1</td>
<td>3.1</td>
<td>7.8929</td>
</tr>
<tr>
<td>$X'_3 + E$</td>
<td>2.2143</td>
<td>3.4798</td>
<td>4.0519</td>
<td>8.3059</td>
<td></td>
</tr>
<tr>
<td>$A + E$</td>
<td>2.2214</td>
<td>3.0475</td>
<td>3.5291</td>
<td>4.0519</td>
<td>8.3503</td>
</tr>
</tbody>
</table>

Table 2.3: Different lower bounds for the eigenvalues of $A + E$ given by the eigenvalues of $X_3 + E$ and $X'_3 + E$

For most eigenvalues, $X'_3$ offers tighter bounds than $X_3$ while still maintaining the size of the small eigenvalue problem unchanged. Note, moreover, that the error matrices satisfy $0.1 = \|S'\|_2 \leq \|S\|_2 = 1$ where $S'_3 = A - X'_3$ and $S_3 = A - X_3$. The idea here is to deviate from the use of multiple eigenvalues located at the bottom right positions of the matrix to the use of multiple eigenvalues located at any positions. Even if in the interior of the spectrum. The existence of eigenvalues with high multiplicity can be exploited for that effect.

In a way similar to the one used in Procedure 2.2.3, also here the matrix $E$ is transformed via a unitary matrix $G$ obtained from a direct sum of unitary submatrices. In turn, these are obtained from QR-decompositions of appropriate partitions of $W$. For instance, with the matrices from (2.10) and assuming that $E$ is of rank-one and that the focus is on the smallest eigenvalue of $A + E$, we may
partition $w$ as

$$w = \begin{bmatrix} w_1 \\ \vec{w}_2 \\ \vec{w}_3 \end{bmatrix}$$

and use the unitary matrix

$$G = \begin{bmatrix} 1 \\ Q_2 \\ Q_3 \end{bmatrix}$$

where $Q_2$ and $Q_3$ are such that $Q_2^* \vec{w}_2 = \| \vec{w}_2 \|_2$ and $Q_3^* \vec{w}_3 = \| \vec{w}_3 \|_2$ to obtain $G^* E G$. Note that $G^* E G$ is of the form

$$\begin{bmatrix} \ast & \ast & \ast \\ \ast & \ast & \ast \\ \ast & \ast & \ast \end{bmatrix}$$

and thus additional permutations of rows and columns are needed to obtain $E_r$. It remains valid, however, that

$$\sigma(Y'_q + E_r) \subset \sigma(X'_q + E').$$

Depending on the eigenvalue for which an improved bound is required, the new matrices should satisfy two conditions: a) $\| S'_q \|_2 \leq \| S_q \|_2$; and b) if a bound for $\lambda_q(A + E)$ is to be determined, then $\lambda_q(X')$ should not belong to a cluster with other (different) eigenvalues. If these conditions are satisfied, both $X'_q$ and $X_q$ approximate $A$ from below. Moreover, under the assumption that all eigenvalues of $A$ are available, the computational effort required to determine the bounds is virtually the same.

**Remark 2.2.7.** Automatic clustering techniques can provide an optimal division of the spectrum. The computational effort required to execute such procedures is,
however, too high and would eliminate all advantage of not computing the exact eigenvalue. Our approach is intended to exploit available information.

2.3 Bordered diagonal matrices

We start this section by introducing the concept of Hermitian $k$-bordered diagonal matrices:

**Definition 2.3.1** (Hermitian $k$-bordered diagonal matrix). A Hermitian matrix is $k$-bordered diagonal if its entries are zero except for the main diagonal and $k$ rows and corresponding columns.

Formally, we pose no restriction on $k$ and, in particular, allow a diagonal matrix and a full matrix to be the 0 and $n-1$-bordered diagonal matrices respectively. Moreover, owing to the invariance of the eigenvalues under similarity transformations, we opt to rearrange the matrix so that the nonzero rows and columns are the first $k$. We are thus interested in matrices of the form

$$
\begin{bmatrix}
C_1 & C_2 & \ldots & C_{n-k+1} \\
C_2^* & d_1 \\
\vdots & \ddots \\
C_{n-k+1}^* & \quad & \quad & \quad & d_{n-k}
\end{bmatrix}
$$

(2.11)

where $C_1 \in \mathbb{C}^{k \times k}$ and for $j \in \{2, \ldots, n-1\}$, $C_j \in \mathbb{C}^k$ and where we assume that $d_1 \leq d_2 \leq \ldots \leq d_{n-k}$. For such matrices, the following result, showing a relation between Hermitian rank-$k$ perturbations of Hermitian matrices and $k$-bordered diagonal matrices, is known (see [83]) but we have not encountered it further in the literature dealing with bordered diagonal matrices.

**Lemma 2.3.2** ([83]). Let $n$ be an integer and let $A, E \in \mathbb{C}^{n \times n}$ be Hermitian matrices and assume further that $E$ is of rank-$k$. Then $A + E$ is unitarily similar to a Hermitian $k$-bordered diagonal matrix.
Proof. Let $U$ be unitary and such that

$$ EU = U \begin{bmatrix} \Theta \\ O_{n-k} \end{bmatrix} $$

where $\Theta \in \mathbb{R}^{k \times k}$ is diagonal and contains the $k$ nonzero eigenvalues of $E$. We may apply the same unitary similarity transformation to $A$, yielding

$$ AU = U \begin{bmatrix} M & B^* \\ B & T \end{bmatrix} $$

for certain $B \in \mathbb{C}^{(n-k) \times k}$ and for Hermitian $M \in \mathbb{C}^{k \times k}$ and $T \in \mathbb{C}^{(n-k) \times (n-k)}$. Now, let $V$ be unitary with $TV = VD$ where $D$ is the diagonal matrix of eigenvalues of $T$. Then, the unitary matrix

$$ S = UG $$

where

$$ G = \begin{bmatrix} I_k \\ V \end{bmatrix} $$

transforms $A + E$ into

$$ S^*(A + E)S = \begin{bmatrix} M + \Theta & B^*V \\ V^*B & D \end{bmatrix}. \quad (2.12) $$

which is $k$-bordered diagonal.

Lemma 2.3.2 is of interest, not only on its own, but also because for $k = 1$, comparing the characteristic polynomial of $S^*(A + E)S$ with that of $A$, provides an alternative (and perhaps more widely known) proof to Proposition 2.2.1 (see [63] or [83]). The connections between rank-$k$ perturbations and bordered diagonal matrices, however, do not end here. The structure of bordered diagonal matrices allows for the computation of lower bounds for its eigenvalues using the principles discussed in Section 2.2.

Remark 2.3.3. From a perturbation point of view one could also treat a $k$-bordered diagonal matrix as a rank-$2k$ perturbation of a diagonal matrix. Such approach
would place us in the same setting of the previous section. However, the perturbation is indefinite and the element located at the tip of the arrow, $c_1$, is not necessarily smaller than $d_1$. A better approach is to treat the matrix on its own.

**Lemma 2.3.4.** Let

$$A = \begin{bmatrix} C_1 & C_2 & \cdots & C_{n-k+1} \\ C_2^* & d_1 \\ \vdots \\ C_{n-k+1}^* & \cdots & d_{n-k} \end{bmatrix}$$

be a given $k$-bordered diagonal matrix in the conditions of (2.11). Moreover, for $q \in \{1, \ldots, n-k\}$ let

$$D_q := \begin{bmatrix} d_1 \\ \cdots \\ d_q \\ d_q \mathbf{I}_{n-q-k} \end{bmatrix}$$

and set

$$A_q := \begin{bmatrix} C_1 & C_2 & \cdots & C_\ell \\ C_2^* & \cdots \cdots \\ \vdots \\ C_{\ell}^* \end{bmatrix}$$

where $\ell = n-k+1$. Then the eigenvalues of $A_q$ are lower bounds for the eigenvalues of $A$.

**Proof.** By direct computation it is easily seen that $A - A_q$ is a real diagonal positive semi-definite matrix.

In view of Procedure 2.2.3 consider, now, Procedure 2.3.5 directed to the nonzero columns (and rows) of the matrix $A_q$.

**Procedure 2.3.5.** Consider the 1-bordered diagonal matrix and let the elements of the nonzero column of $A_q$ be the components of a vector $c$, i.e., let $c$ be the vector $c = [c_2, c_3, \ldots, c_{n-1}, c_n]^*$. For $j \in \{1, \ldots, n-2\}$ partition $c$ as $\tilde{c}_j$ containing the first $j$ rows and the vector $\tilde{c}_{n-1-j}$ with the remaining $n-1-j$ rows:

$$c = \begin{bmatrix} \tilde{c}_j \\ \tilde{c}_{n-1-j} \end{bmatrix}. $$
Then, for a unitary matrix $Q$ such that

$$
Q \begin{bmatrix}
\|\tilde{c}_{n-1-j}\|_2 \\
0
\end{bmatrix} = \tilde{c}_{n-1-j}
$$

is a QR-decomposition of $\tilde{c}_{n-1-j}$ we have with $G = I_j \oplus Q$ that

$$
G^* A_q G = \begin{bmatrix}
c_1 & \cdots & c_j & \|\tilde{c}_{n-1-j}\|_2 & O \\
\vdots & & d_1 & & \\
c_j & & \ddots & & \\
\|\tilde{c}_{n-1-j}\|_2 & & & d_q & \\
O & & & & d_q I_{n-r}
\end{bmatrix} = \begin{bmatrix}
Y_{j+1} \\
& d_q I_{n-r}
\end{bmatrix}.
$$

A procedure for $k$-bordered diagonal matrices with $k > 1$ can be described similarly.

As a consequence of Procedure 2.3.5, the eigenvalues of $A_q$ can be determined at the cost of an eigenvalue problem for a small $(j+1) \times (j+1)$ matrix.

**Theorem 2.3.6.** Let $1 \leq k \leq n$ and for $q \in \{1, \ldots, n-2k\}$ set $r = q + 2k$. Then

$$
\sigma(A_q) = \{\sigma(Y_r), d_q\}.
$$

**Proof.** Applying Procedure 2.3.5 to an appropriate partitioning of the columns of $A_q$ and noting that the resulting matrix is block diagonal gives the statement. \qed

## 2.4 Singular Values

The Singular Value Decomposition (SVD) is a tool that has found several applications in areas such as Image Analysis [16] and Information Retrieval [8]. In both areas, the data is commonly represented in matrix form and the SVD is required. However, due to the dynamic nature of the data, new information can be added (or eventually removed) from the original set. In matrix terms this corresponds to adding or removing a column after which a new SVD is required [14, 16, 78]. In this section we are interested in two situations: a) lower bounds for the singular
values of rank-$k$ perturbations of general $n \times m$ matrices, $B + K$; and b) lower bounds for the singular values of $n \times m$ matrices when columns (or rows) are added.

### 2.4.1 Rank-$k$ perturbations of general matrices

The study of singular values is typically done through the study of the eigenvalues of the Hermitian matrices resulting from right or left multiplication with the adjoint. We will follow the same approach here. The starting point is an $n \times m$ matrix $B$. Without loss of generality we assume that $m \leq n$. This matrix will be perturbed with a rank-$k$ matrix $K$ and we are interested in the singular values of $B + K$.

**Proposition 2.4.1.** Given the $n \times m$ matrix $B$, and a rank-$k$ perturbation $K$ of $B$, the singular values of $B + K$ are the square roots of the nonnegative eigenvalues of

$$
(B + K)^*(B + K) = B^*B + K^*V + V^*K,
$$

where $V = B + \frac{1}{2}K$, which constitutes a Hermitian perturbation of rank at most $2k$ of $B^*B$.

**Proof.** Write $V = B + \frac{1}{2}K$, then

$$
(B + K)^*(B + K) = B^*B + K^*B + B^*K + K^*K
= B^*B + K^*(B + \frac{1}{2}K) + (B^* + \frac{1}{2}K^*)K.
$$

This proves the first part of the statement. To see that the perturbation has rank at most $2k$, note that the rank of each of the products $V^*K$ and $K^*V$ is less than or equal to the rank of $K$, which is $k$. 

The lower bounds for Hermitian rank-$k$ perturbations of Hermitian matrices from Section 2.2 can now be applied to the specific perturbation of $B^*B$ in (2.13). These bounds were in principle developed for perturbations $E$ that were given in the form $E = W\Sigma W^*$ with $W^*W = I$. If the perturbation $K$ of $B$ is given as $K = ZX^*$ for certain $n \times k$ matrix $Z$ and $m \times k$ matrix $X$, this can be exploited
in the sense that, writing $S$ for the $m \times k$ matrix $V^*Z$,

$$K^*V + V^*K = X(Z^*V) + (V^*Z)X^*$$

$$= (S|X) \begin{bmatrix} I & I \\ I & I \end{bmatrix} (S|X)^* = QR \begin{bmatrix} I & I \\ I & I \end{bmatrix} R^*Q^*$$

where $(S|X) = QR$ is a QR-decomposition. Computing an eigenvalue decomposition of the $2k \times 2k$ Hermitian matrix

$$R \begin{bmatrix} I & I \\ I & I \end{bmatrix} R^* = YAY^*$$

then leads to the desired format:

$$E = W\Sigma W^* \quad \text{with} \quad W = QY, \quad W^*W = I, \quad \text{and} \quad \Sigma = \Lambda. \quad (2.14)$$

Observe that the generic situation will be that $R$ is invertible. Then, by Sylvester’s Theorem for Inertia, $\Lambda$ has $k$ negative and $k$ positive eigenvalues, and thus, the perturbation of $B^*B$ is indefinite. This shows that one needs to take care when constructing lower bounds for the singular values of $B + K$. If $B^*B$ is replaced by an approximation from below based on some matrix $X_q$, adding the indefinite perturbation $K^*V + V^*K$ from (2.13) may lead to negative eigenvalues of $X_q + K^*V + V^*K$, rendering some of the lower bounds worthless in the context of singular values, being the square roots of eigenvalues.

Remark 2.4.2. If $K$ is given as its singular value decomposition $K = Z\Gamma X^*$ with $n \times k$ matrix $Z$ and $m \times k$ matrix $X$, both with orthonormal columns, and $\Gamma$ a real nonnegative $k \times k$ diagonal matrix, then by absorbing $\Gamma$ into $X^*$, $K$ will have the form used above.

Proposition 2.4.3. Given the $n \times m$ matrix $B$, and a rank-$k$ perturbation $K$ of $B$. Then the singular values of $B + K$ are the square roots of the $m$ largest nonnegative eigenvalues of the $n \times n$ matrix

$$(B + K)(B + K)^* = BB^* + KV^* + VK^*, \quad (2.15)$$
where \( V = B + \frac{1}{2} K \), which constitutes a Hermitian perturbation of rank at most \( 2k \) of \( BB^* \).

Proof. Write \( V = B + \frac{1}{2} K \), then

\[
(B + K)(B + K)^* = BB^* + BK^* + KB^* + KK^*
\]

\[
= BB^* + (B + \frac{1}{2} K)K^* + K(B^* + \frac{1}{2} K^*).
\]

This proves the first part of the statement. To see that the perturbation has rank at most \( 2k \), note that the rank of each of the products \( KV^* \) and \( VK^* \) is less than or equal to the rank of \( K \), which is \( k \).

In this approach, although seemingly similar to the previous, if \( m < n \), the \( n \times n \) matrix \( BB^* \) has \( (n - m) \) artificial zero eigenvalues. Although this is no restriction to apply the approach of Section 2.2, it may reduce the effectiveness of some of the bounds if.

The same distinction between the different ways to form the product of the matrix and its adjoint will be present in the following section.

Addition of columns or rows

Let \( k, n \) and \( m \) be integers with \( n > m \) and \( k \geq 1 \). Consider again the matrix \( B \in \mathbb{C}^{n \times m} \), and matrices \( C \in \mathbb{C}^{n \times k} \) and \( N \in \mathbb{C}^{m \times k} \). We aim to find lower bounds for the singular values of the matrices \( B_c \) and \( B_r \) defined by

\[
B_c = (B|C) \quad \text{and} \quad B_r = (B^*|N)^*;
\]

and for that we look at the eigenvalues of the \( (m + k) \times (m + k) \) Hermitian matrix \( B_c^*B_c \) and the \( n \times n \) Hermitian matrix \( B_cB_c^* \) and their counterparts for rows.
Case 1: Study of $B_c^*B_c$

The matrix $B_c^*B_c$ equals

$$B_c^*B_c = \begin{bmatrix} B^*B & B^*C \\ C^*B & C^*C \end{bmatrix}.$$ 

Now, let

$$B = U \begin{bmatrix} \Sigma_m \\ O \end{bmatrix} W^*$$ 

be an SVD of $B$, with $U \in \mathbb{C}^{n \times n}, W, \Sigma_m \in \mathbb{C}^{m \times m}$. Then with $G = W \oplus I_k$ we find that

$$G^*(B_c^*B_c)G = \begin{bmatrix} \Sigma^2_m \\ Z^* \\ Z \end{bmatrix} \begin{bmatrix} C_1^* \\ C_2^* \end{bmatrix}, \text{ where } Z = W^*(B^*C).$$

This is (a block permutation of) a $k$-bordered diagonal matrix having no artificial zero eigenvalues. The results from Section 2.3 can now be applied to find lower bounds for its eigenvalues. Again, replacing the matrix by an approximation from below may result in a matrix with negative eigenvalues, thus one needs to be cautious.

Case 2: Study of $B_cB_c^*$

The matrix $B_cB_c^*$ equals

$$B_cB_c^* = BB^* + CC^*.$$ 

Transforming this matrix to the basis given by the columns of $U$ from (2.16) shows (with $C_1$ and $C_2$ chosen appropriately) that

$$U^*(B_cB_c^*)U = \begin{bmatrix} \Sigma^2_m \\ O \end{bmatrix} + \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} [C_1^* \quad C_2^*], \text{ with } \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = U^*C.$$ 

Now, let

$$C_2 = Q \begin{bmatrix} R \\ O \end{bmatrix}$$
be a $QR$-decomposition of $C_2$, where $R$ is $k \times k$, then the at most $m + k$ nonzero eigenvalues of $U^*(B_cB_c^*)U$ are the eigenvalues of the $(m + k) \times (m + k)$ matrix

$$
\begin{bmatrix}
\Sigma_m^2 & O_k \\
O_k & R
\end{bmatrix}
+ \begin{bmatrix}
C_1^* & R^*
\end{bmatrix}.
$$

This is a positive semi-definite perturbation of a diagonal matrix, and again the approach of Section 2.2 may be applied to find lower bounds for its eigenvalues. Since these lower bounds will be nonnegative, their square roots are well-defined, and thus valid lower bounds for the singular values of $B_c$ result.

Similar results can, of course, be obtained for the matrix $B_r$: the matrix $B_rB_r^*$ compares to $B_c^*B_c$ and $B_r^*B_r$ to $B_cB_c^*$.

### 2.5 Numerical illustrations

We now illustrate the techniques developed in the previous sections. Our focus is on the computation of lower bounds for the smallest eigenvalue of Hermitian rank-one perturbations of Hermitian matrices. For these matrices not only are the newly proposed bounds cheaply obtainable, but we are also able to compare them with the existing ones from Weyl and Ipsen and Nadler. A few illustrations are also given for rank-2 perturbations of Hermitian matrices and for the addition of columns to $n \times m$ matrices.

**Example 1** For our first numerical experiment we use the reaction-diffusion example given in §2.2 on $I = [0, 1]$:

\[-u'' + su = \lambda u \quad \text{in } I, \quad \text{with} \quad u(0) = u(1) = 0 \]  \tag{2.17}

where the function $s$ is a local reaction term defined in $\left(\frac{1}{2} - \xi, \frac{1}{2} + \xi\right)$ for $\xi \in [0, \frac{1}{2}]$ by

\[s(x) = e^{-g(x)} > 0 \quad \text{where} \quad g(x) = \frac{\xi^2}{(x - (1/2 - \xi))(x - (1/2 + \xi))} \]  \tag{2.18}
and zero elsewhere. Note that our choice of working in one dimension is not restrictive and similar situations can be set up in higher dimensions. Take \( p \in \mathbb{N} \) and define \( h = 1/p \) and \( x_j = jh \) where \( j \in \{0, \ldots, p\} \). The Finite Difference discretization of problem (2.17) on the grid points \( x_0, \ldots, x_p \) yields two matrices. The first is the typical tridiagonal matrix with 2 in the diagonal and \(-1\) in both co-diagonals which we denote by \( A \). The eigenvalues and eigenvectors of \( A \) are known. The matrix \( E = \sum_j s(x_j)e_j e_j^* \), on the other hand, is a diagonal matrix whose rank is determined by \( h \) and \( \xi \) by truncating \( 2\xi/h \) at integer values.

We wish to determine lower bounds for the eigenvalues of \( A + E \) or, on the basis of eigenvectors of \( A \), those of \( \Lambda + \hat{E} \) where the diagonal matrix \( \Lambda \) and the unitary matrix \( V \) are such that \( A = V\Lambda V^* \) and

\[
\hat{E} = V^*EV = \sum_j s(x_j)v_j v_j^* \quad \text{with} \quad v_j = V^*e_j.
\]

We take \( p = 32, \xi = 1/33 \) such that \( E \) from Equation (2.18) has rank-one. We compute lower bounds for the smallest eigenvalue of \( A + E \) using the approaches of Section 2.2.1 and the technique in Section 2.2.3.

The top plot in Figure 2.1 shows the absolute value of the error in the lower bounds for the smallest eigenvalue of \( A + E \). The lower bounds were obtained from approximations \( X_q \) with \( q \) ranging from 1 to \( n \). Recall that each \( X_q \) uses the \( q \) smallest eigenvalues of \( A \) with the largest of these having multiplicity at least \( n - q + 1 \). The errors for the first ten bounds are displayed again, for convenience, in the bottom plot of Figure 2.1. For notation simplicity we use in the label of the \( y \)-axis \( \lambda_- \) to mean \( \lambda_{\min}(A + E) \) and \( \mu_- \) to denote \( \lambda_{\min}(X_q + E) \).

Recall that, for \( q = 1 \), the bound obtained is the same as the one from Weyl’s inequalities while for \( q = 2 \) the same from Ipsen and Nadler’s results. Figure 2.1 shows that there is a slightly more pronounced error reduction at the beginning and at the end, while with the intermediate matrices we observe only minor accuracy improvements. This feature is shared by most of the matrices we have experimented with in this section. Furthermore, we observe from the bottom plot that although the second bound shows the best relative error reduction, it is still advantageous to use larger matrices if more accuracy is required. We expect this
2.5. NUMERICAL ILLUSTRATIONS

Figure 2.1: Error in the lower bound (and detail for $q \leq 10$) of the smallest eigenvalue of the perturbed matrix $A$ resulting from the Finite Difference discretization ($n = 32$, $\xi = 1/33$) of Problem (2.17) with $E$ a rank-one diagonal matrix resulting from Equation (2.18).

to be more pronounced with larger matrices.

Example 2 For our second example we chose a larger Hermitian matrix $A$ of dimension $n = 1600$ resulting from the standard finite difference discretization of the Poisson equation in a square domain using a $40 \times 40$ grid. We perturb this matrix with a random Hermitian matrix of rank-one. The absolute value of the errors for the lower bounds obtained with different approximating matrices $X_q$ is represented in the two plots of Figure 2.2. For the same problem, Table 2.4 shows the approximate values of the error for the first six lower bounds in comparison to the ones for the matrix of Example 3 to be discussed further on.

We can observe from Table 2.4 and from the second plot in Figure 2.2 that the initial error reduction is, in this case, more pronounced. Consequently, and at a cost of a $6 \times 6$ eigenvalue problem ($q = 5$), Weyl’s bounds and those from [46] can be improved. The additional cost is modest while the absolute value of the
error decreases from $O(10^{-2})$ to $O(10^{-6})$. Another interesting observation is the presence of line segments parallel to the $x$-axis between the values $2-3$, $5-6$, $7-8$ and $9-10$. They correspond to multiple eigenvalues (in this case of multiplicity 2) of the matrix $A$. Note that, according to Proposition 2.2.1 and Section 2.2 these lines should still be visible when perturbing the same matrix with a rank-2 perturbation because the two consecutive matrices $X_q$ are equal. We illustrate this in Figure 2.3 where we represent the absolute value of the error for the lower bounds of a random Hermitian rank-2 perturbation of the same matrix $A$.

$$
\begin{align*}
q & \quad |\lambda_- - \mu_-| \\
1 & \quad 1.756e^{-2} \\
2 & \quad 8.544e^{-3} \\
3 & \quad 8.544e^{-3} \\
4 & \quad 7.572e^{-5} \\
5 & \quad 5.586e^{-6} \\
6 & \quad 5.586e^{-6} \\
\vdots & \quad \vdots \\
\end{align*}
$$

Table 2.4: Error in $\mu_- := \lambda_{\min}(Y_r + \tilde{E})$ for $1 \leq r \leq n$ for a first few lower bounds for the matrix obtained from the discretization of Poisson equation on a $40 \times 40$ grid (left) and Sherman1 from Matrix Market (right). The perturbations are random rank-one matrices.

