Inexact methods for the solution of large scale Hermitian eigenvalue problems

vorgelegt von Diplom-Mathematikerin

Ute Kandler

Von der Fakultät II – Mathematik und Naturwissenschaften der Technischen Universität Berlin zur Erlangung des akademischen Grades Doktor der Naturwissenschaften – Dr. rer. nat. – genehmigte Dissertation.

Promotionsausschuss:

Vorsitzender: Prof. Dr. Martin Henk
1. Gutachter: Prof. Dr. Volker Mehrmann
2. Gutachter: Prof. Dr. Christian Mehl
3. Gutachter: Prof. Dr. Christopher Beattie

Tag der wissenschaftlichen Aussprache: 31. Juli 2019

Berlin 2019
Abstract

This thesis focuses on the solution of high dimensional Hermitian eigenproblems in situations where vector operations cannot be carried out exactly.

To this end an inexact Arnoldi method with the aim to approximate extreme eigenvalues and eigenvectors is developed. This method is particularly well-suited for large scale problems as it efficiently reduces the storage and computational requirements by constructing an orthonormal basis of the associated Krylov subspace. To consider the effect of the inexactness of the operations, a variant of Gram Schmidt orthogonalization is introduced in this thesis. The so called compensated Gram Schmidt orthogonalization (ComGS), if necessary enhanced by reorthogonalization, is used in the inexact Arnoldi method to compute a nearly orthogonal basis of the associated Krylov subspace. The effects of small perturbations on the distance to orthogonality of the resulting basis are studied and the ComGS method is compared to classical methods such as the modified and classical Gram Schmidt method.

To investigate the influence of the inexactness of the vector operations not only on the orthogonality of the Krylov subspace basis but also on the quality of the spectral approximations, the convergence behavior of the inexact Arnoldi method and more general of inexact Krylov subspace methods is analyzed. By using the concept of the angle of inclusion, bounds for the distance of the exact invariant subspace to (i) an inexact Krylov subspace and (ii) a Ritz space therein are found. While based on the first bound an a priori statement about how many more iteration steps of the inexact Krylov method are necessary to ensure convergence to a given tolerance can be derived, the second bound allows for an a posteriori statement regarding the quality of the computed Ritz space. To avoid the sensitivity of both bounds to the gap between the desired and the remaining eigenvalues the concept of nested subspaces is used in the a priori setting whereas in the a posteriori setting the Ritz space is allowed to be of larger dimension than the invariant subspace. Consequently, the bounds can be valid and meaningful even in the presence of perturbations and a small gap between the desired and remaining eigenvalues.

Lastly, the inexact Arnoldi method is generalized to the tensor setting to make storage and computations feasible. Numerical experiments with the YZ-model show that in the operator-tensor setting the inexact Arnoldi method allows for the solution of quantum systems of much larger dimension than it would be possible in the matrix-vector setting.
Acknowledgements

In the last years many great people have supported me and contributed to the research presented in this thesis. I am grateful for that.

First of all, I would like to thank my advisors Volker Mehrmann and Christian Schröder for introducing me into the interesting topic of eigenvalue problems, providing me guidance and supporting me with valuable advice and inspiring discussions. Further, I would like to thank Christian Mehl and Christopher Beattie for their acceptance to referee my thesis. I am grateful to Gernot Schaller, who helped me to bridge the gap between the physical and mathematical language. Thanks goes also to Daniel Kressner for several invitations to the EPF Lausanne, his advice and great hospitality, as well as to Michael Steinlechner for insightful discussions about tensor techniques. I am also grateful to the German research council DFG for supporting my work in the project "Scalable Numerical Methods for Adiabatic Quantum Preparation".

I would like to thank my colleagues and friends at the TU Berlin, in particular Andreas, Agnieszka, André, Ann-Kristin, Benjamin, Lena, Leo, Federico, Jan, Marine, Matthias, Philipp and Robert for all the valuable hints, lively discussions and making my time at the 4th floor so special. I am glad to have met you.

I also would like to thank my colleagues from the Fraunhofer Institute Axel, Denise, Franziska, Jörg and Maximilian for their constant support. My special thanks goes to Denise for trusting in me, driving me forward with loving care and being my friend.

I am especially grateful to my sister Anne for reading my thesis, everlasting encouragement, understanding me without words and providing me with confidence. I would like to thank my parents for supporting my education and offering me so many opportunities. Finally, I would like to thank Sven for his unconditional love, support and faith in me.
Contents

3.5.2 Upper bidiagonal matrix ........................................... 86
3.5.3 Diagonal matrix ..................................................... 87
3.6 Conclusion ............................................................. 89

4 Spectral perturbation .................................................. 90
4.1 Distance from a search space to an invariant subspace ............ 92
4.2 Convergence of the invariant subspace ............................... 103
4.3 Numerical results ..................................................... 110
  4.3.1 Academic example .............................................. 110
  4.3.2 Ising model .................................................... 119
4.4 Conclusion ............................................................. 125

5 Application of the inexact Arnoldi method to the tensor format .... 127
5.1 Preliminaries .......................................................... 128
  5.1.1 Notation and basic operations .................................. 128
  5.1.2 Tensor formats .................................................. 129
  5.1.3 TT format ....................................................... 130
5.2 Inexact Arnoldi method using TT format ............................ 135
5.3 Numerical Example: YZ-model .................................... 138
5.4 Conclusion ............................................................. 153

6 Summary and outlook .................................................. 155
# List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS</td>
<td>Alternating Linear Scheme</td>
</tr>
<tr>
<td>BLAS</td>
<td>Basic Linear Algebra Subprograms</td>
</tr>
<tr>
<td>CGS</td>
<td>Classical Gram Schmidt method</td>
</tr>
<tr>
<td>ComGS</td>
<td>Compensated Gram Schmidt method</td>
</tr>
<tr>
<td>LAPACK</td>
<td>Linear Algebra Package</td>
</tr>
<tr>
<td>MGS</td>
<td>Modified Gram Schmidt method</td>
</tr>
<tr>
<td>MPS</td>
<td>Matrix product states</td>
</tr>
<tr>
<td>MVecOP</td>
<td>Matrix vector operations</td>
</tr>
<tr>
<td>QTT</td>
<td>Quantized Tensor Train</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular value decomposition</td>
</tr>
<tr>
<td>TT</td>
<td>Tensor train</td>
</tr>
</tbody>
</table>
## List of Algorithms

<table>
<thead>
<tr>
<th>Section</th>
<th>Algorithm Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Classical Gram Schmidt method (CGS)</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>Arnoldi method</td>
<td>26</td>
</tr>
<tr>
<td>2.3</td>
<td>Modified Gram Schmidt method (MGS)</td>
<td>26</td>
</tr>
<tr>
<td>2.4</td>
<td>Rayleigh-Ritz method</td>
<td>27</td>
</tr>
<tr>
<td>3.1</td>
<td>Compensated Gram Schmidt method (ComGS)</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>ComGS method with reorthogonalization</td>
<td>43</td>
</tr>
<tr>
<td>3.3</td>
<td>Inexact Arnoldi method</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>Inexact QR decomposition</td>
<td>82</td>
</tr>
<tr>
<td>3.5</td>
<td>Inexact CGS method</td>
<td>82</td>
</tr>
<tr>
<td>3.6</td>
<td>Inexact CGS method with reorthogonalization</td>
<td>82</td>
</tr>
<tr>
<td>3.7</td>
<td>Inexact MGS method</td>
<td>82</td>
</tr>
<tr>
<td>3.8</td>
<td>Inexact MGS method with reorthogonalization</td>
<td>82</td>
</tr>
<tr>
<td>5.1</td>
<td>ComGS method using TT format</td>
<td>136</td>
</tr>
<tr>
<td>5.2</td>
<td>Selective orthogonalization</td>
<td>136</td>
</tr>
<tr>
<td>5.3</td>
<td>Inexact Arnoldi method using TT format</td>
<td>136</td>
</tr>
<tr>
<td>5.4</td>
<td>Rayleigh-Ritz method using TT format</td>
<td>137</td>
</tr>
</tbody>
</table>
List of Figures

3.1 Experiment 1: Loss of orthogonality in the inexact QR decomposition using CGS, ComGS, MGS ........................................... 84
3.2 Experiment 2: Loss of orthogonality in the inexact QR decomposition using CGSre, ComGSre, MGSre ........................................... 85
3.3 Experiment 3: Loss of orthogonality in the inexact QR decomposition using CGS, ComGS, MGS ........................................... 87
3.4 Experiment 4: Backward error of the inexact Arnoldi algorithm and its bounds ................................................................. 88
4.1 Illustration of the nested spectral subsets of Theorem 4.4 ................ 95
4.2 Qualitative shape of the bound (4.3) .................................... 101
4.3 Illustration of the spectral sets in Theorem 4.19 ......................... 106
4.4 Illustration of the spectral sets of Theorem 4.21 ......................... 108
4.5 Experiment 1: $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and bound (4.3) for case (i) .......... 111
4.6 Experiment 1: $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and bound (4.3) for case (ii) .......... 112
4.7 Experiment 2: $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and its bounds (4.3) for $\Lambda_2 = \Lambda_3 = \Lambda_4 = \Lambda_1$ ......................................................... 115
4.8 Experiment 3: $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and its bounds (4.3) for $\Lambda_2 = \Lambda_3 = \Lambda_4 \supset \Lambda_1$ ......................................................... 115
4.9 Experiment 4: $\angle_{\max}(\mathcal{X}_1, \tilde{Y})$ and corresponding bounds (4.13) for $\dim(\mathcal{X}_1) = \dim(\tilde{Y})$ and $\dim(\mathcal{X}_1) < \dim(\tilde{Y})$ ......................... 117
4.10 Experiment 5: Distance between the eigenvalues of interest $\Lambda$ and the corresponding Ritz values for $g = 1$ .................... 118
4.11 Experiment 5: $\angle_{\max}(\mathcal{X}, \tilde{Y})$ and the bounds (4.19) and (4.23) ...... 119
4.12 Linear chain of 10 spins with nearest-neighbor interaction and periodic boundary conditions ............................................... 120
4.13 Eigenvalue distribution of the Ising model (4.27) with $d = 10$, $J^z = 0.4$ and $g = 0.6$ ................................................. 122
4.14 Experiment 5: $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and its bounds (4.3) for $\ell \in \{60, 75, 90\}$ 123
4.15 Experiment 5: $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and its bounds (4.3) for adapted nested subspaces ..................................................... 123
4.16 Experiment 6: $\angle_{\max}(\mathcal{X}_1, \tilde{Y})$ and corresponding bound (4.13) .... 125
4.17 Experiment 7: Illustration of the a posteriori bound (4.23) ....... 126
5.1 Slices of a third order tensor ........................................... 129
5.2 Tensor network diagram of a TT tensor ................................ 131
5.3 Tensor network diagram of a TT operator ................................ 134
5.4 Phase diagram of a spin 1/2 transverse YZ-chain .......................... 139
5.5 **Experiment 1:** Inexact Arnoldi method for $\epsilon_v \in \{10^{-6}, 10^{-8}\}$ applied to the YZ-model with $d = 10$ .......................... 142
5.6 **Experiment 1:** Inexact Arnoldi method for $\epsilon_v \in \{10^{-10}, 10^{-12}\}$ applied to the YZ-model with $d = 10$ .......................... 143
5.7 **Experiment 2:** Residual-after-rounding and Arnoldi-residual of the three smallest Ritz values for $\epsilon_v \in \{10^{-6}, 10^{-8}\}$ and $d = 20$ .......................... 144
5.8 **Experiment 2:** Residual-after-rounding and Arnoldi-residual of the three smallest Ritz values for $\epsilon_v \in \{10^{-10}, 10^{-12}\}$ and $d = 20$ .......................... 144
5.9 **Experiment 2:** Development of the sum of TT ranks of the Arnoldi basis tensors and the time needed for each iteration step of the inexact Arnoldi method for $\epsilon_v \in \{10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\}$ and $d = 20$ .......................... 145
5.10 **Experiment 3:** Arnoldi-residual of the three smallest Ritz values for $\epsilon_v \in \{10^{-4}, 10^{-8}\}$ and $d = 30$ .......................... 147
5.11 **Experiment 3:** Arnoldi-residual of the three smallest Ritz values for $\epsilon_v = 10^{-12}$ and $d = 30$ .......................... 147
5.12 **Experiment 3:** Development of the sum of TT ranks of the Arnoldi basis tensors and the time needed for each iteration step of the inexact Arnoldi method for $\epsilon_v \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\}$ and $d = 30$ .......................... 148
5.13 **Experiment 4:** Arnoldi-residual of the three smallest Ritz values for $\epsilon_v \in \{10^{-4}, 10^{-6}\}$ and $d = 50$ .......................... 149
5.14 **Experiment 4:** Development of the sum of the TT rank of the Arnoldi basis tensors and the time needed for each iteration step of the inexact Arnoldi method for $\epsilon_v \in \{10^{-4}, 10^{-6}, 10^{-8}\}$ and $d = 50$ .......................... 149
5.15 **Experiment 5:** Maximum TT rank of the 70th Arnoldi basis tensor and the corresponding Arnoldi-residual for $d \in \{25, 30, \ldots , 60\}$ and $\epsilon_v = 10^{-6}$ .......................... 150
5.16 **Experiment 6:** Computed ground state energy and the energy of the first and second exited state of the YZ-model (5.7) .......................... 151
5.17 **Experiment 6:** Energy gaps of the two smallest states and the Arnoldi residual of the ground state and first excited state energy .......................... 152
5.18 **Experiment 6:** Sum of the TT ranks of the tensor taken over all computed Arnoldi basis tensors and the execution time of the inexact Arnoldi method for $m = 150$ .......................... 152
5.19 Region of convergence for $d = 20$ .......................... 154
List of Symbols

Basic Sets

\[ \mathbb{N}, \mathbb{N}_0 \]  
set of natural numbers, \( \mathbb{N}_0 = \mathbb{N} \cup \{0\} \), respectively

\[ \mathbb{R}, \mathbb{C} \]  
the fields of real and complex numbers, respectively

\[ \mathbb{F} \]  
set of IEEE 754 floating point numbers

\[ \Lambda \]  
set of eigenvalues

\[ \mathcal{M} \]  
set of Ritz values

\[ \text{eig}(A) \]  
eigenvalues of a matrix \( A \in \mathbb{C}^{n \times n} \)

Tensors, matrices, vectors, scalars

\[ I_n \]  
\( n \times n \) identity matrix (subscripts will be omitted if clear from the context).

\[ 0_{m \times n} \]  
\( m \times n \) zero matrix (subscripts will be omitted if clear from the context).

\[ A^T \]  
transpose of the matrix \( A \).

\[ A^H \]  
conjugate transpose of a matrix \( A \), \( A^H = \bar{A}^T \).

\[ A^{-1} \]  
inverse of a Matrix \( A \).

\[ A^+ \]  
Moore-Penrose-inverse of a Matrix \( A \).

\[ \text{ran}(A) \]  
image of a matrix \( A \in \mathbb{C}^{m \times n} \).

\[ \text{ker}(A) \]  
kernel of a matrix \( A \in \mathbb{C}^{m \times n} \).

\[ \text{rank}(A) \]  
rank of a matrix \( A \in \mathbb{C}^{m \times n} \).

\[ \text{env}(\mathcal{M}) \]  
envelope of a set \( \mathcal{M} \subseteq \mathbb{R} \cup \{\infty, -\infty\} \) is defined by the smallest interval that contains \( \mathcal{M} \).

\[ \sigma_{\max}(A), \sigma_{\min}(A) \]  
maximum, minimum singular value of a matrix \( A \in \mathbb{C}^{m \times n} \), respectively.

\[ \kappa(A) \]  
condition number of the matrix \( A \).
Spaces

\[ \|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\max_{\lambda \in \lambda(A(A^HA))} \lambda} \text{ for } A \in \mathbb{C}^{m \times n} \]

(2-norm, spectral norm)

\[ \|A\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} \text{ for } A \in \mathbb{C}^{m \times n} \]

(Frobenius norm)

\[ \varepsilon_v \] error caused by vector operations

\[ \varepsilon_s \] error caused by scalar operations

\[ \varepsilon_{\text{sum}} \] error caused by the orthogonalization step, i.e., the vector to be orthogonalized can be lengthen by the factor \((1 + \varepsilon_{\text{sum}})\)

\[ |S| \] number of elements in a discrete set \(S\) (cardinality of \(S\))

\[ \text{vec}(A) := (a_1^H, \ldots, a_n^H)^H \]

(vectorization of \(A = [a_1, \ldots, a_n] \in \mathbb{C}^{m \times n}\))

\[ A \otimes B := \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \]

(Kronecker product of \(A \in \mathbb{C}^{m \times n}\) and \(B \in \mathbb{C}^{p \times q}\))

\[ A \oplus B \] direct sum of vector spaces \(A \in \mathbb{C}^n\) and \(B \in \mathbb{C}^m\).

\[ x \] tensor of dimension \(n_1 \times n_2 \times \ldots \times n_d\)

\[ \rho(A) := \max_{i=1,\ldots,n} |\lambda_i| \]

(spectral radius of \(A \in \mathbb{C}^{n \times n}\))

\[ \mathcal{K}_k \]

:= \(\mathcal{K}_k(A, v_1)\) \(k\)-th Krylov subspace generated by the matrix \(A\) and the starting vector \(v_1\).

\[ \tilde{\mathcal{K}}_k \]

:= \(\mathcal{K}_k(A + E_k, \tilde{v}_1)\) \(k\)-th Krylov subspace generated by the matrix \(A + E_k\) and the starting vector \(\tilde{v}_1\).

\[ \mathcal{X} \] invariant subspace of \(A\).

\[ \mathcal{X}^\perp \] orthogonal complement of \(\mathcal{X}\) in \(\mathbb{C}^n\).

\[ \tilde{\mathcal{X}} \] invariant subspace of \(A + E_k\).
Functions

\[ \mathcal{Y} \]
Ritz space of \( A \) in \( K_k \).

\[ \tilde{\mathcal{Y}} \]
Ritz space of \( A + E_k \) in \( \tilde{K}_k \).

Measures

\[
\text{gap}(\Lambda_1, \Lambda_2) := \min_{\lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2} |\lambda_1 - \lambda_2|
\]

\[
\text{sep}_*(A, B) := \min_{Z \in \mathbb{C}^{m \times n}} \|A Z - Z B\|_*, \quad \text{where} \quad * \in \{2, F\}
\]

\[
\text{spread}(\Lambda) := |\min(\Lambda) - \max(\Lambda)|
\]

\[ O - \text{measure} \quad := \|I - V^H V\|_F \]

\[ vO - \text{measure} \quad := \|V^H v\|_2 \]

\[
\angle(x, y) \quad \text{acute angle between the vectors } x, y \in \mathbb{C}^n \text{ with } x, y \neq 0.
\]

\[
\angle(X, Y) \quad \text{largest principal angle between the subspaces } X, Y \subseteq \mathbb{C}^n.
\]

\[
\angle_{\text{max}}(X, Y) \quad \text{angle of inclusion of } X \text{ in } Y.
\]

Functions

\[ \psi(x) \]
Chebychev Polynomial.
List of Tables

3.1 Comparison of computational costs per step of CGS, ComGS and MGS ........................................... 49
3.2 Coefficients of the error bound (3.112) for different choices of $B_k$ ........................................... 76
3.3 Bounds on the norm of the perturbation of CGS and CGSre ........................................... 83
3.4 Bounds on the norm of the perturbation of MGS and MGSre ........................................... 83
4.1 Experiment 2: Constants of the bound (4.3) for $\Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4$, $\Lambda_1 \subset \Lambda_2 = \Lambda_3 = \Lambda_4$, and $\Lambda_1 \subset \Lambda_2 \subset \Lambda_3 = \Lambda_4$ ........................................... 114
5.1 The three smallest Ritz values and the Frobenius norm of the considered YZ-model for $d \in \{10, 20, 30, 50\}$ ........................................... 140
5.2 Gaps between the four smallest Ritz values for $d \in \{10, 20, 30, 50\}$ ........................................... 141
5.3 Experiment 1: Direct-residual after stagnation of the corresponding Ritz value and $\epsilon_v$ ........................................... 143
1 Introduction

The computation of eigenvalues, their corresponding eigenvectors or invariant subspaces of high dimensional matrices is one of the fundamental mathematical problems in science and engineering. The matrix eigenvalue problem consists of solving the equation

\[ Ax = \lambda x \]

for \( \lambda \in \mathbb{C} \), \( x \in \mathbb{C}^n \setminus \{0\} \) and is among the best studied problems in numerical linear algebra. In this context, extreme eigenvalues are of great importance for examining the behavior of high-dimensional problems characterized by a large system matrix \( A \). Such problems arise for example in applications as diverse as the following. We note that this list of applications is by no means exhaustive. In fact, advancements in the solution methodologies of eigenvalue problems increase the number of applications further.

- **Structural Engineering.** Modern structures, such as buildings or bridges, tend to increase in height and/or complexity in design but still have to be resistant for example against vibrations, see, e.g., [116]. The stiffness, mass distribution and damping of the structure considered are described by so called system matrices and the solutions of the corresponding eigenvalue problem, in particular the natural modes and eigenfrequencies, provide crucial stability information of the structure.

- **Ranking of results of search engines.** The PageRank algorithm assigns numerical weights, the PageRank, to each element of a hyperlinked set of documents (such as the world wide web) to reflect its relative importance within the set. To do so the set of documents is described by the rescaled adjacency matrix and the PageRank values of each document are the corresponding entries of the dominant right eigenvector, see, e.g., [124].

- **Acoustic field computations.** The emission of traffic noise by transport vehicles, such as cars, airplanes, etc., can be modeled via discretized partial differential, equations (PDEs) [76]. The analysis of the behavior of the system, i.e., the solution of the discretized PDEs, requires the solution of a nonlinear eigenvalue problem.

The focus of this thesis is the calculation of eigenvalues and eigenvectors of large quantum systems, which are necessary to understand phenomena like quan-
1 Introduction

tum phase transitions [98]. Quantum systems are collections of interacting particles described by a Hamiltonian. The Hamiltonian captures the total energy of the system and we are particularly interested in its ground state which corresponds to the most stable state of the system, i.e., the state with lowest energy. From the linear algebra point of view the ground state is represented by the eigenvector corresponding to the smallest eigenvalue of the Hamiltonian [52].

Unfortunately, quantum systems can exhibit phase transitions which are characterized by a potentially drastic change in the ground state even though the external control parameter moves smoothly across an interval. At a critical point, defined as the value of the control parameter where the phase transition occurs, the gap between the smallest and larger eigenvalues, the so called energy of excited states, cf., [52, 95], becomes small. For this purpose we aim to develop an algorithm which allows for the tracing of near-degenerated eigenvalues at critical points.

Further, an important characteristic of those quantum systems is that the dimension of the Hamiltonian, i.e., the matrix describing the energy of the corresponding quantum system, scales exponentially with the number of considered particles. Hence, calculating properties of a meaningfully sized quantum system is a computationally demanding task. In particular, to identify the ground states of quantum systems of 50 particles one has to solve eigenvalue problems of dimension $n = 2^{50}$. This so called curse of dimensionality makes the explicit storage of a vector in standard format unfeasible. To overcome this obstacle, the underlying structure of the eigenvalue problem has to be exploited, i.e., it is essential to assume that the vector can be represented in a tensor format of low rank. For example, if a vector $x$ of dimension $n^d = 2^{50}$ can be represented as a rank 1 product, i.e.,

$$x = x_d \otimes x_{d-1} \otimes \ldots \otimes x_1,$$

where $x_i \in \mathbb{R}^n, i = 1, \ldots, d$

the storage requirements are reduced from $n^d = 2^{50}$ entries to $nd = 100$ entries. Of course, for most problems classes this simple approximation is not exact enough. However, more advanced and sophisticated low rank tensor approximation techniques, such as the tensor train format, see, e.g., [51, 81, 85], or the hierarchical tensor formats, see, e.g., [40, 46, 68], have been developed. These techniques allow for the storage of vectors but at the costs that vector operations cannot be carried out exactly anymore. Instead only approximations of the desired quantities are available, which means by each vector operation an additional error of small norm is introduced. In such an inexact setting we aim to compute some extreme eigenvalues and corresponding eigenvectors of a large sparse matrices $A$.

One of the most important classes of eigenvalue solvers for high dimensional problems are iterative methods that search in a Krylov subspace.
1 Introduction

\[ \mathcal{K}_k := \mathcal{K}_k(A, v_1) := \text{span}(v_1, Av_1, A^2v_1, \ldots, A^{k-1}v_1), \]

cf., [74, 92, 119]. In particular, we focus in this thesis on Arnoldi’s method, see, e.g., [4, 5, 73], which is an orthogonal projection method, onto \( \mathcal{K}_k \). In more detail, in the \( j \)-th step the new basis vector \( v_j \) is constructed by orthonormalizing the vector \( Av_{j-1} \) against the previous basis vectors of \( \mathcal{K}_{j-1} \). Thereby, standard orthogonalization schemes like the classical Gram Schmidt method are very sensitive to perturbations. In particular, the presence of inexact vector operations may cause a drastic loss of orthogonality of the basis vectors, see, e.g., [37, 104], which can affect the accuracy of the computed spectral quantities.

In order to deal with situations where those perturbations occur we develop in this thesis a method to compute extreme eigenvalues and eigenvectors in the presence of inexact vector operations. To achieve this we introduce a variant of Gram Schmidt orthogonalization, called compensated Gram Schmidt method, which improves the orthogonality of the computed Krylov subspace basis in comparison to standard methods, such as the classical Gram Schmidt method. This is proven via bounds on the distance to orthogonality of computed Krylov subspace bases.

This compensated Gram Schmidt orthogonalization is embedded into the Arnoldi method to compute eigenvalues and eigenvectors in the presence of perturbations. Since the computed basis is not exactly orthogonal we are particularly interested in how the perturbations affect the quality of the approximated subspace. To address this issue, we derive bounds on the distance between an exact invariant subspace and a Krylov subspace as well as between an exact invariant subspace and a Ritz space. To measure these distances we use the concept of the angle of inclusion. In contrast to traditional bounds for these spectral quantities which are particularly sensitive to the gap between the wanted eigenvalues and the remaining spectrum, the derived bounds allow this gap to be small, which is achieved by considering how well the exact invariant subspace is contained in a slightly larger approximated invariant subspace.

Inspired by these results of the inexact Arnoldi method in the matrix-vector setting we generalize the ideas to the tensor setting, where we use the TT format to make storage and computations feasible.

All programs presented within this thesis are written in MATLAB. Numerical results based in the MATLAB implementations have been obtained on an infiniband cluster with 17 Opteron nodes. Each node contains of two DualCore-Opteron 2218 processors with 2.6 GHz connected to 16GB of shared memory.

In the following, an overview of this thesis is provided where new contributions are highlighted.
1 Introduction

Outline. Chapter 2 contains the theoretical basis of this thesis. We introduce the used notation and well-known results of spectral perturbation and approximation theory are recalled. The notation does not differ significantly from standard linear algebra notation used, e.g., in [38, 92], for a brief overview see the list of symbols on pages v-vii. To evaluate the quality of the result of the central algorithm developed in this thesis, the inexact Arnoldi’s method, we stated some well-established results regarding the approximation quality of eigenvalues and eigenvectors. For this we describe measures on the distance between two spectra respectively the nearness of two subspaces. In particular, we recall the concept of the principal angle and introduce the unsymmetric and more intuitive measure angle of inclusion and state some properties which are heavily used in this thesis. In order to set the baseline for Chapter 3 we recall Krylov subspace methods and focus in particular on the Arnoldi method that computes the search space $K_k$. Further, we introduce the Rayleigh-Ritz method that calculates based on $K_k$ the approximations of the wanted eigenvalues and corresponding eigenvectors.

Focusing on a scenario where all vector operations cannot be carried out exactly we investigate in Chapter 3 a modification of the Arnoldi method tailored to this scenario.

Since the orthogonality of the Arnoldi basis is essential for the accuracy of the projection and might affect the reliability of the computed eigenpair, we start our analysis by investigating how the effect of the inexactness on the orthogonality of the computed Arnoldi basis $\tilde{V}_k$ can be reduced. To do so the orthogonalization step is replaced by the compensated Gram Schmidt method respectively the compensated Gram Schmidt method with reorthogonalization, where in the case of inexact operations a non-standard projection is used to compensate the deviation from orthogonality caused by the perturbations. In the presence of inexact vector operations this leads to an improvement of the orthogonality but it does not lead to an exactly orthogonal Arnoldi basis. Therefore, the focus in the first part of this chapter is on the evaluation of the distance to orthogonality of the computed Arnoldi basis $\tilde{V}_k$. This is quantified by the distance of the cross product matrix to the identity $\| I - \tilde{V}^H \tilde{V} \|_F$ and the distance from a vector to a set of basis vectors $\| \tilde{V}^H \tilde{v} \|_2$.

Not achieving an exactly orthogonal basis implies that the Arnoldi relation is not valid for the computed subspace and it might even occur that the generated vectors stay out side of the Krylov subspace. However, in the second part of this chapter we investigate how the computed subspace basis $\tilde{V}$ can be interpreted as an exact subspace basis of a matrix close to $A$, i.e., for the computed subspace basis the Arnoldi relation is satisfied for a perturbed matrix $A + E_k$, where $E_k$ is small in norm. As the system matrix $A$ is Hermitian it is desirable to preserve this property, i.e., the backward error $E_k$ should be Hermitian too. However, this can be only guaranteed if we relax some conditions on the Arnoldi relation

$$(A + E_k)\tilde{V}_k = \tilde{V}_k \tilde{H}_k + \tilde{e}_{k+1} \tilde{h}_{k+1,k} e_k^T,$$
i.e., we replace the Hessenberg matrix $\tilde{H}_k$ by an Hermitian matrix $B_k$. To validate suitable choices of $B_k$ we investigate bounds for the norm of the backward error $E_k$. The results of this chapter are published to a great extent in [59].

In Chapter 4 we focus on the impact of the perturbations on the convergence of the Krylov subspace method. In particular, we investigate the quality of the computed spectral approximations and aim to answer the central question:

**How well is a desired invariant subspace $X$ of the matrix $A$ approximated by**

(i) the Krylov subspace $\tilde{K}_k$ of a perturbed matrix $A + E_k$ or

(ii) a Ritz space $\tilde{Y}$ which is a subspace of $\tilde{K}_k$?

The first part (i) of the question refers to the issue if a few iteration steps of a Krylov subspace method have been carried out how many more have to be performed to ensure convergence. Therefore, we derive in Section 4.1 an a priori bound on the angle of inclusion of the invariant subspace $X$ and $\tilde{K}_k$ that needs only information of the $\ell$-th iteration step and is able to handle perturbations and a small gap between the desired and the remaining eigenvalues.

The second part (ii) of the question is answered in an a posteriori setting, where information about the search space and the residual are explicitly available. For this we derive in Section 4.2 bounds on the angle of inclusion of the invariant subspace $X$ and the Ritz space $\tilde{Y}$ in terms of the corresponding Krylov subspace and in terms of the residual.

Assuming $\tilde{K}_k$ contains increasingly accurate approximations of the desired invariant subspace $X$, the bound in terms of the Krylov subspace indicates the convergence of the Ritz space to the invariant subspace. On the other hand the a posteriori bound depending in the residual provides information when to stop the iteration process. An important feature of the presented bounds is that they are still valid and meaningful in the presence of perturbations and a small gap between the wanted and the remaining spectrum, which is achieved by considering a slightly larger Ritz space. The spectral error bounds presented in this chapter have been published to a large extent in [60].

Finally, we show in Chapter 5 the applicability of inexact Arnoldi methods in a tensor setting. In particular, using the TTeMPS Toolbox [106] for the TT-format, makes the storage of the Arnoldi basis and the required computations possible. But the summation of tensors as well as the operator-tensor multiplication lead to an increase of the tensor rank that needs to be avoided. Therefore, this operations are followed by an approximation procedure where the accuracy can be chosen, which enables us to tune the algorithm to an either a more storage-efficient or more accurate solution. In the numerical examples we focus on the YZ-model to analyze the convergence behavior of the inexact Arnoldi method via considering the residuals of the three smallest Ritz values. We investigate the YZ-model for different problem sizes, i.e., for different number of particles and analyze the influence of the accuracy of the approximation procedure.
2 Preliminaries

In this chapter we introduce existing concepts of spectral approximation theory which provide the foundation of our approximation results. More precisely, we start by introducing the notation used in this thesis and state basic properties of the Hermitian eigenvalue problem. Since the numerical computation of eigenvalues is affected by rounding or truncation errors we review in Section 2.2 relevant concepts and results of the perturbation analysis of Hermitian eigenvalue problems. In Section 2.3 we state some results on the convergence behavior of Krylov subspace methods and close this chapter by the introduction of the Arnoldi method.

2.1 The Hermitian eigenvalue problem

The Hermitian eigenvalue problem consists of solving the equation

\[ Ax = \lambda x \]

for \( \lambda \in \mathbb{R}, x \in \mathbb{C}^n \setminus \{0\} \), where \( A = A^H \in \mathbb{C}^{n \times n} \) is given. We call \( \lambda \) an eigenvalue and \( x \) the eigenvector corresponding to \( \lambda \). The multiset of all eigenvalues of \( A \), denoted by \( \text{eig}(A) \), is called the spectrum of \( A \). The eigenvalues of \( A \) are not necessarily distinct. A distinct eigenvalue of \( A \) is called a simple eigenvalue, whereas a “repeated” eigenvalue is called a multiple eigenvalue. The eigenvectors corresponding to a multiple eigenvalue are not unique, but the subspace spanned by these vectors is \([38,92,119]\).

The mapping \( B = U^{-1}AU \), where \( B \) is a nonsingular matrix, is called a similarity transformation. The matrix \( B \) has the same spectrum as \( A \) and for each eigenvector \( x \) of \( A \) the corresponding eigenvector of \( B \) is given by \( U^{-1}x \).

For any matrix \( A \in \mathbb{C}^{n \times n} \) there exists a unitary matrix \( U \) such that

\[ \Lambda = U^H AU, \]

where \( \Lambda \) is an upper triangular matrix, see \([38,91]\). This transformation is called Schur decomposition.

If the matrix \( A \) is Hermitian then the Schur decomposition simplifies. More precisely, in this situation the matrix \( \Lambda \) is Hermitian as well and therefore diagonal. Moreover, it follows from

\[ \bar{\Lambda} = \Lambda^H = (U^H AU)^H = U^H A^HAU = U^H AU = \Lambda \]
that for each diagonal element $\lambda_i$ of $\Lambda$ we have $\tilde{\lambda}_i = \lambda_i$. Hence, the eigenvalues of an Hermitian matrix $A$ are real, i.e., $\text{eig}(A) \subseteq \mathbb{R}$, see, e.g., [92].

2.2 Perturbation of eigenvectors and eigenvalues

This section introduces some elementary aspects of perturbation theory. In particular we focus on the question of how the eigenvalues and corresponding eigenvectors vary if the matrix $A$ is perturbed. In the following, perturbations are considered as small variations of the original entries which are mainly caused by the rounding and truncation errors in numerical eigenvalue solvers. The rounding error results from using finite precision arithmetic whereas the truncation error result from using an approximation in place of an exact mathematical procedure. As we will demonstrate in Chapter 5 the latter error plays also an important role in tensor computation.

Given these error sources, the best we can hope for is that the eigenvalue solver computes exactly the eigenvalues of a slightly perturbed matrix, i.e., of $A + E$ with $\|E\|_2 \leq \varepsilon \|A\|_2$, where $\varepsilon$ is close to the machine precision $\varepsilon_{\text{mach}}$. Such an algorithm is called backward stable and all algorithms considered in this thesis share this characteristic. This naturally raises the question of how well the eigenvalues of the perturbed problem coincide with the eigenvalues of the original problem. In order to explore this we introduce in the following measures on the distance of different eigenvalues and eigenvectors. A more detailed discussion can be found for example in [112].

A subspace $\mathcal{X} \subseteq \mathbb{C}^n$ is called an invariant subspace of $A \in \mathbb{C}^{n \times n}$ if $A \mathcal{X} \subseteq \mathcal{X}$. If $A$ is Hermitian and $\mathcal{X}$ is an invariant subspace then the block spectral decomposition of $A$ is of the form

$$\begin{bmatrix} X, & X^\perp \end{bmatrix}^H A \begin{bmatrix} X, & X^\perp \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix},$$

(2.1)

with $A_{11} \in \mathbb{C}^{m \times m}$, $A_{22} \in \mathbb{C}^{(n-m) \times (n-m)}$ and $\text{eig}(A) = \text{eig}(A_{11}) \cup \text{eig}(A_{22})$, whereby $X \in \mathbb{C}^{n \times m}$, $X^\perp \in \mathbb{C}^{n \times (n-m)}$ describe the orthonormal bases of $\mathcal{X}$ and its orthogonal complement $\mathcal{X}^\perp$, respectively. The invariant subspace $\mathcal{X}$ is called simple if $\text{eig}(A_{11}) \cap \text{eig}(A_{22}) = \emptyset$.

The following result is known as the Hoffman-Wielandt Theorem for the Frobenius norm, e.g., [53, 122] and as Weyl’s theorem for the 2-norm, see [112, Corollary 4.10] and [53].

**Theorem 2.1.** Let $A, E \in \mathbb{C}^{n \times n}$ be Hermitian. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of $A$ and let $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \ldots \geq \tilde{\lambda}_n$ the eigenvalues of $A + E$. For $* \in \{2, F\}$ we have

$$\|\text{diag}(\lambda_1, \ldots, \lambda_n) - \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n)\|_* \leq \|E\|_*.$$

15
Theorem 2.2. [113, Theorem 3.3] Let $A$ be a Hermitian matrix, $\mathcal{X}$ an $m$-dimensional invariant subspace of $A$ and $\tilde{\mathcal{X}}$ an approximation of $\mathcal{X}$, so that $\mathcal{X} = \text{ran}(X)$, $\tilde{\mathcal{X}} = \text{ran}(\tilde{X})$ and $X^H X = \tilde{X}^H \tilde{X} = I$. Let

$$L_1 = X^H A X, \quad \tilde{L}_1 = \tilde{X}^H A \tilde{X} \quad \text{and} \quad R = A \tilde{X} - \tilde{X} \tilde{L}_1.$$ 

and suppose that the corresponding eigenvalues

$$\text{eig}(L_1) = \{\lambda_i\}_{i=1}^m \quad \text{and} \quad \text{eig}(\tilde{L}_1) = \{\tilde{\lambda}_i\}_{i=1}^m$$

are ordered non-increasingly. Then for any unitarily invariant norm $\|\cdot\|$ we have

$$\| \text{diag}(\lambda_1, \ldots, \lambda_m) - \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m) \| \leq \tan \angle_{\text{max}}(\mathcal{X}, \tilde{\mathcal{X}}) \|R\|,$$

with $\angle_{\text{max}}(\mathcal{X}, \tilde{\mathcal{X}}) = \max_{x \in \mathcal{X}, \tilde{x} \in \tilde{\mathcal{X}}} \min_{x \neq 0, \tilde{x} \neq 0} \angle(x, \tilde{x}).$

Further information regarding the concept of principal angles can be found in Section 2.2.2.

Remark 2.3. In [113, Theorem 3.3] the angle $\angle_{\text{max}}(\mathcal{X}, \tilde{\mathcal{X}})$ was assumed to be less than $\frac{\pi}{2}$. However, using that this angle cannot excess $\frac{\pi}{2}$ and interpreting $\tan\left(\frac{\pi}{2}\right)$ as being larger than every finite number lets the result be valid without this restriction.

In the following we will frequently make use of norms to quantify the consequences inflicted by perturbations. Therefore, in the following we quantify the change in the inverse of $A$ as a function of change in $A$.

Lemma 2.4. [38, Lemma 2.3.3] Let $\star \in \{2, F\}$ and $P \in \mathbb{R}^{n \times n}$ with $\|P\|_{\star} < 1$. Then

(i) $I - P$ is nonsingular,

(ii) $(I - P)^{-1} = \sum_{i=0}^{\infty} P^i$,

(iii) $\|(I - P)^{-1} - I\|_{\star} \leq \frac{\|P\|_{\star}}{1 - \|P\|_{\star}}$, and

(iv) $\|(I - P)^{-1}\|_{\star} \leq \frac{1}{1 - \|P\|_{\star}}$.

Using $P := -A^{-1}\delta A$ this result is extended to general matrices by the following lemma.

Lemma 2.5. (Perturbation Lemma) [38, Lemma 2.3.4] Let $\star \in \{2, F\}$, $A \in \mathbb{R}^{n \times n}$ nonsingular, and $\delta A \in \mathbb{R}^{n \times n}$ satisfy $\|A^{-1}\|_2 \|\delta A\|_2 < 1$, then $A + \delta A$ is nonsin-
2.2 Perturbation of eigenvectors and eigenvalues

...regular and

\[ \| (A + \delta A)^{-1} \|_* \leq \frac{\| A^{-1} \|_*}{1 - \| A^{-1} \|_2 \| \delta A \|_2}, \]

\[ \| (A + \delta A)^{-1} - A^{-1} \|_2 \leq \frac{\| A^{-1} \|_2^2 \| \delta A \|_2}{1 - \| A^{-1} \|_2 \| \delta A \|_2}. \]

A consequence of Lemma 2.5 is that under small perturbations, more precisely if \( \| \delta A \|_* < 1/\| A^{-1} \|_* \), the matrix \( A \) remains nonsingular.

2.2.1 Measurement of the distance between two spectra

The sensitivity of an invariant subspace to perturbation depends on the separation of the eigenvalues of interest from the remaining ones of the spectrum. To measure the distance between two spectra, we introduce the following two quantities.

First, the gap between closed sets \( \Lambda_1, \Lambda_2 \subset \mathbb{C} \) respectively square matrices \( A, B \) is defined by

\[ \text{gap}(\Lambda_1, \Lambda_2) := \min_{\lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2} |\lambda_1 - \lambda_2| \quad \text{and} \quad \text{gap}(A, B) := \text{gap}(\text{eig}(A), \text{eig}(B)). \]

Note that with (2.1) \( \text{gap}(A_{11}, A_{22}) = \text{gap}(\text{eig}(A_{11}), \text{eig}(A) \setminus \text{eig}(A_{11})) \) if \( \text{ran}(X) \) is a simple invariant subspace. The gap is a straightforward way to describe the distance between two spectra.

Second, the separation measures in some sense the distance between two matrices. More precisely, let \( A \in \mathbb{C}^{n \times n}, B \in \mathbb{C}^{m \times m} \) be two arbitrary square matrices, the separation is defined by

\[ \text{sep}_*(A, B) := \min_{\|Z\|_* = 1} \|ZA - BZ\|_* , \quad \text{where} \quad * \in \{2, F\}. \]

Elsewise, following [55], let \( S \) be the Sylvester operator defined by

\[ S(Z) = ZA - BZ. \]

Hence, the eigenvalues of the linear operator \( S \) are the pairwise differences of the eigenvalues of \( A \) and \( B \), i.e.,

\[ \Lambda(S) = \Lambda(A) - \Lambda(B). \]

If we assume that the physical separation \( \text{gap}(A, B) = \min \|\Lambda(S)\| \) is greater than...
2.2 Perturbation of eigenvectors and eigenvalues

zero then \( S \) is nonsingular. Thus, it follows from (2.2) that

\[
\text{sep}^{-1}(A, B) \geq \max_{Z \in \mathbb{R}^{n \times m}} \frac{\|ZA - BZ\|_F^{-1}}{\|Z\|_2} = \|S^{-1}\|_F.
\]

Using that the spectral radius \( \rho(S^{-1}) \) is bounded by \( \|S^{-1}\|_F \), cf. [112, Theorem II.2.6], and equation (2.3) it follows that

\[
\text{sep}^{-1}(A, B) = \rho(S^{-1}) = \max |\Lambda(A) - \Lambda(B)|^{-1},
\]

and thus

\[
\text{sep}(A, B) \leq \min |\Lambda(A) - \Lambda(B)|.
\]

In other words, the separation \( \text{sep}(A, B) \) is not greater than the physical separation \( \text{gap}(A, B) \). The great advantage of the function \( \text{sep}(A, B) \) compared to \( \text{gap}(A, B) \) is the Lipschitz continuity. Furthermore, the separation between two subspaces \( \mathcal{X}, \mathcal{Y} \subseteq \mathbb{C}^n \) with respect to a matrix \( A \in \mathbb{C}^{n \times n} \) is defined by

\[
\text{sep}_{A,*}(\mathcal{X}, \mathcal{Y}) := \text{sep}_{*}(X^HA, Y^HAY),
\]

where \( X \) and \( Y \) are any orthonormal bases for \( \mathcal{X} \) and \( \mathcal{Y} \). This quantity is well defined because the norms used in (2.2) are unitarily invariant, cf. [109]. Further details of the separation can be found in, e.g., [55, 112]. In the case of Hermitian matrices the sep and the gap operators are related as follows.

**Lemma 2.6.** Let \( A_{11} \in \mathbb{C}^{n \times n} \) and \( A_{22} \in \mathbb{C}^{m \times m} \) be Hermitian. Then

\[
\frac{2}{\pi} \text{gap}(A_{11}, A_{22}) \leq \text{sep}_2(A_{11}, A_{22}) \leq \text{gap}(A_{11}, A_{22}) = \text{sep}_F(A_{11}, A_{22}).
\]

**Proof.** The first inequality is discussed in [11, p.15], where the proof is based on a function-theoretical result shown in [115]. The second inequality is obtained by restricting the matrix \( Z \) in (2.2) to the form \( Z = u_i v_j^H \), where \( u_i \in \mathbb{C}^n \) and \( v_j \in \mathbb{C}^m \) are a unit (right) eigenvector of \( A_{11} \) and a unit (left) eigenvector of \( A_{22} \), respectively. More precisely, it follows from \( A_{11}u_i = u_i \lambda_i \) and \( v_j^HA_{22} = \mu_j v_j^T \) with \( \|u_i\|_2 = 1 \) and \( \|v_j\|_2 = 1 \) that

\[
\text{sep}_2(A_{11}, A_{22}) = \min_{\|Z\|_2=1} \|A_{11}Z - ZA_{22}\|_2 \\
\leq \min_{i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}} \|A_{11}u_i v_j^H - u_i v_j^H A_{22}\|_2 = \min_{i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}} \|\lambda_i u_i v_j^H - \mu_j u_i v_j^H\|_2 \\
\leq \min_{i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}} |(\lambda_i - \mu_j)\|u_i v_j^H\|_2 = \text{gap}(A_{11}, A_{22}).
\]

The last relation is stated in [112, Theorem 3.1].

\[\square\]
2.2 Perturbation of eigenvectors and eigenvalues

**Remark 2.7.** The second inequality of Lemma 2.6 is also valid in the case of non-Hermitian matrices $A_{11}$ and $A_{22}$. In that case the nature of the eigenvalues becomes more important: as indicated in the proof above $u_i$ has to be the right and $v_j$ the left eigenvector. Moreover, the first and the last relation in Lemma 2.6 also hold for normal $A_{11}$ and $A_{22}$ if $\frac{2}{\pi}$ is replaced by $\frac{4}{\pi}$, but are not valid in the non-normal case.

### 2.2.2 Measurement of the nearness of two subspaces

If we want to assess the quality of an approximate invariant subspace, we need a measure of the distance between two subspaces. In particular, we are interested in a measure that is independent of the choice of the bases of the two subspaces.

#### 2.2.2.1 Principal Angles

According to [112, p.45] and [88] the concept of angles between subspaces goes back to [58]. In numerical analysis the concept of principal angles has been shown to be essential for the perturbation theory of invariant subspaces, e.g., [27, 89, 92, 112]. Moreover, principal angles appear naturally to estimate how close an approximate eigenspace is to the true eigenspace, cf. [64].

In detail, the acute angle between two nonzero vectors $x, y \in \mathbb{C}^n$ is defined by

$$
\angle(x, y) = \arccos \frac{|(x, y)|}{\|x\|_2 \|y\|_2}.
$$

It is important to note that by definition $0 \leq \angle(x, y) \leq \frac{\pi}{2}$. To define the concept of principal or canonical angles between subspaces we recursively extend the definition of the acute angle between two vectors, see, e.g., [15, 39].

**Definition 2.8.** Let $\mathcal{X} \subset \mathbb{C}^n$ and $\mathcal{Y} \subset \mathbb{C}^n$ be subspaces with $\dim(\mathcal{X}) = p$ and $\dim(\mathcal{Y}) = q$. Let $p \geq q \geq 1$. The principal angles between $\mathcal{X}$ and $\mathcal{Y}$

$$
\Theta(\mathcal{X}, \mathcal{Y}) = [\theta_1, \ldots, \theta_q], \quad \text{where} \quad \theta_k \in \left[0, \frac{\pi}{2}\right], \quad k = 1, \ldots, q,
$$

are recursively defined by

$$
\cos(\theta_k) = \max_{x \in \mathcal{X}, y \in \mathcal{Y}} \max_{i,k} |x^Hy| = |x^Hy_k|,
$$

subject to

$$
\|x\|_2 = \|y\|_2 = 1, \quad x^Hx_i = 0, \quad y^Hy_i = 0, \quad , i = 1, \ldots, k - 1.
$$

The vectors $\{x_1, \ldots, x_q\}$ and $\{y_1, \ldots, y_q\}$ are called the principal vectors.
2.2 Perturbation of eigenvectors and eigenvalues

Definition 2.8 allows for a recursive determination of the complete sets of principal angles and principal vectors and in order to construct these sets we carry out the following steps. We assume that in the first step the maximum is obtained for the principal angle $\theta_1$ and the corresponding vectors $x_1$ and $y_1$. In the second step the search space is orthogonal to $x_1$ and $y_1$, respectively. Continuing in this way until one of the search spaces is empty leads to a complete set of principal angle and vectors. We note that in contrast to the principal vectors the principal angles are always uniquely defined.

Since we are interested in a measure for the quality of an approximate invariant subspace we aim to find the largest principal angle $\theta_\phi$. According to Definition 2.8 the largest principal angle is referred to as $\theta_\phi = \max_{(X, Y)}$. Considering the case where the dimensions of the subspaces $X$ and $Y$ are equal, i.e., $p = q$, the definition above is equivalent to the more intuitive representation

$$\max_{(X, Y)} = \max_{x \in X, \ y \in Y} \min_{x \neq 0, \ y \neq 0} \angle(x, y)$$

(2.4)

see, [110,112]. In the following we develop this measure further in order to make it suitable for subspaces of different dimension. More precisely, we change the question from “How close is $X$ to $Y$?” into “How well is $X$ contained in $Y$?”.

2.2.2.2 The angle of inclusion

**Definition 2.9.** For nonzero subspaces $X, Y \subset \mathbb{C}^n$ the angle of inclusion of $X$ in $Y$ is defined by

$$\angle_{\text{max}}(X, Y) := \max_{x \in X, \ y \neq 0} \angle(x, y)$$

where $\angle(x, y) := \min_{y \in Y, \ y \neq 0} \angle(x, y)$. (2.5)

For matrices $X, Y$ we define $\angle_{\text{max}}(X, Y) := \angle_{\text{max}}(\text{ran}(X), \text{ran}(Y))$. Intuitively, the angle of inclusion (2.5) provides a measure of how well $X$ is contained in $Y$. If this angle is small then for every $x \in X$ there exists a $y \in Y$ that is close to $x$, i.e., the distance $\|x - y\|_2 / \|x\|_2$ is small. Especially, we have $\angle_{\text{max}}(X, Y) = 0$ if and only if $X \subset Y$ and $\angle_{\text{max}}(X, Y)$ reaches $\pi/2$ if $X$ contains a direction orthogonal to $Y$. The latter is the case whenever $\dim(X) > \dim(Y)$. Using this motivation, the angle of inclusion can be extended to zero-dimensional spaces, i.e., $\angle_{\text{max}}(\{0\}, Y) := 0$ for any (zero or non-zero) $Y$, and $\angle_{\text{max}}(X, \{0\}) := \pi/2$ for any non-zero $X$.

We also stress that in general the angle of inclusion is non-symmetric, i.e., $\angle_{\text{max}}(X, Y) \neq \angle_{\text{max}}(Y, X)$. This is an appropriate property as for example a 2-dimensional subspace can be contained in a 3-dimensional subspace but obviously not vice versa. However, the angle of inclusion is symmetric whenever $\dim(X) = \dim(Y)$, i.e., when it coincides with the maximal principal angle (2.4). More precisely, the angle of inclusion and the principal angle between subspaces

20
are related by $\angle_{\max}(\mathcal{X}, \mathcal{Y}) = \min\{\angle_{\max}(\mathcal{X}, \mathcal{Y}), \angle_{\max}(\mathcal{Y}, \mathcal{X})\}$, see, e.g., [120].

As a side note, there is also a minimal angle, which is defined by

$$\angle_{\min}(\mathcal{X}, \mathcal{Y}) := \min_{x \in \mathcal{X}, x \neq 0} \angle(x, \mathcal{Y}),$$

but this concept does not play a role in this thesis. For more details on subspace angles, see, e.g., [15, 29, 63, 65, 112, 120].

In the following lemma we summarize the properties of the angle of inclusion that are used in this thesis.

**Lemma 2.10.** Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be subspaces of $\mathbb{C}^n$. Then

i). $\angle_{\max}(\mathcal{X}, \mathcal{Z}) \leq \angle_{\max}(\mathcal{X}, \mathcal{Y}) + \angle_{\max}(\mathcal{Y}, \mathcal{Z})$

(triangle inequality);

ii). $\angle_{\max}(\mathcal{X}, \mathcal{Y}) \geq \angle_{\max}(\mathcal{X}, \mathcal{Z})$

whenever $\mathcal{Y} \subset \mathcal{Z}$;

iii). $\angle_{\max}(\mathcal{X}, \mathcal{Y}) = \angle_{\max}(\mathcal{Y}^\perp, \mathcal{X}^\perp)$

where $\mathcal{X}^\perp, \mathcal{Y}^\perp \subset \mathbb{C}^n$ are the orthogonal complements of $\mathcal{X}, \mathcal{Y}$ in $\mathbb{C}^n$;

iv). $\angle_{\max}(\mathcal{X}, \mathcal{Y}) = \begin{cases} 0 & \text{if } \dim(\mathcal{X}) = 0 \\ \arccos\left(\sigma_{\dim(\mathcal{X})}(X^H Y)\right) & \text{if } 1 \leq \dim(\mathcal{X}) \leq \dim(\mathcal{Y}), \\ \frac{\pi}{2} & \text{if } \dim(\mathcal{X}) > \dim(\mathcal{Y}) \end{cases}$

where $X, Y$ are any orthonormal bases of the subspaces $\mathcal{X}, \mathcal{Y}$ respectively, i.e., $\mathcal{X} := \text{ran}(X), \mathcal{Y} := \text{ran}(Y)$, and $\sigma_i(X^H Y)$ denotes the $i$-th largest singular value of $X^H Y$;

v). $\angle_{\max}(\mathcal{X}, \mathcal{Y}^\perp \cap \mathcal{Z}) \leq \angle_{\max}(\mathcal{X} \oplus \mathcal{Y}, \mathcal{Z})$

whenever $\mathcal{X} \perp \mathcal{Y}$.

