Recycling Krylov subspace methods for sequences of linear systems

Analysis and applications

vorgelegt von Diplom-Mathematiker

André Gaul

geboren in Köln

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:

Vorsitzende:	Prof. Dr. Noemi Kurt (TU Berlin)
1. Gutachter:	Prof. Dr. Jörg Liesen (TU Berlin)
2. Gutachter:	Prof. Dr. Reinhard Nabben (TU Berlin)
3. Gutachter:	Prof. Dr. Kees Vuik (TU Delft)

Tag der wissenschaftlichen Aussprache: 3. Juli 2014

Berlin 2014
D 83

Abstract

In several applications, one needs to solve a sequence of linear systems with changing matrices and right hand sides. This thesis concentrates on the analysis and application of recycling Krylov subspace methods for solving such sequences efficiently. The well-definedness of deflated CG, MINRES and GMRES methods and their relationship to augmentation is analyzed. Furthermore, the effects of perturbations on projections, spectra of deflated operators and Krylov subspace methods are studied. The analysis leads to convergence bounds for deflated methods which provide guidance for the automatic selection of recycling data. A novel approach is based on approximate Krylov subspaces and also gives valuable insight in the case of non-normal operators. Numerical experiments with nonlinear Schrödinger equations show that the overall time consumption is reduced by up to 40% by the derived automatic recycling strategies.

Acknowledgements

I deeply wish to thank Jenny for all the love and joy we experienced and will experience in the future – this thesis would not have been written without you!

This work greatly benefited from an enormous number of people whom I wish to thank. First of all, my advisors Jörg Liesen and Reinhard Nabben deserve special thanks for their support, research ideas and inspiring discussions. Together with Volker Mehrmann, I also thank them for keeping away from me most of the administrative tasks and I appreciate the complete freedom for my research. I am also very grateful that the DFG Forschungszentrum MATHEON supported my work in the MATHEON project C29. Furthermore, I thank Martin Gutknecht for the fruitful collaboration and Kees Vuik for accepting to referee my thesis. Special thanks go to my friend Nico Schlömer not only for the collaboration which had a tremendous impact on this work.

I am glad to have met my friends and colleagues who made my time at TU Berlin so amazing, in particular Daniel Arroyo, Manuel Baumann, Carlos Echeverría Serur, Luis García Ramos, Jan Heiland & Xiao Ai Zhou, Sebastian Holtz & family, Elisabeth Kamm & Piero and Martin Sassi, Ute Kandler, Sophia Kohle, Robert Luce, Sonja Peschutter & Henning Möldner, Daniel Pfisterer, Federico Poloni, Timo Reis, Thorsten Rohwedder & family, Olivier Sète and Patrick Winkert. Also the activists of the warm-hearted Freifunk and c-base families deserve a thank you for welcoming me and for making such important technical and political progress. Finally, I wish to thank my parents Jutta and Franz-Josef, my brother Mirko, Anna West and my whole family for the lasting support.

This thesis and the accompanying software package KryPy [60] benefited from the universe of free software packages around the Python programming language, in particular NumPy, SciPy [85], matplotlib [81], matplotlib2tikz [150], shapely [68], IPython [139], FEniCS/DOLFIN [108, 109] and PyAMG [9]. Furthermore, the LaTeX typesetting language and the PGFplots package helped to shape the appearance of this work.

Contents

Notation

$\mathbb{N}$, $\mathbb{R}$, $\mathbb{C}$	Set of non-negative integers, real numbers and complex numbers.
$\mathbb{N}_+$, $\mathbb{R}_\geq$, $\mathbb{R}_+$	Set of positive integers, non-negative real numbers and positive real numbers.
$\mathbb{i}$	Imaginary unit.
Re, Im	Real and imaginary part.
$\mathbb{P}_n$	Set of polynomials of degree at most n.
$\mathbb{P}_{n,0}$	Set of polynomials of degree at most n with value 1 at the origin: $\mathbb{P}_{n,0} = \{p \in \mathbb{P}_n \mid p(0) = 1\}$.
$\mathbb{P}_{n,\infty}$	Set of monic polynomials of degree n: $\mathbb{P}_{n,\infty} = \{p \in \mathbb{P}_n \mid p(\lambda) = \lambda^n + \sum_{i=0}^{n-1} \alpha_i \lambda^i,\ \alpha_0, \dots, \alpha_{n-1} \in \mathbb{C}\}$.
α	Scalar (Greek lowercase).
$\mathcal{V}$	Vector space (calligraphic uppercase).
v	Element of vector space $v \in \mathcal{V}$ (lowercase).
V	Tuple of vectors $V = [v_1, \dots, v_k] \in \mathcal{V}^k$ (uppercase). Also used as linear operator $V : \mathbb{C}^k \longrightarrow \mathcal{V},\ Vx \coloneqq \sum_{i=1}^{k} v_i x_i$.
$V^\star$	Adjoint of V, i.e., $V^\star : \mathcal{V} \longrightarrow \mathbb{C}^k,\ V^\star v \coloneqq \langle V, v \rangle$.
$[\![V]\!] = [\![v_1, \dots, v_k]\!]$	Linear span of $V = [v_1, \dots, v_k] \in \mathcal{V}^k$: $[\![V]\!] = [\![v_1, \dots, v_k]\!] = \operatorname{span}\{v_1, \dots, v_k\}$.
$\langle v, w \rangle$	Inner product with $\langle \alpha v, w \rangle = \overline{\alpha} \langle v, w \rangle$ and $\langle v, \alpha w \rangle = \alpha \langle v, w \rangle$.
$\langle V, W \rangle$	Block inner product for $V \in \mathcal{H}^m$ and $W \in \mathcal{H}^n$: $\langle V, W \rangle = \begin{bmatrix} \langle v_1, w_1 \rangle & \cdots & \langle v_1, w_n \rangle \\ \vdots & \ddots & \vdots \\ \langle v_m, w_1 \rangle & \cdots & \langle v_m, w_n \rangle \end{bmatrix} \in \mathbb{C}^{m,n}.$
$\mathcal{V} \oplus \mathcal{W}$	Direct sum of two subspaces $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ with $\mathcal{V} \cap \mathcal{W} = \{0\}$.
$\mathcal{L}(\mathcal{V}, \mathcal{W})$	Vector space of bounded linear operators between Hilbert spaces $\mathcal{V}$ and $\mathcal{W}$: $\mathcal{L}(\mathcal{V}, \mathcal{W}) = \{\mathsf{L} : \mathcal{V} \longrightarrow \mathcal{W} \text{ linear and bounded}\}$.
L	Linear operator between Vector spaces (sans-serif uppercase).
$\mathbf{L}$	Matrix (bold uppercase).
id	Identity operator.
$\mathbf{I}_n$	Identity matrix $\mathbf{I}_n \in \mathbb{C}^{n,n}$.
$\mathsf{L}^\star$, $\mathbf{L}^\mathsf{H}$, $\mathbf{L}^\mathsf{T}$	Adjoint operator, Hermitian transpose matrix and transpose matrix.
$\mathsf{P}_{\mathcal{V},\mathcal{W}}$	Projection onto $\mathcal{V}$ along $\mathcal{W}$.
$\mathcal{N}(\mathsf{L})$	Null space of L.
$\mathcal{R}(\mathsf{L})$	Range of L.

$\Lambda(\mathsf{L})$	Spectrum of $\mathsf{L} \in \mathcal{L}(\mathcal{V}, \mathcal{V})$.
$\Sigma(\mathsf{L})$	Singular values of L.
$\kappa(\mathsf{L})$	Condition number of linear operator $\mathsf{L} \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ with bounded inverse: $\kappa(\mathsf{L}) = \lVert \mathsf{L} \rVert \left\lVert \mathsf{L}^{-1} \right\rVert$.

1. Introduction

In a wide range of applications, one needs to solve a sequence of linear systems

$$\mathbf{A}^{(i)}x^{(i)} = b^{(i)}, \quad i \in \{1, \ldots, M\}, \tag{1.1}$$

with matrices $\mathbf{A}^{(i)} \in \mathbb{C}^{N,N}$ and right hand sides $b^{(i)} \in \mathbb{C}^{N}$. Such sequences arise, e.g., in the solution of nonlinear equations with Newton's method, time-stepping schemes for time-dependent partial differential equations and parameter optimization. For very large and sparse matrices, Krylov subspace methods are often an attractive choice for solving linear systems, e.g., when direct methods fail because of exceeding requirements of storage or computation time, and when other iterative schemes such as multigrid methods are not applicable.

This thesis is devoted to the exploration and analysis of possibilities to improve the convergence behavior of Krylov subspace methods in situations where a sequence of linear systems has to be solved. A natural approach for sequential tasks is *recycling* which – in the context of Krylov subspace methods – means to re-use information that has been computed in the solution process of a previous linear system in order to speed up the solution process for subsequent linear systems.

The idea of recycling for sequences of linear systems is not new. Here, only a brief and incomplete overview on the literature is given and it is referred to chapter 3 for more details and pointers to further literature. Recycling Krylov subspace methods for sequences of linear systems have been pushed forward since 2006 most notably by Kilmer and de Sturler [94], Parks et al. [135] and Wang, de Sturler and Paulino [181]. The history of the employed techniques goes back to the mid 1990s when restarted and nested Krylov subspace methods have been proposed, e.g., by Morgan [118, 122], de Sturler [169] and Erhel, Burrage and Pohl [47]. Two ideas predominate in these works: *augmentation* and *deflation*. In augmented methods, the search space is enlarged by a subspace that is supposed to contain useful information about the solution. In deflated methods, the matrix of the linear system is modified with a projection in order to achieve faster convergence. The origins of the latter approach are to be found in 1987 and 1988 in the works of Nicolaides [128] and Dostál [34].

A multitude of algorithmic realizations of augmented and deflated Krylov subspace methods has been proposed since then. Yet, a large number of questions concerning the *mathematical* principles and the relationship between these methods remains open.

Scope and goals. The intention of this thesis is to close several gaps both on the side of mathematical theory and on the side of practical applications of recycling

Krylov subspace methods. This includes a thorough analysis of the building blocks deflation and augmentation. A special focus lies on perturbation theory for projections, deflated matrices and Krylov subspace methods in general. For practical applications, the results provide guidance for the automatic selection of recycling data. In this work, the CG, MINRES and GMRES methods are considered because of their attractive optimality and finite termination properties (in exact arithmetic). In practice, a preconditioner is usually used in order to accelerate a Krylov subspace method. Deflation and augmentation techniques can often only unfold their potential if they are used in addition to preconditioning. Here, the topic of preconditioning is left out in large parts since, in general, it requires a problem-dependent approach. Therefore, it is assumed that the occurring linear systems already are in preconditioned form. Although special emphasis is put on mathematical aspects, this thesis is accompanied by the free software Python package KryPy [60] which offers easy-to-use and extensively tested implementations of deflated and recycling CG, MINRES and GMRES methods as well as all algorithms that are discussed in this thesis. In the following, an overview of this thesis is provided where new contributions are highlighted.

Outline. In chapter 2, the notation is introduced and well-known results about projections, angles between subspaces and Krylov subspace methods are recalled in the setting of a possibly infinite-dimensional Hilbert space and possibly singular operators. The matrix $\mathbf{A}^{(i)}$ in (1.1) may thus be replaced by a linear operator $\mathsf{A}^{(i)}$ throughout this thesis. The notation does not differ significantly from standard linear algebra notation and a brief overview of the notation in section 2.1 is provided on page 1. With the general Hilbert space presentation, the author wishes to make the contributions readily applicable to a broader range of problems such as the numerical solution of PDEs where Krylov subspace methods can operate in the natural function space and inner product of the problem. The inclusion of singular operators is of importance in the analysis of deflated methods in chapter 3. Because treatments of projections, angles between subspaces and Krylov subspace methods in general Hilbert spaces with possibly singular operators are rare or scattered across the literature on (numerical) linear algebra and functional analysis, the presentation in chapter 2 is quite extensive in order to ease the reading of the main part in chapter 3. Experienced readers may safely skip chapter 2 and only use it as a reference on demand.

Chapter 3 constitutes the main part of this thesis. The following questions are addressed:

1. How can data be incorporated in Krylov subspace methods in order to accelerate them? (Section 3.1)

2. When are deflated methods well defined and what is the relationship to augmentation? (Section 3.2)

3. How are projections, spectra of deflated operators and Krylov subspace methods affected by perturbations? (Section 3.3)
4. What is the optimal choice of deflation vectors? (Section 3.4)
5. How do the convergence bounds derived in this chapter perform in numerical experiments with non-normal operators? (Section 3.5)

In section 3.1, an overview is given of existing strategies to speed up Krylov subspace methods for sequences of linear systems by means of adapting initial guesses, updating preconditioners or augmenting search spaces.

Section 3.2 expands on deflated Krylov subspace methods based on CG, MINRES and GMRES. For two popular choices of projections, the question of well-definedness, i.e., if the method yields well-defined iterates and terminates with a solution, is systematically answered in sections 3.2.4 and 3.2.5. Theorem 3.3 was shown by the author, Gutknecht, Liesen and Nabben [62] and extends a result of Brown and Walker [18] by characterizing all linear systems with singular operators for which GMRES is well defined. For all considered deflated variants of CG, MINRES and GMRES, the same condition on the deflation subspace guarantees well-definedness. Examples 3.6 and 3.7 illustrate the possibility of breakdowns of deflated GMRES and MINRES methods if the condition on the deflation subspace is violated. In section 3.2.6, a general equivalence theorem by the author, Gutknecht, Liesen and Nabben [62] is provided which shows that two widely used variants of deflated CG and GMRES/MINRES implicitly perform augmentation. Furthermore, it is shown in section 3.2.7 that the necessary correction of iterates is equivalent to the correction of the initial guess if an appropriate projection is used as a right “preconditioner”. An overview of equivalent variants of well-defined deflated CG, MINRES and GMRES methods is provided in corollary 3.15.