**Example 3** For the third example we retrieve the matrix Sherman1 from the Matrix Market repository [1]. This is a real symmetric matrix of size $n = 1000$ with all eigenvalues situated on the negative real line. The perturbation was again a random Hermitian matrix of rank-one. We observe both from the second plot of Figure 2.4 as well as from the second column of Table 2.4 that the results for this matrix are not as good as the ones from Example 2. However, this example is interesting for just that. The stagnation period between $q = 200$ and $q = 500$ is due to the existence of an eigenvalue of multiplicity 300. We have mentioned in Section 2.2.4 that multiple eigenvalues are the most simple clusters and that these can be exploited to obtain improved bounds without increasing the size of the approximating matrix.

We illustrate this situation in Figure 2.4 where to construct the approximations of the matrix $A$ we make use of the knowledge of an eigenvalue with large
2.5. NUMERICAL ILLUSTRATIONS

Figure 2.2: Error in the lower bound (and detail for \( q \leq 10 \)) of the smallest eigenvalue of a random rank-one perturbation \( E \) of the matrix \( A \) obtained from the discretization of Poisson equation on a \( 40 \times 40 \) grid.

In the bottom plot of this figure we see the absolute value of the error for lower bounds for the smallest eigenvalue using the standard approach of Section 2.2.1. We observe also the long stagnation period. In the top plot we have computed the bounds using the approach of Section 2.2.4. The stagnation stage is now gone and the exact eigenvalues are obtained using at most a matrix with \( n = 700 \).

Example 4 The final example illustrates the theory of Section 2.4.1. Here we compute the error for the lower bounds of the smallest singular value of a matrix \( B_c \) resulting from the addition of a columns to an \( n \times m \) matrix \( B \). The matrix \( B \) is \textit{ILLC1850} of dimensions \( 1850 \times 712 \), also retrieved from the Matrix Market [1]. Figure 2.5 plots the absolute value of the error in the lower bounds for the smallest singular value of \( B_c \) given by different approximating matrices. In the bottom plot we display the first twenty elements. Once again, a few small eigenproblems suffice to obtain reasonable approximations for the singular values. For this matrix and due to the fast error reduction stage at the beginning, using more than the second
lower bounds is certainly worthwhile.

2.6 Final notes

The previous sections have dealt with Hermitian matrices. Some of the theory developed, however, can be used with matrices other than Hermitian. In the next few pages we comment on examples of these situations.

Remark 2.6.1. If $A$ and $E$ are Hermitian matrices, $iA$ and $iE$ are skew-Hermitian and we can use the previous theory to study skew-Hermitian perturbations of skew-Hermitian matrices. In fact, for $\theta \in [0, 2\pi)$ this is true for any matrices $e^{2\pi i \theta} A$ and $e^{2\pi i \theta} E$. In the next chapter we investigate a more general class of normal matrices of which these are a particular case.

2.6.1 Non-Hermitian perturbations of Hermitian matrices

The notions of lower or upper bounds for eigenvalues loose their meaning for most non-Hermitian matrices. Some of the techniques described earlier, however, are
Figure 2.4: Error in the lower bound of the smallest eigenvalue of the perturbed matrix $A$ for the matrix Sherman1 when a multiple eigenvalue is clustered (below) and comparison with the naive approach (above).

Figure 2.5: Error in the lower bound (and detail for $q \leq 20$) of the smallest singular value of $B_c$ where $B$ is the matrix ILLC1850 of dimensions $1850 \times 712$ from the Matrix Market and $c$ is a random $n$-dimensional vector.
still valid. The next result shows that when a Hermitian matrix $A$ is perturbed by a non-Hermitian matrix $E$, we can still use similarity transformations to reduce the size of the eigenvalue problem for $A + E$. We begin by presenting a useful procedure.

**Procedure 2.6.2.** Given $v, w \in \mathbb{C}^n$ define the matrix $Z = \begin{bmatrix} v & w \end{bmatrix}$. Partition each column of $Z$ in $p$ blocks (which, for notational simplicity, we assume now to be two) as

$$Z = \begin{bmatrix} \vec{v}_1 & \vec{w}_1 \\ \vec{v}_2 & \vec{w}_2 \end{bmatrix}.$$ 

For $p \in \{1, 2\}$ let $Q_p$ and $R_p$ be such that

$$Q_p \begin{bmatrix} R_p \\ O \end{bmatrix}$$

is a QR-decomposition of each partition $[\vec{v}_p \ \vec{w}_p]$ and denote by $Q$ and $R$ the matrices

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} R_1^* & O \\ R_2^* & O \end{bmatrix}^*$$

respectively. Note now that

$$Q^*vw^*Q = \begin{bmatrix} * \\ 0_{\ell_1}^* \\ * \end{bmatrix} \begin{bmatrix} * & * & 0^*_{\ell_1 - 1} & * & 0^*_{\ell_2 - 1} \\ * & * & * & * & * \\ 0_{\ell_2} \end{bmatrix} = zy^*.$$

We now use Procedure 2.6.2 to prove the following theorem.

**Theorem 2.6.3.** Given $v, w \in \mathbb{C}^n$ set $E := vw^*$. Let $X \in \mathbb{R}^{n \times n}$ be diagonal with exactly $t$ distinct eigenvalues denoted by $\lambda_p$ ($1 \leq p \leq t$) with multiplicity $m_p$. Let $z$ and $y$ be obtained from $v$ and $w$ using Procedure 2.6.2 with $p = t$ and $v_p, w_p \in \mathbb{C}^{m_p}$. Then the eigenvalues of $X + E$ are the eigenvalues of $X$, of which double copies
of multiple eigenvalues are replaced by the eigenvalues of the $2t \times 2t$ matrix

$$Y = \begin{bmatrix} \lambda_1 I_2 & & \\ & \ddots & \\ & & \lambda_t I_2 \end{bmatrix} + zy^*.$$

Moreover,

$$\{\lambda_1, \ldots, \lambda_t\} \subset \sigma(Y).$$

**Proof.** We give the details of the proof for $t = 2$ after which the proof for $t > 2$ should be clear. Let

$$X = \begin{bmatrix} \lambda_1 I_{m_1} \\ \lambda_2 I_{m_2} \end{bmatrix}$$

where for $m_p > 1$ is the multiplicity of $\lambda_1$ and $m_2 > 1$ is the multiplicity of $\lambda_2$ and $m_1 + m_2 = n$. Using Procedure 2.6.2 we obtain a unitary matrix $Q$ such that $Q^*vw^*Q = zy^*$ and $Q^*XQ = X$. Therefore,

$$Q^*(X + E)Q = X + Q^*vw^*Q = X + zy^*.$$

By appropriate permutations we may write $X + zy^*$ as $Y \oplus \lambda_1 I_{m_1 - 2} \oplus \lambda_2 I_{m_2 - 2}$ where

$$Y = \begin{bmatrix} \lambda_1 \\ \lambda_2 & \lambda_1 \end{bmatrix} + \begin{bmatrix} * \\ * & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} * & * & * & * \end{bmatrix}.$$

Note that $Y$ can be seen as the sum of an upper triangular matrix with a Hermitian rank-one and thus the eigenvalues on the two last diagonal positions of the diagonal matrix are still eigenvalues of $Y$. Therefore, $m_p - 1$ copies of the eigenvalues of $X$ are still eigenvalues of $X + E$. This proves the statement for $t = 2$. For $t > 2$ the proof is similar. \qed

When compared with Proposition 2.2.1 the consequences of the loss of symmetry are clear. The resulting matrix is now of size $2t \times 2t$ instead of size $t \times t$ for the Hermitian case.
2.6.2 Non-Hermitian $k$-bordered diagonal matrices

In Lemma 2.3.2 we have related Hermitian rank-$k$ perturbations of Hermitian matrices to Hermitian $k$-bordered diagonal matrices. When the perturbation is non-Hermitian, however, we are still able to show a similar result.

**Lemma 2.6.4.** Let $n$ be an integer and let $A \in \mathbb{C}^{n \times n}$ be Hermitian and $E \in \mathbb{C}^{n \times n}$ be a non-Hermitian matrix. If $E$ is of rank-$k$ then $A + E$ is unitarily similar to a non-Hermitian $k$-bordered diagonal matrix.

**Proof.** Let $U \in \mathbb{C}^{n \times n}$, $\Theta \in \mathbb{C}^{k \times k}$ and $L \in \mathbb{C}^{(n-k) \times k}$ be such that

$$U^*EU = \begin{bmatrix} \Theta & L^* \\ O_{n-k} \end{bmatrix}$$

is a Schur decomposition of $E$ and consider the matrix

$$U^*AU = \begin{bmatrix} M & B^* \\ B & T \end{bmatrix},$$

for certain $M \in \mathbb{C}^{k \times k}$, $B \in \mathbb{C}^{(n-k) \times k}$ and (Hermitian) $T \in \mathbb{C}^{(n-k) \times (n-k)}$. Now, let $V$ be unitary and satisfy $V^*TV = D$ with $D$ the diagonal matrix of eigenvalues of $T$. Defining the unitary matrices

$$G = \begin{bmatrix} I_k \\ V \end{bmatrix}$$

and

$$S = UG,$

$A + E$ can now be transformed as

$$S^*(A + E)S = \begin{bmatrix} M + \Theta & (B^* + L^*)V \\ V^*B & D \end{bmatrix}. \tag{2.19}$$

which is $k$-bordered diagonal.

In the same way, it is possible (in the presence of a diagonal element of large multiplicity) to reduce the size of a non-Hermitian $k$-bordered diagonal eigenproblem by unitary similarity transformations. This is described in Procedure 2.6.5.
Procedure 2.6.5. Consider the non-Hermitian 1-bordered diagonal matrix $A \in \mathbb{C}^{n \times n}$

$$A = \begin{bmatrix}
a_1 & w_2 & \ldots & w_q & \tilde{w}^* \\
v_1 & d_1 & \\
\vdots & \ddots & \\
v_q & d_q & \\
\tilde{v} & d_{q+1}I_{n-q-1}
\end{bmatrix}$$

where $\tilde{v} = [v_{q+1}, \ldots, v_{n-1}]^*$ and $\tilde{w} = [w_{q+1}, \ldots, w_{n-1}]^*$. Let $QR$ be a QR-decomposition of $Z = [\tilde{v} \quad \tilde{w}]$. The matrix $G = I_q \oplus Q$ is then unitary and transforms $A$ into

$$G^*AG = \begin{bmatrix}
D_1 \\
d_{q+1}I_{n-q-3}
\end{bmatrix}$$

where

$$D_1 = \begin{bmatrix}
a_1 & w_2 & \ldots & w_q & * & * \\
v_1 & d_1 & \\
\vdots & \ddots & \\
v_q & d_q & \\
* & d_{q+1} \\
0 & d_{q+1}
\end{bmatrix}.$$ 

Therefore, the eigenvalues of $A$ are the eigenvalues of $D_1$ together with $d_{q+1}$ with multiplicity $n - q - 3$.

Even though the eigenvalues of $X_q + E$ no longer bound the eigenvalues of $A + E$, they can still be considered approximations for them. We end this chapter with an illustration of this situation.

2.6.3 Example

Let $w = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^*$ and $v = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 \end{bmatrix}^*$ and consider the matrix $E = wv^*$ as well as the matrix $A$ from Example 2.1.1. The approximate eigenvalues of $X_3 + E$ and $A + E$ are displayed in Table 2.5.

Note that by applying the Bauer-Fike Theorem (Theorem 1.1.1) to the matrix $(A+E) + (X_3-A) = X_3 + E$ we would only expect that for at least one eigenvalue
<table>
<thead>
<tr>
<th></th>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( \lambda_4 )</th>
<th>( \lambda_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_3 + E )</td>
<td>2.76-0.34i</td>
<td>2.76+0.34i</td>
<td>3.1</td>
<td>3.1</td>
<td>3.58</td>
</tr>
<tr>
<td>( A + E )</td>
<td>2.77-0.36i</td>
<td>2.77+0.36i</td>
<td>3.59-0.25i</td>
<td>3.59+0.25i</td>
<td>4.48</td>
</tr>
</tbody>
</table>

Table 2.5: Eigenvalues of \( X_3 + E \) and \( A + E \) where \( E \) is a non-Hermitian perturbation.

\[
\lambda \text{ of } A + E \quad \quad |\lambda - \mu| \leq \kappa_2(V)\|X_3 - A\|_2 \approx 12
\]

where \( V \) diagonalizes \( A + E \) and \( \mu \) is an eigenvalue of \( X_3 + E \). This shows the potential of approximating the eigenvalues of \( A \) in this manner.
Chapter 3

Normal matrices

When $A$ and $E$ are not Hermitian and even when the analysis is restricted to normal matrices the study of perturbations is difficult. One important reason is that for normal matrices many fundamental questions remain unanswered. For instance, given $A$ and $E$ normal matrices, it is still not fully understood under which conditions is $A + E$ also normal.

In this chapter we deal with normality preserving perturbations and augmentations of normal matrices and their consequences to the eigenvalues. We revisit the normality preserving augmentation of normal matrices studied by Ikramov and Elsner [45] in 1998 and complement their results by showing how the eigenvalues of the original matrix are perturbed by the augmentation. Moreover, we construct all augmentations that result in normal matrices with eigenvalues on a quadratic curve in the complex plane, using the stratification of normal matrices presented by Huhtanen [42] in 2001. To make this construction feasible, but also for its own sake, we study normality preserving normal perturbations of normal matrices. For $2 \times 2$ and for rank-one matrices, the analysis is complete. For higher rank, all essentially Hermitian normality perturbations are described. In all cases, the effect of the perturbation on the eigenvalues of the original matrix is given.
3.1 Introduction

A complex number $z \in \mathbb{C}$ is often split up as $z = \Re(z) + \Im(z)$, where $2\Re(z) = z + \bar{z}$ and $2\Im(z) = z - \bar{z}$. This interprets the complex plane as a two-dimensional real vector space with as basis the numbers 1 and $i$. In this chapter, it will be convenient to decompose $z$ differently. For this, we introduce a family of decompositions parametrized in $\theta \in \mathbb{T} \subset \mathbb{C}$, the circle group of unimodular numbers. For a given $\theta \in \mathbb{T}$ we let

$$z = \Theta(z) + \Theta^\perp(z), \quad \text{where} \quad \Theta(z) = \Re(\bar{\theta}z) \theta \quad \text{and} \quad \Theta^\perp(z) = \Im(\bar{\theta}z) \theta. \quad (3.1)$$

Moreover, apart from the standard Toeplitz or Cartesian decomposition of a square matrix, $A$, into its Hermitian and skew-Hermitian parts,

$$A = \mathcal{H}(A) + \mathcal{S}(A), \quad \text{where} \quad \mathcal{H}(A) = \frac{1}{2}(A + A^*) \quad \text{and} \quad \mathcal{S}(A) = \frac{1}{2}(A - A^*),$$

in accordance with (3.1) we consider the family of matrix decompositions

$$A = \Theta(A) + \Theta^\perp(A), \quad \text{where} \quad \Theta(A) = \mathcal{H}(\bar{\theta}A) \theta \quad \text{and} \quad \Theta^\perp(A) = \mathcal{S}(\bar{\theta}A) \theta. \quad (3.2)$$

We call this decomposition the $\theta$-Toeplitz decomposition of $A$. The matrix $\Theta(A)$ is then the $\theta$-Hermitian part of $A$ and $\Theta^\perp(A)$ its $\theta$-skew-Hermitian part and subsequently, $A$ is $\theta$-Hermitian if $A = \Theta(A)$ and $\theta$-skew-Hermitian if $A = \Theta^\perp(A)$. Note that $\theta$-skew-Hermitian matrices are $i\theta$-Hermitian. We now generalize a well-known result to the $\theta$-Toeplitz decomposition of normal matrices.

**Lemma 3.1.1.** Let $\theta \in \mathbb{T}$ be arbitrarily given. For any normal matrix $A$ we have that

$$Av = \lambda v \iff \Theta(A)v = \Theta(\lambda)v \quad \text{and} \quad \Theta^\perp(A)v = \Theta^\perp(\lambda)v,$$

where $\Theta$ and $\Theta^\perp$ and their relation to $\theta$ are defined in (3.2).

**Proof.** Let $Au = \lambda u$ for some $u \neq \bar{0}$. Since $A$ is normal, there exists a unitary matrix $U$ with $u$ as first column and $U^*AU = \Lambda$ diagonal. But then $U^*A^*U = \Lambda^*$, showing that $A^*u = \bar{\lambda}u$. Since the argument can be repeated with $A^*$ instead of $A$, that yields that $A$ and $A^*$ have the same eigenvectors and that the corresponding
3.1. INTRODUCTION

eigenvalues are each other’s complex conjugates. Therefore,
\[ 2\Theta(A)v = (\bar{\theta}A + \theta A^*)\theta v = (\bar{\theta}\lambda + \theta \bar{\lambda})\theta v = 2\Theta(\lambda)v \]
and, similarly, also \( \Theta^\perp(A)v = \Theta^\perp(\lambda)v \). The reverse implication is trivial. \( \square \)

**Corollary 3.1.2.** Let \( \lambda_1 \) be an eigenvalue of a normal matrix \( A \). For given \( \theta \in \mathbb{T} \), consider the line \( \ell \subset \mathbb{C} \) through \( \lambda_1 \) defined by
\[ \ell : \{ \lambda_1 + \rho \theta \mid \rho \in \mathbb{R} \}. \]
Assume that \( \lambda_1, \ldots, \lambda_p \) are all eigenvalues of \( A \) that lie on \( \ell \). Then the eigenspace \( U \) of the eigenvalue \( \Theta^\perp(\lambda_1) \) of \( \Theta^\perp(A) \) equals the invariant subspace of \( A \) spanned by \( u_1, \ldots, u_p \). Restricted to \( U \), the matrix \( A - \lambda_1 I_p \) is \( \theta \)-Hermitian.

**Proof.** Let \( \lambda_a \) and \( \lambda_b \) be eigenvalues of \( A \), then we have that
\[ \Theta^\perp(\lambda_a) = \Theta^\perp(\lambda_b) \iff \Im(\bar{\theta}\lambda_a) = \Im(\bar{\theta}\lambda_b) \iff \bar{\theta}(\lambda_a - \lambda_b) = \rho \in \mathbb{R} \iff \lambda_a - \lambda_b = \theta \rho. \]
Thus, \( \Theta^\perp(A)u_j = \Theta^\perp(\lambda_1)u_j \) for all \( j \in \{1, \ldots, p\} \), and conversely, if we have \( \Theta^\perp(A)u = \Theta^\perp(\lambda_1)u \) then \( u \) is a linear combination of \( u_1, \ldots, u_p \). Writing \( U_p \) for the matrix with columns \( u_1, \ldots, u_p \) we moreover find that
\[ AU_p = U_p \Lambda_p \quad \text{with} \quad \Lambda_p = \theta R + \lambda_1 I_p \]
where \( \Lambda_p \) is the \( p \times p \) diagonal matrix whose eigenvalues are \( \lambda_1, \ldots, \lambda_p \) and \( R \) is real diagonal. Thus, \( \Theta^\perp(U^*_p A U_p - \lambda_1 I_p) = O_p \) proving the last statement. \( \square \)

### 3.1.1 Eigenvalues on polynomial curves

If \( A \) is \( \theta \)-Hermitian then \( A \) is normal. Moreover, by the spectral theorem for Hermitian matrices, all eigenvalues of \( A \) lie on the line \( \ell : \{ \rho \theta \mid \rho \in \mathbb{R} \} \). In the literature, for instance [7, 23, 27], the matrix \( A \) is called **essentially Hermitian** if there exists an \( \alpha \in \mathbb{C} \) such that \( A - \alpha I \) is \( \theta \)-Hermitian for some \( \theta \in \mathbb{T} \). Clearly, the spectrum of an essentially Hermitian matrix lies on an affine line shifted over
α ∈ C. Conversely, if a normal matrix has all its eigenvalues on a line ℓ ⊂ C, it is essentially Hermitian. This includes all normal 2 × 2 matrices and all normal rank-one perturbations of αI for α ∈ C. Larger and higher rank normal matrices have their eigenvalues on a polynomial curve C ⊂ C of higher degree.

Polynomial curves of degree \( k \geq 2 \)

Each matrix \( A ∈ \mathbb{C}^{n×n} \) has its eigenvalues on a polynomial curve \( C ⊂ \mathbb{C} \) of degree \( k ≤ n − 1 \). This can be explained as follows (see also [12]). First, fix \( θ ∈ \mathbb{T} \) such that for each pair \( λ_p, λ_q \) of eigenvalues of \( A \)

\[
λ_p ≠ λ_q ⇒ \Theta(λ_p) ≠ \Theta(λ_q).
\]  

(3.3)

Note that there exists at most \( n(n−1) \) values of \( θ \) for which this cannot be realized. These values correspond to the at most \( \frac{1}{2}n(n−1) \) lines going through each pair of distinct eigenvalues of \( A \). Once (3.3) is satisfied, the points

\[
(\theta \Theta(λ_1), \overline{iθ} \Theta^+(λ_1)), \ldots, (\theta \Theta(λ_n), \overline{iθ} \Theta^+(λ_n)) \in \mathbb{R} \times \mathbb{R}
\]

form a feasible set of points in \( \mathbb{R} \times \mathbb{R} \) through which a Lagrange interpolation polynomial \( ϕ ∈ \mathcal{P}^{n−1}(\mathbb{R}) \) can be constructed that satisfies

\[
Θ^+(λ_j) = i · θϕ(\overline{θ}(λ_j)) \quad \text{for all} \ j ∈ \{1, \ldots, n\}.
\]

If \( A \) is normal, however, we may draw additional consequences. We summarize these in the following lemma.

Lemma 3.1.3. Let \( A \) be normal and let \( θ \) be such that (3.3) is satisfied. Then there is a \( ϕ ∈ \mathcal{P}^{n−1}(\mathbb{R}) \) such that \( Θ^+(A) = i · θϕ(\overline{θ}(A)) \), and thus the \( θ \)-Toeplitz decomposition of \( A \) can be written as

\[
A = Θ(A) + i · θϕ(\overline{θ}(A)),
\]
or, in terms of $\bar{\theta}A$ and its classical Toeplitz decomposition,

$$\bar{\theta}A = \mathcal{H}(\bar{\theta}A) + i\varphi(\mathcal{H}(\bar{\theta}A)).$$

Moreover, the eigenvalues of $A$ lie on the image $\mathcal{C}$ of the function

$$c : \mathbb{R} \to \mathbb{C} : \rho \mapsto \theta \rho + i \theta \varphi(\rho).$$

Proof. The statement is proved by applying the spectral theorem for normal and Hermitian matrices.

Remark 3.1.4. Note that $A$ is essentially Hermitian if and only if there exists a $\theta \in \mathbb{T}$ such that the interpolating polynomial $\varphi \in \mathcal{P}^1(\mathbb{R})$. In fact, $\varphi$ may even be in $\mathcal{P}^0(\mathbb{R})$.

Remark 3.1.5. If for some normal matrix $A$ the degree of the interpolation polynomial equals, say, two for some value of $\theta$, it may well be of degree $n - 1$ for almost all other values of $\theta$, since in that case the eigenvalues lie on a rotated parabola (see Figure 3.1).

Remark 3.1.6. There does not seem to be an easy way to determine $\theta$ for which the polynomial degree is minimal, although for given $\theta$, the polynomial can be computed in a finite number of arithmetic operations without knowing the eigenvalues. This is explained in the following section.