**Proof.** i). To prove the first property, we distinguish between three cases. First we assume that at least one of $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ is zero. If $\mathcal{X} = \{0\}$ we obtain $0 \leq \angle_{\max}(\mathcal{Y}, \mathcal{Z})$. For $\mathcal{Y} = \{0\}$ it follows that $\angle_{\max}(\mathcal{X}, \mathcal{Z}) \leq \frac{\pi}{2}$ and finally if $\mathcal{Z} = \{0\}$ we have $\frac{\pi}{2} \leq \angle_{\max}(\mathcal{X}, \mathcal{Y}) + \frac{\pi}{2}$. Hence, if one of the subspaces $\mathcal{X}, \mathcal{Y}$ or $\mathcal{Z}$ is zero the triangle inequality i) is valid. Second we consider the case that $\dim(\mathcal{X}) > \dim(\mathcal{Y})$ or $\dim(\mathcal{Y}) > \dim(\mathcal{Z})$. In this case at least one angle on the right-hand side reaches $\frac{\pi}{2}$ and there is nothing to prove. Third we assume that $1 \leq \dim(\mathcal{X}) \leq \dim(\mathcal{Y}) \leq \dim(\mathcal{Z})$. Thus the definition of $\angle_{\max}(\mathcal{X}, \mathcal{Y})$ coincides with the definition of angles between subspaces in [120, p. 275]. Hence in this case the proof of the metric property 3 in [120, p. 276] applies to our definition as well.

ii). This property follows from the triangle inequality i), with $\angle_{\max}(\mathcal{Y}, \mathcal{Z}) = 0$.

iii). To prove Property iii) we again distinguish between three cases. First, if $\mathcal{X} \subset \mathcal{Y}$ or, equivalently, $\mathcal{Y}^\perp \subset \mathcal{X}^\perp$, then it follows that both angles are zero.
This covers also the cases that $\mathcal{X} = 0$ or $\mathcal{Y}^\perp = 0$. Second, we assume that $\dim(\mathcal{X}) > \dim(\mathcal{Y})$ or, equivalently, $\dim(\mathcal{Y}^\perp) > \dim(\mathcal{X}^\perp)$. Then both angles are $\pi/2$ which covers also the cases that $\mathcal{Y} = 0$ or $\mathcal{X}^\perp = 0$. Third, considering the case $1 \le \dim(\mathcal{X}) \le \dim(\mathcal{Y})$ the proof can be found in [63, p.382].

iv). The first case $\dim(\mathcal{X}) = 0$ of property iv) holds by definition. The middle case ($1 \le \dim(\mathcal{X}) \le \dim(\mathcal{Y})$) can be found for example in [65, p.2010] and the last case $\dim(\mathcal{X}) > \dim(\mathcal{Y})$ follows by the fact that if $\mathcal{X}$ contains a direction orthogonal to $\mathcal{Y}$ the angle $\angle_{\max}(\mathcal{X}, \mathcal{Y})$ reaches $\frac{\pi}{2}$.

v). To prove this property, we assume w.l.o.g. that $\mathcal{X}$ is non-zero (otherwise the angle on the left hand side is zero and there is nothing to prove). Also, w.l.o.g. we assume that $\dim(\mathcal{X}) \le \dim(\mathcal{Y}^\perp \cap \mathcal{Z})$ (otherwise $\dim(\mathcal{X}) > \dim(\mathcal{Y}^\perp \cap \mathcal{Z})$ implies

$$
\dim(\mathcal{X} \oplus \mathcal{Y}) = \dim(\mathcal{X}) + \dim(\mathcal{Y}) > \dim(\mathcal{Y}^\perp \cap \mathcal{Z}) + \dim(\mathcal{Y}) \\
\ge (\dim(\mathcal{Y}^\perp) + \dim(\mathcal{Z}) - n) + \dim(\mathcal{Y}) \\
= \dim(\mathcal{Z}),
$$

and both angles in the claimed inequality are $\pi/2$ and there is nothing to prove). So we have $1 \le \dim(\mathcal{X}) \le \dim(\mathcal{Y}^\perp \cap \mathcal{Z})$ which implies that $\mathcal{Z}$ and $\mathcal{Y}^\perp$ are non-zero. W.l.o.g., also $\mathcal{Y}$ is non-zero (otherwise both angles in the claimed inequality reduce to $\angle_{\max}(\mathcal{X}, \mathcal{Z})$ and there is nothing to prove). Finally, we can assume w.l.o.g. that $\dim(\mathcal{X} \oplus \mathcal{Y}) = \dim(\mathcal{X}) + \dim(\mathcal{Y}) \le \dim(\mathcal{Z})$ (otherwise the angle on the right-hand side in the claimed inequality is $\pi/2$ and there is nothing to prove).

Let $X \in \mathbb{C}^{n \times \dim(\mathcal{X})}$, $Y \in \mathbb{C}^{n \times \dim(\mathcal{Y})}$, $Z \in \mathbb{C}^{n \times \dim(\mathcal{Z})}$ be orthonormal bases of $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, respectively, where $Z = [Z_1, Z_2]$ is chosen such that $\text{ran}(Z_1) = \mathcal{Y}^\perp \cap \mathcal{Z}$. Then $[X, Y]^H[Z_1, Z_2]$ is an orthonormal basis of $\mathcal{X} \perp \mathcal{Y}$, since $\mathcal{X} \perp \mathcal{Y}$. Moreover, $W := [X, Y]^H[Z_1, Z_2]$ is of the form $[W_{11} \quad W_{12} \cr W_{21} \quad W_{22}]$, where $W_{11} = X^H Z_1$ does not have fewer columns than rows (since $1 \le \dim(\mathcal{X}) \le \dim(\mathcal{Y}^\perp \cap \mathcal{Z})$) and $W_{22} = Y^H Z_2$ has full column rank. Also $W$ does not have less columns than rows (since $1 \le \dim(\mathcal{X} \oplus \mathcal{Y}) \le \dim(\mathcal{Z})$). Hence, with part iv). we have

$$
\cos \angle_{\max}(\mathcal{X}, \mathcal{Y}^\perp \cap \mathcal{Z}) = \sigma_{\dim(\mathcal{X})}(W_{11}) \quad \text{and} \quad \sigma_{\dim(\mathcal{X} \oplus \mathcal{Y})}(W) = \cos \angle_{\max}(\mathcal{X} \oplus \mathcal{Y}, \mathcal{Z}).
$$

Using the monotonically decreasing behavior of the cosine in the interval $[0, \frac{\pi}{2}]$ all that remains to prove is that $\sigma_{\dim(\mathcal{X})}(W_{11}) \ge \sigma_{\dim(\mathcal{X} \oplus \mathcal{Y})}(W)$. We distinguish two cases. If $W_{22}$ has more rows than columns, then the rows of $W$ must be linearly dependent, i.e., $\sigma_{\dim(\mathcal{X} \oplus \mathcal{Y})}(W) = \sigma_{\min}(W) = 0$. Otherwise $W_{22}$ is square and we have by the interlacing property of eigenvalues of Hermitian matrices [89] that

$$
\sigma_{\dim(\mathcal{X})}(W_{11})^2 = \lambda_{\dim(\mathcal{X})}(W_{11}^H W_{11}) \ge \lambda_{\dim(\mathcal{X}) + \dim(\mathcal{Y})}(W) \\
\ge \lambda_{\dim(\mathcal{X}) + \dim(\mathcal{Y})}(W) \left( \begin{bmatrix} W_{11}^H & \ast \\
\ast & \ast \end{bmatrix} \right).
$$
2.3 Approximation of eigenvalues and eigenvectors

\[ \lambda = \lambda_{\dim(\mathcal{X})+\dim(\mathcal{Y})}(W^H W) \]
\[ = \sigma_{\dim(\mathcal{X} \oplus \mathcal{Y})}(W)^2. \]

In the following we review results on perturbation theory of invariant subspaces. Note that any matrix \( X \in \mathbb{C}^{n \times m} \) with orthonormal columns can be completed by (a non unique) \( X_\perp \in \mathbb{C}^{n \times (n-m)} \) with orthonormal columns such that \( [X, X_\perp] \) is unitary.

**Theorem 2.11.** [107, Theorem 2.7] Let \( X \in \mathbb{C}^{n \times m}, X_\perp \in \mathbb{C}^{n \times (n-m)}, Z \in \mathbb{C}^{(n-m) \times m} \) such that \( [X, X_\perp] \) is unitary. Let \( \mathcal{X} := \text{ran}(X) \) and \( \tilde{\mathcal{X}} := \text{ran}(X + X_\perp Z) \). Then
\[ \tan \angle_{\max}(\mathcal{X}, \tilde{\mathcal{X}}) = \|Z\|_2. \]

The next theorem is a generalization of the Davis-Kahan \( \tan(\theta) \) theorem which imposes, compared to the original formulation [27, Theorem 6.3], relaxed conditions on the spectrum.

**Theorem 2.12.** [78, Theorem 2] Let \( A \in \mathbb{C}^{n \times n} \) be an Hermitian matrix and let \( X = [X_1, X_2, X_3] \) be unitary so that \( X^H A X = \text{diag}(A_{11}, A_{22}, A_{33}) \) is block diagonal, where \( X_i \in \mathbb{C}^{n_i \times n_i}, A_{ii} \in \mathbb{C}^{n_i \times n_i} \) for \( i = 1, 2, 3 \) with \( n_1 + n_2 + n_3 = n \). Let \( \tilde{X}_3 \in \mathbb{C}^{n \times n_3} \) have orthogonal columns and let \( R = A \tilde{X}_3 - \tilde{X}_3 A_3 \), where \( A_3 = \tilde{X}_3^H A \tilde{X}_3 \). Suppose that eig\((A_{11})\) lies in \( [a, b] \) and eig\((A_3)\) lies in the union of \( (-\infty, a-\delta] \) and \( [b+\delta, \infty) \). Then
\[ \tan \angle_{\max}(\tilde{X}_3, [X_2, X_3]) \leq \frac{\|R\|_2}{\delta}. \]

**Remark 2.13.** In [78, Theorem 2] \( X^H A X \) was assumed diagonal (instead of just block diagonal). Our formulation holds, because the change of bases that diagonalizes \( A_{11}, A_{22}, \) and \( A_{33} \) does not influence the subspace angles. The standard Davis-Kahan \( \tan(\theta) \) theorem is obtained for \( n_2 = 0 \).

2.3 Approximation of eigenvalues and eigenvectors

Typically, algorithms for solving high dimensional eigenproblems are iteratively in nature and consist of two steps. In the first step a search space is generated and in the second step the approximation of the eigenvectors or invariant subspaces are extracted from the subspace generated in step one. If the required accuracy of the approximation is not achieved, i.e., if the distance between the approximated and the original subspace is too large, then the approximation may be used as input for step one and both steps are carried out again. This procedure continues until a certain approximation accuracy is obtained. Consequently, those methods generate a sequence of subspaces \( \mathcal{V}_k \) that contain increasingly accurate
approximations of the desired eigenvectors or invariant subspaces. There exists a large variety of numerical methods differing in the identification of the search space, e.g., the Jacobi-Davidson method, subspace iteration and Krylov subspace methods. The latter are described in the following section. For an overview of the other methods, see, e.g., [30,38,92].

2.3.1 Krylov subspace methods

For $A \in \mathbb{C}^{n \times n}$ and $v_1 \in \mathbb{C}^n$ with $v_1 \neq 0$ a Krylov subspace method extracts the eigenvector or invariant subspace approximation from a subspace of the form

$$K_k(A, v_1) = \text{span}(v_1, Av_1, A^2v_1, \ldots, A^{k-1}v_1)$$

referred to as the $k$-th Krylov subspace associated with $A$ and $v_1$. Accordingly, we choose in the following the $k$-th Krylov subspace as search space, i.e., $V_k = K_k$.

Note that for each iteration step the dimension of the Krylov subspace increases by one until it is an invariant subspace than it stagnates.

A Krylov relation of $A$ of order $k$ is a relation of the form

$$AV_k = V_k B_k + v_{k+1} b_{k+1}^H,$$ (2.6)

where $B_k \in \mathbb{C}^{k \times k}$, $b_{k+1} \in \mathbb{C}^k$, and the columns of $[V_k, v_{k+1}] \in \mathbb{C}^{n \times k+1}$ are orthonormal. Relation (2.6) implies that $\text{ran}(V_k)$ is a Krylov subspace of $A$. But it does not imply that $\text{ran}(V_i) = \text{ran}(\begin{bmatrix} v_1, \ldots, v_i \end{bmatrix})$ is a Krylov subspace of $A$ for $i < k$. It does if $B_k$ is Hessenberg and $b_{k+1} = b_{k+1,k}e_k$.

The following theorem is used to examine convergence properties of Krylov subspace methods and it provides the basis for the convergence analysis in Section 4.1. Note that the following theorem makes use of Chebychev polynomials, which are defined by

$$\psi_k(x) = \cosh(k \cdot \arccosh(x)) \quad \text{for} \quad |x| > 1. \quad (2.7)$$

For further information and details about these polynomials, see, e.g., [30,38,92].

**Theorem 2.14.** [92, Theorem 6.3], [90] Let the eigenvalues $\lambda_i$ of the Hermitian matrix $A$ be ordered decreasingly. Then the angle between the exact eigenvector $z_i$ associated with $\lambda_i$ and the $k$-th Krylov subspace $K_k$ satisfies the inequality

$$\tan \langle z_i, K_k \rangle \leq \frac{\theta_i}{\psi_{k-i}(1 + 2\eta_i)} \tan \langle v_1, z_i \rangle, \quad (2.8)$$

where

$$\theta_1 = 1, \quad \theta_i = \prod_{j=1}^{i-1} \frac{\lambda_j - \lambda_n}{\lambda_j - \lambda_i} \quad \text{for} \quad i > 1, \quad \eta_i = \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_n}$$

and $\psi_{k-i}$ is the Chebychev polynomial of degree $k - i$. 

24
2.3 Approximation of eigenvalues and eigenvectors

If the vectors \(v_1, A v_1, A^2 v_1, \ldots, A^{k-1} v_1\) are linear independent, then they form a basis of the Krylov subspace \(K_k(A, v_1)\). However, from a numerical point of view this is in general not a good choice, because in finite precision for increasing \(k\) the basis vectors \(A^i v_1\) point more and more in the direction of the dominant eigenvector, i.e., the basis becomes ill conditioned. One possibility to circumvent this problem is to choose an orthogonal basis which leads us to Arnoldi’s method.

2.3.2 Arnoldi’s method

In [4] a procedure was introduced to reduce a dense matrix into Hessenberg form and it was suggested that the algorithm could provide approximations to some eigenvalues. Later on it was discovered that the procedure is suitable to approximate eigenvalues of large, sparse matrices.

Arnoldi’s method is an orthogonal projection onto the \(k\)-th Krylov subspace \(K_k\). More precisely, at the \(j\)-th step the method generates a new basis vector \(v_j\), such that \(v_j\) is normalized, orthogonal to \(K_i\) for all \(i < j\), and \(v_j \in K_j\). Taking into account these properties \(v_j\) has to fulfill

\[
A v_j = \sum_{i=1}^{j} h_{ij} v_i + h_{j+1,j} v_{j+1},
\]

with \(h_{ij} = v_i^H A v_j\). Thus, after \(k\) steps the Arnoldi method constructs an orthonormal basis \(V_k = [v_1, \ldots, v_k]\) and a Hessenberg matrix \(H_k \in \mathbb{C}^{k \times k}\) such that the so-called Arnoldi relation

\[
AV_k = V_k H_k + v_{k+1} h_{k+1,k} e_k^T \tag{2.9}
\]

holds. Premultiplying (2.9) by \(V_k^H\) results in

\[
V_k^H AV_k = H_k.
\]

According to [110, Theorem 2.6], this Hessenberg matrix is optimal in a sense that it minimizes the norm of \(AV_k - V_k H_k\) for all \(H_k\). If the subdiagonal element \(h_{k+1,k}\) equals zero the basis \(V_k\) spans an invariant subspace of \(A\) and if the \(k\)-th subdiagonal element of \(H_k\) is small with respect to \(\|A\|_2\) then \(V_k\) is a good approximation on an eigenspace of \(A\).

Using the classical Gram Schmidt method (CGS) stated in Algorithm 2.1, one variant for the Algorithm of Arnoldi’s method is given by Algorithm 2.2.

It is well known that the CGS method has very poor numerical properties. But a rearrangement of the CGS method, such that as soon as \(v_i\) is computed, all the remaining vectors are orthogonalized against \(v_i\), leads to a sounder computation. This results in the Modified Gram Schmidt method (MGS) stated in Algorithm 2.3.
2.3 Approximation of eigenvalues and eigenvectors

Algorithm 2.1 Classical Gram Schmidt

**Input:** \( w \in \mathbb{C}^n, V \in \mathbb{C}^{n \times k} \),  
**Output:** \( v \in \mathbb{C}^n, h \in \mathbb{C}^k, \eta \in \mathbb{R} \)

1: \( h = V^H w \)  \hspace{1cm} \text{(orthogonalization coefficients)}
2: \( \ell = w - Vh \)  \hspace{1cm} \text{(orthogonalization)}
3: \( \eta = \|\ell\|_2 \)
4: if \( \eta = 0 \) then stop end
5: \( v = \ell / \eta \)  \hspace{1cm} \text{(normalization)}

Algorithm 2.2 Arnoldi method

**Input:** \( A \in \mathbb{C}^{n \times n}, v_1 \in \mathbb{C}^n, \|v_1\|_2 = 1, m \in \mathbb{N} \),  
**Output:** \( V_{m+1} \in \mathbb{C}^{n \times m+1}, H_m \in \mathbb{C}^{m \times m}, h_{m+1,m} \)

1: \( V_1 = v_1, H_0 = [ ] \in \mathbb{C}^{0 \times 0} \)  \hspace{1cm} \text{(initialization)}
2: for \( k = 1, 2, 3, \ldots, m \) do
3: \( w_{k+1} = Av_k \)  \hspace{1cm} \text{(matrix-vector mult.)}
4: \( [v_{k+1}, h_{1:k,k}, h_{k+1,k}] = \text{CGS}(w_{k+1}, V_k) \)  \hspace{1cm} \text{(orthogonalization)}
5: \( H_k = \begin{bmatrix} H_{k-1} & h_{1:k-1,k} \\ h_{k,k-1} & h_{k,k} \end{bmatrix} \)  \hspace{1cm} \text{(update \( H_k \))}
6: \( V_{k+1} = [V_k, v_{k+1}] \)  \hspace{1cm} \text{(update \( V_k \))}
7: end for

Algorithm 2.3 Modified Gram Schmidt

**Input:** \( w \in \mathbb{C}^n, V \in \mathbb{C}^{n \times k} \),  
**Output:** \( v \in \mathbb{C}^n, h \in \mathbb{C}^k, \eta \in \mathbb{R} \)

1: \( \ell^{(0)} = w \)
2: for \( i = 1, \ldots, k \) do
3: \( h_i = v_i^H \ell^{(i-1)} \)
4: \( \ell^{(i)} = \ell^{(i-1)} - v_i h_i \)
5: end for
6: \( \eta = \|\ell^{(k)}\|_2 \)
7: if \( \eta = 0 \) then stop end
8: \( v = \ell^{(k)} / \eta \)
2.3 Approximation of eigenvalues and eigenvectors

In the Hermitian case the Arnoldi method reduces to the Lanczos method which features a short recurrence relation, see, e.g., [38, 69, 92]. Unfortunately this short recurrence may lead to extreme sensitivity to perturbations. More precisely, a rapid loss of the orthogonality of the basis and spurious eigenvalues [23] occur. To avoid these drawbacks, there is a variant that uses full recurrence even in the Hermitian case, which is used, e.g., in the ARPACK package [73]. However, for the ease of notation, whenever we use full recurrence we speak in the following of Arnoldi’s method.

The last problem that we still need to investigate is: how do we extract the approximations on the eigenvalues and eigenvectors from the generated subspace. A widely used technique for this task is introduced in the next subsection.

2.3.3 Rayleigh-Ritz method

Let $A \in \mathbb{C}^{n \times n}$, $V \subset \mathbb{C}^{n}$ be a subspace, $Y \in \mathbb{C}^{n \times k}$, and $M \in \mathbb{C}^{k \times k}$. The pair $(M, Y)$ is called a Ritz pair of $A$ with respect to $V$ if it satisfies the Galerkin condition

$$AY - YM \perp V \quad \text{and} \quad \text{ran}(Y) \subset V.$$ 

A subspace $\mathcal{Y}$ is called a Ritz space of $A$ in $V$ if for some basis $Y \in \mathbb{C}^{n \times k}$ of $\mathcal{Y}$ there is an $M \in \mathbb{C}^{\dim(\mathcal{Y}) \times \dim(\mathcal{Y})}$ such that $(M, Y)$ is a Ritz pair of $A$ with respect to $\mathcal{V}$. The eigenvalues of $M$ are well-defined and are called Ritz values of $A$ corresponding to $\mathcal{Y}$. Algorithm 2.4 describes the Rayleigh-Ritz method for obtaining an approximation on the eigenpair $(L, X)$ of $A$ with $L \in \mathbb{C}^{k \times k}$ and $X \in \mathbb{C}^{n \times k}$.

**Algorithm 2.4 Rayleigh-Ritz method**

**Input:** $A \in \mathbb{C}^{n \times n}$, $V \subset \mathbb{C}^{n}$

**Output:** $AY - YM \perp V$ and $\text{ran}(Y) \subset V$

1. Compute an orthonormal basis $V$ of the subspace $V$.
2. Compute $B = V^H A V$.
3. Compute the eigenpair $(M, Z)$ of $B$.
4. Take $(M, Y) = (M, VZ)$ as Ritz pair of $A$.

A Ritz pair is optimal in the sense, that it minimizes the norm of $AY - YM$. After the $k$-th step of Arnoldi’s method, we obtain from

$$AY - YM = AV_k Z - V_k Z M$$

$$= V_k H_k Z + h_{k+1, k} e_k^T Z V_k Z$$

$$= V_k Z M + h_{k+1, k} e_k^T Z V_k Z$$

$$= h_{k+1, k} e_k^T Z$$
2.3 Approximation of eigenvalues and eigenvectors

that \( \|A Y - Y M\|_2 = |h_{k+1,k}| \|e_k^T Z\|_2 \), thus the Ritz pair \((M,Y)\) is likely to be a good approximation of the eigenpair \((L,X)\) whenever the product of \(|h_{k+1,k}|\) and \(\|e_k^T Z\|_2\) is small.

The following theorem is a generalization of [92, Theorem 4.6] and it relates an eigenspace of \(A\) to the corresponding Ritz space.

**Theorem 2.15.** [109, Theorem 2] Let \(X\) be an eigenspace of \(A \in \mathbb{C}^{n \times n}\) and let \(V\) be a subspace in \(\mathbb{C}^n\). Let \(Y\) be a Ritz space in \(V\) and let \(Y^\perp\) be the orthogonal complement of \(Y\) in \(K\). Then

\[
\sin \angle_{\text{max}}(X,Y) \leq \sin \angle_{\text{max}}(X,V) \sqrt{1 + \frac{\|P A (I - P)\|_2^2}{\text{sep}_{A,2}(Y^\perp, X)^2}},
\]

where \(P\) is the orthogonal projector onto \(V\).

**Remark 2.16.** In the original formulation of Theorem 2.15 in [109] it is required that \(X\) and \(Y\) are of the same dimension. However, the proof given there is actually valid for \(\dim(X) \neq \dim(Y)\).

The next lemmas provide bounds on the norm of the residual of the Ritz pair, where we use the \(k\)-th Krylov subspace as search space, i.e., \(V = K_k\).

**Lemma 2.17.** Let (2.6) be a Krylov relation and let \((M, Y) = (M, V_k Z)\) be a Ritz pair of \(A\) in \(\text{ran}(V_k)\). Then for \(* \in \{2, F\}\) the \(*\)-norm of the residual is given by

\[
\|A Y - Y M\|_* = \|b_{k+1}^H Z\|_2.
\]  

**Proof.** This result is common knowledge, but rarely formulated stand-alone. We provide a short proof for completeness. We have

\[
A Y - Y M = A V_k Z - V_k Z M = (V_k B_k + v_{k+1}^H b_{k+1}) Z - V_k Z M = v_{k+1} b_{k+1}^H Z,
\]

where we used \(B_k Z = Z M\). The result follows upon taking norms. \(\square\)

**Lemma 2.18.** Let \(A, E \in \mathbb{C}^{n \times n}\). Consider the Krylov relation (2.6) of the perturbed matrix \(A + E\). Let \((\tilde{M}, \tilde{Y}) = (\tilde{M}, V_k Z)\) be a Ritz pair of \(A + E\) in \(\text{ran}(V_k)\). Then for \(* \in \{2, F\}\) the \(*\)-norm of the residual is bounded by

\[
\|A \tilde{Y} - \tilde{Y} \tilde{M}\|_* \leq \|b_{k+1}^H Z\|_2 + \|E\|_* \|Z\|_2.
\]

**Proof.** Rewriting the residual in term of \(A + E\) gives

\[
\|A \tilde{Y} - \tilde{Y} \tilde{M}\|_* = \|(A + E) \tilde{Y} - \tilde{Y} \tilde{M} + E \tilde{Y}\|_* \leq \|(A + E) \tilde{Y} - \tilde{Y} \tilde{M}\|_* + \|E \tilde{Y}\|_*.
\]
2.3 Approximation of eigenvalues and eigenvectors

The proof is completed by applying Lemma 2.17 to the first term and using
\[ \|E\tilde{Y}\|_* \leq \|E\|_*\|\tilde{Y}\|_2 \leq \|E\|_*\|Z\|_2. \]

Comparing Lemmas 2.17 and 2.18 we see that the norm of the residual in the perturbed case can be bounded by the norm of the residual in the unperturbed case \(\|b_{k+1}^H Z\|_2\) and the additional term \(\|E\|_*\|Z\|_2\) which mainly depends on the norm of the perturbation \(E\). Whenever \(b_{k+1}\) is a multiple of \(e_k\) the residual (2.10) just depends on the last row of the matrix \(Z\).
3 Inexact Arnoldi method

As explained in the introduction the identification of ground states of quantum systems involves the solution of eigenvalue problems

\[ Ax = \lambda x, \]

for \( A \in \mathbb{C}^{n \times n}, \lambda \in \mathbb{C} \) and \( x \in \mathbb{C}^n \setminus \{0\} \), that are of extremely large dimension. More precisely, a problem dimension of \( n = 2^d \) where \( d \) is the number of particles of the considered system with \( d > 50 \) is not uncommon [52, 95, 98]. Thus, the dimension of the underlying vector space grows exponentially in the number of particles. In those situations, the memory capacity of even large computing clusters is not sufficient to store a single vector in standard format. To nevertheless allow the storage of high-dimensional vectors and matrices, data sparse tensor formats, like the tensor train [51, 81, 85] or the hierarchical tensor formats [40, 46, 68] are typically used. A more detailed discussion on this storage problem will follow in Chapter 5. The use of tensor formats makes storing vectors possible but it entails the drawback that vector operations cannot be carried out exactly. Instead only approximations of the intended quantities are available.

Applied to our eigenvalue problem this means that the matrix \( A \) is given in a low rank tensor operator format and the vector \( x \) in a tensor format. Tensor operations like tensor summation or operator-tensor multiplication would result in a growth of the tensor rank and therefore an unnecessary increase in the storage requirements. To avoid this increase any operation is followed by an approximation procedure, i.e., the operation can only be evaluated approximately. In the following analysis we model the tensor operations by perturbed vector operations. For example the operator-tensor multiplication is modeled by the perturbed matrix-vector multiplication \( Av_i + f_i \), where \( f_i \) is some small unknown vector.

The aim of this chapter is to develop a method which can tackle the problem of inexact operations. To do so we introduce a modification of the well known Arnoldi Algorithm. As mentioned in Section 2.3.2, Arnoldi’s method constructs an orthonormal basis \( V_{k+1} = [v_1, \ldots, v_{k+1}] \) and a Hessenberg matrix \( H_k \in \mathbb{C}^{k \times k} \) such that the Arnoldi relation

\[ AV_k = V_k H_k + v_{k+1} h_{k+1,k} e_k^T \]

holds. Since the eigenvalues of \( H_k \) approximate some eigenvalues of \( A \) in a sense
3 Inexact Arnoldi method

that if \( H_k z = \mu_k z \), then

\[
\| A(V_k z) - \mu(V_k z) \|_2 \leq |h_{k+1,k}| e_k^T z,
\]

the Arnoldi method is used for the solution of high dimensional and sparse eigenvalue problems. In particular, the basis \( V_k \) is calculated by the classical Gram Schmidt method but it is widely known that especially in situations of inexact vector operations this method often computes a set of vectors that are far from orthogonal. Even linear independence may be lost. However, in the framework of Arnoldi’s method, the orthogonality of the computed basis \( V_k \) is essential for the accuracy of the projection onto the corresponding subspace. In other words the loss of orthogonality (or independence) can affect the reliability of the computed eigenpair, see, e.g., [14,17,38].

To address this problem we develop the compensated Gram Schmidt method (ComGS) and the compensated Gram Schmidt method with reorthogonalization (ComGSre). These methods use in the case of inexact operations a non-standard projection in comparison to classic Gram Schmidt methods to compensate the deviation from orthogonality caused by the perturbations. The compensated Gram Schmidt method preserves the orthogonality property better than the classical Gram Schmidt method. Nevertheless, exact orthogonality cannot be achieved. Therefore, a central element in the thesis is the quantification of the distance to orthogonality. Denoted by \( O\)-measure we use in the following \( \| V^H V - I \|_2 \), i.e., the distance of the cross-product matrix from the identity as a measure of orthogonality. This measure was also used for example in [13,36,49,104]. Additionally, we determine in Section 3.1.2 the distance of a vector to a set of basis vectors, denoted as \( vO\)-measure.

An advantage of ComGS and ComGSre is that both methods are parallelizable and therefore are favored here over other orthogonalization schemes, e.g., modified Gram Schmidt method or Householder transformation.

Embedding ComGS or ComGSre into the Arnoldi method still results in not exactly orthogonal basis vectors. In other words, the Arnoldi relation (2.9) is not valid, but importantly we have improved the distance to orthogonality. We will prove that the residual of the Arnoldi relation \( A\tilde{V}_k - \tilde{V}_k \tilde{H}_k = \tilde{v}_{k+1} h_{k+1,k} e_k^T \) is small in norm. Further, using backward error analysis we show that also the computed matrices \( \tilde{V}_{k+1} \) and \( \tilde{H}_k \) satisfy an exact Arnoldi relation, however not for the matrix \( A \) but for a nearby matrix \( A + E_k \), where \( E_k \) is small in norm.

Since the matrix \( A \) is Hermitian it would be desirable that the backward error \( E_k \) possesses the same property so that the sum \( A + E_k \) is Hermitian as well. However, it turns out that an Hermitian \( E_k \) that satisfies the Arnoldi relation

\[
(A + E_k)\tilde{V}_k = \tilde{V}_k \tilde{H}_k + \tilde{v}_{k+1} h_{k+1,k} e_k^T
\]

does not exist. To ensure the above relation for an Hermitian \( E_k \) we relax the condition on \( H_k \) and replace the Hessenberg matrix \( \tilde{H}_k \) by a Hermitian matrix.
Inexact Arnoldi method

$B_k$. Further, we provide bounds on the norm of $E_k$ for our suitable choices of $B_k$.

The problem of inexact vector operations is contained in a wide range of applications. The analysis shown in this section is independent of the actual source of perturbation and thus applies to a broader context. Examples of the other applications include the following problems.

- **Rounding:** Rounding (roundoff) error is caused by the finite precision nature of computer arithmetic. More precisely, computer arithmetic can only represent a finite subset of the natural and real numbers. Thereby, the computer representation of the natural numbers $\mathbb{N}$ introduces only a wrap-around error and preserves almost all algebraic properties. A standard representation of the real numbers $\mathbb{R}$ was established in 1985 by the ANSI/IEEE standards 754 and 854, where the set $\mathbb{F}$, so called floating-point numbers, is used to represent $\mathbb{R}$. A number in $\mathbb{F}$ is stored in a fixed bit-length array which includes the information of the sign, the mantissa and the exponent. This array requires 32 bits in single and 64 bits in double precision arithmetic.

Unfortunately, it is not always the case that the result of an mathematical operation is part of $\mathbb{F}$. This introduces an additional error type as the result has to be rounded to a floating point number. Further, if the result is too small or too large to be representable in $\mathbb{F}$ we are speaking about an underflow and overflow, respectively. More detailed information can be found in, e.g., [30,38,48,123].

- **Mixed precision arithmetic:** Considering an Arnoldi method where the basis vectors of the Krylov space are stored in single precision whereas all remaining quantities and computations are handled in double precision. This approach effectively halves the memory requirements of Arnoldi’s method which are dominated by the need to store the basis vectors. However every computed basis vector has to be rounded to single precision, which introduces a perturbation to the operations.

- **Sparse eigenvectors:** If the eigenvectors of interest are known to be well approximable by sparse vectors, i.e., vectors with only a few non-zeros, then it is natural to ask for sparse Arnoldi basis vectors. This can be achieved by neglecting elements smaller than a certain threshold in the computed basis vectors. This thresholding introduces a perturbation. To be more precise, matrix-vector-multiplication and vector addition (including sparsification) are inexact whereas vector scaling and scalar products are exact, for a more detailed discussion, see, e.g., [19,22,26].

The results shown in this chapter complement earlier work in this field. A background on Arnoldi’s method without perturbations was stated in, e.g., [74, 89,91], whereas the method with perturbations was considered in [100–102,117]. However, in these works only the matrix-vector product where assumed to be
inexact. Effects of inaccuracies on the result of the Gram Schmidt orthogonalization process was analyzed in [13,37,49,89].

This chapter is structured as follows: In Section 3.1 we introduce a Gram Schmidt method that uses in the case of inexact operations a non-standard projection to alleviate the deviation from orthogonality of the computed vector caused by the perturbations and to provide a measure of the orthogonality of the computed vector. In Section 3.2 we investigate how the distance to orthogonality behaves if we add a reorthogonalization step to the ComGS method. In Section 3.3 we present the inexact Arnoldi method and investigate the orthogonality of the Arnoldi basis. More precisely we bound the norm of the distance of the whole cross product matrix $\tilde{V}^H \tilde{V}$ from the identity. In addition, we compare the new variant with the well known classical and modified Gram Schmidt methods. Then in Section 3.4 we show that the obtained subspace can be interpreted as an exact subspace of a matrix close to $A$ and establish bounds for the backward error. Numerical experiments confirm the applicability of the method and illustrate the theoretical results in Section 3.5. Finally, we offer some concluding remarks in Section 3.6.

3.1 Compensated Gram Schmidt method (ComGS)

The orthogonality of the classical Gram Schmidt method can be improved by computing $\tilde{V} = \tilde{V} G^{-1}$, where $G$ is the Cholesky factor of the cross product matrix $D := \tilde{V}^H \tilde{V}$. Embedded in the algorithm, this idea leads to the new Gram Schmidt variant described in this section. It uses a different projector to alleviate the deviation from orthogonality, inflicted by the perturbations. This projector compensates the perturbation caused by the inexact operations and therefore the method is called compensated Gram Schmidt method (ComGS).

Note that we add tildes “∼” to the variables that also appear in the classic Gram Schmidt method in order to emphasize that their computations entail perturbations. Variables that do not appear in the classic algorithm, i.e., $D$ and its Cholesky factor $G$ are named without tildes. In the following we describe the compensated Gram Schmidt method in more detail.

3.1.1 Algorithm

The compensated Gram Schmidt method is initialized with a vector $\tilde{w} \in \mathbb{C}^n$, a matrix $\tilde{V} \in \mathbb{C}^{n \times k}$ that does not have to be orthonormal, and the nonsingular cross product matrix $D = \tilde{V}^H \tilde{V} \in \mathbb{C}^{k \times k}$. Due to the fact that neither of the vector operations addition, scaling and inner products is evaluated exactly, the method computes a vector $\tilde{v} \in \mathbb{C}^n$ that is only nearly orthonormal with respect to the columns of the matrix $\tilde{V}$. Moreover, ComGS determines the orthogonalization coefficients $\tilde{\mu} \in \mathbb{C}^k$, $\tilde{\eta} \in \mathbb{R}$ and allows the computation of the updated cross
product matrix $\tilde{D}_+$. We indicate by “≈” that we consider inexact operations.

Algorithm 3.1 Compensated Gram Schmidt

**Input:** $\tilde{w} \in \mathbb{C}^n, \tilde{V} \in \mathbb{C}^{n \times k}, D \in \mathbb{C}^{k \times k}$, $D$ nonsingular

**Output:** $\tilde{v} \in \mathbb{C}^n, \tilde{h} \in \mathbb{C}^k, \tilde{\eta} \in \mathbb{R}, D_+ \in \mathbb{C}^{k+1 \times k+1}$

1: $\tilde{h} \approx D^{-1}\tilde{V}^H\tilde{w}$
2: $\tilde{\ell} \approx \tilde{w} - \tilde{V}\tilde{h}$
3: $\tilde{\eta} \approx ||\tilde{\ell}||_2$
4: if $\tilde{\eta} = 0$ then stop end if
5: $\tilde{v} \approx \tilde{\ell}/\tilde{\eta}$
6: $D_+ \approx \begin{bmatrix} D & \tilde{V}^H\tilde{v} \\ \tilde{v}^H\tilde{v} & \tilde{v}^H\tilde{\eta} \end{bmatrix}$

The Algorithm 3.1 has the following basic properties.

**Remark 3.1.** The classical Gram Schmidt method is obtained upon omitting the perturbations, i.e., replacing “≈” by “=”, and the computation of the matrix $D$.

**Remark 3.2.** For $k = 0$ the matrices $\tilde{V}$ and $D$ have zero columns and steps 1 and 2 reduce to $\tilde{h} = [\ ]$ and $\tilde{\ell} = \tilde{w}$, respectively.

**Remark 3.3.** The algorithm will not break down, i.e., $\tilde{\eta} > 0$, if and only if $\tilde{w}$ is linearly independent of the columns of $\tilde{V}$, also under small perturbations.

**Remark 3.4.** To analyze the efficiency we use in the following the number of matrix-vector operations per iteration. In this thesis an operation is called matrix-vector operation (MVecOP) if the matrix $\tilde{V}$ is involved, e.g., for the ComGS method a vector operation occurs in steps 1, 2, and 6.

In Algorithm 3.1 we use in the case of inexact operations a non-standard projection to orthogonalize $\tilde{w}$ against $\tilde{V}$. Since we do not assume that the basis spanned by the columns of $\tilde{V}$ is orthonormal, we use the projector $I - \tilde{V}D^{-1}\tilde{V}^H$ with $D \approx \tilde{V}^H\tilde{V}$ (instead of $I - \tilde{V}\tilde{V}^H$ for an orthonormal $\tilde{V}$). To do this we need to construct the cross product matrix $D = \tilde{V}^H\tilde{V}$. But, as we consider the case where inner products cannot be evaluated exactly, we only have the approximated matrix $D \approx \tilde{V}^H\tilde{V}$. The cross product matrix $D$ can be updated during the iteration (Step 6 of Algorithm 3.1). The non-standard projector is chosen in order to reduce the impact of the perturbations caused by the inexact vector operations on the orthogonality of $\tilde{V}$.

ComGS is a modification of classical Gram Schmidt (CGS) which does not require orthonormal bases. It inherits the property of CGS that all scalar products $\tilde{v}_1^H\tilde{w}, \ldots, \tilde{v}_k^H\tilde{w}$ may be computed in parallel, whereas in modified Gram Schmidt
(MGS) they have to be evaluated sequentially. We note that also MGS can deal to some extent with non-orthogonal bases, as it implicitly solves a linear system with the matrix $D$ (as ComGS does in Step 1), see [13, page 308 f.].

Additionally ComGS determines the Cholesky factor of $D = G^T G$. The factor $G$ is beneficial for solving the linear systems with $D$ that occur in Step 1 of Algorithm 3.1. The matrix $G$ can be updated along with $D$, via [48, Proof of Theorem 10.1],

$$G = \begin{bmatrix} G & g \\ 0 & \gamma \end{bmatrix} \quad \text{with} \quad g = G^{-H}(\tilde{V}^H \tilde{v}) \quad \text{and} \quad \gamma = \sqrt{\|\tilde{v}\|^2_2 - \|g\|^2_2}.$$ 

Further, using the Cholesky factor $G$ the basis $\tilde{V} = \tilde{V}G^{-1}$ is implicitly available and has orthonormal columns.

**Remark 3.5.** The basis $\tilde{V} = \tilde{V}G^{-1}$ should never be explicitly formed, since that would cause further perturbations. Instead only the factorized form should be used.

The inexactness in all operations is modeled by adding a perturbation vector $f^{(*)}$ of small norm each time an operation is performed. The perturbations occurring in Algorithm 3.1 are defined by

1: $f^{(P)} := D^{-1}\tilde{V}^H \tilde{w} - \tilde{h}, \quad (3.1)$  
2: $f^{(O)} := \tilde{w} - \tilde{V}^H h - \tilde{\ell}, \quad (3.2)$  
3: $f^{(N)} := \|\tilde{\ell}\|_2 - \tilde{\eta}, \quad (3.3)$  
5: $f^{(S)} := \tilde{\ell} - 2 \tilde{\eta}, \quad (3.4)$

respectively. The superscripts $(P)$, $(O)$, $(N)$, and $(S)$ indicate the source of the perturbation, i.e., $f^{(P)}$ arises from the inexactness of the inner product, $f^{(O)}$ from the inexactness of the orthogonalization, $f^{(N)}$ from the inexactness of the normalization, and $f^{(S)}$ from the inexactness of the scaling. Further, the number before each perturbation (3.1)-(3.4) refers to the step of Algorithm 3.1 where the perturbation occurs.

In some cases inner products can be evaluated more precisely than additions and scalings. Therefore we distinguish between perturbations caused by vector operations, such as addition and scaling, and perturbations caused by scalar operations, such as inner products and norm computations. The magnitude of the perturbation is quantified by the parameters $\varepsilon_v$ and $\varepsilon_s$, where $\varepsilon_v$ models the error caused by vector operations and $\varepsilon_s$ models perturbations induced by scalar operations.

In more detail, in the case of tensor approximation it is even assumed that the scalar products can be computed to the order of machine precision, see [80]. In contrast, the addition of two tensors increases the tensor ranks and consequently increases the storage requirements. Thus, an approximation procedure should
3.1 Compensated Gram Schmidt method (ComGS)

be applied that truncates the tensor back to lower tensor ranks within a specified tolerance. This truncation threshold is denoted by $\varepsilon_v$. A detailed consideration of basic operations in a tensor format is provided in Chapter 5. In the case of mixed precision arithmetic $\varepsilon_v$ can be interpreted as single machine precision, i.e., $\varepsilon_v \approx 6 \cdot 10^{-8}$ and $\varepsilon_s$ as double machine precision, i.e., $\varepsilon_s \approx 10^{-16}$.

In order to evaluate the effects of the perturbations caused by inexact operations we need to determine error bounds for $f^{(\text{in})}$. Conceptually, these bounds are derived from the following error model. We assume that for the inexact inner product we have

$$ f^{(\text{in})} := \tilde{V}^H \tilde{w} - \text{rd}(\tilde{V}^H \tilde{w}), $$

i.e., with (3.5) its norm is bounded by

$$ \| f^{(\text{in})} \|_2 = \| \tilde{V}^H \tilde{w} - \text{rd}(\tilde{V}^H \tilde{w}) \|_2 \leq \varepsilon_s \sqrt{k} \| \tilde{V}^H \tilde{w} \|_2, $$

(3.7)

where we have used that $\| \tilde{V} \|_2 \leq \| \tilde{V} \|_2$. Furthermore, the residual of the linear system that we have to solve in Step 1 of Algorithm 3.1 is defined by

$$ r_{\text{LES}} := \tilde{D} \tilde{h} - \text{rd}(\tilde{V}^H \tilde{w}). $$

Thus, it follows from the definition of $f^{(\text{in})}$ that

$$ \tilde{h} = D^{-1} \left( \text{rd}(\tilde{V}^H \tilde{w}) + r_{\text{LES}} \right) = D^{-1} \left( \tilde{V}^H \tilde{w} - f^{(\text{in})} + r_{\text{LES}} \right). $$

(3.8)

Moreover, [48, Theorem 7.1] states that the norm of this residual is bounded by

$$ \| r_{\text{LES}} \|_2 \leq \varepsilon_s \left( \| D \|_2 \| \tilde{h} \|_2 + \| \text{rd}(\tilde{V}^H \tilde{w}) \|_2 \right). $$

(3.9)

Note that in the original formulation of [48, Theorem 7.1] $\varepsilon_s$ was chosen to be the machine precision $\varepsilon_M$. However, because of $\varepsilon_M \leq \varepsilon_s$ we use $\varepsilon_s$ to simplify the
3.1 Compensated Gram Schmidt method (ComGS)

derivation. Taking norms of equation (3.1) and inserting (3.8) for \( \tilde{h} \) results in

\[
\|f^{(P)}\|_2 = \|D^{-1} \tilde{V}^H \tilde{w} - \tilde{h}\|_2 = \|D^{-1}(f^{(In1)} - r_{LES})\|_2 \\
\leq \|D^{-1}\|_2 \left( \|f^{(In1)}\|_2 + \|r_{LES}\|_2 \right)
\]

and together with the definition of \( f^{(In1)} \), (3.7), and (3.9) we obtain

\[
\|f^{(P)}\|_2 \leq \varepsilon_s \|D^{-1}\|_2 \left( \sqrt{k} \|\tilde{V}\|_2 \|\tilde{w}\|_2 + \|D\|_2 \|\tilde{h}\|_2 + \|r_d(\tilde{V}^H \tilde{w})\|_2 \right) \\
\leq \varepsilon_s \|D^{-1}\|_2 \left( (1 + \varepsilon_s \sqrt{k}) \|\tilde{V}\|_2 \|\tilde{w}\|_2 + \|D\|_2 \|\tilde{h}\|_2 \right).
\]

Finally, it follows from (3.1) that \( \tilde{h} = D^{-1} \tilde{V}^H \tilde{w} - f^{(P)} \), i.e., inserting in the inequality above and solving for \( \|f^{(P)}\|_2 \) results in

\[
\|f^{(P)}\|_2 \leq \frac{\varepsilon_s \|D^{-1}\|_2}{1 - \varepsilon_s \kappa(D)} \left( 1 + 2 \sqrt{k} + \kappa(D) \right) \|\tilde{V}\|_2 \|\tilde{w}\|_2,
\]

(3.10)

where we used that \((1 + \varepsilon_s) < 2\) and \(1 - \varepsilon_s \kappa(D) > 0\) with \(\kappa(D) = \|D^{-1}\|_2 \|D\|_2\).

The perturbation (3.2) is caused by \( k \) additions, i.e., it depends on

\[
\|\tilde{w}\|_2 + \|\tilde{v}_1 h_1\|_2 + \|\tilde{v}_2 h_2\|_2 + \ldots + \|\tilde{v}_k h_k\|_2 + \|\tilde{\ell}\|_2.
\]

Since, \( \|\tilde{\ell}\|_2 \) is tiny compared to the other summands and \( \|\tilde{h}\|_1 \leq \sqrt{k} \|\tilde{h}\|_2 \) it follows from (3.6) that

\[
\|f^{(0)}\|_2 \leq \varepsilon_v \left( \|\tilde{w}\|_2 + \sqrt{k} \|\tilde{V}\|_2 \|\tilde{h}\|_2 \right).
\]

(3.11)

The perturbation (3.3), occurring in the normalization Step 5 of Algorithm 3.1, is determined by a perturbation in the computation of the inner product \( \tilde{\ell}^H \tilde{\ell} \) and the square root of \( r_d(\tilde{\ell}^H \tilde{\ell}) \). According to (3.5) the norm of \( f^{(In2)} := \tilde{\ell}^H \tilde{\ell} - r_d(\tilde{\ell}^H \tilde{\ell}) \) is bounded by

\[
\|f^{(In2)}\|_2 \leq \varepsilon_s \|\tilde{\ell}\|_2^2
\]

and the norm of the error, occurring in the computation of the square root, can be approximated by

\[
\|f^{(Sqrt)}\|_2 \leq \varepsilon_s \sqrt{r_d(\|\tilde{\ell}\|_2^2)} = \varepsilon_s \sqrt{\|\tilde{\ell}\|_2^2 - f^{(In2)}} \leq \varepsilon_s \|\tilde{\ell}\|_2 \sqrt{(1 - f^{(In2)})},
\]

with \( f^{(In2)} := f^{(In2)} / \|\tilde{\ell}\|_2^2 \) and \( |f^{(In2)}| \leq \varepsilon_s \). Because of

\[
|1 - \sqrt{(1 + f)}| \leq \frac{|f|}{2} + O(f^2)
\]

(3.12)
3.1 Compensated Gram Schmidt method (ComGS)

for all \( f \in (-1, 1) \) we get

\[
\| f^{(\text{sqrt})} \|_2 \leq \varepsilon_s \| \tilde{\ell} \|_2 \left( 1 + \frac{\varepsilon_s}{2} + O(\varepsilon_s^2) \right).
\]

From (3.3) it follows that

\[
f^{(N)} = \| \tilde{\ell} \|_2 - \bar{\eta} = \| \tilde{\ell} \|_2 - \left( \sqrt{\| \tilde{\ell} \|_2^2 - f^{(\text{ln2})}} - f^{(\text{sqrt})} \right).
\]

Using again the normalized perturbation \( \tilde{f}^{(\text{ln2})} \), we get

\[
f^{(N)} = \| \tilde{\ell} \|_2 \left( 1 - \sqrt{1 - \tilde{f}^{(\text{ln2})}} \right) + f^{(\text{sqrt})}.
\]

Thus, taking norms, using (3.12) and inserting the bound on \( \| f^{(\text{sqrt})} \|_2 \) for \( \varepsilon_s < 1/2 \) we obtain that

\[
\| f^{(N)} \|_2 \leq \left( \frac{3}{2} \varepsilon_s + O(\varepsilon_s^2) \right) \| \tilde{\ell} \|_2 \leq \varepsilon_s 2 \| \tilde{\ell} \|_2.
\]  

Finally, \( f^{(S)} \), defined in (3.4), models the error that arises in the scaling of \( \tilde{\ell} \), i.e., it follows from (3.6) that

\[
\| f^{(S)} \|_2 \leq \varepsilon_v \| \tilde{\ell} \|_2.
\]  

In order to evaluate the orthogonality properties of ComGS in the next section we derive the \( vO \)-measure, i.e., a measure of the orthogonality of the computed vector \( \tilde{v} \) to the subspace spanned by \( \tilde{V} \).

3.1.2 Distance to orthogonality of \( \tilde{v} \)

We approximate the distance to orthogonality of the vector \( \tilde{v} \), computed by Algorithm 3.1, to the matrix \( \tilde{V} \) by calculating the bound of the norm of the normalized inner product of \( \tilde{V} \) and \( \tilde{v} \), denoted as \( vO \)-measure. In the presence of perturbations, we should at best expect that the \( vO \)-measure is of the order of \( (\varepsilon_v + \varepsilon_s) \). The accuracy of ComGS and ComGSre can never exceed the accuracy of the input data. The following theorem verifies this expectation.

Assumption 3.6. The perturbations (3.1)-(3.4) in Algorithm 3.1 are bounded by (3.10)-(3.14) for some \( \varepsilon_s \), \( \varepsilon_v \in \mathbb{R} \) such that \( 0 \leq \varepsilon_v < 1 \), \( 0 \leq \varepsilon_s < 1 \), \( 4(\varepsilon_v + \varepsilon_s) + \varepsilon_s \| D^{-1} \|_2 (k + \| D \|_2) < 1 \), \( 2(\varepsilon_s + \varepsilon_v \sqrt{k\theta}) < 1 \) with \( \theta := \| D^{-1} \|_2 \| \tilde{V} \|_2^2 \) and \( \| \tilde{V}^H \tilde{V} - D \|_2 < 1/\| D^{-1} \|_2 \).

Theorem 3.7. Let \( \tilde{w} \), \( \tilde{V} \), \( D \), \( \tilde{v} \), \( \tilde{h} \), and \( \bar{\eta} \) be as in Algorithm 3.1 and suppose that no breakdown occurs, i.e., \( \bar{\eta} \neq 0 \) in Step 4. Let \( \rho := \| \tilde{h} \|_2 / \bar{\eta} \). If Assumption 3.6
3.1 Compensated Gram Schmidt method (ComGS)

holds, then there exists an \( \varepsilon_{\text{sum}} < 1 \) with

\[
\varepsilon_{\text{sum}} \leq \varepsilon_v (1 + \sqrt{k}\theta) + \varepsilon_s \theta \frac{k \|D^{-1}\|_2 + 2 \left( 1 + 2\sqrt{k} + \|D\|_2 \|D^{-1}\|_2 \right)}{1 - 4(\varepsilon_v + \varepsilon_s) - \varepsilon_s \|D^{-1}\|_2 (k + \|D\|_2)} \tag{3.15}
\]

such that

\[
\left\| \bar{w} - [\bar{V}, \bar{v}] \begin{bmatrix} \bar{h} \\ \bar{\eta} \end{bmatrix} \right\|_2 \leq \varepsilon_v \left( 2(1 + \theta \sqrt{k}) + \varepsilon_{\text{sum}} \right) \|\bar{w}\|_2, \tag{3.16}
\]

\[
\frac{1}{\|\bar{V}\|} \leq \varepsilon_{\text{sum}} \left( 1 + \|\bar{V}\|_2 \rho \right) \frac{\varepsilon_v}{1 - 2(\varepsilon_s + \varepsilon_v \sqrt{k}\theta)} + \frac{\varepsilon_v}{1 - 2\varepsilon_s}. \tag{3.18}
\]

**Proof.** We start by deriving bound (3.17). From the perturbations (3.3) and (3.4) and its bounds (3.13) and (3.14) it follows that

\[
\frac{1}{\|\bar{V}\|} \leq \frac{1}{1 - 2(\varepsilon_s + \varepsilon_v \sqrt{k}\theta)} = 1 + \frac{\varepsilon_v + 2\varepsilon_s}{1 - 2\varepsilon_s}, \tag{3.19}
\]

and analogously it holds that

\[
\frac{1}{\|\bar{V}\|} \geq 1 - \frac{(\varepsilon_v + 2\varepsilon_s)}{(1 - 2\varepsilon_s)}. \tag{3.20}
\]

Combining this two inequalities proves bound (3.17).

In the second step we prove inequality (3.16). From the perturbations (3.2) and (3.4) we get that

\[
\frac{1}{\|\bar{V}\|} = \frac{1}{\|\bar{V}\|} \leq \frac{\|\bar{V}\|_2 + \|f(S)\|_2}{\|\bar{V}\|_2 - |f(N)|} \leq \frac{1 + \varepsilon_v}{1 - 2\varepsilon_s} = 1 + \frac{\varepsilon_v + 2\varepsilon_s}{1 - 2\varepsilon_s}, \tag{3.19}
\]

Inserting the corresponding bounds (3.11) and (3.14) results in

\[
\left\| \bar{w} - [\bar{V}, \bar{v}] \begin{bmatrix} \bar{h} \\ \bar{\eta} \end{bmatrix} \right\|_2 \leq \varepsilon_v \left( \|\bar{w}\|_2 + \sqrt{k} \|\bar{V}\|_2 \|\bar{h}\|_2 + \|\bar{\eta}\|_2 \right). \tag{3.20}
\]

We define

\[
K := \bar{V}^H \bar{V} - D \quad \text{and} \quad J := D^{-1} - (\bar{V}^H \bar{V})^{-1}. \tag{3.21}
\]
With (3.1), (3.2) and the above definition of $J$ it follows that
\[
\begin{align*}
\tilde{\ell} &= \tilde{w} - \tilde{V} D^{-1} \tilde{V}^H \tilde{w} + \tilde{V} f^{(P)} - f^{(0)} \\
&= \tilde{w} - \tilde{V} \left( (\tilde{V}^H \tilde{V})^{-1} + \ell \right) \tilde{V}^H \tilde{w} + \tilde{V} f^{(P)} - f^{(0)} \\
&= \ell - \tilde{V} J \tilde{V}^H \tilde{w} + \tilde{V} f^{(P)} - f^{(0)},
\end{align*}
\]  
(3.22)
where $\tilde{\ell} := \tilde{w} - \tilde{V} (\tilde{V}^H \tilde{V})^{-1} \tilde{V}^H \tilde{w}$ is the vector exactly orthogonalized to $\tilde{V}$. Note that, since the inner products cannot be evaluated exactly, this vector is not available in practice, but it satisfies $\|\tilde{\ell}\|_2 \leq \|\tilde{w}\|_2$. Taking norms in (3.22) results in
\[
\|\tilde{\ell}\|_2 \leq (1 + \|\tilde{V}\|_2^2 \|J\|_2) \|\tilde{w}\|_2 + \|\tilde{V}\|_2 \|f^{(P)}\|_2 + \|f^{(0)}\|_2. \tag{3.23}
\]
Moreover, from (3.1) and (3.10) we obtain that
\[
\|\tilde{h}\|_2 \leq \left( 1 + \frac{\varepsilon_s}{1 - \varepsilon_s \kappa(D)} \left( 1 + 2\sqrt{k} + \kappa(D) \right) \right) \|D^{-1}\|_2 \|\tilde{V}\|_2 \|\tilde{w}\|_2. \tag{3.24}
\]
Thus, inserting (3.10) and (3.11) into (3.23) and using (3.24) leads to
\[
\|\tilde{\ell}\|_2 \leq (1 + \varepsilon_{\text{sum}}) \|\tilde{w}\|_2, \tag{3.25}
\]
where
\[
\varepsilon_{\text{sum}} := \|J\|_2 \|\tilde{V}\|_2^2 + \frac{\varepsilon_s (1 + 2\sqrt{k} + \kappa(D))}{1 - \varepsilon_s \kappa(D)} (1 + \varepsilon_v \sqrt{k}) \theta + \varepsilon_v (1 + \sqrt{k} \theta). \tag{3.26}
\]
The introduced $\varepsilon_{\text{sum}}$ can be interpreted as the "effective" $\varepsilon$ of the orthogonalization step, i.e., the orthogonalization step can lengthen $\tilde{w}$ by the factor $(1 + \varepsilon_{\text{sum}})$. Definition (3.26) shows that $\varepsilon_{\text{sum}}$ depends on $\varepsilon_w$, $\varepsilon_v$, and the norm of $D$. By Assumption 3.6 it holds that $\varepsilon_{\text{sum}}$ is small.