Several new results for perturbed projections, deflated operators and Krylov subspace methods are presented in section 3.3. In section 3.3.1, a new bound on the normwise difference $\|\mathsf{P} - \mathsf{Q}\|$ of two *arbitrary* projections P and Q in terms of angles between the ranges and null spaces of the projections is presented. The result is stated for a general Hilbert space and also holds for infinite-dimensional subspaces. The new bound generalizes and sharpens a bound by Berkson [13] where the null spaces of both projections have to coincide.

For a finite-dimensional Hilbert space, section 3.3.2 analyzes the spectrum of deflated operators. Theorem 3.24 is a new result that characterizes the *entire* spectrum of a deflated operator for *any* invertible operator A by the behavior of A^{-1} on the orthogonal complement of the deflation subspace. The theorem does not require A to be self-adjoint or positive definite and thus generalizes a result of Nicolaides [128] who characterized the smallest nonzero and largest eigenvalue of a deflated self-adjoint and positive-definite operator A. The remaining part of the section concentrates on the case of self-adjoint but not necessarily positive-definite operators. The inertia of a deflated operator is characterized by theorem 3.27. Two further theorems provide inclusion intervals for the eigenvalues of deflated operators in the case where the deflation subspace is an approximate invariant subspace. In theorem 3.33, the new angle-based bound on the normwise difference of projections yields a bound on the deviation of the eigenvalues of the deflated operator from certain eigenvalues of the original operator. The bound involves a spectral interval gap condition and depends linearly on the norm of the residual of the approximate eigenpairs, e.g., the Ritz residual. However, a quadratic dependence on the Ritz residual can be observed in numerical experiments, see example 3.34 and figure 3.4. A quadratic Ritz residual bound for each individual eigenvalue of the deflated operator is provided in theorem 3.36. Furthermore, the bound only depends on a spectral gap which is more permissive and yields sharper bounds than the spectral interval gap. The theorem draws on a quadratic residual bound for eigenvalues by Mathias [113] and apparently is the first result that can explain why relatively poor approximations to invariant subspaces perform well as deflation subspaces in certain situations.

Section 3.3.3 discusses the behavior of Krylov subspace methods for solving a linear system $\mathsf{A}x = b$ under perturbations of the operator A and the right hand side b. Example 3.38 shows that a widespread opinion is not true in general, namely that small perturbations of the operator and the right hand side only lead to small perturbations of the convergence behavior of Krylov subspace methods. In theorem 3.41, a recently published result by Sifuentes, Embree and Morgan [153] on the behavior of GMRES with perturbed matrices is generalized in order to allow perturbations in the right hand side and the initial guess. The theorem is based on the pseudospectrum of A and is of importance in the later section 3.4.3 for the evaluation of deflation vector candidates.

The actual selection of recycling vectors in the situation of a sequence of linear systems is treated in section 3.4. Having solved a linear system (possibly with deflation), a natural question for the selection of deflation vectors is: which choice of deflation vectors performs best for the *same* linear system? The influence of changes in the operator and right hand side can then be examined in a second step.

The first subsection 3.4.1 provides formulas for the efficient computation of Ritz and harmonic Ritz pairs with the corresponding residual norms from the subspaces that are available after a run of the deflated Krylov subspace methods from section 3.2. This subsection can be seen as a reference for implementations and can be skipped if implementational details are not of interest.

In section 3.4.2, the quadratic residual bound from theorem 3.36 is used to derive inclusion intervals for the eigenvalues of the next deflated operator in a sequence of linear systems if the deflation vectors are chosen as Ritz vectors from the subspaces that were constructed during the solution process of the current linear system, cf. theorem 3.50. For a given set of Ritz vectors, the eigenvalue inclusion bounds can be used with a priori bounds in order to estimate the convergence behavior of the deflated Krylov subspace method for the next linear system.

Section 3.4.3 proposes a novel approach for assessing a given set of Ritz vectors that is based on the construction of approximate Krylov subspaces and is also valid for non-self-adjoint operators. Given a deflation subspace candidate, the question is: how can the convergence behavior of a Krylov subspace method with this deflation subspace be characterized by only using already computed data? The Krylov subspace that would be constructed is not contained in the available subspaces in general. However, a Krylov subspace for a nearby operator and right hand side can be constructed. Given any subspace $\mathcal{V}$ and the action of A on this subspace, theorem 3.53 shows how the operator A can be optimally perturbed *in each step* such that the provided subspace $\mathcal{V}$ constitutes a Krylov subspace of the perturbed operator with respect to a *specific* initial vector $v \in \mathcal{V}$. Theorem 3.59 shows that an Arnoldi relation for such an approximate Krylov subspace is readily available for the given deflation subspace candidate. In theorem 3.64, a new residual bound for deflated GMRES/MINRES is presented based on the approximate Krylov subspace technique. The resulting bound features a residual norm that is efficiently computable and a pseudospectral term that takes account of the perturbations that arise from the approximate Krylov subspace and the fact that the next linear system has to be considered. Example 3.72 shows that – unlike asymptotic bounds – the new bound is able to capture different phases of the GMRES convergence, e.g., from a transient phase of stagnation to a fast residual norm reduction.

Numerical experiments with a non-normal operator resulting from a finite element discretization of the convection-diffusion equation confirm the usefulness of the new approximate Krylov subspace bound in section 3.5. It is observed that the cheaply computable part of the bound is able to capture the actual convergence behavior way beyond the point where the full bound is not able to provide significant information due to the growth of the pseudospectral term.

While the experiments in chapter 3 are mostly based on academic examples and model problems in order to illustrate the mathematical theory, chapter 4 presents an application of the proposed recycling strategies to a more realistic problem. The experiments show how the numerical solution of nonlinear Schrödinger equations can benefit from the use of recycling MINRES methods to solve linear systems that arise from Newton's method. The results in this chapter as well as the underlying code PyNosh [64] have been developed jointly by Schlömer and the author and have been published to a great extent in [63]. The numerical experiments are shown for the Ginzburg–Landau equation which is an important instance of nonlinear Schrödinger equations. In addition to a fixed choice of deflation vectors that was used in [63], automatic recycling strategies are employed here that are based on the new bounds in this thesis. The recycling strategies do not require user interaction and yield the same reduction of the overall solution time as the manually optimized number of deflation vectors.

All experiments in this thesis can be reproduced with the available source code[1] and the free software packages KryPy [60], PseudoPy [61] and PyNosh [64] that emerged from this thesis and the related work in [63].

[1] `https://github.com/andrenarchy/phdthesis`

2. Background: projections and Krylov subspace methods

This chapter gives an overview of Krylov subspace methods and recalls important results that are helpful in the analysis in chapter 3. Projections play a major role in this thesis due to the fact that the studied Krylov subspace methods CG, MINRES and GMRES fulfill an optimality condition which can be characterized mathematically by projections. Furthermore, the deflation techniques that are the subject of chapter 3 are also based on projections. For this reason the first sections of this chapter are dedicated to the characterization of projections and angles between subspaces of Hilbert spaces. The author found that many important results on this topic are scattered across the literature on (numerical) linear algebra and functional analysis. In order to keep the later sections and chapters comprehensible, the presentation is a bit more extensive than in most treatises of Krylov subspace methods. Furthermore, the results in sections 2.1–2.7 are stated for a general (possibly infinite-dimensional) Hilbert space $\mathcal{H}$ for the sake of generality and reusability. Especially the arbitrary inner product is of relevance in the application to the nonlinear Schrödinger equations in chapter 4. When it comes to practical numerical aspects in later sections, a transition to the finite-dimensional case $\dim \mathcal{H} < \infty$ is made. Krylov subspaces in the setting of a possibly infinite-dimensional Hilbert space with an arbitrary inner product have been considered in a similar manner in the works of Eiermann, Ernst and Schneider [43] and Eiermann and Ernst [42]. In contrast to these works, the presentation in this chapter does not assume that the operators at hand are nonsingular. Taking into account possibly singular operators is of major importance in the analysis of deflated Krylov subspace methods in chapter 3. Iterative methods have also been considered in the context of Hilbert spaces by Kirby [95] and for the derivation of stopping criteria by Arioli, Noulard and Russo [4] and Arioli, Loghin and Wathen [3]. A recent work by Málek and Strakoš [110] stresses the need for a unified treatment of functional analysis and numerical linear algebra in the solution of partial differential equations.

2.1. Preliminaries

This section introduces the basic notation and recalls basic properties of bounded linear operators between Hilbert spaces. If questions concerning the notation arise while reading, a glance at the notational overview on page 1 may be worthwhile. In the following, $\mathcal{H}$ denotes a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H}\times\mathcal{H} \longrightarrow \mathbb{C}$ where $\langle v, \alpha w\rangle_{\mathcal{H}} = \alpha\langle v, w\rangle_{\mathcal{H}}$ and $\langle \alpha v, w\rangle_{\mathcal{H}} = \overline{\alpha}\langle v, w\rangle_{\mathcal{H}}$ for $v, w \in \mathcal{H}$ and $\alpha \in \mathbb{C}$. The induced norm $\|\cdot\|_{\mathcal{H}}$ is defined by $\|v\|_{\mathcal{H}} := \sqrt{\langle v, v\rangle_{\mathcal{H}}}$ for $v \in \mathcal{H}$. If $\mathcal{G}$ is also a Hilbert space, then a linear operator $\mathsf{L} : \mathcal{H} \longrightarrow \mathcal{G}$ is called bounded if there exists an $\alpha \in \mathbb{R}$ with $\|\mathsf{L}v\|_{\mathcal{G}} \le \alpha \|v\|_{\mathcal{H}}$ for all $v \in \mathcal{H}$. By $\mathcal{L}(\mathcal{H}, \mathcal{G})$, the vector space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$ is denoted and $\mathcal{L}(\mathcal{H}) := \mathcal{L}(\mathcal{H}, \mathcal{H})$. For an $\mathsf{L} \in \mathcal{L}(\mathcal{H}, \mathcal{G})$ the sets $\mathcal{R}(\mathsf{L})$ and $\mathcal{N}(\mathsf{L})$ denote the range and the null space of the operator and the operator norm is defined by $\|\mathsf{L}\| := \sup_{0 \neq v \in \mathcal{H}} \frac{\|\mathsf{L}v\|_{\mathcal{G}}}{\|v\|_{\mathcal{H}}}$. The identity operator is denoted by id. For $v, w \in \mathcal{H}$, the condition $v \perp_{\mathcal{H}} w$ is equivalent to $\langle v, w\rangle_{\mathcal{H}} = 0$. For a subspace $\mathcal{V} \subseteq \mathcal{H}$, the orthogonal complement of $\mathcal{V}$ is defined by $\mathcal{V}^{\perp} := \{z \in \mathcal{H} \mid z \perp_{\mathcal{H}} v \text{ for all } v \in \mathcal{V}\}$ and the closure is defined by $\overline{\mathcal{V}} := \{z \in \mathcal{H} \mid \text{there exists a sequence } (v_n)_{n\in\mathbb{N}} \text{ in } \mathcal{V} \text{ with } v_n \to z\}$. The subspace $\mathcal{V}$ is called closed if $\mathcal{V} = \overline{\mathcal{V}}$. A linear operator $\mathsf{L} \in \mathcal{L}(\mathcal{H})$ can be applied to a subspace with $\mathsf{L}\mathcal{V} := \{\mathsf{L}v \mid v \in \mathcal{V}\}$. Two subspaces $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ form a direct sum $\mathcal{V} \oplus \mathcal{W}$ if $\mathcal{V} \cap \mathcal{W} = \{0\}$. For tuples of vectors $V = [v_1, \ldots, v_m] \in \mathcal{H}^m$ and $W = [w_1, \ldots, w_n] \in \mathcal{H}^n$, the block inner product is defined as the matrix

$$\langle V, W\rangle_{\mathcal{H}} = \begin{bmatrix} \langle v_1, w_1\rangle_{\mathcal{H}} & \cdots & \langle v_1, w_n\rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle v_m, w_1\rangle_{\mathcal{H}} & \cdots & \langle v_m, w_n\rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{C}^{m,n}$$

and the subspace that is spanned by the elements of a tuple is denoted by $\llbracket V \rrbracket = \llbracket v_1, \ldots, v_m \rrbracket = \operatorname{span}\{v_1, \ldots, v_m\}$. If $V \in \mathcal{H}^m$ is interpreted as a linear operator $V : \mathbb{C}^m \longrightarrow \mathcal{H}$, its adjoint $V^\star : \mathcal{H} \longrightarrow \mathbb{C}^m$ is defined by $V^\star z = \langle V, z\rangle$ for $z \in \mathcal{H}$. For two tuples $V \in \mathcal{H}^m$ and $W \in \mathcal{H}^n$, the adjoint of $[V, W]$ is given by $[V, W]^\star = [V^\star, W^\star]^{\mathsf{T}}$. Subscripts are dropped for reasons of readability if it is unambiguous.