Computing the polynomial $\varphi$ for given $\theta \in \mathbb{T}$

For almost any fixed value of $\theta$, the interpolation polynomial $\varphi$ belonging to a normal matrix $A$ can be computed, without knowing the eigenvalues of $A$, in a finite number of arithmetic operations. This provides us with a curve $\mathcal{C} \subset \mathbb{C}$ on which all eigenvalues of $A$ lie. Indeed, if the degree of $\varphi$ equals $k$ then $A - \Theta(A)$ is a linear combination of

$$i\theta I, \ i\theta \Theta(A), \ i(\theta \Theta(A))^2, \ldots, \ i(\theta \Theta(A))^k.$$
Making the combination explicit is equivalent to finding the coefficients of $\varphi$. To obtain $\varphi$ in practice, notice that for any $v \in \mathbb{C}^n$,

$$\varphi \left( \overline{\Theta}(A) \right) v \in \mathcal{K}^k \left( \overline{\Theta}(A), v \right) = \text{span}\{v, \overline{\Theta}(A)v, \ldots, \left( \overline{\Theta}(A) \right)^k v\},$$

the Krylov subspace of dimension $k$ generated by the matrix $\overline{\Theta}(A)$ and the vector $v$. Since $\overline{\Theta}(A)$ is Hermitian, an orthonormal basis for $\mathcal{K}^k \left( \overline{\Theta}(A), v \right)$ can be constructed using a three-term recursion, and solving the linear system can be done cheaply. These and other considerations led Huhtanen to the development of efficient structure preserving eigensolvers \cite{42} and linear systems \cite{43} for problems involving normal matrices.

### 3.1.2 Commuting normal matrices

We begin by recalling a well-known result \cite[§1.3]{39} on commuting normal matrices to which we shall refer often. Moreover, because we will also need to draw conclu-
sions about commuting normal matrices with distinct or with multiple eigenvalues, we opt to give its complete proof.

**Lemma 3.1.7.** Normal matrices \( A \in \mathbb{C}^{n \times n} \) and \( E \in \mathbb{C}^{n \times n} \) commute if and only if they are simultaneously unitarily diagonalizable.

**Proof.** If both \( \Lambda = W^*AW \) and \( \Delta = W^*EW \) are diagonal for a unitary matrix \( W \), then clearly \( [A,E] = W[\Lambda, \Delta]W^* = O_n \). Conversely, assume that \( [A,E] = O_n \). Let \( U \) be a unitary matrix such that \( \Delta = U^*EU \) is diagonal with multiple eigenvalues being neighbors on the diagonal of \( \Delta \). Thus

\[
\Delta = \begin{bmatrix}
\delta_1 I_{m_1} \\
& \ddots \\

& & \delta_\ell I_{m_\ell}
\end{bmatrix},
\]

where \( m_j \) denotes the multiplicity of \( \delta_j \). Let \( S = U^*AU \) and write \( s_{pq} \) for its entries. Then equating the entries of \( S\Delta \) and \( \Delta S \) in view of the relation

\[
[S,\Delta] = U^*[A,E]U = O_n,
\]

shows that \( s_{pq} = 0 \) whenever \( \delta_p \neq \delta_q \). Thus, \( S \) is block diagonal with respective blocks \( S_1, \ldots, S_\ell \) of sizes \( m_1, \ldots, m_\ell \). For each \( j \in \{1, \ldots, \ell\} \), \( S_j \) is normal. Let \( S_j = Q_j \Lambda_j Q_j^* \) for some unitary \( Q_j \) and diagonal \( \Lambda_j \). Now, with

\[
Q = \begin{bmatrix}
Q_1 \\
& \ddots \\

& & Q_\ell
\end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix}
\Lambda_1 \\
& \ddots \\

& & \Lambda_\ell
\end{bmatrix} \quad \text{and} \quad W = UQ
\]

we find that \( \Delta = Q^*\Delta Q = W^*EW \) and \( \Lambda = Q^*SQ = W^*AW \).

**Remark 3.1.8.** In case all eigenvalues of \( E \) are distinct, then the matrix \( S \) in the proof is itself diagonal and the proof is complete. In case \( E \) has an eigenvalue, say \( \delta_1 \), of multiplicity \( m_1 > 1 \), then there is freedom in the choice of the first \( m_1 \) columns \( u_1, \ldots, u_{m_1} \) of \( U \) that correspond to \( \delta_1 \). Even though each choice diagonalizes \( E \), not each choice diagonalizes \( A \) as well. This is expressed as \( S \)
having a diagonal block $S_1$ of size $m_1$. Writing $U_1$ for the matrix with columns $u_1, \ldots, u_{m_1}$ we have that

$$AU_1 = U_1 S_1,$$

hence the column span of $U_1$ is an invariant subspace $V$ of $A$. The matrix $Q_1$ determines, through the transformation $W_1 = U_1 Q_1$, an orthonormal basis for $V$ of eigenvectors of $A$. If $S_1$ has multiple eigenvalues, again there may be much freedom in the choice of $Q_1$.

### 3.2 Normality preserving augmentation

In this section we revisit, from an alternative point of view, a problem studied by Ikramov and Elsner in [45]. It concerns the augmentation by a number $m$ of rows and columns of a normal matrix in such a way that normality is preserved. Our analysis differs from the one in [45], and we add details on the eigendata of $A_+$ in terms of those of $A$.

**Normality preserving augmentation.** Let $A \in \mathbb{C}^{n \times n}$ be normal. Characterize all $n \times m$ matrices $V, W \in \mathbb{C}^{n \times m}$, and all $\Gamma \in \mathbb{C}^{m \times m}$ such that

$$A_+ = \begin{bmatrix} A & V \\ W^* & \Gamma \end{bmatrix},$$

is normal, too. In other words, characterize all normality preserving augmentations of $A$.

**Remark 3.2.1.** Note that Hermicity, $\theta$-Hermicity and essentially Hermicity preserving augmentation problems are all trivial, because each of these properties is inherited by principal submatrices. For unitary matrices this does not hold. However, if $A$ and $A_+$ are both unitary, their rows and columns all have length one. Thus $V = W = O$ and $\Gamma$ is unitary. This solves the unitarity preserving augmentation problem.
3.2. NORMALITY PRESERVING AUGMENTATION

3.2.1 Normality preserving augmentation for $m = 1$

First consider the case $m = 1$, and write $v, w$ and $\gamma$ instead of $V, W$ and $\Gamma$. It is easily verified that $[A_+, A^*_+] = O_n$ if and only if

$$ww^* = vv^*, \quad w^*w = v^*v \quad \text{and} \quad A^*v + \gamma w = Aw + \gamma v.$$  \hfill (3.5)

The two leftmost relations hold if and only if $v = \phi w$ for some $\phi \in \mathbb{T}$. The rightmost relation may add further restrictions on $v, w$ and $\phi$. Before studying these, however, note that $\phi$ is the square of a unique $\theta \in \mathbb{T}_U \subset \mathbb{T}$, where

$$\mathbb{T}_U = \{ \tau \in \mathbb{T} \mid \arg(\tau) \in [0, \pi) \}.$$ 

This yields a reformulation of $v = \phi w$ that better reveals the underlying structure,

$$v = \phi w \iff v = \theta^2 w \iff \theta v = \theta w = u \iff v = \theta u \quad \text{and} \quad w = \theta u$$

for some $u \in \mathbb{C}^n$. Further restrictions on the vector $u$ and the scalar $\theta \in \mathbb{T}_U$ follow from substituting $v = \theta u$ and $w = \theta u$ into the rightmost equation in (3.5). After some rearrangements we obtain the condition

$$\Theta^\perp(A)u = \Theta^\perp(\gamma)u.$$ 

Because the eigenpairs of $\Theta^\perp(A)$ were already characterized in Corollary 3.1.2, we have solved the augmentation problem for $m = 1$. The theorem below summarizes this solution, constructively, in terms of the eigendata of $A$. Note that this result was proved already in [45], though in a different manner.

**Theorem 3.2.2.** Let $A$ be normal. Let $\gamma \in \mathbb{C}$, and let $\ell : \{ \gamma + \rho \theta \mid \rho \in \mathbb{R} \}$ be a line in $\mathbb{C}$ through $\gamma$ with slope $\theta \in \mathbb{T}_U$. Moreover, let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $A$ that lie on $\ell$. Then, the matrix

$$A_+ = \begin{bmatrix} A & v \\ w^* & \gamma \end{bmatrix}$$
is normal if \( v = \theta u \) and \( w = \theta \bar{u} \), where \( u \) is a linear combination of eigenvectors corresponding to \( \lambda_1, \ldots, \lambda_p \), with the convention that \( u = \vec{0} \) if \( p = 0 \). Conversely, if \( A_+ \) is normal then \( v = \theta u \) and \( w = \theta \bar{u} \), for some \( \theta \) and the vector \( u \) is a linear combination of eigenvectors of \( A \) whose corresponding eigenvalues all lie on a line \( \ell : \{ \gamma + \rho \theta \mid \rho \in \mathbb{R} \} \) for some fixed \( \gamma \in \mathbb{C} \).

**Proof.** Corollary 3.1.2 shows the relation between the eigendata of \( \Theta^\perp(A) \) and \( A \), and together with the derivation in this Section this proves the statement.

### 3.2.2 Eigenvalues of the augmented matrix

We now augment the analysis in [45] with a study of the eigenvalues of \( A_+ \) in relation to those of \( A \). Let \( \Lambda_p \in \mathbb{C}^{p \times p} \) be the diagonal matrix whose eigenvalues are the \( p \) eigenvalues of \( A \) that lie on \( \ell : \mathbb{R} \to \mathbb{C} : \gamma + \rho \theta \). Then, as already mentioned in the proof of Corollary 3.1.2

\[
\Lambda_p = \theta R + \gamma I_p
\]

for some real diagonal matrix \( R \). Let \( U \in \mathbb{C}^{n \times n} \) be any unitary matrix whose last \( p \) columns are eigenvectors of \( A \) belonging to \( \lambda_1, \ldots, \lambda_p \) and let \( U_p \) contain those last \( p \) columns of \( U \). Then, assuming that \( A \) and \( A_+ \) are normal, Theorem 3.2.2 shows, with \( r = U_p^* u \), that

\[
\begin{bmatrix}
U \\ 1
\end{bmatrix}^* \begin{bmatrix}
A & v \\
w^* & \gamma
\end{bmatrix} \begin{bmatrix}
U \\ 1
\end{bmatrix} = \begin{bmatrix}
B \\
\Lambda_p & \theta r \\
\theta r^* & \gamma
\end{bmatrix}.
\] (3.6)

Moreover,

\[
\begin{bmatrix}
\Lambda_p & \theta r \\
\theta r^* & \gamma
\end{bmatrix} = \theta R_+ + \gamma I_{p+1}, \quad \text{where} \quad R_+ = \begin{bmatrix}
R & r \\
r^* & 0
\end{bmatrix}.
\] (3.7)

The above observations reveal some additional features of the solution of the augmentation problem, that we formulate as a new theorem.
Theorem 3.2.3. The only normality preserving 1-augmentations of $A$ are the ones that, on an orthonormal basis of eigenvectors of $A$, augment a $p \times p$ essentially Hermitian submatrix of $A$. Hence, $n - p$ eigenvalues of $A$ are also eigenvalues of $A_+$. The remaining $p + 1$ eigenvalues of $A_+$ lie on the same line as, and are interlaced by, the remaining $p$ eigenvalues of $A$.

Proof. The block form in Equation (3.6) shows that the eigenvalues of $B$ are eigenvalues of both $A$ and $A_+$, whereas (3.7) shows that to locate the remaining $p + 1$ eigenvalues of $A_+$, one only needs to observe that, by Cauchy Interlace Theorem (Theorem 1.1.10) the (real) eigenvalues of $R$ interlace those of $R_+$. 

Remark 3.2.4. Note that the case $p = 0$, covered by Theorem 3.2.2, is also included in the above analysis if one is willing to interpret the remaining $p$ eigenvalues of $A$ as any line in $\mathbb{C}$. This just reflects that the additional eigenvalue, $\gamma \in \mathbb{C}$, of $A_+$ can lie anywhere.

For an illustration of the constructions described in Theorems 3.2.2 and 3.2.3 we refer the reader to Section 3.5.1. There, we augment a given $3 \times 3$ matrix $A$ in two different ways and compute the eigenvalues of the augmented matrix, $A_+$.

3.2.3 Normal matrices with normal principal submatrices

By applying the procedure for $m = 1$ several times consecutively, we may also construct $m$-augmentations with $m > 1$. In particular, all normal matrices having the property that all their leading principal submatrices are normal can be constructed. Since, generally, normal matrices do not have normal principal submatrices, this shows that the $m$-augmentation for $m > 1$ has not yet been completely solved. In Section 3.4 we investigate the principal submatrices of normal matrices from the point of view of Section 3.1.1. This study will also give more insight into the $m$-augmentation for $m = 2$. In [45], this case already proved to be quite complicated. In particular, we give a procedure to augment $A$ that does not reduce to $m$-fold application of the 1-augmentation. Before that, we investigate normality preserving normal perturbations. Apart from being of interest on its own, we shall make use of some of the conclusions obtained here in Section 3.4.
3.3 Normality preserving normal perturbations

In this section we consider a question related to the augmentation problem, and we study it using the same techniques as the ones from the previous sections. In particular, we investigate normality preserving $\theta$-Hermitian perturbations. These type of perturbations play a role also in the augmentation problem of Section 3.4.1.

Normality preserving normal perturbation. Let $A \in \mathbb{C}^{n \times n}$ be normal. Characterize all normal $E$ such that $A^* = A + E$ is normal. In other words, characterize the normality preserving normal perturbations $E$ of $A$.

Remark 3.3.1. Recall that any matrix $A$ may be written as the sum of two normal matrices, for instance, as $A = \Theta(A) + \Theta^\perp(A)$, where both $\Theta(A)$ and $\Theta^\perp(A)$ are normal. This illustrates why the problem above is non-trivial: the sum of normal matrices can be, literally, any matrix.

We begin by formulating a multi-functional lemma summarizing the technicalities of writing out commutators of linear combinations of matrices.

Lemma 3.3.2. Let $A, E \in \mathbb{C}^{n \times n}$ and $\gamma, \mu \in \mathbb{C}$. Then, with $\theta = \frac{\gamma \mu}{|\gamma \mu|}$,

$$[\gamma A + \mu E, (\gamma A + \mu E)^*] = |\gamma|^2 [A, A^*] + 2 \gamma \mu \Theta ([A, E]) + |\mu|^2 [E, E^*], \quad (3.8)$$

and

$$\Theta ([A, E^*]) = \Theta(A), \Theta^\perp(E^*) \quad (3.9)$$

Therefore, if $A$ and $E$ are normal, then $\gamma A + \mu E$ is normal if and only if

$$\Theta ([A, E^*]) = O_n, \quad (3.10)$$

or, in other words, if and only if $[A, E^*]$ is $\gamma \mu$-skew-Hermitian.

Proof. The statements are obtained from straight-forward manipulations with the commutator.

Corollary 3.3.3. Let $A, E \in \mathbb{C}^{n \times n}$ be normal. Then $A + E$ is normal if and only if $\gamma A + \mu E + \alpha I$ is normal for all $\gamma, \mu, \alpha \in \mathbb{C}$ with the restriction that $\gamma \mu \in \mathbb{R}$.
Corollary 3.3.4. If $A, E \in \mathbb{C}^{n \times n}$ are normal then $[A, E^*] = O_n \iff [A, E] = O_n$. Thus, if either term vanishes, both $A + E^*$ and $A + E$ are normal.

Proof. As was shown in the proof of Lemma 3.1.1 $E$ and $E^*$ have the same eigenvectors. Thus, $A$ and $E^*$ are simultaneously unitarily diagonalizable if and only if $A$ and $E$ are. Lemma 3.1.7 now proves that $[A, E^*] = O_n \iff [A, E] = O_n$, and Lemma 3.3.2 proves the conclusion.

Corollary 3.3.5. The matrix $\gamma A + \mu E$ with $\gamma, \mu \in \mathbb{C}$ and $A$ and $E$ Hermitian is normal if and only if $\gamma \mu \in \mathbb{R}$ or $[A, E] = O_n$.

Proof. The commutator of Hermitian matrices is always skew-Hermitian. Thus, for $[A, E]$ to be $\theta$-Hermitian in Equation (3.10), $\gamma \mu$ must be real, or $[A, E]$ should vanish.

3.3.1 Normality preserving normal rank-one perturbations

This section aims to show the similarities between the normality preserving normal rank-one perturbation problem and the $m$-augmentation problem of Section 3.2.1. Indeed, let $E = vw^*$ with $v, w \in \mathbb{C}^n$. Then $E$ is normal if and only if

$$\|w\|^2 vv^* = \|v\|^2 ww^*,$$

and thus if and only if $v = zw$ for some $z \in \mathbb{C}$. Write $z = \theta \rho$ with $\theta \in \mathbb{T}$ and $0 \leq \rho \in \mathbb{R}$. This shows that a rank-one matrix $E$ is normal if and only if $E$ is $\theta$-Hermitian,

$$E = \theta uu^*, \quad \theta \in \mathbb{T}.$$

With $A \in \mathbb{C}^{n \times n}$ normal, we look for the conditions on $u \in \mathbb{C}^n$ and $\theta \in \mathbb{T}$ such that $A + \theta uu^*$ is normal. Since $\theta uu^*$ is $\theta$-Hermitian, from Equations (3.8) and (3.9) in Lemma 3.3.2 we obtain

$$[(A + \theta uu^*), (A + \theta uu^*)^*] = 2\theta [\Theta^\perp(A), uu^*].$$

Therefore, $A + \theta uu^*$ is normal if and only if $\Theta^\perp(A)$ and $uu^*$ commute. According to Lemma 3.1.7 this is true if and only if they are simultaneously unitarily diagonalizable.
diagonalizable. For this, it is necessary and sufficient that \( u \) be an eigenvector of \( \Theta^\perp(A) \). As in Theorem 3.2.2 we formulate this result constructively in terms of the eigendata of \( A \).

**Theorem 3.3.6.** Let \( A \) be normal and let \( \gamma \in \mathbb{C} \) and \( \ell : \{ \gamma + \rho \theta \mid \rho \in \mathbb{R} \} \) be a line in \( \mathbb{C} \) through \( \gamma \) with slope \( \theta \in \mathbb{T} \). Let \( \lambda_1, \ldots, \lambda_p \) be the eigenvalues of \( A \) that lie on \( \ell \). Then, the matrix

\[
A^+ = A + \theta uu^*
\]

is normal if and only if \( u \) is a linear combination of eigenvectors corresponding to \( \lambda_1, \ldots, \lambda_p \), with the convention that \( u = \vec{0} \) if \( p = 0 \).

**Proof.** Corollary 3.1.2 shows the relation between the eigendata of \( \Theta^\perp(A) \) and \( A \), and together with the derivation above this proves the statement. \( \square \)

**Remark 3.3.7.** An interesting consequence of adding the normality preserving normal rank-one perturbation \( E = \theta uu^* \) is that

\[
\Theta^\perp(A + \theta uu^*) = \Theta^\perp(A) + \Theta^\perp(\theta uu^*) = \Theta^\perp(A),
\]

because \( \theta uu^* \) is \( \theta \)-Hermitian. Thus, the conditions under which adding another \( \theta \)-Hermitian normal rank-one perturbation \( F = \theta ww^* \) to \( A + E \) lead to a normal \( A + E + F \) are identical to the conditions just described for \( E \). We shall get back to this observation in Section 3.3.3.

### 3.3.2 Eigenvalues of the perturbed matrix

To study the eigenvalues of \( A^+ \) in relation to those of \( A \), let \( \Lambda_p \in \mathbb{C}^{p \times p} \) be the diagonal matrix whose eigenvalues are the \( p \) eigenvalues of \( A \) that lie on the line \( \ell : \{ \gamma + \rho \theta \mid \rho \in \mathbb{R} \} \). Then

\[
\Lambda_p = \theta R + \gamma I_p
\]

for some real diagonal matrix \( R \). Let \( U \in \mathbb{C}^{n \times n} \) be any unitary matrix whose first \( p \) columns are eigenvectors of \( A \) belonging to \( \lambda_1, \ldots, \lambda_p \). Then, assuming that \( A \)
and $A^+$ are normal, Theorem 3.3.6 shows that

$$U^*A^+U = U^*(A + \theta uu^*)U = \begin{bmatrix} \Lambda_p \\ B \end{bmatrix} + \begin{bmatrix} \theta rr^* \\ O_{n-p} \end{bmatrix}$$

because $u$ is a linear combination of the first $p$ columns of $U$. This leads to the following theorem, in which we summarize the above analysis.

**Theorem 3.3.8.** The only normality preserving normal rank-one perturbations of $A$ are the ones that, on an orthonormal basis of eigenvectors of $A$, are $\theta$-Hermitian rank-one perturbations of a $p \times p$, $\theta$-Hermitian submatrix. Hence, $n-p$ eigenvalues of $A$ are also eigenvalues of $A^+$. The remaining $p$ eigenvalues of $A^+$ are the eigenvalues of

$$\Lambda_p + \theta rr^* = \theta (R + rr^*) + \gamma I_p. \quad (3.11)$$

These interlace the $p$ eigenvalues of $A$ on $\ell$ with the additional $(p+1)$st point $+\infty \theta$.

**Proof.** The eigenvalues of $A^+$ are the eigenvalues of $B$ together with the eigenvalues of the matrix in Equation (3.11). Obviously, all eigenvalues of $B$ are eigenvalues of $A$ as well. Since $rr^*$ is a positive semi-definite rank-one perturbation of $R$, the eigenvalues $\rho_1 \leq \ldots \leq \rho_p$ of $R + rr^*$ and the eigenvalues $r_1 \leq \ldots \leq r_p$ of $R$ satisfy

$$r_1 \leq \rho_1 \leq r_2 \leq \ldots \leq \rho_{p-1} \leq r_p \leq \rho_p$$

as a result of Weyl’s Theorem (Theorem 1.1.7). Multiplying by $\theta$ and shifting over $\gamma$ yields the proof. □

**Remark 3.3.9.** Note that if $p = 1$, only one eigenvalue is perturbed, and we have $\rho_1 = r_1 + \|r\|^2_2$. In terms of the original perturbation $E = \theta uu^*$ this becomes $\tilde{\lambda} = \lambda + \|u\|^2_2$, where $\lambda$ is the eigenvalue of $A$ belonging to the eigenvector $u$ (see also Corollary 1.1.8).

As a consequence of the following theorem, it is possible to indicate where the eigenvalues of the family of matrices $A + tE$ are located. This can only be done for *normal* normality preserving perturbations.
Theorem 3.3.10. Let $A, B \in \mathbb{C}^{n \times n}$ be normal. Consider the line $\ell$ through $A$ and $B$,

$$\ell : \mathbb{R} \to \mathbb{C}^{n \times n} : t \mapsto tA + (1 - t)B.$$ 

If $E = B - A$ is normal, all matrices on $\ell$ are normal; if $E$ is not normal, $A$ and $B$ are the only normal matrices on $\ell$.

Proof. Observe that $\ell(t) = A + (1 - t)E$. If $E$ is normal, then Corollary 3.3.3 shows that all matrices $\gamma A + \mu E$ with $\gamma, \mu \in \mathbb{R}$ are normal, which includes the line $\ell$. Assume now that $E$ is not normal. Because $A$ and $B = A + E$ both are normal, Equation (3.3.2) in Lemma 3.8 gives that

$$(1 - t) \left( 2\mathcal{H}([A, E^*]) + (1 - t)[E, E^*] \right) = O_n \quad \text{with} \quad [E, E^*] \neq O_n.$$ 

The solution $t = 1$ confirms the normality of $A$, and the linear matrix equation

$$2\mathcal{H}([A, E^*]) + (1 - t)[E, E^*] = O_n \quad \text{with} \quad [E, E^*] \neq O_n$$

allows at most one solution in $t$ which, by assumption, is $t = 0$.

Thus, any line in $\mathbb{C}^{n \times n}$ parametrized by a real variable that does not lie entirely in the set of normal matrices, contains at most two normal matrices.