In the next step of the proof we determine the bound of $\varepsilon_{\text{sum}}$. Using that
\[
|K_{ij}| = |d_{ij} - \bar{v}_i^H \bar{v}_j| \leq \varepsilon_s \|\bar{v}_i\|_2 \|\bar{v}_j\|_2
\]
and (3.19) the norm of $K$ is bounded by
\[
\|K\|_2 \leq \|K\|_F = \|D - \tilde{V}^H \tilde{V}\|_F \leq \varepsilon_s k \left( \frac{1 + \varepsilon_v}{1 - 2\varepsilon_s} \right)^2 \leq \frac{\varepsilon_s k}{1 - 4(\varepsilon_v + \varepsilon_s)}. \tag{3.27}
\]
With the assumption $\|\tilde{V}^H \tilde{V} - D\|_2 < 1/\|D^{-1}\|_2$, Lemma 2.5 is applicable and the norm of $J$ can be bounded by
\[
\|J\|_2 \leq \frac{\|K\|_2 \|D^{-1}\|_2^2}{1 - \|K\|_2 \|D^{-1}\|_2}. \tag{3.28}
\]

40
3.1 Compensated Gram Schmidt method (ComGS)

Inserting (3.27) into (3.28) leads to

$$\|f\|_2 \leq \frac{\varepsilon_s k \|D^{-1/2}\|_2^2}{1 - 4(\varepsilon_v + \varepsilon_s) - \varepsilon_s k \|D^{-1/2}\|_2}. \tag{3.29}$$

Thus, together with (3.26) and using the fact that $1 + \varepsilon_v \sqrt{k} < 2$ (see Assumption 3.6), $\varepsilon_{\text{sum}}$ can be bounded by (3.15). Finally, inserting (3.24) and (3.25) into (3.20) and using that by Assumption 3.6 we have

$$1 + \frac{\varepsilon_s}{1 - \varepsilon_s \kappa(D)} \left(1 + 2\sqrt{k} + \kappa(D)\right) < 2, \tag{3.30}$$

proves inequality (3.16).

To derive the remaining bound (3.18), we pre-multiply equation (3.4) from the left with $e^V_H$ and take norms. For $\tilde{\eta} > 0$ this implies that

$$\|\tilde{V}^H \tilde{f}\|_2 \leq \left(\|\tilde{V}^H \tilde{f}\|_2 + \|\tilde{V}\|_2 \|f(S)\|_2\right) / \tilde{\eta}. \tag{3.31}$$

For $\|\tilde{V}^H \tilde{f}\|_2$ it follows from equation (3.22) and the orthogonality of $\tilde{e}$ to $\tilde{V}$ that

$$\|\tilde{V}^H \tilde{f}\|_2 = \|\tilde{V}^H \left(\tilde{V} J \tilde{V}^H \tilde{w} + \tilde{V} f(p) - f(0)\right)\|_2$$

and with (3.10), (3.11), and (3.24) we get

$$\|\tilde{V}^H \tilde{f}\|_2 \leq \varepsilon_{\text{sum}} \|\tilde{V}\|_2 \|\tilde{w}\|_2. \tag{3.32}$$

Furthermore, taking norms of (3.3) gives

$$\|\tilde{f}\|_2 \leq \tilde{\eta} + \|f(N)\|_2.$$  

Thus, inserting (3.13) and solving for $\|\tilde{f}\|_2$, where we assume that $1 - 2\varepsilon_s > 0$, results in

$$\|\tilde{f}\|_2 \leq \frac{\tilde{\eta}}{1 - 2\varepsilon_s}. \tag{3.33}$$

Inserting (3.14) and (3.32) into (3.31) and using (3.33) yields

$$\frac{\|\tilde{V}^H \tilde{f}\|_2}{\|\tilde{V}\|_2} \leq \frac{\varepsilon_{\text{sum}} \|\tilde{w}\|_2}{\tilde{\eta}} + \frac{\varepsilon_v}{1 - 2\varepsilon_s}. \tag{3.34}$$

To bound $\|\tilde{w}\|_2$, we rearrange equation (3.2) and take norms, thus

$$\|\tilde{w}\|_2 \leq \|\tilde{V} h\|_2 + \|\tilde{f}\|_2 + \|f(0)\|_2.$$
3.2 ComGS with reorthogonalization (ComGSre)

Inserting (3.11), (3.24), (3.30), and (3.33) with \( \rho = \| \tilde{h} \|_2 / \tilde{\eta} \) we get that

\[
\| \tilde{w} \|_2 \leq \tilde{\eta} \| \tilde{V} \|_2 \rho + \frac{\tilde{\eta}}{1 - 2\epsilon_s} + \epsilon_v 2\sqrt{k \theta} \| \tilde{w} \|_2.
\]

Thus, solving for \( \| \tilde{w} \|_2 \) results in

\[
\| \tilde{w} \|_2 \leq \frac{(1 + \| \tilde{V} \|_2 \rho) \tilde{\eta}}{1 - 2 (\epsilon_s - \epsilon_v \sqrt{k \theta})}.
\]

Finally, (3.18) follows by inserting the inequality above into (3.34).

The bound (3.18) of Theorem 3.7 describes the degree of orthogonality of \( \tilde{v} \) to the columns of the matrix \( \tilde{V} \). We have shown that the vO-measure is small, i.e., on the order of \( \epsilon_{\text{sum}} \), if the factor \( (1 + \| \tilde{V} \|_2 \rho) \) is not too large. In this sense \( (1 + \| \tilde{V} \|_2 \rho) \) can be interpreted as a condition number, with \( (1 + \| \tilde{V} \|_2 \rho) \in [1, \infty) \). The factor \( \rho \) becomes large only if \( \tilde{\eta} \) is tiny compared to the elements in \( \tilde{h} \). In this case the bound (3.18) becomes large and we loose orthogonality of \( \tilde{v} \) to \( \tilde{V} \). Note that in this case only \( \tilde{v} \) is far from being orthogonal to \( \tilde{V} \). More precisely, the bound (3.18) is not recursive, i.e., even if \( \tilde{V} \) is highly non-orthogonal the bound for the next \( \tilde{v} \) may be small (provided that the corresponding \( \rho \) is moderate). This behavior of the ComGS method provides a significant improvement compared to the CGS method when applied to inexact problems. Indeed, CGS does not recover once orthogonality is lost, but on the contrary only loses orthogonality even faster, see, e.g., [36].

**Remark 3.8.** In order to formulate Theorem 3.7 as general as possible we do not assume that \( \tilde{V} \) forms orthonormal basis. However, if \( \tilde{V} \) then is non-orthonormal the bounds of Theorem 3.7 becomes less meaningful.

**Remark 3.9.** If the norm of \( D \) is close to the norm of \( \tilde{V}^H \tilde{V} \), then \( \theta = \| D^{-1} \|_2 \| V \|_2^2 \) can be interpreted as \( \kappa(\tilde{V})^2 \), the square of the condition number of \( \tilde{V} \), where \( \kappa(\tilde{V}) = \| \tilde{V} \|_2 \| \tilde{V}^+ \|_2 \).

The bound 3.18 on the vO-measure indicates that if \( \rho \) is of the order \( \epsilon_{\text{sum}}^{-1} \), then the bound (3.18) is of order one, i.e., orthogonality is lost. In order to address this problem, we develop in the following an improved orthogonalization scheme.

### 3.2 ComGS with reorthogonalization (ComGSre)

While performing the ComGS method in the presence of rounding errors it is unlikely that \( \tilde{\eta} \) is zero, but it can be very small. The ComGS method is designed such that \( \tilde{V}^H \tilde{\ell} \) is small relative to \( \| \tilde{w} \|_2 \), which we confirmed in the analysis.
3.2 ComGS with reorthogonalization (ComGSre)

by (3.32). However, even if \( \| \hat{V} H \hat{\ell} \|_2 \) is small, if \( \tilde{\eta} \) is very small the normalized vector \( \tilde{v} = \hat{\ell} / \tilde{\eta} \) satisfies only (3.18). Thus, for situations where \( \rho = \| \tilde{h} \|_2 / \tilde{\eta} \) is large, i.e., of the order \( \varepsilon_{\text{sum}}^{-1} \), the orthogonality of \( \tilde{v} \) could be lost completely.

To improve this situation we investigate that if \( \| \hat{\ell} \|_2 / \| \hat{w} \|_2 \) is small, the numerical cancellation must occur while computing \( \hat{\ell} \) which might be inaccurate relative to its length. To correct this we add a reorthogonalization step to Algorithm 3.1. This strategy was already pursued in [25], for the CGS method. We investigate in the following how reorthogonalization improves orthogonality of \( \tilde{v} \) by projecting the result of the first orthogonalization step \( \hat{\ell} \) a second time, such that \( f^{(0)} \) is orthogonalized to \( \tilde{V} \).

3.2.1 Algorithm

The compensated Gram Schmidt method with reorthogonalization is initialized with a vector \( \tilde{w} \in \mathbb{C}^n \), a matrix \( \tilde{V} \in \mathbb{C}^{n \times k} \), and the nonsingular cross product matrix \( D = \tilde{V} H \tilde{V} \in \mathbb{C}^{k \times k} \). Due to the inexact vector operations, the method computes a vector \( \tilde{v} \in \mathbb{C}^n \) that is a nearly orthonormal to the columns of the matrix \( \tilde{V} \). Furthermore, ComGS determines the orthogonalization coefficients \( \tilde{h} \in \mathbb{C}^k, \tilde{\eta} \in \mathbb{R} \) and computes the updated cross product matrix \( \tilde{D}_+ \). Again, we indicate by "\( \approx \)" that we consider inexact operations.

**Algorithm 3.2** Compensated Gram Schmidt with reorthogonalization

**Input:** \( \tilde{w} \in \mathbb{C}^n, \tilde{V} \in \mathbb{C}^{n \times k}, D \in \mathbb{C}^{k \times k}, D \) nonsingular

**Output:** \( \tilde{v} \in \mathbb{C}^n, \tilde{h} \in \mathbb{C}^k, \tilde{\eta} \in \mathbb{C}, D_+ \in \mathbb{C}^{k+1 \times k+1} \)

1: \( \tilde{s}^{(0)} \approx D^{-1} \tilde{V} H \tilde{w} \) (orthogonalization)

2: \( \tilde{\ell}^{(0)} \approx \tilde{w} - \tilde{V} \tilde{s}^{(0)} \)

3: \( \tilde{s}^{(1)} \approx D^{-1} \tilde{V} H \tilde{\ell}^{(0)} \)

4: \( \tilde{\ell}^{(1)} \approx \tilde{\ell}^{(0)} - \tilde{V} \tilde{s}^{(1)} \) (reorthogonalization)

5: \( \tilde{\eta} \approx \| \tilde{\ell}^{(1)} \|_2 \)

6: if \( \tilde{\eta} = 0 \) then stop end if

7: \( \tilde{v} \approx \tilde{\ell}^{(1)} / \tilde{\eta} \) (normalization)

8: \( D_+ \approx \begin{bmatrix} D & \tilde{V} H \tilde{v} \\ \tilde{v} H \tilde{V} & \tilde{v} H \tilde{v} \end{bmatrix} \) (update \( D \))

9: \( \tilde{h} = \tilde{s}^{(0)} + \tilde{s}^{(1)} \)

As in Section 3.1.1 neither of the operations can be evaluated exactly and the perturbations that appear from Step (1) to Step (7) of Algorithm 3.2 are defined
3.2 ComGS with reorthogonalization (ComGSre)

by

\begin{align*}
1: & \quad f^{(P)} := D^{-1} \tilde{V}^H \tilde{w} - \tilde{s}^{(0)}, & (3.35) \\
2: & \quad f^{(0)} := \tilde{w} - \tilde{V} \tilde{s}^{(0)} - \tilde{\ell}^{(0)}, & (3.36) \\
3: & \quad f^{(P1)} := D^{-1} \tilde{V}^H \tilde{s}^{(0)} - \tilde{\ell}^{(1)}, & (3.37) \\
4: & \quad f^{(1)} := \tilde{\ell}^{(0)} - \tilde{V} \tilde{s}^{(1)} - \tilde{\ell}^{(1)}, & (3.38) \\
5: & \quad f^{(N)} := \|\tilde{\ell}^{(1)}\|_2 - \tilde{\eta}, & (3.39) \\
7: & \quad f^{(S)} := \tilde{\ell}^{(1)} - \tilde{v} \tilde{\eta}, & (3.40) \\
\end{align*}

respectively. As for the ComGS method, the superscripts \((P), (0), (P1), (1), (N),\) and \((S)\) indicate the actual source of the perturbation, i.e., the additional \(f^{(P1)}\) arises from in the inner product of the first reorthogonalization step and \(f^{(1)}\) from the orthogonalization of the first reorthogonalization step. Again, the number before each perturbation refers to the step of Algorithm 3.2 where the corresponding perturbation occurs.

Additional to the perturbations (3.35), (3.36), (3.39), and (3.40) that arise already in Algorithm 3.1, we introduce the perturbations (3.37) and (3.38). Similarly to (3.1), the norm of (3.37) is induced by the inner product \(\tilde{V}^H \tilde{\ell}^{(0)}\) and solving the linear system \(\tilde{V}^H \tilde{\ell}^{(0)} \approx D \tilde{s}^{(1)},\) i.e.,

\[\|f^{(P1)}\|_2 \leq \frac{\varepsilon_s}{1 - \varepsilon_s \kappa(D)} \left(1 + 2 \sqrt{k} + \kappa(D)\right) \|D^{-1}\|_2 \|\tilde{V}\|_2 \|\tilde{\ell}^{(0)}\|_2.\] (3.41)

The vector sum in the reorthogonalization step is inexact which causes the perturbation (3.38). Following the arguments for determining the bound for (3.2) it follows that

\[\|f^{(1)}\|_2 \leq \varepsilon_v \left(\|\tilde{\ell}^{(0)}\|_2 + \sqrt{k} \|\tilde{V}\|_2 \|\tilde{s}^{(1)}\|_2\right).\] (3.42)

Moreover, replacing \(\tilde{\ell}\) by \(\tilde{\ell}^{(1)}\) and \(\tilde{h}\) by \(\tilde{s}^{(0)}\) the inequalities (3.10)-(3.14) still hold, i.e.,

\[\|f^{(0)}\|_2 \leq \varepsilon_v \left(\|\tilde{w} + \sqrt{k} \|\tilde{V}\|_2 \|\tilde{s}^{(0)}\|_2\right),\] (3.43)

\[\|f^{(P)}\|_2 \leq \frac{\varepsilon_s}{1 - \varepsilon_s \kappa(D)} \left(1 + 2 \sqrt{k} + \kappa(D)\right) \|\tilde{V}\|_2 \|\tilde{w}\|_2,\] (3.44)

\[\|f^{(N)}\|_2 \leq \varepsilon_s 2 \|\tilde{\ell}^{(1)}\|_2,\] and

\[\|f^{(S)}\|_2 \leq \varepsilon_v \|\tilde{\ell}^{(1)}\|_2.\] (3.45)

The effect on the orthogonality property of the reorthogonalization step in Algorithm 3.2 is analyzed in the following section.
3.2 ComGS with reorthogonalization (ComGSre)

3.2.2 Distance to orthogonality of \( \tilde{v} \)

Similar to Theorem 3.7 we compute in the following a bound on the vO-measure to analyze the distance to orthogonality of the computed vector \( \tilde{v} \) to the column space of \( \tilde{V} \).

Assumption 3.10. The perturbations in the steps 1-5 and 7-8 of Algorithm 3.2 are bounded by (3.41)-(3.46) for some \( 0 \leq \varepsilon_v < 1 \), \( 0 \leq \varepsilon_s < 1 \), and \( \varepsilon_{\text{sum}} + 2\varepsilon_s + 3\varepsilon_v(1 + 2\sqrt{k}\theta) < 1 \) with \( \theta = \|D^{-1}\|_2 \|\tilde{V}\|_2^2 \). It holds that \( \|\tilde{V}^H\tilde{V} - D\|_2 < 1/\|D^{-1}\|_2 \).

Theorem 3.11. Let \( \tilde{w}, \tilde{V}, D, \tilde{v}, \tilde{h} \), and \( \tilde{\eta} \) be as in Algorithm 3.2 and assume that no breakdown occurs, i.e., \( \tilde{\eta} \neq 0 \) in Step 6. Let \( \rho := \|\tilde{h}\|_2/\|\tilde{\eta}\|_2 \). If Assumption 3.10 holds, then there exists an \( \varepsilon_{\text{sum}} < 1 \) with

\[
\varepsilon_{\text{sum}} \leq \varepsilon_v(1 + \sqrt{k}\theta) + \varepsilon_s \frac{k\|D^{-1}\|_2 + 2\left(1 + 2\sqrt{k} + \|D\|_2\|D^{-1}\|_2\right)}{1 - 4(\varepsilon_v + \varepsilon_s) - \varepsilon_s\|D^{-1}\|_2(k + \|D\|_2)}
\]

such that

\[
\begin{align*}
\|\tilde{\eta} - [\tilde{V}, \tilde{v}] [\tilde{h} \tilde{\eta}] \|_2 & \leq \varepsilon_v \left((1 + 2\sqrt{k}\theta)(2 + \varepsilon_{\text{sum}}) + (1 + \varepsilon_{\text{sum}})^2\right) \|\tilde{w}\|_2, \\
\|\tilde{\eta}\|_2 - 1 & \leq \frac{\varepsilon_v + 2\varepsilon_s}{1 - 2\varepsilon_s}, \\
\frac{\|\tilde{V}^H\tilde{v}\|_2}{\|\tilde{V}\|_2} & \leq \frac{\varepsilon_{\text{sum}}^2(1 + \theta)(1 + \|\tilde{V}\|_2\rho)}{1 - \varepsilon_{\text{sum}} - 2\varepsilon_s - 3\varepsilon_v(1 + 2\sqrt{k}\theta)} + \frac{\varepsilon_{\text{sum}} + \varepsilon_v}{1 - 2\varepsilon_s - \varepsilon_{\text{sum}}}
\end{align*}
\]

Proof. The proof of the bound (3.48) is analogous to the proof of (3.17) in Theorem 3.7. The first orthogonalization step of ComGSre (i.e., steps 1. and 2. of Algorithm 3.2) is equivalent to the orthogonalization step of ComGS (steps 1. and 2. of Algorithm 3.1), the bound on \( \|\tilde{\ell}^{(0)}\|_2 \) coincides with the bound (3.25) on \( \|\tilde{\ell}\|_2 \), i.e.,

\[
\|\tilde{\ell}^{(0)}\|_2 \leq (1 + \varepsilon_{\text{sum}})\|\tilde{w}\|_2,
\]

where \( \varepsilon_{\text{sum}} \) is defined in (3.26). Thus the first orthogonalization step can lengthen a vector by a factor of \( (1 + \varepsilon_{\text{sum}}) \) and after a second step of orthogonalization we have

\[
\|\tilde{\ell}^{(1)}\|_2 \leq (1 + \varepsilon_{\text{sum}})\|\tilde{\ell}^{(0)}\|_2 \leq (1 + \varepsilon_{\text{sum}})^2\|\tilde{w}\|_2.
\]

Further, we conclude from (3.32) that

\[
\|\tilde{V}^H\tilde{\ell}^{(0)}\|_2 \leq \varepsilon_{\text{sum}}\|\tilde{V}\|_2\|\tilde{w}\|_2.
\]
Step 9 of Algorithm 3.2 and the perturbations (3.36) and (3.38) lead to
\[
\tilde{w} = \tilde{V}h + \tilde{\ell}^{(1)} + f^{(0)} + f^{(1)}. \tag{3.53}
\]

Using (3.40) and (3.53) we obtain
\[
\left\| \tilde{w} - [\tilde{V}, \tilde{v}] \left[ \frac{\tilde{h}}{\tilde{\eta}} \right] \right\|_2 = \| \tilde{w} - \tilde{V}h - \tilde{\eta} \|_2 = \| f^{(0)} + f^{(1)} + f^{(S)} \|_2 \\
\leq \| f^{(0)} \|_2 + \| f^{(1)} \|_2 + \| f^{(S)} \|_2.
\]

From (3.42), (3.43), and (3.46) we get
\[
\left\| \tilde{w} - [\tilde{V}, \tilde{v}] \left[ \frac{\tilde{h}}{\tilde{\eta}} \right] \right\|_2 \leq \varepsilon_v \left( \| \tilde{w} \|_2 + \| \tilde{\ell}^{(0)} \|_2 + \| \tilde{\ell}^{(1)} \|_2 \\
+ \sqrt{k} \| \tilde{V} \|_2 \left( \| \tilde{s}^{(0)} \|_2 + \| \tilde{s}^{(1)} \|_2 \right) \right). \tag{3.54}
\]

Step 9 of Algorithm 3.2 and the perturbations (3.35) and (3.37) lead to
\[
\| \tilde{h} \|_2 \leq \| \tilde{s}^{(0)} \|_2 + \| \tilde{s}^{(1)} \|_2 \\
\leq \| D^{-1} \tilde{V}^H \tilde{w} \|_2 + \| f^{(P)} \|_2 + \| D^{-1} \tilde{V}^H \tilde{\ell}^{(0)} \|_2 + \| f^{(P1)} \|_2.
\]

Applying (3.41) and (3.44) to (3.50) and using (3.30) results in
\[
\| \tilde{h} \|_2 \leq \| \tilde{s}^{(0)} \|_2 + \| \tilde{s}^{(1)} \|_2 \leq 2 \| D^{-1} \|_2 \| \tilde{V} \|_2 \left( \| \tilde{\ell}^{(0)} \|_2 + \| \tilde{w} \|_2 \right) \\
\leq 2(2 + \varepsilon_{sum}) \| D^{-1} \|_2 \| \tilde{V} \|_2 \| \tilde{w} \|_2. \tag{3.55}
\]

Inserting the second inequality of (3.55) and the bounds (3.50) and (3.51) into inequality (3.54) proves (3.47).

Taking norms in (3.53) and inserting (3.43) and (3.42) implies that
\[
\| \tilde{w} \|_2 \leq \| \tilde{\ell}^{(1)} \|_2 + \| \tilde{V} \|_2 \| \tilde{h} \|_2 \\
+ \varepsilon_v \left( \| \tilde{w} \|_2 + \| \tilde{\ell}^{(0)} \|_2 + \sqrt{k} \| \tilde{V} \|_2 \left( \| \tilde{s}^{(0)} \|_2 + \| \tilde{s}^{(1)} \|_2 \right) \right).
\]

Thus, with (3.50), (3.55), and (3.33) (replacing \( \tilde{\ell} \) by \( \tilde{\ell}^{(1)} \)) we obtain
\[
\| \tilde{w} \|_2 \leq \left( \frac{1}{1 - 2\varepsilon_s} + \| \tilde{V} \|_2 \rho \right) \tilde{\eta} + \varepsilon_v \left( 2(2 + \varepsilon_{sum}) + 2(2 + \varepsilon_{sum}) \sqrt{k} \theta \right) \| \tilde{w} \|_2 \\
\leq \frac{\tilde{\eta}}{1 - 2\varepsilon_s} \left( 1 + \| \tilde{V} \|_2 \rho \right) + \varepsilon_v \left( 2 + \varepsilon_{sum} \right) \left( 1 + 2\sqrt{k} \theta \right) \| \tilde{w} \|_2.
\]
3.2 ComGS with reorthogonalization (ComGSre)

Approximating \((2 + \varepsilon_{\text{sum}})\) by 3 and solving for \(\|\tilde{w}\|_2\) results in
\[
\|\tilde{w}\|_2 \leq \frac{\tilde{\eta}}{1 - 2\varepsilon_s - 3\varepsilon_v \left(1 + 2\sqrt{k}\theta\right)} (1 + \|\tilde{V}\|_2 \rho).
\] (3.56)

Note that performing Steps 3–8 of Algorithm 3.2 is equivalent to \([\tilde{\nu}, \tilde{\eta}, \tilde{\eta}, D_+] = \text{ComGS}(\tilde{\ell}^{(0)}, \tilde{V}, D)\). Hence, for \(\tilde{\eta} > 0\) it follows from (3.31) (with the replacement of \(\tilde{\ell}\) by \(\tilde{\ell}^{(1)}\)) that
\[
\|\tilde{V}^H \tilde{\ell}\|_2 = \|\tilde{V}^H \left(\tilde{\ell}^{(1)} - f^{(S)}\right)\|_2 / \tilde{\eta} \leq \left(\|\tilde{V}^H \tilde{\ell}^{(1)}\|_2 + \|\tilde{V}^H f^{(S)}\|_2\right) / \tilde{\eta}.
\] (3.57)

To bound the first summand \(\|\tilde{V}^H \tilde{\ell}^{(1)}\|_2\), we conclude from (3.22) and the equivalence of the first two steps of the Algorithms 3.1 and 3.2 that
\[
\tilde{\ell}^{(1)} = \tilde{\ell}^{(1)} - \tilde{V} J \tilde{V}^H \tilde{\ell}^{(0)} + \tilde{V} f^{(P1)} - f^{(1)},
\]
where \(\tilde{\ell}^{(1)} := \tilde{\ell}^{(0)} - \tilde{V} (\tilde{V}^H \tilde{V})^{-1} \tilde{V}^H \tilde{\ell}^{(0)}\). Using the orthogonality of \(\tilde{\ell}^{(1)}\) to \(\tilde{V}\) and taking norms we get
\[
\|\tilde{V}^H \tilde{\ell}^{(1)}\|_2 \leq \|\tilde{V}\|_2 \|J\|_2 \|\tilde{V}^H \tilde{\ell}^{(0)}\|_2 + \|\tilde{V}\|_2 \left(\|f^{(1)}\|_2 + \|\tilde{V}\|_2 \|f^{(P1)}\|_2\right).
\] (3.58)

From (3.37) and (3.41) we obtain
\[
\|\tilde{s}^{(1)}\|_2 \leq \left(1 + \frac{\varepsilon_s}{1 - \varepsilon_s \kappa(D)} \left(1 + 2\sqrt{k} + \kappa(D)\right)\right) \|D^{-1}\|_2 \|\tilde{V}\|_2 \|\tilde{\ell}^{(0)}\|_2.
\]

Together with (3.41) and (3.42) it follows that
\[
\|f^{(1)}\|_2 + \|\tilde{V}\|_2 \|f^{(P1)}\|_2 \leq \left(\frac{\varepsilon_s \left(1 + 2\sqrt{k} + \|D\|_2 \|D^{-1}\|_2\right)}{1 - \varepsilon_s \kappa(D)} (1 + \varepsilon_v \sqrt{k})\theta + \varepsilon_v (1 + \sqrt{k}\theta)\right) \|\tilde{\ell}^{(0)}\|_2
\]
\[
\leq \varepsilon_{\text{sum}} \|\tilde{\ell}^{(0)}\|_2,
\] (3.59)

where the last approximation step is rather conservative in order to keep the structure of the bound simple. With (3.37) and (3.38) we have
\[
\tilde{\ell}^{(0)} = \tilde{V} D^{-1} \tilde{V}^H \tilde{\ell}^{(0)} + \tilde{\ell}^{(1)} + f^{(1)} - \tilde{V} f^{(P1)}.
\]

Inserting this into (3.59), taking norms, and solving for \(\|\tilde{V}\|_2 \|f^{(P1)}\|_2 + \|f^{(1)}\|_2\)
3.2 ComGS with reorthogonalization (ComGSre)

leads to
\[ \|f^{(1)}\|_2 + \|\tilde{V}\|_2\|f^{(p)}\|_2 \leq \frac{\varepsilon_{\text{sum}}}{1 - \varepsilon_{\text{sum}}} \left( \|\tilde{V}\|_2 \|D^{-1}\|_2 \|\tilde{V}^H \bar{e}^{(0)}\|_2 + \|\tilde{e}^{(1)}\|_2 \right). \]

Using this bound in (3.58) and dividing by the norm of \( \tilde{V} \) results in
\[ \frac{\|\tilde{V}^H \bar{e}^{(1)}\|_2}{\|\tilde{V}\|_2} \leq \|\tilde{V}\|_2 \left( \frac{\varepsilon_{\text{sum}}\|D^{-1}\|_2}{1 - \varepsilon_{\text{sum}}} + \|J\|_2 \right) \|\tilde{V}^H \bar{e}^{(0)}\|_2 + \frac{\varepsilon_{\text{sum}}\bar{e}_{\text{sum}}}{1 - \varepsilon_{\text{sum}}}. \]

Thus, with (3.33), the replacement of \( \bar{e} \) by \( \bar{e}^{(1)} \), and (3.52) we have
\[ \frac{\|\tilde{V}^H \bar{e}^{(1)}\|_2}{\|\tilde{V}\|_2} \leq \frac{\varepsilon_{\text{sum}}^2 (1 + \theta)}{1 - \varepsilon_{\text{sum}}} \|\bar{w}\|_2 + \frac{\varepsilon_{\text{sum}}\bar{\eta}}{1 - 2\varepsilon - \varepsilon_{\text{sum}}}, \]

where \( \|J\|_2 \|\tilde{V}\|_2^2 \) is approximated by \( \varepsilon_{\text{sum}} \). Inserting (3.56) we get
\[ \frac{\|\tilde{V}^H \bar{e}^{(1)}\|_2}{\|\tilde{V}\|_2} \leq \frac{\varepsilon_{\text{sum}}^2 (1 + \theta)(1 + \|\tilde{V}\|_2\bar{\rho})\bar{\eta}}{1 - \varepsilon_{\text{sum}} - 2\varepsilon - 3\varepsilon \left( 1 + 2\sqrt{k}\theta \right)} + \frac{\varepsilon_{\text{sum}}\bar{\eta}}{1 - 2\varepsilon - \varepsilon_{\text{sum}}}. \]

Further, from (3.46) and (3.33) by replacing \( \bar{e} \) by \( \bar{e}^{(1)} \) we obtain that
\[ \frac{\|\tilde{V}^H f^{(s)}\|_2}{\|\tilde{V}\|_2} \leq \varepsilon_{\text{v}} \|\tilde{e}^{(1)}\|_2 \leq \frac{\varepsilon_{\text{v}}}{1 - 2\varepsilon - \bar{\eta}}. \]

Finally, inserting (3.60) and (3.61) into (3.57) proves (3.49).

Remark 3.12. A comparison of Theorem 3.7 and Theorem 3.11 leads to
\[ \frac{\|\tilde{V}^H \bar{v}\|_2}{\|\tilde{V}\|_2} \leq \frac{\varepsilon_{\text{sum}}^{i+1} (1 + \theta)^i (1 + \|\tilde{V}\|_2\bar{\rho})}{1 - i\varepsilon_{\text{sum}} - 2\varepsilon - \varepsilon_{\text{v}} 3^i \left( 1 + 2\sqrt{k}\theta \right)} + \frac{\varepsilon_{\text{v}} + i\varepsilon_{\text{sum}}}{1 - 2\varepsilon - i\varepsilon_{\text{sum}}}. \]

for some \( \varepsilon_{\text{sum}} \) satisfying (3.15), where \( i \) is the number of reorthogonalization steps used in ComGS.

Remark 3.12 shows that the reorthogonalization step decreases the error. The dominant difference is the power of \( \varepsilon_{\text{sum}} \) in the “pre-factor” of \( (1 + \|\tilde{V}\|_2\bar{\rho}) \). This implies little change for moderate values of \( \bar{\rho} \), i.e., the distance from orthogonality is still on the order of \( \varepsilon_{\text{sum}} \). For large \( \bar{\rho} \), on the other hand, we observe a huge improvement. Roughly speaking, \( \bar{v} \) will be almost orthogonal for values of \( \bar{\rho} \) up to the order of \( \varepsilon_{\text{sum}}^{-1} \).

The properties of Algorithm 3.2 naturally raise the question whether results can be further improved by repeated reorthogonalization. In the vector case this situation is analyzed in [89, pp. 113–118], where the “twice is enough” algorithm is presented for the case of two vectors and it is referred to Kahan for a
corresponding analysis. An extension to several nearly orthonormal vectors is given in [25]. This "twice is enough" rule of thumb also holds in our situation under mild assumptions. We have already established that $\rho$ can reach the order $O(\varepsilon_{\text{sum}}^{-1})$. In order to grow even further, the perturbation $f^{(0)}$ would have to lie almost completely in the image of $V$. This is very unlikely as long as the perturbations can be interpreted as being generic.

Considering the computational costs of the different Gram Schmidt variants in terms of matrix-vector operations (MVecOPs), as defined in Remark 3.4, leads to the results stated in Table 3.1. In contrast to the standard schemes CGS and MGS, stated in the Algorithms 2.1 and 2.3, the ComGS method implicitly provides the basis $\hat{V} = V G^{-1}$. This basis is even closer to orthogonality. One additional matrix-vector operation in each orthogonalization step is needed to achieve this, as it is displayed in Table 3.1.

Table 3.1: Comparison of computational costs per step of CGS, ComGS and MGS

<table>
<thead>
<tr>
<th></th>
<th>CGS</th>
<th>ComGS</th>
<th>MGS</th>
<th>CGSre</th>
<th>ComGSre</th>
<th>MGSre</th>
</tr>
</thead>
<tbody>
<tr>
<td>vector operations</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>building $\hat{V} = V G^{-1}$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>parallel</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

However, if we would additionally compute $\hat{V}$ after the CGS and the MGS method we would end up with the same number of vector operations as for the ComGS method.

So far we have only considered the distance to orthogonality of just one vector $\tilde{v}$ to the previous basis $\tilde{V}$. Now we shift our attention to the whole cross product matrix $\tilde{V}^H \tilde{V}$ aiming for bounds on the O-measure, i.e., the distance of the cross product matrix from the identity. For this, in the following section we introduce the inexact Arnoldi method which produces the nearly orthonormal basis $\bar{V}$ of the corresponding Krylov subspace, taking into account that all vector operations can only be evaluated approximately. Further, we derive bounds on the distance to orthogonality of the computed Arnoldi basis $\bar{V}$.

### 3.3 Inexact Arnoldi’s method

In exact arithmetic Arnoldi’s method is stated in Algorithm 2.2. It is typically used to compute a few of the exterior eigenvalues of a large, sparse matrix. Assuming that $A \in \mathbb{C}^{n \times n}$ and $v_1 \in \mathbb{C}^n$ is normalized, the $k$-th Krylov subspace
3.3 Inexact Arnoldi’s method

\( \mathcal{K}_k(A, v_1) \), generated by the matrix \( A \) and the starting vector \( v_1 \), is defined by

\[
\mathcal{K}_k := \mathcal{K}_k(A, v_1) := \text{span}(v_1, Av_1, A^2v_1, \ldots, A^{k-1}v_1).
\]

In exact arithmetic the Arnoldi process computes the orthonormal basis \( V_k \) of the Krylov subspace, via the so-called Arnoldi relation

\[
AV_k = V_kH_k + v_{k+1}h_{k+1,k}e_k^T,
\]

where \( H_k \in \mathbb{C}^{k \times k} \) is a Hessenberg matrix with elements \( h_{ij} \) and \( e_k \) denotes the \( k \)-th column of the identity matrix. If we obtain \( h_{k+1,k} = 0 \), then we have \( AV_k = V_kH_k \), i.e., all eigenvalues of \( H_k \) are eigenvalues of \( A \). Otherwise, the eigenvalues of \( H_k \) are approximations, called Ritz values, to a few eigenvalues of \( A \).

Coming back to the case where each vector operation introduces an inaccuracy, we investigate in the following the inexact Arnoldi method. Thus, considering inexact vector operations we use the ComGS method of Section 3.1.1 in the orthogonalization step of Arnoldi’s method.

3.3.1 Inexact Arnoldi Algorithm

Initialized with a matrix \( A \in \mathbb{C}^{n \times n} \) and a normalized vector \( \tilde{v}_1 \in \mathbb{C}^n \), the inexact Arnoldi method constructs a search space basis \( \tilde{V}_k \in \mathbb{C}^{n \times k} \) and a Hessenberg matrix \( \tilde{H}_k \in \mathbb{C}^{k \times k} \) consisting of the orthogonalization coefficients. We add tildes “\( \sim \)” to the variables that appear in the standard Arnoldi method to emphasize that their computations entail perturbations. The variable \( D_k \) that does not appear in the standard algorithm is named without a tilde.

Algorithm 3.3 Inexact Arnoldi method

<table>
<thead>
<tr>
<th>Input:</th>
<th>( A \in \mathbb{C}^{n \times n} ), ( \tilde{v}_1 \in \mathbb{C}^n ), ( \parallel \tilde{v}_1 \parallel_2 = 1 ), ( m \in \mathbb{N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>( \tilde{V}<em>{m+1} \in \mathbb{C}^{n \times m+1} ), ( \tilde{H}<em>m \in \mathbb{C}^{m \times m} ), ( \tilde{h}</em>{m+1,m} ), ( D</em>{m+1} \in \mathbb{C}^{m+1 \times m+1} )</td>
</tr>
</tbody>
</table>

1. \( \tilde{V}_1 = \tilde{v}_1, D_1 = 1, H_0 = [ ] \in \mathbb{C}^{0 \times 0} \) (initialization)
2. for \( k = 1, 2, 3, \ldots, m \) do
3. \( \tilde{w}_{k+1} \approx A\tilde{v}_k \) (matrix mult.)
4. \( [\tilde{v}_{k+1}, \tilde{h}_{1:k,k}, \tilde{h}_{k+1:k}, D_{k+1}] = \text{ComGS}(\tilde{w}_{k+1}, \tilde{V}_k, D_k) \) (orthogonalization)
5. \( \tilde{H}_k = \begin{bmatrix}
\tilde{h}_{k-1}\tilde{k-1} & \tilde{h}_{1:k-1,k} \\
\tilde{h}_{k-1}\tilde{k-1} & \tilde{h}_{k,k}
\end{bmatrix} \) (update \( H_k \))
6. \( \tilde{V}_{k+1} = [\tilde{V}_k, \tilde{v}_{k+1}] \) (update \( V_k \))
7. end for

The perturbation of the matrix-vector multiplication in Step 3 of Algorithm 3.3
is indicated by \( \sim \). According to (3.6) this perturbation is modeled by

\[ f_{k+1}^{(M)} := A \tilde{v}_k - \tilde{w}_{k+1}, \quad (3.63) \]

where we assume that

\[ \|f_{k+1}^{(M)}\|_2 \leq \varepsilon_v \|\tilde{w}_{k+1}\|_2. \quad (3.64) \]

**Remark 3.13.** In our considerations we exclude the case that the Algorithm 3.3 breaks down. A breakdown occurs whenever the norm of orthogonalized vector \( \tilde{e} \) (see Algorithm 3.1 or 3.2) equals zero at a certain iteration step. This implies that the subdiagonal element \( h_{k+1,k} \) is zero and \( \tilde{e}_{k+1} \) cannot be computed. In this case the corresponding Krylov subspace is invariant and the eigenvalues and eigenvectors of \( \tilde{H}_k \) are exact eigenvalues and eigenvectors after lifting of a perturbed matrix \( \tilde{A} \). However, these eigenvalues and eigenvectors may differ from the exact eigenvalues and eigenvectors of the matrix \( A \), for further details on the relation between \( A \) and \( \tilde{A} \) see Section 3.4.1.

**Remark 3.14.** In the unperturbed case, i.e., all perturbations are zero, the matrix \( D_k \) becomes the identity matrix and Algorithm 3.3 reduces to the classic Arnoldi method stated in Algorithm 2.2.

In the following we derive bounds on the distance to orthogonality of the perturbed Arnoldi basis \( \tilde{V}_k \) and on the norm of the error of the Arnoldi relation (3.62).

### 3.3.2 Distance to orthogonality

As inner products cannot be evaluated exactly, the matrix \( D_k \) is actually computed cross product matrix. The matrices \( C_k \) and \( G_k \) are the Cholesky factors of \( \tilde{V}_k^H \tilde{V}_k \) and \( D_k \), respectively. Using these matrices we define the measures \( \| \tilde{V}_k^H \tilde{V}_k - I \|_F \), \( \| C_k - I \|_F \), \( \| D_k - I \|_F \), and \( \| \tilde{V}_k^H \tilde{V}_k - I \|_F \), where \( \tilde{V}_k = \tilde{V}_kG_k^{-1} \) for the distance to orthornormality. Furthermore, we use \( \| A\tilde{V}_{k-1} - \tilde{V}_k \tilde{H}_{k-1} \|_F \) to assess the quality of the Arnoldi relation.

In the literature especially the measure that we introduced as O-measure, i.e., \( \| \tilde{V}_k^H \tilde{V}_k - I \|_2 \), is widely used, see, e.g., in [36,38,42,104], and is a good estimator for the canonical distance \( \delta_{\text{orth}} := \min_{U \in \mathbb{C}^{n \times k}} \{ \| \tilde{V}_k - U \|_2 : U^H U = I \} \). More precisely, we have [48, Problem 19.14] \( \| \tilde{V}_k^H \tilde{V}_k - I \|_2 / (\| \tilde{V}_k \|_2 + 1) \leq \delta_{\text{orth}}(\tilde{V}_k) \leq \| D_k - I \|_2 \).

In the following we subdivide our derivation of the measures on the distance to orthogonality of the Arnoldi basis into two cases. First we consider the case without reorthogonalization, i.e., we use Algorithm 3.1 in the orthogonalization Step 4 of Algorithm 3.3 and in the second case we analyze the same measures using reorthogonalization in the orthogonalization Step, i.e., we use Algorithm 3.2 in Step 4 of Algorithm 3.3.
3.3 Inexact Arnoldi's method

3.3.2.1 Distance to orthogonality using ComGS

To derive error bounds on the quantities \( ||\hat{V}_k^H\hat{V}_k - I||_F, ||C_k - I||_F, ||D_k - I||_F, \)
\( ||\hat{V}_k^H\hat{V}_k - I||_F, \) and \( ||A\hat{V}_{k-1} - \hat{V}_k\hat{H}_{k-1}||_F \) we first introduce some useful tools.

In the following derivations we will make frequent use of the relation

\[
\frac{\alpha \varepsilon^2 + \beta \varepsilon}{1 - \gamma \varepsilon^2 - \delta \varepsilon} \leq \frac{\beta \varepsilon}{1 - (\gamma + \delta + \frac{\alpha}{\beta}) \varepsilon} \tag{3.65}
\]

for arbitrary \( \alpha, \beta, \gamma, \delta, \varepsilon > 0. \) Further, to prove the bounds on the measures \( ||\hat{V}_k^H\hat{V}_k - I||_F, ||C_k - I||_F, ||D_k - I||_F, ||\hat{V}_k^H\hat{V}_k - I||_F, \) and \( ||A\hat{V}_{k-1} - \hat{V}_k\hat{H}_{k-1}||_F, \) we use the following lemma and theorem.

**Lemma 3.15.** Let \( \{a_k\}, \{b_k\} \) be non-negative sequences and \( m \in \mathbb{N}. \) For \( 0 < \varepsilon < \frac{1}{1 - m \delta_{k-1}} \) define

\[
\delta_k := \delta_{k-1} + \frac{\varepsilon a_{k-1}}{1 - m \delta_{k-1} - \varepsilon b_{k-1}} \tag{3.66}
\]

with \( \delta_1 = \frac{2\varepsilon}{1 - 2\varepsilon} \) and

\[
a_k := \sum_{i=1}^{k-1} \alpha_i + 2, \tag{3.67}
\]

\[
b_k := \sum_{i=1}^{k-1} (\beta_i + m(k - i)\alpha_i + 2m) + 2, \tag{3.68}
\]

then

\[
\delta_k \leq \frac{\varepsilon a_k}{1 - \varepsilon b_k} \tag{3.69}
\]

for all \( k \geq 2. \)

**Proof.** We proceed by induction over \( k. \) With inequality (3.69) and the assumed \( \delta_1 = 2\varepsilon/(1 - 2\varepsilon) \) it follows that \( a_1 = b_1 = 2. \) For \( k = 2 \) we obtain that

\[
\delta_2 = \delta_1 + \frac{\varepsilon \alpha_1}{1 - m \delta_1 - \varepsilon \beta_1} \leq \frac{\varepsilon (\alpha_1 + 2)}{1 - \varepsilon (\beta_1 + 2m + 2)},
\]

i.e., the assertion holds for \( k = 2. \) Suppose that (3.69) holds for some positive integer \( k - 1. \) Equation (3.66) and the hypothesis

\[
\delta_{k-1} \leq \frac{\varepsilon a_{k-1}}{1 - \varepsilon b_{k-1}},
\]

lead to

\[
\delta_k \leq \frac{\varepsilon a_{k-1}}{1 - \varepsilon b_{k-1}} + \frac{\varepsilon a_{k-1}}{1 - m \left(\frac{\varepsilon a_{k-1}}{1 - \varepsilon b_{k-1}}\right) - \varepsilon \beta_{k-1}}
\]

52
3.3 Inexact Arnoldi’s method

\[
\frac{\varepsilon(a_{k-1} + a_{k-1})}{1 - \varepsilon(b_{k-1} + \beta_{k-1} + m\alpha_{k-1})}.
\]

It follows recursively that

\[
\delta_k \leq \frac{\varepsilon \left( \sum_{i=1}^{k-1} a_i + 2 \right)}{1 - \varepsilon \left( \sum_{i=1}^{k-1} \left( \frac{\beta_i + m \sum_{j=1}^{i-1} a_j + 2}{m} \right) + 2 \right)}.
\]

Together with (3.67) and (3.68) this proves (3.69).

**Theorem 3.16.** [114, Theorem 1.4] Let \( A_0 \) be an \( n \times n \) positive-definite matrix with the Cholesky factorization \( A_0 = C_0^H C_0 \). If \( \Delta A_0 \) is an Hermitian \( n \times n \) matrix satisfying

\[
\| A_0^{-1} \|_2 \| \Delta A_0 \|_F < 1,
\]

then there exists a unique Cholesky factorization

\[
A_0 + \Delta A_0 = (C_0 + \Delta C_0)(C_0 + \Delta C_0)^H,
\]

and

\[
\frac{\| \Delta C_0 \|_F}{\| C_0 \|_2} \leq \frac{\kappa(A_0) \varepsilon}{\sqrt{2} \left( 1 - \kappa(A_0) \varepsilon \right)},
\]

where \( \varepsilon = \| \Delta A_0 \|_F \) and \( \kappa(A_0) = \| A_0 \|_2 \| A_0^{-1} \|_2 \).

Now the following corollary quantifies bounds on the measures \( \| \tilde{V}_k^H \tilde{V}_k - I \|_F, \| C_k - I \|_F, \| D_k - I \|_F, \| \tilde{V}_k^H \tilde{V}_k - I \|_F, \) and \( \| A \tilde{V}_{k-1} - \tilde{V}_k \tilde{H}_{k-1} \|_F \).

**Corollary 3.17.** Let \( \tilde{V}_k, \tilde{H}_{k-1} \) and \( D_k \) result from Algorithm 3.3 applied to \( A \in \mathbb{C}^{n \times n} \). Let \( G_k \) and \( C_k \) be the Cholesky factors of \( D_k \) and \( \tilde{V}_k^H \tilde{V}_k \), respectively. Let \( \rho_k := \| \tilde{h}_{1,k} \|_2 / \tilde{h}_{k+1,k} \) and \( \rho_{\text{max}} := \max_{i=1,\ldots,k} \rho_i \). Let

\[
(\varepsilon_v + \varepsilon_s) \leq \frac{1}{k^6(1 + \rho_{\text{max}})}
\]

and \( \rho_{\text{max},0} = 0 \). Then, if Assumption 3.6 and bound (3.64) are fulfilled and \( k \geq 3 \),
the computed matrices $\tilde{V}_k, C_k, \tilde{H}_{k-1}, D_k,$ and $\tilde{V}_k = \tilde{V}_k G_k^{-1}$ satisfy

$$
\|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \frac{(\varepsilon_v + \varepsilon_s) p_{\tilde{V}}(k)(1 + \rho_{\max})}{1 - (\varepsilon_v + \varepsilon_s) q_{\tilde{V}}(k)(1 + \rho_{\max})},
$$

(3.71a)

$$
\|C_k - I\|_F \leq \frac{(\varepsilon_v + \varepsilon_s) p_C(k)(1 + \rho_{\max})}{1 - (\varepsilon_v + \varepsilon_s) q_C(k)(1 + \rho_{\max})},
$$

(3.71b)

$$
\|D_k - I\|_F \leq \frac{(\varepsilon_v + \varepsilon_s) p_D(k)(1 + \rho_{\max})}{1 - (\varepsilon_v + \varepsilon_s) q_D(k)(1 + \rho_{\max})},
$$

(3.71c)

$$
\|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \frac{\varepsilon_s p_{\tilde{V}}(k)}{1 - (\varepsilon_v + \varepsilon_s) q_{\tilde{V}}(k)(1 + \rho_{\max})},
$$

(3.71d)

$$
\|A \tilde{V}_{k-1} - \tilde{V}_k \tilde{H}_{k-1}\|_F \leq \frac{\varepsilon_v p_{A \tilde{V}}(k)\|A\|_2}{1 - (\varepsilon_v + \varepsilon_s) q_{A \tilde{V}}(k)(1 + \rho_{\max})},
$$

(3.71e)

where the low degree polynomials $p_s(k)$ and $q_s(k)$ are defined by

$$
p_{\tilde{V}}(k) = 4(k-1)^2k^2, \quad q_{\tilde{V}}(k) = 2k^5,
$$

$$
p_C(k) = 2\sqrt{2}(k-1)^2k^2, \quad q_C(k) = 2k^3(k+1)^2,
$$

$$
p_D(k) = 4(k-1)k^3, \quad q_D(k) = 2k^5,
$$

$$
p_{\tilde{V}}(k) = k, \quad q_{\tilde{V}}(k) = 2k^2(k+2)^2,
$$

$$
p_{A \tilde{V}}(k) = \frac{3}{2}(k-1)k^2, \quad \text{and} \quad q_{A \tilde{V}}(k) = 2k^5.
$$

Proof. We define the two non-negative sequences $\{\delta_k\}$ and $\{\xi_k\}$ by

$$
\delta_k := \delta_{k-1} + \frac{8(\varepsilon_v + \varepsilon_s) \tau_{k-1}^4(k-1)(1 + \rho_{k-1})}{1 - 2\delta_{k-1} - (\varepsilon_v + \varepsilon_s)(26 + 2(k-1)^2)},
$$

(3.72a)

$$
\xi_k := \frac{\delta_k}{\sqrt{2}(1 - \delta_k)}, \quad \text{for } k = 1, 2, \ldots .
$$

(3.72b)

The bounds (3.71a) and (3.71b) are proven in two steps. First we show that $\delta_k$ and $\xi_k$ are smaller than the right hand side of (3.71a) and (3.71b) and then that $\|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \delta_k$ and $\|C_k - I\|_F \leq \xi_k$, i.e., we prove that

$$
\|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \delta_k \leq \frac{4(\varepsilon_v + \varepsilon_s)(k-1)^2k^2(1 + \rho_{\max})}{1 - 2(\varepsilon_v + \varepsilon_s)k^5(1 + \rho_{\max})}
$$

and

$$
\|C_k - I\|_F \leq \xi_k \leq \frac{2\sqrt{2}(\varepsilon_v + \varepsilon_s)(k-1)^2k^2(1 + \rho_{\max})}{1 - 2(\varepsilon_v + \varepsilon_s)k^3(k+1)^2(1 + \rho_{\max})}.
$$

(3.73b)
3.3 Inexact Arnoldi’s method

We obtain the right inequalities of (3.73a) and (3.73b) by using

$$\tau_{k-1} = \min \{ \sqrt{k-1}, 1 + \zeta_{k-1} \} \leq \sqrt{k-1}$$

in the definition of \{\delta_k\}. More precisely, \{\delta_k\} in (3.72a) is of the form (3.66) with

$$a_{k-1} = 8(k-1)\tau_{k-1}^4(1 + \rho_{k-1}), \quad \beta_{k-1} = 2(k-1)^2 + 26 \quad \text{and} \quad m = 2.$$  

Consequently, using that \(\tau_{k-1} \leq \sqrt{k-1}\), the sequences \(a_k\) and \(b_k\) of Lemma 3.15 are bounded by

$$a_k \leq 8(k-1)(1 + \rho_{\max}) \sum_{i=1}^{k-1} i^2 + 2$$

with \(\frac{4}{3}k(2k-1) + 2 \leq 4k^2\) and

$$b_k \leq \sum_{i=1}^{k-1} \left( 26 + 2(k-1)^3 + 16(k-i)(k-1)i^2(1 + \rho_{\max}) + 4 \right) + 2$$

$$= 30(k-1) + 2(k-1)^3 + \left( \frac{8}{3}k^2(2k-1)^2 - 4(k-1)^3k^2 \right) (1 + \rho_{\max}) + 2$$

$$= 30(k-1) + 2(k-1)^3 + \frac{4}{3}(k+1)k^2(k-1)^2(1 + \rho_{\max}) + 2$$

$$\leq 2k^5(1 + \rho_{\max}).$$

In the last step of the approximation of \(b_k\) we replace \(\frac{4}{3}(k+1)k^2(k-1)^2\) by \(2k^5\), which more than compensates the omission of \(30(k-1) + 2(k-1)^3 + 2\). Finally, the bound (3.69) of Lemma 3.15 results to the bound (3.73a) for \(\delta_k\).

Note that, with the bound (3.70) on \((\varepsilon_v + \varepsilon_s)\), we have \(\delta_i < 1\) for all \(i = 1, \ldots, k\). Thus, inserting the right inequality of (3.73a) into (3.72b) leads to

$$\|C_k - I\|_F \leq \frac{2\sqrt{2}(\varepsilon_v + \varepsilon_s)(k-1)^2k^2(1 + \rho_{\max})}{1 - 2(\varepsilon_v + \varepsilon_s)k^2(k^3 + 2(k-1)^2)(1 + \rho_{\max})}.$$  

Further, using \(k^3 + 2(k-1)^2 \leq k(k+1)^2\) in the denominator results to the right inequality of (3.73b).

The second step, i.e., the proof of the left inequalities in (3.73a) and (3.73b) is done by induction over \(k\). For \(k = 1\) we have

$$\|\tilde{V}_1^H \tilde{V}_1 - I_1\|_2 = \|\tilde{V}_1^H \tilde{v}_1 - 1\|_2 \leq \left( \frac{1+\varepsilon_v}{1-2\varepsilon_s} \right)^2 - 1 \leq \frac{2(\varepsilon_s + \varepsilon_v)}{1 - 4(\varepsilon_s + \varepsilon_v)} = \delta_1.$$  

55
Moreover, \( \|C_1 - I_1\|_2 = \| \sqrt{\|\tilde{v}_1\|_2^2 - 1}\|_2 \leq (\epsilon_s + \epsilon_v) / (1 - 2(\epsilon_s + \epsilon_v)) \leq \zeta_1 \), i.e.,
the assertion holds for \( k = 1 \).