Lemma 2.1. *Let $\mathcal{G}$ and $\mathcal{H}$ be two Hilbert spaces and let $\mathsf{L} \in \mathcal{L}(\mathcal{G}, \mathcal{H})$. Then the following holds:*

1. *There exists a unique bounded linear operator $\mathsf{L}^\star \in \mathcal{L}(\mathcal{H}, \mathcal{G})$, the adjoint of L, with $\langle \mathsf{L}v, w\rangle_{\mathcal{H}} = \langle v, \mathsf{L}^\star w\rangle_{\mathcal{G}}$ for all $v \in \mathcal{G}$ and $w \in \mathcal{H}$ and $\|\mathsf{L}\| = \|\mathsf{L}^\star\|$.*

2. *$\mathcal{N}(\mathsf{L}^\star) = \mathcal{R}(\mathsf{L})^\perp$ and $\overline{\mathcal{R}(\mathsf{L}^\star)} = \mathcal{N}(\mathsf{L})^\perp$.*

Proof. Cf. [182, Satz V.5.2]. □

Lemma 2.2. *Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces and $\mathsf{L}, \mathsf{M} \in \mathcal{L}(\mathcal{G}, \mathcal{H})$ with closed ranges such that $\mathcal{R}(\mathsf{L}) \perp_{\mathcal{H}} \mathcal{R}(\mathsf{M})$ and $\mathcal{N}(\mathsf{L})^\perp \perp_{\mathcal{G}} \mathcal{N}(\mathsf{M})^\perp$. Then the sum $\mathsf{L} + \mathsf{M}$ satisfies $\|\mathsf{L} + \mathsf{M}\| = \max\{\|\mathsf{L}\|, \|\mathsf{M}\|\}$.*

Proof. It is first shown that $\|\mathsf{L} + \mathsf{M}\| \le \max\{\|\mathsf{L}\|, \|\mathsf{M}\|\}$.

$$\begin{aligned} \|\mathsf{L} + \mathsf{M}\|^2 &= \sup_{\substack{z \in \mathcal{G} \\ \|z\|_{\mathcal{G}} = 1}} \|(\mathsf{L} + \mathsf{M})z\|_{\mathcal{H}}^2 = \sup_{\substack{v \in \mathcal{N}(\mathsf{L})^\perp, w \in \mathcal{N}(\mathsf{M})^\perp \\ \|v + w\|_{\mathcal{G}} = 1}} \|(\mathsf{L} + \mathsf{M})(v + w)\|_{\mathcal{H}}^2 \\ &= \sup_{\substack{v \in \mathcal{N}(\mathsf{L})^\perp, w \in \mathcal{N}(\mathsf{M})^\perp \\ \|v + w\|_{\mathcal{G}} = 1}} \|\mathsf{L}v + \mathsf{M}w\|_{\mathcal{H}}^2 = \sup_{\substack{v \in \mathcal{N}(\mathsf{L})^\perp, w \in \mathcal{N}(\mathsf{M})^\perp \\ \|v + w\|_{\mathcal{G}} = 1}} \left(\|\mathsf{L}v\|_{\mathcal{H}}^2 + \|\mathsf{M}w\|_{\mathcal{H}}^2\right) \end{aligned}$$

$$\begin{aligned}
&\leq \sup_{\substack{v\in\mathcal{N}(\mathsf{L})^\perp, w\in\mathcal{N}(\mathsf{M})^\perp\\ \|v+w\|_{\mathcal{G}}=1}} \left(\|\mathsf{L}\|_{\mathcal{H}}^2 \|v\|_{\mathcal{G}}^2 + \|\mathsf{M}\|_{\mathcal{H}}^2 \|w\|_{\mathcal{G}}^2\right)\\
&\leq \max\{\|\mathsf{L}\|^2, \|\mathsf{M}\|^2\} \sup_{\substack{v\in\mathcal{N}(\mathsf{L})^\perp, w\in\mathcal{N}(\mathsf{M})^\perp\\ \|v+w\|_{\mathcal{G}}=1}} \left(\|v\|_{\mathcal{G}}^2 + \|w\|_{\mathcal{G}}^2\right)\\
&\leq \max\{\|\mathsf{L}\|^2, \|\mathsf{M}\|^2\} \sup_{\substack{v\in\mathcal{N}(\mathsf{L})^\perp, w\in\mathcal{N}(\mathsf{M})^\perp\\ \|v+w\|_{\mathcal{G}}=1}} \|v+w\|_{\mathcal{G}}^2 = \max\{\|\mathsf{L}\|^2, \|\mathsf{M}\|^2\}.
\end{aligned}$$

Without loss of generality $\|\mathsf{L}\| = \max\{\|\mathsf{L}\|, \|\mathsf{M}\|\}$. Then the proof is complete with the following computation:

$$\|\mathsf{L}+\mathsf{M}\|^2 = \sup_{\substack{z\in\mathcal{G}\\ z\neq 0}} \frac{\|(\mathsf{L}+\mathsf{M})z\|_{\mathcal{H}}^2}{\|z\|_{\mathcal{G}}^2} = \sup_{\substack{z\in\mathcal{G}\\ z\neq 0}} \frac{\|\mathsf{L}z\|_{\mathcal{H}}^2 + \|\mathsf{M}z\|_{\mathcal{H}}^2}{\|z\|_{\mathcal{G}}^2} \geq \|\mathsf{L}\|^2.$$

□

2.2. Projections

In this section, the basic properties of projections on a Hilbert space are presented. Most books on (numerical) linear algebra and functional analysis include projections [90, 189, 147, 182, 148], but many are restricted to orthogonal projections or only cover a subset of characterizations that are of importance in later sections. All results in this and the following section are known but scattered in the literature and are repeated here with short proofs for convenience. A quite general treatment of projections in Hilbert spaces can be found in chapter 7 of the book by Galántai [59].

Definition 2.3. A linear operator $\mathsf{P} \in \mathcal{L}(\mathcal{H})$ is called a *projection* if $\mathsf{P}^2 = \mathsf{P}$.

The next two lemmas show that a projection $\mathsf{P} \in \mathcal{L}(\mathcal{H})$ can be characterized completely by its range and null space.

Lemma 2.4. *Let $\mathsf{P} \in \mathcal{L}(\mathcal{H})$ be a projection. Then $\mathcal{H} = \mathcal{R}(\mathsf{P}) \oplus \mathcal{N}(\mathsf{P})$, where both $\mathcal{R}(\mathsf{P})$ and $\mathcal{N}(\mathsf{P})$ are closed subspaces.*

Proof. Each $z \in \mathcal{H}$ can be decomposed as $z = \mathsf{P}z + (z - \mathsf{P}z)$, where $\mathsf{P}z \in \mathcal{R}(\mathsf{P})$ and $z - \mathsf{P}z \in \mathcal{N}(\mathsf{P})$ because $\mathsf{P}(z - \mathsf{P}z) = \mathsf{P}z - \mathsf{P}^2 z = \mathsf{P}z - \mathsf{P}z = 0$. Let $v \in \mathcal{R}(\mathsf{P}) \cap \mathcal{N}(\mathsf{P})$ arbitrary. Then $v = \mathsf{P}v$ and $\mathsf{P}v = 0$ yield $v = 0$ which shows that $\mathcal{R}(\mathsf{P}) \cap \mathcal{N}(\mathsf{P}) = \{0\}$ and the decomposition is thus unique.

The null space of a bounded linear operator is always closed. In order to see that also its range is closed one can proceed as in [59, Theorem 7.13]. Let $w = \lim_{n\to\infty} \mathsf{P}v_n \in \overline{\mathcal{R}(\mathsf{P})}$ with $v_n \in \mathcal{H}$. Then $\mathsf{P}w = \lim_{n\to\infty} \mathsf{P}^2 v_n = \lim_{n\to\infty} \mathsf{P}v_n = w$ and thus $w \in \mathcal{R}(\mathsf{P})$. □

Lemma 2.5. *Let $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ be two closed subspaces such that $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$. Then there exists a unique projection $\mathsf{P} \in \mathcal{H}$ with $\mathcal{R}(\mathsf{P}) = \mathcal{V}$ and $\mathcal{N}(\mathsf{P}) = \mathcal{W}$.*

Proof. Cf. [90, Chapter III §5.4] or [59, Theorem 7.17]. □

In the light of these two lemmas the following notation for projections is introduced:

Definition 2.6. For two closed subspaces $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ with $\mathcal{V} \oplus \mathcal{W} = \mathcal{H}$ the operator $\mathsf{P}_{\mathcal{V},\mathcal{W}} \in \mathcal{L}(\mathcal{H})$ is defined as the unique projection with range $\mathcal{R}(\mathsf{P}_{\mathcal{V},\mathcal{W}}) = \mathcal{V}$ and null space $\mathcal{N}(\mathsf{P}_{\mathcal{V},\mathcal{W}}) = \mathcal{W}$. The projection $\mathsf{P}_{\mathcal{V},\mathcal{W}}$ is called the *projection onto $\mathcal{V}$ along $\mathcal{W}$*.

The adjoint and complementary operator of a projection are investigated in the following lemma.

Lemma 2.7. *Let $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ be two closed subspaces such that $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$. Then:*

1. *Also the orthogonal complements form a direct sum: $\mathcal{H} = \mathcal{V}^\perp \oplus \mathcal{W}^\perp$.*
2. *The adjoint operator of the projection $\mathsf{P}_{\mathcal{V},\mathcal{W}}$ is given by $\mathsf{P}^\star_{\mathcal{V},\mathcal{W}} = \mathsf{P}_{\mathcal{W}^\perp,\mathcal{V}^\perp}$ and satisfies $\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathcal{V}^\perp}\right\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|$.*
3. *The complementary operator of the projection $\mathsf{P}_{\mathcal{V},\mathcal{W}}$ is given by $\mathsf{id} - \mathsf{P}_{\mathcal{V},\mathcal{W}} = \mathsf{P}_{\mathcal{W},\mathcal{V}}$. If $\mathcal{V}, \mathcal{W} \neq \mathcal{H}$, then $\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|$.*

Proof. 1. It is first shown that $\mathsf{P}^\star_{\mathcal{V},\mathcal{W}}$ is a projection and then 1. and 2. are obtained by determining its range and null space. In order to see that $\mathsf{P}^\star_{\mathcal{V},\mathcal{W}}$ actually is a projection, note that

$$\begin{aligned}\left\|\mathsf{P}^\star_{\mathcal{V},\mathcal{W}} - (\mathsf{P}^\star_{\mathcal{V},\mathcal{W}})^2\right\|^2 &= \sup_{\|z\|=1} \left\|\left(\mathsf{P}^\star_{\mathcal{V},\mathcal{W}} - (\mathsf{P}^\star_{\mathcal{V},\mathcal{W}})^2\right) z\right\|^2 \\ &= \sup_{\|z\|=1} \left\langle z, (\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}^2_{\mathcal{V},\mathcal{W}}) \left(\mathsf{P}^\star_{\mathcal{V},\mathcal{W}} - (\mathsf{P}^\star_{\mathcal{V},\mathcal{W}})^2\right) z\right\rangle = 0.\end{aligned}$$

It follows from lemma 2.1 that $\left\|\mathsf{P}^\star_{\mathcal{V},\mathcal{W}}\right\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|$ and thus the adjoint operator $\mathsf{P}^\star_{\mathcal{V},\mathcal{W}}$ is a projection. Furthermore, the lemma reveals that $\mathcal{N}(\mathsf{P}^\star_{\mathcal{V},\mathcal{W}}) = \mathcal{R}(\mathsf{P}_{\mathcal{V},\mathcal{W}})^\perp = \mathcal{V}^\perp$ and $\overline{\mathcal{R}(\mathsf{P}^\star_{\mathcal{V},\mathcal{W}})} = \mathcal{N}(\mathsf{P}_{\mathcal{V},\mathcal{W}})^\perp = \mathcal{W}^\perp$. Because $\mathsf{P}^\star_{\mathcal{V},\mathcal{W}}$ is a projection it can be deduced from lemma 2.4 that $\mathcal{R}(\mathsf{P}^\star_{\mathcal{V},\mathcal{W}}) = \overline{\mathcal{R}(\mathsf{P}^\star_{\mathcal{V},\mathcal{W}})} = \mathcal{W}^\perp$ and that $\mathcal{H} = \mathcal{V}^\perp \oplus \mathcal{W}^\perp$. Thus $\mathsf{P}^\star_{\mathcal{V},\mathcal{W}} = \mathsf{P}_{\mathcal{W}^\perp,\mathcal{V}^\perp}$.

3. Let $z = v + w \in \mathcal{H}$ be arbitrary with $v \in \mathcal{V}$ and $w \in \mathcal{W}$. Then $v = \mathsf{P}_{\mathcal{V},\mathcal{W}} v$ and $w = \mathsf{P}_{\mathcal{W},\mathcal{V}} w$ and the following equation can be obtained:

$$\begin{aligned}(\mathsf{id} - \mathsf{P}_{\mathcal{V},\mathcal{W}}) z &= z - \mathsf{P}_{\mathcal{V},\mathcal{W}}(v + w) = z - \mathsf{P}_{\mathcal{V},\mathcal{W}} v = z - v = w \\ &= \mathsf{P}_{\mathcal{W},\mathcal{V}} w = \mathsf{P}_{\mathcal{W},\mathcal{V}}(v + w) = \mathsf{P}_{\mathcal{W},\mathcal{V}} z.\end{aligned}$$

For the norm equality $\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|$ see, e.g., Szyld [171].

□

The norm equality $\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|$ in lemma 2.7 has been proven in many ways, e.g., by Del Pasqua [30] in 1955, Ljance [107] in 1958 and Kato [89]. An extensive and readable overview of several proofs is given by Szyld in [171, Theorem 2.1].

The next lemma characterizes projections where the range and null space are given as sums of subspaces that satisfy certain orthogonality constraints.