Remark 3.3.11. Let $A$ be normal. Lemma 3.1.1 shows that if $E$ is such that $A + E$ is normal, then

$$\sigma(A + E) \subset \sigma \left( \Theta(A) + \Theta(E) \right) \times \sigma \left( \Theta^\perp(A) + \Theta^\perp(E) \right).$$

and perturbation theory for $\theta$-Hermitian matrices can be used to derive statements about the eigenvalues of $A + E$. According to Theorem 3.3.10, if $E$ itself is normal too, this relation is valid continuously in $t$ along the line $A + tE$:

$$\sigma(A + tE) \subset \sigma \left( \Theta(A) + t\Theta(E) \right) \times \sigma \left( \Theta^\perp(A) + t\Theta^\perp(E) \right).$$

For non-normal $E$ this is, generally, not true, as is illustrated in Section 3.5.2.
Corollary 3.3.12. Let \( A \) be normal. As a result of Theorem 3.3.10 and Remark 3.3.11, the perturbed eigenvalues of

\[ t \mapsto A + t\theta uu^*, \quad 0 \leq t \leq 1 \]

seen as functions of \( t \), form line segments that all lie on the same line with slope \( \theta \).

3.3.3 \( \theta \)-Hermitian rank-\( k \) perturbations of normal matrices

Consider for given \( k \) with \( 1 \leq k \leq n \) the \( \theta \)-Hermitian rank-\( k \) matrix

\[ E = \theta H, \quad \text{where} \quad \theta \in \mathbb{T}_U \quad \text{and} \quad H = H^*. \tag{3.12} \]

Let \( A \) be normal. Then, since \( \Theta^+(E^*) = O_n \), Lemma 3.3.2 shows that \( A + E \) is normal if and only if

\[ [\Theta^+(A), H] = O_n. \tag{3.13} \]

By Lemma 3.1.7 this is equivalent to \( \Theta^+(A) \) and \( H \) being simultaneously diagonalizable by a unitary transformation \( U \). Thus, \( H \) needs to be of the form

\[ H = U\Delta U^* \tag{3.14} \]

where \( \Delta \in \mathbb{R}^{k \times k} \) is diagonal with diagonal entries \( \delta_1, \ldots, \delta_k \) and the columns \( u_1, \ldots, u_k \) of \( U \) are orthonormal eigenvectors of \( \Theta^+(A) \). But then, writing

\[ E = E_1 + \ldots + E_k, \quad \text{where for all} \quad j \in \{1, \ldots, k\}, \quad E_j = \theta \delta_j u_j u_j^*, \]

the observation in Remark 3.3.7 reveals that perturbing \( A \) by \( E \) is equivalent to perturbing \( A \) consecutively by the rank-one matrices \( E_1, \ldots, E_k \). The above analysis is summarized in the following theorem. An illustration of this theorem is provided in Section 3.5.2.

Theorem 3.3.13. Let \( E = \theta H \) be a \( \theta \)-Hermitian rank-\( k \) perturbation of a normal
matrix $A$. Then $E$ is normality preserving if and only if $E$ can be decomposed as

$$E = E_1 + \ldots + E_k,$$

where $E_1, \ldots, E_k$ are all normality preserving $\theta$-Hermitian rank-one perturbations of $A$. In fact, for each permutation $\sigma$ of $\{1, \ldots, k\}$ and each $m \in \{1, \ldots, k\}$, the partial sum

$$A + \sum_{j=1}^{m} E_{\sigma(j)}$$

is normal, too.

**Remark 3.3.14.** In accordance with Remark 3.1.8, if the matrix $\Delta$ in Equation (3.14) has multiple eigenvalues, there exist non-diagonal unitary matrices $Q$ such that $\Delta = Q\Delta Q^\ast$. As a result, $H$ can be written as $Z\Delta Z^\ast$ where the orthonormal columns of $Z$ span an invariant subspace of $\Theta^\perp(A)$. This implicitly writes the perturbation $\theta H$ as the sum of rank-one normal perturbations that do not necessarily preserve normality. This aspect is also illustrated in Section 3.5.2.

**Theorem 3.3.15.** The only normality preserving $\theta$-Hermitian rank-$k$ perturbations of $A$ are the ones that, on an orthonormal basis of eigenvectors of $A$, are $\theta$-Hermitian perturbations of $\theta$-Hermitian submatrices of size $s_1 \times s_1, \ldots, s_m \times s_m$ of ranks $k_1, \ldots, k_m$, where $k_1 + \ldots + k_m = k$. As a result of this perturbation, at most $s_1 + \ldots + s_m$ eigenvalues of $A$ are perturbed, which are located on at most $m$ distinct parallel lines $\ell_1, \ldots, \ell_m$, defined by $\theta \in \mathbb{T}$ and $\gamma_1, \ldots, \gamma_m \in \mathbb{C}$ as

$$\ell_j : \{\gamma_j + \theta \rho \mid \rho \in \mathbb{R}\}.$$

Moreover, the eigenvalues of $A + tE$ with $t \in [0, 1]$ connect the eigenvalues of $A$ with those of $A^+$ by line segments that lie on $\ell_1, \ldots, \ell_m$.

**Proof.** Write $\theta H = \theta(\delta u_1 u_1^* + \ldots + \delta_k u_k u_k^*)$, where $u_1, \ldots, u_k$ are eigenvectors $\Theta^\perp(A)$, and repeatedly apply Theorem 3.3.8 The statement about the eigenvalues of $A + tE$ follows from Corollary 3.3.12.

For a qualitative illustration of the effect on the eigenvalues due to a rank-one $\theta$-Hermitian perturbation and a rank-$k$ $\theta$-Hermitian perturbation of a normal
matrix, see Figure 3.2. The asterisks in the pictures are eigenvalues of $A$, the circles represent different choices for $\gamma$, and the boxes are the perturbed eigenvalues.

Figure 3.2: Eigenvalue perturbation by a rank-one matrix (left) and a rank-$k$ matrix (right).

**Remark 3.3.16.** If $H$ in (3.12) is semi-definite, then the eigenvalues of $A+tE$ all move in the same direction over the parallel lines $\ell_1, \ldots, \ell_m$ from Theorem 3.3.15.

### 3.3.4 Normality preserving normal perturbations

The above analysis of $\theta$-Hermitian normality preserving perturbations also gives sufficient conditions for when normal perturbations $E$ of the form $E = \theta_1 H_1 + \theta_2 H_2$ with $H_1, H_2$ Hermitian and $\theta_1, \theta_2 \in \mathbb{T}$ are normality preserving in case $\theta_2 \neq \pm \theta_1$. Notice that in order for $E$ to be normal itself, $[H_1, H_2] = O_n$ by Corollary 3.3.5. Obviously, $E$ is, in general, not $\theta$-Hermitian for some value of $\theta$. Nevertheless, the following holds.

**Theorem 3.3.17.** Let $A$ be normal, $H_1, H_2$ Hermitian with $[H_1, H_2] = O_n$, and $\theta_1, \theta_2 \in \mathbb{T}$ with $\theta_1 \theta_2 \notin \mathbb{R}$. Then $E = E_1 + E_2 = \theta_1 H_1 + \theta_2 H_2$ is a normality preserving normal perturbation of $A$ if $E_1$ and $E_2$ both are normality preserving perturbations of $A$.

**Proof.** Corollary 3.3.5 covers the normality of $E$. Furthermore, assuming that $A + E_1$ is normal, Equation (3.13) gives that $A + E_1 + E_2$ is normal if and only if $[\Theta^\perp (A + E_1), H_2] = O_n$, where $\Theta^\perp$ refers to the $\theta_2$-skew-Hermitian part. But
then
\[
[\Theta^\perp(A + E_1), H_2] = [\Theta^\perp(A), H_2] + [\Theta^\perp(\theta_1 H_1), H_2] = [\Theta^\perp(A), H_2],
\]
because \([H_1, H_2] = O_n\), proving the statement.

**Remark 3.3.18.** A similar result holds for normal perturbations \(E_1 + \ldots + E_k\) where \(E_j = \theta_j H_j\) with \(H_j\) Hermitian and \(\theta_j \in \mathbb{T}\) for each \(j \in \{1, \ldots, k\}\). Moreover, each normal perturbation \(E\) can be written in this form in several different ways.

We are now ready to return to the augmentation problem of Section 3.3 and to study augmentations of normal matrices \(A\) that result in an augmented matrix \(A_+\) whose eigenvalues all lie on the graph of a quadratic polynomial.

### 3.4 Further augmentations

We now return to the \(m\)-augmentation problem of Section 3.3 and concentrate on the case \(m > 1\). If all eigenvalues of \(A_+\) lie on a line, then \(A_+\) is essentially Hermitian, a property that is inherited by principal submatrices. In that case it is clear which matrices \(A\) can be augmented into \(A_+\). The next simplest case is the case where all eigenvalues of \(A_+\) lie on a curve \(C\) that is the image of a quadratic function in a rotated complex plane.

#### 3.4.1 Matrices \(A_+\) with all eigenvalues on a quadratic curve

Assume that for \(\theta \in \mathbb{T}\) there exists a \(\varphi \in \mathcal{P}^2(\mathbb{R})\) such that

\[
A_+ = \Theta(A_+) + i \cdot \theta \varphi(\overline{\Theta}(A_+)), \quad A_+ = \begin{bmatrix} A & V \\ W^* & \Gamma \end{bmatrix}. \tag{3.15}
\]

Then, \(A_+\) is normal with all its eigenvalues on the rotated parabola \(C \subset \mathbb{C}\) defined as the image of \(q\), where

\[
q : \mathbb{R} \to \mathbb{C} : \rho \mapsto \theta \rho + i \cdot \theta \varphi(\rho)
\]
3.4. FURTHER AUGMENTATIONS

and

\[ \varphi(x) = r_0 + r_1 x + r_2 x^2 \quad \text{with} \quad r_0, r_1, r_2 \in \mathbb{R}. \]

**Remark 3.4.1.** Throughout this section we assume, without loss of generality, that \( r_2 > 0 \). The case \( r_2 = 0 \), as argued above, is trivial and concerns essentially Hermitian matrices, whereas the case \( r_2 < 0 \) can be avoided by replacing \( \theta \) by \(-\theta\), which is nothing else than a trivial change of coordinates that transforms the polynomial \( \varphi \) into \(-\varphi\).

The curve \( C \) now divides the complex plane \( \mathbb{C} \) in three disjoint parts

\[ C = C_+ \cup C \cup C_- \tag{3.16} \]

where \( C_+ \) is the open part of \( \mathbb{C} \) that lies on the one side of \( C \) that is convex.

**The principal submatrices of \( A_+ \) and their eigenvalues**

For convenience, write

\[ X = \overline{\vartheta} \Theta(A), \quad M = \overline{\vartheta} \Theta(\Gamma) \quad \text{and} \quad 2Z = \overline{\vartheta} V + \vartheta W; \]

then,

\[ \overline{\vartheta} \Theta(A_+) = \begin{bmatrix} X & Z \\ Z^* & M \end{bmatrix} \quad \text{and} \quad \left( \overline{\vartheta} \Theta(A_+) \right)^2 = \begin{bmatrix} X^2 + ZZ^* & XZ + ZM \\ Z^*X + MZ^* & M^2 + Z^*Z \end{bmatrix} \tag{3.17} \]

Thus, explicitly evaluating \( \varphi \) at \( A_+ \) using the block forms in (3.17), and comparing the result with the block form of \( A_+ \) displayed in (3.15) yields

\[ A = \Theta(A) + i \cdot \vartheta \varphi \left( \overline{\vartheta} \Theta(A) \right) + i \cdot \vartheta r_2 ZZ^* \tag{3.18} \]

and

\[ \Gamma = \Theta(\Gamma) + i \cdot \vartheta \varphi \left( \overline{\vartheta} \Theta(\Gamma) \right) + i \cdot \vartheta r_2 Z^*Z. \]

The results that follow will sometimes be stated for \( A \) only, even though similar statements obviously hold for \( \Gamma \). The first proposition simply translates Equation
Proposition 3.4.2. The $n \times n$ principal submatrix $A$ of $A_+$ is a $\theta$-skew-Hermitian rank-$k$ (with $k \leq \min(m, n)$) perturbation of a normal matrix that has all the eigenvalues on $C$.

Lemma 3.4.3. If $A$ in Equation (3.18) is normal then $\sigma(A) \subset C \cup C_+$. 

Proof. If $A$ is normal, then $i \cdot \theta r_2 ZZ^*$ is a normality preserving $\theta$-skew-Hermitian matrix perturbation of the normal matrix $\Theta(A) + i \cdot \theta \wp (\Theta(A))$ that has all its eigenvalues on $C$. By Theorem 3.3.15 each perturbed eigenvalue $\lambda \in C$ of $A$ moves along a line $\ell : \{\lambda + i \cdot \theta \rho \mid \rho \in \mathbb{R}\}$. Note that $\ell$ is vertical in the $\theta$-rotated complex plane. By Remark 3.3.16 and because $ZZ^*$ is positive semi-definite, the direction is the same for each perturbed eigenvalue and is determined by the sign of $r_2$. In Remark 3.4.1 we assumed that $r_2 > 0$, and thus the direction is directed into $C_+$ defined in (3.16). \hfill\Box

Remark 3.4.4. Note that a multiple eigenvalue $\lambda$ of $\Theta(A) + i \cdot \theta \wp (\Theta(A))$, located on $C$, may be perturbed by $i \cdot \theta r_2 ZZ^*$ into several distinct eigenvalues of $A$. Those will all be located on $\ell : \mathbb{R} \to C : \rho \mapsto \lambda + i \cdot \theta \rho$ with $\rho > 0$.

Corollary 3.4.5. Assume that $A$ in (3.18) is normal. Then $\sigma(A) \subset C$ if and only if $A_+$ is block diagonal with blocks $A$ and $\Gamma$.

Proof. If $Z \neq O$ then $\text{trace}(ZZ^*) \neq 0$ and at least one eigenvalue is perturbed. Lemma 3.4.3 shows that a perturbed eigenvalue cannot stay on $C$ and necessarily moves from $C$ into $C_+$. \hfill\Box

Augmentations with eigenvalues on a quadratic curve

We may now reverse the previous observations. Given $A$, we choose a parabolic curve $C$ and construct $Z \in C^{n \times m}$ such that $i \cdot \theta r_2 ZZ^*$ perturbs the eigenvalues of $A$ onto $C$. We then use $C$ to define the corresponding $m$-augmentation $A_+$ of $A$.

Corollary 3.4.6. Necessary for a normal matrix $A$ to be $m$-augmentable into a normal matrix $A_+$ with all eigenvalues on a quadratic curve $C$ is that $\sigma(A) \subset C \cup C_+$.

Proof. This is just another corollary of Lemma 3.4.3. \hfill\Box
Clearly, for any given finite set of points in \( \mathbb{C} \), there are infinitely many candidates for such quadratic curves \( C \). It is the purpose of this section to show that each of this candidates can be used, and to construct essentially all possible corresponding augmentations \( A_+ \).

**Theorem 3.4.7.** Let \( A \in \mathbb{C}^{n \times n} \) be normal, and let \( \theta \in \mathbb{T} \) and \( \varphi \in \mathcal{P}^2(\mathbb{R}) \) be such that \( \sigma(A) \subset \mathcal{C} \cup \mathcal{C}_+ \) where \( \mathcal{C} \) is the graph of

\[
q : \mathbb{R} \to \mathbb{C} : \rho \mapsto \theta \rho + i \cdot \theta \varphi(\rho).
\]

Then there exist \( p \)-augmentations \( A_+ \) of \( A \) such that

\[
\sigma(A_+) \subset \mathcal{C},
\]

where \( p \) is the number of eigenvalues of \( A \) in \( \mathcal{C}_+ \).

**Proof.** Write \( \Lambda_p \in \mathbb{C}^{p \times p} \) for the diagonal matrix with precisely the eigenvalues \( \lambda_1, \ldots, \lambda_p \) of \( A \) that do not lie on \( \mathcal{C} \) and let \( U_p \in \mathbb{C}^{n \times p} \) have corresponding orthonormal eigenvectors \( u_1, \ldots, u_p \) as columns. Since \( \sigma(\Lambda_p) \subset \mathcal{C}_+ \), for each \( j \in \{1, \ldots, p\} \) there exists a positive real number \( \psi_j \) such that

\[
(\lambda_j - i \cdot \theta \psi_j) \in \mathcal{C}.
\]

Write \( \Psi \in \mathbb{C}^{p \times p} \) for the diagonal matrix with \( \sigma(\Psi) = \{\sqrt{\psi_1}, \ldots, \sqrt{\psi_p}\} \) and set

\[
Z = U_p \Psi.
\]

By Lemma 3.1.1 the columns of \( U_p \) are also eigenvectors of \( \Theta(A) \) and thus \( i \cdot \theta ZZ^* \) is a \( \theta \)-skew-Hermitian normality preserving perturbation of \( A \). Moreover,

\[
u^*_j (A - i \cdot \theta ZZ^*) u_j = (\lambda - i \cdot \theta \psi) \in \mathcal{C}.
\]

Because the matrix \( A - i \cdot \theta ZZ^* \) is normal with all eigenvalues in \( \mathcal{C} \), the equality
\( \Theta(A - i \cdot \theta ZZ^*) = \Theta(A) \) leads to

\[
A - i \cdot \theta ZZ^* = \tilde{\theta} \Theta(A) + i \cdot \theta \varphi \left( \tilde{\theta} \Theta(A) \right).
\]

The assumption \( r_2 > 0 \), justified in Remark 3.4.1, now gives that

\[
A = \tilde{\theta} \Theta(A) + i \cdot \theta \varphi \left( \tilde{\theta} \Theta(A) \right) + i \cdot \theta r_2 \left( \frac{Z}{\sqrt{r_2}} \right) \left( \frac{Z}{\sqrt{r_2}} \right)^*,
\]

and according to Equation (3.18) this is precisely the \( n \times n \) leading principal submatrix of \( A_+ \), where \( A_+ \) is defined as \( A_+ = \tilde{\theta} H + i \cdot \theta \varphi(H) \), where

\[
H = \begin{bmatrix}
\tilde{\theta} \Theta(A) & \hat{Z} \\
\hat{Z}^* & M
\end{bmatrix}, \quad \text{with} \quad \hat{Z} = \frac{ZQ}{\sqrt{r_2}} \quad (3.19)
\]

and \( M, Q \in \mathbb{C}^{m \times m} \) are arbitrary Hermitian and unitary matrices.

For a given \( n \times n \) normal matrix \( A \), the typical situation is that after selecting \( \theta \) suitably, at least \( p = n - 3 \) of its eigenvalues do not lie on a quadratic curve, and a matrix \( Z \) of rank \( p \) is needed to push those outliers onto \( \mathbb{C} \). This is illustrated in Figure 3.3.

**Remark 3.4.8.** It is, of course, possible to move each eigenvalue of \( A \) from \( \mathbb{C}_+ \) onto \( \mathbb{C} \) as a result of an arbitrary amount of rank-one perturbations. This would increase the number of columns of \( Z \), and give \( m \)-augmentations of \( A \) with \( m > p \). However, this would not increase the rank of \( Z \), and \( ZZ^* \) would remain the same. Together with the analysis of Section 3.3.3 that shows which \( \theta \)-Hermitian perturbations are normality preserving, this shows that in essence, each \( p \)-augmentation \( A_+ \) with \( Z \) of full rank, is of the form (3.19). In Section 3.5.3 we give an explicit example of the construction in the proof of Theorem 3.4.7.

**Augmentations without computing eigenvalues**

So far, explicit knowledge about the eigenvalues and eigenvectors of \( A \) was used to construct augmentations \( A_+ \). There are, however, cases in which it is sufficient to know the polynomial curve \( \mathcal{E} \) on which the eigenvalues of \( A \) lie. To see this,
assume that $A$ is normal and

$$A = \Theta(A) + i \cdot \theta \varphi \left( \overline{\Theta}(A) \right), \quad \varphi \in \mathcal{P}^{2k}(\mathbb{R}) \quad \text{and} \quad \varphi \notin \mathcal{P}^{2k-1}(\mathbb{R})$$

for some integer $k \geq 1$. Since $\varphi$ has even degree, there exist polynomials $p \in \mathcal{P}^2(\mathbb{R})$ such that

$$\varphi(x) - p(x) \geq 0 \quad \text{for all} \quad x \in \mathbb{R}.$$ 

This implies that the matrix $(\varphi - p) \left( \overline{\Theta}(A) \right)$, is positive semi-definite and, hence, it can be factorized as

$$(\varphi - p) \left( \overline{\Theta}(A) \right) = ZZ^*, \quad (3.20)$$

after which we have that

$$A = \Theta(A) + i \cdot \theta p \left( \overline{\Theta}(A) \right) + i \cdot \theta ZZ^*.$$
It is trivial that the matrix $i \cdot \theta ZZ^*$ is a normality preserving perturbation of $\Theta(A) + i \cdot \theta (\overline{\Theta}(A))$ and by choosing between $\theta$ and $-\theta$, as explained in Remark 3.4.1, this leads to an $m$-augmentation of $A$, with generally $m = n - 1$. Section 3.1.1 explained how $\varphi$ can be computed in a finite number of arithmetic operations, and the same is valid for the factorization (3.20). Of course, the problem of finding a minorizing polynomial $p \in P(\mathbb{R})$ may prove to be difficult in specific situations.

### 3.4.2 Polynomial curves of higher degree

If one tries to generalize the approach of Section 3.4.1 to polynomial curves $\mathcal{C}$ of higher degree, the situation rapidly becomes more difficult. As an illustration, consider the cubic case. The third power of the matrix $\overline{\Theta}(A_+)$ in (3.17) equals

$$
\begin{bmatrix}
X^3 + XZZ^* + ZZ^*X + ZMZ^* & X^2Z + XZM + ZM^2 + ZZ^*Z \\
Z^*X^2 + Z^*ZZ + MZ^*X + M^2Z & M^3 + Z^*ZM + MZ^*Z + Z^*XZ
\end{bmatrix},
$$

and thus, comparing the leading principal submatrices,

$$A = \Theta(A) + i \cdot \theta \varphi(\overline{\Theta}(A)) + i \cdot \theta r_2 ZZ^* + i \cdot \theta r_3 (XZZ^* + ZZ^*X + ZMZ^*),$$

where $r_3$ is the coefficient of $x^3$ of $\varphi$. Thus, $A$ is a rank-$k$ (with $k \leq 2m$), $\theta$-skew-Hermitian perturbation of the normal matrix $\Theta(A) + i \cdot \theta (\overline{\Theta}(A))$. Of course, if $ZZ^*$ commutes with $\Theta(A)$, then it commutes with $X$, and this may help the analysis. However, it becomes much harder to control the perturbation in such a way, that $A$ will be augmented into a matrix $A_+$ with $\sigma(A_+) \subset \mathcal{C}$. Therefore, we will not pursue this idea any further.