Let \( \tilde{V}_{k+1} = [\tilde{V}_k, \tilde{v}_{k+1}] \) and suppose that \( \|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \delta_k \) and \( \|C_k - I\|_F \leq \zeta_k \)
holds for some positive integer \( k \), then we have
\[
\|\tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\|_F \leq \|\tilde{V}_k^H \tilde{V}_k - I\|_F + 2\|\tilde{V}_k^H \tilde{v}_{k+1}\|_2 + (\|\tilde{v}_{k+1}\|_2^2 - 1). \tag{3.74}
\]
Inserting the induction hypothesis and (3.18) implies that
\[
\|\tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\|_F \leq \|\tilde{V}_k^H \tilde{V}_k - I\|_F + 2\|\tilde{V}_k^H \tilde{v}_{k+1}\|_2 + (\|\tilde{v}_{k+1}\|_2^2 - 1).
\tag{3.75}
\]
Using (3.19), the rightmost term is bounded by
\[
\|\tilde{v}_{k+1}\|_2^2 - 1 \leq (1 + \epsilon_v)^2/(1 - 2\epsilon_s)^2 - 1 \leq 2(\epsilon_s + \epsilon_v)/(1 - 4(\epsilon_s + \epsilon_v)). \tag{3.76}
\]
For \( \|\tilde{V}_k\|_2 \) we find
\[
\|\tilde{V}_k\|_2 = \|C_k\|_2 \leq \|I\|_2 + \|C_k - I\|_2 \leq 1 + \zeta_{k-1},
\]
because the matrix \( \tilde{V}_k C_k^{-1} \) has orthonormal columns. Moreover, (3.19) leads to
\[
\|\tilde{V}_k\|_2^2 \leq \|\tilde{V}_k\|_F^2 \leq k \left( \frac{1 + \epsilon_v}{1 - 2\epsilon_s} \right)^2,
\]
such that
\[
\|\tilde{V}_k\|_2 \leq \min \left\{ \frac{1 + \epsilon_v}{1 - 2\epsilon_s} \sqrt{k}, 1 + \zeta_k \right\}
\leq \frac{1}{1 - 2(\epsilon_s + \epsilon_v)} \min \{ \sqrt{k}, 1 + \zeta_k \}
\leq \frac{1}{1 - 2(\epsilon_s + \epsilon_v)} \tau_k. \tag{3.77}
\]
The assumed bound (3.70) on \( \epsilon_v + \epsilon_s \) and the right hand site of (3.73a) implies that \( \|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \delta_k < 1 \). Thus, Lemma 2.5 is applicable and leads to
\[
\| (\tilde{V}_k^H \tilde{V}_k)^{-1} \|_F \leq \frac{1}{1 - \|\tilde{V}_k^H \tilde{V}_k - I\|_F} \leq \frac{1}{1 - \delta_k}.
\]
3.3 Inexact Arnoldi’s method

Further, with (3.21) we get

\[ \|D_k - I\|_F = \|V_k^H \tilde{V}_k - K_k - I\|_F \leq \|V_k^H \tilde{V}_k - I\|_F + \|K_k\|_F. \] (3.78)

Consequently, by the right hand sight of (3.73a), the bound (3.27) and the assumed bound (3.70) it follows that \( \|D_k - I\|_F < 1 \). Hence the condition of Lemma 2.5 is satisfied such that the norm of \( D_k^{-1} \) is bounded by

\[ \|D_k^{-1}\|_2 \leq \|D_k\|_2 \leq \frac{1}{1 - \|D_k - I\|_F} \leq \frac{1}{1 - \|V_k^H \tilde{V}_k - I\|_F - \|K_k\|_F} \]

\[ \leq \frac{1}{1 - \delta_k - 4(\varepsilon_v + \varepsilon_s) - \varepsilon_s k}. \] (3.79)

Inserting (3.79) into (3.29) results to

\[ \|J_k\|_2 \leq \frac{\varepsilon_s k}{1 - \delta_k - 4(\varepsilon_v + \varepsilon_s) - 2\varepsilon_s k}. \] (3.80)

From (3.21) it follows that

\[ \|D_k\|_2 \leq \|\tilde{V}_k\|_2^2 + \|K_k\|_2, \]

i.e., using (3.77) and (3.27) we have

\[ \|D_k\|_2 \leq \frac{\tau_k^2}{1 - 4(\varepsilon_v + \varepsilon_s)} + \frac{\varepsilon_s k}{1 - 4(\varepsilon_v + \varepsilon_s)} \leq \frac{\tau_k^2 + \varepsilon_s k}{1 - 4(\varepsilon_v + \varepsilon_s)}. \]

Together with (3.79) and \( \tau_k \leq \sqrt{k} \) it follows that

\[ \kappa(D) = \|D_k\|_2 \|D_k^{-1}\|_2 \leq \frac{(1 + \varepsilon_s)k - 4(\varepsilon_v + \varepsilon_s)k}{1 - \delta_k - 8(\varepsilon_v + \varepsilon_s) - \varepsilon_s k}. \]

Further, inserting (3.77) and (3.79) into the definition of \( \theta_k \) results in

\[ \theta_k \leq \frac{(1 - 4(\varepsilon_v + \varepsilon_s)) \tau_k^2}{1 - \delta_k - 8(\varepsilon_v + \varepsilon_s) - \varepsilon_s k}. \] (3.81)

With the bound on \( \kappa(D) \), (3.77), (3.80), and (3.81), inequality (3.15) results in

\[ \varepsilon_{\text{sum}} \leq \varepsilon_s \left( k + 2 \left( 1 + 2\sqrt{k} + (1 + \varepsilon_s)k \right) \right) \tau_k^2 \]

\[ \frac{1}{(1 - \delta_k - 12(\varepsilon_v + \varepsilon_s) - 2\varepsilon_s k - \varepsilon_s (1 + \varepsilon_s)k - (1 - \delta_k - 8(\varepsilon_v + \varepsilon_s) - \varepsilon_s k) \varepsilon_v (1 + \sqrt{k} \tau_k^2)} \]

\[ \frac{1}{1 - \delta_k - 8(\varepsilon_v + \varepsilon_s) - \varepsilon_s k}. \] (3.82)
Using for $k \geq 2$ the rough approximation $1 + 2\sqrt{k} \leq 2k$ and the relation (3.65) it follows with $\epsilon_s < \epsilon_v$ that

$$\epsilon_{\text{sum}} \leq \frac{7\epsilon_s k \tau_k^2 + \epsilon_v k \tau_k^2}{1 - 2\delta_k - 20(\epsilon_v + \epsilon_s) - 5\epsilon_s k - 2\frac{2}{7}\epsilon_s} \leq \frac{4(\epsilon_v + \epsilon_s)k \tau_k^2}{1 - 2\delta_k - 21(\epsilon_v + \epsilon_s) - 5\epsilon_s k}. \quad (3.83)$$

Finally, inserting (3.76), (3.77), (3.81), and (3.83) into (3.75) results in

$$\| \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I \|_F \leq \delta_k + \frac{8(\epsilon_v + \epsilon_s)k \tau_k^3 (1 + \tau_k \rho_k)}{1 - 2\delta_k - 26(\epsilon_v + \epsilon_s) - 5\epsilon_s k - 2\epsilon_v \tau_k^2 \sqrt{k}} + \frac{2\epsilon_v \tau_k + 2(\epsilon_v + \epsilon_s)}{1 - 4(\epsilon_v + \epsilon_s)},$$

where we have used that $\tau_k \leq \sqrt{k}$. Because of $\epsilon_s < \epsilon_v$, the replacement of $(1 + \tau_k \rho_k)$ by $\tau_k (1 + \rho_k)$ more than compensates for $k \geq 2$ the neglection of the last summand, i.e.,

$$\| \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I \|_F \leq \delta_k + \frac{8(\epsilon_v + \epsilon_s)k \tau_k^4 (1 + \rho_k)}{1 - 2\delta_k - 26(\epsilon_v + \epsilon_s) - 5\epsilon_s k - 2\epsilon_v \tau_k^2 \sqrt{k}}.$$

To obtain a well structured bound, we use that for $\epsilon_s \leq \epsilon_v$, $\tau_k \leq \sqrt{k}$, and $k \geq 2$ we have $5\epsilon_s k + 2\epsilon_v \sqrt{k} \tau_k^2 \leq 2(\epsilon_v + \epsilon_s)k^2$ such that

$$\| \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I \|_F \leq \delta_k + \frac{8(\epsilon_v + \epsilon_s)k \tau_k^4 (1 + \rho_k)}{1 - 2\delta_k - (\epsilon_v + \epsilon_s) (26 + 2k^2)} = \delta_{k+1}.$$

In order to treat the Cholesky factor of $\tilde{V}_{k+1}^H \tilde{V}_{k+1}$, we consider Theorem 3.16 with $A_0 = I$, $\Delta A_0 = \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I$, $C_0 = I$, and $\Delta C_0 = C_{k+1} - I$. With the bound (3.70) on $(\epsilon_v + \epsilon_s)$, we have

$$\| I^{-1} \|_2 \| \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I \|_F < 1 \cdot \delta_{k+1} < 1,$$

thus Theorem 3.16 is applicable and results in

$$\| C_{k+1} - I \|_F \leq \frac{\| \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I \|_F}{\sqrt{2}(1 - \| \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I \|_F)} \leq \frac{\delta_{k+1}}{\sqrt{2}(1 - \delta_{k+1})} = \zeta_{k+1}.$$

Hence, the induction proof is complete.

Further, (3.70) implies that $\delta_{k+1} < 1$ and consequently ensures non-negativity of $\{ \zeta_{k+1} \}$. Finally, from (3.73a) and (3.73b) we obtain bounds (3.71a) and (3.71b) with the polynomials $p_\varphi(k)$, $q_\varphi(k)$, $p_C(k)$, and $q_C(k)$ stated in Corollary 3.17.
3.3 Inexact Arnoldi’s method

To prove the bound (3.71c) we insert (3.73a) and (3.27) into (3.78) and obtain

\[ \|D_k - I\|_F \leq \frac{4(\varepsilon_v + \varepsilon_s)(k-1)^2 k^2 (1 + \rho_{\text{max}})}{1 - 2(\varepsilon_v + \varepsilon_s)k^2 (1 + \rho_{\text{max}})} + \frac{\varepsilon_s k}{1 - 4(\varepsilon_s + \varepsilon_v)} \]

\[ \leq \frac{4(\varepsilon_v + \varepsilon_s)(k-1)^3 (1 + \rho_{\text{max}})}{1 - 2(\varepsilon_v + \varepsilon_s)k^3 (1 + \rho_{\text{max}})}. \]

To verify the bound (3.71d) we use that \( \hat{\nu}_k = \tilde{\nu}_k G^{-1}_k \) and (3.21), such that

\[ \| \hat{\nu}_k^H \hat{\nu}_k - I \|_F = \| \left( \tilde{\nu}_k G^{-1}_k \right)^H \tilde{\nu}_k G^{-1}_k - I \|_F \]

\[ = \| G^{-1}_k \left( \tilde{\nu}_k^H \tilde{\nu}_k - G_k^H G_k \right) G^{-1}_k \|_F \]

\[ = \| G^{-1}_k \tilde{\nu}_k \|_F \leq \| G^{-1}_k \|_F \| \tilde{\nu}_k \|_F. \quad (3.84) \]

The norm of the inverse of \( G_k \) is obtained by Theorem 3.16 with \( A_0 = I, \Delta A_0 = D_k - I, C_0 = I, \) and \( \Delta C_0 = G_k - I. \) More precisely, the assumption (3.70) ensures the applicability of Theorem 3.16 which provides, together with (3.71c), that

\[ \| G_k - I \|_F \leq \frac{2(\varepsilon_v + \varepsilon_s)(k-1)^2 k^2 (1 + \rho_{\text{max}})}{1 - 2(\varepsilon_v + \varepsilon_s)k^3 (1 + \rho_{\text{max}})}. \]

(3.85)

Hence, together with the bound (3.70) we have \( \| G_k - I \|_F < 1, \) i.e., Lemma 2.5 is applicable and implies

\[ \| G_k^{-1} \|_F \leq \frac{1}{1 - \| G_k - I \|_F} \leq \frac{1 - 2(\varepsilon_v + \varepsilon_s)k^4 (k+2) (1 + \rho_{\text{max}})}{1 - 2(\varepsilon_v + \varepsilon_s)k^4 (k+4) (1 + \rho_{\text{max}})}. \]

(3.86)

Finally, inserting (3.27) and (3.86) into (3.84) yields

\[ \| \hat{\nu}_k^H \hat{\nu}_k - I \|_F \leq \frac{\varepsilon_s k}{1 - 2(\varepsilon_v + \varepsilon_s)k^4 (k+3) (1 + \rho_{\text{max}}) - 4(\varepsilon_v + \varepsilon_s)}. \]

Replacing \( k(k+3) \) by \( (k+2)^2 \) in the denominator compensates the omission of the last summation \( -4(\varepsilon_v + \varepsilon_s) \) and proves (3.71d).

It remains to show (3.71e). From Step 3 of Algorithm 3.3 and the inequalities (3.16) and (3.64) we get

\[ \left\| A \tilde{\nu}_{k-1} - \begin{bmatrix} \tilde{h}_{1:k-1}^{-1} & \tilde{h}_{k}^{-1} \end{bmatrix} \begin{bmatrix} \hat{h}_{1:k-1}^{-1} & \hat{h}_{k}^{-1} \end{bmatrix} \right\|_2 \leq \varepsilon_v \left( 3 + \varepsilon_{\text{sum}} + 2\sqrt{k-1}\theta_{k-1} \right) \| \tilde{w}_k \|_2. \]

Inserting the bounds (3.81) and (3.83), leads to
Finally, by summing over all vectors of the basis \( e_k \) for which more than compensates the omission of \( (\cdot) \).

To simplify the denominator we replace 4 with 3.

Thus, inserting the right inequality of (3.73a) for \( \delta_{k-1} \) gives

\[
\| A\tilde{v}_{k-1} - [\tilde{V}_{k-1}, \tilde{v}_k] \left[ \frac{\tilde{h}_{1,k-1}}{\tilde{h}_{k,k-1}} \right] \|_2 \\
\leq \frac{3\varepsilon_v \sqrt{k} \tau_{k-1}^2 \| \tilde{w}_k \|_2}{1 - 2\delta_{k-1} - (\varepsilon_v + \varepsilon_s)(21 + 2\sqrt{k - 1}) - 5\varepsilon_s(k - 1)}.
\]

(3.87)

From (3.63), (3.64) and (3.19) we obtain

\[
\| \tilde{w}_k \|_2 \leq \frac{\| A\tilde{v}_{k-1} \|_2}{1 - \varepsilon_v} \leq \frac{\| A \|_2}{1 - (\varepsilon_v + \varepsilon_s)}.
\]

(3.88)

Thus, inserting the right inequality of (3.73a) for \( \delta_{k-1} \) and (3.88) into (3.87) gives

\[
\| A\tilde{v}_{k-1} - [\tilde{V}_{k-1}, \tilde{v}_k] \left[ \frac{\tilde{h}_{1,k-1}}{\tilde{h}_{k,k-1}} \right] \|_2 \leq \frac{3\varepsilon_v \sqrt{k} \tau_{k-1}^2 \| A \|_2}{1 - n(k)},
\]

with

\[
n(k) = (\varepsilon_v + \varepsilon_s) \left( 2(k - 1)^2(1 + \rho_{\text{max}})(4k - 2)^2 + (k - 1)^3 \right)
\]

\[
+ (\varepsilon_v + \varepsilon_s) \left( 23 + 2\sqrt{k - 1} \right) - 5\varepsilon_s(k - 1).
\]

To simplify the denominator we replace 4 with 3 by \( (k - 1)^2(k + 3) \), which more than compensates the omission of \( (\varepsilon_v + \varepsilon_s)(\frac{3}{3}\sqrt{k - 1}) + 5\varepsilon_s(k - 1) \) for \( k \geq 1 \), i.e.,

\[
\| A\tilde{v}_{k-1} - [\tilde{V}_{k-1}, \tilde{v}_k] \left[ \frac{\tilde{h}_{1,k-1}}{\tilde{h}_{k,k-1}} \right] \|_2 \leq \frac{3\varepsilon_v \sqrt{k} \tau_{k-1}^2 \| A \|_2}{1 - (\varepsilon_v + \varepsilon_s)(2(k - 1)^4(k + 3) + 23)(1 + \rho_{\text{max}})}.
\]

Finally, by summing over all vectors of the basis \( \tilde{V}_k \) and using (3.77), the norm of
the error in the Arnoldi relation is bounded by

\[ \| \mathbf{A} \overline{\mathbf{V}}_{k-1} - \overline{\mathbf{V}}_{k} \mathbf{H}_{k-1} \|_F \leq \sum_{i=1}^{k-1} \frac{3 \epsilon_v \sqrt{k_i}}{1 - (\epsilon_v + \epsilon_s) (2(k-1)^4(k+3) + 23)(1 + \rho_{\text{max}})} \| \mathbf{A} \|_2 \]

\[ \leq \frac{3}{2} \epsilon_v (k-1) k^2 \frac{1}{1 - 2(\epsilon_v + \epsilon_s) k^5(1 + \rho_{\text{max}})} \| \mathbf{A} \|_2, \]

where we used \((k-1)^4(k+3) + 23 \leq k^5\) for \(k \geq 2\) to simplify the structure of the denominator.

As expected, Corollary 3.17 provides similar results regarding the distance to orthogonality for the whole cross product matrix as Theorem 3.7 for a single vector. More precisely, we can ensure near orthogonality of the Krylov subspace basis \(\mathbf{e}_{V_k}\), as long as the factor \((1 + \rho_{\text{max}})\) stays moderate. This is the case whenever the subdiagonal element \(h_{k+1,k}\) of the Hessenberg matrix is not too small compared to the remaining elements in the \(k\)-th column of \(\mathbf{H}_k\). In other words as long as \(\mathbf{A} \mathbf{e}_{V_k}\) is securely linearly independent of the previous basis vectors, \(\rho_{\text{max}}\) will be moderate and the upper bound (3.71a) will be on the order of \((\epsilon_v + \epsilon_s)\).

However, \(\rho_{\text{max}}\) can become arbitrarily large, if the subdiagonal element \(h_{k+1,k}\) is very small. Usually, this situation is called a "lucky breakdown" as it indicates convergence of the search space to an invariant subspace of \(\mathbf{A}\). In the inexact case tiny subdiagonal elements cannot be considered as "lucky", since in this case the O-measure and its bound (3.71a) may get large, i.e., the orthogonality of \(\overline{\mathbf{V}}_{k+1}\) may be drastically perturbed. Note that in this case only \(\mathbf{e}_{V_{k+1}}\) is far from being orthogonal to \(\overline{\mathbf{V}}_k\). However, \(\overline{\mathbf{V}}_k\) itself will be an almost orthogonal basis of an almost invariant subspace.

**Remark 3.18.** Analyzing the distance from orthogonality, only the perturbations during the orthonormalization scheme, i.e., in Algorithm 3.1 or Algorithm 3.2, play a role; those occurring during the matrix-vector multiplication in Algorithm 3.3 Step 4 are insignificant for the deviation. Indeed, in the special case where perturbations occur only in Algorithm 3.3, but not in Algorithm 3.1 or Algorithm 3.2, the columns of \(\overline{\mathbf{V}}_k\) would be orthonormal. However, the spanned space would nevertheless cease to be a Krylov subspace of \(\mathbf{A}\).

**Remark 3.19.** Algorithm 3.1 additionally provides the cross product matrix \(\mathbf{D}_k\) and the corresponding Cholesky factor \(\mathbf{G}_k\). The bound (3.71d) indicates that the implicitly available basis \(\hat{\mathbf{V}}_k = \overline{\mathbf{V}}_k \mathbf{G}_k^{-1}\) maintains the orthogonality much better.

**Remark 3.20.** In the context of the closely related GMRES method it was noticed in [87, 118] and used in [34] that the Arnoldi relation (3.62) can be seen as QR decomposition of the matrix \([\overline{\mathbf{v}}_1 \mathbf{A} \overline{\mathbf{V}}_k]\). Hence, the Arnoldi process could be also
analyzed by the investigation of the QR decomposition

\[
[\tilde{v}_1 A \tilde{V}_k] = \tilde{V}_{k+1} \left[ e_1 \ \tilde{H}_k \right] = \tilde{V}_{k+1} \tilde{R}_{k+1},
\]

with the upper triangular matrix \( \tilde{R}_{k+1} = [e_1 \ \tilde{H}_k] \) and \( e_1 \) the first column of the \( k + 1 \) dimensional identity matrix.

Considering the case where \( A \tilde{v}_k \) is almost linearly dependent to the previous basis vectors \( \tilde{V}_k \), the following Corollary illustrates the improvement by using reorthogonalization in the ComGS method. More precisely, we investigate the same quantities as in Corollary 3.17, but now \( \tilde{V}_k, D_k, \) and \( \tilde{H}_{k-1} \) are obtained from Algorithm 3.3 where in Step 4 Algorithm 3.2 is used instead of Algorithm 3.1.

**Corollary 3.21.** Let \( \tilde{V}_k, H_{k-1} \) and \( D_k \) result from Algorithm 3.3, using Algorithm 3.2 in Step 4, applied to \( A \in \mathbb{C}^{n \times n} \). Let \( G_k \) and \( C_k \) be the Cholesky factor of \( D_k \) and \( V_k^H \tilde{V}_k \), respectively. Let \( \rho_k := \|\hat{h}_{1:k,k}\|_2 / \hat{h}_{k+1,k} \) and \( \rho_{\text{max}} := \max_{i=1,...,k} \rho_i \). Let \( \tau_{k-1} := \min\{\sqrt{k-1}, 1 + \xi_{k-1}\} \). Let

\[
(\varepsilon_v + \varepsilon_s) < \frac{1}{2(k+1)^6(1 + \rho_{\text{max}})}
\]

and \( \rho_{\text{max,0}} = 0 \). Then, if Assumption 3.10 is fulfilled and \( k \geq 4 \), the computed matrices \( \tilde{V}_k, C_k, \tilde{H}_{k-1}, D_k, \) and \( \tilde{V}_k = \tilde{V}_k G_k^{-1} \) satisfy

\[
\|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \frac{(\varepsilon_v + \varepsilon_s)p_\tilde{V}(k)}{1 - (\varepsilon_v + \varepsilon_s)q_\tilde{V}(k)(1 + \rho_{\text{max}})},
\]

\[
\|C_k - I\|_F \leq \frac{(\varepsilon_v + \varepsilon_s)p_C(k)}{1 - (\varepsilon_v + \varepsilon_s)q_C(k)(1 + \rho_{\text{max}})},
\]

\[
\|D_k - I\|_F \leq \frac{(\varepsilon_v + \varepsilon_s)p_D(k)}{1 - (\varepsilon_v + \varepsilon_s)q_D(k)(1 + \rho_{\text{max}})},
\]

\[
\|\tilde{V}_k^H \tilde{V}_k - I\|_F \leq \frac{\varepsilon_s p_\tilde{V}(k)}{1 - (\varepsilon_v + \varepsilon_s)q_\tilde{V}(k)(1 + \rho_{\text{max}})},
\]

\[
\|A \tilde{V}_{k-1} - \tilde{V}_k \tilde{H}_{k-1}\|_F \leq \frac{\varepsilon_v p_{A \tilde{V}}(k)}{1 - (\varepsilon_v + \varepsilon_s)q_{A \tilde{V}}(k)(1 + \rho_{\text{max}})} \|A\|_2,
\]

where \( p_\ast(k) \) and \( q_\ast(k) \) are low degree polynomials in \( k \) given by

\[
p_\tilde{V}(k) = 5(k - 1)k^3,
\]

\[
p_C(k) = \frac{5}{\sqrt{2}}(k - 1)k^3,
\]

\[
p_D(k) = 5k^4,
\]

\[
p_\tilde{V}(k) = k,
\]

\[
q_\tilde{V}(k) = 25(k - 1)^2k^3,
\]

\[
q_C(k) = 25(k - 1)k^4,
\]

\[
q_D(k) = 25(k - 1)^2k^3,
\]

\[
q_\tilde{V}(k) = 25k^5,
\]

62
3.3 Inexact Arnoldi's method

\[ p_{AV}(k) = 2(k-1)(2 + \sqrt{k-1}) \] and \[ q_{AV}(k) = 5(k-2)(k-1)^4. \]

Proof. The proof can be carried out similarly to the proof of Corollary 3.17. We define the two non-negative sequences \{\delta_k\} and \{\zeta_k\} by \[ \delta_1 = 2(5s + 5v), \] and \[ \delta_k = \frac{5(\epsilon_v + \epsilon_s)(k-1)k^2}{1 - 5\delta_{k-1} - (\epsilon_v + \epsilon_s)\left(54 + 6(k-1) + 20(k-1)^2\frac{\tau_k^2}{1 + \rho_k-1}\right)}, \] for \( k = 1, 2, \ldots \). (3.91a)

\[ \zeta_k = \frac{\delta_k}{\sqrt{2(1 - \delta_k)}} \] for \( k = 1, 2, \ldots \). (3.91b)

The proof of (3.90a) and (3.90b) is divided into two steps, i.e., from right to left we show that

\[ \|\tilde{V}_k^H\tilde{V}_k - I\|_2 \leq \delta_k \leq \frac{5(\epsilon_v + \epsilon_s)(k-1)k^3}{1 - 25(\epsilon_v + \epsilon_s)(k-1)^2k^3(1 + \rho_{\text{max}})} \] and (3.92a)

\[ \|C_k - I\|_2 \leq \zeta_k \leq \frac{5\sqrt{2}(\epsilon_v + \epsilon_s)(k-1)k^3}{1 - 25(\epsilon_v + \epsilon_s)(k-1)^2k^3(1 + \rho_{\text{max}})}. \] (3.92b)

In the first step the right inequalities of (3.92a) and (3.92b) are proven by the application of Lemma 3.15. In more detail, the sequence (3.91a) is of the structure (3.66) with

\[ a_{k-1} = 5(k-1)k^2, \]

\[ \beta_{k-1} = 54 + 6(k-1) + 20(k-1)^2\frac{\tau_k^2}{1 + \rho_{k-1}}, \]

and \( m = 5 \). Thus, with \( \min\{\sqrt{k}, 1 + \zeta_k\} \leq \sqrt{k} \) it follows from (3.67) and (3.68) that

\[ a_k = \sum_{i=1}^{k-1} 5(k-1)ki + 2 \]

\[ \leq 5(k-1)^2k^2 + 2 \]

\[ \leq 5(k-1)k^3 \]

and

\[ b_k = \sum_{i=1}^{k-1} \left(64 + 6(k-1) + 20(k-1)^2(1 + \rho_{\text{max}})i + 25(k-i)(k-1)ki\right) + 2 \]

\[ \leq 10(k-1)^3k(1 + \rho_{\text{max}}) + \frac{25}{2}k^3(k-1)^2 + \frac{25}{6}k^2(k-1)^2(2k-1) + 6(k-1)^2 + 64(k-1) + 2 \]

63
3.3 Inexact Arnoldi’s method

\[ \leq 25k^3(k - 1)^2(1 + \rho_{\text{max}}) \]

for \( k \geq 2 \). The last step of the of approximation of \( b_k \) is rather conservative in order to keep the structure of the bound simple.

Finally, the application of Lemma 3.15 leads to the right inequality of (3.92a). Note that by the assumed bound on \( \epsilon_{\text{sum}} \), we have \( \delta_i < 1 \) for all \( i = 1, \ldots, k \). Thus, inserting the right inequality of (3.92a) into (3.91b) leads to

\[
\frac{\delta_{k+1}}{\sqrt{2}(1 - \delta_{k+1})} \leq \frac{5}{\sqrt{2}}(\epsilon_v + \epsilon_s)k(k + 1)^3 \frac{1 - (\epsilon_v + \epsilon_s)(25k^2(k + 1)^3(1 + \rho_{\text{max}}) + 5k(k + 1)^3)}{1 - 25(\epsilon_v + \epsilon_s)k(k + 1)^4(1 + \rho_{\text{max}})},
\]

where the replacement of \( 25k^2(k + 1)^3 \) by \( 25k(k + 1)^4 \) in the last step compensates the negligence of \( 5k(k + 1)^3 \) in the denominator. Hence, we proved the right inequality of (3.92b).

The second step, the proof of the left inequalities of (3.92a) and (3.92b), is done by induction over \( k \). The case \( k = 1 \) is equivalent to the case \( k = 1 \) in the proof of Corollary 3.17. In the induction step we assume that \( \|\bar{V}_k^H\bar{V}_k - I\|_2 \leq \delta_k \) and \( \|C_k - I\|_2 \leq \zeta_k \) for some positive integer \( k \). Inserting the hypothesis, (3.48) and (3.49) into (3.74) implies that

\[
\|\bar{V}_{k+1}^H\bar{V}_{k+1} - I\|_F \leq \delta_k + \frac{2\epsilon_{\text{sum}}^2(1 + \theta_k)(1 + \|\bar{V}_k\|_2\rho_k)\|\bar{V}_k\|_2}{1 - \epsilon_{\text{sum}} - 2\epsilon_s - 3\epsilon_v(1 + 2\sqrt{k}\theta_k)} + \frac{2(\epsilon_v + \epsilon_s)\|\bar{V}_k\|_2}{1 - \epsilon_{\text{sum}} - 2\epsilon_s} + \frac{2(\epsilon_v + \epsilon_s)}{1 - 4(\epsilon_v + \epsilon_s)},
\]

With \( (1 + \|\bar{V}_k\|_2\rho_k)\|\bar{V}_k\|_2 \leq (1 + \rho_k)\|\bar{V}_k\|_2^2 \) and \( 2\epsilon_s + 3\epsilon_v(1 + 2\sqrt{k}\theta_k) \leq 3\epsilon_{\text{sum}} \) we get

\[
\|\bar{V}_{k+1}^H\bar{V}_{k+1} - I\|_F \leq \delta_k + \frac{2\epsilon_{\text{sum}}^2(1 + \theta_k)(1 + \rho_k)\|\bar{V}_k\|_2^2}{1 - 4\epsilon_{\text{sum}}} + \frac{2((\epsilon_v + \epsilon_{\text{sum}})\|\bar{V}_k\|_2 + (\epsilon_v + \epsilon_s))}{1 - \epsilon_{\text{sum}} - 2\epsilon_s}.
\]

(3.93)

Using \( 1 + 2\sqrt{k} \leq 2k \) for \( k \geq 2 \) in (3.83) as well as (3.77) and (3.81) the second summand of the right hand side of (3.93) can be bounded by

\[
\frac{2\epsilon_{\text{sum}}^2(1 + \theta_k)\|\bar{V}_k\|_2^2(1 + \rho_k)}{1 - 4\epsilon_{\text{sum}}} \leq \frac{2(16(\epsilon_v + \epsilon_s)^2k^2\tau_k^6(1 + \tau_k^2)(1 + \rho_k))}{1 - 5\delta_k - 54(\epsilon_v + \epsilon_s) - 11\epsilon_s k - 16(\epsilon_v + \epsilon_s)k\tau_k^2}.
\]
3.3 Inexact Arnoldi’s method

\[
\leq \frac{32(\varepsilon_v + \varepsilon_s)^2 k^2 (k + 1) \tau_k^k (1 + \rho_k)}{1 - 5\delta_k - (\varepsilon_v + \varepsilon_s)(54 + 6k + 16k\tau_k^2)}, \tag{3.94}
\]

where we have used in the last approximation step that \(\varepsilon_s < \varepsilon_v\) and \(\tau_k \leq \sqrt{k}\), i.e., \(1 + \tau_k^2 \leq k + 1\). Furthermore, with (3.77) and (3.82) the third summand of (3.93) is bounded by

\[
\frac{2 \left( (\varepsilon_v + \varepsilon_{\text{sum}}) \|\tilde{V}_k\|_2 + (\varepsilon_v + \varepsilon_s) \right)}{1 - \varepsilon_{\text{sum}} - 2\varepsilon_s} \leq \frac{2 \left( (\varepsilon_s (7k\tau_k^2 + \varepsilon_s k\tau_k^2) + 2\varepsilon_v k\tau_k^2) \tau_k\right)}{1 - 2\delta_k - (\varepsilon_v + \varepsilon_s)(21 + 3k + 5k\tau_k^2)} \\
\leq \frac{10(\varepsilon_v + \varepsilon_s)k\tau_k^3}{1 - 2\delta_k - (\varepsilon_v + \varepsilon_s)(22 + 3k + 5k\tau_k^2)}, \tag{3.95}
\]

where we have used (3.65), \(\tau_k \leq \sqrt{k}\), and \(\varepsilon_s < \varepsilon_v\) in the last inequality. Inserting (3.94) and (3.95) into (3.93) and using that \(\sqrt{k} \leq \frac{1}{2}(k + 1)\) for \(k \geq 2\) leads to

\[
\|\tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\|_F \leq \delta_k + \frac{50(\varepsilon_v + \varepsilon_s)^2 k^2 (k + 1) \tau_k^{10}(1 + \rho_k) + 5(\varepsilon_v + \varepsilon_s)k(k + 1)\tau_k^2}{1 - 5\delta_k - (\varepsilon_v + \varepsilon_s)(54 + 6k + 20k\tau_k^2)}.
\]

Hence, from (3.65) it follows that

\[
\|\tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\|_F \leq \delta_k + \frac{5(\varepsilon_v + \varepsilon_s)k(k + 1)\tau_k^2}{1 - 5\delta_k - (\varepsilon_v + \varepsilon_s)(54 + 6k + 20k^2\tau_k^2(1 + \rho_k))} = \delta_{k+1},
\]

with \(10(\varepsilon_v + \varepsilon_s)k\tau_k^4(1 + \rho_k) + 20k\tau_k^2 \leq 20(\varepsilon_v + \varepsilon_s)k^2\tau_k^2(1 + \rho_k)\) for \(k \geq 2\). The Cholesky factor of \(\tilde{V}_{k+1}^H \tilde{V}_{k+1}\) can be treated analogously to Corollary 3.17, thus with the bound (3.89) we have

\[
\|I^{-1}\|_2 \|\tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\|_F < 1 \cdot \delta_{k+1} < 1,
\]

which ensures the applicability of Theorem 3.16. For \(A_0 = I\), \(\Delta A_0 = \tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\), \(C_0 = I\), and \(\Delta C_0 = C_{k+1} - I\) Theorem 3.16 results in

\[
\|C_{k+1} - I\|_F \leq \frac{\|\tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\|_F}{\sqrt{2}(1 - \|\tilde{V}_{k+1}^H \tilde{V}_{k+1} - I\|_F)} \leq \frac{\delta_{k+1}}{\sqrt{2}(1 - \delta_{k+1})} = \zeta_{k+1},
\]

which completes the induction proof of the left inequalities of (3.92a) and (3.92b).

Further, the assumed bound (3.89) on \((\varepsilon_v + \varepsilon_s)\) implies that \(\delta_{k+1} < 1\), which ensures the non-negativity of \(\{\zeta_{k+1}\}\). Finally, from (3.92a) and (3.92b) we obtain the polynomials \(p_\gamma(k), q_\gamma(k), p_C(k),\) and \(q_C(k)\), stated in Corollary 3.21 and the bounds (3.90a) and (3.90b).
3.3 Inexact Arnoldi’s method

To prove the bound (3.90c) we insert (3.92a) and (3.27) into (3.78) such that

\[ \| D_k - I \|_F \leq \frac{5(\varepsilon_v + \varepsilon_s)(k - 1)k^3}{1 - 25(\varepsilon_v + \varepsilon_s)(k - 1)^2k^3(1 + \rho_{\text{max}})} + \frac{\varepsilon_sk}{1 - 4(\varepsilon_v + \varepsilon_s)}. \]

Substituting \( 5(k - 1)k^3 \) by \( 5k^4 \) more then compensates the omission of the last summand and proves (3.90c).

To derive the bound (3.71d) we start with the inequality (3.84). By (3.89), the assumed bounds on \( (\varepsilon_v + \varepsilon_s) \), we can ensure the applicability of Theorem 3.16. Consequently, the left inequality of (3.85) holds and inserting (3.90c) into (3.85) results in

\[ \| G_k - I \|_F \leq \frac{5\sqrt{2}(\varepsilon_v + \varepsilon_s)k^4}{1 - 5(\varepsilon_v + \varepsilon_s)k^4 \left( 5(k - 1)(1 + \rho_{\text{max}}) + \frac{1}{\sqrt{2}} \right)}, \]

(3.96)

with \( 5(k - 1)^2 + k < 5(k - 1)k \) for \( k \geq 2 \). By (3.89) we have \( \| G_k - I \|_F < 1 \), thus Lemma 2.5 is applicable and results in the left inequality of (3.86). By inserting (3.96) into the left inequality of (3.86) we get

\[ \| G_k^{-1} \|_F \leq \frac{1 - 25(\varepsilon_v + \varepsilon_s)(k - 1)k^4(1 + \rho_{\text{max}})}{1 - 5(\varepsilon_v + \varepsilon_s)k^4 \left( 5(k - 1)(1 + \rho_{\text{max}}) + \frac{1}{\sqrt{2}} \right) + 4}. \]

(3.97)

Further, inserting (3.27) and (3.97) into (3.84) yields

\[ \| \hat{\mathcal{V}}_k^H \hat{V}_k - I \|_F \leq \frac{\varepsilon_sk}{1 - (\varepsilon_v + \varepsilon_s) \left( 5k^4 \left( 5(k - 1)(1 + \rho_{\text{max}}) + \frac{1}{\sqrt{2}} \right) + 4 \right)}. \]

Finally, substituting in the denominator \( 5k^4 \) by \( 5k \) compensates the omission of \( \frac{5}{\sqrt{2}}k^4 + 4 \) and proves (3.90d).

To verify the remaining bound (3.90e), we obtain from (3.63) that

\[ A\tilde{\mathbf{v}}_{k-1} - [\tilde{\mathbf{V}}_{k-1}, \tilde{\mathbf{v}}_k] \begin{bmatrix} \hat{h}_{1:k-1,k-1} \\ \hat{h}_{k,k-1} \end{bmatrix} = \tilde{\mathbf{w}}_k - [\tilde{\mathbf{V}}_{k-1}, \tilde{\mathbf{v}}_k] \begin{bmatrix} \hat{h}_{1:k-1,k-1} \\ \hat{h}_{k,k-1} \end{bmatrix} + f_k^{(M)}. \]

Taking norms and inserting the bounds (3.47) and (3.64) gives

\[ \| A\tilde{\mathbf{v}}_{k-1} - [\tilde{\mathbf{V}}_{k-1}, \tilde{\mathbf{v}}_k] \begin{bmatrix} \hat{h}_{1:k-1,k-1} \\ \hat{h}_{k,k-1} \end{bmatrix} \|_2 \leq \varepsilon_v \left( (1 + 2\sqrt{k - 1}\theta_{k-1})(2 + \varepsilon_{\text{sum}}) + (1 + \varepsilon_{\text{sum}})^2 + 1 \right) \| \tilde{\mathbf{w}}_k \|_2. \]
Moreover, it follows from $\varepsilon_{\text{sum}} \leq 1$ and (3.65) that
\[
\left\| A\tilde{v}_{k-1} - \left[ \tilde{V}_{k-1}, \tilde{v}_k \right] \begin{bmatrix} \tilde{h}_{1:k-1,k-1} \\ \tilde{h}_{k,k-1} \end{bmatrix} \right\|_2 \\
\leq \varepsilon_v \left( 4(1 + \sqrt{k-1} \theta_{k-1}) + 2\varepsilon_{\text{sum}}(2 + \sqrt{k-1} \theta_{k-1}) \right) \|\tilde{w}_k\|_2 \\
\leq \frac{4\varepsilon_v(1 + \sqrt{k-1} \theta_{k-1})}{1 - 2\varepsilon_{\text{sum}}} \|\tilde{w}_k\|_2,
\]
where we used that $4(1 + \sqrt{k-1} \theta_{k-1})/(2(2 + \sqrt{k-1} \theta_{k-1})) \leq 2$. Inserting the bounds (3.81), (3.83), and (3.88) and using that $\varepsilon_s < \varepsilon_v$ gives
\[
\left\| A\tilde{v}_{k-1} - \left[ \tilde{V}_{k-1}, \tilde{v}_k \right] \begin{bmatrix} \tilde{h}_{1:k-1,k-1} \\ \tilde{h}_{k,k-1} \end{bmatrix} \right\|_2 \\
\leq \frac{4\varepsilon_v(1 + \sqrt{k-1} \tau_{k-1}^2) \|A\|_2}{1 - 3\delta_{k-1} - (\varepsilon_v + \varepsilon_s) \left( 31 + 3(k-1) + 8(k-1) \tau_{k-1}^2 \right)}.
\]

Further, inserting the right inequality of (3.92a) for $\delta_{k-1}$ results in
\[
\left\| A\tilde{v}_{k-1} - \left[ \tilde{V}_{k-1}, \tilde{v}_k \right] \begin{bmatrix} \tilde{h}_{1:k-1,k-1} \\ \tilde{h}_{k,k-1} \end{bmatrix} \right\|_2 \\
\leq \frac{4\varepsilon_v(1 + \sqrt{k-1} \tau_{k-1}^2) \|A\|_2}{1 - (\varepsilon_v + \varepsilon_s) \left( 5(k-2)(k-1)^3(1+\rho_{\text{max}})(5k-7) + 31 + 3(k-1) + 8(k-1) \tau_{k-1}^2 \right)}.
\]

To obtain a simplified structure of the bound we use that the replacement of $5k - 7$ by $5(k - 1)$ in the denominator compensates the omission of $(31 + 3(k - 1) + 8(k-1) \tau_{k-1}^2)$ for $k \geq 3$. Thus we get
\[
\left\| A\tilde{v}_{k-1} - \left[ \tilde{V}_{k-1}, \tilde{v}_k \right] \begin{bmatrix} \tilde{h}_{1:k-1,k-1} \\ \tilde{h}_{k,k-1} \end{bmatrix} \right\|_2 \\
\leq \frac{4\varepsilon_v(1 + \sqrt{k-1} \tau_{k-1}^2) \|A\|_2}{1 - 5(\varepsilon_v + \varepsilon_s)(k-2)(k-1)^4(1 + \rho_{\text{max}})}.
\]

Finally, the summation over the $k$ basis vectors of $\tilde{V}_k$ gives
\[
\left\| A\tilde{V}_{k-1} - \tilde{V}_k \tilde{H}_{k-1} \right\|_2 \leq \sum_{i=1}^{k-1} \frac{4\varepsilon_v \left( 1 + \sqrt{k-1} i \right) \|A\|_2}{1 - 5(\varepsilon_v + \varepsilon_s)(k-2)(k-1)^4(1 + \rho_{\text{max}})} \\
\leq \frac{2\varepsilon_v(k-1) \left( 2 + \sqrt{k-1} k \right) \|A\|_2}{1 - 5(\varepsilon_v + \varepsilon_s)(k-2)(k-1)^4(1 + \rho_{\text{max}})}.
\]
3.3 Inexact Arnoldi’s method

which proves (3.90e).

Comparing the two Corollaries 3.17 and 3.21 we see that reorthogonalization decreases the distance to orthogonality. To be more specific, the factor \( (1 + \rho_{\text{max}}) \) vanishes in the numerator of all quantities, i.e., even in the case of tiny subdiagonal elements \( \bar{h}_{k+1,k} \) (up to the order of \((\varepsilon_{V} + \varepsilon_{s})\)) Algorithm 3.3 nearly keeps the orthogonality of the basis \( \bar{V}_{k+1} \).

Remark 3.22. We have presented two bounds for \( \| D_{k+1} - I \|_{2} \) and \( \| C_{k+1} - I \|_{2} \) for both cases without reorthogonalization, i.e., in (3.73a) and (3.73b) and with reorthogonalization, i.e., in (3.92a) and (3.92b). The first bounds are sharper, whereas the second ones are more compact. The sharper bounds, being recursively defined, are useful in practice, but do not convey much theoretical insight. Both compact bounds are explicit and of the form \( a_{1}(\#v + \#s)/\min\{a_{1}, a_{2}\} \), where \( a_{1} \), \( a_{2} \) are functions depending on the dimension of \( V_{k} \). So, they are useful if \( \min\{a_{1}, a_{2}\} \ll (\varepsilon_{V} + \varepsilon_{s})^{-1} \). This is the case whenever \( (\varepsilon_{V} + \varepsilon_{s})^{2} \rho_{\text{max}} \) is not large, where \( i \) is the number of orthogonalization steps. Then, \( V_{k}^{H}V_{k} \) differs from the identity by a term of the order \((\varepsilon_{V} + \varepsilon_{s})\). This means that our algorithms generate an almost orthonormal basis \( \bar{V}_{k} \).

In (3.71a) we have given a bound on the O-measure, where the basis \( \bar{V}_{k} \) was computed by the inexact Arnoldi Algorithm 3.3. In the following we relate this result to well known bounds on the loss of orthogonality of the CGS and the MGS method.

3.3.3 Comparison of the orthogonalization schemes

To compare the distance of orthogonality of the ComGS method with the distance to orthogonality of the CGS and the MGS method, we first introduce some well known bounds on the measure \( \| V_{k}^{H}V_{k} - I \|_{2} \) for the CGS and the MGS method.

It is well known that from the numerical point of view the CGS method may be unstable and may produce a set of vectors which is far from being orthogonal and in some cases even loses orthogonality completely, i.e., the value of the O-measure becomes larger than one, cf. for example, [12, 63]. To be more precise, the loss of orthogonality can be bounded in terms of the condition number \( \kappa(A) \) of \( A \), see, e.g., [37, 104], by

\[
\| V_{k}^{H}V_{k} - I \|_{2} \leq c_{1}(n,k)\varepsilon_{M}\kappa^{2}(A) + O\left(\varepsilon_{M}^{2}\right),
\]

(3.98)

where \( \varepsilon_{M} \) is the machine precision and \( c_{1} \) is a function depending on the dimension of \( V_{k} \). A simple change in the sequence of the computations in CGS leads to the MGS method. In exact arithmetic both algorithms are equivalent. However, if the computations are inexact, then MGS shows a better numerical behavior,
which is validated by the following bound on the loss of orthogonality

\[ \| \tilde{V}_k^H \tilde{V}_k - I \|_2 \leq c_2(n,k) \varepsilon_M \kappa(A) + O(\varepsilon_M^2), \]  

(3.99)

where again \( \varepsilon_M \) is the machine precision and \( c_2 \) is a function depending on the dimension of \( \tilde{V}_k \). This bound was proved in [12,16].

Since the bound (3.98) for the CGS method is less strict than the bound (3.99) for the MGS method, we compare in the following the result of Corollary 3.17 with the tighter bound (3.99). The dominant difference lies in the quantities \( \kappa(A) \) and \( \rho_{\text{max}} \). Considering Remark (3.20) the Arnoldi basis could be also computed via the QR decomposition, i.e.,

\[ [\tilde{e}_1 A \tilde{V}_{k-1}] = \tilde{V}_k [e_1 \ \tilde{H}_{k-1}] = \tilde{V}_k \tilde{R}_k, \]

with the upper triangular matrix \( \tilde{R}_k := [e_1 \ \tilde{H}_{k-1}] \) and \( e_1 \) the first column of the \( k \) dimensional identity matrix. For \( k \) sufficiently large we obtain

\[ \kappa(A) \approx \kappa(\tilde{R}_k) = \frac{\sigma_{\text{max}}(\tilde{R}_k)}{\sigma_{\text{min}}(\tilde{R}_k)} = \frac{\max_x \| \tilde{R}_k x \|_2}{\min_x \| \tilde{R}_k x \|_2}, \]

Assuming that \( \rho_{\text{max}} = \frac{\| \tilde{h}_{1:k-1} \|_2}{|\tilde{h}_{k:k-1}|} \), we choose \( x \) in the numerator to be the \( k \)-th column of the identity matrix and \( x \) in the denominator to be the \( k \)-th eigenvector of \( \tilde{R}_k \), denoted by \( y_k \). Then, we obtain

\[ \kappa(A) \geq \frac{\| \tilde{R}_k e_k \|_2}{\| \tilde{R}_k y_k \|_2} \geq \frac{\| \tilde{h}_{1:k,k-1} \|_2}{|\tilde{h}_{k:k-1}|} \geq \frac{\| \tilde{h}_{1:k-1,k-1} \|_2}{|\tilde{h}_{k,k-1}|} = \rho_{\text{max}}. \]

Consequently, there are some cases where the MGS method loses orthogonality much faster than the ComGS which is demonstrated in Section 3.5 by Example 3. However, in many cases the MGS method and the ComGS method will behave similar.

Concluding this section, we show a relation between Algorithms 3.1 and 3.2, more precisely, that Algorithm 3.2 is a special case of Algorithm 3.1.

**Theorem 3.23.** With \( P_{\tilde{V}_k} := \tilde{V}_k D_k^{-1} \tilde{V}_k^H \) the results of Algorithm 3.3 using Algorithm 3.1 and Algorithm 3.3 using Algorithm 3.2 coincide if the perturbations are related via

\[
\begin{align*}
f_{k+1}^{(P,\text{no})} &= \left( I - P_{\tilde{V}_k} \right) \left( f_{k+1}^{(0,\text{re})} - \tilde{V}_k f_{k+1}^{(P,\text{re})} \right) + \tilde{V}_k f_{k+1}^{(P,\text{re})} + \tilde{V}_k f_{k+1}^{(P,1,\text{re})}, \\
f_{k+1}^{(0,\text{no})} &= f_{k+1}^{(0,\text{re})} + f_{k+1}^{(1,\text{re})}, \\
f_{k+1}^{(S,\text{no})} &= f_{k+1}^{(S,\text{re})}.
\end{align*}
\]

(3.100)
3.3 Inexact Arnoldi’s method

\[ f_{k+1}^{(N, \text{no})} = f_{k+1}^{(N, \text{re})}, \]
\[ f_{k+1}^{(M, \text{no})} = f_{k+1}^{(M, \text{re})}. \]

Here, \( f_{k+1}^{(M, \text{no})} \), \( f_{k+1}^{(0, \text{no})} \), \( f_{k+1}^{(P, \text{no})} \), \( f_{k+1}^{(S, \text{no})} \), and \( f_{k+1}^{(N, \text{no})} \) are the perturbations occurring in Algorithm 3.1 whereas \( f_{k+1}^{(M, \text{re})} \), \( f_{k+1}^{(0, \text{re})} \), \( f_{k+1}^{(1, \text{re})} \), \( f_{k+1}^{(P, \text{re})} \), \( f_{k+1}^{(S, \text{re})} \), and \( f_{k+1}^{(N, \text{re})} \) are the perturbations occurring in Algorithm 3.2. The superscripts (re) and (no) at the perturbations indicate that reorthogonalization was used (re) or was not used (no).

**Proof.** The \( k \)-th iteration of Algorithm 3.3 using Algorithm 3.1 generates

\[ \tilde{w}_{k+1} = Av_k + f_{k+1}^{(M, \text{no})}, \]
\[ \tilde{V}_{k+1} = P_{\hat{V}} \tilde{w}_{k+1} - \tilde{V}_{f_{k+1}^{(P, \text{no})}}, \]
\[ \tilde{h}_{k+1} = \left\| (I - P_{\hat{V}}) \tilde{w}_{k+1} - \left( f_{k+1}^{(0, \text{no})} - \tilde{V}_{f_{k+1}^{(P, \text{no})}} \right) \right\|_2 - f_{k+1}^{(N, \text{no})}, \]
\[ \tilde{v}_{k+1} = \left( I - P_{\hat{V}} \right) \tilde{w}_{k+1} - \left( f_{k+1}^{(0, \text{no})} - \tilde{V}_{f_{k+1}^{(P, \text{no})}} \right) - f_{k+1}^{(S, \text{no})}, \]

whereas the \( k \)-th iteration of Algorithm 3.3 using Algorithm 3.2 yields

\[ \tilde{w}_{k+1} = Av_k + f_{k+1}^{(M, \text{re})}, \]
\[ \tilde{V}_{k+1} = P_{\hat{V}} \tilde{w}_{k+1} - P_{\hat{V}} \left( f_{k+1}^{(0, \text{re})} - \tilde{V}_{f_{k+1}^{(P, \text{re})}} \right) - \tilde{V}_{f_{k+1}^{(P, \text{re})}} - \tilde{V}_{f_{k+1}^{(P, \text{re})}}, \]
\[ \tilde{h}_{k+1} = \left\| \left( I - P_{\hat{V}} \right) \left( \tilde{w}_{k+1} - \left( f_{k+1}^{(0, \text{re})} - \tilde{V}_{f_{k+1}^{(P, \text{re})}} \right) \right) - \left( f_{k+1}^{(1, \text{re})} - \tilde{V}_{f_{k+1}^{(P, \text{re})}} \right) \right\|_2 - f_{k+1}^{(N, \text{re})}, \]
\[ \tilde{v}_{k+1} = \left( I - P_{\hat{V}} \right) \left( \tilde{w}_{k+1} - \left( f_{k+1}^{(0, \text{re})} - \tilde{V}_{f_{k+1}^{(P, \text{re})}} \right) \right) - \left( f_{k+1}^{(1, \text{re})} - \tilde{V}_{f_{k+1}^{(P, \text{re})}} \right) - f_{k+1}^{(S, \text{re})}. \]

Both sets of results coincide if (3.100) holds.

Algorithm 3.2 can be interpreted as a special case of Algorithm 3.1, i.e., in the following analysis we only consider Algorithm 3.3 using Algorithm 3.1.

**Remark 3.24.** Due to the fact that reorthogonalization requires extra scalar products and is only necessary for large \( \rho_{\text{max}} \), we propose to only carry out the reorthogonalization step if \( \rho_{\text{max}} \) exceeds a certain threshold. Also this hybrid technique fits in the framework of Algorithm 3.1.
3.4 Backward error analysis

So far we have provided bounds on the distance to orthogonality of the basis \( \tilde{V}_k \), computed via the ComGS method. Due to the non-orthogonality of \( \tilde{V}_k \) the Arnoldi relation (2.9) is not valid anymore. However, we show in the following that the relation (2.9) holds for the computed matrices \( \tilde{V}_{k+1} \) and \( \tilde{H}_k \) if we replace \( A \) by a nearby matrix \( A + E \), where \( E \) is small in norm.

3.4.1 Krylov-like relations

As mentioned in Section 2.3.2 Arnoldi's method without perturbations produces an orthonormal basis \( V_k \) and a Hessenberg matrix \( H_k \) satisfying an Arnoldi relation

\[
AV_k = V_k H_k + v_{k+1} h_{k+1,k} e_k^T.
\]

Due to the perturbations, this is no longer true for the results of Algorithm 3.3. Here we want to derive a relation between \( A, \tilde{V}_k \), and \( \tilde{H}_k \) that is close to an Arnoldi relation in order to admit a backward error analysis of our method. From Step 3 of Algorithm 3.3 and Steps 2 and 5 of Algorithm 3.1 for \( j = 1, \ldots, k \) we have

\[
A \tilde{v}_j = \tilde{w}_{j+1} + f^{(M)}_{j+1} \\
= \tilde{V}_j h_{1:j} + \tilde{\ell}_{j+1} + f_0^{(0)} + f^{(M)}_{j+1} \\
= \tilde{V}_j h_{1:j} + \| \tilde{\ell}_{j+1} \| \tilde{v}_{j+1} + f_0^{(0)} + f^{(S)}_{j+1} + f^{(M)}_{j+1} \\
= \tilde{V}_{k+1} \left[ \tilde{h}_{1:j} \| \tilde{\ell}_{j+1} \| \right] + f_0^{(0)} + f^{(S)}_{j+1} + f^{(M)}_{j+1}.
\]

Hence, after \( k \) steps we obtain the relation

\[
A \tilde{V}_k = \tilde{V}_{k+1} \tilde{H}_k + \tilde{v}_{k+1} \tilde{h}_{k+1,k} e_k^T + F_k.
\]

Here \( F_k = [f^{(M)}_2 + f^{(0)}_2 + f^{(S)}_2, \ldots, f^{(M)}_{k+1} + f^{(0)}_{k+1} + f^{(S)}_{k+1}] \) is the matrix consisting of the individual perturbations occurring throughout Algorithms 3.1 and 3.3. In order to bound the norm of \( F_k \), we assume that the corresponding perturbations are bounded by Assumption 3.6 and bound (3.64), i.e., we assume

\[
\| f^{(0)}_{j+1} \|_2 \leq \varepsilon_v \left( \| \tilde{w}_{j+1} \|_2 + \sqrt{j} \| \tilde{V}_j \|_2 \| \tilde{h}_{1:j} \|_2 \right), \\
\| f^{(S)}_{j+1} \|_2 \leq \varepsilon_v \| \tilde{\ell}_{j+1} \|_2 \text{ and} \\
\| f^{(M)}_{j+1} \|_2 \leq \varepsilon_v \| \tilde{w}_{j+1} \|_2.
\]
3.4 Backward error analysis

Then each column of $F_k$ can be bounded by

$$
\|F_k e_j\|_2 \leq \|f_{j+1}^{(M)}\|_2 + \|f_{j+1}^{(0)}\|_2 + \|f_{j+1}^{(S)}\|_2
\leq \varepsilon_v \left( \|\tilde{V}_{j+1}\|_2 + 2\|\tilde{w}_{j+1}\|_2 + \sqrt{\tilde{v}_j}\|\tilde{v}_j\|_2 + \|\tilde{h}_{1,j}\|_2 \right).
$$

Inserting (3.24), (3.25), (3.81), (3.88) and using the assumption $\varepsilon_{\text{sum}} < 1$ results in

$$
\|F_k e_j\|_2 \leq \varepsilon_v \left( \frac{1 + \varepsilon_{\text{sum}}}{1 - 2(\varepsilon_v + \varepsilon_s)} \|A\|_2 + \frac{2\|A\|_2}{1 - 2(\varepsilon_v + \varepsilon_s)} + \sqrt{\tilde{v}_j}(1 + \varepsilon_{\text{sum}})\theta_j\|A\|_2 \right)
\leq \varepsilon_v \left( \frac{1 - \delta_j - 6(\varepsilon_v + \varepsilon_s) - \varepsilon_s j}{4 + 2\sqrt{k_j}} \|A\|_2 \right)
\leq \varepsilon_v (4 + 2\sqrt{k_j})\|A\|_2 + \mathcal{O}((\varepsilon_v + \varepsilon_s)^2).
$$

Thus,

$$
\|F_k\|_2 \leq \|F_k\|_F \leq 3\varepsilon_v\|A\|_2 \sum_{j=1}^{k} (4 + 2\sqrt{k_j}) + \mathcal{O}((\varepsilon_v + \varepsilon_s)^2)
\leq \varepsilon_v\|A\|_2 (4k + k^2(k + 1)) + \mathcal{O}((\varepsilon_v + \varepsilon_s)^2)
\leq \varepsilon_v\|A\|_2 k^2(k + 1) + \mathcal{O}((\varepsilon_v + \varepsilon_s)^2).
$$  (3.103)

Since $\tilde{H}_k$ is a Hessenberg matrix and the norm of $F_k$ is small, (3.102) is close to an Arnoldi relation. Consequently, $\tilde{K}_k = \text{ran}(\tilde{V}_k)$ is close to a Krylov subspace of $A$. However, (3.102) fails to be an Arnoldi relation for two reasons:

(i) the presence of $F_k$ and
(ii) the non-orthogonality of $\tilde{V}_k$.