Lemma 2.8. *Let $\mathcal{V},\mathcal{W},\mathcal{X},\mathcal{Y} \subseteq \mathcal{H}$ be closed subspaces such that $\mathcal{H} = \mathcal{V}\oplus\mathcal{W}^\perp = \mathcal{X}\oplus\mathcal{Y}^\perp$, $\mathcal{V} \perp \mathcal{Y}$ and $\mathcal{X} \perp \mathcal{W}$. Then also $\mathcal{H} = \mathcal{V} \oplus \mathcal{X} \oplus (\mathcal{W} \oplus \mathcal{Y})^\perp$ and*

$$\mathsf{P}_{\mathcal{V}+\mathcal{X},(\mathcal{W}+\mathcal{Y})^\perp} = \mathsf{P}_{\mathcal{V},\mathcal{W}^\perp} + \mathsf{P}_{\mathcal{X},\mathcal{Y}^\perp}.$$

If furthermore $\mathcal{V} \perp \mathcal{X}$ and $\mathcal{W} \perp \mathcal{Y}$, then

$$\left\|\mathsf{P}_{\mathcal{V}+\mathcal{X},(\mathcal{W}+\mathcal{Y})^\perp}\right\| = \max\{\left\|\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}\right\|, \left\|\mathsf{P}_{\mathcal{X},\mathcal{Y}^\perp}\right\|\}.$$

Proof. From the orthogonality conditions it can be seen that $\mathcal{H} = \mathcal{V}\oplus\mathcal{W}^\perp = \mathcal{V}\oplus(\mathcal{W}^\perp\cap\mathcal{X})\oplus(\mathcal{W}^\perp\cap\mathcal{X}^\perp) = \mathcal{V}\oplus\mathcal{X}\oplus(\mathcal{W}^\perp\cap\mathcal{X}^\perp)$ and thus $\mathcal{V}+\mathcal{X} = \mathcal{V}\oplus\mathcal{X}$ indeed is a direct sum. Analogously, $\mathcal{W}+\mathcal{Y} = \mathcal{W}\oplus\mathcal{Y}$ holds because $\mathcal{H} = \mathcal{W}\oplus\mathcal{V}^\perp = \mathcal{W}\oplus(\mathcal{V}^\perp\cap\mathcal{Y})\oplus(\mathcal{V}^\perp\cap\mathcal{Y}^\perp) = \mathcal{W} \oplus \mathcal{Y} \oplus (\mathcal{V}^\perp \cap \mathcal{Y}^\perp)$. Now it is verified that the operator $\mathsf{Q} \coloneqq \mathsf{P}_{\mathcal{V},\mathcal{W}^\perp} + \mathsf{P}_{\mathcal{X},\mathcal{Y}^\perp}$ is a projection by computing

$$\mathsf{Q}^2 = \mathsf{P}^2_{\mathcal{V},\mathcal{W}^\perp} + \mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}\mathsf{P}_{\mathcal{X},\mathcal{Y}^\perp} + \mathsf{P}_{\mathcal{X},\mathcal{Y}^\perp}\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp} + \mathsf{P}^2_{\mathcal{X},\mathcal{Y}^\perp} = \mathsf{P}_{\mathcal{V},\mathcal{W}^\perp} + \mathsf{P}_{\mathcal{X},\mathcal{Y}^\perp} = \mathsf{Q}.$$

Obviously $(\mathcal{W}\oplus\mathcal{Y})^\perp \subseteq \mathcal{N}(\mathsf{Q})$. However, $\mathsf{Q}z \neq 0$ for every nonzero $z \in \mathcal{W}\oplus\mathcal{Y}$ because $\mathcal{V}$ and $\mathcal{X}$ form a direct sum $\mathcal{V} \oplus \mathcal{X}$. Thus the null space of Q is $\mathcal{N}(\mathsf{Q}) = (\mathcal{W} \oplus \mathcal{Y})^\perp$. Furthermore, it can be seen that $\mathsf{Q}\mathcal{W} = \mathcal{V}$ and $\mathsf{Q}\mathcal{Y} = \mathcal{X}$ and thus $\mathcal{R}(\mathsf{Q}) = \mathcal{V} \oplus \mathcal{X}$ and $\mathsf{Q} = \mathsf{P}_{\mathcal{V}+\mathcal{X},(\mathcal{W}+\mathcal{Y})^\perp}$.

The norm property directly follows from lemma 2.2. □

An important class of projections are orthogonal projections where the null space is the orthogonal complement of the range:

Definition 2.9. For a closed subspace $\mathcal{V} \subseteq \mathcal{H}$ the *orthogonal projection* onto $\mathcal{V}$ is defined by $\mathsf{P}_{\mathcal{V}} \coloneqq \mathsf{P}_{\mathcal{V},\mathcal{V}^\perp}$. A projection that is not orthogonal is called an *oblique projection.*

Orthogonal projections have attractive properties: they are self-adjoint and for any vector $z \in \mathcal{H}$ and a closed subspace $\mathcal{V} \subseteq \mathcal{H}$, the orthogonal projection $\mathsf{P}_{\mathcal{V}}$ provides the vector in $\mathcal{V}$ that is closest to z:

Theorem 2.10. *Let $\mathcal{V} \subseteq \mathcal{H}$ be a closed subspace. Then*

1. *$\mathsf{P}_{\mathcal{V}}$ is self-adjoint, i.e., $\mathsf{P}^\star_{\mathcal{V}} = \mathsf{P}_{\mathcal{V}}$.*

2. *If $z \in \mathcal{H}$, then*
$$\|z - \mathsf{P}_{\mathcal{V}}z\| = \inf_{v\in\mathcal{V}} \|z - v\|.$$

Proof. The first statement immediately follows from the definition of an orthogonal projection and statement 2 of lemma 2.7. For the second statement the following holds for $v \in \mathcal{V}$

$$\|z - v\|^2 = \|z - \mathsf{P}_{\mathcal{V}}z + \mathsf{P}_{\mathcal{V}}z - v\|^2 = \|z - \mathsf{P}_{\mathcal{V}}z\|^2 + \|\mathsf{P}_{\mathcal{V}}z - v\|^2 \geq \|z - \mathsf{P}_{\mathcal{V}}z\|^2,$$

where the second equality holds because $z - \mathsf{P}_{\mathcal{V}}z = \mathsf{P}_{\mathcal{V}^\perp}z \perp \mathsf{P}_{\mathcal{V}}z - v = \mathsf{P}_{\mathcal{V}}(z - v)$. □

The orthogonal projection onto the sum of two subspaces can be represented as the sum of two projections as demonstrated in the next lemma.

Lemma 2.11. *Let $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ be closed subspaces. Then*

$$\mathsf{P}_{\mathcal{V}+\mathcal{W}} = \mathsf{P}_{\mathcal{V}} + \mathsf{P}_{\mathsf{P}_{\mathcal{V}^\perp}\mathcal{W}}$$

Proof. The subspace $\mathcal{V} + \mathcal{W}$ can be decomposed into

$$\mathcal{V} + \mathcal{W} = \mathsf{P}_{\mathcal{V}}(\mathcal{V} + \mathcal{W}) + \mathsf{P}_{\mathcal{V}^\perp}(\mathcal{V} + \mathcal{W}) = \mathcal{V} + \mathsf{P}_{\mathcal{V}^\perp}\mathcal{W}.$$

The result then follows with lemma 2.8 because $\mathcal{V} \perp \mathsf{P}_{\mathcal{V}^\perp}\mathcal{W}$. □

In later sections and chapter 3, the ranges and null spaces of projections are defined via the action of an operator on subspaces. In these cases, the following lemma can be helpful.

Lemma 2.12. *Let $\mathsf{L} \in \mathcal{L}(\mathcal{H})$ and let $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ be two closed subspaces such that $\mathsf{L}\mathcal{V} \oplus \mathcal{W}^\perp = \mathcal{H}$ and $\mathsf{L}|_{\mathcal{V}}$ is invertible with bounded inverse. Then*

$$\mathcal{V} \oplus (\mathsf{L}^\star\mathcal{W})^\perp = \mathcal{H} \qquad \text{and} \qquad \mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L} = \mathsf{L}\mathsf{P}_{\mathcal{V},(\mathsf{L}^\star\mathcal{W})^\perp}.$$

Proof. Let $\mathsf{Q} := (\mathsf{L}|_{\mathcal{V}})^{-1}\mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L}$ and first note that $\mathsf{L}\mathsf{Q} = \mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L}$ and that

$$\mathsf{Q}^2 = (\mathsf{L}|_{\mathcal{V}})^{-1}\mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L}(\mathsf{L}|_{\mathcal{V}})^{-1}\mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L} = (\mathsf{L}|_{\mathcal{V}})^{-1}\mathsf{P}^2_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L} = (\mathsf{L}|_{\mathcal{V}})^{-1}\mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L} = \mathsf{Q}$$

is a projection because $(\mathsf{L}|_{\mathcal{V}})^{-1}$, $\mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}$ and L are bounded. Now it is shown that $\mathcal{V} \oplus (\mathsf{L}^\star\mathcal{W})^\perp$, $\mathcal{R}(\mathsf{Q}) = \mathcal{V}$ and $\mathcal{N}(\mathsf{Q}) = \mathcal{R}(\mathsf{id} - \mathsf{Q}) = (\mathsf{L}^\star\mathcal{W})^\perp$ which can be seen from the decomposition of any $z \in \mathcal{H}$ into $z = \mathsf{Q}z + (\mathsf{id} - \mathsf{Q})z$. Then $\mathsf{Q}z = (\mathsf{L}|_{\mathcal{V}})^{-1}\mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp}\mathsf{L}z \in \mathcal{V}$ and $(\mathsf{id} - \mathsf{Q})z \in (\mathsf{L}^\star\mathcal{W})^\perp$ can be concluded from the fact that for any $w \in \mathcal{W}$

$$\langle \mathsf{L}^\star w, (\mathsf{id} - \mathsf{Q})z \rangle = \langle w, \mathsf{L}(\mathsf{id} - \mathsf{Q})z \rangle = \left\langle w, (\mathsf{id} - \mathsf{P}_{\mathsf{L}\mathcal{V},\mathcal{W}^\perp})\mathsf{L}z \right\rangle = \left\langle w, \mathsf{P}_{\mathcal{W}^\perp,\mathsf{L}\mathcal{V}}\mathsf{L}z \right\rangle = 0$$

holds. Thus $\mathsf{Q} = \mathsf{P}_{\mathcal{V},(\mathsf{L}^\star\mathcal{W})^\perp}$. □

An interesting question is: what is the inverse of a projection if it is restricted to the orthogonal complement of its null space? The answer is trivial for orthogonal projections since the restriction is just the identity operator. In the general case, the answer reveals an interesting link between oblique and orthogonal projections that is used in the analysis of perturbed projections in section 3.3.1. The next lemma provides the inverse and some other fundamental relationships between oblique and orthogonal projections.

Lemma 2.13. *For two closed subspaces $\mathcal{V},\mathcal{W} \subseteq \mathcal{H}$ with $\mathcal{V} \oplus \mathcal{W} = \mathcal{H}$, the following statements hold:*

1. *The operator*

$$\mathsf{Q}_{\mathcal{V},\mathcal{W}} := \begin{cases} \mathcal{V} \to \mathcal{W}^{\perp} \\ v \mapsto \mathsf{P}_{\mathcal{W}^{\perp}} v \end{cases} \quad \textit{with norm} \quad \|\mathsf{Q}_{\mathcal{V},\mathcal{W}}\| = \|\mathsf{P}_{\mathcal{W}^{\perp}}\mathsf{P}_{\mathcal{V}}\|$$

is invertible and its inverse is given by

$$\mathsf{Q}_{\mathcal{V},\mathcal{W}}^{-1} : \begin{cases} \mathcal{W}^{\perp} \to \mathcal{V} \\ w \mapsto \mathsf{P}_{\mathcal{V},\mathcal{W}} w \end{cases} \quad \textit{with norm} \quad \left\|\mathsf{Q}_{\mathcal{V},\mathcal{W}}^{-1}\right\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \,.$$

2. $\mathsf{P}_{\mathcal{V},\mathcal{W}} = \mathsf{Q}_{\mathcal{V},\mathcal{W}}^{-1}\mathsf{P}_{\mathcal{W}^{\perp}}$ *with* $\mathsf{Q}_{\mathcal{V},\mathcal{W}}^{-1}$ *as in 1.*

3. $\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{V}} = \mathsf{P}_{\mathcal{V}}$.

4. $\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}^{\perp}} = \mathsf{P}_{\mathcal{V},\mathcal{W}}$.

Proof. 1. Let $v \in \mathcal{V}$ be arbitrary. Then the statement follows from

$$v - \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{Q}_{\mathcal{V},\mathcal{W}} v = v - \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}^{\perp}} v = \mathsf{P}_{\mathcal{V},\mathcal{W}}(\mathsf{id} - \mathsf{P}_{\mathcal{W}^{\perp}})v = \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}} v = 0.$$

2. It can be seen with statement 1 that

$$\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{Q}_{\mathcal{V},\mathcal{W}}^{-1}\mathsf{P}_{\mathcal{W}^{\perp}} = \mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}^{\perp}} = \mathsf{P}_{\mathcal{V},\mathcal{W}}(\mathsf{id} - \mathsf{P}_{\mathcal{W}^{\perp}}) = \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}} = 0.$$

3. $\mathsf{P}_{\mathcal{V}} - \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{V}} = (\mathsf{id} - \mathsf{P}_{\mathcal{V},\mathcal{W}})\mathsf{P}_{\mathcal{V}} = \mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{V}} = 0.$

4. $\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}^{\perp}} = \mathsf{P}_{\mathcal{V},\mathcal{W}}(\mathsf{id} - \mathsf{P}_{\mathcal{W}^{\perp}}) = \mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}} = 0.$

□

The following results by Buckholtz [20, 19] give necessary and sufficient conditions for two subspaces to be complementary and thus for the existence of projections, cf. lemmas 2.4 and 2.5. These conditions are stated in terms of orthogonal and oblique projections onto the subspaces and their orthogonal complements.

Theorem 2.14. *Let $\mathcal{X},\mathcal{Y} \subseteq \mathcal{H}$ be two closed subspaces. Then the following statements are equivalent:*

1. $\mathcal{H} = \mathcal{X} \oplus \mathcal{Y}$.

2. *The operator* $\mathsf{P}_{\mathcal{X}} - \mathsf{P}_{\mathcal{Y}}$ *is invertible and its inverse is given by* $\mathsf{P}_{\mathcal{X},\mathcal{Y}} + \mathsf{P}_{\mathcal{Y}^{\perp},\mathcal{X}^{\perp}} - \mathsf{id}$.

3. $\|\mathsf{P}_{\mathcal{X}} + \mathsf{P}_{\mathcal{Y}} - \mathsf{id}\| < 1$.

Proof. The proof of the equivalences was given in [19] while the inverse in statement 2 was presented in [20]. □

For bases of finite-dimensional subspaces $\mathcal{V}$ and $\mathcal{W}$, the following lemma provides a way to check if $\mathcal{V}$ and $\mathcal{W}^\perp$ form a direct sum and provides an explicit representation of the corresponding projection.