### 3.5 Illustrations

In this section we present some illustrations of the main constructions and theorem of this chapter. By making them explicit, we hope to create more insight in their structure.
3.5. ILLUSTRATIONS

3.5.1 Illustrations belonging to Section 3.2

This example, illustrates Theorem 3.2.2 and Theorem 3.2.3. Let $A$ be the matrix

$$A = \begin{bmatrix}
  2i & 2 + i \\
  2 + i & -3
\end{bmatrix}.$$ 

Take $\gamma = 1$ and choose $\ell$ the line through $\gamma = 1$ and the eigenvalue $2 + i$ of $A$,

$$\ell : \mathbb{R} \to \mathbb{C} : \rho \mapsto 1 + \rho \theta, \quad \text{with} \quad \theta = e^{i \frac{\pi}{4}} = \frac{1 + i}{\sqrt{2}}.$$ 

Thus, in order for $A_+$ to be normal, $u$ must be a multiple of $e_2$, $v = \theta u$ and $w^* = \theta u^*$, which yields that for all $\mu \in \mathbb{C}$,

$$A_+ = \begin{bmatrix}
  2i & 2 + i & \theta \mu \\
  2 + i & -3 & \theta \mu \\
  \theta \mu & \theta \mu & 1
\end{bmatrix}$$

is a normal augmentation of $A$. Moreover, for the choice $\gamma = 1$ and $\ell$, these are all the normal augmentations of $A$. The eigenvalues of $A_+$ that are not eigenvalues of $A$ are the eigenvalues of

$$\begin{bmatrix}
  2 + i & \theta \mu \\
  \theta \mu & 1
\end{bmatrix} = \begin{bmatrix}
  1 + \theta \sqrt{2} & \theta \mu \\
  \theta \mu & 1
\end{bmatrix} = \theta \begin{bmatrix}
  \sqrt{2} & \mu \\
  \mu & 0
\end{bmatrix} + I,$$

and thus equal to

$$\lambda = \theta \left( \frac{\sqrt{2}}{2} \pm \sqrt{2} \mu^2 + 1 \right) + 1,$$

which lie on $\ell$ and have the eigenvalue $2 + i$ (that was perturbed) as average, as is depicted in the left in Figure 3.4. Here, the stars represent eigenvalues of $A$, the circles depict different choices for $\gamma$ and the squares indicate the perturbed eigenvalues.
A second option is to choose \( \gamma = -2 + 3i \) instead of \( \gamma = 1 \). This gives other possibilities to construct normality preserving augmentations. The first one is to choose \( \gamma = -2 + 3i \) instead of \( \gamma = 1 \). This gives other possibilities to construct normality preserving augmentations. The first one is to choose \( \ell \) through \( \gamma \) and \( -3 \), which shows that \( u \) must be a multiple of \( e_3 \) and \( v = \theta u \) and \( w^* = \theta u^* \) where \( \theta = (1 + 3i)/\sqrt{10} \), which is a similar situation as for \( \gamma = 1 \). The second non-trivial option is to choose \( \ell \) through \( \gamma \) and both \( 2i \) and \( 2 + i \). Then with \( \theta = (-2 + i)/\sqrt{5} \), we may take \( u \) as a linear combination of \( e_1 \) and \( e_2 \), showing that

\[
A_+ = \begin{bmatrix}
2i & \theta \alpha \\
2 + i & \theta \mu \\
\theta \alpha & 0 \\
\theta \mu & -3
\end{bmatrix}.
\]

is normal for all \( \alpha, \mu \in \mathbb{C} \). For the given value of \( \gamma \) those two options are the only possible normality preserving augmentations. The perturbed eigenvalues are depicted in the right picture in Figure 3.4.

Figure 3.4: Normality preserving augmentations of a 3 \( \times \) 3 matrix for two different values of \( \gamma \).
3.5. ILLUSTRATIONS

3.5.2 Illustrations belonging to Section 3.3

Normality preserving non-normal perturbations

First we illustrate Theorem 3.3.10 by presenting an example of a line through two normal matrices that contains only two normal matrices. For this, let

\[ A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \text{and} \quad E = B - A = \begin{bmatrix} 0 & 0 \\ -2 & 0 \end{bmatrix}. \]

Thus, \( E \) is a non-normal normality preserving perturbation of \( A \). Hence, apart from \( A \) and \( B \), no other matrix of the form \( A + tE \) with \( t \in \mathbb{R} \) is normal. Indeed,

\[ [A + tE, (A + tE)^*] = \begin{bmatrix} 1 - (1 - 2t)^2 & 0 \\ 0 & (1 - 2t)^2 - 1 \end{bmatrix}, \]

and this matrix is only zero for \( t = 0 \) and \( t = 1 \). Moreover, the eigenvalues of \( A + tE \) are

\[ \sqrt{1 - 2t} \quad \text{and} \quad -\sqrt{1 - 2t} \]

whereas the corresponding sums of the eigenvalues of \( \mathcal{H}(A + tE) \) and \( \mathcal{S}(A + tE) \) equal

\[ (1 - t) + it \quad \text{and} \quad -(1 - t) - it. \]

Thus, for instance at \( t = \frac{1}{2} \), the eigenvalue zero of \( A + tE \) is not the sum of the eigenvalues of the the Hermitian and skew-Hermitian parts of \( A + tE \).

Normality preserving normal perturbations

Now we illustrate Theorem 3.3.13. This concerns normality preserving normal perturbations. For this, we take \( \theta = 1 \) and consider the matrix \( A + E \), where

\[ A = \begin{bmatrix} 1 \\ i \\ 1 + i \end{bmatrix} \quad \text{and} \quad E = u_1 u_1^* + 2u_2 u_2^*. \]
with \( u_1, u_2 \in \mathbb{C}^3 \) mutually orthonormal. Then \( A + E \) is normal if and only of both \( u_1 \) and \( u_2 \) are eigenvectors of

\[
2\Theta^\perp(A) = 2S_A = \begin{bmatrix} 0 & i \\ i & i \end{bmatrix}.
\]

However, \( A + E \) where \( E = u_1 u_1^* + u_2 u_2^* \) with \( u_1, u_2 \in \mathbb{C}^3 \) mutually orthonormal, is normal if and only if both \( u_1 \) and \( u_2 \) are linear combinations of the same two eigenvectors \( v_1 \) and \( v_2 \) of \( 2\Theta^\perp(A) \). Thus, with

\[
\begin{align*}
u_1 &= \frac{1}{4} \begin{bmatrix} \sqrt{2} \\ 1 \\ 1 \end{bmatrix}, \quad &u_2 &= \frac{1}{4} \begin{bmatrix} -\sqrt{2} \\ 1 \\ 1 \end{bmatrix}, \quad &v_1 &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad &v_2 &= \begin{bmatrix} 0 \\ \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{bmatrix},
\end{align*}
\]

we have that

\[
E_1 + E_2 = \frac{1}{16} \begin{bmatrix} 2 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & 1 & 1 \\ \sqrt{2} & 1 & 1 \end{bmatrix} + \frac{1}{16} \begin{bmatrix} 2 & -\sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & 1 & 1 \\ -\sqrt{2} & 1 & 1 \end{bmatrix} = \frac{1}{16} \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}
\]

is a normality preserving rank-two perturbation of \( A \), written as the sum of two rank-one perturbations that individually do not preserve normality. However, we also have that

\[
\frac{1}{4} v_1 v_1^* + \frac{1}{2} v_2 v_2^* = \frac{1}{16} \begin{bmatrix} 4 & 0 \\ 0 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix} + \frac{1}{16} \begin{bmatrix} 0 & 2 & 2 \\ 2 & 2 \\ 2 & 2 & 2 \end{bmatrix} = \frac{1}{16} \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}
\]

and this expresses the perturbation as a sum of two normality preserving normal rank-one perturbations. The eigenvalues of \( A + tE \) are \( 1 + \frac{1}{4}t \) due to the term \( \frac{1}{4} v_1 v_1^* \), together with the eigenvalues of

\[
\begin{bmatrix} i & 1 \\ 1 + i & 1 \end{bmatrix} + \frac{1}{8} t \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \text{which are} \quad i + \frac{1}{2} \left( \frac{t}{4} \pm \sqrt{1 + \frac{t^2}{16}} \right),
\]
which are due to the term \( \frac{1}{4} v_2 v_2^* \). As stated in Theorem 3.3.15, the rank-2 perturbation moves the eigenvalues of \( A \) in the horizontal direction. Since the eigenvalues \( i \) and \( 1 + i \) of \( A \) are on the same horizontal line, they can be simultaneously perturbed by a rank-one perturbation. For \( t \in [0, 4] \) those eigenvalues are plotted by circles in Figure 3.5. We also computed the eigenvalues of \( A + tE_1 \) and of \( A + 4E_1 + tE_2 \) for \( t \in [0, 4] \) and indicated them by asterisks and boxes, respectively. As is visible in Figure 3.5, the eigenvalues leave the straight line before returning to the eigenvalues of the normal matrix \( A + 4E_1 + 4E_2 = A + 4E \).

Figure 3.5: Eigenvalue trajectories of a normality preserving perturbation, and of the same perturbation written as the sum of non-normality preserving normal perturbations.

### 3.5.3 Illustrations belonging to Section 3.4

We now illustrate Theorem 3.4.7. The starting point is a \( 3 \times 3 \) matrix \( A \), chosen so that the conditions of the theorem are easy to satisfy. Let

\[
A = \begin{bmatrix}
5i \\
1 \\
2 + 2i
\end{bmatrix}
\]

and thus, \( \mathcal{H}(A) = X = \begin{bmatrix} 0 & 1 \\ & \end{bmatrix} \).
With $\theta = 1$, the eigenvalues already lie on a parabolic curve, and thus also with the trivial choice $Z = O_n$ augmentations $A_+$ can be constructed having eigenvalues on the same curve. More interesting is to choose a $Z \neq O_n$ such that $A - iZZ^*$ is normal. Since $\mathcal{H}(A)$ has distinct eigenvalues, $Z$ needs to have eigenvectors of $\mathcal{H}(A)$ as columns. Take for example

$$Z = \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \text{ hence } ZZ^* = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } A - iZZ^* = \begin{bmatrix} i \\ 1 \\ 2 + i \end{bmatrix}.$$ 

The eigenvalues of $A - iZZ^*$ lie on the curve $C$ that is the image of

$$q : \mathbb{R} \to \mathbb{C} : \rho \mapsto \rho + i(1 - \rho)^2.$$ 

Augmentation $A_+$ can now be constructed by choosing an arbitrary Hermitian $2 \times 2$ matrix $M$ and an arbitrary unitary matrix $Q$, for instance

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \text{ and } Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix},$$

and then to form the Hermitian part of $A_+$ as

$$\mathcal{H}(A_+) = \begin{bmatrix} Z & ZQ \\ Q^*Z^* & M \end{bmatrix} = \begin{bmatrix} 0 & \sqrt{2} & \sqrt{2} \\ 1 & 0 & 0 \\ \sqrt{2} & 0 & 1 \\ \sqrt{2} & \frac{1}{2} \sqrt{2} & 1 \\ \sqrt{2} & 0 & 1 \\ 1 & 0 & 2 \end{bmatrix}.$$ 

Finally, $A_+$ itself can be formed as $A_+ = \mathcal{H}(A) + iq(\mathcal{H}(A)) = \mathcal{H}(A) + i(I - \mathcal{H}(A))^2$, resulting in
Indeed, $A_+$ is a 2-augmentation of $A$. From this example we observe that if $Z, M, Q$ are chosen real, then $\mathcal{H}(A_+)$ is real symmetric and $A_+$ complex symmetric, being the sum of a real symmetric matrix and $i$ times a polynomial of this real symmetric matrix. Note that not all complex symmetric matrices are normal. In fact, the leading $4 \times 4$ principal submatrix of $A_+$ in the above example is not normal, nor is the trailing $2 \times 2$ principal submatrix. Thus $A_+$ could not have been constructed using the procedure for $m = 1$ twice.
Part II

Iterative methods
Chapter 4

The Subspace Projected Approximate Matrix method

The Subspace Projected Approximate Matrix method [69] is a fairly recent (2001) and unobserved method for the computation of eigenvalues of large Hermitian matrices. It is known by the acronym of SPAM and is interpreted by its inventors as a modification of Davidson’s method [22]. Its most interesting feature is the way it proposes to reduce the costs of matrix-vector products with the original matrix by using an approximation $A_0$, acting only on a part of the space.

In this chapter, we extend the work initiated in [65] and investigate the Subspace Projected Approximated Matrix method in more detail. We begin with a brief contextualization of SPAM within the class of iterative methods for eigenvalue problems. Next, we show how, for certain special choices of the approximation $A_0$, SPAM turns out to be mathematically equivalent to known eigenvalue methods. More sophisticated approximations $A_0$ turn SPAM into a boosted version of the Lanczos method [49], although, it can also be interpreted as an attempt to enhance a certain instance of the preconditioned Jacobi-Davidson method [72]. The chapter ends with a set of numerical experiments tailored to illustrate certain aspects of SPAM and its variations.
4.1 Introduction

The Subspace Projected Approximate Matrix (SPAM) method belongs to the category of inner-outer iteration methods and is interpreted by its inventors as a modification of the Davidson method [22]. It is based on the following observation. Even when sparse, the computational effort necessary to carry out a matrix-vector multiplication with $A$ can be significant and often represents the bottleneck of the total computational effort. SPAM reduces the costs of matrix-vector products with $A$ by replacing its action within the inner iteration of the algorithm with a sparser or more structured approximation. By doing so, it attempts to slash the overall computational cost.

The idea is not altogether new and is related to a certain type of preconditioning, called one-step approximation in the Jacobi-Davidson method [72, Section 4.1]. There too, the matrix $A$ is, in the inner iteration, replaced by a preconditioner. The originality of the approach in [69] lies in the fact that the action of the preconditioner is only applied to the subspace in which the action of $A$ has not yet been computed in the outer iteration of the method. The resulting approximation is then different after each outer iteration step, even though only one approximate matrix $A_0$ needs to be provided. Consequently, the approximated action of $A$ in the inner iterations becomes more and more accurate as the number of outer iterations increases. As such, SPAM may be interpreted as a discrete homotopy method. Intuitively, one would expect SPAM to outperform Jacobi-Davidson with one-step approximation.

The expansion vector (see line 15 in Algorithm 1) is itself a suitable eigenvector of the approximation of $A$. Thus, the matrix-vector products within the inner iteration are cheaper than for instance in the Jacobi-Davidson [72] and Riccati [9] methods. These methods, which we explain in more detail in Section 4.2, both use $A$ itself in their inner iteration.
4.1. INTRODUCTION

4.1.1 Preliminaries

Before beginning, we assume that $A$ is a Hermitian $n \times n$ matrix with eigenvalues

$$\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{n-1} \leq \lambda_n = \lambda_{\max}. \quad (4.1)$$

Moreover, we recall from Section 1.2.1 that for the unitary matrix $(V|V_{\perp})$, with the columns of $V$ forming an orthonormal basis for the current search space $V$, we have

$$\hat{A} = (V|V_{\perp})^* A (V|V_{\perp}) = \begin{bmatrix} M & R^* \\ R & S \end{bmatrix} \quad (4.2)$$

where $M = V^* AV$ is $k \times k$ and $S = V_{\perp}^* AV_{\perp}$. Recall also that the eigenvalues of $M$ are denoted by $\mu_j$ for $j \in \{1, \ldots, k\}$ and satisfy

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_{k-1} \leq \mu_k. \quad (4.3)$$

4.1.2 General description of SPAM.

Let $A_0$ be an approximation of $A$. In Section 4.3 we will comment on how this approximation can be chosen. For now, assume only that $A_0$ is available and define, in view of (4.2)

$$\hat{S} = V_{\perp} A_0 V_{\perp}^* \quad (4.4)$$

and

$$A_k = (V|V_{\perp}) \hat{A}_k (V|V_{\perp})^*, \quad \text{where} \quad \hat{A}_k = \begin{bmatrix} M & R^* \\ R & \hat{S} \end{bmatrix}. \quad (4.5)$$

The subscript $k$ of $A_k$ refers to the number of columns of $V$. In [69], the matrix $A_k$ is called a subspace projected approximate matrix. This is motivated by the fact that $A_k V = AV$ and $V^* A_k = V^* A$. In particular, since $M = V^* AV = V^* A_k V$, both $A_k$ and $A$ have the same Ritz pairs in $V$ (see Section 1.2.1). This will be exploited to derive bounds for the eigenvalues of $A_k$ in Section 4.3. This is of interest since the search space $V$ in the outer iteration is expanded with an eigenvector of $A_k$. Note that the action of $A_k$ on $V_{\perp}$ does not equal the action of $A_0$ on $V_{\perp}$. Since $A^* = A$, the action of $A_k$ on $V_{\perp}$ equals, in fact, the action of $A$
in \( k \) linearly independent functionals on \( \mathbb{C}^n \).

With the convention that \( \Pi_0 = I \), write

\[
\Pi_k = V_\perp V_\perp^* = I - VV^*.
\] (4.6)

This shows, together with equations (4.4) and (4.2), that

\[
A_k = -VV^*AVV^* + AVV^* + VV^*A + \Pi_k A_0 \Pi_k.
\] (4.7)

Thus, the action of \( A_k \) can benefit from the stored action of \( A \) on \( V \) as follows. With \( W = AV \) we have that

\[
A_k v = -VMV^*v + WV^*v + VW^*v + \Pi_k A_0 \Pi_k v
\] (4.8)

where we have used (4.6) to avoid the numerically infeasible formation of \( V_\perp \). Note that if \( v \in V_\perp \), the first two terms vanish. We now observe the following.

**Remark 4.1.1.** If SPAM runs for the full \( n \) iterations, it produces a unitary matrix \( U_n \) that depends only on \( A \) and \( A_0 \). The \( k \)th leading principal submatrix of the \( n \times n \) matrix \( M = U_n^* AU_n \) then contains the Rayleigh-Ritz approximations after \( k \) steps of the outer iteration.

For theoretical purposes and without loss of generality, we assume that \( V_\perp \) in (4.5) contains precisely the basis for the orthogonal complement of \( V \) that SPAM is about to produce in future iterations. With respect to this basis, \( \hat{A}_k \) is an update of \( \hat{A}_{k-1} \) of *arrowhead type*, in the sense that

\[
\hat{A}_k - \hat{A}_{k-1} = \begin{bmatrix}
O_{k-1} & \tau & t^* \\
\tau & t & O_k
\end{bmatrix} = \begin{bmatrix}
0 & e_k^* + e_k \begin{bmatrix}
0^* & \tau & t^*
\end{bmatrix} - e_k \tau e_k^*
\end{bmatrix},
\] (4.9)

where each entry in the arrowhead formed by \( t \in \mathbb{C}^{n-k}, \tau \in \mathbb{R} \) and \( t^* \), is the difference between the corresponding entries of the matrix \( \hat{A} \) from (4.2) and the matrix \( \hat{A}_0 = (V|V_\perp)_* A_0 (V|V_\perp) \) from (4.5). Thus, with respect to the basis defined by the columns of \( (V|V_\perp) \), the matrix \( \hat{A}_0 \) simply transforms step by step into \( \hat{A} \) in
the sense that after \( k \) steps, the first \( k \) columns and rows have changed into those of \( \hat{A} \), and the resulting matrix is called \( \hat{A}_k \). This is visualized in Figure 1.

On the original basis, this transformation is described in the next proposition. Note that in this proposition, \( v \) is the eigenvector of interest of \( A_{k-1} \), orthonormalized to \( V_{k-1} \).

### Proposition 4.1.2

Let \( k \geq 1 \). Write \((V_{k-1}|v|V_{\perp}) = (V|V_{\perp})\), thus \( v = Ve_k \).

Then \( A_k \) is the following indefinite Hermitian rank-2 update of \( A_{k-1} \),

\[
A_k = A_{k-1} + uv^* + vu^* = A_{k-1} + (u|v) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (u|v)^*,
\]

(4.10)

where

\[
u = \left( \Pi_{k-1} - \frac{1}{2} vv^* \right) (A - A_0)v = \left( \Pi_k + \frac{1}{2} vv^* \right) (A - A_0)v.
\]

(4.11)

**Proof.** Combining (4.9) with \((V|V_{\perp})(\hat{A} - \hat{A}_0)(V|V_{\perp})^* = A - A_0\) we find

\[
\begin{bmatrix} 0 \\ t \end{bmatrix} = \begin{bmatrix} O & e_k & \cdots & e_n \end{bmatrix}^* (\hat{A} - \hat{A}_0)e_k = (O|v|V_{\perp})^*(A - A_0)v
\]

(4.12)
where \( O \) stands for the \( n \times (k - 1) \) zero matrix. Therefore, substituting (4.12) into

\[
A_k - A_{k-1} = (V_{k-1}|v|V_\perp) \begin{pmatrix}
\vec{0} \\
t \\
\tau \\
e_k^* + e_k \begin{bmatrix}
\vec{0}^* & t^* \\
\tau 
\end{bmatrix} - e_k \tau e_k^*
\end{pmatrix} (V_{k-1}|v|V_\perp)^*
\]

we arrive, using (4.12) and \( v = (V_{k-1}|v|V_\perp)e_k \), at

\[
A_k - A_{k-1} = \Pi_{k-1}(A - A_0)vv^* + vv^*(A - A_0)\Pi_{k-1} - vv^*(A - A_0)vv^*.
\]

(4.13)

Splitting the most right term in two equal parts and rearranging the terms proves the formula for \( u \) in (4.11). Since by (4.6)

\[
\Pi_{k-1} - vv^* = \Pi_k,
\]

(4.14)

also the second equality is proved.

\[\square\]

**Remark 4.1.3.** Note that the result of Proposition 4.1.2 can also be derived in an alternative way, starting from (4.7). The proof presented here also emphasizes the validity of (4.9).

From (4.2) and (4.5) we see that

\[
\text{rank}(A - A_k) = \text{rank}(\hat{A} - \hat{A}_k) \leq n - k.
\]

Thus, even though \( A_k - A_{k-1} \) has rank-2, the update - or maybe better downdate

\[
A - A_k = A - A_{k-1} + (A_{k-1} - A_k)
\]

will generically decrease the rank of \( A - A_{k-1} \) with at most one. This remark goes hand in hand with the observation that even though the approximations \( A_k \) of the matrix \( A \) may seem unusual, the viewpoint of considering the reverse sequence \( O_n = A - A_n, A - A_{n-1}, \ldots, A - A_1, A - A_0 \) as increasingly better approximations of \( A - A_0 \) is very natural indeed: they form a sequence of Rayleigh-Ritz approximations to \( A - A_0 \) in the orthogonal complements of the spaces \( V_k \).
In the outer iteration, the products $Av$ and $V^*Av$ were computed. Thus, in (4.11), both $v^*Av$ and $\Pi_{k-1}Av$ are available without additional computational costs. Furthermore, since $v$ is orthogonal to the $(k-1)$-dimensional search space, we have that $v^*A_{k-1}v = v^*A_0v$. Now, because $v$ is the result of orthogonalization of an eigenvector of $A_{k-1}$ to $V_{k-1}$, also $v^*A_0v$ can be retrieved from the inner iteration. Thus, the vectors $u$ and $v$ in the updating procedure (4.10) are cheaply available.

Remark 4.1.4. Of course, the update (4.10) itself should not be performed explicitly because it will generally result in fill-in of originally sparse matrices.

4.1.3 Choice of the method for the inner iteration of SPAM

In [69], the authors suggest to use Davidson’s method [22] to solve the eigenvalue problem for $A_k$ but, of course, any other method can be adopted. Apart from the Lanczos method, also the Generalized Davidson method from [57] was tested as inner method in [69]. Other possibilities include, for instance, the Jacobi-Davidson method [72]. The latter can be a good option because it often needs only a few iterations to converge if a good start vector is available. This start vector may be either the eigenvector approximation of $A_{k-1}$, or the current eigenvector approximation of $A$ from the outer iteration. We study this choice in Section 4.2.

Remark 4.1.5. SPAM should first of all perform well under the assumption that the eigenproblem for $A_k$ is solved exactly. This will be investigated in the numerical illustrations in Section 4.4.

In [69] it is also noted that SPAM itself can be chosen in the inner iteration. This leads to a recursive multilevel version of the method, and assumes that a sequence of approximating matrices of $A$ is available, each having a cheaper action than its predecessor. The eigenvector computed at a given level is then used as expansion vector at the first higher level.

1M. Hochstenbach presented this option at the 2006 GAMM/SIAM Applied Linear Algebra conference.
4.1.4 Practical implementation

The implementation of the Subspace Projected Approximate Matrix method is, in theory, given by Algorithm 1 where on line 14 the matrix $A_k$ could be formed explicitly and its target eigenpair determined. Such implementation, however, does not take into account the fact that part of the information needed for the computations with $A_k$ is already available. The subspace projected approximate matrix $A_k$ need not be explicitly assembled. An algorithm closer to the practical implementation is given in Algorithm 3 (see also Algorithm/Figure 2 in [69]).

Remark 4.1.6. In Algorithm 3 we define $V_q := [v_1, \ldots, v_q]$ and $W_q := [w_1, \ldots, w_q]$ and use the indexes $k$ and $p$ to keep track of the dimensions of the inner and outer search subspaces. Moreover, due to the characteristics of the SPAM method, inner and outer iterations are only distinguishable from the type of matrix-vector products. In Algorithm 3 this difference is showed by the variable $\text{it}$ which may assume the value inner or outer. When $\text{it}=\text{inner}$ the Algorithm computes an approximation for a target eigenpair of $A_k$ and when $\text{it}=\text{outer}$ for one of $A$.

4.2 SPAM and other subspace methods

The philosophy of SPAM is that only $S$ needs to be approximated. Therefore, all available (computed) information from (4.2), i.e., $M, R$ and $R^*$, is used to determine the expansion vector. The Jacobi-Davidson [72] and Riccati [9] methods partially share this philosophy. Instead of $R$, however, they only use the residual corresponding to the selected eigenvector approximation. On the other hand, in their most simple forms, they do not approximate the matrix $S$ but use its full action. In this section we outline their similarities and differences.