We address the second issue by switching from $\tilde{V}_k$ to $\hat{V}_k = \tilde{V}_k G_k^{-1}$, the known orthonormal basis of $\tilde{K}_k$. Post-multiplication of Equation (3.102) by $G_k^{-1}$ yields

$$
A\hat{V}_k G_k^{-1} = (\hat{V}_k \tilde{H}_k + \hat{v}_{k+1} \hat{h}_{k+1,k}^T \tilde{e}_k^T) G_k^{-1} + F_k G_k^{-1}
= [\hat{V}_k, \hat{v}_{k+1}] G_{k+1}^{-1} G_k^{-1} \begin{bmatrix} \hat{H}_k \\ \hat{h}_{k+1,k} \end{bmatrix} G_k^{-1} + F_k G_k^{-1}.
$$

Exploiting the upper triangular structure of $G_k$ we define a Hessenberg matrix $\hat{H}_k$ and a scalar $\hat{h}_{k+1,k}$ by

$$
\begin{bmatrix} \hat{H}_k \\ \hat{h}_{k+1,k}^T \end{bmatrix} := G_{k+1} \begin{bmatrix} \hat{H}_k \\ \hat{h}_{k+1,k} \end{bmatrix} G_k^{-1}.
$$

72
3.4 Backward error analysis

Introducing further \( \hat{F}_k := F_k G_k^{-1} \) and using that
\[
[\hat{V}_k, \hat{a}_{k+1}] = \bar{V}_{k+1} = \bar{V}_{k+1} G_{k+1}^{-1} = [\bar{V}_k, \bar{a}_{k+1}] G_{k+1}^{-1}
\]
results in
\[
A \hat{V}_k = \bar{V}_k \hat{H}_k + \hat{a}_{k+1} \hat{h}_{k+1, k} e_k^T + \hat{F}_k.
\]  
(3.104)

This would be an Arnoldi relation if the matrix \( \hat{F}_k \) would not be present. In order to keep the structure of (3.101), we interpret \( \hat{F}_k \) as backward error of \( A \). In fact, (3.104) is equivalent to
\[
(A + E_k) \hat{V}_k = \bar{V}_k \hat{H}_k + \hat{a}_{k+1} \hat{h}_{k+1, k} e_k^T,
\]  
(3.105)

whenever \( E_k \in \mathbb{C}^{n \times n} \) fulfills
\[
E_k \hat{V}_k = -\hat{F}_k.
\]  
(3.106)

One possibility to satisfy (3.106) would be to choose \( E_k \) as \( E_k = -\hat{F}_k \hat{V}_k^H \). This choice was already analyzed in, e.g., \([100, 108, 111]\). With \( E_k = -\hat{F}_k \hat{V}_k^H \) we have managed to arrive at a correct Arnoldi relation of the form (3.105). However, in this thesis we investigate Hermitian matrices \( A \) and therefore, choosing \( E_k = -\hat{F}_k \hat{V}_k^H \) is not satisfactory, since this choice is in general not Hermitian.

Unfortunately, the following lemma rules out the existence of a Hermitian matrix \( E_k \) satisfying the condition (3.106).

**Lemma 3.25.** Let \( V \in \mathbb{C}^{n \times k} \) have orthonormal columns and \( F \in \mathbb{C}^{n \times k} \). Then there is an Hermitian \( E \) with \( EV = F \) if and only if \( V^HF \) is Hermitian.

**Proof.** For \( k = 1 \) the proof is contained in [75]. We give it for completeness. Let \( E \) be any matrix such that \( EV = F \). Now, if \( E \) is Hermitian, then so is \( V^HEV = V^HF \). Hence, if \( V^HF \) is not Hermitian, then there is no Hermitian \( E \) with \( EV = F \). On the other hand, let \( V^HF \) be Hermitian, i.e., \( V^HF = F^HV \). Then \( E = FV^H + V^HF - VF^HV^H \) is Hermitian and \( EV = F \). \( \square \)

A solution to this dilemma is to replace the non-Hermitian Hessenberg matrix \( \hat{H}_k \) by an Hermitian matrix \( B_k \). Note that any Hermitian \( B_k \) matrix is feasible. Therefore, we relax the concept of the Arnoldi relation by allowing a non-Hessenberg \( H_k \). Thus, the matrices \( A \in \mathbb{C}^{n \times n}, V_{k+1} = [V_k, v_{k+1}] \in \mathbb{C}^{n \times k+1}, \) and \( H_k \in \mathbb{C}^{k \times k} \) satisfy a so called Krylov relation, introduced by Stewart in [108], if (3.101) holds and \( V_{k+1} \) has orthonormal columns. For this Krylov relation we have the following result.

**Theorem 3.26.** Let \( A \in \mathbb{C}^{n \times n} \) be Hermitian, let \( \hat{V}_{k+1} = [\hat{V}_k, \hat{a}_{k+1}] \in \mathbb{C}^{n \times k+1} \) have orthonormal columns, and suppose that \( \hat{H}_k \in \mathbb{C}^{k \times k}, \hat{h}_{k+1, k} \in \mathbb{C} \) and \( \hat{F}_k \in \mathbb{C}^{n \times k} \) are such that (3.104) holds.

Then for every Hermitian matrix \( B_k \in \mathbb{C}^{k \times k} \) there exists an Hermitian matrix \( E_k \in \mathbb{C}^{n \times n} \) such that
\[
(A + E_k) \hat{V}_k = \hat{V}_k B_k + \hat{a}_{k+1} \hat{h}_{k+1, k} e_k^T
\]  
(3.107)

73
\section{Backward error analysis}

is a Krylov relation. For $* \in \{2, F\}$ \(\|E_k\|_*\) is bounded by

\[
\alpha_*\|(I - P_{\tilde{K}_k})\tilde{F}_k\|_* \leq \|E_k\|_* \leq \|B_k - \tilde{S}_k\|_* + \alpha_*\|(I - P_{\tilde{K}_k})\tilde{F}_k\|_*
\]

\[
\leq \|B_k - \tilde{S}_k\|_* + \alpha_*\|\tilde{F}_k\|_*
\]

(3.108)

with \(\tilde{S}_k := \tilde{\mathcal{V}}_k^H A \tilde{V}_k, P_{\tilde{K}_k} := \tilde{V}_k D^{-1} \tilde{\mathcal{V}}_k^H = \tilde{\mathcal{V}}_k \tilde{\mathcal{V}}_k^H\), and \(\alpha_2 = 1, \alpha_F = \sqrt{2}\). The lower bound holds for any Hermitian \(E_k \in \mathbb{C}^{n \times n}\) for which a \(B_k \in \mathbb{C}^{k \times k}\) exists such that (3.107) holds.

\textbf{Proof.} We prove the lower bound, i.e., \(\alpha_*\|(I - P_{\tilde{K}_k})\tilde{F}_k\|_* \leq \|E_k\|_*\), first. Assume that \(E_k\) and \(B_k\) are such that (3.107) holds and let \(\tilde{V}_\perp \in \mathbb{C}^{n \times n - k}\) be any matrix such that \([\tilde{V}_k, \tilde{V}_\perp]\) is unitary. Then pre-multiplicating (3.107) by \(\tilde{V}_\perp\) gives \(\tilde{\mathcal{V}}_k^H (A + E_k) \tilde{V}_k = 0\), i.e., \(\tilde{\mathcal{V}}_k^H A \tilde{V}_k = -\tilde{\mathcal{V}}_k^H E_k \tilde{V}_k\) and pre-multiplicating (3.104) by \(\tilde{V}_\perp\) gives \(\tilde{\mathcal{V}}_k^H A \tilde{F}_k = \tilde{\mathcal{V}}_k^H \tilde{F}_k\). Together this leads to \(\tilde{\mathcal{V}}_k^H E_k \tilde{V}_k = -\tilde{\mathcal{V}}_k^H \tilde{F}_k\). Consequently, because \(E_k\) is Hermitian, the matrix must be of the form

\[
E_k = [\tilde{V}_k, \tilde{V}_\perp] \begin{bmatrix} E_{11} & -\tilde{F}_k^H \tilde{V}_\perp \\ -\tilde{V}_\perp^H \tilde{F}_k & E_{22} \end{bmatrix} [\tilde{V}_k, \tilde{V}_\perp]^H,
\]

(3.109)

where \(E_{11}, E_{22}\) are still undetermined. Let \(E^{(0)}\) be the matrix that is obtained by setting \(E_{11}\) and \(E_{22}\) in (3.109) to zero. Then

\[
\|E_k\|_* \geq \|E^{(0)}\|_* = \alpha_*\|\tilde{V}_\perp^H \tilde{F}_k\|_* = \alpha_*\|(I - P_{\tilde{K}_k})\tilde{F}_k\|_*,
\]

proves the lower bound on \(\|E_k\|_*\) in (3.108).

To prove the upper bound, i.e., \(\|E_k\|_* \leq \|B_k - \tilde{S}_k\|_* + \|\tilde{F}_k\|_*\), we assume that \(B_k \in \mathbb{C}^{k \times k}\) is Hermitian. Pre-multiplying equation (3.104) with \(\tilde{V}_k^H\) shows that \(\tilde{S}_k, \tilde{H}_k,\) and \(\tilde{F}_k\) are related via

\[
\tilde{S}_k = \tilde{H}_k + \tilde{V}_k^H \tilde{F}_k.
\]

(3.110)

Note that \(A\) and \(B_k\) being Hermitian implies that \(\tilde{S}_k\) and \(E_k\) are Hermitian as well.

The following proof is constructive. We choose \(E_k = E^{(0)} + \tilde{V}_k (B_k - \tilde{S}_k) \tilde{V}_k^H\), then

\[
(A + E_k) \tilde{V}_k = A \tilde{V}_k + E^{(0)} \tilde{V}_k + \tilde{V}_k (B_k - \tilde{S}_k) \tilde{V}_k^H \tilde{V}_k
\]

\[
= (\tilde{V}_k \tilde{H}_k + \tilde{o}_{k+1,k} e_k^T + \tilde{F}_k) - \tilde{V}_\perp \tilde{V}_k^H \tilde{F}_k + \tilde{V}_k (B_k - (\tilde{H}_k + \tilde{V}_k^H \tilde{F}_k))
\]

\[
= \tilde{V}_k B_k + \tilde{o}_{k+1,k} \tilde{H}_k + \tilde{o}_{k+1,k} e_k^T,
\]

where we have used that \(\tilde{V}_k \tilde{V}_k^H + \tilde{V}_\perp \tilde{V}_\perp^H = I\). This proves (3.107). For the norm of \(E_k\) we have

\[
\|E_k\|_* = \|E^{(0)} + \tilde{V}_k (B_k - \tilde{S}_k) \tilde{V}_k^H\|_* \leq \|E^{(0)}\|_* + \|B_k - \tilde{S}_k\|_*
\]

\[
\leq \alpha_*\|\tilde{F}_k\|_* + \|B_k - \tilde{S}_k\|_*
\]

74
which concludes the proof of the upper bound.

It follows from equation (3.107) that \( \tilde{K}_k \) is a Krylov subspace for the Hermitian matrix \( A + E_k \).

**Remark 3.27.** The perturbation \( E_k \) depends on \( k \). Hence, although \( \tilde{K}_j \) is a Krylov subspace of \( A + E_j \), in general it is not a Krylov subspace of \( A + E_k \) for \( j < k \).

So far we have shown that for a Hermitian matrix \( A \) the inexact Arnoldi method yields an exact Krylov relation of a nearby Matrix \( A + E_k \). In the following section we are interested in upper bounds on the norm of \( E_k \) for different choices of \( B_k \).

### 3.4.2 Bounds on the backward error

Theorem 3.26 opens up some freedom in the choice of \( B_k \). This leads us to the question, which matrix should be used in practice?

One criterion could be that for positive definite \( A \) also \( B_k \) should be definite. By left multiplication of (3.107) by \( \hat{V}_k \) we see that \( B_k = \hat{V}_k (A + E_k) \hat{V}_k \), implying that \( B_k \) is guaranteed to be definite, whenever \( A + E_k \) is. Hence, \( \| E_k \| \) should be small.

In view of our bound (3.108), a perfect choice is \( B_k = \hat{S}_k = \hat{V}_k A \hat{V}_k \), as it minimizes this bound on \( \| E_k \|_\ast \). Unfortunately, constructing \( \hat{S}_k \) requires matrix-vector-products with \( A \), which are not possible without perturbations. In the special case when \( f_k^{(M)} = 0 \) for all \( k \), \( \hat{S}_k \) can be computed as \( \hat{S}_k = G_k^{-H} \tilde{S}_k G_k^{-1} \), where \( \tilde{S}_k := \tilde{V}_k A \tilde{V}_k \) can be updated along \( \tilde{V}_k \) and \( \tilde{H}_k \) in Algorithm 3.3 by

\[
\tilde{S}_k = \begin{bmatrix}
\tilde{S}_{k-1} & \tilde{V}_k^{-1} \tilde{w}_{k+1} \\
\tilde{w}_{k+1}^{-1} \tilde{V}_{k-1} & \tilde{V}_k^{-1} \tilde{w}_{k+1}
\end{bmatrix}.
\]

When the matrix-vector-multiplication is inexact, other choices for \( B_k \) have to be found. Considering the two terms in the bound (3.108), we want to find a \( B_k \) that is close to the optimal \( \hat{S}_k \), i.e., \( \| B_k - \hat{S}_k \|_\ast \) should be not much larger than \( \| E_k \|_\ast \). By (3.110) \( \hat{H}_k \) is close to \( \tilde{S}_k \). Since \( \hat{H}_k \) itself is non-Hermitian, we propose to use its Hermitian part, i.e., \( \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) \).

In some situations it might be even beneficial if \( B_k \) is of a tridiagonal structure. Reasons for this restriction may be on the one hand of theoretical nature. Since, in contrast to Remark 3.27, the Krylov subspace corresponding to a tridiagonal \( B_k \) is nested, i.e., \( \tilde{K}_j \) is a Krylov subspace for \( A + E_k \) for \( j \leq k \). On the other hand a reason for this restriction could be just of practical nature, since the computation of eigenvalues for tridiagonal matrices is faster and more accurate than for general Hermitian matrices [31].

To cover all these situations we will analyze in the following the choices \( B_k = T_{\tilde{S}_k} \) and \( B_k = T_{\tilde{H}_k} \), where \( T_H \) denotes the tridiagonal part of the Hermitian part.
3.4 Backward error analysis

Table 3.2: Coefficients of the error bound (3.112) for different choices of $B_k$

<table>
<thead>
<tr>
<th>$B_k$</th>
<th>$\alpha_{2,B_k}$</th>
<th>$\beta_{2,B_k}$</th>
<th>$\alpha_{F,B_k}$</th>
<th>$\beta_{F,B_k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{S}_k$</td>
<td>1</td>
<td>0</td>
<td>$\sqrt{2}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{\tau} (\hat{H}_k + \hat{H}_k^H)$</td>
<td>2</td>
<td>0</td>
<td>$1 + \sqrt{2}$</td>
<td>0</td>
</tr>
<tr>
<td>$\frac{1}{\tau} (\hat{H}_k + \hat{H}_k^H)$</td>
<td>2</td>
<td>1</td>
<td>$1 + \sqrt{2}$</td>
<td>1</td>
</tr>
<tr>
<td>$T_{\hat{S}_k}$</td>
<td>$1 + \sqrt{2}k$</td>
<td>0</td>
<td>$2 \sqrt{2}$</td>
<td>0</td>
</tr>
<tr>
<td>$T_{\hat{H}_k}$</td>
<td>$2 + \sqrt{k}$</td>
<td>0</td>
<td>$2 + \sqrt{2}$</td>
<td>0</td>
</tr>
<tr>
<td>$T_{\hat{H}_k}$</td>
<td>$2 + \sqrt{k}$</td>
<td>1 + $\sqrt{k}$</td>
<td>$2 + \sqrt{2}$</td>
<td>2</td>
</tr>
</tbody>
</table>

of $H$, i.e.,

$$(T_H)_{i,j} := \begin{cases} \frac{1}{2} (h_{i,j} + \bar{h}_{i,j})/2, & \text{for } |i - j| \leq 1 \\ 0, & \text{otherwise.} \end{cases} (3.111)$$

Finally, we will also look at $B_k = \frac{1}{\tau} (\hat{H}_k + \hat{H}_k^H)$ and $B_k = T_{\hat{H}_k}$.

The reasoning behind these choices for $B_k$ is that without perturbations, Algorithm 3.3 would reduce to the Lanczos method and $\hat{H}_k$ would be Hermitian and tridiagonal. With perturbations, $\hat{H}_k$ is neither, but should still be close. Thus, its Hermitian or Hermitian tridiagonal parts should be good approximations of the original tridiagonal matrix. In particular, the choice $B_k = T_{\hat{H}_k}$ is used in ARPACK [73] for Hermitian matrices $A$.

In the following Corollary we state bounds on the norm of $E_k$ corresponding to the different choices if $B_k$.

**Corollary 3.28.** Let $A \in \mathbb{C}^{n \times n}$ be Hermitian and let $\hat{V}_{k+1} = [\hat{V}_k, \hat{v}_{k+1}] \in \mathbb{C}^{n \times k+1}$ have orthonormal columns. Suppose that $\hat{H}_k \in \mathbb{C}^{k \times k}$ is Hessenberg, $\hat{h}_{k+1,k} \in \mathbb{C}$ and $\hat{F}_k \in \mathbb{C}^{n \times k}$ is such that (3.104) holds. Let $\hat{S}_k := \hat{V}_k^H A \hat{V}_k$ and $\hat{H}_k := [\hat{H}_k^T, \hat{h}_{k+1,k} \hat{v}_k]$. Let $G_k \in \mathbb{C}^{k \times k}$ and $G_{k+1} \in \mathbb{C}^{k+1 \times k}$ be invertible upper triangular matrices such that $\|G_k - I_k\|_2 \leq \epsilon_k$ and $\|G_{k+1} - I_{k+1}\|_2 \leq \epsilon_{k+1} < 1$. Moreover, define $\tilde{H}_k := [I_{k+1}, 0] G_{k+1}^{-1} \hat{H}_k G_k$ and $S_k, T_{\hat{S}_k}, T_{\hat{H}_k}$ as in (3.111).

Then, for $B_k \in \{S_k, T_{\hat{S}_k}, T_{\hat{H}_k}/2, (\hat{H}_k + \hat{H}_k^H), T_{\hat{H}_k}/2, (\hat{H}_k + \hat{H}_k^H), T_{\hat{H}_k} \}$ there exists an Hermitian matrix $E_k \in \mathbb{C}^{n \times n}$ such that (3.107) is a Krylov relation and for $* \in \{2, F\}$ $\|E_k\|_*$ is bounded by

$$\|E_k\|_* \leq \alpha_{*,B_k} \|\hat{F}_k\|_* + \beta_{*,B_k} \|\hat{H}_k\|_* \frac{\epsilon_k + \epsilon_{k+1}}{1 - \epsilon_{k+1}} (3.112)$$

with constants $\alpha_{*,B_k}, \beta_{*,B_k}$ given in Table 3.2.

**Proof.** The error bounds of the different choices of $B_k$ are proved separately.

- Case $B_k = \hat{S}_k$:
  This case follows directly from (3.108).
Case $B_k = \frac{1}{2}(\hat{H}_k + \hat{H}_k^H)$:

Using relation (3.110) and that $\hat{S}_k$ is Hermitian, we have

$$\frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k = \frac{1}{2}(\hat{S}_k - \hat{V}_k^H \hat{F}_k + \hat{S}_k^H - (\hat{V}_k^H \hat{F}_k)^H) - \hat{S}_k$$

implying that

$$\|\frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k\|_* = \|\frac{1}{2}\hat{V}_k^H \hat{F}_k + \hat{F}_k^H \hat{V}_k\|_* \leq \|\hat{V}_k^H \hat{F}_k\|_* \leq \|P_{\hat{K}_k} \hat{F}\|_*. \tag{3.113}$$

Hence by (3.108), the backward error corresponding to $B_k = \frac{1}{2}(\hat{H}_k + \hat{H}_k^H)$ is bounded by

$$\|E_k\|_* \leq \|\frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k\|_* + \alpha_* \|(I - P_{\hat{K}_k}) \hat{F}_k\|_*$$

$$\leq \|P_{\hat{K}_k} \hat{F}_k\|_* + \alpha_* \|(I - P_{\hat{K}_k}) \hat{F}_k\|_*$$

$$\leq \alpha_* \|\frac{1}{2}(\hat{H}_k + \hat{H}_k^H)\| S_k_2$$

with $\alpha_*$ as in Theorem 3.26.

Case $B_k = T_{\hat{H}_k}$:

For $B_k = T_{\hat{H}_k}$ we consider a splitting $\hat{H}_k = \hat{T} + \hat{U}$, where $\hat{T}$ is the tridiagonal part of $\hat{H}_k$ and $\hat{U} = \hat{H}_k - \hat{T}$ is strictly upper triangular. Then

$$T_{\hat{H}_k} = \frac{1}{2}(\hat{T} + \hat{T}^H) = \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \frac{1}{2}(\hat{U} + \hat{U}^H). \tag{3.114}$$

From (3.110) we have

$$\hat{H}_k - \hat{H}_k^H = \hat{S}_k - \hat{V}_k^H \hat{F}_k - (\hat{S}_k^H - \hat{F}_k^H \hat{V}_k) = \hat{F}_k^H \hat{V}_k - \hat{V}_k^H \hat{F}_k. \tag{3.115}$$

Hence, we have

$$\|\hat{U} + \hat{U}^H\|_* \leq \|\hat{U} + \hat{U}^H\|_F = \|\hat{U} - \hat{U}^H\|_F$$

$$\leq \|\hat{H}_k - \hat{H}_k^H\|_F$$

$$= \|\hat{F}_k^H \hat{V}_k - \hat{V}_k^H \hat{F}_k\|_F \leq 2\|\hat{V}_k^H \hat{F}_k\|_F.$$

Thus, together with (3.114) and (3.113) we have

$$\|T_{\hat{H}_k} - \hat{S}_k\|_* \leq \|\frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k\|_* + \frac{1}{2}\|\hat{U} + \hat{U}^H\|_*$$

$$\leq \|P_{\hat{K}_k} \hat{F}_k\|_* + \|P_{\hat{K}_k} \hat{F}_k\|_F.$$

Hence, with (3.108), the backward error $E_k$ corresponding to $B_k = T_{\hat{H}_k}$ is
3.4 Backward error analysis

bounded by

\[
\| E_k \|_* \leq \| T_{\tilde{F}_k} - \tilde{S}_k \|_* + \alpha_* \| (I - P_{\tilde{K}_k}) \tilde{F}_k \|_* \\
\leq \| P_{\tilde{K}_k} \tilde{F}_k \|_* + \| P_{\tilde{K}_k} \tilde{F}_k \|_F + \alpha_* \| (I - P_{\tilde{K}_k}) \tilde{F}_k \|_* \\
\leq \alpha_* T_{\tilde{F}_k} \| \tilde{F}_k \|_*. 
\]

- **Case** \( B_k = T_{\tilde{S}_k} \):
  For \( B_k = T_{\tilde{S}_k} \) we consider a splitting \( \tilde{S}_k = T_{\tilde{S}_k} + U + U^H \), where \( U \) denotes the upper triangular part of \( \tilde{S}_k - T_{\tilde{S}_k} \) (note that the main and the first super diagonals of \( U \) are zero). Since \( \hat{H}_k \) is a Hessenberg matrix it follows from (3.110) that \( \hat{S}_k \) and \( \hat{V}_k^H \hat{F}_k \) coincide below the first subdiagonal. Hence, \( \| U^H \|_F \leq \| \hat{V}_k^H \hat{F}_k \|_F \) and it follows that

\[
\| \tilde{S}_k - T_{\tilde{S}_k} \|_* \leq \| \tilde{S}_k - T_{\tilde{S}_k} \|_F = \| U + U^H \|_F \\
= \sqrt{2} \| U^H \|_F \leq \sqrt{2} \| \hat{V}_k^H \hat{F}_k \|_F = \sqrt{2} \| P_{\tilde{K}_k} \tilde{F}_k \|_F.
\]

Thus, using (3.108), the backward error \( E_k \) corresponding to \( B_k = T_{\tilde{S}_k} \) is bounded by

\[
\| E_k \|_* \leq \| \tilde{S}_k - T_{\tilde{S}_k} \|_* + \alpha_* \| (I - P_{\tilde{K}_k}) \tilde{F}_k \|_* \\
\leq \sqrt{2} \| P_{\tilde{K}_k} \tilde{F}_k \|_F + \alpha_* \| (I - P_{\tilde{K}_k}) \tilde{F}_k \|_* \\
\leq \alpha_* T_{\tilde{F}_k} \| \tilde{F}_k \|_*.
\]

- **Case** \( B_k = \frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H) \):
  Define \( \tilde{H}_k := G_{k+1}^{-1} \tilde{H}_k G_k \) and note that \( \tilde{H}_k \) consists of the top \( k \) rows of \( \tilde{H}_k \). Then we have \( \tilde{H}_k = (G_{k+1}^{-1} - I_{k+1} + I_{k+1}) \tilde{H}_k (G_k - I_k + I_k) \) implying that

\[
\tilde{H}_k = \tilde{H}_k + (G_{k+1}^{-1} - I_{k+1}) \tilde{H}_k + \tilde{H}_k (G_k - I_k) + (G_{k+1}^{-1} - I_{k+1}) \tilde{H}_k (G_k - I_k).
\]

With \( \| G_k - I_k \|_2 = \zeta_k \) and using Lemma 2.4 iii) with \( I - P = G_{k+1} \) we get \( \| G_{k+1}^{-1} - I_{k+1} \|_2 \leq \frac{\zeta_{k+1}}{1 - \zeta_{k+1}} \), thus we have

\[
\| \tilde{H}_k - \hat{H}_k \|_* \leq \| \tilde{H}_k - \hat{H}_k \|_* \\
\leq \| G_{k+1}^{-1} - I_{k+1} \|_2 \| \tilde{H}_k \|_* + \| \hat{H}_k \|_* \| G_k - I_k \|_2 \\
\leq \| \hat{H}_k \|_* \left( \frac{\zeta_{k+1}}{1 - \zeta_{k+1}} + \zeta_k + \frac{\zeta_k \zeta_{k+1}}{1 - \zeta_{k+1}} \right)
\]

78
3.4 Backward error analysis

\[ = \| \hat{H}_k \| \frac{\hat{c}_k + \hat{c}_{k+1}}{1 - \hat{c}_{k+1}}. \]  

(3.116)

Consequently, with (3.113) and

\[ \| \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) \|_* \leq \| \hat{H}_k - \hat{H}_k \|_*, \]

we have

\[ \| \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k \|_* \leq \| \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) \|_* + \| \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \hat{S}_k \|_* \]

\[ \leq \| \hat{H}_k \|_* \frac{\hat{c}_k + \hat{c}_{k+1}}{1 - \hat{c}_{k+1}} + \| P_{\hat{c}_k} S_k \|_. \]

(3.117)

Thus, using (3.108), the backward error \( E_k \) corresponding to \( B_k = \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) \) is bounded by

\[ \| E_k \|_* \leq \| \hat{S}_k - \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) \|_* + \alpha_* \| (I - P_{\hat{c}_k}) \hat{S}_k \|_* \]

\[ \leq \| \hat{H}_k \|_* \frac{\hat{c}_k + \hat{c}_{k+1}}{1 - \hat{c}_{k+1}} + \| P_{\hat{c}_k} S_k \|_* + \alpha_* \| (I - P_{\hat{c}_k}) \hat{S}_k \|_* \]

\[ \leq \| \hat{H}_k \|_* \frac{\hat{c}_k + \hat{c}_{k+1}}{1 - \hat{c}_{k+1}} + \alpha_* \| \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) \| \hat{S}_k \|_* \].

- **Case \( B_k = T_{\hat{H}_k} \):**
  Using a splitting \( \hat{H}_k = \hat{T} + \hat{U} \), where \( \hat{T} \) is the tridiagonal part of \( \hat{H}_k \) and \( \hat{U} = \hat{H}_k - \hat{T} \) is strictly upper triangular, leads to

\[ T_{\hat{H}_k} = \frac{1}{2}(\hat{T} + \hat{T}^H) = \frac{1}{2}(\hat{H}_k + \hat{H}_k^H) - \frac{1}{2}(\hat{U} + \hat{U}^H). \]  

(3.118)

Using (3.115) and (3.116) we have

\[ \| \hat{H}_k - \hat{H}_k^H \|_* \leq \| \hat{H}_k - \hat{H}_k^H \|_* + 2\| \hat{H}_k - \hat{H}_k \|_* \]

\[ \leq \| \hat{H}_k \hat{H}_k^H \|_* + 2\| \hat{H}_k \hat{H}_k^H \|_* + \| \hat{H}_k \|_* \frac{\hat{c}_k + \hat{c}_{k+1}}{1 - \hat{c}_{k+1}}. \]

Hence, with \( \gamma_2 := \sqrt{k}, \gamma_F := 1 \) it follows that

\[ \| \hat{U} + \hat{U}^H \|_* \leq \| \hat{U} + \hat{U}^H \|_F = \| \hat{U} - \hat{U}^H \|_F \]

\[ \leq \| \hat{H}_k - \hat{H}_k^H \|_F \leq \gamma_* \| \hat{H}_k - \hat{H}_k^H \|_* \]

\[ \leq \gamma_* \left( \| \hat{H}_k \hat{H}_k^H \|_* + 2\| \hat{H}_k \|_* \frac{\hat{c}_k + \hat{c}_{k+1}}{1 - \hat{c}_{k+1}} \right) \]

79
3.4 Backward error analysis

\[ \leq 2\gamma_s \|P_{\tilde{K}_k} \hat{F}_k\|_* + 2\gamma_s \|\tilde{H}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}}. \]

(3.119)

Consequently, with (3.118), (3.117), and (3.119)

\[ \|T_{\tilde{H}_k} - \hat{S}_k\|_* \leq \|\frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H) - \hat{S}_k\|_* + \frac{1}{2}\|U + \tilde{U}\|_* \]

\[ \leq \|\tilde{H}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}} + (1 + \gamma_s)\|P_{\tilde{K}_k} \hat{F}_k\|_* + \gamma_s \|\tilde{H}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}} \]

\[ = (1 + \gamma_s)\|P_{\tilde{K}_k} \hat{F}_k\|_* + (1 + \gamma_s)\|\tilde{H}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}}. \]

Hence, using (3.108), the backward error corresponding to \(B_k = T_{\tilde{H}_k}\) is bounded by

\[ \|E_k\|_* \leq \|T_{\tilde{H}_k} - \hat{S}_k\|_* + \alpha_s \|(I - P_{\tilde{K}_k}) \hat{F}_k\|_* \]

\[ \leq (1 + \gamma_s)\|P_{\tilde{K}_k} \hat{F}_k\|_* + (1 + \gamma_s)\|\tilde{H}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}} + \alpha_s \|(I - P_{\tilde{K}_k}) \hat{F}_k\|_* \]

\[ \leq (1 + \gamma_s + \alpha_s)\|\hat{F}_k\|_* + (1 + \gamma_s)\|\tilde{H}_k\|_* \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}} \]

concluding the proof.

As expected, the bound is best for \(\hat{S}_k\), but it is not much worse for \(\frac{1}{2}(\tilde{H}_k + \tilde{H}_k^H)\) making it a good candidate when \(\hat{S}_k\) is not available. All the bounds for tridiagonal \(B_k\) involve a \(\sqrt{k}\) factor. This stems from the use of the Frobenius norm and is likely to be an overestimation in our case. Moreover, it is noteworthy that for those \(B_k\) involving \(\tilde{H}_k\) the bounds contain a term \(\|\tilde{H}_k\|_2\|G_k - I\|_2\).

**Remark 3.29.** By combining the results of Corollary 3.28 with those of Corollary 3.17 and using (3.103), we can show that \(E_k\) can be bounded in terms of \(\|A\|_2(e_v + \varepsilon_s)\).

For example, for \(B_k = T_{\tilde{f}_k}\) we obtain

\[ \|E_k\|_F \leq (2 + \sqrt{2})\|\hat{F}_k\|_F \leq (2 + \sqrt{2})\|\hat{F}_k\|_F(1 + \mathcal{O}(e_v + \varepsilon_s)) \]

\[ \leq (2 + \sqrt{2})k^2(k + 1)\|A\|_2\varepsilon_v + \mathcal{O}((e_v + \varepsilon_s)^2), \]

where we have used that \(\|G_k^{-1}\|_2 = 1 + \mathcal{O}((e_v + \varepsilon_s))\) because of Lemma 2.4 and Corollary 3.17.

For \(B_k = T_{\tilde{f}_k}\) we obtain

\[ \|E_k\|_F \leq (2 + \sqrt{2})\|\hat{F}_k\|_F + 2\|\tilde{H}_k\|_F \frac{\zeta_k + \zeta_{k+1}}{1 - \zeta_{k+1}} \]
3.5 Numerical results

\[
\leq (2 + \sqrt{2}) \| A \|_F (1 + \mathcal{O}(\varepsilon_v + \varepsilon_s)) \\
+ 2 \| H_k \|_2 \left( 2 \sqrt{2} (\varepsilon_v + \varepsilon_s) k^2 (k^2 + 1) (1 + \rho_{\text{max},k+1}) \varepsilon + \mathcal{O}(\varepsilon_v + \varepsilon_s)^2 \right) \\
\leq (2 + \sqrt{2}) \| A \|_2 \varepsilon_v k^2 (k + 1) \\
+ 4 \sqrt{2} \| A \|_2 (\varepsilon_v + \varepsilon_s) k^2 (k^2 + 1) (1 + \rho_{\text{max},k+1}) \varepsilon + \mathcal{O}(\varepsilon_v + \varepsilon_s)^2 \\
\leq 4 k^2 \left( (k + 1) + (k^2 + 1)(1 + \rho_{\text{max},k+1}) \right) \| A \|_2 (\varepsilon_v + \varepsilon_s) + \mathcal{O}(\varepsilon_v + \varepsilon_s)^2,
\]

where we have used that \( \| H_k \|_2 \leq \| A + E_k \|_2 = \| A \|_2 (1 + \mathcal{O}(\varepsilon_v + \varepsilon_s)) \) and \( \| G_k^{-1} - I \|_2 = 2 \sqrt{2} (k - 1)^2 k^2 (1 + \rho_{\text{max},k})(\varepsilon_v + \varepsilon_s) + \mathcal{O}((\varepsilon_v + \varepsilon_s)^2) \), which follows from Lemma 2.4 iv and the coarse bound (3.73b).

Comparing these results suggests that \( T_{\hat{H}_k} \) should be preferred over \( T_{\hat{H}_k} \). Experiment 4 below shows that this is indeed the case when \( (1 + \rho_{\text{max},k+1}) \) is large. Otherwise these two choices of \( B_k \) actually perform rather similar.

3.5 Numerical results

In this section we present numerical experiments to illustrate the theoretical results obtained in the Sections 3.1–3.4. The numerical experiments consists of two parts. In the first part, more precisely in the first three experiments, we quantify the \( O \)-measure, i.e., the level of orthogonality, achieved with ComGS compared to the classical and modified Gram Schmidt schemes with and without reorthogonalization. The second part, i.e., Experiment 4, analyzes the norm of the backward error matrices \( E_k \) that emerges from the translation of the inexact Krylov relation (3.107) into a exact Krylov relation of the perturbed matrix \( A + E_k \) for given \( \hat{V}_{k+1} \) and several choices of Hermitian \( B_k \). We quantify the error by the Frobenius norm of \( E_k \) and verify its bound obtained from Corollary 3.28.

To evaluate the \( O \)-measure we use the different orthogonalization schemes ComGS, CGS, and MGS in an inexact QR decomposition, as stated in Algorithm 3.4 below. We have chosen the inexact QR decomposition because it only consists of orthonormalization steps. Thus any deviation from orthogonality in the numerical results is caused by the used orthonormalization scheme. So far we just introduced the exact CGS and MGS method. In order to compare the standard methods with the ComGS method in an inexact setting, we state here for completeness what we mean by inexact CGS and MGS with and without reorthogonalization, see Algorithms 3.5–3.8. The basic structure of all Gram Schmidt variants can be summarized as follows. The input values

\[ w_k \in \mathbb{C}^n, \tilde{Q}_{k-1} \in \mathbb{C}^{n \times k-1}, \varepsilon_s \geq 0, \text{ and } \varepsilon_v \geq 0 \]

are transformed into the output values

\[ \tilde{q}_k \in \mathbb{C}^n, \tilde{r}_{1:k,k} \in \mathbb{C}^k \text{ such that } \tilde{Q}_{k-1}^H \tilde{q}_k \approx 0 \text{ and } w_k \approx [\tilde{Q}_{k-1}, \tilde{q}_k][\tilde{r}_{1:k,k}]. \]
As before mentioned, all vector operations are inexact and the inexactness is modeled by adding a perturbation vector \( f^{(s)} \) of small norm. The corresponding bounds on the perturbations follow from the error model given in (3.5) and (3.6).
3.5 Numerical results

<table>
<thead>
<tr>
<th></th>
<th>CGS</th>
<th>CGSre</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_k^{(S)} )</td>
<td>( \varepsilon_v | \tilde{e}_k |_2 )</td>
<td>( \varepsilon_v | \tilde{e}^{(1)}_k |_2 )</td>
</tr>
<tr>
<td>( f_k^{(N)} )</td>
<td>( \varepsilon_s 2 | \tilde{e}_k |_2 )</td>
<td>( \varepsilon_s 2 | \tilde{e}^{(1)}_k |_2 )</td>
</tr>
<tr>
<td>( f_k^{(P)} )</td>
<td>( \varepsilon_s | \tilde{Q}_{k-1} |_2 | w_k |_2 )</td>
<td>( \varepsilon_s | \tilde{Q}_{k-1} |_2 | w_k |_2 )</td>
</tr>
<tr>
<td>( f_k^{(P1)} )</td>
<td>( \varepsilon_s | \tilde{Q}_{k-1} |_2 | \tilde{e}^{(0)}_k |_2 )</td>
<td>( \varepsilon_s | \tilde{Q}_{k-1} |_2 | \tilde{e}^{(1)}_k |_2 )</td>
</tr>
</tbody>
</table>

Table 3.3: Bounds on the norm of the perturbation of CGS and CGSre

<table>
<thead>
<tr>
<th></th>
<th>MGS</th>
<th>MGSre</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_k^{(S)} )</td>
<td>( \varepsilon_v | \tilde{e}_k |_2 )</td>
<td>( \varepsilon_v | \tilde{e}^{(1)}_k |_2 )</td>
</tr>
<tr>
<td>( f_k^{(N)} )</td>
<td>( \varepsilon_s 2 | \tilde{e}_k |_2 )</td>
<td>( \varepsilon_s 2 | \tilde{e}^{(1)}_k |_2 )</td>
</tr>
<tr>
<td>( f_k^{(P,i)} )</td>
<td>( \varepsilon_s | q_i |_2 | \tilde{e}^{(i-1)}_k |_2 )</td>
<td>( \varepsilon_s | q_i |_2 | \tilde{e}^{(i-1)}_k |_2 )</td>
</tr>
<tr>
<td>( f_k^{(P1,i)} )</td>
<td>( \varepsilon_s | q_i |_2 | \tilde{e}^{(k+i-2)}_k |_2 )</td>
<td>( \varepsilon_s | q_i |_2 | \tilde{e}^{(k+i-2)}_k |_2 )</td>
</tr>
<tr>
<td>( f_k^{(0,i)} )</td>
<td>( \varepsilon_v \left( | e^{(i-1)}_k |_2 + | q_i |<em>2 | \tilde{r}</em>{i,k} |_2 \right) )</td>
<td>( \varepsilon_v \left( | e^{(i-1)}_k |_2 + | q_i |_2 | s^{(0)}_i |_2 \right) )</td>
</tr>
<tr>
<td>( f_k^{(1,i)} )</td>
<td>( \varepsilon_v \left( | e^{(k+i-2)}_k |_2 + | q_i |_2 | s^{(1)}_i |_2 \right) )</td>
<td>( \varepsilon_v \left( | e^{(k+i-2)}_k |_2 + | q_i |_2 | s^{(1)}_i |_2 \right) )</td>
</tr>
</tbody>
</table>

Table 3.4: Bounds on the norm of the perturbation of MGS and MGSre

We summarize the assumed error bounds for the CGS and the MGS method with and without reorthogonalization in the Tables 3.3 and 3.4. In Table 3.4 the perturbations \( f_k^{(0,i)} \), \( f_k^{(1,i)} \), \( f_k^{(P,i)} \), and \( f_k^{(P1,i)} \) are bounded for each step, i.e., for \( i = 1, \ldots, k \).

To analyze the robustness of the ComGS method we compare the ComGS method with the standard CGS and MGS method. To be specific, we use Algorithm 3.4 to compute an inexact QR decomposition column by column, i.e., at the \( k \)-th step we compute \( \tilde{Q}_k \tilde{R}_k = [\tilde{q}_1, \ldots, \tilde{q}_k] [\tilde{r}^{(k)}_{ij} \}_{i,j=1}^k \) of the first \( k \) columns of \( A \). Thus, we obtain a sequence of nested inexact QR decompositions. To illustrate how the quality evolves with the number of processed columns \( k \) we present two different measures of the loss of orthogonality of the basis \( \tilde{Q}_k \) computed by the different Gram Schmidt variants. In particular, we illustrate the vO-measure, i.e., the distance to orthogonality of the computed vector to the previous basis.
3.5 Numerical results

![Graphs illustrating numerical results](image)

Figure 3.1: Experiment 1: Loss of orthogonality illustrated by the vO-measure (left) and the O-measure (right) in the inexact QR decomposition using the Gram Schmidt variants: CGS, ComGS, MGS with and without reorthogonalization vectors and the O-measure, i.e., the distance to orthogonality of the whole cross product matrix for increasing $k$.

3.5.1 Vandermonde matrix

*Our first test matrix $A$ is a Vandermonde matrix, which is elementwise defined by*

$$a_{ij} = (j/m)^{i-1}$$

*for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, see, e.g., [48, Section 20.7].*

Let $A_k$ be the matrix consisting of the leading $k$ columns of $A$. In this case the condition number of $A_k$ grows rapidly with $k$, e.g., already for $k = 9$ the condition number reaches $\kappa(A_9) \approx 10^{16}$. Hence, for sufficiently large $k$ our test matrix is ill-conditioned.

**Experiment 1**  In our first experiment we choose a Vandermonde matrix with $n = 300$, $m = 180$ and specify our error model by $\varepsilon_v = 10^{-6}$ and $\varepsilon_s = 10^{-12}$.

Figure 3.1 illustrates how the vO-measure (left figure) and the O-measure (right figure) of the ComGS, CGS, and MGS method evolve with increasing $k$. We see that both measures behave similar. In particular, it is apparent that, as expected, the CGS method performs worst. Its O-measure is larger than one after just three processed columns, thus the orthogonality of the corresponding basis is lost. Perhaps somewhat more surprising is that for all other variants without reorthogonalization the O-measure is also larger than one already after five pro-
3.5 Numerical results

Figure 3.2: Experiment 2: Loss of orthogonality illustrated by the vO-measure (left) and the O-measure (right) in the inexact QR decomposition using the Gram Schmidt variants with reorthogonalization: CGSre, ComGSre, MGSre.

cessed columns. On the other hand, for every variant with reorthogonalization the O-measure is on the level of $(\epsilon_v + \epsilon_s)$ until the first 100 columns, afterwards CGSre loses orthogonality as well. However, the O-measure of ComGSre and MGSre stays on the order $(\epsilon_v + \epsilon_s)$ up to the last column of $A$.

The advantage of ComGS(re) compared to the other methods is the availability of the cross product matrix $D_k \approx \bar{Q}_k^H\bar{Q}_k$ and its corresponding Cholesky factor $G_k$, which implicitly provides a nearly orthonormal basis $\hat{Q}_k = \bar{Q}_k G_k^{-1}$. This is fact is shown by our next experiment.

**Experiment 2** Here we assume a mixed precision setting, which means that the computations are performed in double precision and the result is stored in single precision. The used test matrix is a Vandermonde matrix specified by $n = 100$ and $m = 50$.

In Figure 3.2 only the Gram Schmidt variants with reorthogonalization are considered and all variants without reorthogonalization are omitted because they lose orthogonality already after three steps. In the right figure of Figure 3.2 we see that with reorthogonalization the O-measure of the CGS method is larger than one after seven steps and the O-measure of the MGS method exceeds one after 12 steps. Also for the ComGS method the O-measure starts to increase but after 17 steps it tends to level of $10^{-2}$. Thus, the cross product matrix $D_k = \bar{Q}_k^H \bar{Q}_k$ is still positive definite and we can use the available Cholesky factor of $D_k$ to construct the nearly orthogonal basis $\hat{Q}_k = \bar{Q}_k G_k^{-1}$, introduced in the Remark 3.5. This leads us to the question whether we could use the same strategy to improve
3.5 Numerical results

the bases resulting from the CGSre or MGSre method, i.e., could we use the inverse of the Cholesky factor of the corresponding cross product matrix $\tilde{Q}_k^H \tilde{Q}_k$ to obtain a better orthogonal basis? But, caused by the perturbations in the CGS and MGS methods, the cross product matrices $\tilde{Q}_k^H \tilde{Q}_k$ are not positive definite anymore. Hence, the computation of the corresponding Cholesky factor $G_k$ is not possible and $\tilde{Q}_k$ is not available for MGS and CGS method. In other words, even without a second step of reorthogonalization ComGS gives as good results as CGS and MGS with two steps of reorthogonalization. Thus, in this case the ComGS method avoids an additional step of reorthogonalization.

Note that the left figure of Figure 3.2 displays that for ComGS and MGS the vO-measure is not recursive, i.e., even if $\tilde{q}_k$ is far from being orthogonal to $\tilde{Q}_{k-1}$, this does not need to be the case for the next iterate $\tilde{q}_{k+1}$ with respect to $\tilde{Q}_k$. Furthermore, the left figure of Figure 3.2 illustrates that the CGS does not have this fortunate behavior, i.e., it does not recover ones orthogonality is lost.

The third experiment is chosen to verify the result of Section 3.3.3.

3.5.2 Upper bidiagonal matrix

In this section the test matrix $A \in \mathbb{R}^{k \times k}$ is an upper bidiagonal matrix with ones on the diagonal and twos on the upper off diagonal.

**Experiment 3** In this experiment we analyze the influence of the condition number of the test matrix on the orthogonality measure for the different orthogonalization schemes. The assumed error model is defined by $\epsilon_v = 10^{-6}$ and $\epsilon_s = 10^{-12}$.

Considering the bounds (3.98) and (3.99) the loss of orthogonality in the CGS and the MGS method depends on the condition number of $A$. The condition number of the upper bidiagonal test matrix increases as $k$ evolves, more precisely already for $k = 40$ the condition number $A$ is at the level $10^{12}$. However, according to (3.71a) for the ComGS method the quantity $\rho_{\text{max}}$ is responsible for the loss of orthogonality and this quantity equals to two in this experiment for all $k$.

This behavior is also illustrated in Figure 3.3, where both measures show that the ComGS method keeps the orthogonality of $\tilde{Q}_k$ much better then CGS and MGS method. More precisely, in the right figure of Figure 3.3 the O-measure of the ComGS method tends to $10^{-5}$. In contrast, caused by the large condition number of $A$, the O-measure of the CGS and MGS method exceed one already after 20 steps. In other words, it would be necessary to perform one additional step of reorthogonalization for the CGS and MGS method to obtain the same level of orthogonality as the ComGS method.

Further, we have checked in the previous three experiments that $\| A_k - \tilde{Q}_k \tilde{R}_k \|_F$ stays on the order $(\epsilon_v + \epsilon_s)$ for every Gram Schmidt variant.
3.5 Numerical results

Figure 3.3: Experiment 3: Loss of orthogonality illustrated by the vO-measure (left) and O-measure (right) in the inexact QR decomposition using the Gram Schmidt variants: CGS, ComGS, MGS

Finally, in our fourth experiment we aim to verify bounds given in Corollary 3.17. We focus on the tridiagonal matrices $T_{k}$ and $T_{\hat{k}}$.

### 3.5.3 Diagonal matrix

*In this section the test matrix $A \in \mathbb{R}^{n \times n}$ built by the MATLAB command*

$$A = \text{diag}([10, 9, 8, 7, 0.1+0.9\times\text{rand}(1,n)]).$$

**Experiment 4** In this experiment we applied the inexact Arnoldi method, see Algorithm 3.3, to a diagonal matrix which has four large eigenvalues at 7, 8, 9, 10 and the remaining eigenvalues are between 0.1 and 1. We chose $n = 10^5$ and $\varepsilon_v = 10^{-9}$, $\varepsilon_s = 10^{-16}$. We perform 10 steps of the Algorithm 3.3 with

$$v_1 = [\text{randn}(4,1);\text{zeros}(n-4,1)]$$

that is normalized to unit norm, i.e., $\tilde{v}_1$ is nonzero in the first four components only. Note that a diagonal matrix represents the general case, since Arnoldi’s method is invariant under unitary similarity transformations of $A$, see, e.g., [13, 308f.].

By construction, the vector $v_1$ is in the invariant subspace of $A$ which corresponds to the four largest eigenvalues of $A$. Thus, the exact Arnoldi method will experience a lucky breakdown after four iterations, whereas the inexact Arnoldi method will experience a near breakdown at that point. A near breakdown is described by a near singularity in the generated basis, for more details, see [18].

87
3.5 Numerical results

Figure 3.4 shows the results for the choices $B_k = T_{\tilde{H}}$ and $B_k = T_{\hat{H}}$ and their respective bounds from (3.112) as they evolve with increasing $k$.

Since the matrix $E_k$ is never actually formed, we specify in the following how its Frobenius norm can be obtained. From (3.109) we have $\|E_k\|_F = (2\|E_k\tilde{V}_k\|_F^2 + \|B_k - S_k\|_F^2)^{1/2}$, where the term $\|E_k\tilde{V}_k\|_F$ can be evaluated by using (3.107), i.e.,

$$E_k\tilde{V}_k = A\tilde{V}_k - \hat{V}_kB_k - \tilde{\delta}_k + 1 e_k^T h_{k+1,k}.$$

![Figure 3.4](image-url)

Figure 3.4: Experiment 4: Backward error of the inexact Arnoldi algorithm and its bounds without (left) and with (right) reorthogonalization

Again, the results illustrated in Figure 3.4 depend on whether reorthogonalization is used or not. Without reorthogonalization, as shown in the left figure of Figure 3.4, $\|E_k\|_F$ is of order $(\varepsilon_v + \varepsilon_s)$ for $B_k = T_{\tilde{H}}$. The bound of $T_{\tilde{H}}$ correctly captures the trends of the curve and overestimates the actual value by an almost constant factor of just three.

The left figure of Figure 3.4 also illustrates that using $\tilde{H}_k$ instead of $\hat{H}_k$ yields similar results for the first four steps. Starting from step five, i.e., after the near breakdown, this choice requires a much larger perturbation $E_k$. The bound also captures the trends, and overestimates by a factor of $\approx 20$. However, the bound jumps one step earlier than the actual value which can be explained by the occurrence of the term $\zeta_{k+1}$ in (3.112) that seems to be too conservative.

With reorthogonalization, as shown in the right figure of Figure 3.4, there is no relevant difference between $T_{\tilde{H}}$ and $T_{\hat{H}}$, neither in $\|E_k\|_F$, nor in its bounds. All these quantities are of order $(\varepsilon_v + \varepsilon_s)$.
3.6 Conclusion

In this chapter we have investigated the behavior of Arnoldi’s method in a scenario where all vector operations, i.e., matrix-vector multiplication, scalar products, vector addition and scalings entail perturbations. We have proposed a Gram Schmidt orthogonalization variant, called compensated Gram Schmidt (ComGS), tailored to this scenario.

In the first part, i.e., in the Subsections 3.1–3.3, we have shown that using the ComGS method, if needed enhanced by reorthogonalization, in the inexact Arnoldi method computes a nearly orthogonal basis of a sequence of the associated Krylov subspaces. We have shown that the bases are orthogonal to the same level of accuracy as the vector operations themselves.

We have seen that there are some cases where ComGS maintains the orthogonality much better than MGS and in the other cases it keeps the orthogonality as good as MGS. In other words, in some cases we would have to perform an additional step of reorthogonalization in MGS in order to achieve the same level of orthogonality as in ComGS without reorthogonalization.

In the parallel computing environment the ComGS method should be the method of choice, because in contrary to the MGS method, the ComGS method is parallelizable. Hence, taking the advantage of BLAS 2 or parallel facilities into account, reduces the computation time significantly. This was verified by several experiments where it was shown that using CGS with reorthogonalization in a parallel setting can be faster than MGS, even so twice as many operations are needed for the CGS method with reorthogonalization, see, e.g., [35,72].

Finally, the numerical results confirm that without reorthogonalization MGS and ComGS behave similar but there are some cases where ComGS behaves much better, see Example 1. Using reorthogonalization MGS and ComGS behave similar, however, there are cases, e.g., Example 2, where only ComGS re gives a nearly orthogonal basis. Thus, computing the implicitly available basis \( \tilde{V} = \tilde{V}G^{-1} \) avoids an additional orthogonalization step. This additional step would have been necessary for the CGS and the MGS method.

In the second part, i.e., in Section 3.4, we showed that for a Hermitian matrix \( A \) the inexact Arnoldi method ensures an exact Krylov relation of a nearby matrix \( A + E \). To provide that the perturbed matrix \( A + E \) is also Hermitian we replaced the non-Hermitian Hessenberg matrix \( H \) by an Hermitian matrix \( B \). We have provided several choices for \( B \) and proved bounds for the corresponding norm of \( E \). In Experiment 4 the bounds have been confirmed to correctly predict the trends of the curve within one order of magnitude.
Central to the inexact Arnoldi algorithm considered in the last chapter are inexact vector operations, i.e., summations, inner products and matrix-vector multiplications that can only be evaluated approximately. These inexact operations used in a Krylov subspace method, however, affect the orthogonality of the subspace basis, as analyzed in Chapter 3, but also the accuracy of the approximated eigenvalues and invariant subspaces. In this chapter we focus on the evaluation of the quality of the spectral approximation. More precisely, we are interested in how well the target subspace $\mathcal{X}$ is approximated by the computed Krylov subspace $\mathcal{K}_k$ or by the Ritz space $\mathcal{Y}$ in $\mathcal{K}_k$.

As starting point for our considerations we use the inexact Arnoldi method, stated in Algorithm 3.3. As shown in Section 3.4.1 and, e.g., in [74, 97, 102, 117] the resulting perturbed Arnoldi relation (3.102) can be interpreted as an exact Krylov relation of a perturbed matrix $(A + E_k)$. In particular, the inexact Arnoldi method will generate an orthonormal $\hat{V}_k = [\hat{v}_1, \ldots, \hat{v}_{k+1}] \in \mathbb{C}^{n \times k+1}$, $b_{k+1} \in \mathbb{C}^k$, and an Hermitian $B_k \in \mathbb{C}^{k \times k}$ such that a Krylov relation of the form

$$(A + E_k)\hat{V}_k = \hat{V}_k B_k + \hat{v}_{k+1}b^H_{k+1} \tag{4.1}$$

holds. The backward error matrix $E_k = E_k^H \in \mathbb{C}^{n \times n}$ on the left-hand side of equation (4.1) is unknown, but usually a bound on its norm $\|E_k\|_2$ is available, as shown in Section 3.4.2.

In this chapter a tilde indicates a quantity that corresponds to a perturbed matrix $A + E_i$ for $i = 1, \ldots, k$. For example $\tilde{K}_k$ is a Krylov subspace of $A + E_k$ and $\tilde{Y}$ is a Ritz space of $A + E_k$ in $\tilde{K}_k$. Relation (4.1) implies that (for some starting vector $\tilde{v}_1 \in \text{ran}(\hat{V}_i)$) the matrix $\hat{V}_i \in \mathbb{C}^{n \times i}$ is a basis of a Krylov subspace

$$\tilde{K}_i := \text{ran}(\hat{V}_i)$$

of the Hermitian matrix $(A + E_i)$ that is close to $A$. The sequence of generated subspaces $\tilde{K}_i$ contains increasingly accurate approximations to the desired eigenvectors.

To extract approximations to an eigenspace $\mathcal{X}$ of the matrix $A$ from the subspace $\mathcal{K}_k$ of the matrix $A + E_k$ we use the Rayleigh-Ritz method, as stated in Algorithm 2.4. In more detail, we compute the eigenpair $(\tilde{M}, \tilde{Z})$ of $B_k$ and take $(\tilde{M}, \tilde{Y}) = (\tilde{M}, \hat{V}_k \tilde{Z})$ as Ritz pair of the matrix $A + E_k$. For convenience we call the subspace generated by an exact Krylov subspace method exact Krylov subspace and inexact Krylov subspace otherwise. In order to specify the quality
of the approximation on the exact subspace $\mathcal{X}$ of the original matrix $A$ we first have to find a suitable measure for the quality of the approximate invariant subspace. We propose to use the concept of the angle of inclusion introduced in Section 2.2.2. As stated in (2.5) for nonzero subspaces $\mathcal{X}, \mathcal{Y} \subset \mathbb{C}^n$ the angle of inclusion is defined by

$$\angle_{\text{max}}(\mathcal{X}, \mathcal{Y}) := \max_{x \in \mathcal{X}, x \neq 0} \angle(x, \mathcal{Y}).$$

A small angle $\angle_{\text{max}}(\mathcal{X}, \mathcal{Y})$ does not mean that $\mathcal{X}$ is close to $\mathcal{Y}$, but it indicates that $\mathcal{X}$ is almost contained in $\mathcal{Y}$. In numerous applications that is all that is needed [3]. For $\dim(\mathcal{X}) \leq \dim(\mathcal{Y})$ the angle $\angle_{\text{max}}(\mathcal{X}, \mathcal{Y})$ coincides with the well-known maximal canonical angle [14, 89, 120]. Otherwise, i.e., for $\dim(\mathcal{X}) > \dim(\mathcal{Y})$ the angle is $\pi/2$. This unsymmetric formulation is reasonable, because for example a 2-dimensional space can be approximately contained in a 3-dimensional space but obviously not vice versa. As a consequence we obtain that the bounds hold also in the trivial case when $\dim(\mathcal{X}) > \dim(\mathcal{Y})$.