Theorem 2.15. *Let $V, W \in \mathcal{H}^n$ be such that $\mathcal{V} = [\![V]\!]$ and $\mathcal{W} = [\![W]\!]$ have dimension $n < \infty$. Then the following statements are equivalent:*

1. $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}^\perp$.
2. $\langle W, V\rangle \in \mathbb{C}^{n,n}$ *is nonsingular and the projection* $\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}$ *can be represented by* $\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp} z = V\langle W, V\rangle^{-1}\langle W, z\rangle$ *for* $z \in \mathcal{H}$.
3. $\|\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{W}\| < 1$.

Proof. 1.$\Longrightarrow$2.: Assume that $\langle W, V\rangle$ is singular. Then there exists a nonzero $t \in \mathbb{C}^n$ such that $\langle W, V\rangle t = 0$ which is equivalent to $x := Vt \in \mathcal{W}^\perp$. Thus there is a nonzero $x \in \mathcal{V} \cap \mathcal{W}^\perp$ which is a contradiction to $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}^\perp$. The representation of $\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}$ can be verified by a trivial calculation.

2.$\Longrightarrow$1.: Follows directly from the fact that the projection $\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}$ is well defined.

3.$\Longleftrightarrow$1.: Follows from theorem 2.14. □

In the next section, it is shown that the norm $\|\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{W}\|$ in condition 3 can be characterized as the sine of the maximal canonical angle $\theta_{\max}(\mathcal{V}, \mathcal{W})$ between the subspaces $\mathcal{V}$ and $\mathcal{W}$. Because the maximal canonical angle adds some intuition and also does not depend on the basis of the subspaces, the condition $\theta_{\max}(\mathcal{V}, \mathcal{W}) < \frac{\pi}{2}$ is used extensively in chapter 3 as an equivalent condition to the ones in theorem 2.15.

Remark 2.16 (Matrix representation of projections)**.** If $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{m,n}$ are two matrices such that $\mathbf{W}^\mathsf{H}\mathbf{V}$ is nonsingular, then the projection onto $\mathcal{V} := [\![\mathbf{V}]\!]$ along $\mathcal{W}^\perp := [\![\mathbf{W}]\!]^\perp$ with respect to the Euclidean inner product is well defined and can be represented by

$$\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp} = \mathbf{V}(\mathbf{W}^\mathsf{H}\mathbf{V})^{-1}\mathbf{W}^\mathsf{H}. \tag{2.1}$$

Note that care has to be taken when implementing the application of a projection to a vector. The representations in theorem 2.15 and equation (2.1) may be problematic in the presence of round-off errors and it is referred to section 2.10 for a brief discussion of numerically sound algorithms for the application of projections.

2.3. Angles and gaps between subspaces

In a Hilbert space, the inner product $\langle\cdot,\cdot\rangle$ can be used to measure the angle between two vectors and the concept of angles can be extended to subspaces. It is well-known in the literature that certain properties of projections on Hilbert spaces can be characterized by angles between the involved subspaces, i.e., the ranges and null spaces of the projections. Besides adding a helpful geometric intuition to subspaces, angles play an important role in the analysis of approximate invariant subspaces such as in the theory of Davis and Kahan [26] which is used in section 3.3.2 in order to analyze the spectrum of deflated operators. Furthermore, the basic results of this section pave the way for an apparently new result on the behavior of projections with perturbed ranges and null spaces in section 3.3.1.

The next definition introduces the minimal canonical angle between two closed subspaces of a Hilbert space.

Definition 2.17. For two closed and nonzero subspaces $\mathcal{V},\mathcal{W}\subseteq\mathcal{H}$ the *minimal canonical angle* $\theta_{\min}(\mathcal{V},\mathcal{W})\in[0,\frac{\pi}{2}]$ between $\mathcal{V}$ and $\mathcal{W}$ is defined by

$$\cos\theta_{\min}(\mathcal{V},\mathcal{W})=\sup_{\substack{v\in\mathcal{V},\|v\|=1\\ w\in\mathcal{W},\|w\|=1}}|\langle v,w\rangle|.$$

The well-definedness of the minimal canonical angle directly follows from the Cauchy–Schwarz inequality. The following lemma gathers basic properties of the minimal canonical angle and shows that it can be expressed as the norm of the product of orthogonal projections.

Lemma 2.18. *Let $\mathcal{V},\mathcal{W}\subseteq\mathcal{H}$ be closed and nonzero subspaces. Then:*

1. $\theta_{\min}(\mathcal{V},\mathcal{W})=\theta_{\min}(\mathcal{W},\mathcal{V})$.
2. $|\langle v,w\rangle|\le\|v\|\,\|w\|\cos\theta_{\min}(\mathcal{V},\mathcal{W})$ *for all* $v\in\mathcal{V}$ *and* $w\in\mathcal{W}$ *(Cauchy–Schwarz inequality for subspaces).*
3. $\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\|=\|\mathsf{P}_{\mathcal{W}}\mathsf{P}_{\mathcal{V}}\|=\cos\theta_{\min}(\mathcal{V},\mathcal{W})$.
4. $\|\mathsf{L}\mathsf{P}_{\mathcal{V}+\mathcal{W}}\|\le\|\mathsf{L}\mathsf{P}_{\mathcal{V}}\|+\|\mathsf{L}\mathsf{P}_{\mathcal{W}}\|$ *for* $\mathsf{L}\in\mathcal{L}(\mathcal{H})$.
5. *If* $\mathcal{V}\neq\mathcal{H}$ *and* $\dim\mathcal{W}=1$ *then* $\cos^2\theta_{\min}(\mathcal{V},\mathcal{W})+\cos^2\theta_{\min}(\mathcal{V}^\perp,\mathcal{W})=1$.

Proof. 1. Is a direct consequence of the inner product's symmetry.

2. If $v=0$ or $w=0$ the statement is clear. For $v\neq0$ and $w\neq0$ the following holds:

$$\begin{aligned}|\langle v,w\rangle|&=\|v\|\,\|w\|\left|\left\langle\frac{v}{\|v\|},\frac{w}{\|w\|}\right\rangle\right|\le\|v\|\,\|w\|\sup_{\substack{x\in\mathcal{V},\|x\|=1\\ y\in\mathcal{W},\|y\|=1}}|\langle x,y\rangle|\\&=\|v\|\,\|w\|\cos\theta_{\min}(\mathcal{V},\mathcal{W}).\end{aligned}$$

3. Cf. Szyld [171, Lemma 5.1].

4. $$\begin{aligned}\|\mathsf{L}\mathsf{P}_{\mathcal{V}+\mathcal{W}}\|&=\|\mathsf{L}|_{\mathcal{V}+\mathcal{W}}\|=\sup_{\substack{v\in\mathcal{V},w\in\mathcal{W}\\ \|v+w\|=1}}\|\mathsf{L}(v+w)\|\le\sup_{\substack{v\in\mathcal{V},w\in\mathcal{W}\\ \|v+w\|=1}}(\|\mathsf{L}v\|+\|\mathsf{L}w\|)\\&\le\sup_{\substack{v\in\mathcal{V},w\in\mathcal{W}\\ \|v+w\|=1}}\|\mathsf{L}v\|+\sup_{\substack{v\in\mathcal{V},w\in\mathcal{W}\\ \|v+w\|=1}}\|\mathsf{L}w\|=\sup_{\substack{v\in\mathcal{V}\\ \|v\|=1}}\|\mathsf{L}v\|+\sup_{\substack{w\in\mathcal{W}\\ \|w\|=1}}\|\mathsf{L}w\|\\&=\|\mathsf{L}|_{\mathcal{V}}\|+\|\mathsf{L}|_{\mathcal{W}}\|=\|\mathsf{L}\mathsf{P}_{\mathcal{V}}\|+\|\mathsf{L}\mathsf{P}_{\mathcal{W}}\|.\end{aligned}$$

5. Let $w \in \mathcal{H}$ such that $\|w\| = 1$ and $\mathcal{W} = [\![w]\!]$. Then

$$\begin{aligned} 1 = \|w\|^2 &= \|(\mathsf{P}_\mathcal{V} + \mathsf{P}_{\mathcal{V}^\perp})w\|^2 = \|\mathsf{P}_\mathcal{V} w\|^2 + \|\mathsf{P}_{\mathcal{V}^\perp} w\|^2 = \|\mathsf{P}_\mathcal{V}\mathsf{P}_\mathcal{W}\|^2 + \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_\mathcal{W}\|^2 \\ &= \cos^2\theta_{\min}(\mathcal{V},\mathcal{W}) + \cos^2\theta_{\min}(\mathcal{V}^\perp,\mathcal{W}). \end{aligned}$$

□

The following definition introduces the widely used concept of the gap between two closed subspaces of a Hilbert space, see also [90, 59, 86, 99].

Definition 2.19. For two closed subspaces $\mathcal{V},\mathcal{W} \subseteq \mathcal{H}$ the *gap* (also *maximal gap*, *aperture* or *opening*) $\Theta(\mathcal{V},\mathcal{W})$ between $\mathcal{V}$ and $\mathcal{W}$ is defined by

$$\Theta(\mathcal{V},\mathcal{W}) = \|\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{W}\|.$$

In the next lemma, basic properties of the gap between two subspaces are gathered.

Lemma 2.20. *Let $\mathcal{V},\mathcal{W},\mathcal{Z} \subseteq \mathcal{H}$ be closed subspaces. Then:*

1. $\Theta(\mathcal{V},\mathcal{W}) = \Theta(\mathcal{W},\mathcal{V})$.
2. $\Theta(\mathcal{V},\mathcal{W}) \in [0,1]$.
3. $\Theta(\mathcal{V},\mathcal{W}) \le \Theta(\mathcal{V},\mathcal{Z}) + \Theta(\mathcal{Z},\mathcal{W})$.
4. $\Theta(\mathcal{V},\mathcal{W}) = \max\{\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_\mathcal{W}\|, \|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_\mathcal{V}\|\}$.
5. $\Theta(\mathcal{V},\mathcal{W}) = \Theta(\mathcal{V}^\perp,\mathcal{W}^\perp)$.

Proof. 1. Follows directly from the definition.

2. $\Theta(\mathcal{V},\mathcal{W}) \ge 0$ also follows from the definition and for $z \in \mathcal{H}$

$$\begin{aligned} \|(\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{W})z\|^2 &= \|\mathsf{P}_\mathcal{V}(\mathsf{id} - \mathsf{P}_\mathcal{W})z - (\mathsf{id} - \mathsf{P}_\mathcal{V})\mathsf{P}_\mathcal{W} z\|^2 \\ &= \|\mathsf{P}_\mathcal{V}\mathsf{P}_{\mathcal{W}^\perp} z\|^2 + \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_\mathcal{W} z\|^2 \qquad (2.2) \\ &\le \|\mathsf{P}_\mathcal{V}\|^2 \|\mathsf{P}_{\mathcal{W}^\perp} z\|^2 + \|\mathsf{P}_{\mathcal{V}^\perp}\|^2 \|\mathsf{P}_\mathcal{W} z\|^2 \\ &\le \|\mathsf{P}_{\mathcal{W}^\perp} z\|^2 + \|\mathsf{P}_\mathcal{W} z\|^2 = \|z\|^2 \end{aligned}$$

and thus $\Theta(\mathcal{V},\mathcal{W}) = \sup_{\|z\|=1} \|(\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{W})z\| \le 1$.

3. The norm triangle inequality yields:

$$\begin{aligned} \Theta(\mathcal{V},\mathcal{W}) = \|\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{W}\| &= \|\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{Z} + \mathsf{P}_\mathcal{Z} - \mathsf{P}_\mathcal{W}\| \le \|\mathsf{P}_\mathcal{V} - \mathsf{P}_\mathcal{Z}\| + \|\mathsf{P}_\mathcal{Z} - \mathsf{P}_\mathcal{W}\| \\ &= \Theta(\mathcal{V},\mathcal{Z}) + \Theta(\mathcal{Z},\mathcal{W}). \end{aligned}$$

4. The proof can be found in [53, theorem 1.2] and is given here for the sake of completeness. For $z \in \mathcal{H}$ it can be concluded from equation (2.2) that

$$\max\{\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}^\perp}z\|, \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|\} \leq \|(\mathsf{P}_{\mathcal{V}} - \mathsf{P}_{\mathcal{W}})z\|$$

and

$$\begin{aligned}\|(\mathsf{P}_{\mathcal{V}} - \mathsf{P}_{\mathcal{W}})z\|^2 &\leq \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}^\perp}\|^2 \|\mathsf{P}_{\mathcal{W}^\perp}z\|^2 + \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}\|^2 \|\mathsf{P}_{\mathcal{W}}z\|^2 \\ &\leq \max\{\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}^\perp}\|^2, \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}\|^2\}(\|\mathsf{P}_{\mathcal{W}^\perp}z\|^2 + \|\mathsf{P}_{\mathcal{W}}z\|^2) \\ &= \max\{\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}^\perp}\|^2, \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}\|^2\} \|z\|^2.\end{aligned}$$

However, $\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}^\perp}\| = \|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{V}}\|$ holds according to statement 3 of lemma 2.18 and the above inequalities yield combined:

$$\max\{\|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{V}}z\|, \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|\} \leq \max\{\|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{V}}\|, \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}\|\} \|z\|.$$

Employing the supremum over all $z \in \mathcal{H}$ completes the proof.

5. $\Theta(\mathcal{V}, \mathcal{W}) = \|\mathsf{P}_{\mathcal{V}} - \mathsf{P}_{\mathcal{W}}\| = \|\mathsf{P}_{\mathcal{V}^\perp} - \mathsf{P}_{\mathcal{W}^\perp}\| = \Theta(\mathcal{V}^\perp, \mathcal{W}^\perp)$.

□

Statement 2 of lemma 2.20 allows the gap between two subspaces to be interpreted as the maximal canonical angle between two subspaces:

Definition 2.21. For two closed subspaces $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ the *maximal canonical angle* $\theta_{\max}(\mathcal{V}, \mathcal{W}) \in [0, \frac{\pi}{2}]$ between $\mathcal{V}$ and $\mathcal{W}$ is defined by

$$\sin\theta_{\max}(\mathcal{V}, \mathcal{W}) = \Theta(\mathcal{V}, \mathcal{W}) = \|\mathsf{P}_{\mathcal{V}} - \mathsf{P}_{\mathcal{W}}\|.$$

The following lemma shows that the notation is not misleading because the minimal canonical angle is actually less or equal to the maximal canonical angle.