Remark 4.2.1. Since in the Lanczos method all residuals are linearly dependent, also the Lanczos method uses all information about the residual. However, Lanczos has no inner iteration. This will make sense from the point of view taken in Section 4.3. There we will show that for the choice $A_0 = O_n$, also SPAM needs no inner iteration, and that this choice makes SPAM mathematically equivalent to Lanczos.
Algorithm 3 Subspace Projected Approximate Matrix method

1: Input: $A$, $A_0$, $v_0$, $\epsilon$;
2: Set $k = 0; p = 1$;
3: $\text{it}=\text{inner}$;
4: loop
5: $v_{k+p} = v_{k+p}/\|v_{k+p}\|_2$;
6: if $\text{it}=\text{inner}$ then
7: $w_{k+p} = A_0v_{k+p}$;
8: else
9: $w_k = Av_k$;
10: end if
11: for $j = 1, \ldots, k + p - 1$ do
12: $M_{j,k+p} = v_j^*w_{k+p}; M_{k+p,j} = M_{j,k+p}^*$;
13: end for
14: $M_{k+p,k+p} = v_{k+p}^*w_{k+p}$;
15: Compute the target eigenpair $(\mu, z)$ of $M_{k+p,k+p}$;
16: Compute $r = W_{k+p}z - \mu V_{k+p}z$;
17: if $\|r\|_2 \leq \epsilon$ and $\text{it} = \text{outer}$ then
18: stop; \{(\mu, Vz) is the desired eigenpair of $A$\}
19: else if $\|r\|_2 \leq \epsilon$ and $\text{it} = \text{inner}$ then
20: Set $\text{it} = \text{outer}$ \{The target eigenvector $Vz$ of $A_k$ was found.\}
21: $v_{k+1} = V_{k+p}z$;
22: $k = k + 1; p = 0$;
23: Orthogonalize $v_{k+1}$ with respect to the columns of $V_k = [v_1, \ldots, v_k]$;
24: else
25: $v_{k+p+1} = V_{k+p}z$
26: Orthogonalize $v_{p+1}$ with respect to the columns of $V_{k+p} = [v_1, \ldots, v_{k+p}]$;
27: $p = p + 1; \text{it} = \text{inner}$;
28: end if
29: end loop

Recall from Section 1.2.4 that Jacobi-Davidson is simpler than the Riccati method in the sense that only a linear system for $\hat{t}$ needs to be solved. Jacobi-Davidson can be interpreted as an accelerated Newton method [73]. However, much more than the Riccati method, Jacobi-Davidson suffers from stagnation in case the term $t\hat{t}^*t$ from (1.9) is not small. On the other hand, if $t\hat{t}^*t$ is small enough, Jacobi-Davidson converges quadratically, as one would expect from a Newton type method. This shows that one should be careful in proposing alternatives to the correction equation (1.10). For instance, in [25], the authors investigated the effect
of solving the following alternative correction equation,

\[ \tilde{t} \perp V \quad \text{and} \quad (I - VV^*)(A - \mu I)(I - VV^*)\tilde{t} = -\hat{r}. \]  

(4.15)

At first sight this seems to make sense, because it directly looks for a correction orthogonal to \( V \). Also, the conditioning of the linear equation (4.15) may be better than (1.10) in case the search space contains good approximations of eigenvectors belonging to eigenvalues close to \( \mu \). However, orthogonalizing \( \hat{t} \) from (1.10) to \( V \) generally does not result in \( \tilde{t} \) from (4.15) and the price to pay is that \( \| t - \tilde{t} \|_2 \) is not of higher order, as is \( \| t - \hat{t} \|_2 \). Indeed, in [28] it is shown explicitly that, apart from some exceptional cases, the quadratic convergence of Jacobi-Davidson is lost, whereas in those exceptional cases, both expansions are equivalent. Numerical experiments in [28] confirm the above observations.

Preconditioning in the Jacobi-Davidson method.

In the original paper on the Jacobi-Davidson method [72], reprinted as [74], preconditioning is discussed as follows. Suppose that an approximation \( A_0 \) of \( A \) is available. It is shown in [72] how to apply such a preconditioner to (1.10), which is a linear equation, though not with system matrix \( A \). To be explicit, since \( (I - uu^*)\hat{t} = \hat{t} \), we have that

\[ (A - \mu I)\hat{t} = -\varepsilon u - \hat{r}. \]

where \( \varepsilon \) is such that \( \hat{t} \perp u \). Or equivalently, written as an augmented system,

\[
\begin{bmatrix}
A - \mu I & u \\
u^* & 0
\end{bmatrix}
\begin{bmatrix}
\hat{t} \\
\varepsilon
\end{bmatrix}
=
\begin{bmatrix}
-\hat{r} \\
0
\end{bmatrix}.
\]

(4.16)

Thus, an approximation \( \hat{t}_0 \) of \( \hat{t} \), together with an approximation \( \varepsilon_0 \) of \( \varepsilon \) can be obtained by replacing \( A \) by \( A_0 \) in (4.16). The pair \( \hat{t}_0, \varepsilon_0 \) can be computed as

\[ \hat{t}_0 = -\varepsilon_0 (A_0 - \mu I)^{-1} u - (A_0 - \mu I)^{-1} \hat{r}, \]

(4.17)
with
\[ \varepsilon_0 = -u^* (A_0 - \mu I)^{-1} \hat{r} - u^*(A_0 - \mu I)^{-1} u. \] (4.18)

This approximation is called a one step approximation in [72]. It was observed that setting \( \varepsilon_0 = 0 \) in (4.17), the Davidson method [22] results. With \( A_0 = A \), which corresponds to Jacobi-Davidson with full accuracy solution of the correction equation, (4.17) becomes
\[ \hat{t} = -\varepsilon (A - \mu I)^{-1} u - u, \] (4.19)

and since \( \hat{t} \) is then orthogonalized to \( u \), the method is mathematically equivalent to an accelerated shift and invert iteration that works with \( (A - \mu I)^{-1} u \). It is argued, and demonstrated by experiments in [72], that Jacobi-Davidson combines the best of those two methods. Of course, a natural next stage in preconditioning is to use the matrix
\[ A_0^n = \begin{bmatrix} A_0 - \mu I & u \\ u^* & 0 \end{bmatrix} \]
as preconditioner within the iterative method that aims to solve (4.16). In each step of such a method one would need to solve a system with \( A_0^n \). This can be done by solving two systems as in (4.16)-(4.17) in the first step of the inner iteration. In each consecutive step, only one system of the form \( (A_0 - \mu I)z = y \) would need to be solved.

**One step approximation of the SPAM eigenproblem for \( A_k \).**

In the SPAM method, the expansion vector for the Ritz Galerkin subspace in the outer iteration is a relevant eigenvector \( v_k \) of \( A_k \). In principle, any eigenvalue method can be used to compute an approximation for \( v_k \), but observe that the starting point is as follows. In the outer iteration, we have just solved a \( k \times k \) eigenproblem for \( M = V^*AV \), and a Ritz pair \((\mu, u)\) with \( \|u\|_2 = 1 \) and with residual \( \hat{r} = Au - \mu u \) has been selected. The matrix \( A_k \) is now available, either explicitly or implicitly. Since \( A_k V = AV \) and \( V^*A_k = V^*A \), the Ritz pair \((\mu, u)\) for \( A \) with respect to the current search space \( \mathcal{V}_k \) is also a Ritz pair for \( A_k \). Thus we can exploit the fact that \( \mathcal{V}_k \) contains, by definition, good approximations of
the relevant eigenvectors of $A_j$ with $j < k$, and use it as initial search space for a
Ritz Galerkin method applied to $A_k$ to approximate $v_k$. Since $V_k$ is generally not
a Krylov subspace, the Lanczos method is not a feasible candidate. The Jacobi-
Davidson method is. The correction equation for the first step of the Jacobi-
Davidson method in the inner iteration can be set up without any additional
computations:
\[
\begin{bmatrix}
A_k - \mu I & u \\
u^* & 0
\end{bmatrix}
\begin{bmatrix}
t_k \\
x_k
\end{bmatrix}
= 
\begin{bmatrix}
-\hat{r} \\
0
\end{bmatrix}.
\] (4.20)

Since quadratic convergence in the outer iteration cannot be expected even if an
exact eigenvector of $A_k$ would be computed, we study the effect of applying only
one iteration of Jacobi-Davidson in the inner iteration. This is also motivated by
the fact that the initial search space $V_k$ for Jacobi-Davidson applied to $A_k$ may be
relatively good and may result in quadratic convergence in the inner iteration.

**Remark 4.2.2.** If only one step of Jacobi-Davidson is applied, then after solving $t_k$
from (4.20), the new approximation $v$ for the eigenvector $v_k$ of $A_k$ would lie in the
space $V_k \oplus \langle t_k \rangle$. It would not be necessary to actually compute this approximation,
because
\[
V_k \oplus \langle v \rangle \subset V_k \oplus \langle t_k \rangle.
\] (4.21)

Thus, instead of computing the eigendata of a $(k+1) \times (k+1)$ matrix, the expansion
of the Ritz-Galerkin space in the outer iteration can be done immediately with $t_k$.

SPAM, in which the eigenproblem for $A_k$ is approximated with one iteration
of Jacobi-Davidson, we will refer to as one step SPAM, abbreviated by SPAM(1). We will talk about Full SPAM if the eigenproblem for $A_k$ is solved to full precision.

**Comparing one step Jacobi-Davidson with SPAM(1).**

SPAM(1) can best be compared with preconditioned Jacobi-Davidson with one
step approximation, as described in Section 4.2. The only difference between the
two methods is that in iteration $k$ of SPAM(1) the preconditioner $A_k$ is used,
whereas in one-step Jacobi-Davidson this is $A_0$. As such, SPAM(1) can be seen as
an attempt to enhance this type of preconditioned Jacobi-Davidson. We investigate
the effect of this attempt.
Lemma 4.2.3. Assume that $V_{k+1} = V_k \oplus \langle v \rangle$ with $v \perp V_k$ and $\|v\|_2 = 1$. Then

$$A_{k+1} = -vv^*Avv^* + Avv^* + vv^*A + (I - vv^*)A_k(I - vv^*). \quad (4.22)$$

Proof. By substitution of the defining relation $[4.7]$ for $A_k$. \qed

Corollary 4.2.4. Let $V_1$ be the span of the relevant eigenvector $u$ of $A_0$ and let $\mu = u^*Au$ and $\hat{r} = Au - \mu u$. Then the solution $t_1$ from the system

$$\begin{bmatrix} A_1 - \mu I & u \\ u^* & 0 \end{bmatrix} \begin{bmatrix} t_1 \\ \xi_1 \end{bmatrix} = \begin{bmatrix} -\hat{r} \\ 0 \end{bmatrix},$$

in the first iteration of SPAM(1), to be used to expand $V_1$, coincides with the solution $t_0$ from the system

$$\begin{bmatrix} A_0 - \mu I & u \\ u^* & 0 \end{bmatrix} \begin{bmatrix} t_0 \\ \xi_0 \end{bmatrix} = \begin{bmatrix} -\hat{r} \\ 0 \end{bmatrix},$$

solved in the first iteration of Jacobi-Davidson with one-step approximation using the matrix $A_0$.

Proof. The linear system $[4.20]$ for SPAM(1) with $k = 1$ is equivalent to

$$t_1 \perp u, \quad (I - uu^*)(A_1 - \mu I)(I - uu)t_1 = -\hat{r}, \quad (4.23)$$

where $u$ is a unit vector spanning $V_1$. Substituting the expression $[4.22]$ for $A_1$ immediately proves the statement. \qed

Thus, there is no difference in the first iteration. There is, however, a difference in further iterations. To study this difference we take a different viewpoint. Above, we approximated the SPAM inner eigenvalue problem by a linear correction equation which made it suitable for comparison with one step Jacobi-Davidson. The opposite viewpoint is also possible, which is to interpret the Jacobi-Davidson correction equation with one-step approximation as an exact correction equation of a perturbed eigenproblem.
Lemma 4.2.5. Given \( u \) with \( \|u\|_2 = 1 \) and \( \mu = u^*A u \) and \( r = A u - \mu u \). Define for a given approximation \( A_0 \) of \( A \) the matrix

\[
A_u = -u u^* A u u^* + u u^* A + A u u^* + (I - u u^*) A_0 (I - u u^*)
\]

\[
= u \mu u^* + u r^* + r u^* + (I - u u^*) A_0 (I - u u^*).
\] (4.24)

Then \((u, \mu)\) is a Ritz pair of \( A_u \) in the one-dimensional span of \( u \) with residual \( r = A_u u - \mu u \) and with the equation

\[
(I - u u^*) (A_0 - \mu I) (I - u u^*) t = -r
\] (4.25)

as its exact (i.e., without preconditioning) Jacobi-Davidson correction equation.

Proof. It is easily verified that \( u^*A_u u = \mu \) and \( A_u u - \mu u = r \) and thus \((\mu, u)\) is a Ritz pair for \( A_u \) with residual \( r \). Since moreover,

\[
(I - u u^*) A_u (I - u u^*) = (I - u u^*) A_0 (I - u u^*)
\]

its correction equation is precisely (4.25), or equivalently, (4.17).

Note that \( A_u \) from (4.24) is the subspace projected approximate matrix for the one dimensional subspace spanned by \( u \). Now, with

\[
V_k = U \oplus \langle u \rangle,
\]

where \( U \) is the orthogonal complement of the span of the relevant Ritz vector \( u \) in the current search space \( V_k \), we have, similar as in (4.22) and (4.24), that

\[
A_k = u \mu u^* + u r^* + r u^* + (I - u u^*) A_U (I - u u^*).
\] (4.26)

Here, \( A_U \) is the subspace projected approximated matrix corresponding to the subspace \( U \). Now, the opposite viewpoint mentioned above, is the observation that in one step Jacobi-Davidson, the expansion vector is (an arguably good approximation of) the relevant eigenvector of \( A_u \) in (4.24), whereas in SPAM, it is (an
3. Selecting the matrix in the SPAM method

A arguably good approximation of) the relevant eigenvector of $A_k$ in (4.26). Both matrices have now an appearance that is suitable for studying their differences and similarities.

The most important observation is that neither correction will lead to the unique correction that results in quadratic convergence in the outer iteration. Second, since both matrices $A_u$ and $A_k$ differ only in their restriction to the orthogonal complement $u^\perp$ of $u$, the difference in the methods they represent will be marginal if the residual is already small. Since, as already mentioned in Corollary 4.2.4, not only at the start but also in the first iteration both methods coincide, the difference from the second iteration onwards will probably be very small, especially if $A_0$ provides a good initial approximation of the relevant eigenvector. Finally, even though $A_u$ uses less information of $A$ than $A_k$, it does use the optimal information in some sense. It may even be a disadvantage to use more information, because this involves approximations of eigenvectors orthogonal to the eigenvector of interest. The above observations will be tested, and confirmed by our numerical illustrations in Section 4.4.

4. Selecting the matrix $A_0$ in the SPAM method

Here we study some of the effects of the choice for $A_0$ on the iterative approximation process in SPAM. We assume, for simplicity, that the largest eigenvalue of $A$ is the target, although by replacing $A$ by $-A$, we might as well have set the smallest eigenvalue of $A$ as a target.

4.3.1 Some simple choices for $A_0$

It is instructive to study the consequences of the choice $A_0 = O_n$. This generic parameter-free choice may seem dubious at first sight, but it is not. First note that since the start vector of SPAM is the relevant eigenvector of $A_0 = O_n$, this is a random vector $v \in \mathbb{C}^n$, with $\|v\|_2 = 1$. Thus, we set $V_1 = \text{span}\{v\}$. Write $\mu = v^*Av$ and $\hat{r} = Av - v\mu$. Then, with $V_\perp$ such that $(v|V_\perp)$ is orthogonal, and
$r = V^*_r \hat{r}$ we find that

$$A = (v | V) \begin{bmatrix} \mu & r^* \\ r & S \end{bmatrix} (v | V^*)_r,$$

(4.27)

and consequently, replacing $S$ by the zero matrix, the next approximating matrix $A_1$ from (4.5) is defined by

$$A_1 = (v | V) \begin{bmatrix} \mu & r^* \\ r & O_{n-1} \end{bmatrix} (v | V^*)_r.$$

(4.28)

As shown already in Proposition 4.1.2, $A_1$ is a simple rank-two matrix, that on the basis defined by the columns of $(v | V^*_r)$ is of arrow-head type. It has two nontrivial eigenpairs $A_1 w_\pm = \theta_\pm w_\pm$, where

$$w_\pm = \theta_\pm v + \hat{r} \quad \text{and} \quad \theta_\pm = \frac{1}{2} \mu \pm \sqrt{\frac{1}{4} \mu^2 + \|\hat{r}\|^2}.$$

(4.29)

Assuming that $A$ is positive definite will lead to the selection of $w_+$ for the expansion of the search space. Since $w_+$ is a linear combinations of $v$ and $\hat{r}$, we find that

$$V_2 = \text{span}\{v, \hat{r}\} = \mathcal{K}^2(A, v),$$

and the two eigenvalue approximations computed in the outer loop of the SPAM method are the same as in the Lanczos method. This is, of course, not a coincidence.

**Theorem 4.3.1.** If the goal is to find the largest eigenvalue of a positive definite matrix $A$, the SPAM method with $A_0 = O_n$ is mathematically equivalent to the Lanczos method.

**Proof.** Let $A_0 = O_n$. Then the eigenvalues of $A_k$ in (4.5) are those of the $n \times n$ Hermitian arrowhead

$$\begin{bmatrix} M & R^* \\ R & O_{n-k} \end{bmatrix}.$$

The Cauchy Interlace Theorem (Theorem 1.1.10) immediately gives that the largest $k$ of them are each larger than or equal to the corresponding eigenvalue of $M$. This
assures that the eigenvector of $A_k$ that is selected for expansion is not from its null space but from its column span. From (4.7) we see that if $A_0 = O_n$ this column span is the span of $V$ and $AV$. A simple induction argument shows that this span equals $K^{k+1}(A,v)$.

**Remark 4.3.2.** If $A$ is indefinite and we would like to find the eigenvalue closest to zero, the choice $A_0 = O_n$ would lead to expansion vectors from the null space of $A_0$, and the method would be worthless. As a solution, $A$ may be shifted by a suitable multiple $\alpha$ times the identity $I$ to make it positive semi-definite. Equivalently, instead of using $A_0 = O_n$ we may choose $A_0 = \alpha I$ as approximating matrix. In both cases, it is easy to verify that SPAM will still be equal to Lanczos.

The observant reader may have noticed that a peculiar situation has arisen. In the inner loop of SPAM, a better approximation of the largest eigenvalue of $A$ was computed than the Ritz values from the outer loop. In view of the philosophy of inner-outer iterations, this in itself is not out of the ordinary, but its computation did not require any additional matrix-vector multiplication with $A$, nor with an elaborate approximation $A_0$ of $A$. The following proposition, which uses the notation (4.1) and (4.3), makes this explicit.

**Proposition 4.3.3.** With $\theta_+$ as in (4.29) we have that $\mu \leq \|Av\|_2 \leq \theta_+$. Moreover, if $A$ is positive semi-definite, we find that $\theta_+ \leq \lambda_+$, where $\lambda_+$ is the largest eigenvalue of $A$.

**Proof.** Because $\hat{r} \perp v$ and $Av - \mu v = \hat{r}$, by Pythagoras’ Theorem we have

\[
\mu^2 \|v\|_2^2 + \|\hat{r}\|_2^2 = \mu^2 + \|\hat{r}\|_2^2 = \|Av\|_2^2,
\]

hence $\mu \leq \|Av\|_2$. Squaring $\theta_+$ from (4.29) gives

\[
\theta_+^2 = \|\hat{r}\|_2^2 + \frac{1}{2} \mu^2 + \mu \sqrt{\frac{1}{4} \mu^2 + \|\hat{r}\|_2^2},
\]

which together with (4.30) shows that $\|Av\|_2 \leq \theta_+$. Since $S$ in (4.27) is positive definite whenever $A$ is, combining (4.27) and (4.28) with Weyl’s bound (Theorem 1.1.7) yields $\theta_+ \leq \lambda_+$. \qed
The key issue here is that the inner eigenvalue approximations are much related to the so-called harmonic Ritz values \([6, 71]\) of \(A\). Indeed, assuming that \(A\) itself is positive semi-definite, these are the \(k\) positive eigenvalues of the at most rank-2\(k\) matrix

\[
\tilde{A}_k = \begin{bmatrix} M & R^* \\ R & T \end{bmatrix}, \quad \text{where} \quad T = RM^{-1}R^*.
\] (4.31)

They can be computed without additional matrix vector multiplications with \(A\). Note that we introduced in Section 1.2.1 harmonic Ritz values as the reciprocals of the Rayleigh-Ritz approximations of \(A^{-1}\) in the space \(AV\). It is well known that for positive semi-definite matrices \(A\), the harmonic Ritz values are better approximations of the larger eigenvalues of \(A\) than the standard Ritz values. We provide the short argument in Lemma 4.3.5. See also \([6]\).

**Proposition 4.3.4.** The matrix \(\tilde{A}_k\) can be decomposed as

\[
\begin{bmatrix} M & R^* \\ R & T \end{bmatrix} = \begin{bmatrix} M \\ R \end{bmatrix}M^{-1}\begin{bmatrix} M & R^* \end{bmatrix}.
\]

The blocks

\[
\begin{bmatrix} M \\ R \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -M^{-1}R^* \\ I \end{bmatrix}
\]

span the range and null space, respectively, and the nonzero eigenvalues are the eigenvalues of the \(k \times k\) matrix

\[
UM^{-1}U^* \quad \text{where} \quad \begin{bmatrix} M \\ R \end{bmatrix} = QU
\]

is a QR-decomposition. In particular, those eigenvalues are positive.

**Proof.** The statements are all easy to verify. The positivity of the \(k\) nonzero eigenvalues follows from Sylvester’s Theorem of Inertia. \(\square\)

The \(k\) eigenpairs \((\theta_j, w_j)\) of \(\tilde{A}_k\) in (4.31) with positive eigenvalues we label as

\[
0 < \tilde{\theta}_1 \leq \tilde{\theta}_2 \leq \ldots \leq \tilde{\theta}_{k-1} \leq \tilde{\theta}_k.
\]
The proposition shows that they can be computed by solving a $k \times k$ eigenproblem and that no additional matrix-vector products with $A$ are needed. We can now easily prove the following bounds. See also [6, 63].

**Lemma 4.3.5.** For all $j \in \{1, \ldots, k\}$, we have that

$$\mu_j \leq \tilde{\theta}_j \leq \lambda_j.$$  \hfill (4.32)

**Proof.** The left inequalities follow from the Cauchy Interlace Theorem (Theorem 1.1.10) applied to $\tilde{A}_k$. Now, with $\hat{A}$ as in (4.2), recognizing the Schur complement $\hat{A}/M$ shows that,

$$\hat{A} - \tilde{A}_k = \begin{bmatrix} O_k \\ \hat{A}/M \end{bmatrix}$$  \hfill (4.33)

is positive semi-definite, hence the right inequalities follow from Weyl’s bound (Theorem 1.1.7). []

Observe that, as was the case with the choice $A_0 = O_n$, assuming that from the start of the SPAM method the eigenvector $w_1$ belonging to $\theta_1$ was selected for expansion, the eigenvectors $w_j$, called the harmonic Ritz vectors, lie in the column span of $AV$ and hence, in $K^{k+1}(A,v)$.

Thus, even though we may have improved the inner loop eigenvalue approximations, the SPAM method is still equal to the Lanczos method. It does give, however, valuable insight into SPAM: Lanczos results from the choice $A_0 = O_n$, even after modification of $A_k$ into the positive semi-definite matrix $\tilde{A}_k$. In order to get a method different than Lanczos, we should use less trivial approximations that are based on the structure and properties of $A$ itself, while aiming to retain similar inequalities as in Lemma 4.3.5. For this, we would need the matrices $A - A_k$ to be positive semi-definite. We will investigate this in the following sections.

### 4.3.2 One-sided approximations

Having seen that the trivial choice $A_0 = O_n$, even after after a correction that turns all approximate matrices positive semi-definite, will generally lead to the Lanczos
method, we now turn our attention to approximations from below, by which we mean \( A_0 \) such that \( A - A_0 \) is positive semi-definite.