Using this measure we can now turn to the central question of this chapter:

How well is a desired invariant subspace $\mathcal{X}$ of the matrix $A$ approximated by the Krylov subspace $\mathcal{K}_k$ of a perturbed matrix $A + E_k$ or by a Ritz space $\mathcal{Y}$ in $\mathcal{K}_k$?

We present a priori and a posteriori bounds of the quality of these approximations. In the a priori setting we want to analyze the question: if a few iteration steps have been taken without convergence, how many more iterations have to be performed to achieve a specific tolerance. In the a posteriori setting we aim to derive error bounds on the angle between the exact invariant subspace and its approximation.

The theory in this chapter is strongly based on spectral perturbation theory. Classical works in this field include [27, 53, 89, 92, 112], where the lists of references is fare away from being complete. For the a priori setting we will generalize a classic result [90, Theorem 6.3] that bounds the angle between an eigenvector and the $k$-th Krylov subspace. This bound was enhanced in [56] to treat defective eigenvalues and generalized for block Krylov methods in [71, 99]. Further, bounds for the angle between an invariant subspace of $A$ and the Krylov subspace $\mathcal{K}_k(A, v_1)$ are presented in [9], where the subspace is generated by a Krylov method with polynomial restarts. In the a posteriori setting the distance of an approximate subspaces from the invariant subspace was considered in [55, 78] where the derived bound depends on the residual and in [109] where the bound is based on the quality of a surrounding search space. Usually traditional bounds, see, e.g., [27, 55, 92, 109], are problematic when there exists a small gap between the desired and remaining eigenvalues. In the following we are particularly interested in how the sensitivity of the derived spectral error bounds with respect to a small gap can be avoided.
4.1 Distance from a search space to an invariant subspace

The spectral bounds introduced in this chapter are divided into bounds on the distance to Krylov subspaces and bounds on the distance to Ritz spaces. Bounds on the distance to a Krylov subspace are analyzed in Section 4.1 where we present an a priori result, i.e., only knowledge about \( \| E_k \|_2 \) is required. The latter bounds, i.e., bounds on the distance to Ritz spaces are considered in Section 4.2 where we provide a posteriori results that require in addition to knowledge about \( \| E_k \|_2 \) also information about \( \hat{V}_k, B_k, b_{k+1} \) and the residual \( \tilde{R} \). In both cases knowledge about the exact eigenvalues of \( A \) are assumed. The presented bounds are verified by numerical experiments, including an academic example and an one-dimensional quantum Ising model, in Section 4.3. Lastly, we state some concluding remarks in Section 4.4.

4.1 Distance from a search space to an invariant subspace

In this section we analyze how well an invariant subspace \( \mathcal{X} \) is approximated by the search space \( \tilde{\mathcal{K}}_k \), i.e., by the Krylov subspace of a perturbed matrix \( A + E_k \). More specifically, we investigate a situation where \( \ell \) iterations of an inexact Krylov subspace method have been carried out but the desired eigenpair approximations cannot yet be considered to have converged and we are interested in how many more iteration steps have to be carried out until we achieve convergence. The spectral perturbation analysis presented in this section is applicable to a wide range of inexact Krylov subspace methods and in particular also to the inexact Arnoldi method stated in Section 3.3.

The following lemma lays the foundation for the proof of the main theorem of this section. It provides a measure how well the invariant subspace \( \text{ran}(X_1) \) of \( A \) is contained in the invariant subspace of larger dimension \( \text{ran}([\tilde{X}_1, \tilde{X}_2]) \) of a perturbed matrix \( A + E \) and is a corollary of Theorem 2.12. Note that Lemma 4.1 is even applicable if the eigenvalues corresponding to \( \text{ran}(X_1) \) are not well separated from the remaining ones.

**Lemma 4.1.** Let \( A, E \in \mathbb{C}^{n \times n} \) be Hermitian. Let \( X_i, \tilde{X}_i \in \mathbb{C}^{n \times n} \) for \( i = 1, 2, 3 \) with \( n_1 + n_2 + n_3 = n \) be such that \( X := [X_1, X_2, X_3], \tilde{X} := [\tilde{X}_1, \tilde{X}_2, \tilde{X}_3] \) are unitary and

\[
X^HAX = \text{diag}(A_{11}, A_{22}, A_{33}), \quad \tilde{X}^H(A + E)\tilde{X} = \text{diag}(\tilde{A}_{11}, \tilde{A}_{22}, \tilde{A}_{33})
\]

are block diagonal.

i) If \( \text{gap}(\text{env}(\text{eig}(A_{11})), \text{eig}(\tilde{A}_{33}))) > \| E \|_2 \) then

\[
\tan \gamma_{\text{max}}(X_1, [\tilde{X}_1, \tilde{X}_2]) \leq \frac{\| E \|_2}{\text{gap}(A_{11}, \tilde{A}_{33}) - \| E \|_2}.
\]

92
4.1 Distance from a search space to an invariant subspace

ii) If \( \text{gap}(\text{env}(\text{eig}(A_{11})), \text{eig}(A_{33})) > 2\|E\|_2 \) and

\[
\max_{\lambda \in \text{eig}(A_{33})} \min_{\lambda \in \text{eig}(A_{33})} |\bar{\lambda} - \lambda| \leq \|E\|_2
\]

then

\[
\tan \angle_{\text{max}}^\circ \left( X_1, [\bar{X}_1, \bar{X}_2] \right) \leq \frac{\|E\|_2}{\text{gap}(A_{11}, A_{33}) - 2\|E\|_2}.
\]

iii) If \( \text{gap}(\text{env}(\text{eig}(\bar{A}_{11})), \text{eig}(\bar{A}_{33})) > 2\|E\|_2 \) and

\[
\max_{\lambda \in \text{eig}(A_{11})} \min_{\lambda \in \text{eig}(A_{11})} |\bar{\lambda} - \lambda| \leq \|E\|_2
\]

then

\[
\tan \angle_{\text{max}}^\circ \left( X_1, [\bar{X}_1, \bar{X}_2] \right) \leq \frac{\|E\|_2}{\text{gap}(A_{11}, A_{33}) - 2\|E\|_2}.
\]

Proof. Defining \( \bar{A}_3 := \bar{X}_3^H A \bar{X}_3 = \bar{X}_3^H (A + E) \bar{X}_3 - \bar{X}_3^H E \bar{X}_3 = \bar{A}_{33} - \bar{X}_3^H E \bar{X}_3 \) we get

\[
R := A \bar{X}_3 - \bar{X}_3 \bar{A}_3 = (A + E) \bar{X}_3 - E \bar{X}_3 - \bar{X}_3 (\bar{A}_{33} - \bar{X}_3^H E \bar{X}_3) = -E \bar{X}_3 + \bar{X}_3 \bar{X}_3^H E \bar{X}_3 = -(I - \bar{X}_3 \bar{X}_3^H) E \bar{X}_3,
\]

which implies that \( \|R\|_2 \leq \|E\|_2 \). Since \( \bar{A}_3 \) and \( \bar{A}_{33} \) differ by \( \bar{X}_3^H E \bar{X}_3 \), by Theorem 2.1 their eigenvalues differ by at most \( \|E\|_2 \), thus

\[
\text{gap}(\text{eig}(A_{11}), \text{eig}(\bar{A}_3)) \geq \text{gap}(\text{eig}(A_{11}), \text{eig}(\bar{A}_{33})) - \|E\|_2.
\]

Together with the assumption

\[
\text{gap} \left( \text{env}(\text{eig}(A_{11})), \text{eig}(\bar{A}_{33}) \right) > \|E\|_2
\]

we conclude that there is no eigenvalue of \( \bar{A}_3 \) in the interval

\[
[\lambda_{\min}(A_{11}) - \text{gap}(A_{11}, \bar{A}_3), \lambda_{\max}(A_{11}) + \text{gap}(A_{11}, \bar{A}_3)],
\]

where \( \lambda_{\min}(A_{11}) := \min(\text{eig}(A_{11})) \) and \( \lambda_{\max}(A_{11}) := \max(\text{eig}(A_{11})) \). Hence Theorem 2.12 is applicable and leads to

\[
\angle_{\text{max}}^\circ (\bar{X}_3, [X_2, X_3]) \leq \frac{\|R\|_2}{\text{gap}(A_{11}, \bar{A}_3)} \leq \frac{\|E\|_2}{\text{gap}(A_{11}, \bar{A}_3)} \leq \frac{\|E\|_2}{\text{gap}(A_{11}, \bar{A}_{33}) - \|E\|_2}.
\]
4.1 Distance from a search space to an invariant subspace

From Lemma 2.10.iii it follows that

\[ \angle_{\text{max}}(X_1, [\tilde{X}_1, \tilde{X}_2]) = \angle_{\text{max}}([\tilde{X}_1, \tilde{X}_2], X_1^\perp) = \angle_{\text{max}}(\tilde{X}_3, [X_2, X_3]), \]

which concludes the proof of part i).

For part ii) we use that by assumption \( \max_{\lambda \in \text{eig}(\tilde{A}_{33})} \min_{\lambda \in \text{eig}(A_{33})} |\tilde{\lambda} - \lambda| \leq \|E\|_2 \) for every eigenvalue of \( \tilde{A}_{33} \) there exists an eigenvalue of \( A_{33} \) that is no further away than \( \|E\|_2 \). This implies \( \text{gap}(A_{11}, A_{33}) \geq \text{gap}(A_{11}, \tilde{A}_{33}) - \|E\|_2 \). Together with part i) we obtain the assertion. Analogously, also part iii) follows from part i).

\[ \square \]

Remark 4.2. The assumption in part ii) that \( \max_{\lambda \in \text{eig}(\tilde{A}_{33})} \min_{\lambda \in \text{eig}(A_{33})} |\tilde{\lambda} - \lambda| \leq \|E\|_2 \) and the similar condition in part iii) are not very restrictive. By Theorem 2.1, the eigenvalues of \( A \) differ from the corresponding ones of \( A + E \) by at most \( \|E\|_2 \). So the assumption is mainly a restriction on the distribution of \( \text{eig}(A + E) \) into \( \text{eig}(\tilde{A}_{11}), \text{eig}(\tilde{A}_{22}), \) and \( \text{eig}(\tilde{A}_{33}) \).

In the main theorem of this section we make extensive use of nested subsets of eigenvalues, that are defined by

\[ \Lambda_{n+1} \subseteq \Lambda_n \quad \forall n \in 1, 2, \ldots \]

An illustration of the nested spectral subsets used in Theorem 4.4 and an introduction of the corresponding notation is given in Figure 4.1. The nested subsets are denoted by \( \Lambda_1, \ldots, \Lambda_4 \), where the set of eigenvalues possibly increases with increasing subscripts. Further, the spectral complement of the subset \( \Lambda_i \) is denoted by \( \Lambda_{-i} \). The subscripts of the gaps in Figure 4.1 indicate whether we consider the left or right gap between the subsets, such that

\[ \text{gap}(\Lambda_{i'}, \Lambda_j) = \min\{\text{gap}_{i',j'}, \text{gap}_{j,i}\}. \]

Using this notation, the following theorem provides a bound on the angle of inclusion of the invariant subspace \( X_1 \) of \( A \) and the search space \( \tilde{K}_k \).

Assumption 4.3. Denote by \( J_L := \{1, 2, \ldots, \min(J_3) - 1\} \) the leading and by \( J_T := \{\max(J_3) + 1, \ldots, n\} \) the trailing indices of \( J_3 \).

Let \( \Lambda_i := \{\lambda_j : j \in J_i\} \) and \( \Lambda_{-i} := \{\lambda_j : j \in \{1, \ldots, n\} \setminus J_i\} \) for \( i \in \{1, \ldots, 4, L, T\} \).

Let \( X_1 \) and \( X_4 \) be invariant subspaces of \( A \) corresponding to \( \Lambda_1 \) and \( \Lambda_4 \), respectively.

For \( j = 1, \ldots, k \) let \( \tilde{K}_j := K_j(A + E_j, \tilde{v}_1) \) for some Hermitian \( E_j \in \mathbb{C}^{n \times n} \) and some \( \tilde{v}_1 \in \mathbb{C}^n \).

For \( i \in \{2, 3\} \) let \( \tilde{X}_i \) be an invariant subspace of \( A + E_k \) corresponding to \( \{\lambda_j(A + E_k) : j \in J_i\} \) such that \( \tilde{X}_2 \subset \tilde{X}_3 \).
4.1 Distance from a search space to an invariant subspace

Let $A \in \mathbb{C}^{n \times n}$ be Hermitian with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Let $J_1 \subseteq J_2 \subseteq J_3 \subseteq J_4$ be nested nonempty subsets of $\{1, 2, \ldots, n\}$ so that $\max(J_3) < n$ and $J_2, J_3$ consist of consecutive integers. If Assumption 4.3 holds and $2\|E_k\|_2 < \min \{\text{gap} (\Lambda_1, \Lambda_{-2}), \text{gap} (\Lambda_2, \Lambda_{-3}), \text{gap} (\Lambda_3, \Lambda_{-4}) \}$ then for every $\ell = 1, \ldots, k - |J_L|$, we have

\[
\angle_{\max}(X_1, \tilde{K}_k) \leq \angle_{\max}(X_2, \tilde{K}_k) + \delta_{1,2} \\
\leq \arctan \left( q_{k,\ell} \cdot \tan \angle_{\max}(X_3, \tilde{K}_\ell) \right) + \delta_{1,2} \\
\leq \arctan \left( q_{k,\ell} \cdot \tan \angle_{\max}(X_4, \tilde{K}_\ell) + \delta_{3,4} \right) + \delta_{1,2}
\]

with $\tan_{\frac{\pi}{2}}(a) := \tan(\min\{a, \frac{\pi}{2}\})$ and

\[
q_{k,\ell} := \left( \sum_{i \in J_2} \frac{\tilde{\theta}_i^2}{\psi_{k-\ell-|\Lambda_1|}(1 + 2\tilde{\eta}_i)^2} \right)^{\frac{1}{2}}, \quad \delta_{i,j} := \arctan \left( \frac{\|E_k\|_2}{\text{gap}(\Lambda_i, \Lambda_{-j}) - 2\|E_k\|_2} \right),
\]

\[
\tilde{\theta}_i := \prod_{j \in J_L} \frac{\lambda_j - \lambda_n}{|\lambda_j - \lambda_i| - 2\|E_k\|_2} > 0, \quad \tilde{\eta}_i := \frac{\text{gap}(\Lambda_i, \Lambda_T) - 2\|E_k\|_2}{\text{spread}(\Lambda_T) + 2\|E_k\|_2} > 0,
\]

where $\psi_j$ denotes the Chebychev polynomial of degree $j$.

**Proof.** The proof of Theorem 4.4 can be divided into three steps, i.e., the claim follows by proving

i) $\angle_{\max}(X_1, \tilde{K}_k) \leq \angle_{\max}(X_2, \tilde{K}_k) + \delta_{1,2},$

ii) $\tan \angle_{\max}(X_2, \tilde{K}_k) \leq q_{k,\ell} \cdot \tan \angle_{\max}(X_3, \tilde{K}_\ell),$ and

![Figure 4.1: Illustration of the nested spectral subsets](image)

\[\Box\]
4.1 Distance from a search space to an invariant subspace

iii) \( \angle_{\max} (\mathcal{X}_3, \mathcal{K}_\ell) \leq \min \left( \frac{\pi}{2}, \angle_{\max} (\mathcal{X}_4, \mathcal{K}_\ell) + \delta_{3,4} \right) \),

respectively and finally the exploitation of the monotonicity of the tangent in the interval \([0, \frac{\pi}{2}]\) in Step ii) and iii).

In order to prove Step i) we use Theorem 4.1 ii) and the definition of \( \delta_{1,2} \) such that

\[
\tan \angle_{\max} (\mathcal{X}_1, \mathcal{X}_2) \leq \frac{\|E_k\|_2}{\text{gap}(\Lambda_1, \Lambda_{-2}) - 2\|E_k\|_2} = \tan \delta_{1,2}.
\]

Hence, by the triangle inequality, stated in Lemma 2.10.i), we get

\[
\angle_{\max} (\mathcal{X}_1, \mathcal{K}_k) \leq \angle_{\max} (\mathcal{X}_1, \mathcal{X}_2) + \angle_{\max} (\mathcal{X}_2, \mathcal{K}_k) \leq \delta_{1,2} + \angle_{\max} (\mathcal{X}_2, \mathcal{K}_k).
\]

To prove Step ii) we distinguish between two cases:

\[
\text{dim}(\mathcal{K}_\ell) < \text{dim}(\mathcal{X}_3) \quad \text{and} \quad \text{dim}(\mathcal{K}_\ell) \geq \text{dim}(\mathcal{X}_3).
\]

If \( \text{dim}(\mathcal{K}_\ell) < \text{dim}(\mathcal{X}_3) = |J_3| \) then \( \angle_{\max} (\mathcal{X}_3, \mathcal{K}_\ell) = \frac{\pi}{2} \) and the right-hand side is infinite. Thus there is nothing to prove in this case and we continue by investigating the second case, i.e., \( \text{dim}(\mathcal{K}_\ell) \geq \text{dim}(\mathcal{X}_3) = |J_3| \). We start by considering the angles between \( \mathcal{K}_k \) and eigenvectors of \( A + E_k \). So, let \( \tilde{x}_1, \ldots, \tilde{x}_n \) be unit eigenvectors of \( A + E_k \) corresponding to the eigenvalues \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \) such that \( \mathcal{X}_3 = \text{span}\{ \tilde{x}_j : j \in J_3 \} \) and \( \mathcal{X}_2 = \text{span}\{ \tilde{x}_j : j \in J_2 \} \). Let \( \tilde{\Lambda}_T := \{ \tilde{\lambda}_j : j \in J_T \} \). For \( i \in J_2 \) we define

\[ V_i := \text{span}\{ \tilde{x}_j : j \in J_3, j \neq i \} = \mathcal{X}_3 \cap \text{span}(\tilde{x}_i)^\perp. \]

Note that \( \mathcal{K}_\ell \cap V_i^\perp \) is an intersection of a \((|J_3|)\)-dimensional and an \((n - |J_3| + 1)\)-dimensional subspace, i.e., it is at least one-dimensional. Consequently, there exists a nonzero vector \( v_i \in \mathcal{K}_\ell \cap V_i^\perp \) such that \( \angle(\tilde{x}_i, v_i) = \angle(\tilde{x}_i, \mathcal{K}_\ell \cap V_i^\perp) \). We define

\[
A_i := A + E_k - \sum_{j \in J_3 \setminus i} (\tilde{\lambda}_j - \tilde{\lambda}_n) \tilde{x}_j \tilde{x}_j^H.
\]

Note that the matrices \( A_i \) and \( A + E_k \) have the same eigenvectors. Moreover, the eigenvalues \( \tilde{\lambda}_i \) of both matrices for \( j \in (\{1, \ldots, n\} \setminus J_3) \cup \{i\} \) coincide. By construction of \( A_i \) the remaining eigenvalues of \( A + E_k \) are shifted to the smallest eigenvalue \( \tilde{\lambda}_n \in \tilde{\Lambda}_T \). Since, the eigenvalue \( \tilde{\lambda}_i \) of \( A_i \) is well separated from the rest of the spectrum, the bound of Theorem 2.14 is meaningful. The application of Theorem 2.14 to matrix \( A_i \) and the vector \( v_i \) yields

\[
\tan \angle(\tilde{x}_i, \mathcal{K}_{k-\ell+1}(A_i, v_i)) \leq \rho_i := \frac{\theta_i}{\psi_{k-\ell-|J_T|}(1 + 2\eta_i)} \tan \angle(\tilde{x}_i, v_i), \tag{4.6}
\]

96
4.1 Distance from a search space to an invariant subspace

with

\[
\theta_i := \prod_{j \in J_k} \left| \frac{\tilde{\lambda}_j - \tilde{\lambda}_n}{\tilde{\lambda}_j - \tilde{\lambda}_i} \right|, \quad \eta_i := \frac{|\tilde{\lambda}_i - \tilde{\lambda}_{\text{min}(f_T)}|}{|\tilde{\lambda}_{\text{min}(f_T)} - \tilde{\lambda}_n|} = \text{gap}(\tilde{\lambda}_i, \tilde{\Lambda}_T), \quad (4.7)
\]

Let \( g_i \in \mathcal{K}_{k-\ell+1}(A_i, v_i) \) be a vector such that \( \angle(\tilde{x}_i, g_i) = \angle(\tilde{x}_i, \mathcal{K}_{k-\ell+1}(A_i, v_i)) \). W.l.o.g. we assume that \( g_i \) is scaled such that \( \tilde{x}_i^T g_i = 1 \). Otherwise, we distinguish between two cases. In the one hand if \( \tilde{x}_i^T g_i \neq 0 \) then \( g_i \) can be rescaled and on the other hand if \( \tilde{x}_i^T g_i = 0 \) then it follows with Lemma 2.10.v) that

\[
\frac{\pi}{2} = \angle(\tilde{x}_i, g_i) = \angle(\tilde{x}_i, \bar{\mathcal{K}}_\ell \cap \mathcal{V}_i) \leq \angle_{\text{max}}(\bar{x}_3, \bar{\mathcal{K}}_\ell).
\]

Thus, \( \angle_{\text{max}}(\bar{x}_3, \bar{\mathcal{K}}_\ell) = \pi/2 \) and thus \( q_{k,\ell} \cdot \tan \angle_{\text{max}}(\bar{x}_3, \bar{\mathcal{K}}_\ell) \), the right-hand side of Step ii), is infinite and nothing is to prove in this case. Note that by the construction of \( \mathcal{V}_i \) we have \( v_i \perp \mathcal{V}_i \), thus \( A_i v_i = (A + E_k) v_i = A_i v_i \perp \mathcal{V}_i \). Thus by induction it follows that \( \mathcal{K}_j(A_i, v_i) = \mathcal{K}_j(A + E_k, v_i) \) and \( \mathcal{K}_j(A_i, v_i) \perp \mathcal{V}_i \) for every \( j = 1, 2, \ldots \). Due to \( g_i \in \mathcal{K}_{k-\ell+1}(A_i, v_i) \) it holds that \( g_i \perp \mathcal{V}_i \) and consequently there exist \( g_{L,i} \in \mathbb{C}^{\|I_i\|} \) and \( g_{T,i} \in \mathbb{C}^{\|J_i\|} \) such that

\[
g_i = [\tilde{x}_1, \ldots, \tilde{x}_J] \|g_{L,i}\| + \tilde{x}_i + [\tilde{x}_{\text{min}(f_T)}, \ldots, \tilde{x}_n] \|g_{T,i}\|.
\]

Using Theorem 2.11, the above definition of \( g_i \), and inequality (4.6) leads to

\[
\|([g_{L,i}^T g_{T,i}^T]^T)\|_2 = \tan \angle(g_i, \tilde{x}_i) \leq \rho_i. \quad (4.8)
\]

Since the subspaces \( \mathcal{K}_j \) are nested and \( v_i \in \mathcal{K}_\ell \), we have

\[
\mathcal{K}_{k-\ell+1}(A + E_k, v_i) \subset \mathcal{K}_k(A + E_k, \bar{v}_1) = \bar{\mathcal{K}}_k.
\]

Therefore, \( g_i \in \bar{\mathcal{K}}_k \). In the next step we generalize equation (4.8) to obtain a bound on the angle of inclusion of \( \bar{x}_2 \) to \( \bar{\mathcal{K}}_k \). Since \( \bar{\mathcal{K}}_k \) is independent of \( i \), it contains \( g_i \) for all \( i \in J_2 \). Thus it contains the subspace \( \text{ran}(G) \) where

\[
G := [g_{\text{min}(J_2)}, \ldots, g_{\text{max}(J_2)}] = [\tilde{x}_1, \ldots, \tilde{x}_n] \|I_{\|J_2\|}\| + [0, \ldots, 0] \|G_T\|
\]

with \( G_L := [g_{L,\text{min}(J_2)}, \ldots, g_{L,\text{max}(J_2)}] \in \mathbb{C}^{\|I_2\| \times \|J_2\|} \), \( G_T := [g_{T,\text{min}(J_2)}, \ldots, g_{T,\text{max}(J_2)}] \in \mathbb{C}^{\|J_2\| \times \|J_2\|} \), and the upper and lower zero blocks are of the format...
4.1 Distance from a search space to an invariant subspace

\[(\min(J_2) - \min(J_3)) \times |J_2| \text{ and } (\max(J_3) - \max(J_2)) \times |J_2|, \text{ respectively. Using Lemma 2.10.ii) and Theorem 2.11 we have} \]

\[
\tan \angle_{\text{max}}(\tilde{X}_2, \tilde{K}_k) \leq \tan \angle_{\text{max}}(\tilde{X}_2, \text{ran}(G)) = \left\| \begin{bmatrix} G_L \\ G_T \end{bmatrix} \right\|_2.
\] (4.9)

Further transformations yield

\[\left\| \begin{bmatrix} G_L \\ G_T \end{bmatrix} \right\|_2 \leq \left\| \begin{bmatrix} G_L \\ G_T \end{bmatrix} \right\|_F \leq \left( \sum_{i \in J_2} \rho_i^2 \right)^{\frac{1}{2}},\]

such that with (4.9)

\[
\tan \angle_{\text{max}}(\tilde{X}_2, \tilde{K}_k) \leq \left( \sum_{i \in J_2} \rho_i^2 \right)^{\frac{1}{2}}.
\] (4.10)

In order to determine a bound for \(\rho_i\), we apply Theorem 2.1 and obtain for \(i \neq j\) that

\[|\lambda_i - \lambda_j| - 2\|E_k\|_2 \leq |\bar{\lambda}_i - \bar{\lambda}_j| \leq |\lambda_i - \lambda_j| + 2\|E_k\|_2.\]

Hence we get \(\theta_i \leq \tilde{\theta}_i\) and \(\eta_i \geq \tilde{\eta}_i\), where \(\theta_i\) and \(\eta_i\) are defined in (4.7) and \(\tilde{\theta}_i\) and \(\tilde{\eta}_i\) are defined in (4.5). We note that by the assumed bound on \(\|E_k\|_2\) the scalars \(\tilde{\theta}_i\) and \(\tilde{\eta}_i\) are positive. Moreover, we have

\[\psi_{k - \ell - |J_2|}(1 + 2\eta_i) > \psi_{k - \ell - |J_2|}(1 + 2\tilde{\eta}_i),\]

because from (2.7) we see that Chebychev polynomials are positive and monotonically increasing in \([1; \infty)\). Furthermore, since \(\tilde{X}_3 = \text{span}(\bar{x}_i) \oplus V_i\), it follows from Lemma 2.10.v) that \(\angle(\tilde{x}_i, \tilde{K}_\ell \cap V_i) \leq \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell), \text{ hence } \angle(\tilde{x}_i, v_i) \leq \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell).\) Thus, insertion of the above bounds into (4.6) results in

\[\rho_i \leq \tilde{\theta}_i \psi_{k - \ell - |J_2|}(1 + 2\tilde{\eta}_i)^{-1} \tan \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell).
\] (4.11)

Finally, inserting (4.11) into (4.10) and using the definition (4.4) of \(\varrho_{k,\ell}\) results in

\[
\tan \angle_{\text{max}}(\tilde{X}_2, \tilde{K}_k) \leq \left( \sum_{i \in J_2} \psi_{k - \ell - |J_2|}(1 + 2\tilde{\eta}_i)^2 \varrho_i^2 \right)^{\frac{1}{2}} \tan \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell) = \varrho_{k,\ell} \tan \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell),
\]

which concludes the proof of Step ii).

The last Step iii) is proven similar to Step i). We use a triangle inequality, this
4.1 Distance from a search space to an invariant subspace

time in the form of
\[ \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell) \leq \angle_{\text{max}}(\tilde{X}_3, \mathcal{X}_4) + \angle_{\text{max}}(\mathcal{X}_4, \tilde{K}_\ell) \]
and apply Theorem 4.1 iii) such that
\[ \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell) \leq \delta_{3,4} + \angle_{\text{max}}(\mathcal{X}_4, \tilde{K}_\ell) . \]
This concludes the proof of the theorem.

The following remarks highlight some additional insights that Theorem 4.4 offers.

**Remark 4.5.** Theorem 4.4 is a generalization of Theorem 2.14 with the following improvements.

i) In contrast to Theorem 2.14, where the search space is a Krylov subspace of the matrix \( A \), we chose the search space to be a Krylov subspace of a perturbed matrix \( A + E_k \), such that for \( E_k \neq 0 \) additionally the setting of inexact Krylov methods is covered by Theorem 4.4.

ii) Instead of the angle between just one eigenvector and the search space we consider the angle of inclusion of an invariant subspace of the matrix \( A \) and the search space. This offers the possibility to treat clusters of eigenvalues as a whole.

iii) The angle \( \angle(v_1, z_i) \) on the right-hand side of (2.8) can be rewritten as \( \angle(v_1, \mathcal{K}_1) \). In Theorem 4.4 the dimension \( \ell \) of the Krylov subspace on the right-hand side of (4.2) and (4.3) is allowed to be larger than one. This improvement is necessary for the theorem to be meaningful as for \( \ell < \dim(\mathcal{X}_3) \) the right-hand side of (4.2) and for \( \ell < \dim(\mathcal{X}_4) \) the right-hand side of (4.3) is infinite.

Moreover, this may be useful if information about the angles \( \angle_{\text{max}}(\tilde{X}_3, \tilde{K}_\ell) \) or \( \angle_{\text{max}}(\mathcal{X}_4, \tilde{K}_\ell) \) are available for some \( \ell < k \).

iv) Theorem 4.4 provides meaningful results even if the eigenvalues \( \Lambda_1 \) are not well separated from the rest of the spectrum, i.e., if \( \text{gap}(\Lambda_1, \Lambda_{-1}) - 2\|E_k\|_2 \) is very small or even negative. This is achieved by considering four possibly different but nested sets of eigenvalues, \( \Lambda_1, \ldots, \Lambda_4 \). In fact what is important is that the gaps between \( \Lambda_i \) and the complement of \( \Lambda_{i+1} \), i.e., \( \text{gap}(\Lambda_1, \Lambda_{-2}) \), \( \text{gap}(\Lambda_2, \Lambda_{-3}) \), and \( \text{gap}(\Lambda_3, \Lambda_{-4}) \), are sufficiently larger than \( 2\|E_k\|_2 \). We stress that \( \text{gap}(\Lambda_i, \Lambda_{-i}) \) may even be zero, i.e., \( \chi_i \) is not required to be simple. However, if \( \text{gap}(\Lambda_1, \Lambda_{-1}) - 2\|E_k\|_2 \) is sufficiently large thus there is no need to choose \( \Lambda_2, \Lambda_3, \Lambda_4 \) differently from \( \Lambda_1 \).

In summary, Theorem 4.4 is a generalization of Theorem 2.14 and reduces to the latter for \( \|E_k\|_2 = 0, \ell = 1, J_1 = J_2 = J_3 = J_4 = \{i\} \).
4.1 Distance from a search space to an invariant subspace

**Remark 4.6.** If the dimension $k$ of the search space $\mathcal{K}_k$ grows, the constant $q_{k,\ell}$ on the right-hand side of bound (4.3) decreases exponentially. In fact, from the definition of the Chebychev polynomial (2.7) and the trigonometric functions $\cosh(x) = \frac{1}{2}(e^x + e^{-x}) > \frac{1}{2}e^x$ and $\arccosh(x) = \ln(x + \sqrt{x^2 - 1})$ for $x \geq 1$ we deduce that

$$\psi_k(1 + 2\eta) > \frac{1}{2} \left(1 + 2\eta + 2\sqrt{\eta + \eta^2}\right)^k > \frac{1}{2} (1 + 2\sqrt{\eta})^k.$$  

This indicates at least linear convergence of $q_{k,\ell}$ towards zero.

**Remark 4.7.** In order to analyze the qualitative behavior of bound (4.3) we consider the following. For small arguments $0 < a \ll 1$ we have $\arctan(a) \approx a$ and $\arctan(a) \leq a$. Since in the later stages of the iteration $q_{k,\ell}$ is small we may replace $\arctan(\cdot)$ by the identity in (4.3). Doing so, taking logarithms, and substituting $y_i := 1 + 2\tilde{\eta}_i + 2\sqrt{\tilde{\eta}_i + \tilde{\eta}_i^2}$, $\tau_\ell := \tan(\frac{1}{2} (\angle_{\max}^\circ (\mathcal{X}_1, \mathcal{K}_k) + \delta_{3,4}))$ results in

$$\log(\angle_{\max}^\circ (\mathcal{X}_1, \mathcal{K}_k)) \leq \log \left( \tau_\ell \left( \frac{1}{\delta_1,2} \sum_{i \in I_2} \frac{1}{\gamma_i^2} \right)^{\frac{1}{2}} + \delta_{1,2} \right).$$

Moreover, using that $\log(a + b) \approx \max(\log(a), \log(b))$ results in

$$\log(\angle_{\max}^\circ (\mathcal{X}_1, \mathcal{K}_k)) \approx \max \left( \log(\delta_1,2), \log \left( \tau_\ell \left( \frac{1}{\delta_1,2} \sum_{i \in I_2} \frac{1}{\gamma_i^2} \right)^{\frac{1}{2}} \right) \right),$$

$$\approx \max \left( \log(\delta_1,2), \log(\tau_\ell) + \max_{i \in I_2} \left( \log(\tilde{\theta}_i) - \log(\frac{1}{2}) \right. \right.$$

$$\left. \left. - (k - \ell - |\Lambda_L|) \log(\gamma_i) \right) \right).$$

Note that $a \approx b$ results from the assumption

$$\log(a + b) \approx \max(\log(a), \log(b))$$

and should be interpreted as “$a$ is usually smaller than $b$, however it can be a bit larger, but not by much”.

During the first iteration steps the argument of $\arctan$ in (4.3) is usually large and the only safe bound for $\angle_{\max}^\circ (\mathcal{X}_1, \mathcal{K}_k)$ is $\frac{\pi}{2}$. Together with the previous considerations we get

$$\log(\angle_{\max}^\circ (\mathcal{X}_1, \mathcal{K}_k)) \approx \min \left( \log \left( \frac{\pi}{2} \right), \max \left( \log(\delta_1,2), \max_{i \in I_2} (a_i - k\beta_i) \right) \right),$$

(4.12)

with $a_i = \log(\tau_\ell) + \log(\tilde{\theta}_i) - \log(\frac{1}{2}) + (\ell + |\Lambda_L|) \log(\gamma_i)$ and $\beta_i = \log(\gamma_i) > 0$. 

100
Due to the structure of (4.12) the logarithm of $\zeta_{\text{max}}(X_1, \tilde{K}_k)$ can be approximately bounded by a set of straight lines, more precisely by two horizontal lines at $\log(p_2)$ and $\log(\delta_{1,2})$ as well as by a small number of decreasing lines defined by $a_i + k\beta_i$. For an illustration see Figure 4.2, where we can identify three phases of the iteration process. In the first phase, i.e., during the first iteration steps, the maximum of all the decreasing lines lies above the $\frac{p_2}{2}$-level. Hence the minimum on the right hand side of (4.12) is taken by $\log(p_2)$. We call this phase the initialization phase. In the second phase, which is denoted by convergence phase, the pointwise maximum of the decreasing lines $a_i + k\beta_i$ falls below the $\frac{p_2}{2}$-level and consequently defines the bound. In this phase we obtain a strictly monotonically decreasing, piecewise linear, convex shape of the bound. It is well possible that only one line dominates the whole convergence phase, i.e., there is no bend. This can be also distinguished by the numerical examples Experiment 1 and 2 in Section 4.3. In the third phase, the maximum of $\left(\log(\delta_{1,2}), \max_{i \in J_2} (a_i - k\beta_i)\right)$ in (4.12) is taken by $\log(\delta_{1,2})$. Because the bound on angle of inclusion $\zeta_{\text{max}}(X_1, \tilde{K}_k)$ does not change anymore we denote this last section as stagnation phase.

![Figure 4.2: Qualitative shape of the bound (4.3), see Remark 4.7](image)

**Remark 4.8.** Theorem 4.4 offers some freedom in choosing the index sets $J_i$. The index set $J_1$ is fixed by the indices of the desired eigenvalues. The remaining
index sets $J_2, J_3, J_4$ are free and could be chosen to minimize the bounds (4.2) and (4.3). Here the various relations between the index sets $J_1, J_2, J_3, J_4$ influence the constants $\delta_{12}, \bar{\theta}_i, \bar{\eta}_i$ and $\delta_{34}$ in different ways. For example extending $J_2$ leads to a larger gap $(\Lambda_1, \Lambda_{-2})$ and thus to a decreased $\delta_{12}$. Similar extending $J_3$ with respect to $J_2$ will improve $\bar{\theta}_i$ and $\bar{\eta}_i$. Finally extending $J_4$ with respect to $J_3$ will improve $\delta_{34}$.

**Remark 4.9.** Although this case is irrelevant in practice, we mention that Theorem 4.4 still holds if the denominator of $\bar{\eta}_i$ is zero, i.e., if $\|E_k\|_2 = 0 = \text{spread}(\Lambda_T)$, where $\text{spread}(\Lambda_T) := |\min(\Lambda_T) - \max(\Lambda_T)|$. In this case $\bar{\eta}_i = \infty$ for all $i \in J_2$ and $q_{k,\ell} = 0$. Then $\theta_{\max}(\mathcal{X}_1, \mathcal{K}_k) \leq \delta_{12}$ for all $k > \ell \in |J_L|$.

**Remark 4.10.** Theorem 4.4 relates four angles to each another: $\theta_{\max}(\mathcal{X}_1, \mathcal{K}_k), \theta_{\max}(\mathcal{X}_2, \mathcal{K}_k), \theta_{\max}(\mathcal{X}_3, \mathcal{K}_\ell)$, and $\theta_{\max}(\mathcal{X}_4, \mathcal{K}_\ell)$. The relation “looks nicest” when restricted to the first and the last of these. Especially since the middle two angles involve invariant subspaces of $A + E_k$ which are hard to obtain in practice. However, the reason to include all four angles in Theorem 4.4 is that the first angle $\theta_{\max}(\mathcal{X}_2, \mathcal{K}_k)$ appears in the following Theorem 4.15 and the second $\theta_{\max}(\mathcal{X}_3, \mathcal{K}_\ell)$ may actually be bounded in practice, see Remark 4.25.

**Remark 4.11.** The formulation of Theorem 4.4 and Figure 4.1 may seem to suggest that the theorem is also meaningful if the desired eigenvalues lie in the center of the spectrum. But, the constants $\bar{\theta}_i, i \in J_2$ grow rapidly with the number of leading eigenvalues $|J_L|$. Hence the theorem is useful only if fairly exterior eigenvalues are desired, more precisely if $|J_L|$ is not larger than two or three.

This fact is illustrated by the following example.

**Example 4.12.** Consider the matrix constructed by the MATLAB command

$$A = \text{diag}(100 : -1 : 1)$$

and choose $\|E_k\|_2 = 10^{-10}, \Lambda_2 = \Lambda_3 = \Lambda_4 = \Lambda_1$. Then for

- $\Lambda_1 = \{90, 91, \ldots, 100\}$ we get $\bar{\theta}_{\max} = 1$,
- $\Lambda_1 = \{90, 91\}$ we get $\bar{\theta}_{\max} = 1.731 \cdot 10^{12}$,
- $\Lambda_1 = \{50, 51\}$ we get $\bar{\theta}_{\max} = 5.045 \cdot 10^{28}$

with $\bar{\theta}_{\max} := \max_{i \in J_2} \bar{\theta}_i$. Hence, even if we are only interested in the 9th and 10th largest eigenvalue of $A$ Theorem 4.4 becomes less meaningful.

**Remark 4.13.** Theorem 4.4 still holds if the eigenvalues are sorted in increasing order, $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. This can be seen by applying the theorem to $-A$ and using that all the constants $\bar{\theta}_i, \bar{\eta}_i, \delta_{i,j}$ and $q_{k,\ell}$ are invariant under of negation the eigenvalues.
4.2 Convergence of the invariant subspace

So far we have considered how well the exact eigenspace $X_1$ is contained in the search space $\tilde{K}_k$. If $\angle_{\max}(X_1, \tilde{K}_k)$ is small, then there is a subspace of $\tilde{K}_k$ that is close to $X_1$. However, in practice this subspace in $\tilde{K}_k$ is not known and $X_1$ is approximated by a Ritz space $\tilde{Y}$ that is a subspace of $\tilde{K}_k$. Accordingly, in the following section we investigate how much is the distance to a Ritz space growing in comparison to the distance to the whole search space.

4.2 Convergence of the invariant subspace

Assuming that the search spaces $K_k$ contain increasingly accurate approximations of the invariant subspace of interest $X_1$ in this section we evaluate how well is this subspace $X_1$ contained in the Ritz space $\tilde{Y}$ which is a subspace of $\tilde{K}_k$. Therefore, in the following we bound the quality measure $\angle_{\max}(X_1, \tilde{Y})$ in terms of the angle between a slightly enlarged, perturbed subspace $\tilde{X}_2$ and the search space $\tilde{K}_k$.

Assumption 4.14. Let $A$, $E_k$, $\Lambda_1$, $\Lambda_2$, $\Lambda_{-2}$, $X_1$, $\tilde{X}_2$ and $\delta_{1,2}$ be defined as in Theorem 4.4.
Let $B_k \in \mathbb{C}^{k \times k}$, $b_{k+1} \in \mathbb{C}^k$ and $\tilde{V}_k \in \mathbb{C}^{n \times k}$ be such that the Krylov relation (4.1) is satisfied and $\tilde{K}_k = \text{ran}(\tilde{V}_k)$.
Let $\tilde{Y}$ be a Ritz space of $A + E_k$ in $\tilde{K}_k$.
Let $\tilde{M}$ be the Ritz values corresponding to $\tilde{Y}$ and let $\tilde{M}_{-}$ be the set of $k - \text{dim}(\tilde{Y})$ remaining Ritz values.

Theorem 4.15. If Assumption 4.14 is satisfied and $\|E_k\|_2 < \text{gap}(\tilde{M}_{-}, \Lambda_2)$ then
\[
\angle_{\max}(X_1, \tilde{Y}) \leq \delta_{1,2} + \arcsin_{\leq 1}\left(\sqrt{1 + \frac{\pi^2 \|b_{k+1}\|^2}{4(\text{gap}(\tilde{M}_{-}, \Lambda_2) - \|E_k\|^2)} \sin \angle_{\max}(\tilde{X}_2, \tilde{K}_k)}\right)
\] (4.13)

where $\arcsin_{\leq 1}(x) := \arcsin(\min\{1, x\})$.

Proof. By Lemma 2.10.i) we have
\[
\angle_{\max}(X_1, \tilde{Y}) \leq \angle_{\max}(X_1, \tilde{X}_2) + \angle_{\max}(\tilde{X}_2, \tilde{Y}).
\] (4.14)

Now, we treat each term of the right-hand side separately. By using Theorem 4.1 ii) and the definition of $\delta_{1,2}$ given in (4.4) the first angle on the right-hand side is bounded by
\[
\angle_{\max}(X_1, \tilde{X}_2) \leq \arctan\left(\frac{\|E_k\|_2}{\text{gap}(\Lambda_1, \Lambda_{-2}) - 2\|E_k\|_2}\right) = \delta_{1,2}.
\] (4.15)
Applying Theorem 2.15 to \( \alpha_{\max}(\tilde{X}_2, \tilde{Y}) \) results in

\[
\alpha_{\max}(\tilde{X}_2, \tilde{Y}) \leq \arcsin \left( 1 + \frac{\pi^2 \|b_{k+1}\|^2}{4 \left( \text{gap}(\tilde{\Lambda}_-, \tilde{\Lambda}_2) - \|E_k\|^2 \right)^2} \sin \alpha_{\max}(\tilde{X}_2, \tilde{K}_k) \right),
\]

where \( P \) is the orthogonal projector onto \( \tilde{K}_k \) and \( \tilde{Y} \) is the orthogonal complement of \( \tilde{Y} \) in \( \tilde{K}_k \). Choosing \( \tilde{V}_\perp \in \mathbb{C}^{n \times n-k} \) such that \( |\tilde{V}_k, \tilde{V}_\perp| \) is unitary, the numerator \( \|P(A + E_k)(I - P)\|_2^2 \) can be rewritten as

\[
\|P(A + E_k)(I - P)\|_2^2 = \|V_\perp^H(A + E_k)V_\perp\|_2^2 = \|V_\perp^H(A + E_k)k_{V_k}\|_2^2 = \|V_\perp^Hk_{V_k} + V_\perp^Hv_{k+1}b_{k+1}^H\|_2^2 = \|b_{k+1}\|^2, \tag{4.17}
\]

by using relation (4.1) and the fact that \( A + E_k \) is Hermitian. Moreover, analyzing the denominator \( \text{sep}_{A+E_k,2}(\tilde{Y}_\perp, \tilde{X}_2)^2 \), we obtain from Lemma 2.6 that

\[
\text{sep}_{A+E_k,2}(\tilde{Y}_\perp, \tilde{X}_2)^2 \geq \frac{4}{\pi^2} \text{gap}(\tilde{\Lambda}_-, \tilde{\Lambda}_2)^2 \geq \frac{4}{\pi^2} \left( \text{gap}(\tilde{\Lambda}_-, \tilde{\Lambda}_2) - \|E_k\|^2 \right)^2,
\]

with \( \tilde{\Lambda}_2 := \{\lambda_j(A + E_k) : j \in J_2\} \) where \( J_2 \) is defined as in Theorem 4.4. Inserting the above approximation and (4.17) into (4.16) we get

\[
\alpha_{\max}(\tilde{X}_2, \tilde{Y}) \leq \arcsin \left( 1 + \frac{\pi^2 \|b_{k+1}\|^2}{4 \left( \text{gap}(\tilde{\Lambda}_-, \tilde{\Lambda}_2) - \|E_k\|^2 \right)^2} \sin \alpha_{\max}(\tilde{X}_2, \tilde{K}_k) \right). \tag{4.18}
\]

Finally, inserting (4.15) and (4.18) into (4.14) completes the proof. \( \square \)

Traditional bounds, like the one presented in Theorem 2.15, are particularly sensitive regarding the separation of the desired and remaining eigenvalues. This restriction was relaxed in Theorem 4.15.

**Remark 4.16.** In order to deal with small gaps between the desired and remaining eigenvalues, we allow the Ritz space \( \tilde{Y} \) to be of larger dimension than the invariant subspace \( \tilde{X}_1 \) of interest.

Since a bound on the angle of the right hand side of (4.18) was derived in Theorem 4.4 the following remark provides a combination of Theorem 4.15 and 4.4.

**Remark 4.17.** Combining Theorem 4.15 and 4.4 as well as using that \( \sin(\alpha) < \tan(\alpha) \) for \( \alpha \in [0, \frac{\pi}{2}] \) results in

\[
\sin \alpha_{\max}(\tilde{X}_1, \tilde{Y}) \leq \sqrt{1 + \frac{\pi^2 \|b_{k+1}\|^2}{4 \left( \text{gap}(\tilde{\Lambda}_-, \tilde{\Lambda}_2) - \|E_k\|^2 \right)^2} \sin \alpha_{\max}(\tilde{X}_2, \tilde{K}_k)}
\]
4.2 Convergence of the invariant subspace

\[
\leq \sqrt{1 + \frac{\pi^2 \|b_{k+1}\|^2}{4 \left( \text{gap}(\mathcal{M}_-, \Lambda_2) - \|E_k\|_2 \right)^2} \tan \angle_{\max}(\tilde{X}_2, \tilde{K}_k)}
\]

\[
\leq \sqrt{1 + \frac{\pi^2 \|b_{k+1}\|^2}{4 \left( \text{gap}(\mathcal{M}_-, \Lambda_2) - \|E_k\|_2 \right)^2}}.
\]

\[
\varrho_{k, \ell} \tan \frac{\pi}{2} \angle_{\max}(\tilde{X}_4, \tilde{K}_\ell) + \delta_{3,4}
\]

where the constants \( \varrho_{k, \ell} \) and \( \delta_{3,4} \) are defined in Theorem 4.4.

Consequently, in comparison to Theorem 4.15 less information is needed for this bound. To be more precise, the angle \( \angle_{\max}(\tilde{X}_1, \tilde{Y}) \) is bounded in terms of \( \angle_{\max}(\tilde{X}_4, \tilde{K}_\ell) \), i.e., only the Krylov subspace after \( \ell \) steps, with \( \ell \ll k \), is needed. Nevertheless, in comparison to (4.18) the bound derived in Remark 4.17 is less precise.

Using that the search spaces contain an increasingly accurate approximation of the desired invariant subspace, Theorem 4.15 ensures the convergence of the Ritz space to an eigenspace, however it does not tell us when to stop the iteration. Therefore, in the following we discuss bounds for the approximation quality of Ritz spaces in terms of their residual. Again we are in an a posteriori setting, i.e., search space and residual are explicitly available. Moreover, the search space is no longer restricted to a Krylov space, but is assumed to be an arbitrary subspace of \( \mathbb{C}^n \). As before, by looking at Ritz spaces of the matrix \( A + E_k \) instead of \( A \), we consider a setting with inexact operations. Further, we allow small gaps between the desired and remaining eigenvalues.

The following considerations are divided into two parts. First we investigate the approximation quality if the Ritz space \( \tilde{Y} \) is of the same dimension like the invariant subspace \( X \) of \( A \) and second we analyze the situation where the Ritz space can be of larger dimension than the invariant subspace. To specify the setting in the case of equal dimensions we provide an illustration of the spectral sets involved in Theorem 4.19 in Figure 4.3. However, because it is difficult to present the relation of subspaces to each other we depict the situation of the corresponding eigen- and Ritz values instead.
4.2 Convergence of the invariant subspace

Assumption 4.18. Let \( A \in \mathbb{C}^{n \times n} \) be Hermitian with eigenvalues \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \).

Let \( J_A \subset \{1, 2, \ldots, n\} \) and \( X \) be a \(|J_A|\)-dimensional invariant subspace of \( A \) corresponding to the eigenvalues \( \Lambda := \{\lambda_j : j \in J_A\} \subset \text{eig}(A) \).

Let \( \Lambda_- := \{\lambda_j : j \in \{1, \ldots, n\} \setminus J_A\} \) denote the remaining eigenvalues of \( A \).

Let \( E_k \in \mathbb{C}^{n \times n} \) be Hermitian and let \( V_k \) be a \( k \)-dimensional subspace in \( \mathbb{C}^n \).

Let \( \tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \ldots \geq \tilde{\mu}_k \) be the Ritz values of \( A + E_k \) with respect to \( V_k \).

Let the elements of \( J_{\tilde{\mathcal{M}}} \) be consecutive integers in \( \{1, 2, \ldots, k - 1\} \), \( \tilde{\mathcal{M}} := \{\tilde{\mu}_j : j \in J_{\tilde{\mathcal{M}}}\} \) and \( \tilde{Y} \) be the Ritz space of \( A + E_k \) in \( V_k \) corresponding to the Ritz values in \( \tilde{\mathcal{M}} \).

Let the columns of \( \tilde{Y} \in \mathbb{C}^{n \times |J_{\tilde{\mathcal{M}}}|} \) form an orthonormal basis of \( \tilde{Y} \) and

\[
\tilde{R} := (A + E_k)\tilde{Y} - \tilde{Y}\tilde{M} \quad \text{with} \quad \tilde{M} := \tilde{Y}H(A + E_k)\tilde{Y}.
\]

Let \( \dim(X) = \dim(\tilde{Y}) \).

Theorem 4.19. If Assumption 4.18 holds and \( \|E_k\|_2 + \|\tilde{R}\|_2 < \text{gap}(\text{env}(\tilde{\mathcal{M}}), \Lambda_-) \), then we have

\[
\tan \Theta_{\text{max}}(X, \tilde{Y}) \leq \frac{\|E_k\|_2 + \|\tilde{R}\|_2}{\text{gap}(\tilde{\mathcal{M}}, \Lambda_-) - \|E_k\|_2 - \|\tilde{R}\|_2} \quad \text{and}
\]

\[
\|\text{diag}(\lambda_1, \ldots, \lambda_m) - \text{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_m)\|_* \leq \frac{\|\tilde{R}\|_2 \|\tilde{R}\|_*}{\text{gap}(\tilde{\mathcal{M}}, \Lambda_-) - \|E_k\|_2 - \|\tilde{R}\|_2} + \|E_k\|_*
\]

for \( * \in \{2, F\} \).

Proof. Since \( \tilde{Y} \) is a Ritz space in \( V_k \) and \( \tilde{Y} \subset V_k \), we have \( \tilde{R}H\tilde{Y} = 0 \). Further, with \( \tilde{E} := -\tilde{Y}\tilde{R}H - \tilde{Y}\tilde{R}H \), we get

\[
(A + E_k + \tilde{E})\tilde{Y} = (A + E_k)\tilde{Y} + (-\tilde{R}H - \tilde{Y}\tilde{R}H)\tilde{Y}
\]

106
4.2 Convergence of the invariant subspace

\[ = \hat{Y}M - \hat{Y}R^H\hat{Y} \]
\[ = \hat{Y}M. \]

Hence \( \hat{Y} \) is an invariant subspace of \( A + E_k + \hat{E} \) corresponding to the eigenvalues eig(\( \hat{M} \)) = \( \hat{M} \subset \) eig(\( A + E_k + \hat{E} \)). Note that \( \| \hat{E} \|_2 = \| \hat{Y}R^H + \hat{Y} \hat{Y}^H \|_2 = \| \hat{R} \|_2 \), because \( \hat{R}^H \hat{Y} = 0 \). By the assumed bound \( \| E_k \|_2 + \| \hat{R} \|_2 < \text{gap}(\text{env}(\hat{M}), \Lambda_-) \) Theorem 4.1.i) is applicable and together with the symmetry property of the maximum angle between subspaces of equal dimensions we obtain

\[
\tan \angle_{\text{max}}(\mathcal{X}, \hat{Y}) = \tan \angle_{\text{max}}(\hat{Y}, \mathcal{X}) \leq \frac{\| E_k + \hat{E} \|_2}{\text{gap}(\hat{M}, \Lambda_-) - \| E_k \|_2}. \]

With \( \| E_k + \hat{E} \|_2 \leq \| E_k \|_2 + \| \hat{E} \|_2 = \| E_k \|_2 + \| \hat{R} \|_2 \) this results in bound (4.19).

In order to prove bound (4.20) the eigenvalues of \( A + E_k \) are assumed to be \( \bar{\lambda}_1 \geq \ldots \geq \bar{\lambda}_n \). Let \( \mathcal{X} \) be an invariant subspace of \( A + E_k \) corresponding to the eigenvalues \( \{ \bar{\lambda}_j : j \in I_A \} \) and the remaining eigenvalues are defined by \( \bar{\lambda}_- := \{ \bar{\lambda}_j : j \in \{1, \ldots, n \} \setminus I_A \} \). It follows from Theorem 2.1 that

\[
\text{gap}(\hat{M}, \bar{\lambda}_-) \geq \text{gap}(\hat{M}, \Lambda_-) - \| E_k \|_2. \]

Thus, Theorem 4.1.i) is applicable and yields

\[
\tan \angle_{\text{max}}(\mathcal{X}, \hat{Y}) \leq \frac{\| \hat{E} \|_2}{\text{gap}(\hat{M}, \Lambda_-) - \| \hat{E} \|_2} \leq \frac{\| \hat{E} \|_2}{\text{gap}(\hat{M}, \Lambda_-) - \| \hat{E} \|_2 - \| E_k \|_2}. \]

Together with Theorem 2.2 this implies that

\[
\| \text{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_m) - \text{diag}(\bar{\mu}_1, \ldots, \bar{\mu}_m) \|_* \leq \tan \angle_{\text{max}}(\mathcal{X}, \hat{Y}) \| \hat{R} \|_2 \leq \frac{\| \hat{R} \|_2 \| \hat{R} \|_*}{\text{gap}(\hat{M}, \Lambda_-) - \| E_k \|_2 - \| \hat{R} \|_2}. \]

(4.21)

Moreover, it follows from Theorem 2.1 that

\[
\| \text{diag}(\lambda_1, \ldots, \lambda_m) - \text{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_m) \|_* \leq \| E_k \|_* \tag{4.22} \]

Inserting bound (4.21) and (4.22) into the triangle inequality

\[
\| \text{diag}(\lambda_1, \ldots, \lambda_m) - \text{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_m) \|_* \leq \| \text{diag}(\lambda_1, \ldots, \lambda_m) - \text{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_m) \|_* + \| \text{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_m) - \text{diag}(\bar{\mu}_1, \ldots, \bar{\mu}_m) \|_* \]

proves (4.20).
4.2 Convergence of the invariant subspace

\[ \mathcal{M}_\varepsilon \]

\[ \Lambda \]

\[ \mathcal{M}_\perp \]

\[ \mathcal{M} \]

\[ \mathcal{M}_\perp \]

- eigenvalues of \( A \)
- eigenvalues of \( A + E_k + \tilde{E}_\varepsilon \)
- Ritz values of \( A + E_k \)

Figure 4.4: Illustration of the spectral sets of Theorem 4.21

In the next step we generalize Theorem 4.19 by allowing the dimension of the approximated subspace to be larger than the dimension of the exact subspace. More precisely, we analyze how well the invariant subspace \( X \) is contained in the Ritz space \( \tilde{Y} \) of possibly larger dimension. To illustrate this spectral situation and to clarify the notation of Theorem 4.21, we show in Figure 4.4 the relation between the eigen- and Ritz values considered in Theorem 4.21. Note that the eigenvalues of \( A + E_k + \tilde{E}_\varepsilon \) are illustrated additionally in Figure 4.4, since they appear in the proof of Theorem 4.21.