Lemma 2.22. *Let $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ be closed subspaces. Then*

$$\theta_{\min}(\mathcal{V}, \mathcal{W}) \leq \theta_{\max}(\mathcal{V}, \mathcal{W}).$$

Proof. Because the angles satisfy $\theta_{\min}(\mathcal{V}, \mathcal{W}), \theta_{\max}(\mathcal{V}, \mathcal{W}) \in [0, \frac{\pi}{2}]$, the proof is complete if $\sin\theta_{\min}(\mathcal{V}, \mathcal{W}) \leq \sin\theta_{\max}(\mathcal{V}, \mathcal{W})$. From statement 3 of lemma 2.18 it is known that

$$\sin^2\theta_{\min}(\mathcal{V}, \mathcal{W}) = 1 - \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\|^2 = 1 - \|\mathsf{P}_{\mathcal{W}}\mathsf{P}_{\mathcal{V}}\|^2.$$

For $z \in \mathcal{W}$ with $\|z\| = 1$ the following equality holds:

$$1 = \|z\|^2 = \|\mathsf{P}_{\mathcal{W}}z\|^2 = \|(\mathsf{P}_{\mathcal{V}} + \mathsf{P}_{\mathcal{V}^\perp})\mathsf{P}_{\mathcal{W}}z\|^2 = \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}z\|^2 + \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|^2$$

and thus

$$\sin^2\theta_{\min}(\mathcal{V}, \mathcal{W}) = 1 - \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\|^2 = 1 - \sup_{\substack{z\in\mathcal{W}\\ \|z\|=1}} \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}z\|^2 = \inf_{\substack{z\in\mathcal{W}\\ \|z\|=1}} \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|^2$$

$$\leq \sup_{\substack{z\in\mathcal{W}\\ \|z\|=1}} \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|^2 = \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}\|^2 .$$

Analogously, $\sin\theta_{\min}(\mathcal{V},\mathcal{W}) \leq \|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{V}}\|$ and the following statement is obtained:

$$\sin\theta_{\min}(\mathcal{V},\mathcal{W}) \leq \max\{\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}\|, \|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{V}}\|\} = \Theta(\mathcal{V},\mathcal{W}) = \sin\theta_{\max}(\mathcal{V},\mathcal{W}).$$

□

So far angles between arbitrary closed subspaces have been discussed. In the case of complementary subspaces, i.e., $\mathcal{V}\oplus\mathcal{W}=\mathcal{H}$, certain angles can be expressed by norms of projections and relations between the minimal and maximal canonical angles can be established.

Lemma 2.23. *For two closed and nonzero subspaces $\mathcal{V},\mathcal{W}\subseteq\mathcal{H}$ with $\mathcal{V}\oplus\mathcal{W}=\mathcal{H}$ the following holds:*

1. $\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \frac{1}{\sqrt{1-\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\|^2}}$.
2. $\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\| = \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}^\perp}\| = \Theta(\mathcal{V},\mathcal{W}^\perp)$.
3. $\theta_{\min}(\mathcal{V},\mathcal{W}) = \theta_{\min}(\mathcal{V}^\perp,\mathcal{W}^\perp)$.
4. $\theta_{\min}(\mathcal{V},\mathcal{W}) + \theta_{\max}(\mathcal{V},\mathcal{W}^\perp) = \frac{\pi}{2}$.
5. $\cos\theta_{\min}(\mathcal{V},\mathcal{W}) = \sin\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)$ *and* $\sin\theta_{\min}(\mathcal{V},\mathcal{W}) = \cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)$.
6. $\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \dfrac{1}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)}$.
7. *If* $\mathsf{L}\in\mathcal{L}(\mathcal{H})$ *then*
$$\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\,\|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \leq \|\mathsf{L}\mathsf{P}_{\mathcal{V}}\| \leq \cos\theta_{\min}(\mathcal{V},\mathcal{W}^\perp)\,\|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\| .$$

Proof. 1. Similar to the proof of lemma 2.22, the fact is used that for $z\in\mathcal{W}$ with $\|z\|=1$ the equality $1=\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}z\|^2+\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|^2$ holds. With statement 1 of lemma 2.13 (note that also $\mathcal{V}^\perp\oplus\mathcal{W}^\perp=\mathcal{H}$, cf. statement 1 of lemma 2.7), the following equation is obtained:

$$1-\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\|^2 = 1-\sup_{\substack{z\in\mathcal{W}\\ \|z\|=1}}\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}z\|^2 = \inf_{\substack{z\in\mathcal{W}\\ \|z\|=1}}\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|^2 = \frac{1}{\sup_{\substack{z\in\mathcal{W}\\ z\neq0}}\frac{\|z\|^2}{\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}}z\|^2}}$$
$$= \frac{1}{\sup_{\substack{z\in\mathcal{W}\\ z\neq0}}\frac{\|z\|^2}{\|\mathsf{Q}_{\mathcal{W},\mathcal{V}}z\|^2}} = \frac{1}{\sup_{\substack{z\in\mathcal{V}^\perp\\ z\neq0}}\frac{\|\mathsf{Q}^{-1}_{\mathcal{W},\mathcal{V}}z\|^2}{\|z\|^2}} = \frac{1}{\left\|\mathsf{Q}^{-1}_{\mathcal{W},\mathcal{V}}\right\|^2} = \frac{1}{\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\|^2}.$$

The proof is complete by recognizing that $\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|$, cf. lemma 2.7.

2. By exchanging $\mathcal{V}$ with $\mathcal{V}^\perp$ and $\mathcal{W}$ with $\mathcal{W}^\perp$ in statement 1 the equation
$$\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}^\perp}\|^2 = 1 - \frac{1}{\|\mathsf{P}_{\mathcal{V}^\perp,\mathcal{W}^\perp}\|^2}$$
follows. However, $\mathsf{P}_{\mathcal{V}^\perp,\mathcal{W}^\perp}$ is the complementary adjoint projection of $\mathsf{P}_{\mathcal{V},\mathcal{W}}$ with $\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \|\mathsf{P}_{\mathcal{V}^\perp,\mathcal{W}^\perp}\|$ (cf. lemma 2.7) and thus $\|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\| = \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}^\perp}\|$. The last equality follows from lemma 2.20.

3. Follows directly from the previous statement and statement 3 of lemma 2.18.

4. This statement has been shown by Ljance [107]. Here, a different proof is given. Due to the fact that $\theta_{\min}(\mathcal{V},\mathcal{W}), \theta_{\max}(\mathcal{V},\mathcal{W}) \in [0, \frac{\pi}{2}]$, it suffices to show that
$$\sin\theta_{\max}(\mathcal{V},\mathcal{W}^\perp) = \sin\left(\frac{\pi}{2} - \theta_{\min}(\mathcal{V},\mathcal{W})\right) = \cos\theta_{\min}(\mathcal{V},\mathcal{W}).$$
This can now easily be shown with statement 4 of lemma 2.20, statement 2 of this lemma and statement 3 of lemma 2.18:
$$\sin\theta_{\max}(\mathcal{V},\mathcal{W}^\perp) = \max\{\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{W}^\perp}\|, \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\|\} = \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\| = \cos\theta_{\min}(\mathcal{V},\mathcal{W}).$$

5. Follows from the previous result.

6. From statements 1 and 2 of this lemma the equation can be shown by
$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \frac{1}{\sqrt{1 - \|\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{W}}\|^2}} = \frac{1}{\sqrt{1-\sin^2\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)}} = \frac{1}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)}.$$

7. The first inequality follows from statement 6 and
$$\|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \|\mathsf{L}\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \le \|\mathsf{L}\mathsf{P}_{\mathcal{V}}\|\,\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \|\mathsf{L}\mathsf{P}_{\mathcal{V}}\|\frac{1}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)}.$$
The second inequality can be verified with statements 3 and 4:
$$\begin{aligned}\|\mathsf{L}\mathsf{P}_{\mathcal{V}}\| &= \|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{V}}\| = \|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{V}}\| \\ &\le \|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\|\,\|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{V}}\| = \cos\theta_{\min}(\mathcal{V},\mathcal{W}^\perp)\,\|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\|.\end{aligned}$$

□

In addition to the well-known results in this and the preceding section, an apparently new result on the norm of the difference of two arbitrary projections is presented in section 3.3.1.

2.4. Galerkin and Petrov–Galerkin methods

The characterization and analysis of Krylov subspace methods in this thesis rely to a great extent on the projection framework for solving a linear system which was introduced by Saad [142]. This projection framework is in fact a Petrov–Galerkin method which is defined in this section. The notation is mostly adopted from the book of Liesen and Strakoš [105]. As before, instead of the Euclidean inner product, a Hilbert space with arbitrary inner product is used. This setting has also been discussed in the extensive article by Eiermann and Ernst [42]. However, in contrast to [105, 42], most statements made in this and the following sections do not require that the linear operator is nonsingular. Therefore, proofs are given for theorems that are well-known for the nonsingular case in the literature. The issue of singular operators reappears in the context of deflated Krylov subspace methods in chapter 3.

For this and the following sections, let a linear system

$$\mathsf{A}x = b \tag{2.3}$$

be given with $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $b \in \mathcal{H}$. It is assumed that the linear system (2.3) is consistent, i.e., $b \in \mathcal{R}(\mathsf{A})$. Then a solution x of (2.3) can be approximated by

$$y = x_0 + s \quad \text{with} \quad s \in \mathcal{S}, \tag{2.4}$$

where $x_0 \in \mathcal{H}$ is a given initial approximation and $\mathcal{S} \subseteq \mathcal{H}$ is an n-dimensional subspace, also referred to as the *search space*. Now one has n degrees of freedom in the choice of s in (2.4) and thus a constraint is imposed on the residual

$$r := b - \mathsf{A}y \perp \mathcal{C}, \tag{2.5}$$

where $\mathcal{C} \subseteq \mathcal{H}$ also is an n-dimensional subspace, the *constraint space*.

Definition 2.24 ((Petrov–)Galerkin method)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $x_0, b \in \mathcal{H}$. For two n-dimensional subspaces $\mathcal{S}, \mathcal{C} \subseteq \mathcal{H}$ the method described by (2.4) and (2.5) is called a *Galerkin method* if $\mathcal{S} = \mathcal{C}$ and a *Petrov–Galerkin method* otherwise. A (Petrov–) Galerkin method is called *well defined* if there exists a unique approximate solution y satisfying (2.4) and (2.5).

The question of well-definedness of (Petrov–)Galerkin methods is answered in the following theorem:

Theorem 2.25. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $x_0, b \in \mathcal{H}$ and two n-dimensional subspaces $\mathcal{S}, \mathcal{C} \subseteq \mathcal{H}$ be given. Then the following statements are equivalent:*

1. *The (Petrov–)Galerkin method with search space $\mathcal{S}$ and constraint space $\mathcal{C}$ is well defined.*

2. *$\langle C, \mathsf{A}S \rangle$ is non-singular for any $S, C \in \mathcal{H}^n$ with $[\![S]\!] = \mathcal{S}$ and $[\![C]\!] = \mathcal{C}$.*

3. $\mathsf{A}\mathcal{S} \oplus \mathcal{C}^\perp = \mathcal{H}$.

Proof. 1.$\Longleftrightarrow$2.: Let $s = St \in \mathcal{S}$ fulfill (2.4) and (2.5) and let $S, C \in \mathcal{H}^n$ with $[\![S]\!] = \mathcal{S}$ and $[\![C]\!] = \mathcal{C}$. Condition (2.5) is equivalent to

$$\langle C, \mathsf{A}S\rangle t = \langle C, b - \mathsf{A}x_0\rangle.$$

Thus the uniqueness of s and y is equivalent to the nonsingularity of $\langle C, \mathsf{A}S\rangle$.

2.$\Longleftrightarrow$3.: Has been proven in theorem 2.15. □

Corollary 2.26. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $x_0, b \in \mathcal{H}$ and two n-dimensional subspaces $\mathcal{S}, \mathcal{C} \subseteq \mathcal{H}$ be given such that the Petrov–Galerkin method with search space $\mathcal{S}$ and constraint space $\mathcal{C}$ is well defined.*

Then the approximate solution y and the corresponding residual r that satisfy (2.4) *and* (2.5) *are given by*

$$y = x_0 + S\langle C, \mathsf{A}S\rangle^{-1}\langle C, r_0\rangle \tag{2.6}$$

$$\text{and} \quad r = b - \mathsf{A}y = \mathsf{P}_{\mathcal{C}^\perp, \mathsf{A}\mathcal{S}} r_0, \tag{2.7}$$

where $r_0 = b - \mathsf{A}x_0$ is the initial residual. Furthermore, the linear system (2.3) *is solved by y if and only if $r_0 \in \mathsf{A}\mathcal{S}$.*

Proof. Let $S, C \in \mathcal{H}^n$ with $[\![S]\!] = \mathcal{S}$ and $[\![C]\!] = \mathcal{C}$. From the proof of theorem 2.25 it follows that $y = x_0 + S\langle C, \mathsf{A}S\rangle^{-1}\langle C, r_0\rangle$ and thus

$$r = b - \mathsf{A}y = r_0 - \mathsf{A}S\langle C, \mathsf{A}S\rangle^{-1}\langle C, r_0\rangle = (\mathsf{id} - \mathsf{P}_{\mathsf{A}\mathcal{S}, \mathcal{C}^\perp})r_0 = \mathsf{P}_{\mathcal{C}^\perp, \mathsf{A}\mathcal{S}} r_0.$$

□

From corollary 2.26, the link between Petrov–Galerkin methods and projections becomes apparent. Equation (2.7) shows that the residual is $r = \mathsf{P}_{\mathcal{C}^\perp, \mathsf{A}\mathcal{S}} r_0$ and because of the definition of orthogonal and oblique projections (cf. definition 2.9) a projection method is called *orthogonal* if $\mathsf{A}\mathcal{S} = \mathcal{C}$ and *oblique* otherwise.