**Lemma 4.3.6.** If \( A_0 \) approximates \( A \) from below, then so does each matrix \( A_k \).

**Proof.** Combining (4.2) and (4.4) with (4.5) we see that for all \( x \in \mathbb{C}^n \),

\[
x^* (\hat{A} - \hat{A}_k) x = x^* \begin{bmatrix} O_k & \end{bmatrix} x = x^* V_\perp^*(A - A_0) V_\perp x \geq 0 \tag{4.34}
\]

where the last inequality holds because \( A - A_0 \) is positive semi-definite. And thus, \( \hat{A} - \hat{A}_k \) is positive semi-definite, and hence, so is \( A - A_k \).

By Proposition 4.1.2, \( A_k - A_{k-1} \) is an indefinite rank-2 matrix, hence it will generally not be true that \( A_k^-1 \) approximates \( A_k \) from below.

**Lemma 4.3.7.** The following inequalities are valid generally,

\[
\theta_{j-1} \leq \mu_j \leq \theta_j \quad \text{for all } j \in \{1, \ldots, k\},
\]

whereas, if \( A_0 \) approximates \( A \) from below, additionally

\[
\theta_j \leq \lambda_j \quad \text{for all } j \in \{1, \ldots, n\}.
\]

**Proof.** The first set of inequalities applies because \( A_k \) has the same Ritz values as \( A \). See also Section 4.1.2. It is well known that the Ritz values interlace the exact eigenvalues [63] (see also Sections 1.1.4 and 1.2.1). Since \( A - A_k \) is positive semi-definite due to Lemma 4.3.6, the equality \( A_k + (A - A_k) = A \) together with Weyl’s bound (Theorem 1.1.7) proves the second set of inequalities.

Lemma 4.3.7 shows that if \( A_0 \) approximates \( A \) from below, the approximations for the larger eigenvalues of \( A \) that are produced in the inner iteration, will never be worse than the ones obtained in the outer iteration. Moreover, they will never be larger than the corresponding exact eigenvalues. Thus, it indeed makes sense to expand the search space with the eigenvector that is computed in the inner
4.3. SELECTING THE MATRIX $A_0$ IN THE SPAM METHOD

Iteration. Question that remains is how to obtain matrices $A_0$ that approximate $A$ from below.

### 4.3.3 Algebraic construction of approximations from below

Clearly, for any positive definite matrix $H$ we have that $A_0 = A - H$ approximates $A$ from below, even though $A_0$ itself may not be positive definite. The problem is of course how to choose $H$ such that $A_0$ is close to $A$ in an appropriate sense, while its action is considerably less expensive than that of $A$.

If $A$ itself is a positive definite matrix, a purely algebraic option is at hand. Given an index set $\mathcal{I} \subset \{1, \ldots, n\}$ of cardinality $m$, let $E_{\mathcal{I}}$ be the matrix with the standard basis vectors $e_j, j \in \mathcal{I}$ as columns. Then set

$$A_0 = A - H,$$

where

$$H = E_{\mathcal{I}}E_{\mathcal{I}}^* A E_{\mathcal{I}}E_{\mathcal{I}}^*.$$

The matrix $H$ is the result of a Ritz Galerkin procedure with as search space the column span of $E_{\mathcal{I}}$. For a randomly chosen index set, this space has no a priori relation with $A$ and thus $H$ is probably a relatively poor approximation of $A$ in comparison with, for example, a Krylov subspace approximation.

**Remark 4.3.8.** In this particular situation, it is an advantage if $H$ does not approximate $A$ very well, because it is $A_0$ that should be close to $A$, not $H$. Notice also that $A_0$ has zero entries at positions $(j,k)$ for all $j,k \in \mathcal{I}$ and is, thus, always more sparse than $A$. A priori knowledge of $A$ may lead to a more sophisticated choice of the index set(s) $\mathcal{I}$. If the goal is to approximate the largest eigenvalues of $A$, the index set $\mathcal{I}$ could be chosen such that the smallest diagonal entries of $A$ are selected to put in $H$. Consequently, $A_0$ will share with $A$ the largest diagonal entries, and this may increase its approximation quality. This is illustrated in Figure 4.2.

**Remark 4.3.9.** Notice that $\text{rank}(A_0) \leq 2(n - m)$. For large $m$ this may greatly simplify the computation of the eigendata of $A_0$ and of $A_k$ for small values of $k$.

**Remark 4.3.10.** Although $A$ itself is positive definite, $A_0$ generally is not. It can be made positive definite by adding a Schur complement at the position of
Figure 4.2: Approximating a matrix $A$ from structural engineering from below by a sparser matrix $A_0$ by subtracting from $A$ a definite matrix $H$; the sparsity plot of $A$ (with 3648 nonzero entries) is on the left and of $A_0$ (with 968 nonzero entries) on the right. See the experiments with this pair of matrices in Section 4.4.4.

The zero block, as was done in (4.31). Since the Schur complement block is the smallest correction to $A_0$ that makes it positive semi-definite, the result would still approximate $A$ from below. However, its computation involves the evaluation of the inverse of an $(n - m) \times (n - m)$ matrix. It is not clear if the additional computational effort is well spent. Also without the Schur complement, $A_0$ has as many positive as negative eigenvalues and is likely to perform better than $A_0 = O_n$.

The rather crude approach that we just described is easy to implement and can be applied in the context of the multilevel version of SPAM, mentioned in Section 4.1.3. The solution of the inner iteration is done using SPAM itself, with an approximation of $A_0$ that is based on a larger index set than the one that was used to construct $A_0$ itself.

### 4.3.4 Natural construction of approximations from below

A situation in which approximations from below are naturally available is the setting of discretized partial differential equations including an elliptic term, either by the finite difference method or the finite element method \[18, 19\]. Then removing the positive definite discrete elliptic operator, either completely or partially, from the total discrete operator, results in an approximation from below. Indeed, let
Ω ⊂ \mathbb{R}^3 be a domain, and consider as an example the problem of finding eigen-modes for the linear operator \( L \), defined by

\[
L(u) = \lambda u \quad \text{where} \quad L(u) = -\varepsilon \text{div}(K\nabla u) + cu, \quad \text{and} \quad u = 0 \text{ on } \partial \Omega. \tag{4.35}
\]

Here \( \varepsilon > 0 \) is a parameter, \( K : \Omega \to M^{3\times3}(\mathbb{R}) \) maps into the symmetric positive definite matrices, and \( c \in C(\overline{\Omega}) \) is nonnegative. Discretizing this equation with the finite difference method will lead to an algebraic eigenvalue problem of the form \( \varepsilon Kx + Mx = \xi x \). The matrix \( K \) that represents the discretized diffusion is positive definite. Although sparse, it will generally have more fill-in than the diagonal matrix \( M \) that represents the reaction term, and, if \( \varepsilon \) is small enough, have smaller entries. Thus, the total discretized operator \( A = K + M \) has \( A_0 = M \) as a candidate for the approximation matrix: its action is cheaper than the action of \( A \) and \( A - A_0 = K \) is positive definite. A similar strategy can also be employed in the finite element method, when so-called mass lumping is used in the assembly of the mass matrices.

**Remark 4.3.11.** In this context, the algebraic method can be used to approximate the smallest eigenvalue of \( A \) by applying it to \( \alpha I - A \) with \( \alpha \) such that \( \alpha I - A \) is positive semi-definite. Even though the largest eigenvalue usually has no physical relevance, together they would provide good estimates for the condition number of the \( A \), which is indeed of interest.

### 4.3.5 Cutting off the bandwidth

Here we describe an obvious choice of approximating matrices that was used in [69] to illustrate the effectiveness of their method. It concerns symmetric banded matrices. Apart from the approach that we will now describe, in the numerical illustrations of Section 4.4 we also intend to apply our algebraic approximations from below to these matrices.
Given $0 \leq \epsilon \leq 1$ and $1 \leq q \leq n - 1$, define a matrix $A = (A_{jk})$ by

$$A_{jk} = \begin{cases} j & \text{if } j = k; \\ \epsilon |j-k| & \text{if } 1 \leq |j-k| \leq q; \\ 0 & \text{otherwise}. \end{cases} \quad (4.36)$$

In [69] it is proposed to choose as approximating matrix for $A$ a matrix $A_0$ of the same type with a smaller half-bandwidth $\tilde{q} < q$. For instance, for

$$A = \begin{pmatrix} 1 & \epsilon & \epsilon^2 \\ \epsilon & 2 & \epsilon \epsilon^2 \\ \epsilon^2 & \epsilon & 3 \epsilon \\ \epsilon^2 & \epsilon & 4 \end{pmatrix}, \quad \text{the matrix } \quad A_0 = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 2 \epsilon \\ \epsilon^2 & 3 \epsilon \\ \epsilon^2 & 4 \epsilon \end{pmatrix}.$$ 

For each eigenvalue $\theta$ of $A_0$ there is an eigenvalue $\lambda$ of $A$ with

$$|\lambda - \theta| \leq \frac{\epsilon^{\tilde{q}+1} - \epsilon^{q+1}}{1 - \epsilon} \sqrt{n}. \quad (4.37)$$

Indeed, the difference $A - A_0$ is zero except for the bands $\tilde{q} + 1$ to $q$. Each non-zero row contains at most the numbers $\epsilon^{\tilde{q}+1}$ to $\epsilon^q$, and (4.37) follows from the the Bauer-Fike (Theorem 1.1.1) theorem and the finite geometric series sum formula. Thus, for values of $\epsilon$ small enough, $A_0$ may be a good candidate to approximate $A$ even though it is generally not an approximation from below. Nevertheless, its eigenvalues are close to the eigenvalues of $A$ and its action is cheaper than the action of $A$. The number of floating point operations required to compute $Av$ is approximately twice as much as for an approximation $A_0$ having half the bandwidth of $A$. In other words, each two matrix-vector products $A_0$ are approximately equally costly as a single product with $A$.

Cutting off the bandwidth in this fashion makes sense especially if the decay of the size of the off-diagonal entries in relation to their distance to the main diagonal is quick enough. Apart from the example above, this is the case in applications where the boundary element method [84] is used to approximate integral equations. Notice that for such applications also the approach from Section 4.3.3 can be
applied, resulting in both sparse and low rank approximating matrices $A_0$.

4.4 Numerical illustrations

In the previous sections, we have compared SPAM with both the Jacobi-Davidson method and the Lanczos method. We have also studied ways to define an appropriate approximate matrix $A_0$ in the context of the SPAM method. Since the choice $A_0 = O_n$ will effectively lead to the Lanczos method, our starting point in the upcoming numerical illustrations will be to consider SPAM as a boosted version of the Lanczos method. This is, of course, particularly justified if $A_0$ is an approximation of $A$ from below. As discussed in Section 4.2, SPAM can also be considered as an attempt to enhance the Jacobi-Davidson method with one-step approximation as preconditioning. Therefore, we will also present a comparison of SPAM and this version of Jacobi-Davidson. We end this section with a discussion of the numerical results.

4.4.1 Objectives

First, we list the methods and abbreviations that we use to describe them.

- Lanczos (see Section 1.2.3);
- $\text{JD}(\ell)$: Jacobi-Davidson using $\ell$ steps of MinRES to approximate the solution of the exact correction equation (4.16) in augmented form;
- $\text{JD}(1,\ell)$: Jacobi-Davidson with one step approximation as preconditioner (see Section 4.2) using the matrix $A_0$, and $\ell$ steps of MinRES to approximate the solution of the correction equation (4.16), with $A$ replaced by $A_0$;
- Full SPAM: each eigenproblem for $A_k$ solved in full precision;
- SPAM(1): eigenproblem for $A_k$ approximated with one step of the Jacobi-Davidson method, correction equation (4.20) in augmented form solved to full precision (see Section 4.2);
- SPAM(1,$\ell$): using $\ell$ steps of MinRES to approximate the solution of the correction equation for SPAM(1).
**Remark 4.4.1.** To minimize the size of the legends in the pictures, we sometimes write LZS for Lanczos, FSP for Full SPAM, SP(1) for SPAM(1), S12 for SPAM(1,2), JD13 for JD(1,3), etcetera.

In the experiments, we give illustrations of the following aspects of SPAM.

- When a nonzero approximation $A_0$ of $A$ from below is used, less outer iterations of Full SPAM are needed to arrive close to the dominant eigenvalue of $A$ than with the choice $A_0 = O_n$, which is equivalent to the Lanczos method with a random start vector.

- Even if the Lanczos method is started with the same approximation of the dominant eigenvector of $A_0$ as Full SPAM, Full SPAM method will still outperform Lanczos in terms of the number of outer iterations.

- Also for other eigenvalues, Full SPAM outperforms Lanczos; this may be expected because in Full SPAM the Ritz Galerkin subspace will be forced in the direction of the appropriate eigenvector in every iteration. In Lanczos, this is done only by means of the start vector. Of course, Lanczos allows efficient implicit restart strategies [75], but also Full SPAM may be restarted. We feel that the comparison would become diffuse and thus we refrain from incorporating restarts.

- We investigate the effect of approximating the desired eigenvector of $A_k$ with just one step of the Jacobi-Davidson method, i.e., we will be comparing Full SPAM with SPAM(1).

- SPAM(1,\ell) will be compared with JD(1,\ell), both started with the same initial vector, the relevant eigenvector of $A_0$; i.e., both methods will spend the same number \ell of matrix-vector products in their inner iteration, where in SPAM(1,\ell), the matrix will be $A_k$, and in JD(1,\ell), the matrix will be $A_0$. From the viewpoint of this chapter, this is the comparison that is of most interest. Not only is JD(1,\ell) the closest related to SPAM(1,\ell), the difference between the two is solely the fact that the action of $A$ from the outer loop is taken into the inner iteration of SPAM(1,\ell), whereas in JD(1,\ell), this is not done. See the discussion in, and particularly at the end of Section 4.2.
Finally, we compare SPAM(1, ℓ) with JD(ℓ). This is perhaps the comparison that the authors of SPAM [69] had in mind: in the inner iteration of JD(ℓ) the original matrix $A$ is used, whereas in SPAM(1, ℓ) it will be $A_k$.

We will comment on the computational costs of an inner iteration in comparison to having no such costs (as in Lanczos) or the full costs (Jacobi-Davidson), although these costs may depend very much on the specific problem and the available approximations and even on hardware parameters like available memory.

### 4.4.2 Lanczos versus full SPAM

**Reaction-Diffusion problem, various eigenvalues**

In this section we compare Lanczos with Full SPAM. Our comparison is, for the time being, only in terms of the number of outer iterations. A first naive comparison uses a random start vector for Lanczos, but from then onwards, we start Lanczos with the appropriate eigenvector of $A_0$. The approximate matrix $A_0$ is constructed using both approaches described in Sections 4.3.3 and 4.3.4. For this, we discretized the one-dimensional version of (4.35) on $\Omega = [0, 1]$ using finite differences with grid size $h = 1/33$ with the parameters

$$\varepsilon = \frac{1}{33^2}, \quad K = 1, \quad \text{and} \quad c(x) = x(1 - x)e^{3x}.$$  

The resulting $32 \times 32$ algebraic eigenproblem is of the form $Ax = \lambda x$, where $A = D + R$ is the sum of the tridiagonal discretized diffusion $D$ and the diagonal discretized reaction $R$. With the approach from Section 4.3.4 we approximate the largest eigenvalue and with the approach from Section 4.3.3 the smallest eigenvalue.

**Natural approximation from below** The left picture of Figure 4.3 illustrates the typical convergence of the Lanczos method with a random start vector. We display the $k$ Ritz values at outer iteration $k$ as circles above the value $k$ of the iteration number on the horizontal axis. Due to the interlacing property, eigenvalue approximations “converge” to either side of the spectrum, and this is emphasized by the connecting lines between Ritz values for different values of $k$, both upwards
and downwards. In view of Theorem 4.3.1 we could also say that this picture belongs to SPAM with choice $A_0 = O_n$. In the right picture of Figure 4.3 we show in a similar fashion the convergence of SPAM, using $A_0 = R$ as the approximate matrix, as suggested in Section 4.3.4. We see that the convergence towards the largest eigenvalues is stimulated. The costs for this faster convergence is solving a diagonal plus rank $2k$ eigenproblem in iteration step $k$. There exist efficient methods for such eigenproblems based on the secular equation and Newton’s method.

**Algebraic approximation from below** We also tested the algebraic approach from Section 4.3.3 to construct approximations from below. We created a rank-12 approximation $A_0$ of $A$ based on the largest diagonal elements of $A$. Full SPAM and Lanczos were started with the dominant eigenvector of $A_0$ in order to approximate its dominant eigenvalue. The leftmost picture in Figure 4.4 shows that the incorporation of an approximation $A_0$ into Full SPAM has an effect that carries beyond that of creating a better start vector for Lanczos. In the rightmost picture of Figure 4.4, the same was done for the matrix $6I - A$, which is positive definite and has the smallest eigenvalue of $A$ as dominant eigenvalue. Also here, SPAM outperforms Lanczos in terms of the number of outer iterations. In the middle two pictures of Figure 4.4, we plotted the absolute error in the second and fifth largest eigenvalue of $A$. The results are shown from iteration 2 and 5 onwards, respectively, because the Ritz-Galerkin method produces approximations of the second and fifth largest eigenvalues from that iteration onwards. Again, Full SPAM clearly outperforms Lanczos in both cases. Note that, when using Full SPAM to approximate the $p$-th largest eigenvalue, the Ritz-Galerkin subspace in the outer iteration is in each step expanded with the eigenvector belonging to the $p$-th largest eigenvalue of $A_k$. For a fair comparison, we also started Lanczos with the eigenvector of $A_0$ belonging to the $p$-th largest eigenvalue.

### 4.4.3 Lanczos versus Full SPAM and SPAM(1)

**Banded matrices**

In this section we not only compare Lanczos with Full SPAM, but also with SPAM(1), by which we mean that the eigenproblem for $A_k$ in Full SPAM is ap-
4.4. NUMERICAL ILLUSTRATIONS

Figure 4.3: Lanczos method (left) with a random start vector, versus SPAM (right) with the discretized reaction term as approximating matrix $A_0$, and the largest eigenvalue of $A$ as target.

proximated using one step of the Jacobi-Davidson method, as explained in Section 4.2. The correction equation that results is still solved to full precision. For our experiments, we took the banded matrix from Section 4.3.5 of size $32 \times 32$ and with $q = 5$ and $\epsilon = 0.5$. In [69], it was suggested to take for $A_0$ the matrix $A$ with a number of its outer diagonals put to zero. We repeated that experiment with $A_0$ equal to the diagonal of $A$, such that here too, $A_k$ is diagonal plus a rank-$2k$ perturbation. The comparison between Lanczos, Full SPAM and SPAM(1) is depicted in the left graph in Figure 4.5. This comparison is initially in favor of Lanczos. This may be due to the fact that the difference $A - A_0$ is indefinite: it has only eleven nonnegative eigenvalues. In the middle left picture we took the tridiagonal part as approximation. This approximation is indeed positive definite and gives better results. Taking the simple algebraic approximation $A_0$ from below from Section 4.3.3 based on the three largest diagonal entries of $A$, which is of rank 6, gives comparable results, but its low rank and higher sparsity make this choice more interesting. In the right graph in Figure 4.5, the largest eigenvalue of $\alpha I - A$ was approximated with the approximation from below that kept the largest three diagonal entries of $\alpha I - A$, and thus the smallest three diagonal entries of $A$. In all cases, the positive effect of incorporating an approximation $A_0$ goes beyond delivering a good start vector for Lanczos. Also in all cases, there is virtually no difference between Full SPAM and SPAM(1).
Figure 4.4: From left to right: Lanczos and Full SPAM approximating the largest, the second, the fifth, and the smallest eigenvalue of $A$, using algebraic rank-12 approximation from below (for the smallest eigenvalue, we applied both the methods to $6I - A$). Lanczos and Full SPAM used in each experiment the same start vector.

4.4.4 Lanczos versus Full SPAM and SPAM(1)

Matrices from structural engineering

As a next set of experiments, we took some matrices from the Harwell-Boeing collection of test matrices. They have their origin in the area of structural engineering and are called bcsstk04, bcsstk07 and bcsstk10 and have respective sizes $132 \times 132$, $420 \times 420$ and $1024 \times 1024$. As approximating matrix $A_0$ we took approximations from below keeping respectively the largest 12, 20 and 180 diagonal entries. Recall that in Figure 4.2 we displayed the sparsity plots of $A$ and $A_0$ for bcsstk04. As was the case for the banded matrix in the previous section, Full SPAM and SPAM(1) behave virtually the same, and need less outer iterations than the Lanczos method to arrive at a given accuracy.

Conclusions The main goal of the experiments so far was, first of all, to investigate if Full SPAM is competitive with the Lanczos method in terms of the number of iterations in the outer loop. Solving the eigenproblems for $A_k$ in the inner loop, even if this would not be done to full precision, makes SPAM always more expensive than Lanczos. The experiments in Figures 4.3, 4.4, 4.5 and 4.6 show that
4.4. NUMERICAL ILLUSTRATIONS

Figure 4.5: Lanczos versus Full SPAM and SPAM(1) with diagonal (left), tridiagonal (middle left), and with algebraic rank-6 approximation from below (both middle right and right). In the three leftmost pictures, the target was the largest eigenvalue, at the right it was the smallest eigenvalue (i.e., the largest of $\alpha I - A$).

this is the case for different matrices $A$ and different types of approximations $A_0$. Not only for the largest, but also for other eigenvalues of $A$. Secondly, we have illustrated that the use of one step of the Jacobi-Davidson method to approximate the eigenvalue problem for $A_k$ in the inner iteration for SPAM hardly influences its behavior. We will now investigate what happens if the linear system in SPAM(1) is approximated with a small number of steps of the Minimal Residual method, MinRES [61].

4.4.5 Lanczos versus SPAM(1, $\ell$)

Approximating the correction equation of SPAM(1) using MinRES

In this section we investigate the use of $\ell$ iterations of the MinRES method [61] for approximating the solution of the Jacobi-Davidson correction equation (4.20) for the eigenproblem for $A_k$ in the inner iteration of SPAM(1). Each iteration of MinRES requires one matrix vector product with $A_k$, which, for small values of $k$, will be approximately as costly as a matrix vector product with $A_0$. The initial vector of all methods to which the comparison is applied, i.e., the eigenvector of interest of $A_0$, will still be computed in full precision. The resulting method is
Figure 4.6: Lanczos versus Full SPAM and SPAM(1) for bcsstk04, bcsstk07 and bcsstk10, with approximating matrices from below. All three methods used the same start vector.

Remark 4.4.2. In SPAM(1,1), MinRES produces an approximation of the Jacobi-Davidson correction equation (4.20) for $A_k$ in the one-dimensional Krylov subspace with the right-hand side of that correction equation as start vector. Since this is the current residual from the outer iteration, the expansion is the same as for the Lanczos method. We will therefore not display the convergence curves of SPAM(1,1).

In the light of the previous remark, it is reasonable to expect that SPAM(1,$\ell$) will represent a transition between Lanczos and SPAM(1). Thus, for reference, in the experiments to come, we displayed the convergence graphs of Lanczos and SPAM(1) in solid black lines without any additional symbols. The experiments concern four of the situations that we have already studied. First, we approximated the smallest eigenvalue of the reaction-diffusion problem. For the other three experiments we approximated the largest eigenvalue: in the second, we took the banded matrix with low rank approximation from below, and in the third and fourth the matrices bcsstk07 and bcsstk10 with the respective approximations from the previous section. The results are displayed in Figure 4.7 and confirm the expectations. Even for $\ell = 2$ and $\ell = 3$, SPAM(1,$\ell$) resembles SPAM(1) much.
more than it resembles Lanczos. It depends, however, very much on the actual application if the gain in the number of iterations is not undone by the costs of \( \ell \) steps of MinRES per outer iteration with the matrix \( A_k \). For instance, in the banded matrix example (second picture in Figure 4.7), the matrix \( A \) itself has 322 nonzero entries and is of full rank, whereas \( A_0 \) only has 33 nonzero elements and is of rank 6, and especially when \( k \) is small, the action of \( A_k \) will not be very much more expensive than the action of \( A_0 \). This, however, brings us to the next question that we would like to investigate, which is, if using \( A_k \) in the \( k \)-th inner iteration instead of \( A_0 \) all along, is going to make any difference, because this is what distinguishes SPAM from Jacobi-Davidson with one-step approximation as preconditioner. As argued in Section 4.2, if SPAM is going to be better, then probably not by very much.