**Assumption 4.20.** Let \( A \in \mathbb{C}^{n \times n} \) be Hermitian with eigenvalues \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \). Let \( I_A \subset \{1, 2, \ldots, n\} \) and \( \mathcal{X} \) be a \( |I_A| \)-dimensional invariant subspace of \( A \) corresponding to the eigenvalues \( \Lambda := \{\lambda_j : j \in I_A\} \subset \text{eig}(A) \).

Let the elements of \( I_{\tilde{\mathcal{M}}} \) be consecutive in \( \{1, 2, \ldots, k - 1\} \).

Let \( E_k \in \mathbb{C}^{n \times n} \) be Hermitian and \( \mathcal{V}_k \) be a \( k \)-dimensional subspace in \( \mathbb{C}^n \).

Let \( \tilde{\mathcal{M}} := \{\tilde{\mu}_j : j \in I_{\tilde{\mathcal{M}}}\} \) and \( \tilde{Y} \) be the Ritz space of \( A + E_k \) in \( \mathcal{V}_k \) corresponding to \( \tilde{\mathcal{M}} \) with \( \dim(\mathcal{X}) \leq \dim(\tilde{Y}) \).

Let \( \tilde{\mathcal{M}}_n := \{\tilde{\mu}_{\min(I_{\tilde{\mathcal{M}}}) - 1}, \tilde{\mu}_{\max(I_{\tilde{\mathcal{M}}}) + 1}\} \) be the neighboring Ritz values of \( \tilde{\mathcal{M}} \) (with \( \tilde{\mu}_0 := +\infty \) if needed), \( \tilde{\mathcal{M}}_\varepsilon := \tilde{\mathcal{M}} \cup \tilde{\mathcal{M}}_n \), and \( \tilde{Y}_\varepsilon \supset \tilde{Y} \) be a Ritz space of \( A + E_k \) in \( \mathcal{V}_k \) corresponding to the finite values in \( \tilde{\mathcal{M}}_\varepsilon \).

Let the columns of \( \tilde{Y}_\varepsilon \) form an orthonormal basis of \( \tilde{Y}_\varepsilon \) such that

\[ \tilde{R}_\varepsilon := (A + E_k)\tilde{Y}_\varepsilon - \tilde{Y}_\varepsilon \tilde{M}_\varepsilon \quad \text{with} \quad \tilde{M}_\varepsilon := \tilde{Y}_\varepsilon^H (A + E_k)\tilde{Y}_\varepsilon \]

and \( \tilde{E}_\varepsilon := -\tilde{Y}_\varepsilon \tilde{R}_\varepsilon^H - \tilde{R}_\varepsilon \tilde{Y}_\varepsilon^H \).

**Theorem 4.21.** If Assumption 4.20 holds, \( \|E_k\|_2 + \|\tilde{E}_\varepsilon\|_2 < \text{gap}(\Lambda, \tilde{\mathcal{M}}_n) \)

\ \text{env}(\Lambda) \subset \text{env}(\tilde{\mathcal{M}}_\varepsilon) \text{ and all eigenvalues of } A + (E_k + \tilde{E}_\varepsilon) \text{ in } \text{env}(\tilde{\mathcal{M}}_\varepsilon) \text{ are in}
4.2 Convergence of the invariant subspace

\begin{align*}
\mathcal{M}_e, \text{ then} \\
\tan \angle_{\max} (\mathcal{X}, \mathcal{Y}) \leq \frac{\|E_k\|_2 + \|\hat{R}_e\|_2}{\text{gap}(\Lambda, \mathcal{M}_n) - \|E_k\|_2 - \|\hat{R}_e\|_2}.
\end{align*}

(4.23)

Proof. Let \( \hat{R}_n := (A + E_k)\hat{Y}_n - \hat{Y}_n\hat{\mathcal{M}}_n \) and \( \hat{E}_n := -\hat{Y}_n\hat{R}_n^H - \hat{R}_n\hat{Y}_n^H \). Analogously to the derivation for \( \mathcal{Y} \) in the prove of Theorem 4.19 it can be shown that \( \hat{Y}_e \) and \( \hat{\mathcal{M}}_e \) are invariant subspaces of \( A + E_k + \hat{E}_e \) and \( A + E_k + \hat{E}_n \) corresponding to \( \hat{\mathcal{M}}_e \) and \( \hat{\mathcal{M}}_n \), respectively. Note that \( \|\hat{E}_e\|_2 = \|\hat{Y}_e\hat{R}_e^H + \hat{R}_e\hat{Y}_e^H\|_2 = \|\hat{R}_e\|_2 \), because \( \hat{R}_e^H\hat{Y}_e = 0 \). Hence, with the assumption \( \|E_k\|_2 + \|\hat{R}_e\|_2 < \text{gap}(\Lambda, \mathcal{M}_n) \) it follows that

\begin{align*}
\|E_k + \hat{E}_e\|_2 \leq \|E\|_2 + \|\hat{R}_e\|_2 < \text{gap}(\Lambda, \mathcal{M}_n).
\end{align*}

(4.24)

Since all eigenvalues of \( A + E_k + \hat{E}_n \) in \( \text{env}(\mathcal{M}_e) \) are in \( \mathcal{M}_e \) and \( \text{env}(\Lambda) \subset \text{env}(\mathcal{M}_e) \) we obtain \( \text{gap}(\Lambda, \mathcal{M}_-) = \text{gap}(\Lambda, \mathcal{M}_n) \), where \( \mathcal{M}_- := \text{eig}(A + E_k + \hat{E}_e) \setminus \hat{\mathcal{M}} \). Hence, together with (4.24) we obtain

\begin{align*}
\|E_k + \hat{E}_e\|_2 \leq \text{gap}(\Lambda, \mathcal{M}_-),
\end{align*}

which ensures the applicability of Theorem 4.1.i). This guarantees that

\begin{align*}
\tan \angle_{\max} (\mathcal{X}, \mathcal{Y}) \leq \frac{\|E_k + \hat{E}_e\|_2}{\text{gap}(\Lambda, \mathcal{M}_-) - \|E\|_2 - \|\hat{R}_e\|_2} \leq \frac{\|E_k\|_2 + \|\hat{E}_e\|_2}{\text{gap}(\Lambda, \mathcal{M}_n) - \|E_k\|_2 - \|\hat{E}_e\|_2}.
\end{align*}

Finally, using \( \|\hat{E}_e\|_2 = \|\hat{R}_e\|_2 \) concludes the proof of (4.23).

Comparing Theorem 4.19 and 4.21 shows that both a posteriori spectral error bounds depend only on the norm of \( E_k \) and the corresponding residual norm as well as the exact eigenvalues of \( A \), and Ritz values of \( A + E_k \). The condition \( \|E_k\|_2 + \|\hat{R}\|_2 < \text{gap}(\text{env}(\mathcal{M}), \Lambda_-) \) of Theorem 4.19 on the distribution of the eigenvalues and Ritz values is much more restrictive than the corresponding condition \( \|E_k\|_2 + \|\hat{R}_e\|_2 < \text{gap}(\Lambda, \mathcal{M}_n) \) of Theorem 4.21. If the gap between the computed Ritz values \( \mathcal{M} \) and the remaining eigenvalues \( \Lambda_- \) is too small, Theorem 4.19 is not applicable. However, the bound of Theorem 4.21 then depends on the gap between the desired eigenvalues and the neighbors of the Ritz values \( \mathcal{M}_n \). Thus, this gap can be increased by choosing a larger number of computed Ritz values. This is verified by the Experiments 5 and 8 in Section 4.3.

The following remarks provide some more insight into the Theorems 4.19 and 4.21 and point out useful features.

Remark 4.22. The effects of the residuals \( \hat{R} \) and \( \hat{R}_e \) can be interpreted as the effects of the backward error by \( \hat{E} \) and \( \hat{E}_e \), respectively, as it has been done in earlier works, cf. [111].

Remark 4.23. The condition, that all eigenvalues of \( A + (E_k + \hat{E}_e) \) in \( \text{env}(\mathcal{M}_e) \) are in \( \mathcal{M}_e \) means that for every desired eigenvalue there is a corresponding Ritz value, i.e. the Krylov method "does not miss" a desired eigenvalue.
Remark 4.24. If the search space $V_k$ is a Krylov space and a Krylov relation is available then the norm residuals $\tilde{R}$ and $\tilde{R}_e$ can be bounded by Lemma 2.18.

Remark 4.25. Let $\tilde{Y}_l$ be a Ritz space of $A + E_k$ in $\tilde{K}_l := K(A + E_k, \tilde{v}_1)$ for some Hermitian $E_k \in \mathbb{C}^{n \times n}$ and some $\tilde{v}_1 \in \mathbb{C}^n$. Because of $\tilde{Y}_l \subset \tilde{K}_l$, we obtain

$$\zeta_{\text{max}}(\mathcal{X}, \tilde{K}_l) \leq \zeta_{\text{max}}(\mathcal{X}, \tilde{Y}_l).$$

Hence, an additional a posteriori bound of $\zeta_{\text{max}}(\mathcal{X}, \tilde{Y}_l)$ could be used to bound the right-hand side of the a priori bound (4.3).

Note that regarding the amount of necessary information there is a major difference between the results of Section 4.1 and 4.2, which is highlighted in the following remark.

Remark 4.26. Theorem 4.4 is an a priori result as it only includes the information of the $\ell$-th step (with $\ell < k$). In contrast, the Theorems 4.15, 4.19 and 4.21 are a posteriori results, as the vector $b_{k+1}$, the Ritz values $\tilde{M}$ and the residual $\tilde{R}$ at the $k$-th step are required. Of course it would be of interest to derive also an a priori result for the distance between the subspace of interest $\mathcal{X}_1$ and the Ritz space $\tilde{Y}$, but so far it is unclear how this result could look like.

4.3 Numerical results

In this section we verify the spectral error bounds presented in the previous Sections 4.1 and 4.2 by analyzing two different examples. The first, academic example uses a diagonal test matrix. It is constructed to evaluate how changes in the eigenvalue distribution influence the spectral bounds. The second example analyzes a one-dimensional Ising model that describes the dynamics of quantum phase transitions, cf. [21, 77]. The choice of this model is motivated by the distribution of the eigenvalues in the Ising model. The ground state and the first excited states (i.e. the eigenvalues of interest) are not well separated from the remaining spectrum. This situation characterizes a special case of the spectral error bounds presented in this chapter.

The spectral error bounds presented in Sections 4.1 and 4.2 are based on perturbed Krylov subspaces whereby the subspaces result from a Krylov relation of the form (4.1). To obtain this relation we apply in the following numerical examples an exact Arnoldi method to a perturbed matrix $A + E_k$, where $E_k$ is chosen as a Hermitian, random matrix of prescribed norm.

4.3.1 Academic example

Our first test matrix is a diagonal matrix $A$ built by the MATLAB command

$$A = \text{diag} \left([9,9-g,7,6,5,4,4-g,2,\text{rand}(1,992)])\right),$$

(4.25)
4.3 Numerical results

where \( g \in [0, 2] \) is a parameter.

The matrix is of dimension \( 1000 \times 1000 \) and possesses eight eigenvalues larger than 1 (to be precise, \( 9, 9 - g, 7, 6, 5, 4, 4 - g, 2 \)) and 992 eigenvalues in the interval \((0, 1)\). We are interested in an eigenspace corresponding to the eigenvalues \( \Lambda_1 = \{9 - g, 7, 6, 5, 4\} \) because for these the parameter \( g \) is equivalent to the gap between desired and remaining eigenvalues. Consequently, for small values of \( g \) the separation between \( \Lambda_1 \) and the remaining spectrum of \( A \) is poor.

**Experiment 1. (Illustrating the influence of \( \|E_k\|_2 \) on the result of Theorem 4.4)** To explore the influence of the norm of the perturbation \( \|E_k\|_2 \) on the bound derived in Theorem 4.4, we analyze the development of the angle of inclusion \( \angle_{\max}(X_1, \overline{K}_k) \) and its bound (4.3) over the course of 45 Arnoldi iteration steps. We choose that the dimension \( \ell \) of the Krylov subspace on the right-hand side of (4.3) equals 10 and take \( g = 1 \) in (4.25), i.e., the gap \((\Lambda_1, \Lambda_{-1})\) is large and the desired eigenvalues in \( \Lambda_1 \) are well separated from the remaining ones. In the following considerations we distinguish between two cases

(i) \( \Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4 \) and (ii) \( \Lambda_1 = \Lambda_2 \subset \Lambda_3 = \Lambda_4 \).

(i) \( \Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4 \) For \( \|E_k\|_2 \in \{10^{-4}\|A\|_2, 10^{-7}\|A\|_2, 10^{-11}\|A\|_2\} \) the angle of inclusion \( \angle_{\max}(X_1, \overline{K}_k) \) is plotted in Figure 4.5 as dotted, dashed, and solid lines without marks, respectively. Note that, although we are interested in eigenspaces of \( A \) we search in Krylov subspaces of \( A + E_k \) instead of \( A \).

![Figure 4.5: Experiment 1 (case (i)): Angle of inclusion \( \angle_{\max}(X_1, \overline{K}_k) \) and bound (4.3) with \( \|E_k\|_2 \in \{10^{-4}\|A\|_2, 10^{-7}\|A\|_2, 10^{-11}\|A\|_2\} \) ](image)

We see in Figure 4.5 that the method converges, i.e., the angle of inclusion \( \angle_{\max}(X_1, \overline{K}_k) \) decreases during the iteration. After an initial phase, the angle starts to converge linearly with a convergence rate independent of \( \|E_k\|_2 \). How-
4.3 Numerical results

ever, the angle \( \angle_{\text{max}}(X_1, K_k) \) decreases not to zero but only to a limiting accuracy that is of the order of \( \|E_k\|_2 / \|A\|_2 \). The lines in Figure 4.5 with marks depict the bound of Theorem 4.4 for each \( \|E_k\|_2 \in \{10^{-4} \|A\|_2, 10^{-7} \|A\|_2, 10^{-11} \|A\|_2 \} \). We see that the curves of the bounds reflect the behavior of the angles and we can identify the different convergence phases mentioned in Remark 4.7. In more detail, the bound (4.3) starts to exist after the initialization phase and correctly predicts the convergence and stagnation phase. In the stagnation phase the bound overestimates the limiting accuracy only by a value of about 6. However, during the convergence phase the true value converges faster than the bound. Defining the convergence rate \( \alpha \) by the following relation

\[
\angle_{\text{max}}(X_1, K_{k+1}) \leq \alpha \cdot \angle_{\text{max}}(X_1, K_k).
\]

reveals a “true” rate \( \alpha \) of 0.12 of the angle \( \angle_{\text{max}}(X_1, K) \) during the convergence phase. The bound (4.3) overestimates this convergence rate to be 0.41. Both values are determined from the observed values. The corresponding slow down factor \( \log(0.12) / \log(0.41) \approx 2.5 \) indicates that reducing the angle by a certain amount takes two and a half times less iterations than predicted by the bound.

The convergence rate of bound (4.3) is determined by \( \bar{\eta} \) that is defined in (4.7). According to Remark 4.8 the constant \( \bar{\eta} \) can be improved by extending the subset \( J_3 \) with respect to \( J_2 \), which is now analyzed in case (ii).

(ii) \( \Lambda_1 = \Lambda_2 \subset \Lambda_3 = \Lambda_4 \) Here we choose the setting \( \Lambda_1 = \Lambda_2 \) and \( \Lambda_3 = \Lambda_4 = \Lambda_1 \cup \{9, 4 - 8, 2\} \). Thus, we consider a large gap between \( \Lambda_2 \) and \( \Lambda_{-3} \).

![Figure 4.6: Experiment 1 (case (ii)): Angle of inclusion \( \angle_{\text{max}}(X_1, K) \) and bound (4.3) with \( \|E_k\|_2 \in \{10^{-4} \|A\|_2, 10^{-7} \|A\|_2, 10^{-11} \|A\|_2 \} \)](image)

We see in Figure 4.6 that during the convergence phase the convergence rate of the true value and the bound almost coincide. More precisely, the convergence
4.3 Numerical results

rate $\alpha$ of the true angle amounts to 0.11 and the convergence rate of the bound to 0.12. Therefore, the adjustment of the subsets leads to a significant improvement of the bound during the convergence phase.

Experiment 2. (Step by step adjustment of the nested subspaces to improve the bound (4.3)) As stated in Remark 4.5 part iv) Theorem 4.4 is meaningful even for a small gap between $\Lambda_1$ and the rest of the spectrum if the appropriately nested subspaces $\Lambda_2, \Lambda_3$ and $\Lambda_4$ are chosen. To illustrate how the choice of the nested subspaces influences the bound we distinguish in this experiment between three cases

(i) $\Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4,$

(ii) $\Lambda_1 \subset \Lambda_2 = \Lambda_3 = \Lambda_4,$

(iii) $\Lambda_1 \subset \Lambda_2 \subset \Lambda_3 = \Lambda_4.$

We are interested in the eigenvalues $\Lambda_1 = \{9 - g, 7, 6, 5, 4\}$ and vary the parameter $g$ in order to analyze the sensitivity of bound (4.3) regarding the gap between the desired eigenvalues $\Lambda_1$ and the remaining ones for these three nested subspace constellations. We choose $g \in \{10^{-2} \| E_k \|_2, 2.1 \| E_k \|_2, 10^2 \| E_k \|_2, 10^4 \| E_k \|_2\}$ and as before $g$ equals the gap between the desired and remaining eigenvalues. Further, we choose $\| E_k \|_2 = 10^{-11} \| A \|_2$ and consider the case where 20 steps of the inexact Krylov subspace method have been already carried out, i.e., $\ell = 20.$

(i) $\Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4$ In the case of equal subspaces bound (4.3) can be expected to be very loose for a small gap $g.$ This is confirmed by Figure 4.7, where the true angle $\chi_{\max}(\chi_1, \tilde{K}_k)$ and the bound (4.3) are depicted for $g \in \{2.1 \| E_k \|_2, 10^2 \| E_k \|_2, 10^4 \| E_k \|_2\}.$ We did not consider $g = 10^{-2} \| E_k \|_2$ because in this situation the assumptions of Theorem 4.4 are not fulfilled and the bound (4.3) does not exist.

Note, as we consider three different matrices we would have to plot three different “true” angles $\chi_{\max}(\chi_1, \tilde{K}_k).$ However, apart from a minimal deviation in the starting point of the convergence phase by two or three iteration steps, these three “true” angles behave in the same way. Therefore, only the “true” angle for $g = 2.1 \| E_k \|_2$ is shown in Figure 4.7.

Figure 4.7 illustrates that the smaller the gap between desired and remaining eigenvalues, $g,$ the less meaningful the bound (4.3) becomes. This result is also reflected in the first part of Table 4.1 (rows indicated by (i)). We see that the smaller the gap $g$ the larger the constants $\delta_{1,2}, \delta_{3,4}$ and $\tilde{\delta}_{\max}$ and the smaller $\tilde{\eta}_{\min},$ i.e., the less significant bound (4.3) becomes. The subscript max on $\tilde{\delta}$ and min on $\tilde{\eta}$ in the header of Table 4.1 indicates that the maximum of $\tilde{\theta}_i$ for $i \in \mathcal{J}_2$ and the minimum of $\tilde{\eta}_i$ for $i \in \mathcal{J}_2$ is stated.
4.3 Numerical results

<table>
<thead>
<tr>
<th>case</th>
<th>$g$ in (4.25)</th>
<th>$\delta_{1,2}$ (lower is better)</th>
<th>$\delta_{3,4}$ (lower is better)</th>
<th>$\tilde{\theta}_{\text{max}}$ (lower is better)</th>
<th>$\tilde{\eta}_{\text{min}}$ (higher is better)</th>
<th>over-est. of lim. accuracy</th>
<th>slow down factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>$10^4|E_k|_2$</td>
<td>$10^{-4}$</td>
<td>$10^{-4}$</td>
<td>$9 \cdot 10^6$</td>
<td>$2 \cdot 10^{-7}$</td>
<td>0.33</td>
<td>1.96</td>
</tr>
<tr>
<td>(i)</td>
<td>$10^2|E_k|_2$</td>
<td>$10^{-2}$</td>
<td>$10^{-2}$</td>
<td>$9 \cdot 10^8$</td>
<td>$2 \cdot 10^{-9}$</td>
<td>0.40</td>
<td>2.41</td>
</tr>
<tr>
<td>(i)</td>
<td>$2.1|E_k|_2$</td>
<td>1.47</td>
<td>1.47</td>
<td>$9 \cdot 10^{11}$</td>
<td>$2 \cdot 10^{-12}$</td>
<td>1.00</td>
<td>$1.5 \cdot 10^5$</td>
</tr>
<tr>
<td>(ii)</td>
<td>$2.1|E_k|_2$</td>
<td>$5 \cdot 10^{-11}$</td>
<td>$5 \cdot 10^{-11}$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.23</td>
<td>1.47</td>
</tr>
<tr>
<td>(ii)</td>
<td>$10^{-2}|E_k|_2$</td>
<td>$5 \cdot 10^{-11}$</td>
<td>$5 \cdot 10^{-11}$</td>
<td>1.00</td>
<td>1.00</td>
<td>0.23</td>
<td>1.49</td>
</tr>
<tr>
<td>(iii)</td>
<td>$10^{-2}|E_k|_2$</td>
<td>$5 \cdot 10^{-11}$</td>
<td>$9 \cdot 10^{-11}$</td>
<td>1.00</td>
<td>3.00</td>
<td>0.11</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 4.1: Experiment 2: Constants of the bound (4.3) for case (i) $\Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4$, case (ii) $\Lambda_1 \subset \Lambda_2 = \Lambda_3 = \Lambda_4$, and case (iii) $\Lambda_1 \subset \Lambda_2 \subset \Lambda_3 = \Lambda_4$ (the numbers are correct to leading digit).

The following two cases (ii) and (iii) illustrate that the sets of eigenvalues $\Lambda_2$, $\Lambda_3$ and $\Lambda_4$ can be chosen such that $\delta_{1,2}$, $\delta_{3,4}$, $\tilde{\theta}_{\text{max}}$ and $\tilde{\eta}_{\text{min}}$ in (4.3) are improved (see Remark 4.8).

(ii) $\Lambda_1 \subset \Lambda_2 = \Lambda_3 = \Lambda_4$ We consider again small gaps, i.e., $g \in \{2.1\|E_k\|_2, 10^{-2}\|E_k\|_2\}$, but now we choose $\Lambda_2 = \Lambda_3 = \Lambda_4 = \Lambda_1 \cup \{9, 4 - g\} \subset \Lambda_1$ resulting in a large gap between $\Lambda_1$ and $\Lambda_2$. The constants of bound (4.3) are stated in the rows indicated by (ii) in Table 4.1. We see that bound (4.3) is meaningful even for a matrix where the gap between desired and remaining eigenvalues, $g$, is smaller than the norm of the perturbation, i.e., for $g = 10^{-2}\|E_k\|_2$.

(iii) $\Lambda_1 \subset \Lambda_2 \subset \Lambda_3 = \Lambda_4$ Finally, bound (4.3) is further improved by choosing $\Lambda_3 = \Lambda_4 = \Lambda_2 \cup \{2\}$ and $\Lambda_1, \Lambda_2$ as in case (ii). According to Remark 4.8 this leads to an improved $\tilde{\eta}$ and thus to a more accurate convergence rate of the bound. In particular, we see that for $g = 10^{-2}\|E\|_2$ the convergence rates of the bound and of the true value almost coincide, i.e., the slow down factor is about 1.00.
4.3 Numerical results

**Figure 4.7:** Experiment 2: Angle of inclusion $\max(\mathcal{X}_1, \tilde{K}_k)$ and its bound (4.3) for $\varrho \in \{2.1\|E_k\|_2, 10^2\|E_k\|_2, 10^4\|E_k\|_2\}$, $\ell = 20$ and $\Lambda_2 = \Lambda_3 = \Lambda_4 = \Lambda_1$.

**Figure 4.8:** Experiment 3: Angle of inclusion $\max(\mathcal{X}_1, \tilde{K}_k)$ and its bound (4.3) for $\Lambda_2 = \Lambda_3 = \Lambda_4 \supset \Lambda_1$, $\varrho = 2.1\|E_k\|_2$, and $\ell \in \{15, 22, 30\}$.

Summarizing, the constants stated in the Table 4.1 improve substantially from case (i) to case (iii) and therefore lead to a much sharper bound (4.3).

**Experiment 3. (Illustration of the influence of $\ell$ on the result of Theorem 4.4)**

To investigate the influence of the number of already proceeded iteration steps $\ell$ on bound (4.3), we use the setting of Experiment 2 case (ii). In particular, we choose $\varrho = 10^{-2}\|E_k\|_2$ and $\Lambda_1 \subset \Lambda_2 = \Lambda_3 = \Lambda_4$ with $\Lambda_4 = \Lambda_3 = \Lambda_2 = \Lambda_1 \cup \{9, 4 - \varrho\}$.

For $\ell \in \{15, 22, 30\}$ already performed iteration steps Figure 4.8 shows the bound (4.3) and the corresponding true value $\max(\mathcal{X}_1, \tilde{K}_k)$. We observe that the bound predicts the actual behavior of $\max(\mathcal{X}_1, \tilde{K}_k)$, i.e., all three phases are recognizable. Further, we see that the larger the number of iteration steps $\ell$ that
have been already carried out, the sharper is the corresponding bound in the convergence phase. In this phase the convergence rate of the bound ($\approx 0.23$) is higher than the “true” rate ($\approx 0.11$), which amounts to a slow down factor of 1.47. During the stagnation phase, the bounds overestimate the actual value only by a constant factor of about 3 and this behavior is independent of the already proceeded iteration steps $\ell$.

**Experiment 4. (Illustrating the result of Theorem 4.15)** In this experiment we investigate how well the exact invariant subspace $X_1$ is included in the Ritz space $\hat{Y}$, i.e., we are interested in the quality of the computed subspace $\hat{Y}$.

The subspace $X_1$ corresponds to the eigenvalues $\Lambda_1 = \{9 - \delta, 7, 6, 5, 4\}$. We choose $\Lambda_2 = \Lambda_1 \cup \{9, 4 - \delta\}$, $\|E_k\|_2 = 10^{-13}\|A\|_2$ and $\delta = 10^{-11}$, i.e., the gap between $\Lambda_1$ and the rest of the spectrum is very small. The approximated eigenpair $(\hat{M}, \hat{Y})$ is obtained by calculating a Ritz pair of $A + E_k$. Theorem 4.15 allows the Ritz space $\hat{Y}$ to be of slightly larger dimension than the subspace $X_1$. Therefore we distinguish in this experiment between the two cases

(i) $\dim(X_1) = \dim(\hat{Y})$ and (ii) $\dim(X_1) < \dim(\hat{Y})$.

**Experiment 5. (Illustrating the results of the Theorems 4.19 and 4.21)** In this experiment we investigate the a posteriori error bounds of the Theorems 4.19 and 4.21, i.e., we are interested in the approximation quality of the Ritz space
4.3 Numerical results

\[ \text{Figure 4.9: Experiment 4: Angle of inclusion } \gamma_{\text{max}}(\mathcal{X}, \tilde{\mathcal{Y}}) \text{ and bound (4.13) for the two cases (i) } \dim(\mathcal{X}) = \dim(\tilde{\mathcal{Y}}) \text{ and (ii) } \dim(\mathcal{X}) < \dim(\tilde{\mathcal{Y}}) \]

The main difference between the two theorems is the dimension of the Ritz space. In Theorem 4.19 we assume that the Ritz space \( \tilde{\mathcal{Y}} \) and the invariant subspace \( \mathcal{X} \) are of equal dimension, whereas in Theorem 4.21 the Ritz space can be of larger dimension than \( \mathcal{X} \). In the following computations we choose \( \|E_k\|_2 = 10^{-10} \) and the desired eigenvalues \( \Lambda = \{9 - g, 7, 6, 5, 4\} \). The approximated eigenpair \((\tilde{\mathcal{M}}, \tilde{\mathcal{Y}})\) is obtained by calculating a Ritz pair of \( A + E_k \).

To verify the results of the Theorems 4.19 and 4.21 we distinguish between two settings. First we choose \( g = 1 \), i.e., we investigate a test matrix where the desired eigenvalues \( \Lambda \) are well separated from the rest of the spectrum and second we assume \( g = \|E_k\|_2 = 10^{-10} \), which implies a very small gap between \( \Lambda \) and the remaining eigenvalues of \( A \).

We start by evaluating the bound on the quality of the approximated eigenvalues given by (4.20). Assuming equal dimensions of the computed and desired subspace, we depict in Figure 4.10 for \( g = 1 \) bound (4.20) and the 2–norm of the distance between the eigenvalues contained in \( \Lambda \) and the computed Ritz values in \( \tilde{\mathcal{M}} \). We see that the Ritz values converge after 15 iteration steps. The bound (4.20) captures this behavior correctly but overestimates the actual value after convergence by a constant factor of 60. The plot looks similar for the Frobenius norm and therefore we omit this figure.

To analyze the quality of the computed subspace we consider both bounds (4.19) and (4.23). In Figure 4.11 we depict on the one hand for \( \dim(\mathcal{X}) = \dim(\tilde{\mathcal{Y}}) \) the angle \( \gamma_{\text{max}}(\mathcal{X}, \tilde{\mathcal{Y}}) \) and the corresponding bounds (4.19) and (4.23) and on the other hand for \( \dim(\mathcal{X}) < \dim(\tilde{\mathcal{Y}}) \) the angle \( \gamma_{\text{max}}(\mathcal{X}, \tilde{\mathcal{Y}}) \) and the
4.3 Numerical results

Figure 4.10: Experiment 5: Distance between the eigenvalues of interest $\Lambda$ and the corresponding Ritz values for $g = 1$.

corresponding bound (4.23).

In Figure 4.11(a) the situation for $g = 1$, and therefore for $\Lambda$ being well separated from the remaining eigenvalues, is illustrated. We see that the angle $\angle_{\text{max}}(X, \tilde{Y})$ for $\dim(X) = \dim(\tilde{Y})$ converges after 18 iterations and the corresponding bound (4.19), which is plotted by a solid line with boxes, predicts this behavior correctly. More precisely, after the convergence phase the bound overestimates the actual value by a constant factor of approximately 5. Furthermore, we see that for $g = 1$ and $\dim(X) = \dim(\tilde{Y})$ the bounds (4.19) and (4.23) almost coincide.

However, considering the situation for $g = \|E_k\|_2$, we see in Figure 4.11(b) that the angle $\angle_{\text{max}}(X, \tilde{Y})$ does not converge for $\dim(X) = \dim(\tilde{Y})$. Moreover, the bound (4.19) does not exist, because the condition $\|E_k\|_2 + \|\hat{R}\|_2 < \text{gap}(\text{env}(\tilde{M}), \Lambda_-)$ is violated.

To overcome this convergence problem we follow Theorem 4.21 and enlarge the Ritz space $\tilde{Y}$. The Ritz space corresponds now to the seven Ritz values that are closest to 6, such that the dimension of $X$ is less than the dimension of $\tilde{Y}$. This enlargement has almost no effect for $g = 1$ (see Figure 4.11(a)). For $\dim(X) < \dim(\tilde{Y})$ the angle $\angle_{\text{max}}(X, \tilde{Y})$ converges equally fast compared to the angle $\angle_{\text{max}}(X, \tilde{Y})$ for $\dim(X) = \dim(\tilde{Y})$. Further, bound (4.23) for $\dim(X) < \dim(\tilde{Y})$, which is depicted by the dashed line in Figure 4.11(a), decreases nearly to the limiting accuracy and overestimates the actual value only by a constant factor of 4. In particular, for $\dim(X) < \dim(\tilde{Y})$, the residual is smaller than for $\dim(X) = \dim(\tilde{Y})$, and consequently in the stagnation phase bound (4.23) is sharper than bound (4.19) for equal dimension.

The situation is different for $g = \|E_k\|_2$, where the angle $\angle_{\text{max}}(X, \tilde{Y})$ does not
4.3 Numerical results

Figure 4.11: Experiment 5: Angle of inclusion $\angle_{\text{max}}(\mathcal{X}, \tilde{\mathcal{Y}})$ and bounds (4.19) and (4.23)

converge for $\dim(\mathcal{X}) = \dim(\tilde{\mathcal{Y}})$. However, this changes for the enlarged Ritz space. In Figure 4.11(b) we see that for $\dim(\mathcal{X}) < \dim(\tilde{\mathcal{Y}})$ the angle of inclusion $\angle_{\text{max}}(\mathcal{X}, \tilde{\mathcal{Y}})$ decreases to the limiting accuracy after 26 iteration steps and the bound (4.23) for $\dim(\mathcal{X}) < \dim(\tilde{\mathcal{Y}})$ captures this behavior correctly. In contrast to bound (4.19) that depends on the gap between the Ritz values $\tilde{\mathcal{M}}$ and the remaining eigenvalues $\Lambda$, bound (4.23) depends only on the gap between the eigenvalues of $\Lambda$ and the neighbors of the Ritz values $\tilde{\mathcal{M}}$. Thus, even if there is an insufficient separation between the desired eigenvalues $\Lambda$ and the remaining spectrum, bound (4.23) provides a posteriori information about how well the exact eigenspace is included in an enlarged eigenspace.

4.3.2 Ising model

As outlined in Section 1, motivation for this thesis was to contribute to an improved understanding of quantum phase transitions in the presence of a reservoir. The phase transition is going to be analyzed in Chapter 5. In this section we use the Ising model as an example to test the accuracy of the bounds derived in this section.

The Ising model, first analyzed in [54], is one of the most fundamental models of statistical mechanics [33, 70]. Despite its rather drastic simplification of realistic situations it provides some valuable insights into the physics of magnetic interaction. The model considers a lattice of magnetic moments, where on each lattice site the magnetic moment is represented by a spin. Each spin has just two possible states either pointing up or down. This can be expressed by a spin variable with values $+1$ and $-1$, where $+1$ implies that the spin is pointing up and $-1$ that the spin is pointing down. Further, in the Ising model each spin is only
allowed to interact with its neighbors.

In particular, the Ising model is used to investigate phase transitions and critical points of magnetic systems, for more information see, e.g., [93,95,98] and the references contained therein.

The energy of the Ising model captures the interaction energy between neighboring particles and the effect of an applied magnetic field on each individual particle. In Figure 4.12 we show a one-dimensional system of 10 particles with nearest-neighbor interaction, where the spin of each particle is represented as an arrow in the figure. Note in this section that we investigate the case of periodic boundary conditions, i.e., we assume an interaction between the first and the last spin in the chain. In Figure 4.12 this implies an interaction between spin 1 and 10.

In the following we consider spin $\frac{1}{2}$ particles such as electrons, protons or neutrons, cf., e.g., [32,43]. Their spin angular momentum operator can be described by the Pauli matrices

$$\sigma^x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma^y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \text{and} \quad \sigma^z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \quad (4.26)$$

Those matrices are traceless and Hermitian and together with the $2 \times 2$ identity matrix $I$ they form an orthogonal basis for the complex Hilbert space of all $2 \times 2$ matrices. Further information of Pauli matrices can be found in, e.g., [44].

Considering first the interaction energy, it holds that if two adjacent spins are parallel, i.e., the spin variables of the neighbors are both $+1$ or both $-1$, then the energy of interaction is negative and set equal to $-Jz$, with $Jz$ being a positive
4.3 Numerical results

coefficient determining the interaction strength. Whereas, if the spins point anti-
parallel, i.e., the spin variables of neighboring particles are +1 next to −1, the
energy is positive and set to \( J \). Consequently, for each pair of neighboring par-
ticles \( i \) and \( i + 1 \) the interaction energy can be described by 
\[- J \sum_{i=1}^{d} \sigma_i^z \sigma_{i+1}^z, \]
where each summand is defined by
\[
\sigma_i^z := I_2(2^{i-1}) \otimes \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes I_2(d-(i+1)),
\]
with \( I_x \) denoting the identity matrix of dimension \( x \). It follows that the energy is
lower for a pair of parallel spins than for a pair of anti-parallel spins.

Second, we consider the energy introduced by an external magnetic field \( g \).
Note that if the magnetic field \( g \) is pointing up, it favors spins pointing up and
vice versa. Hence, for each particle \( i \), the field energy can be written as 
\[- g \sigma_i^x, \]
with
\[
\sigma_i^x := I_2(2^{i-1}) \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes I_2(d-i).
\]

Summarizing, the total energy of the one-dimensional Ising model in the pres-
ence of a transverse field becomes
\[
\mathcal{H} := - J \sum_{i=1}^{d} \sigma_i^z \sigma_{i+1}^z - g \sum_{i=1}^{d} \sigma_i^x. \quad (4.27)
\]

In general, the Hamiltonian describes the total energy of a system and therefore
Equation 4.27 indicates that the Hamiltonian of the Ising model is obtained by
summing \( 2d \) terms, where each term is representing a physical interaction. The
dimension of the Hamiltonian \( \mathcal{H} \) is growing exponentially in the number of par-
ticles, e.g., in a system of \( d \) particles the dimension of the Hamiltonian amounts
to \( n = 2^d \). Further we consider periodic boundary conditions, i.e., we assume
that \( \sigma_{d+1}^z := \sigma_1^z \).

In the following we choose \( J = 0.4, g = 0.6 \) and \( d = 10 \), which leads to a
dimension of \( n = 1024 \). The spectrum of \( \mathcal{H} \) is depicted in Figure 4.13. We are
interested in the ground state, first and second excited states of the system as well
as in the associated energies. In mathematical terms, we want to approximate
the subspace \( \mathcal{X}_1 \) corresponding to \( \Lambda_1 \) consisting of the three smallest eigenval-
ues of \( \mathcal{H} \). Thus, we are interested in the eigenvalues associated with the index
set \( J_1 = \{1, 2, 3\} \).

In Figure 4.13(b), we see that \( \mathcal{H} \) has a very dense eigenvalue distribution. Es-
specially, there is no significant gap between the third smallest and the fourth
smallest eigenvalue of \( \mathcal{H} \), i.e., \( \Lambda_1 \) is not well separated from the remaining eigen-
values, pointing to potential difficulties in the convergence analysis. In the following numerical experiments we will explore how well the spectral bounds, derived in Theorems 4.4, 4.15, 4.19, and 4.21 in this section, predict the trends of the outcome in this situation.

**Experiment 6. (Illustration of the results of Theorem 4.4)** In this experiment we illustrate the result of Theorem 4.4 using the Hamiltonian presented in (4.27). Specifically, we investigate how well the subspace $X_1$ corresponding to $\Lambda_1$ is contained in the search space $\tilde{K}_k$. To do so we consider the two cases

(i) $\Lambda_1 \subset \Lambda_2 = \Lambda_3 = \Lambda_4$,
(ii) $\Lambda_1 \subset \Lambda_2 \subset \Lambda_3 \subset \Lambda_4$,

i.e., we change the setting of nested subspaces in Theorem 4.4 in order to analyze the improvement of bound (4.3).

(i) $\Lambda_1 \subset \Lambda_2 = \Lambda_3 = \Lambda_4$ In the first case, $\Lambda_2$ consists of the four smallest eigenvalues of $\mathcal{H}$, i.e., we have $J_2 = J_3 = J_4 = \{1, \ldots, 4\}$. Further, we choose $\|E_k\|_2 = 10^{-11}\|\mathcal{H}\|_2$. In Figure 4.14 the angle of inclusion $\angle_{\max}^\omega (X_1, \tilde{K}_k)$ and the bound (4.3) are plotted for $\ell \in \{60, 75, 90\}$, where $\ell$ denotes the number of iteration steps that have already been carried out.

We see that bound (4.3) predicts all three convergence phases, i.e., initialization, convergence and stagnation, correctly. Moreover, the larger the number of iteration steps $\ell$, the sharper the corresponding bound during the convergence phase. The convergence rate of $\angle_{\max}^\omega (X_1, \tilde{K}_k)$ amounts to 0.55 and is overestimated by the bounds to be 0.77. Thus we obtain a slow down factor of 2.31 during the convergence phase. After the convergence phase all three bounds
4.3 Numerical results

![Graph](image)

Figure 4.14: Experiment 5: Angle of inclusion $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and its bounds (4.3) for $\ell \in \{60, 75, 90\}$

overestimates the true value by a constant factor of approximately 30.

(ii) $\Lambda_1 \subset \Lambda_2 \subset \Lambda_3 \subset \Lambda_4$ In the second case we aim to improve bound (4.3) by choosing more suitable nested subsets. In more detail, we choose $\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4$ to consist of the 3, 4, 7, 11 smallest eigenvalues of $\mathcal{H}$, respectively. For $\ell = 40$ this choice leads to an improvement of the constants (4.4) and (4.5). Figure 4.15 shows clearly that the bound (4.3) becomes sharper in comparison to case (i). Accordingly, the convergence rate of the bound is also improved such that the slow down factor now amounts to 1.31. However, caused by the dense eigenvalue distribution of the matrix $\mathcal{H}$ the algorithm starts to converge only after the 65th iteration step.

![Graph](image)

Figure 4.15: Experiment 5: Angle of inclusion $\angle_{\max}(\mathcal{X}_1, \tilde{K}_k)$ and its bounds (4.3) for adapted nested subspaces
4.3 Numerical results

Experiment 7. (Illustration of the result of Theorem 4.15) Here we analyze how well the exact subspace $\mathcal{X}_1$ of $\mathcal{H}$ is included in the computed Ritz space $\tilde{\mathcal{Y}}$ by using bound (4.13). We distinguish between $\dim(\mathcal{X}_1) = \dim(\tilde{\mathcal{Y}})$ and $\dim(\mathcal{X}_1) < \dim(\tilde{\mathcal{Y}})$, where in the first case the Ritz space corresponds to the three smallest and in the second case to the four smallest Ritz values.

(i) $\dim(\mathcal{X}_1) = \dim(\tilde{\mathcal{Y}})$ In the case of equal dimensions we are facing an insufficient separation between the eigenvalues of interest $\Lambda_1$ and the remaining Ritz values $\mathcal{M}_\perp$. Figure 4.16 shows that for $\dim(\mathcal{X}_1) = \dim(\tilde{\mathcal{Y}})$ the angle $\angle_{\text{max}}(\mathcal{X}_1, \tilde{\mathcal{Y}})$ does not converge and starts to fluctuate after 100 steps of the algorithm. This behavior is caused by the dense eigenvalue distribution of $\mathcal{H}$. To be more specific, the third and the fourth eigenvalue of $\mathcal{H}$ are multiple, such that the exact subspace corresponding to the three smallest eigenvalues can not be approximated by a Ritz space of the perturbed matrix $\mathcal{H} + E_k$.

Since the third and the fourth eigenvalue of $\mathcal{H}$ are multiple it follows that the condition $\|E_k\|_2 < \text{gap}(\mathcal{M}_\perp, \Lambda_2)$ of bound (4.13) is violated. Only during the initialization phase the bound (4.13) can be computed, in this phase the approximation of Ritz values is very inaccurate such that for the first 90 iteration steps the condition $\|E_k\|_2 < \text{gap}(\tilde{\mathcal{M}}_\perp, \Lambda_2)$ is satisfied.

(ii) $\dim(\mathcal{X}_1) < \dim(\tilde{\mathcal{Y}})$ In the second case the subspace $\tilde{\mathcal{Y}}$ is slightly enlarged and we obtain an improvement of the convergence behavior of $\angle_{\text{max}}(\mathcal{X}_1, \tilde{\mathcal{Y}})$. Figure 4.16 shows that the angle $\angle_{\text{max}}(\mathcal{X}_1, \tilde{\mathcal{Y}})$ starts to converge after 83 iterations. The bound (4.13) captures this behavior correctly and overestimates the true value after the convergence phase by a constant factor of 15.

Experiment 8. (Illustration of the results of Theorem 4.19 and Theorem 4.21) Again, we distinguish between the two cases $\dim(\mathcal{X}) = \dim(\tilde{\mathcal{Y}})$ and $\dim(\mathcal{X}) < \dim(\tilde{\mathcal{Y}})$. The eigenvalues of interest contained in $\Lambda$ corresponding to an eigenspace $\mathcal{X}$ consists of the three smallest eigenvalues of $\mathcal{H}$. Further, the approximated eigenpair $(\tilde{\mathcal{M}}, \tilde{\mathcal{Y}})$ is obtained via calculating a Ritz pair of $\mathcal{H} + E_k$, where $\tilde{\mathcal{M}}$ consists of the three smallest Ritz values for $\dim(\mathcal{X}) = \dim(\tilde{\mathcal{Y}})$ and of the four smallest Ritz values for $\dim(\mathcal{X}) < \dim(\tilde{\mathcal{Y}})$. We choose $\|E_k\|_2 = 10^{-10}$.

(i) Since the third eigenvalue of the Hamiltonian (4.27) is multiple, there is no gap between $\Lambda$ and the remaining eigenvalues. Consequently, as we see in Figure 4.17, the angle of inclusion $\angle_{\text{max}}(\mathcal{X}, \tilde{\mathcal{Y}})$ does not converge. Further, the condition $\|E_k\|_2 + \|R\|_2 < \text{gap}(\text{env}(\tilde{\mathcal{M}}), \Lambda_\perp)$ is violated and therefore bound (4.19) does not exist.

(ii) $\dim(\mathcal{X}) < \dim(\tilde{\mathcal{Y}})$ The situation changes if we enlarge the Ritz space slightly, i.e., now we consider the Ritz space corresponding to the four smallest Ritz values. As illustrated in Figure 4.17 the angle of inclusion converges to limiting accuracy after 85 iteration steps and the bound (4.23) predicts the behavior of the actual value. More precisely, the convergence rate of the angle, computed by 0.44, and the estimate generated by the bound, 0.45, differ only slightly. This leads to a slow down factor of 1.03.
4.4 Conclusion

Using the concept of the angle of inclusion we investigated the convergence of inexact Krylov subspace methods. More precisely, we first bounded the distance of the exact invariant subspace $\mathcal{X}$ to an inexact Krylov subspace $\tilde{K}_k$ and second to an inexact Ritz space $\tilde{Y}$ therein. With the first bound given by inequality (4.3) we investigated a situation where a few iteration steps of an inexact Krylov subspace method have been performed without achieving convergence. We analyzed in a priori setting how many more iteration steps of an inexact Krylov subspace method are necessary to ensure convergence to a certain tolerance. More precisely, we bounded how well the exact invariant subspace $\mathcal{X}$ is contained in the search space $\tilde{K}_k$ in terms of the angle of inclusion of an extended invariant subspace and $\tilde{K}_j$.

In the second, a posteriori, setting we analyzed the quality of the approximated subspace $\tilde{Y}$, i.e., we bounded the angle of inclusion of $\mathcal{X}$ and a Ritz space $\tilde{Y}$ that is a subspace of $\tilde{K}_k$. Here we distinguished between bounds for the ap-
4.4 Conclusion

Approximation quality of the inexact Ritz space involving Krylov subspaces stated in inequality (4.13) and bounds in terms of their residual given by the inequalities (4.19) and (4.23). The first bound indicated the convergence of the Ritz space to the invariant subspace and the second bounds provided us with information when to stop the iteration.

Importantly, we showed that even in the presence of perturbations and a small gap between the desired and remaining eigenvalues the presented bounds can be valid and meaningful. In the a priori setting this was achieved by considering nested subsets, i.e., we considered not only the invariant subspace $X$ itself but also subspaces that belong to slightly enlarged nested sets of eigenvalues. By choosing "suitable" nested subsets we guaranteed that even if the invariant subspace $X$ is simple, the presented bound is meaningful. In the a posteriori setting the key idea was to enlarge the Ritz space in order to guarantee a sufficiently large gap between the remaining Ritz values and desired eigenvalues, thus we allow the approximation $\tilde{Y}$ to be of larger dimension than the invariant $X$ and investigate how well is $X$ contained in $\tilde{Y}$.

Finally, we tested the obtained analytical results by analyzing an academic example and an one dimensional Ising model. It was confirmed by both examples that the bounds predict correctly the trends of the convergence curves.

Beside being interesting in their own rights, the derived spectral error bounds may also be used as stopping criterion within an algorithm when combined with certain spectral estimation techniques, cf. [8, 89].

Inspired by these results we transfer in the next chapter the inexact Arnoldi method from the matrix-vector setting to the operator-tensor setting in order to solve high dimensional eigenvalue problems resulting from quantum spin systems.

Figure 4.17: Experiment 7: Illustration of the a posteriori bound (4.23)
5 Application of the inexact Arnoldi method to the tensor format

In this chapter we use the inexact Arnoldi method to describe the quantum dynamics of $d$ interacting particles. More precisely, we study whether and how this method can be applied to the tensor setting in order to compute the ground state, i.e., the smallest eigenvalue, of large physical spin systems. Such physical systems consisting of $d$ particles are modeled by Hamiltonians of the form

$$H = \sum_{k=1}^{R} a_k P_1^{(k)} \otimes P_2^{(k)} \otimes \ldots \otimes P_d^{(k)} \in \mathbb{C}^{n^d \times n^d},$$

where $P_j^{(k)} \in \mathbb{C}^{n \times n}$ and with the coupling constants $a_k \in \mathbb{R}$. Due to the exponential growth of the problem dimension $n^d$ with the number of particles $d$ the computation of the smallest eigenvalue of $H$ becomes challenging already for moderate sizes of $d$. The structure of quantum dynamical systems ensures that its eigenvectors can be well approximated by a low rank tensor format [67]. Thus, by exploiting the underlying problem structure we can use a suitable tensor format to make storage and computation feasible and to overcome the so called curse of dimension.

In the tensor setting, operations like summation of tensors or operator-tensor multiplication result in a growth of the tensor rank and consequently in an increase of the storage requirements. To avoid this growth in the tensor rank, these operations need to be followed by an approximation procedure. Controlling the approximation accuracy of the rounding procedure enables us to tune the algorithm towards a higher accuracy or a higher storage-efficiency.

In the following we provide a brief overview on tensor formats especially on the TT format, cf. [41, 45, 81]. We introduce necessary concepts and notation to apply the inexact Arnoldi method to the TT format. In our numerical experiments we use the YZ-model [57, 79] to verify the applicability of the inexact Arnoldi method to solve eigenproblems of much larger dimension than we are able to in the matrix-vector setting.
5.1 Preliminaries

In this section we provide a brief summary of the mathematical background needed to apply the inexact Arnoldi Algorithm stated in Section 3.3 to tensors. We show why the TT format is a suitable choice for the considered application of quantum many body systems and introduce necessary tensor concepts and notation. For a more comprehensive introduction we refer to the survey papers [41,66] and textbooks [45,61] as well as the references therein.

5.1.1 Notation and basic operations

A general tensor \( x \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \) with \( n_1, \ldots, n_d \in \mathbb{N} \) is a \( d \)-dimensional array with entries \( x_{i_1,\ldots,i_d} \in \mathbb{C} \). Commonly, \( d \) is called the order of the tensor. Due to the application we are interested in tensors of high order, i.e., \( d = 20 \) or even higher. In the following we introduce some reshaping operations for tensors.

**Matricization.** It describes the process of reshaping the elements of a tensor into a matrix. In particular, we consider the \( \mu \)-th mode matricization where the \( \mu \)-th mode is used as row indices and the remaining modes as column indices, i.e.,

\[
x_{(\mu)} \in \mathbb{C}^{n_\mu \times (n_1 \cdots n_{\mu-1} n_{\mu+1} \cdots n_d)}, \quad \text{with} \quad \mu = 1, \ldots, d.
\]

Note that we apply a colexicographic ordering of the indices, however any consistent ordering would also be possible. We can describe the vectorization analogously where we arrange all tensor entries into one long row vector, i.e.,

\[
\text{vec}(x) = x_{(1,\ldots,d)} \in \mathbb{C}^{(n_1 n_2 \cdots n_d)}.
\]

The \( \mu \)-th mode matricization is just one possibility of reordering the elements of a tensor into a matrix, another possibility is called unfolding which is introduced in Section 5.1.3.

**Slice.** A slice of the tensor \( x \) is obtained by fixing one index and varying all others which results in a subarray of dimension \( d-1 \). For a tensor of order three a possible slice for each tensor mode is illustrated in Figure 5.1.

**Inner product and norm.** Based on the above definitions of \( \mu \)-th mode matricization and vectorization the inner product of two tensors \( x \) and \( y \) is related to the matrix or vector case by

\[
x^H y = \text{vec}(x)^H \text{vec}(y) = \text{trace} \left( x_{(\mu)}^H y_{(\mu)} \right) \quad \text{for} \quad \mu = 1, \ldots, d.
\]
5.1 Preliminaries

![Figure 5.1: Slices of a third order tensor (horizontal x(i_1,:), lateral x(:,i_2,:), frontal x(:, :, i_3))](image)

The norm of a tensor \( x \) is defined by the square root of the sum of the squares of all its elements

\[
\|x\| = \sqrt{\sum_{i_1=1}^{n_1} \ldots \sum_{i_d=1}^{n_d} x_{i_1,\ldots,i_d}^2} = \sqrt{x^Hx}.
\]

5.1.2 Tensor formats

In general, the explicit storage of a high-order tensor is not possible. However, we can use the underlying problem structure to approximate this tensor by a data sparse tensor format, which may lead to a drastic reduction of the storage requirements.

For example, a rank-one approximation of the tensor \( x \) is given by

\[
\text{vec}(x) \approx u_d \otimes u_{d-1} \otimes \ldots \otimes u_1, \quad u_1 \in \mathbb{C}^{n_1}, \ldots, u_d \in \mathbb{C}^{n_d},
\]

and reduces the storage requirement from \( n_1 \cdot n_2 \cdot \ldots \cdot n_d \) to \( n_1 + n_2 + \ldots + n_d \) entries. Unfortunately, in most applications the rank-one approximation will lead to a large approximation error. A straightforward generalization of this approximation is to use the sum of \( r \) rank-one approximations

\[
\text{vec}(x) \approx \sum_{i=1}^{r} u_d^{(i)} \otimes u_{d-1}^{(i)} \otimes \ldots \otimes u_1^{(i)}, \quad \text{with} \quad u_1^{(i)} \in \mathbb{C}^{n_1}, \ldots, u_d^{(i)} \in \mathbb{C}^{n_d}.
\]

This format is called CP (canonical polyadic) format, for detailed information, see, e.g., [66]. The CP format scales linearly in \( d \). However, finding a numerical robust algorithm to determine an approximation to a specific accuracy in this format is a difficult issue, see [2, 6] for a detailed derivation.

This problem can be solved by using the Tucker format which is better suited for numerical treatments. Here the tensor is transformed into a so called core
5.1 Preliminaries

tensor which is multiplied by a matrix along each mode, i.e.,

\[ \text{vec}(x) \approx (U_d \otimes U_{d-1} \otimes \cdots \otimes U_1) \text{vec}(c), \tag{5.1} \]

with the factor matrices \( U_1 \in \mathbb{C}^{n_1 \times r_1}, \ldots, U_d \in \mathbb{C}^{n_d \times r_d} \) and the core tensor \( c \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \). The tuple \( (r_1, r_2, \ldots, r_d) \in \mathbb{N}^d \) is called multilinear rank of the Tucker format and is defined by

\[ \text{rank}_{ML}(x) = (r_1, r_2, \ldots, r_d) = (\text{rank}(x(1)), \text{rank}(x(2)), \ldots, \text{rank}(x(d))). \]

To obtain this multilinear rank approximation of a desirable accuracy we can make use of the higher-order-SVD \[28\], a generalization of the singular value decomposition (SVD), to obtain a quasi best approximation of (5.1), cf. \[28,45\].

Assuming small ranks, the Tucker format leads to a drastic reduction of the storage requirements. However, the storage requirements of the core tensor \( c \) still grow exponentially with \( d \). Consequently, even for small ranks \( r_1, \ldots, r_d \) the Tucker format is only applicable for very small \( d \), i.e., \( d < 5 \).

Referring to quantum spin systems we are interested in problem sizes with \( d > 20 \). Hence, we are looking for a format that combines the good storage properties of the CP-format and the numerical behavior, like SVD based compression and closedness, of the Tucker format. This leads us to the TT format.

5.1.3 TT format

The format was first introduced and used in the physics community under the term Matrix product states, see \[86,96,121\]. In the numerical analysis community the TT format was first proposed in \[81,83,84\].