Remark 2.27. By using the above definition for orthogonal and oblique projection methods, the author follows Liesen and Strakoš [105] in breaking with the tradition of using the term *orthogonal* for the case $\mathcal{S} = \mathcal{C}$ and the term *oblique* otherwise.

Given a search space $\mathcal{S}$, there are two popular choices for the constraint space $\mathcal{C}$: $\mathcal{C} = \mathcal{S}$ and $\mathcal{C} = \mathsf{A}\mathcal{S}$. The first choice is used in the CG method if A is self-adjoint and positive semidefinite, see section 2.8, and results in a minimal A-norm of the error. The second choice is used in the GMRES and MINRES methods, see section 2.9, and results in a minimal residual norm. Conditions for the well-definedness of a Petrov–Galerkin method with these spaces are given in the following lemma.

Lemma 2.28 (Well-definedness and optimality)**.** *Consider a linear system $\mathsf{A}x = b$ with $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $x \in \mathcal{H}$ and $b \in \mathcal{R}(\mathsf{A})$. Furthermore, let $x_0 \in \mathcal{H}$ be an initial guess and let $\mathcal{S} \subseteq \mathcal{H}$ be an n-dimensional subspace. The Petrov–Galerkin method with search space $\mathcal{S}$ and constraint space $\mathcal{C}$ is well defined and defines a unique approximate solution $y \in x_0 + \mathcal{S}$ if one of the following conditions holds:*

1. *$\mathcal{C} = \mathcal{S}$, $\mathcal{S} \cap \mathcal{N}(\mathsf{A}) \neq \{0\}$ and A is self-adjoint and positive semidefinite. Then*

$$\|x - y\|_{\mathsf{A}} = \inf_{z \in x_0 + \mathcal{S}} \|x - z\|_{\mathsf{A}},$$

where $\|\cdot\|_{\mathsf{A}}$ is the norm (or semi-norm if A is singular) defined by $\|z\|_{\mathsf{A}} = \sqrt{\langle z, \mathsf{A}z\rangle}$.

2. *$\mathcal{C} = \mathsf{A}\mathcal{S}$ and $\mathcal{S} \cap \mathcal{N}(\mathsf{A}) = \{0\}$. Then*

$$\|b - \mathsf{A}y\| = \inf_{z \in x_0 + \mathcal{S}} \|b - \mathsf{A}z\|.$$

Proof. 1. Let $S \in \mathcal{H}^n$ with $[\![S]\!] = \mathcal{S}$ and assume that $\langle S, \mathsf{A}S\rangle$ is singular, i.e., there exists a nonzero $t \in \mathbb{C}^n$ with $\langle S, \mathsf{A}St\rangle = 0$. Because A is self-adjoint and positive semidefinite, it has a unique self-adjoint and positive-semidefinite square root $\mathsf{A}^{\frac{1}{2}}$, cf. [182, Korollar VII.1.16]. Then with $s := St \neq 0$ the following holds

$$0 = \langle s, \mathsf{A}s\rangle = \left\langle \mathsf{A}^{\frac{1}{2}}s, \mathsf{A}^{\frac{1}{2}}s\right\rangle = \left\|\mathsf{A}^{\frac{1}{2}}s\right\|^2$$

and hence $\mathsf{A}^{\frac{1}{2}}s = 0$. Applying $\mathsf{A}^{\frac{1}{2}}$ yields $\mathsf{A}s = 0$ which is a contradiction to $\mathcal{S} \cap \mathcal{N}(\mathsf{A}) = \{0\}$. The well-definedness follows from the equivalence of items 1. and 2. in theorem 2.25.

It remains to show the optimality. Therefore, let $\mathcal{R} := \mathcal{N}(\mathsf{A})^\perp$ and let the inner product $\langle \cdot, \cdot\rangle_{\mathsf{A}|_{\mathcal{R}}} : \mathcal{R} \times \mathcal{R} \longrightarrow \mathbb{C}$ be defined by $\langle x, y\rangle_{\mathsf{A}|_{\mathcal{R}}} = \langle x, \mathsf{A}|_{\mathcal{R}}y\rangle$. Note that the restriction $\mathsf{A}|_{\mathcal{R}}$ is self-adjoint and positive definite. Then the following holds with the representation of y in (2.6):

$$\begin{aligned}
\|x - y\|_{\mathsf{A}} &= \|\mathsf{P}_{\mathcal{R}}(x - y)\|_{\mathsf{A}|_{\mathcal{R}}} \\
&= \left\|\mathsf{P}_{\mathcal{R}}(x - x_0) - \mathsf{P}_{\mathcal{R}}S\langle \mathsf{P}_{\mathcal{R}}S, \mathsf{P}_{\mathcal{R}}S\rangle_{\mathsf{A}|_{\mathcal{R}}}^{-1}\langle \mathsf{P}_{\mathcal{R}}S, \mathsf{P}_{\mathcal{R}}(x - x_0)\rangle_{\mathsf{A}|_{\mathcal{R}}}\right\|_{\mathsf{A}|_{\mathcal{R}}} \\
&= \left\|\mathsf{P}_{(\mathsf{P}_{\mathcal{R}}\mathcal{S})^{\perp_{\mathcal{A}|_{\mathcal{R}}}}}\mathsf{P}_{\mathcal{R}}(x - x_0)\right\|_{\mathsf{A}|_{\mathcal{R}}}.
\end{aligned}$$

Because $\mathsf{P}_{(\mathsf{P}_{\mathcal{R}}\mathcal{S})^{\perp_{\mathcal{A}|_{\mathcal{R}}}}}$ is the orthogonal projection onto $\mathsf{P}_{\mathcal{R}}\mathcal{S}$ with respect to the inner product $\langle \cdot, \cdot\rangle_{\mathsf{A}|_{\mathcal{R}}}$, theorem 2.10 shows that

$$\begin{aligned}
\|x - y\|_{\mathsf{A}} &= \inf_{z \in \mathsf{P}_{\mathcal{R}}\mathcal{S}} \|\mathsf{P}_{\mathcal{R}}(x - x_0) - z\|_{\mathsf{A}|_{\mathcal{R}}} = \inf_{z \in \mathcal{S}} \|\mathsf{P}_{\mathcal{R}}(x - (x_0 + z))\|_{\mathsf{A}|_{\mathcal{R}}} \\
&= \inf_{z \in \mathcal{S}} \|x - (x_0 + z)\|_{\mathsf{A}}.
\end{aligned}$$

2. The well-definedness follows from the equivalence of items 1. and 3. in theorem 2.25. The optimality follows again from theorem 2.10.

□

2.5. Ritz pairs and harmonic Ritz pairs

Analogous to the Petrov–Galerkin approximation for linear systems in the last section, a projection framework can be used to determine approximations to eigenvalues and eigenvectors. There is a close relationship between these so-called Ritz values and vectors and the convergence of Krylov subspace methods for linear systems. This relationship becomes apparent in subsequent sections. The presentation in this section is based on the contribution of Saad in chapter 3.2 in [8].

If an operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and two n-dimensional subspaces $\mathcal{S},\mathcal{C} \subseteq \mathcal{H}$ are given, then a basis of approximate eigenvectors

$$w_1, \dots, w_n \in \mathcal{S} \tag{2.8}$$

and approximate eigenvalues

$$\mu_1, \dots, \mu_n \in \mathbb{C}$$

can be determined by requiring

$$\mathsf{A}w_i - \mu_i w_i \perp \mathcal{C} \quad \text{for all } i \in \{1, \dots, n\}. \tag{2.9}$$

Note the analogy of the conditions (2.8)–(2.9) to the Petrov–Galerkin approach for linear systems, cf. conditions (2.4)–(2.5). If two bases of the subspaces are given, i.e., $S, C \in \mathcal{H}^n$ with $[\![S]\!] = \mathcal{S}$ and $[\![C]\!] = \mathcal{C}$, then the approximate eigenvectors can be represented by $W = [w_1, \dots, w_n] = S\mathbf{U}$ for a $\mathbf{U} = [u_1, \dots, u_n] \in \mathbb{C}^{n,n}$ and the conditions (2.8)–(2.9) are equivalent to finding an invertible $\mathbf{U} \in \mathbb{C}^{n,n}$ and $\mathbf{D}_\mu = \operatorname{diag}(\mu_1, \dots, \mu_n) \in \mathbb{C}^{n,n}$ that satisfy

$$\langle C, \mathsf{A}S \rangle \mathbf{U} = \langle C, S \rangle \mathbf{U}\mathbf{D}_\mu. \tag{2.10}$$

This essentially means that the generalized eigenvalue problem

$$\langle C, \mathsf{A}S \rangle u = \mu \langle C, S \rangle u$$

has to be solved.

As in the previous section, two cases are of particular importance: $\mathcal{S} = \mathcal{C}$ and $\mathsf{A}\mathcal{S} = \mathcal{C}$. In the case $\mathcal{S} = \mathcal{C}$, the above procedure is known as the *Rayleigh-Ritz* procedure and the eigenvector and eigenvalue approximations are called *Ritz vectors* and *Ritz values.*

Definition 2.29 (Ritz pair)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and let $\mathcal{S} \subseteq \mathcal{H}$ be an n-dimensional subspace. The vector $w \in \mathcal{S} \setminus \{0\}$ is called a *Ritz vector* and $\mu \in \mathbb{C}$ is called a *Ritz value* of A with respect to $\mathcal{S}$ if

$$\mathsf{A}w - \mu w \perp \mathcal{S}.$$

The pair (w, μ) is called a *Ritz pair.*

Remark 2.30. In practice, an orthonormal basis is used for the computation of Ritz pairs. If $S \in \mathcal{H}^n$ with $\langle S, S \rangle = \mathbf{I}_n$, then the standard eigenvalue problem

$$\langle S, \mathsf{A} S \rangle \mathbf{U} = \mathbf{U} \mathbf{D}_\mu$$

has to be solved, cf. equation (2.10).

The following lemma is well-known and characterizes the residual of Ritz pairs.

Lemma 2.31. *Let* A *and* $\mathcal{S}$ *be as in definition 2.29 and let* (w, μ) *be a Ritz pair of* A *with respect to* $\mathcal{S}$ *with* $\|w\| = 1$.

Then

$$\|\mathsf{A}w - \mu w\| = \sqrt{\|\mathsf{A}w\|^2 - |\mu|^2}.$$

Proof. First note that $\langle w, \mathsf{A}w \rangle = \mu$. The statement then follows from $\mathsf{A}w - \mu w \perp \mathcal{S}$ and $w \in \mathcal{S}$ because

$$\|\mathsf{A}w - \mu w\|^2 = \langle \mathsf{A}w, \mathsf{A}w - \mu w \rangle = \|\mathsf{A}w\|^2 - \mu \langle \mathsf{A}w, w \rangle = \|\mathsf{A}w\|^2 - |\mu|^2.$$

□

In the case $\mathsf{A}\mathcal{S} = \mathcal{C}$, the generalized eigenvalue problem (2.10) becomes

$$\langle \mathsf{A}S, \mathsf{A}S \rangle \mathbf{U} = \langle \mathsf{A}S, S \rangle \mathbf{U} \mathbf{D}_\mu, \tag{2.11}$$

which deserves some more attention. If $\dim \mathsf{A}\mathcal{S} = \dim \mathcal{S} = n$, then the matrix $\langle \mathsf{A}S, \mathsf{A}S \rangle$ is Hermitian and positive definite, but $\langle \mathsf{A}S, S \rangle$ may be singular and thus infinite eigenvalues may occur. Instead of (2.11), the generalized eigenvalue problem

$$\langle \mathsf{A}S, S \rangle \mathbf{U} = \langle \mathsf{A}S, \mathsf{A}S \rangle \mathbf{U} \mathbf{D}_\sigma \tag{2.12}$$

can be solved, where $\mathbf{D}_\sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. This generalized eigenvalue problem is equivalent to the standard eigenvalue problem

$$\langle \mathsf{A}S, \mathsf{A}S \rangle^{-1} \langle \mathsf{A}S, S \rangle \mathbf{U} = \mathbf{U} \mathbf{D}_\sigma$$

and thus all $\sigma_1, \ldots, \sigma_n$ are finite. Trivially, $S = \mathsf{A}|_{\mathsf{A}\mathcal{S}}^{-1} \mathsf{A}S$ holds and equation (2.12) is equivalent to

$$\mathsf{A}|_{\mathsf{A}\mathcal{S}}^{-1} \mathsf{A}S u_i - \sigma_i \mathsf{A}S u_i \perp \mathsf{A}\mathcal{S}$$

for all $i \in \{1, \ldots, n\}$. This motivates the following definition:

Definition 2.32 (Harmonic Ritz pair)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and let $\mathcal{S} \subseteq \mathcal{H}$ be an n-dimensional subspace such that $\dim \mathsf{A}\mathcal{S} = n$. Let $(z, \sigma) \in \mathsf{A}\mathcal{S} \times \mathbb{C}$ with $z = \mathsf{A}w \neq 0$ be a Ritz pair of $\mathsf{A}|_{\mathsf{A}\mathcal{S}}^{-1}$ with respect to $\mathsf{A}\mathcal{S}$, i.e.

$$\mathsf{A}|_{\mathsf{A}\mathcal{S}}^{-1} z - \sigma z \perp \mathsf{A}\mathcal{S}.$$

Then w is called a *harmonic Ritz vector* and

$$\mu := \begin{cases} \frac{1}{\sigma} & \text{if } \sigma \neq 0 \\ \infty & \text{else} \end{cases}$$

is called a *harmonic Ritz value* of A with respect to $\mathcal{S}$. The pair (w,μ) is called a *harmonic Ritz pair.*

Definition 2.32 allows for infinite harmonic Ritz values which occur if and only if the matrix $\langle \mathsf{A}S, S\rangle$ is singular. This situation may arise in practical applications, e.g., in the GMRES method the occurrence of an infinite harmonic Ritz value is equivalent to exact stagnation of GMRES, cf. section 2.9.