Figure 4.7: Lanczos and SPAM(1) compared with SPAM(1,\( \ell \)) for small values of \( \ell \). Left: reaction-diffusion problem, smallest eigenvalue. Other pictures: largest eigenvalue. Middle left: banded matrix; middle right: bcsstk07; right: bcsstk10. The graphs for Lanczos and SPAM(1) can also be found in Figures 4.4, 4.5, and 4.6.

### 4.4.6 SPAM(1,\( \ell \)) versus one-step preconditioned JD

In the \( k \)-th inner iteration of SPAM(1,\( \ell \)), the Jacobi-Davidson correction equation (4.20) for \( A_k \) is solved using \( \ell \) steps of MinRES. We will now compare this with the Jacobi-Davidson method with one-step approximation as preconditioner, as
was described in Section 4.2. This means that in each inner iteration, the initial approximation $A_0$ is used instead of $A_k$. We will still apply $\ell$ steps of MinRES to solve the corresponding correction equation and denote the resulting method by JD(1,$\ell$). Since one of the aims of SPAM was to save on the costs of the matrix-vector products in the inner iteration, we will also apply Jacobi-Davidson without preconditioning, and approximate the exact correction equation (4.16) with the matrix $A$ by performing $\ell$ steps of MinRES as well, and denote this method by JD($\ell$). Thus, the differences between these three methods lie in the inner iteration: SPAM(1,$\ell$), JD(1,$\ell$) and JD($\ell$) all apply $\ell$ steps of MinRES per inner iteration step, to a linear equation with matrix $A_k$, $A_0$, and $A$, respectively. Note that in [69], no explicit comparison of SPAM with Jacobi-Davidson was made, even though the methods are so closely related.

![Figure 4.8: Comparing SPAM(1,$\ell$) with JD(1,$\ell$) and JD($\ell$).](image)

The results are shown in Figure 4.8. As expected, JD($\ell$) is the clear winner in all experiments, although the difference with JD(1,\ell) and SPAM(1,\ell) is not enough to automatically disqualify the latter two. Since the matrix-vector products in their inner iterations are, in general, considerably cheaper than in JD($\ell$), both methods could be competitive. Having said this, the difference between JD(1,$\ell$) and SPAM(1,$\ell$) is quite small and not always in favor of SPAM(1,$\ell$), even though...
4.5. CONCLUSIONS

SPAM(1,ℓ), much more so than JD(1,ℓ) uses the information that is available from the outer iteration also in its inner iteration. As argued already in Section 4.2, this may actually be less effective than using only the best information that is available from the outer iteration, as the Jacobi-Davidson method does. So far, the numerical experiments are in favor of using Jacobi-Davidson with one-step preconditioning instead of SPAM(1,ℓ).

4.5 Conclusions

The experiments above illustrate mathematical aspects of SPAM as a method for approximating eigenvalues of a Hermitian matrix. Using approximations from below, SPAM can be seen as a boosted version of the Lanczos method in the sense that convergence towards the largest eigenvalues is stimulated. Since Lanczos itself is often used to provide a good start vector for Jacobi-Davidson, SPAM is a therefore a good candidate for this task, too. Since the difference between SPAM and Jacobi-Davidson with one step approximation is small, it may be preferred to use the latter, especially since the latter is even more easy to use. There does not seem to be a significant gain in re-using the action of A on the orthogonal complement $U$ of the current Ritz vector $u$ within $V$ also in the inner iterations in comparison with only re-using the action of A on $u$, as Jacobi-Davidson with one step approximation does. This does not mean that the original idea of the authors [69] of SPAM, to save on the costs of the inner iterations of for instance Jacobi-Davidson, was incorrect. It may well pay off to do so, but this may be done with Jacobi-Davidson with one step approximation just as well. Thus, the main conclusion of this chapter is that the value of SPAM probably lies in providing good initial approximations for the Jacobi-Davidson method.
4.6 Comments on a SPAM linear system solver

The subspace projected approximate matrices $A_k$ could, at least in theory, also be used in the context of approximating the solution $x$ of a linear system of equations $Ax = b$. Indeed, given a sequence of subspaces $\mathcal{V}_1 \subset \cdots \subset \mathcal{V}_n$ and the corresponding matrices $A_0, A_1, \ldots, A_n$, an obvious idea is to define $x_k$ as the solution of $A_k x_k = b$, assuming that each $A_k$ is invertible. The initial approximation is then, just as in the eigenvalue case, the solution of $A_0 x_0 = b$, and following the approach in the eigenvalue case even further, the expansion vector $v$ should be chosen as the component of the approximation $x_k$ that is orthogonal to the current search space $\mathcal{V}_k$.

Remark 4.6.1. Approximation methods for linear systems are outside the scope of this thesis. For this reason, below we only provide an extended abstract of the paper [13], to which we refer the interested reader for further details.

Approximations based on the subspace $\mathcal{V}_k$

The main challenge is, that in order for the SPAM linear system solver proposed above to be competitive with methods like the Conjugate Gradient (CG) method [34] or the Minimal Residual (MinRES) method [61], it should be possible to solve the approximate systems $A_k x_k = b$ in a very efficient way. By definition of $A_k$, these systems have the following form,

$$
\begin{bmatrix}
V^t A V & V^t A V_{\perp} \\
V_{\perp}^t A V & V_{\perp}^t A_0 V_{\perp}
\end{bmatrix}
\begin{pmatrix}
y_1 \\
y_2
\end{pmatrix}
= 
\begin{pmatrix}
V^t b \\
V_{\perp}^t b
\end{pmatrix},
$$

(4.38)

after which the approximation $x_k$ equals

$$
x_k = V y_1 + V_{\perp} y_2.
$$

Note that $x_k$ is, in general, not an element from the search space $\mathcal{V}_k$, but rather an approximation based on $\mathcal{V}_k$. As such, the selection mechanism is different than
Ritz-Galerkin or a Minimal Residual type of selection, which would solve

\[ V^t A \hat{y}_1 = V^t b \quad \text{and} \quad \begin{bmatrix} V^t A \\ V_\perp^t A \end{bmatrix} (\tilde{y}_1) = \ell.s \begin{bmatrix} V^t b \\ V_\perp^t b \end{bmatrix}, \tag{4.39} \]

where \( \ell.s \) stands for the least-squares solution, and set \( \hat{x}_k = V \hat{y}_1 \) and \( \tilde{x}_k = V \tilde{y}_1 \), respectively. Many iterative solvers, such as CG and MinRES, are based on the selection mechanisms in (4.39). To set up the system (4.38), \( k \) matrix-vector products with \( A \) are needed, which is the same as in the other two approaches displayed in (4.39).

**Employing an \( A_0 \)-orthonormal basis for \( \mathcal{V}_k \)**

Assuming that \( A_0 \) is a symmetric positive definite approximation of \( A \), the basis of the search spaces can be chosen \( A_0 \)-orthogonal. With the \( \perp \)-sign indicating \( A_0 \) orthogonality, the bottom right block \( V_\perp^t A_0 V_\perp \) in (4.38) becomes the identity matrix. Using block elimination, it can be shown that the computation of \( x_k \) then needs additional \( k \) system solves with \( A_0 \). This is not unreasonable, as long as we would use \( A_0 \) also as preconditioner within CG method and MinRES. Alternatively, one could precondition the system \( Ax = b \) before setting up the system (4.38), and then substitute \( A_0 = I \) in the above block system, and use orthonormal bases for the search spaces with respect to the standard inner product. It is then possible to prove that expansion of the spaces with the component of \( x_k \) orthogonal to \( \mathcal{V}_k \) in fact leads to Krylov subspaces. The approximations, however, are not from those spaces, but as already mentioned, rather based on the spaces. They exploit all available information about the blocks \( V_\perp^t A V \) and \( V^t A \perp V \) and doing so, go one step further than CG and MinRES. Since these blocks, for Krylov subspaces, have only one single non-zero entry, the gain, although present, is small. For other subspaces, the gain can be more substantial. Again, for more details on this so-called LinSPAM method, see [13].
Summary

This chapter summarizes the contents of this thesis. The initial sections are intended for the reader not familiar with the concepts of eigenvalues and matrices and the connections between the two. The language used is accessible and accuracy and detail were sacrificed in favor of the understanding of the larger issues. The knowledgeable reader may wish to skip to the last section where a more specialized overview of the main results is given.
5.1 The mathematical model

The *natural frequency* of an object is the frequency at which such object vibrates when hit, touched or disturbed. While some objects might produce sound when vibrating such as the strings in a violin or crystal glasses, others might not. Nevertheless, every object has a natural frequency. Even larger objects such as cars and airplanes, buildings and bridges have natural frequencies. Buildings oscillate when a heavy truck passes down the street or a plane takes off close by. If a structure is not well designed it may vibrate excessively and may even collapse. A famous case was the Tacoma Narrows Bridge in the state of Washington in the United States. The bridge opened on July of 1940 and, not long after, started oscillate excessively. Five months later collapsed. The study of vibrations is, therefore, crucial in civil engineering.

Let us work with a very simple and typical example and see how the study of vibrations relates with the contents of this thesis. Our model is an apartment building with two stories represented in Figure 5.1.

![Figure 5.1: Two stories building. The elasticity of the walls is denoted by the $a_j$s, their displacement by the $\tilde{x}_j$s and the mass of the ceilings are denoted by the $m_j$s for $j \in \{1, 2, 3\}$.](image)

To make the model as simple as possible we neglect the mass of the walls and model only their elasticities $a_1, a_2$ and $a_3$ which are known quantities. Only
the ceilings have mass which we denote by $m_1, m_2$ and $m_3$ and assume to be the same. To simplify even more we assume that $m_1 = m_2 = m_3 = 1$. The horizontal movement of the ceilings (horizontal displacement) is represented by the functions $\tilde{x}_1, \tilde{x}_2$ and $\tilde{x}_3$. We assume further that the foundations of the building are strong enough and thus the base of the building is fixed to the ground. Using the appropriate physical laws governing forces, masses and displacement one can describe the interactions occurring in the building by a system of equations:

\[
\begin{cases}
(a_1 + a_2)x_1 - a_2x_2 &= \lambda m_1x_1 \\
-a_2x_1 + (a_2 + a_3)x_2 - a_3x_3 &= \lambda m_2x_2 \\
-a_3x_2 + a_3x_3 &= \lambda m_3x_3.
\end{cases}
\]  

(5.1)

The variables $a_j$ and $m_j$ in the system (5.1) are known but $\lambda$ and the $x_j$s are unknown. The system turns out to be even slightly more complicated than what it seems as this is a system of three equations but nine unknowns. In fact, there are three possible values for $\lambda$ and, for each of them, there is an associated solution made up of the elements $x_1, x_2$ and $x_3$. The actual solutions, however, depend always on the specific values of $a_1, a_2$ and $a_3$.

The number of equations in the system depends on the underlying problem. Adding a new floor to the building would result in an additional equation and an additional variable $a_4$ but in four extra unknowns (one extra $\lambda$ and an extra element $x_4$ and so to each $\lambda$ now corresponds a combination of $x_1, \ldots, x_4$).

The system we have just described is simple and small. More complicated structures, and different type of problems may take the number of equations to reach the billions! A famous one is the system arising from Google’s method for determining the order of importance of each webpage. It is estimated that such system of equations contains, currently, more that eleven billion ($11 \times 10^9$ or 11 followed by 9 zeros) equations [52]. (Note that the estimated size of the same matrix in 2002 was only about three billion [55]). Although such number of equations is still exceptional, standard problems have now about $10^6$ (1 million) equations.
5.2 From systems to matrices

With very small systems, say two or three equations such as the one showed earlier, it is still possible to keep control of all elements. One could, in fact, find a solution using just pen and paper. If, however, the number of equations or the number of unknowns increase, the task becomes more cumbersome. The need arose, thus, to develop a way to represent systems of equations that did not require one to write every element of the system every time. This was solved by grouping all common elements together and defining some rules for their manipulation. We can, thus, write the previous system in an equivalent manner as

\[
\begin{bmatrix}
  a_1 + a_2 & -a_2 & 0 \\
  -a_2 & a_2 + a_3 & -a_3 \\
  0 & -a_3 & a_3
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{bmatrix}
= \lambda
\begin{bmatrix}
  m_1 & 0 & 0 \\
  0 & m_2 & 0 \\
  0 & 0 & m_3
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{bmatrix}.
\]

(5.2)

The two large square elements are called *matrices* and are usually represented using capital letters of the alphabet, say A and M, respectively. The rectangular elements are named *vectors* and are usually denoted by lower case letters, for instance \( \vec{x} \). Using the new symbols, system (5.1) in the form of equation (5.2) can, indeed, be expressed simply as

\[
A\vec{x} = \lambda M \vec{x}.
\]

The size of a matrix, that is, the number of rows or columns is equal to the number of equations in the original system. Note that this can reach a few millions! Using the rules defined for the sum and product of matrices and vectors one can show that the solution of problem (5.2), assuming that \( m_1 = m_2 = m_3 = 1 \), is equivalent to the solution of the equation

\[
A\vec{x} = \lambda \vec{x}.
\]

(5.3)

The variable \( \lambda \) is called an *eigenvalue* (from the German “eigen” meaning proper, thus “proper value”) of \( A \) and \( \vec{x} \) is the *eigenvector* of \( A \) associated with the eigenvalue \( \lambda \). The number of eigenvalues of a matrix is equal to the number rows and columns of that matrix and, therefore, equal to the number of equations of the
original system. It may happen, however, that some of the eigenvalues are equal. The value of the natural frequency we mentioned in the beginning is obtained from the smallest of those eigenvalues of $A$.

**Perturbation theory** Perturbation theory for eigenvalues is another component of the study of eigenvalues. In practice, it is not possible to compute the exact eigenvalues of a matrix but just a close approximation. It is, therefore, of interest to understand what happens to $\lambda$ and $\vec{x}$ in the presence of errors for each $a_k$. In other words, what are the values of $\lambda$ and $\vec{x}$ if instead of $a_k$ the actual value is $a_k + e_k$ for some error $e_k$? We may pose the same question in terms of determining the eigenvalues and eigenvectors of a matrix closer to $A$, say $A + E$, for some error matrix $E$, i.e., to solve

$$ (A + E)\tilde{x} = \tilde{\lambda}\tilde{x}. $$

If, however, the eigenvalues and eigenvectors of $A$ are already known, the investment in computational effort to determine the eigenvalues of $A + E$ may be unnecessary. Perturbation theory attempts to determine the approximate location of the eigenvalues of $A + E$ without explicitly computing them.

**Iterative methods for eigenvalue problems** Because nowadays the typical matrix is of the order of the millions, computers are needed to determine the eigenvalues and eigenvectors. Iterative methods for eigenvalue problems are computational methods designed to determine a few of the eigenvalues and eigenvectors of a given matrix. These can be a few of the smallest, of the largest or just a few in between. They are called *iterative* because they repeat a certain set of steps over and over. Each step gets us closer to the solution and the method stops once (a very close approximation to) the final solution is found.
5.3 This thesis

Despite the somewhat negative connotation of the word, not every perturbation is a bad perturbation. In fact, while disturbing the matrix entries, many perturbations still preserve useful properties such as the orthonormality of the basis of eigenvectors or the Hermicity of the original matrix.

In the first part of this thesis some of these property preserving perturbations are analyzed. The ultimate goal is to understand how the eigenvalues of the original matrix are influenced by such perturbations. For Hermitian matrices, this resulted in improved Weyl-type bounds for the eigenvalues of Hermitian rank-$k$ perturbations. Their accuracy is dependent on the available eigenvalues and eigenvectors of the original matrix. The same approach is then used to determine bounds for the singular values of perturbations and augmentations of $n \times m$ matrices.

With respect to normal matrices, normality preserving normal perturbations of normal matrices were also considered. For $2 \times 2$ and for rank-one matrices, the analysis is now complete, while for higher rank, all essentially Hermitian normality perturbations are described. The normality preserving augmentation of normal matrices studied by Ikramov and Elsner is revisited and their results are complemented with an analysis on the consequences to the eigenvalues. All augmentations that result in normal matrices with eigenvalues on a quadratic curve in the complex plane are also constructed.

In the second part the Subspace Projected Approximate Matrix (SPAM) method, an iterative method for the Hermitian eigenvalue problem, is investigated from a Numerical Linear Algebra point of view. For certain special choices of the approximation matrix, SPAM turns out to be mathematically equivalent to the Lanczos method. More sophisticated approximations turn SPAM into a boosted version of the Lanczos method. It can also, however, be interpreted as an attempt to enhance a certain instance of the preconditioned Jacobi-Davidson method.
Notation

$I_n$ identity matrix of size $n$
$e_j$ $j$th standard basis vector in $\mathbb{C}^n$ or $\mathbb{R}^n$
$O, O_n$ zero matrix, zero matrix of size $n$
$\vec{0}$ zero vector
$[A, B]$ commutator of $A$ and $B$
$\lambda, \mu$ eigenvalues (usually)
$\rho_A(x)$ Rayleigh quotient
$\kappa_p(A)$ condition number of $A$ w.r.t. the matrix $p$-norm
$\oplus$ direct sum
$\mathcal{H}(A)$ Hermitian part of $A$
$\mathcal{S}(A)$ skew-Hermitian part of $A$
$i$ imaginary unit
$\sigma(A)$ spectrum of $A$
$\wp(z), \wp(A)$ polynomial in $z$ or in $A$ (page 62)
$\text{Co}(S)$ convex hull of the set $S$
$\Theta(z)$ the $\theta$-real part of $z$ (page 60)
$\Theta^\perp(z)$ the $\theta$-imaginary part of $z$ (page 60)
$\Theta(A)$ the $\theta$-Hermitian part of $A$ (page 60)
$\Theta^\perp(A)$ the $\theta$-skew-Hermitian part of $A$ (page 60)
$\mathcal{T}$ circle group of unimodular numbers (page 60)
$\Pi_k$ projector onto $\mathcal{V}^\perp$ (page 98)


Index

$\theta$-Toeplitz decomposition, 60
accelerated Newton method, 103
approximation from below, 27, 114
  algebraic construction, 115
  natural, 116
Bauer-Fike, 3
Cartesian decomposition, 60
Cauchy Interlace Theorem, 12
commutator, 22
condition number, 3
correction equation, 19
  alternative, 104
Courant-Fischer
  , see min-max theorem
divide-and-conquer, 8
field of values, 5
  $q$-field of values, 5
  generalized two-sided, 5
  Bauer, 5
  Givens, 5
generalized algebraic Riccati equation, 19
Harmonic Ritz
  value, 14
  values, 112
  vector, 14
Hoffman-Wielandt, 3
iterative method
  Eigenproblems, 13
  Full SPAM, 106
  Lanczos, 16
  one step SPAM, 106
  Riccati, 18
Linear Systems
  CG, 130
  MinRES, 119
  MinRes, 130
iterative method eigenproblems
  Jacobi-Davidson, 18
Krylov subspace, 17, 64
matrix
  1 bordered diagonal, 13
  \( \theta \)-Hermitian, 60
  \( \theta \)-Hermitian part, 60
  \( \theta \)-skew-Hermitian, 60
  \( \theta \)-skew-Hermitian part, 60
  arrow, 13
  arrowhead, 13
  Augmentation, 11
  Bordered diagonal, 39
  essentially Hermitian, 10, 61
  Hermitian part, 22
  rank-\( k \), 29
  rank-one, 29
  skew-Hermitian part, 22
  min-max theorem, 6
  Rayleigh-Ritz ratio, 5
  , see Rayleigh quotient
  Ritz
    pair, 14
    value, 14
    vector, 14
  search subspace, 13
  secular equation, 9
  subspace projected approximate matrix, 97
  Tacoma Narrows Bridge, 134
  Teoplitz decomposition, 60
  Weyl’s inequalities, 7

natural frequency, 134

one step approximation, 105
orthogonal correction, 18

preconditioner, 19, 104
preconditioning, 19
Principal submatrix, 11
  leading, 12
  trailing, 12

Rayleigh quotient, 4
  boundedness, 4
  homogenity, 4
  minimal residual, 4
  translation inv., 4
Rayleigh-Ritz
Samenvatting

Matrix Perturbaties - eigenwaarden begrenzen en bepalen

Perturbaties zijn verstoringen. Ondanks de wat negatieve klank van dit woord is niet elke perturbatie een slechte perturbatie. Er zijn perturbaties waaronder nuttige eigenschappen van de matrix behouden blijven, zoals bijvoorbeeld de orthonormaliteit van de basis van eigenvectoren of het Hermitisch zijn van de oorspronkelijke matrix.

Het eerste deel van dit proefschrift analyseert een aantal van zulke perturbaties. Het uiteindelijke doel is om te begrijpen hoe de eigenwaarden van de oorspronkelijke matrix door dergelijke perturbaties worden beïnvloed. Voor Hermitische matrizes resulteert het onderzoek in grenzen voor de eigenwaarden van Hermitische perturbaties van rang $k$, die de klassieke grenzen van Weyl en de recentere grenzen van Ipsen en Nadler verbeteren. Hun nauwkeurigheid is afhankelijk van het aantal beschikbare eigenwaarden en eigenvectoren van de oorspronkelijke matrix. Dezelfde aanpak leidt vervolgens tot grenzen voor de singuliere waarden van zowel perturbaties als uitbreidingen van rechthoekige matrices.

Vervolgens richt het onderzoek zich op normale perturbaties waaronder normaliteit behouden blijft. Voor $2 \times 2$ matrices en matrices van rang één is de analyse compleet, terwijl voor hogere rang alle essentieel Hermitische normaliteitsbehoudende perturbaties worden beschreven. Eerder werk van Ikramov en Elsner op het gebied van normaliteitsbehoudende uitbreidingen van matrices krijgt hier een verdieping, alsmede een aanvulling over de gevolgen voor de eigenwaarden. Het hoofdstuk besluit met de constructie van alle matrixuitbreidingen die resulteren in normale matrices met eigenwaarden op een kwadratische kromme in het complexe vlak.

Het tweede deel van dit proefschrift onderzoekt een iteratieve methode voor Hermitische eigenwaardeproblemen, genaamd de Subspace Projected Approximate Matrix (SPAM) methode, vanuit een numeriek lineair algebraisch oogpunt. Voor een aantal bijzondere keuzes van de benaderende matrix blijkt SPAM mathematisch equivalent te zijn aan de Lanczos methode. Meer geavanceerde benaderingen maken SPAM tot een opgevoerde versie van Lanczos, wat echter ook kan worden uitgelegd als een poging om een bepaalde instantie van de gepreconditioneerde Jacobi-Davidson methode te verbeteren.
About the Author

Ricardo Jorge Reis da Silva was born in Maia, Porto, Portugal in August 1982. He grew up in a town in the suburbs of Porto where he did his middle and high school studies.

In 2000, he was accepted for undergraduate studies in Applied Mathematics at the Faculty of Sciences of the University of Porto where he graduated four years later with a degree (Portuguese licenciatura - 4 year degree) in Applied Mathematics. He worked in Sogrape Vinhos as part of the Planning and Logistics department until 2005. In that same year Ricardo moved to Utrecht in the Netherlands to pursue his MSc in Scientific Computing. He completed his master studies in 2006 with a thesis entitled ”Is SPAM helpful?” in Numerical Linear Algebra under the supervision of Dr. Gerard Sleijpen. SPAM here stands for an iterative method for eigenvalue problems, the Subspace Projected Approximate Matrix method.

In 2007 Ricardo was admitted for a PhD position at the Korteweg-de Vries Institute for Mathematics at the University of Amsterdam also in the Netherlands. There, he continued his research leading to this thesis under the advice of Dr. Jan H. Brandts and Prof. Dr. Rob P. Stevenson in Numerical Linear Algebra with a focus in matrix theory, eigenvalue perturbation and iterative methods for eigenvalue problems. Also at the KdV institute, Ricardo was teaching assistant for Linear Algebra, Numerical Mathematics and Complex Analysis courses at the undergraduate level and Numerical Algorithms at the graduate level.

In 2008 and 2009 he co-organized the Phdays meetings for PhD candidates of the Netherlands and Flanders and during 2009 and 2010 organized the Numerical Mathematics Colloquium at the KdV institute.