To define the TT format, we need to find a different way of reshaping a tensor \( u \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \) into a matrix. In contrast to the \( \mu \)-th mode matricization, where we used only the \( \mu \)-th mode as row indices and the remaining modes as column indices, we split the modes of the tensor \( u \) into two disjoint sets and use the first set as row and the second set as column indices. This yields to the matricization

\[ u_{[\mu]} \in \mathbb{C}^{(n_1 \cdots n_\mu) \times (n_{\mu+1} \cdots n_d)}, \]

which is called \( \mu \)-th unfolding of the tensor \( u \). To order the indices we use a colexicographic ordering. Note that the \( \mu - th \) matricization and the \( \mu - th \) unfolding coincide for \( \mu = 1 \). Further, for \( \mu = d \) we have \( u_{(d)} = (u_{[d-1]})^T \).

Using of the \( \mu - th \) unfolding we can define the TT format. To do so, we first define the tensor train rank of \( u \in \mathbb{C}^{(n_1 \times \cdots \times n_d)} \) by the \((d + 1)\) tuple

\[ \text{rank}_{TT}(u) = (r_0, r_1, \ldots, r_d) \]

\[ := (1, \text{rank}(u_{[1]}), \text{rank}(u_{[2]}), \ldots, \text{rank}(u_{[d-1]}), 1), \tag{5.2} \]
5.1 Preliminaries

Figure 5.2: Tensor network diagram of a TT tensor

with \( r_0 = r_d = 1 \). Based on this tuple, each entry of the tensor \( u \in \mathbb{C}^{(n_1 \times \cdots \times n_d)} \) can be written in TT format by the matrix product

\[
U_1(i_1) U_2(i_2) \cdots U_d(i_d),
\]

with \( U_\mu(i_\mu) \in \mathbb{C}^{r_\mu-1 \times r_\mu} \) for \( i_\mu = 1, \ldots, n_\mu \). The matrix \( U_\mu(i_\mu) \) is in fact a lateral slice of the order three tensor \( u_\mu \in \mathbb{C}^{r_\mu \times n_\mu \times r_{\mu+1}} \). Thus, in index form the TT format can also be written as

\[
u_{i_1, \ldots, i_d} = \sum_{a_1=1}^{r_1} \cdots \sum_{a_{d-1}=1}^{r_{d-1}} u_1(1, i_1, a_1) u_2(a_1, i_2, a_2) \cdots u_d(a_{d-1}, i_d, 1),
\]

where the tensors \( u_\mu \) for \( \mu = 1, \ldots, d \) are called core tensors of the TT format. These tensor format can be represented graphically by a linear tensor network, which is illustrated in Figure 5.2 for a tensor of order \( d = 5 \). In the following we use reformulations of (5.3) to be able to access and manipulate single core tensors \( u_\mu \), and follow the notation of [67]. The left and the right unfolding are defined by

\[
u^L_\mu = (u_\mu)_2 \in \mathbb{C}^{r_\mu-1 \times n_\mu \times r_\mu} \quad \text{and} \quad \nu^R_\mu = (u_\mu)_1 \in \mathbb{C}^{r_\mu-1 \times n_\mu r_\mu}.
\]

Further, to divide the tensor into a left and a right part we introduce the so called interface matrices by

\[
u_{\leq \mu} \in \mathbb{C}^{n_1 n_2 \cdots n_\mu \times r_\mu} \quad \text{with} \quad \nu_{\leq}(i_1, \ldots, i_\mu) = U_1(i_1) U_2(i_2) \cdots U_\mu(i_\mu)
\]

and

\[
u_{\geq \mu} \in \mathbb{C}^{n_{\mu+1} n_{\mu+2} \cdots n_d \times r_{\mu-1}} \quad \text{with} \quad \nu_{\geq}(i_\mu, \ldots, i_d) = [U_\mu(i_\mu) U_{\mu+1}(i_{\mu+1}) \cdots U_d(i_d)]^T.
\]

In Figure 5.2 we illustrate the interface matrices \( \nu_{\leq 2} \) and \( \nu_{\geq 4} \) of a TT tensor of order six. This matrices allow us to extract individual cores of the TT format in the form of

\[
u_{\neq \mu} = \nu_{\geq \mu+1} \otimes I_{n_\mu} \otimes \nu_{\leq \mu-1},
\]

for a detailed derivation, see [67]. This enables us to state the decomposition

\[
u = \nu_{\neq \mu} \operatorname{vec}(\nu^R_\mu).
\]
5.1 Preliminaries

In the following we introduce basic tensor operations necessary to implement the inexact Arnoldi method in the TT format.

5.1.3.1 Basic operations in TT format

To analyze the compensated Gram Schmidt method and the inexact Arnoldi method in the tensor setting, we first describe some basic operations such as addition, inner product, and orthogonalization in the TT format. In particular, we investigate whether the operations cause an increase of the TT rank. Naturally, an increase of the TT rank should be avoided as it would imply an increase of the storage requirements. All basic operations mentioned below are implemented in the TTeMPS toolbox [106], which is heavily used in our implementation of the inexact Arnoldi method using the TT format. Further information and implementation details of the TTeMPS toolbox can be found in [81,105].

### Multiplication by a number and addition.

To multiply a TT tensor by a number, one of the cores needs to be multiplied by this number. The addition of two TT tensors

\[ u_{i_1,i_2,...,i_d} = V_1(i_1)V_2(i_2)\ldots V_d(i_d) \quad \text{and} \quad w_{i_1,i_2,...,i_d} = W_1(i_1)W_2(i_2)\ldots W_d(i_d) \]

with \(\text{rank}_{TT}(v) = (r_1, \ldots, r_d)\) and \(\text{rank}_{TT}(w) = (\hat{r}_1, \ldots, \hat{r}_d)\) results in a TT tensor \(u_{i_1,i_2,...,i_d} = U_1(i_1)U_2(i_2)\ldots U_d(i_d)\) with \(\text{rank}_{TT}(u) = (r_1 + \hat{r}_1, \ldots, r_d + \hat{r}_d)\). The slices of the core tensors result in

\[
U_1(i_1) = [V_1(i_1) W_1(i_1)], \quad U_\mu(i_\mu) = \begin{bmatrix} V_\mu(i_\mu) & 0 \\ 0 & W_\mu(i_\mu) \end{bmatrix}, \quad \text{and} \quad U_d(i_d) = \begin{bmatrix} V_d(i_d) \\ W_d(i_d) \end{bmatrix}
\]

for \(\mu = 2, \ldots, d - 1\). In summary, the addition of two tensors requires no operations, but implies an increase of the tensor rank. To reduce this rank increase, the addition needs to be followed by a rounding procedure which is introduced below.

### Orthogonalization of a TT tensor.

A TT tensor \(u\) is called \(\mu\)-orthogonal if

\[
(u^L_\eta)^H u^L_\eta = I_r \quad \text{for} \quad \eta = 1, \ldots, \mu - 1 \quad \text{and} \quad (u^R_\eta)^H (u^R_\eta) = I_r \quad \text{for} \quad \eta = \mu + 1, \ldots, d.
\]

The orthogonalization of a TT tensor is done iteratively core by core. We start with the first core and compute the QR decomposition \(u^L_1 = Q^L R\). Substituting \(u^L_1\) by \(Q^L\) leads to the first orthogonal core. The nonorthogonal part \(R\) is shifted...
5.1 Preliminaries

to the second core, i.e., we update $u_2^R$ by $Ru_2^R$. We continue this iterative procedure by computing a QR decomposition of the updated second core and so forth. This procedure computes the left orthogonal part of the tensor and is called left orthogonalization. The right orthogonal part can be computed in an analogous manner. We compute the QR decomposition of the last core $u_d^R = (QR)^H$ and continue with substituting $u_d^R$ by $Q^H$ as well as $u_{d-1}^L$ by $u_{d-1}^R$. The continuation of this process is referred to as right orthogonalization. A tensor $u$ is called $m$-orthogonal if all cores $u_h$ with $h < m$ are left orthogonal and all cores with $h > m$ are right orthogonal.

The orthogonalization of a TT tensor does not change the tensor, but its internal representation. Thus, the representation of a vector in the TT format is not unique.

**Inner product and Norm.** The inner product of two TT tensors $v, w \in \mathbb{C}^{n_1, n_2, \ldots, n_d}$ is defined by

$$v^H w = \text{vec}(v)^H \text{vec}(w).$$

The complexity of the computation of the inner product can be reduced by exploiting the structure of the TT tensor, for more information, see, e.g., [81, 106]. The norm of $u$ can be computed by

$$\|u\|_F = \sqrt{u^H u}.$$ 

Note that assuming $u$ is $\mu$-orthogonal, the computation of the norm of the tensor $u$ is equal to the computation of the norm of the $\mu$-th core.

**Linear operator format.** So far we have seen how a vector of length $N = n_1n_2\ldots n_d$ can be represented in TT format. To compute eigenvalues of the corresponding high dimensional, sparse matrix $A \in \mathbb{C}^{N \times N}$, we need to compute a "matrix-vector" product in TT format. For that we consider a linear operator $A : \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} \rightarrow \mathbb{C}^{m_1 \times m_2 \times \cdots \times m_d}$, the so called operator TT format or TT matrix that is introduced in [81]. The matrix $A$ is the matrix representation of $A$ if its elements are defined by

$$A(i_1, i_2, \ldots, i_d, j_1, j_2, \ldots, j_d) = A((i_1, i_2, \ldots, i_d), (j_1, j_2, \ldots, j_d)) = A_1(i_1, j_1)A_2(i_2, j_2)\ldots A_d(i_d, j_d) \quad (5.4)$$ 

where $(i_1, i_2, \ldots, i_d)$ enumerate the rows of $A$ and $(j_1, j_2, \ldots, j_d)$ the columns. Further, each $A_\mu(i_\mu, j_\mu)$ is a matrix of size $s_{\mu-1} \times s_\mu$. Note that $s_\mu$ describes the rank of the TT operator. Similarly to the TT format, $A_\mu(i_\mu, j_\mu)$ can be interpreted as a slice of a core tensor

$$a_\mu(:, i_\mu, j_\mu, :) = A_\mu(i_\mu, j_\mu), \quad \text{with} \quad a_\mu \in \mathbb{C}^{s_{\mu-1} \times m_\mu \times n_\mu \times s_\mu},$$ 

133
5.1 Preliminaries

Figure 5.3: Tensor network diagram of a TT operator

but each core tensor is of dimension four. Hence, also the tensor network diagram, illustrated in Figure 5.3 for \( d = 5 \), differs only by an additional "leg" at each core tensor.

**Application of the operator to a TT tensor.** If \( A \) is the matrix representation corresponding to the operator \( A \) in TT format, then we have

\[
y = Ax \iff \text{vec}(y) = A \text{vec}(x).
\]

Further, as stated in [81], the “TT matrix-vector-product” \( y = Ax \) is given by

\[
y(i_1, \ldots, i_d) = \sum_{j_1=1}^{n_1} \cdots \sum_{j_d=1}^{n_d} A(i_1, \ldots, i_d, j_1, \ldots, j_d) x(j_1, \ldots, j_d)
\]

\[
= \sum_{j_1=1}^{n_1} \cdots \sum_{j_d=1}^{n_d} A_1(i_1, j_1) A_2(i_2, j_2) \cdots A_d(i_d, j_d) X_1(j_1) X_2(j_2) \cdots X_d(j_d)
\]

\[
= \sum_{j_1=1}^{n_1} \cdots \sum_{j_d=1}^{n_d} (A_1(i_1, j_1) \otimes X_1(j_1)) \cdots (A_d(i_d, j_d) \otimes X_d(j_d))
\]

\[
= Y_1(i_1) \cdots Y_d(i_d)
\]

with

\[
Y_\mu(i_\mu) = \sum_{j_\mu=1}^{n_\mu} A_\mu(i_\mu, j_\mu) \otimes X_\mu(j_\mu) \quad \text{for} \quad \mu = 1, \ldots, d.
\]

Note that the TT rank of \( y \) equals the TT rank of the operator \( A \) times the TT rank of \( x \), thus in general each application of the operator implies a significant increase in the TT rank.

As already mentioned for the tensor addition operation the increase of the TT rank can be reduced by the application of the rounding procedure described below.
5.2 Inexact Arnoldi method using TT format

**Rounding.** Assuming that \( u \) is a tensor in TT format with increased ranks, we want to determine an approximation of \( u \) with reduced ranks that satisfies

\[
\|u - \hat{u}\|_F \leq \epsilon \|u\|_F.
\]

To do so, we use the TT-SVD as introduced in [81]. Suppose that \( u \) is in TT format with rank \( \text{rank}_{TT}(u) = (r_0, r_1, \ldots, r_d) \) and \( u \) is \( d \)-orthogonal, then we start by computing the SVD of the last core \( u_R^d = U \Sigma V^H \). The orthogonal factor \( V^H \) is assigned as the new core \( u_R^d \) and \( u_{L,1} \) is updated by \( u_{L,1} = U \Sigma \). Keeping \( \tilde{r}_d \) singular values of \( \Sigma \) reduces the rank between the cores \( u_d \) and \( u_{d-1} \) from \( r_d \) to \( \tilde{r}_d \). We continue with this procedure until we reached the core \( u_1 \) and thus have truncated the ranks from \( (r_0, r_1, \ldots, r_d) \) to \( (\tilde{r}_0, \tilde{r}_1, \ldots, \tilde{r}_d) \).

Note that, if we want to compute an approximation in TT format with a given relative accuracy \( \epsilon \) the corresponding TT ranks are bounded by

\[
 r_i \leq \frac{\epsilon}{\sqrt{d-1}} \|u\|_F \quad \text{for all} \quad i = 1, \ldots, d.
\]

A derivation of this bound can be found in [81].

Using these operations we introduce in the next section the inexact Arnoldi method in the TT format.

5.2 Inexact Arnoldi method using TT format

A central element of the Arnoldi method introduced in Section 3.3 is the orthogonalization of the Arnoldi basis. To apply this method to the tensor setting we first need to investigate how to orthogonalize a tensor against other tensors. Similarly to the ComGS method 3.1 we use in the case of inexact operations a non-standard projection to reduce the deviation from orthogonality of the computed tensor caused by perturbations. In the tensor setting we differentiate between two perturbation sources. On the one hand it is assumed that inner products of tensors can be computed exactly, i.e., with an accuracy of the order of machine precision, cf., [80]. Therefore, a lower bound of this perturbation is given by the machine precision and cannot be improved upon. On the other hand the addition of two tensors in TT format causes a drastic increase of the TT rank and consequently of the storage requirements. Hence, after the addition of two tensors a rounding procedure, as described in Section 5.1.3, needs to be applied. This procedure truncates a tensor at a certain rank that relates to a specific accuracy level, i.e., we obtain a lower rank tensor of lower accuracy. This specific accuracy also quantifies the magnitude of the perturbation caused by this rounding step.

The ComGS\(_{TT} \) method, shown in Algorithm 5.1, is initialized by a tensor \( w \in \mathbb{C}^{n_1 \times \cdots \times n_d} \) in TT format, an array \( V_k \) of \( k \) \( d \)-order tensors in TT format, the nonsingular cross product matrix \( D = V_k^H V_k \in \mathbb{C}^{k \times k} \) and the accuracy level \( \varepsilon_v \) of the rounding procedure. Taking this into account the ComGS\(_{TT} \) method computes a
Algorithm 5.1 Compensated Gram Schmidt using TT format

**Input:** $w \in \mathbb{C}^{n_1 \ldots n_d}$ TT tensor, $V_k = [v_1, \ldots, v_k]$ array of $k$ TT tensors with $v_i \in \mathbb{C}^{n_1 \ldots n_d}$ for $i = 1, \ldots, k$, $D \in \mathbb{C}^{k \times k}$, $D$ nonsingular, $\varepsilon_v$, truncation threshold

**Output:** $\tilde{v} \in \mathbb{C}^{n_1 \ldots n_d}$ TT tensor, $h \in \mathbb{R}$, $D_+ \in \mathbb{C}^{k+1 \times k+1}$

1. $h \approx D^{-1}V_k^Hw$ (orthogonalization coefficients)
2. $\ell = \text{round}(w - V_kh, \varepsilon_v)$ (orthogonalization)
3. $\eta \approx \|\ell\|_F$
4. if $\eta = 0$ then stop end if
5. $v \approx \ell / \eta$ (normalization)
6. $D_+ \approx \begin{bmatrix} D & V_k^Hv \\ V_k^HV_v & V_k^HV_v \end{bmatrix}$ (update $D$

$d$-order tensor in TT format $v$ that is nearly orthogonal to the tensors of the array $V_k$.

**Remark 5.1.** The orthogonalization step, i.e., Step 2 of Algorithm 5.1, is a computationally demanding task where $k$ tensor additions and $k$ tensor scalings need to be carried out. To reduce this computational effort we substitute the orthogonalization step, i.e., Step 2 of Algorithm 5.1, by the selective orthogonalization step presented in Algorithm 5.2. According to Algorithm 5.2 the tensor addition and the tensor scaling are only performed if the absolute value of $h_i$ is larger then the truncation threshold $\varepsilon_v$ times the norm of $w$, i.e., if the tensor $w$ is not already nearly orthogonal to the tensor $v_i$ of the tensor array $V_k$.

After introducing the ComGS$_{TT}$ method we can investigate the inexact Arnoldi method for TT tensors. This method is initialized by a tensor operator $A$
5.2 Inexact Arnoldi method using TT format

in TT format of order $d$, a normalized TT tensor $v_1$ of order $d$ and a truncation threshold $\varepsilon_v$. After $m$ iteration steps Algorithm 5.3 computes $V_{m+1}$, an array of $m + 1$ nearly orthogonal TT tensors of order $d$, the Hessenberg matrix $H_{m+1}$ and the cross product matrix $D_{m+1}$. The matrices $H_{m+1}$ and $D_{m+1}$ are of dimension $m + 1 \times m + 1$. Since the operator-tensor multiplication in Step 3 of

**Algorithm 5.3 Inexact Arnoldi method using TT format**

**Input:** A TT operator of order $d$, $v_1 \in C^{n_1,n_2,...,n_d}$ TT tensor, $\|v_1\|_F = 1$, $\varepsilon_v$ truncation threshold, $m \in \mathbb{N}$

**Output:** $V_{m+1} = [v_1,v_2,\ldots,v_{m+1}]$ with $v_i \in C^{n_1,n_2,...,n_d}$ TT tensor for $i = 1,\ldots,m+1$, $H_m \in C^{m \times m}$, $h_{m+1,m}$, $D_{m+1} \in C^{m+1 \times m+1}$

1: $V_1 = v_1$, $D_1 = 1$, $H_0 = [] \in C^{0 \times 0}$ (initialization)
2: for $k = 1,2,3,\ldots,m$ do
3: \hspace{0.5cm} $w = \text{round} (\mathcal{A}v_k,\varepsilon_v)$ (operator-tensor m.)
4: \hspace{0.5cm} $[v_{k+1},h_{1,k,k},h_{k+1,k},D_{k+1}] = \text{ComGS}_{TT}(w,V_k,D_k,\varepsilon_v)$ (orthogonalization)
5: \hspace{0.5cm} $H_k = \begin{bmatrix} H_{k-1} & h_{1,k-1,k} \\ h_{k-1,k} & h_{k,k} \end{bmatrix}$ (update $H_k$)
6: \hspace{0.5cm} $V_{k+1} = [V_k,v_{k+1}]$ (update $V_k$)
7: end for

Algorithm 5.3 increases the TT rank, this operation is followed by a rounding procedure to truncate the tensor rank while keeping the tolerance $\varepsilon_v$.

As we are interested in an approximation of the eigenpair $(L,X)$, with $L \in C^{p \times p}$ and $X$ is an array of $p$ TT tensors of order $d$, we perform $k$ steps of the inexact Arnoldi method presented in Algorithm 5.3. This results into the basis $V_{k+1}$ and the Hessenberg matrix $H_k$ which are used by the Algorithm 5.4 to compute the Ritz pair $(M_k,Y_k)$.

**Algorithm 5.4 Rayleigh-Ritz method using TT format**

**Input:** $H_k \in C^{k \times k}$ upper Hessenberg matrix, $V_k = [v_1,v_2,\ldots,v_k]$ with $v_i \in C^{n_1,n_2,...,n_d}$ TT tensor for $i = 1,\ldots,k$, $p \in \mathbb{N}$ with $p < k$

**Output:** $M_k \in C^{p \times p}$, $Y_k = [y_1,y_2,\ldots,y_p]$ with $y_i \in C^{n_1,n_2,...,n_d}$ TT tensor for $i = 1,\ldots,p$

1: Compute the eigenpair $(M_k,Z_k)$ of $H_k$.
2: Take $(M_k,Y_k) = (M_k,V_kZ_k)$ as Ritz pair of $\mathcal{A}$.

137
The resulting Ritz pair consists of the diagonal matrix $M_k$ of dimension $p \times p$ and the array $Y_k$ of $p$ TT tensors of order $d$.

To demonstrate the applicability of the inexact Arnoldi method in the tensor setting, we are using in the next section Algorithm 5.3 to compute some of the smallest eigenvalues and the corresponding invariant subspace of the YZ-model [47,79].

5.3 Numerical Example: YZ-model

Already the magnetic interaction in simple structures, like the one-dimensional YZ-model, lead to an interesting critical behavior, cf. [7,93]. For this reason, we focus in this section on the YZ-model which is similar to the Ising model introduced in Section 4.3.2. In terms of Pauli matrices, see (4.26), the Hamiltonian of the YZ-model is represented by

$$
\mathcal{H} = -\sum_{i=1}^{d} g_i \sigma_i^x - \sum_{i=1}^{d} I_i^y \sigma_i^y \sigma_{i+1}^y - \sum_{i=1}^{d} I_i^z \sigma_i^z \sigma_{i+1}^z. \quad (5.5)
$$

The difference to the Ising model consists in the additional interaction term $\sum_{i=1}^{d} I_i^y \sigma_i^y \sigma_{i+1}^y$ which allows the spins to tip away from the z-axis. The parameter $g_i$ in (5.5) describes the strength of the external magnetic field that effects the local energy of the $i$-th spin. Further, the coupling constants $I_i^z$ and $I_i^y$ define the interaction energy between next neighbor spins $i$ and $i+1$. In particular, if $I_i^z < 0$ for $\alpha \in \{y, z\}$, then the next neighbor interaction is called ferromagnetic and the spins tend to align in the ground state. If $I_i^z > 0$ the interaction is called anti-ferromagnetic where the spins tend to alternatingly orient themselves in opposite directions in the ground state. For more information on the YZ-model see for example [57,79].

In the homogeneous case, i.e., $g_i = g$, $I_i^y = I^y$ and $I_i^z = I^z \ \forall i$, the Hamiltonian of the YZ-model results in

$$
\mathcal{H} = -g \sum_{i=1}^{d} \sigma_i^x - I^y \sum_{i=1}^{d} \sigma_i^y \sigma_{i+1}^y - I^z \sum_{i=1}^{d} \sigma_i^z \sigma_{i+1}^z. \quad (5.6)
$$

The Hamiltonian in (5.6) describes a closed spin chain model, i.e., there is an interaction between the first and the last spin. If we additionally assume that $I_d^z = I_d^y = 0$ we obtain a homogeneous open YZ-model. To express the model by using only the two dimensionless parameters $s$ and $\gamma$ we rescale the Hamiltonian (5.6) by an overall energy scale $\Omega$, such that

$$
\mathcal{H}(s, \gamma) = -\Omega(1-s) \sum_{i=1}^{d} \sigma_i^x - \Omega s \frac{1-\gamma}{2} \sum_{i=1}^{d} \sigma_i^y \sigma_{i+1}^y - \Omega s \frac{1+\gamma}{2} \sum_{i=1}^{d} \sigma_i^z \sigma_{i+1}^z, \quad (5.7)
$$
with $g = \Omega(1-s)$, $2J^y = \Omega s(1-\gamma)$, and $2J^z = \Omega s(1+\gamma)$. The parameter $s$ in (5.7) describes the transition from a paramagnetic ground state, where the spins are randomly oriented, to a ferromagnetic ground state either along $y$- or $z$-direction, where all spins are aligned parallel, cf. [93]. The corresponding phase diagram of the one dimensional YZ-model is shown in Figure 5.4.

![Phase diagram of a spin 1/2 transverse YZ-chain](image)

**Figure 5.4: Phase diagram of a spin 1/2 transverse YZ-chain**

The system undergoes a transition from a paramagnetic to a ferromagnetic phase at $h = J_y + J_z$ or equivalent at $s = \frac{1}{2}$, which is marked by the vertical red line in Figure 5.4. The horizontal red line at $\gamma = 0$ and $s > \frac{1}{2}$ indicates the anisotropic phase transition between the phases $FM_y$ and $FM_z$ with a magnetic ordering in $y$-direction if $\gamma > 0 (J_y < J_z)$ and in $z$-direction if $\gamma > 0 (J_y < J_z)$.

In the following we choose $\Omega = 1$. In this case the representation of the Hamiltonian (5.7) in the operator TT format (5.4) is given by

$$A_1 = \begin{pmatrix} - (1-s)\sigma_x & s\frac{1-\gamma}{2}\sigma_y & -s\frac{1+\gamma}{2}\sigma_z & I \end{pmatrix}, \quad A_d = \begin{pmatrix} I \\ \sigma_y \\ \sigma_z \\ -(1-s)\sigma_x \end{pmatrix}$$
and
\[
A_\mu = \begin{bmatrix}
\sigma_y & 0 & 0 & 0 \\
\sigma_z & 0 & 0 & 0 \\
-(1 - s)\sigma_x & s\frac{1 - \gamma}{2}\sigma_y & -s\frac{1 + \gamma}{2}\sigma_z & I
\end{bmatrix}, \quad \text{for } \mu = 2, \ldots, d - 1.
\]

Due to the structure of the considered quantum system each tensor in TT format may result from a tensorization of a $2^d$ dimensional vector. Thus, it follows for each TT decomposition that $n_1 = n_2 = \ldots = n_d = 2$. This particular TT format has been introduced as quantized TT format (QTT) and is strongly related to the description of spin-$\frac{1}{2}$ particles, for more information, see, e.g., [61,62].

In the following numerical experiments we aim to analyze the convergence behavior of the inexact Arnoldi method using the TT format on the basis of the YZ-model (5.7). Further, we want to investigate the effects on the inexact Arnoldi method if the quantum system undergoes a phase transition.

To analyze the first aspect, i.e., the convergence behavior of the inexact Arnoldi method, we consider exemplary the YZ-model (5.7) with $s = 0.42$ and $\gamma = -0.49$. We vary the size of the Hamiltonian and consider systems with $d = 10$, $d = 20$, $d = 30$ and $d = 50$ number of particles. We are interested in the ground state energy as well as in the energy of the first and second excited state of the YZ-model. Hence, we investigate in the following examples the convergence behavior of the three smallest Ritz values, which are displayed in Table 5.1 together with the Frobenius norm of the corresponding tensor operator.

<table>
<thead>
<tr>
<th>Number of particles</th>
<th>1st smallest Ritz value</th>
<th>2nd smallest Ritz value</th>
<th>3rd smallest Ritz value</th>
<th>$|A|_F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 10$</td>
<td>-5.9714</td>
<td>-5.6119</td>
<td>-5.4863</td>
<td>$6.6 \cdot 10^1$</td>
</tr>
<tr>
<td>$d = 20$</td>
<td>-11.9678</td>
<td>-11.6408</td>
<td>-11.6005</td>
<td>$3.0 \cdot 10^3$</td>
</tr>
<tr>
<td>$d = 30$</td>
<td>-17.9641</td>
<td>-17.6446</td>
<td>-17.6251</td>
<td>$1.2 \cdot 10^5$</td>
</tr>
<tr>
<td>$d = 50$</td>
<td>-29.9570</td>
<td>-29.6415</td>
<td>-29.6340</td>
<td>$1.6 \cdot 10^8$</td>
</tr>
</tbody>
</table>

Table 5.1: The three smallest Ritz values and the Frobenius norm of the considered YZ-model for $d \in \{10, 20, 30, 50\}$

To provide more insight into the considered Hamiltonians we state in Table 5.2 for the three smallest Ritz values the gap between the considered Ritz value and the next larger Ritz value for the different problem sizes of 10, 20, 30, and 50 particles. We see that the larger the problem dimension the smaller are the gaps.
5.3 Numerical Example: YZ-model

between the considered Ritz values.

<table>
<thead>
<tr>
<th>Number of particles</th>
<th>$\text{gap}(\mu_1,\mu_2)$</th>
<th>$\text{gap}(\mu_2,\mu_3)$</th>
<th>$\text{gap}(\mu_3,\mu_4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 10$</td>
<td>0.3595</td>
<td>0.1256</td>
<td>0.1784</td>
</tr>
<tr>
<td>$d = 20$</td>
<td>0.3270</td>
<td>0.0403</td>
<td>0.1404</td>
</tr>
<tr>
<td>$d = 30$</td>
<td>0.3196</td>
<td>0.0196</td>
<td>0.0719</td>
</tr>
<tr>
<td>$d = 50$</td>
<td>0.3154</td>
<td>0.0076</td>
<td>0.0290</td>
</tr>
</tbody>
</table>

Table 5.2: Gaps between the four smallest Ritz values for $d \in \{10, 20, 30, 50\}$

To measure the quality of the approximated eigenpair $(\tilde{\mu}_k, \tilde{\gamma}_k)$ after $k$ steps of Arnoldi’s method we determine its direct-residual which is defined by

$$r_k^{\text{direct}} = \| \mathcal{A}\tilde{\gamma}_k - \tilde{\gamma}_k\tilde{\mu}_k \|_F,$$

where $\mathcal{A}$ is a tensor operator in TT format of order $d$. The Ritz tensor $\tilde{\gamma}_k$ of order $d$ equals $\tilde{V}_k\tilde{z}_k$, where $\tilde{z}_k$ is the corresponding eigenvector of the Hessenberg matrix $H_k$. Thus, we need to compute the sum of $k$ tensors $\tilde{z}(i)\tilde{v}_i$, which may lead to a large TT rank of $\tilde{\gamma}_k$ for higher $k$. Additionally, taking into account that the dimension of the problem Hamiltonian scales exponentially with the number of considered particles $d$, the computation of the direct-residual (5.8) may be too computational and memory demanding for higher values of $d$ and $k$. Therefore, we introduce the residual-after-rounding, i.e., we compute

$$r_k^{\text{round}} = \| \mathcal{A}\tilde{\gamma}_k - \tilde{\gamma}_k\tilde{\mu}_k \|_F \text{ with } \tilde{\gamma}_k = \text{round}(\tilde{\gamma}_k, \varepsilon_v).$$

(5.9)

The applied rounding procedure reduces the tensor rank of $\tilde{\gamma}_k$ considerably but also introduces an additional perturbation source. However, for large $d$, i.e., $d \geq 20$, it is also unfeasible to compute even the residual-after-rounding as we have to perform the operator-tensor multiplication $\mathcal{A}\tilde{\gamma}_k$ which is computationally very demanding. Therefore, we investigate the Arnoldi-residual

$$r_k^{\text{Arnoldi}} = \| \tilde{H}_{k+1, k} || e_k^T \tilde{z}_k \|_F \| \tilde{v}_{k+1} \|_F$$

(5.10)

which is implicitly given by Arnoldi’s method, i.e., no additional computational effort is needed. While this residual can be obtained for all dimensions it has the disadvantage that it does not stagnate at a certain level like the direct-residual $r_k^{\text{direct}}$ does. This behavior suggests that the convergence can be improved further if more iteration steps are carried out. This phenomenon was already stated in [10, 20]. Thus, the choice of the considered residuals will depend on the di-
5.3 Numerical Example: YZ-model

Figure 5.5: Experiment 1: Inexact Arnoldi method applied to the YZ-model with $d = 10$

Experiment 1. (Convergence of the three smallest Ritz values for $d = 10$)

For the first experiment we use a rather small problem size of 10 particles in order to be able to compute and compare all three residual types, i.e., the direct-residual (5.8), the residual-after-rounding (5.9) and the Arnoldi-residual (5.10). To analyze the influence of $\epsilon_v$ on the accuracy of the computed Ritz values, we investigate for the three smallest Ritz values the norms of the three residual types for $\epsilon_v \in \{10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\}$.

Figure 5.5 shows the behavior of the different residuals for $\epsilon_v = 10^{-6}$ (Figure 5.5(a)) and for $\epsilon_v = 10^{-8}$ (Figure 5.5(b)). We see that all three residuals almost coincide until the stagnation of the direct- and rounded-residual, shown by the solid and dotted lines, respectively, is reached. The stagnation of the direct- and rounded-residual occurs for $\epsilon_v = 10^{-6}$ (Figure 5.5(a)) at iteration steps 33, 45 and 51 for the smallest Ritz value, second smallest Ritz value and third smallest Ritz value, respectively. Thus, the first Ritz values converges much faster than the 2nd and the 3rd Ritz value, which is caused by the larger gap between the considered Ritz value and the remaining spectrum. As shown in Table 5.2 the gap between the first Ritz value and the remaining spectrum is larger than the gap between the second or the third smallest Ritz value and the remaining spectrum. The similar behavior of the direct- and rounded-residual indicates that the additional application of the rounding procedure after the computation of the Ritz values does not lead to a change in the accuracy of the computed values.

Figures 5.5 and 5.6 illustrate that after stagnation the direct- and rounded-residuals of all considered eigenvalues reach approximately the same level of
5.3 Numerical Example: YZ-model

Figure 5.6: Experiment 1: Inexact Arnoldi method applied to the YZ-model with $d = 10$

accuracy which depends on the choice of $\varepsilon_v$. This behavior can also be seen in Table 5.3, where the direct-residual after stagnation is stated for different values of $\varepsilon_v$. As predicted we see in Figures 5.5 and 5.6 that the behavior of the Arnoldi-residual, shown by a dashed line, suggests that the convergence can be improved even further, i.e., the norm of the Arnoldi-residual drops below the accuracy level $\varepsilon_v$.

<table>
<thead>
<tr>
<th>$\varepsilon_v$</th>
<th>Direct-residual of the smallest Ritz value</th>
<th>Direct-residual of the 2nd smallest Ritz value</th>
<th>Direct-residual of the 3rd smallest Ritz value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-6}$</td>
<td>$6.98 \cdot 10^{-7}$</td>
<td>$6.32 \cdot 10^{-7}$</td>
<td>$2.23 \cdot 10^{-6}$</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>$5.47 \cdot 10^{-8}$</td>
<td>$3.89 \cdot 10^{-8}$</td>
<td>$2.74 \cdot 10^{-8}$</td>
</tr>
<tr>
<td>$10^{-10}$</td>
<td>$6.65 \cdot 10^{-10}$</td>
<td>$3.34 \cdot 10^{-10}$</td>
<td>$8.63 \cdot 10^{-10}$</td>
</tr>
<tr>
<td>$10^{-12}$</td>
<td>$6.11 \cdot 10^{-12}$</td>
<td>$8.30 \cdot 10^{-12}$</td>
<td>$1.61 \cdot 10^{-12}$</td>
</tr>
</tbody>
</table>

Table 5.3: Experiment 1: Direct-residual after stagnation of the corresponding Ritz value and $\varepsilon_v$

The results shown in Figures 5.5 and 5.6 point to the fact that more iteration steps are necessary to achieve a higher accuracy. For example, for $\varepsilon_v = 10^{-6}$ (Figure 5.5(a)) the residual of the smallest Ritz value stagnates at iteration step 33 and for $\varepsilon_v = 10^{-12}$ (Figure 5.6(b)) this only happens at iteration step 57.
5.3 Numerical Example: YZ-model

![Graph](image1)

(a) $\epsilon_v = 10^{-6}$

![Graph](image2)

(b) $\epsilon_v = 10^{-8}$

Figure 5.7: Experiment 2: Residual-after-rounding and Arnoldi-residual of the three smallest Ritz values

Summarizing, we have seen in this first experiment that during the inexact Arnoldi method the norm of the direct-residual and the norm of the rounded-residual almost coincide. As the computation of the direct-residual is very memory and computationally demanding, we compute in the next experiment only the residual-after-rounding (5.9) and the Arnoldi-residual (5.10) to focus on a larger problem size of 20 particles.

![Graph](image3)

(a) $\epsilon_v = 10^{-10}$

![Graph](image4)

(b) $\epsilon_v = 10^{-12}$

Figure 5.8: Experiment 2: Residual-after-rounding and Arnoldi-residual of the three smallest Ritz values
5.3 Numerical Example: YZ-model

**Experiment 2. (Convergence of the three smallest Ritz values for \(d = 20\))**

In Figures 5.7 and 5.8 the residual-after-rounding and the Arnoldi-residual are shown by dotted and solid lines, respectively, for the different levels of accuracy \(\varepsilon_v \in \{10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\}\). For each \(\varepsilon_v\), both residuals nearly coincide until the residual-after-rounding stagnates, which happens approximately at the level of \(\varepsilon_v\).

As in the first example we see that also for \(d = 20\) the first Ritz value converges faster than the two others, which again is caused by the different magnitudes of the gap, see Table 5.2.

We mentioned before that the accuracy level \(\varepsilon_v\) can be considered as a tuning parameter either towards a more accurate or a less memory demanding algorithm. The first aspect, the accuracy, is illustrated in Figures 5.7 and 5.8. To visualize the second aspect, the memory requirements, we show in Figure 5.9(a) the development of the sum TT rank of the Arnoldi basis tensors and in Figure 5.9(b) the time needed to carry out each iteration step of the inexact Arnoldi method for each accuracy level \(\varepsilon_v\).

![Figure 5.9: Experiment 2: Development of the sum of TT ranks of the Arnoldi basis tensors and execution time for each iteration step of the inexact Arnoldi method for \(\varepsilon_v \in \{10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\}\) and \(d = 20\)](image)

(a) Sum of TT ranks of the Arnoldi basis tensor

(b) Execution time

- \(\varepsilon_v = 10^{-4}\)
- \(\varepsilon_v = 10^{-6}\)
- \(\varepsilon_v = 10^{-8}\)
- \(\varepsilon_v = 10^{-10}\)
- \(\varepsilon_v = 10^{-12}\)

More precisely, in Figure 5.9(a) we consider for each basis tensor the sum of the TT ranks, i.e., the sum of the \(d + 1\) tuple (5.2). We see that for each accuracy level \(\varepsilon_v\) the sum increases during the iteration process, whereby the increase of the curves depends on the accuracy level \(\varepsilon_v\). This dependency is caused by the rounding procedure applied after each tensor operation, where the tensor rank is truncated such that the approximation of the tensor is of accuracy \(\varepsilon_v\). Thus, the
lower the accuracy level the higher the sum of the TT rank of the first Arnoldi basis tensors, i.e., the higher the memory consumption of the algorithm. The same behavior is illustrated in Figure 5.9(b): the lower the chosen accuracy level \( \varepsilon_v \) the more time is needed to carry out each iteration step. This fact is mainly caused by Step 3 and 4 of Algorithm 5.3 where the basis tensor \( v_k \) is applied to the operator \( A \) and the tensor \( w_k \) is orthogonalized against the basis \( V_k \). Both steps are more time consuming if the involved tensors have higher TT ranks.

Due to the definition of the TT rank (5.2) and \( n_1 = n_2 = \ldots = n_d = 2 \) it follows that the maximum sum of TT ranks for \( d = 20 \) is 3070. We see in Figure 5.9(a) that for each accuracy level \( \varepsilon_v \) this maximum value is reached after a certain number of Arnoldi steps. Thus the truncation of the Arnoldi basis tensors to the accuracy level \( \varepsilon_v \) does not lead to a stagnation of the growth of the sum of the tensor ranks. It only affects the speed of growth, which consequently may cause problems with the analysis of higher dimensions than 20. To investigate this aspect further, we analyze in the next experiment the YZ-model for \( d = 30 \).

**Experiment 3. (Convergence of the three smallest Ritz values for \( d = 30 \))** As mentioned before, the computation of the residual-after-rounding (5.9) is quite memory demanding and not feasible for problem sizes higher than \( d = 20 \). Thus, in the following experiments we only investigate the implicitly given Arnoldi-residual (5.10). The Arnoldi-residuals of the tree smallest Ritz values are illustrated in Figures 5.10 and 5.11 for \( \varepsilon_v \in \{10^{-4}, 10^{-8}, 10^{-12}\} \).

Figure 5.10 and Figure 5.11 show that the convergence behavior of the three smallest Ritz values is similar for the different levels of accuracy \( \varepsilon_v \). For each \( \varepsilon_v \) the smallest Ritz value converges much faster than the second and the third smallest Ritz value, which is caused by the larger gap between the smallest Ritz value and the remaining spectrum, see Table 5.2.

Further, Figures 5.10 and 5.11 demonstrate that the higher the accuracy, i.e., the smaller the value of \( \varepsilon_v \), the less iteration steps can be performed due to the rapid growth of the sum of TT ranks and the resulting increase in memory demand. In particular, we achieve the maximum value of 150 iteration steps defined by the stopping criteria of Algorithm 5.3 for \( \varepsilon_v = 10^{-4} \), 115 iterations steps for \( \varepsilon_v = 10^{-8} \) and 90 iteration steps for \( \varepsilon_v = 10^{-12} \). Thus, for \( \varepsilon_v = 10^{-12} \) we can not perform enough iteration steps of the inexact Arnoldi method to ensure a sufficient accuracy of the considered Ritz values due to storage limitations. The increase in memory requirements are caused by the drastic growth of the sum of TT ranks of the Arnoldi basis tensors which is illustrated in Figure 5.12 for \( \varepsilon_v \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\} \). Summarizing, as in previous experiments a higher accuracy implies a faster growth of the sum of TT ranks. But in contrast to the results for \( d = 20 \), illustrated in Figure 5.9(a), we obtain for \( d = 30 \) no slow down of the increase of the sum of TT ranks for higher iteration steps, on the contrary, the higher the number of already performed iterations the faster the growth the sum of TT ranks of the new basis tensor. Thus, for \( d = 30 \) the
5.3 Numerical Example: YZ-model

Figure 5.10: Experiment 3: Arnoldi-residual of the three smallest Ritz values

Figure 5.11: Experiment 3: Arnoldi-residual of the three smallest Ritz values for $\varepsilon_v = 10^{-12}$
5.3 Numerical Example: YZ-model

Figure 5.12: Experiment 3: Development of the sum of TT ranks of the Arnoldi basis tensors and execution time needed for each iteration step of the inexact Arnoldi method for $\varepsilon_v \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}, 10^{-12}\}$ and $d = 30$.

sum of TT ranks of the computed Arnoldi basis tensor does not converge to a value below the maximum, which amounts to 98302, for any accuracy level $\varepsilon_v$. Especially, this last observation implies that, using the same memory capacity, we should not expect accurate results in high dimensions.

To show that the limitation of the inexact Arnoldi method regarding the problem dimension is caused by the storage restriction we consider in the following experiment the YZ-model with 50 particles.

**Experiment 4. (Convergence of the three smallest Ritz values for $d = 50$)**

In Figure 5.13 the convergence behavior of the three smallest Ritz values is illustrated by the corresponding Arnoldi-residuals (5.10) for $\varepsilon_v = 10^{-4}$ and $\varepsilon_v = 10^{-6}$. We see that all Ritz values start to converge. However, the convergence speed, especially of the second and the third smallest Ritz value, is very low, which is again caused by the small gap between the considered Ritz value to the remaining spectrum, see Table 5.2. Figure 5.13 shows that only for $\varepsilon_v = 10^{-4}$ the smallest Ritz value converges, i.e., the corresponding Arnoldi-residual reaches the accuracy level $\varepsilon_v$. Due to the memory limitation it is not possible to perform enough iterations steps to ensure the convergence of the remaining Ritz values. 

The high memory demand of each iteration step is caused by the drastic rank growth due to the high problem dimension, illustrated in Figure 5.14(a). We see that a lower accuracy level $\varepsilon_v$ implies a faster increase in the sum of the TT
5.3 Numerical Example: YZ-model

ranks of the Arnoldi basis tensors, which in turn leads to a rapid increase of the required memory.

![Graph](image1.png)

(a) $\varepsilon_v = 10^{-4}$

(b) $\varepsilon_v = 10^{-6}$

- • 1st Ritz
- • 2nd Ritz
- • 3rd Ritz
- • Arnoldi

Figure 5.13: Experiment 4: Arnoldi-residual of the three smallest Ritz values

![Graph](image2.png)

(a) Sum of TT rank of the Arnoldi basis tensor

(b) Execution time

- • $\varepsilon_v = 10^{-4}$
- • $\varepsilon_v = 10^{-6}$
- • $\varepsilon_v = 10^{-8}$

Figure 5.14: Experiment 4: Development of the sum of the TT rank of the Arnoldi basis tensors and execution time needed for each iteration step of the inexact Arnoldi method for $\varepsilon_v \in \{10^{-4}, 10^{-6}, 10^{-8}\}$ and $d = 50$

So far we have seen that an increase of the problem dimension, i.e., of the number of particles, implies an increase of the memory demand, which is analyzed in the next experiment further.
5.3 Numerical Example: YZ-model

Experiment 5. (Maximum TT rank for $\epsilon_v = 10^{-6}$) This experiment illustrates that the growth of the memory demand for increasing number of particles is only caused by the increase of the number of cores of the Arnoldi basis tensors, and not by an increase of the maximum TT rank. Therefore, we perform the inexact Arnoldi method for $d \in \{25, 30, 35, 40, 45, 50, 55, 60\}$ and stop always after the same iteration step. The comparison of the maximum TT rank of the last computed Arnoldi basis tensor is illustrated in Figure 5.15(a), where we stopped after the 70th iteration step.

![Graph](image)

(a) Maximum TT rank

![Graph](image)

(b) Arnoldi-residual

Figure 5.15: Experiment 5: Maximum TT rank of the 70th Arnoldi basis tensor and the corresponding Arnoldi-residual for $d \in \{25, 30, \ldots, 60\}$ and $\epsilon_v = 10^{-6}$

Figure 5.15(a) shows that the maximum TT rank of the 70th Arnoldi basis tensor slightly decreases with increasing number of particles. Additionally, in Figure 5.15(b), we see that the corresponding Arnoldi-residual of the smallest Ritz value increases with the number of particles.

In the previous experiments we considered the convergence of the YZ-model for one specific set of parameters $s$ and $\gamma$ to analyze the influence of the problem dimension $d$ and the accuracy level $\epsilon_v$ on the convergence of the inexact Arnoldi method. In the next example we investigate the performance of the inexact Arnoldi method for systems undergoing a phase transition from the paramagnetic to the ferromagnetic phase.

Experiment 6. (Phase transition for $d = 20$) In the YZ-model (5.7) the parameter $s$ describes the transition from a paramagnetic ground state ($s = 0$) to a ferromagnetic ground state ($s = 1$) either along the $y$-direction for $\gamma < 0$ or the $z$-direction for $\gamma > 0$.

In this experiment we want to trace the smallest eigenvalues of the YZ-model through critical points, where for example the energy gap between the ground state and the first excited state becomes very small. Therefore, we follow the
5.3 Numerical Example: YZ-model

Figure 5.16: Experiment 6: Computed ground state energy and the energy of the first and second exited state of (5.7) for $\varphi \in [-\pi, \pi]$

path illustrated by the dashed line in Figure 5.4 by using polar coordinates, i.e., we consider

$$s = r \cos(\varphi) + r \quad \text{and} \quad \gamma = r \sin(\varphi). \quad (5.11)$$

More precisely, in the following we choose $r = \frac{1}{2}$ and select 40 values of $\varphi$ equidistantly distributed between $-\pi$ and $\pi$. In this setting the energy of the ground state as well as of the first and second exited states of the YZ-model are illustrated in Figure 5.16.

Figure 5.17 illustrates that the magnitude of the Arnoldi-residual of the computed Ritz value is strongly related to the size of the gap between the considered Ritz value and the remaining spectrum. In particular, we see in Figure 5.17(b) that the Arnoldi-residual of the ground state energy is small except in a small neighborhood of $\varphi = -\frac{\pi}{2}$ and $\varphi = \frac{\pi}{2}$, where the energy gap between the ground state and the first excited state becomes small as shown in Figure 5.17(a). From the physical perspective, at these points the system undergoes a phase transition. According to Figure 5.4 the system changes from a paramagnetic to a ferromagnetic phase at $\varphi = -\frac{\pi}{2}$ and from a ferromagnetic to a paramagnetic phase at $\varphi = \frac{\pi}{2}$.

If the parameter $s$ is varied from 0 to 1 then the YZ-model (5.7) provides an interpolation between a Hamiltonian without spin-spin interaction and a Hamiltonian with ferromagnetic next-neighbor interaction in $y-$ and $z-$direction. Hence, we obtain in polar coordinates, as defined in (5.11), a simple Hamiltonian for $\varphi$ close to $\pi$ or $-\pi$, whereas a more complex Hamiltonian is obtained for $\varphi$ close to zero. This is also reflected in Figure 5.18(a), where the sum of the TT ranks of a tensor is taken over all computed Arnoldi basis tensors is illustrated. For a simple Hamiltonian, i.e., $\varphi$ around $\pi$ and $-\pi$, the sum of the TT
5.3 Numerical Example: YZ-model

Figure 5.17: Experiment 6: Energy gaps of the two smallest states and the Arnoldi residual of the ground state and first excited state energy for $\varphi \in [-\pi, \pi]$

ranks is low. However, the sum increases drastically up to the maximum value of 409999 when the influence of the next-neighbor interaction becomes more important. This increase in the sum of the TT ranks does not only effect the memory demand of the inexact Arnoldi method but also its execution time as we can see in Figure 5.18(b).

Figure 5.18: Experiment 6: Sum of the TT ranks of the tensor taken over all computed Arnoldi basis tensors and the execution time of the inexact Arnoldi method for $m = 150$ and $\varphi \in [-\pi, \pi]$
5.4 Conclusion

Based on the YZ-model (5.5) we investigated the influence of the problem dimension, i.e., the number of considered particles $d$, the accuracy level $\varepsilon_v$, and the eigenvalue distribution of the considered Hamiltonian on the convergence behavior of the inexact Arnoldi method using the TT format.

We have seen in Experiment 6 that the convergence speed of the method strongly depends on the gap between the considered eigenvalue and the remaining spectrum of the Hamiltonian. More precisely, the smaller the gap, the lower the convergence speed of the inexact Arnoldi method and the more iterations steps have to be performed to ensure convergence, implying a higher memory demand.

Experiment 1 to Experiment 4 lead to the conclusion that the number of particles influences the memory demand of the inexact Arnoldi method: the larger the number of particles the larger the memory demand. Further, Experiment 5 showed that after the same number of performed iteration steps of the inexact Arnoldi method, the maximum TT Rank of the last computed Arnoldi basis tensor does not increase with the number of considered particles. Thus, the memory demand of the inexact Arnoldi method scales only linear with the number of considered particles.

Experiment 1 to Experiment 4 also provide insights into the relationship between accuracy level $\varepsilon_v$ and the memory requirements of the inexact Arnoldi method. These insights are examplarily for $d = 20$ summarized in Figure 5.19 showing the region of convergence. The region of convergence defines the area in which, depending on the accuracy level $\varepsilon_v$, enough memory is available to ensure the convergence of the inexact Arnoldi method, i.e., enough memory for the storage of the Arnoldi basis is available.

We see in Figure 5.19 that a lower accuracy level $\varepsilon_v$ requires higher memory requirements, to ensure convergence.

To summarize, the possible accuracy of the inexact Arnoldi method using the TT format to solve eigenproblems of extremely large dimension depends crucially on the available memory.
Figure 5.19: Region of convergence for $d = 20$
6 Summary and outlook

In this thesis we investigated the behavior of Arnoldi’s method to compute extreme eigenvalues and eigenvectors of a high dimensional matrix in the presence of inexact operations. The problem considered originates from quantum computing, where the dimension of the system matrix scales with the number of particles investigated. In particular, we are interested in the most stable state of the system, which is referred to as ground state, and the ground state energy. In mathematical terms these quantities correspond to the eigenvector of the smallest eigenvalue and the smallest eigenvalue, i.e., we investigate the eigenproblem of the corresponding quantum system. To deal with the extremely large problem dimension for quantum systems of more than 20 particles we need to switch from a matrix-vector setting to a operator-tensor setting where each tensor operation is followed by a truncation procedure. This leads to a situation where operator-tensor multiplication, tensor addition and -scaling are inexact, but scalar products can be evaluated without error. To explore the consequences of the inexactness of the tensor operations we analyzed in the first part of the thesis a matrix-vector setting where a perturbation vector was added each time an operation was performed. To do so, we introduced a variant of Gram Schmidt orthogonalization, called compensated Gram Schmidt (ComGS).

We showed that the ComGS method, if necessary extended by reorthogonalization, produces a basis that is orthogonal to the same level of accuracy as the vector operations themselves. Additionally, ComGS implicitly provides a second basis that is orthogonal to machine precision, even if no reorthogonalization is performed.

We then showed that for a Hermitian $A$ the inexact Arnoldi method yields an exact Krylov relation of a nearby matrix $A + E$, which is in general not Hermitian. The key idea to ensure the hermiticity of $A + E$, was to replace the non-Hermitian Hessenberg $\tilde{H}_k$ matrix by a Hermitian matrix. We investigated several choices for this Hermitian matrix, where some of them have been even tridiagonal and proved bounds for the norm of the corresponding $E$. We concluded that $\tilde{V}_k^HA\tilde{V}_k$ and the Hermitian part of $\tilde{H}_k = G_k\tilde{H}_kG_k^{-1}$, where $G_k$ is the Cholesky factor of the actual computed cross product matrix, would be a good choices.

The inexact operations effect not only the orthogonality but also the accuracy of the computed spectral quantities. Thus, we continued by a convergence analysis of the inexact Arnoldi method or more generally of inexact Krylov subspace methods. We distinguished between an a priori and an a posteriori setting. In the first setting only information of the $l$-th iteration step were available and we assumed that the method has not converged. We analyzed how many more
6 Summary and outlook

iteration steps of an inexact Krylov subspace method are necessary to ensure convergence. Therefore, we provided a bound on how well the exact subspace $\mathcal{X}$ is contained in the search space $\mathcal{K}_k$. The derived bound is valid even in the presents of perturbations and if the gap between the desired and the remaining eigenvalues is small.

In the a posteriori setting, where information about the search space and the residual are explicitly available, the quality of the approximated eigenvalues and corresponding subspace was bounded. It turns out, that these bounds are particular sensitive to the distribution of the eigenvalues. To relax this condition and to provide meaningful bounds on the angle of inclusion of $\mathcal{X}$ and the approximated subspace $\tilde{\mathcal{Y}}$, even if the desired eigenvalues are not well separated from the remaining ones, we enlarged the approximated subspace to guarantee a sufficiently large gap between the remaining Ritz values and the desired eigenvalues.

The convergence behavior of the inexact Arnoldi method even in the presence of perturbations encouraged us to transfer the method into the tensor setting, where the inexact operations result form the application of a truncation procedure in order to avoid an unnecessary growth of the tensor rank. Thus, the truncation accuracy level $\varepsilon$ can be regarded as a tuning parameter either towards a more accurate or a less memory demanding algorithm. The applicability of the inexact Arnoldi method using the TT-format was shown examplarily on the basis of the YZ-model. The experiments clearly indicated that the inexact Arnoldi method is useful for low-rank approximations of large and sparse tensors. However, as the underlying Krylov subspace increases at every iteration step the application of the method is limited due to its memory demands.

In summary we have developed a method to compute extreme eigenvalues of high dimensional Hermitian problems taking into account inexact operations and provided a convergence analysis for this method. We note that the introduced theory on inexact operations covers a wider range of problems classes, e.g., mixed precision arithmetic and rounding.

To improve particularly the memory demand of the method developed in this thesis the following aspects could be considered in future research:

- The number of necessary Arnoldi iterations depends on properties of the operator, in particular on the size of the gap between the eigenvalue of interest and the remaining spectrum, and also on the starting tensor. In the matrix-vector setting, the high storage requirements of the Arnoldi basis and the computational costs can be reduced by using a restarted Arnoldi approach [71, 92], where the dimension of the search space is limited. Preliminary tests indicated that similar procedures applied to the inexact Arnoldi method using the TT format might improve the efficiency slightly. However, due to the rapid growth of the TT rank the method can be only used for small restart parameters.

- The inexact Arnoldi method belongs to the category of classic iterative
methods combined with low rank truncation. A different type of solution approach considers the associated reformulated trace minimization problem, more precisely the Rayleigh-quotient minimization problem, cf., [1,94]. Various heuristic approaches to solve such constrained optimization problems are available, including the Alternating Linear Scheme (ALS). The basic idea of the ALS method is to optimize each core separately and sweep over all cores repeatedly. The application of the TT format to the ALS method is considered in, e.g., [50,82]. It might be beneficial to run the inexact Arnoldi method with a large accuracy level, i.e., with a relatively low memory demand, and use its result as a promising initial guess for the ALS method to reduce the number of necessary sweeps.

- In principle instead of the inexact Arnoldi method any vector iteration method, like, e.g., Jacobi-Davidson method [103], Newton’s method [110] or Divide and Conquer method [24] could be used for solving the eigenproblem. It might be of interest to investigate the efficiency of each method, taking into account the number of operator-tensor multiplications and the convergence rate. However, these methods require a linear solve in the system matrix, which should be avoided for large problem dimensions.
Bibliography


Bibliography


Bibliography