Note that conditions (2.8)–(2.9) with $\mathsf{A}\mathcal{S} = \mathcal{C}$ may be ambiguous in the case of an infinite harmonic Ritz value but definition 2.32 is not.

With respect to the numerical computation of harmonic Ritz values, the generalized eigenvalue problem (2.12) can be used. Zero eigenvalues indicate an infinite harmonic Ritz value and can be treated specially in order to avoid a division by zero.

Harmonic Ritz pairs were introduced in 1991 by Morgan [119] for symmetric matrices under the name *modified Rayleigh-Ritz procedure.* In the literature, harmonic Ritz values have also been called *roots of kernel polynomials* by Freund [57] or *pseudo-Ritz values* by Freund, Golub and Nachtigal [58]. The term *harmonic Ritz pair* originates from [132], where Paige, Parlett and van der Vorst showed for a symmetric matrix, that the harmonic Ritz values are weighted harmonic means of the eigenvalues of A. In [119] and subsequent publications, *harmonic Ritz values* were promoted as better approximations to interior eigenvalues. However, in [132] it was shown in the context of eigenpair approximations for symmetric matrices from Krylov subspaces (cf. section 2.6), that a harmonic Ritz value $\widehat{\mu}^{(n)}$ at iteration n always lies between two (regular) Ritz values $\mu^{(n)}$ and $\mu^{(n+1)}$ at iterations n and $n+1$. Thus, the convergence of regular and harmonic Ritz values to an eigenvalue of A takes place at about the same iteration n and the benefits of harmonic Ritz values seem to be marginal in this respect. Nevertheless, *harmonic Ritz values* play an important role in the analysis of minimal residual methods, cf. section 2.9.

The harmonic Ritz *vectors* deserve some attention and are a common source of confusion. With regard to definition 2.32, $(z,\mu) = (\mathsf{A}w,\mu)$ is first considered as the approximate eigenpair in the original work on harmonic Ritz pairs by Morgan [119]. He then noticed that the pair (w,μ) may be a better approximation to eigenpairs close to the origin because w can be seen as the result of one step of inverse iteration applied to z. However, the implicit application of one step of inverse iteration destroys an important property in the case of a self-adjoint operator A: the orthogonality of the approximate eigenvectors. In order to see this, let $w_1,\ldots,w_n$ be a basis of harmonic Ritz vectors of $\mathcal{S}$. Then the basis $z_1 = \mathsf{A}w_1,\ldots,z_n = \mathsf{A}w_n$ forms an orthogonal basis of $\mathsf{A}\mathcal{S}$ because in the generalized eigenvalue problem (2.12) the columns of the eigenvector matrix $\mathbf{U}$ are $\langle \mathsf{A}S, \mathsf{A}S\rangle$-orthogonal. However, the harmonic Ritz vector basis $w_1,\ldots,w_n$ is not orthogonal in general.

The residuals of harmonic Ritz pairs can be characterized with the following lemma which is due to Morgan [119] and Stewart [164].

Lemma 2.33. *Let* A *and* $\mathcal{S}$ *be as in definition 2.32, let* (w,μ) *be a harmonic Ritz pair of* A *with respect to* $\mathcal{S}$. *Furthermore, assume that* $\|w\| = 1$, $\mu \neq \infty$ *and let* $\rho = \langle w, \mathsf{A}w\rangle$.

Then

$$\|\mathsf{A}w - \mu w\| = \sqrt{|\mu(\mu-\rho)|} \le |\mu|.$$

Proof. The original proofs can be found in [119, 164]. The equality follows from the fact that $\mathsf{A}w - \mu w \perp \mathsf{A}\mathcal{S}$ because $\mathsf{A}w \in \mathsf{A}\mathcal{S}$ and thus

$$\|\mathsf{A}w - \mu w\|^2 = -\langle \mu w, \mathsf{A}w - \mu w\rangle = \overline{\mu}(\mu\langle w,w\rangle - \langle w, \mathsf{A}w\rangle) = \overline{\mu}(\mu-\rho).$$

The inequality then follows with $\overline{\mu}(\mu-\rho) = |\mu|^2 - \overline{\mu}\rho \le |\mu|^2$. Note that $\overline{\mu}\rho \in \mathbb{R}$ and thus μ and ρ have to lie on a line in the complex plane through the origin. □

Note that ρ is the Rayleigh quotient of A with respect to w in lemma 2.33 and can also be characterized as the Ritz value of A with respect to the subspace $[\![w]\!]$. The following theorem appears to be new and establishes a simple yet useful connection between the harmonic Ritz residual norm $\|\mathsf{A}w - \mu w\|$ and the Ritz residual norm $\|\mathsf{A}w - \rho w\|$.

Theorem 2.34. *Let the assumptions of lemma 2.33 hold.*

Then $\frac{\mu}{\rho} \in \mathbb{R}$ *with* $\frac{\mu}{\rho} \ge 0$ *and*

$$\|\mathsf{A}w - \mu w\| = \sqrt{\frac{\mu}{\rho}}\,\|\mathsf{A}w - \rho w\|.$$

Proof. First note that $\rho \neq 0$ (otherwise the corresponding σ in definition 2.32 is zero) and that $\mu = \frac{\|\mathsf{A}w\|^2}{\overline{\rho}}$. Then with $\mathsf{A}w - \mu w \perp \mathsf{A}\mathcal{S}$ and $\mathsf{A}w \in \mathsf{A}\mathcal{S}$ the following equation is obtained:

$$\begin{aligned}\|\mathsf{A}w - \mu w\|^2 &= -\langle \mu w, \mathsf{A}w - \mu w\rangle = \overline{\mu}\langle w, \mu w - \mathsf{A}w\rangle \\ &= \overline{\mu}(\mu-\rho) = \overline{\mu}\left(\frac{\|\mathsf{A}w\|^2}{\overline{\rho}} - \rho\right) = \frac{\overline{\mu}}{\overline{\rho}}(\|\mathsf{A}w\|^2 - |\rho|^2) \\ &= \frac{\mu}{\rho}\,\|\mathsf{A}w - \rho w\|^2.\end{aligned}$$

The last equality holds because $\frac{\overline{\mu}}{\overline{\rho}}$ has to be real and non-negative and $\|\mathsf{A}w\|^2 - |\rho|^2 = \|\mathsf{A}w - \rho w\|^2$ follows from lemma 2.31. □

Remark 2.35. The equations in lemma 2.31 and lemma 2.33 are mathematically elegant and the right hand sides can be computed cheaply but the computations suffer from round-off errors. Even if the difference $\|\mathsf{A}w\|^2 - |\mu|^2$ or $|\mu(\mu-\rho)|$ is small, e.g., at the level of machine precision ε, the square root will only be of order $\sqrt{\varepsilon}$.

In an implementation, a possible workaround is to first compute $\eta_1 = \|\mathsf{A}w\|^2 - |\mu|^2$ or $\eta_2 = |\mu(\mu - \rho)|$ and to compute the residual norm explicitly if the numerically computed values η_1 or η_2 satisfy $\eta_1 \leq \varepsilon \|\mathsf{A}\|^2$ or $\eta_2 \leq \varepsilon \|\mathsf{A}\|^2$. Note that no applications of A are required because if $w = Su$ is the corresponding Ritz or harmonic Ritz vector, then $\|\mathsf{A}w - \mu w\| = \|Tu - \mu Su\|$, where $T = \mathsf{A}S$ has already been computed for setting up the (generalized) eigenvalue problem. Furthermore, the Ritz or harmonic Ritz pairs with small η_1 or η_2 are usually the ones of interest and $w = Su$ has to be computed anyway. The norm of A can often be approximated cheaply from the available quantities, e.g., by $\|\langle S, \mathsf{A}S\rangle\|_2$.

Furthermore, theorem 2.34 allows to switch between the Ritz and harmonic Ritz residual norms and if one of them can be computed accurately, the other one can be obtained with basically the same level of accuracy.

2.6. Krylov subspaces

This section introduces and characterizes Krylov subspaces and shows how they fit naturally into the Petrov–Galerkin framework which was presented in section 2.4. The results of this section are well-known in the finite-dimensional case and can be found, e.g., in the books of Saad [147] and Liesen and Strakoš [105].

Definition 2.36 (Krylov subspace)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $v \in \mathcal{H}$ and $n \in \mathbb{N}$. The n-th Krylov subspace $\mathcal{K}_n(\mathsf{A}, v)$ is defined by

$$\mathcal{K}_n(\mathsf{A}, v) := [\![v, \mathsf{A}v, \mathsf{A}^2 v, \ldots, \mathsf{A}^{n-1} v]\!]$$

for $n \geq 1$ and $\mathcal{K}_0(\mathsf{A}, v) := \{0\}$.

Some basic properties of Krylov subspaces are apparent from the definition:

Proposition 2.37. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $0 \neq v \in \mathcal{H}$. Then the following hold:*

1. *Krylov subspaces form a nested sequence of subspaces, i.e., $\mathcal{K}_{n-1}(\mathsf{A}, v) \subseteq \mathcal{K}_n(\mathsf{A}, v)$ for $n \geq 1$.*

2. *Krylov subspaces and their elements can be represented by polynomials, i.e., $\mathcal{K}_n(\mathsf{A}, v) = \{p(\mathsf{A})v \mid p \in \mathbb{P}_{n-1}\}$ for $n \geq 1$.*

3. *If $N := \dim \mathcal{H} < \infty$, there exists a uniquely defined monic polynomial $p \in \mathbb{P}_d$ of minimal degree $d \in \{1, \ldots, N\}$ such that $p(\mathsf{A})v = 0$. This polynomial is called the* minimal polynomial of v with respect to A.

Proof. Items 1. and 2. are apparent from definition 2.36. Item 3. can be found in [105, chapter 2.2]. □

Definition 2.38 (grade of a vector)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $0 \neq v \in \mathcal{H}$. The *grade of v with respect to* A is defined by

$$d(\mathsf{A}, v) := \begin{cases} d & \text{if } \dim \mathcal{K}_{d+1}(\mathsf{A}, v) = d \text{ for a } d \in \mathbb{N}_+, \\ \infty & \text{if } \dim \mathcal{K}_d(\mathsf{A}, v) = d \text{ for all } d \in \mathbb{N}_+. \end{cases}$$

Note that if $N := \dim \mathcal{H} < \infty$, the grade of a nonzero vector v with respect to A equals the degree of the minimal polynomial of v with respect to A and thus $d(\mathsf{A}, v) \leq N$. The grade $d = d(\mathsf{A}, v)$ describes the index where the Krylov subspace becomes invariant, i.e., $\mathsf{A}\mathcal{K}_d(\mathsf{A}, v) \subseteq \mathcal{K}_d(\mathsf{A}, v)$.

In the following, let the *index of an eigenvalue* λ denote the size of the largest Jordan block corresponding to the eigenvalue λ.

Proposition 2.39. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and let $0 \neq v \in \mathcal{H}$ be of grade $d = d(\mathsf{A}, v) < \infty$ with respect to A. Then the following hold:*

1. $\dim \mathcal{K}_n(\mathsf{A}, v) = n$ *for* $n \leq d$.
2. *The following statements are equivalent:*
   a) $\mathsf{A}\mathcal{K}_d(\mathsf{A}, v) = \mathcal{K}_d(\mathsf{A}, v)$.
   b) $\mathcal{K}_d(\mathsf{A}, v) \cap \mathcal{N}(\mathsf{A}) = \{0\}$.
   c) $v \in \mathsf{A}\mathcal{K}_d(\mathsf{A}, v)$.

   If $\dim \mathcal{H} < \infty$ *then also the following statement is equivalent to the above:*
   d) $v \in \mathcal{R}(\mathsf{A}^m)$, *where m is the index of the zero eigenvalue of* A.
3. *If* A *is nonsingular, then* $\mathsf{A}\mathcal{K}_d(\mathsf{A}, v) = \mathcal{K}_d(\mathsf{A}, v)$.

Proof. 1. Clear from the definition of d.

2. a)$\Longrightarrow$b): If there exists a nonzero $z \in \mathcal{K}_d(\mathsf{A}, v) \cap \mathcal{N}(\mathsf{A})$, then $\dim \mathsf{A}\mathcal{K}_d(\mathsf{A}, v) \leq d - 1$ which is a contradiction to $\dim \mathsf{A}\mathcal{K}_d(\mathsf{A}, v) = \dim \mathcal{K}_d(\mathsf{A}, v) = d$.

   b)$\Longrightarrow$c): From $\mathcal{K}_d(\mathsf{A}, v) \cap \mathcal{N}(\mathsf{A}) = \{0\}$ follows that $\dim \mathsf{A}\mathcal{K}_d(\mathsf{A}, v) = d$ and because of $\mathsf{A}\mathcal{K}_d(\mathsf{A}, v) \subseteq \mathcal{K}_d(\mathsf{A}, v)$ that $\mathsf{A}\mathcal{K}_d(\mathsf{A}, v) = \mathcal{K}_d(\mathsf{A}, v)$. Then $v \in \mathsf{A}\mathcal{K}_d(\mathsf{A}, v)$ holds.

   c)$\Longrightarrow$a): If $v \in \mathsf{A}\mathcal{K}_d(\mathsf{A}, v)$ then $\mathsf{A}\mathcal{K}_d = [\![v]\!] + [\![\mathsf{A}v, \ldots, \mathsf{A}^d v]\!] = \mathcal{K}_d(\mathsf{A}, v)$ because $v, \mathsf{A}v, \ldots, \mathsf{A}^{d-1}v$ are by the definition of $d = \dim \mathcal{K}_d(\mathsf{A}, v)$ linearly independent.

   a)$\Longrightarrow$d): It follows from a) that $\mathsf{A}^i\mathcal{K}_d(\mathsf{A}, v) = \mathcal{K}_d(\mathsf{A}, v)$ for any $i \in \mathbb{N}$. Thus also $v \in \mathcal{K}_d(\mathsf{A}, v) = \mathsf{A}^m\mathcal{K}_d(\mathsf{A}, v) \subseteq \mathcal{R}(\mathsf{A}^m)$.

   d)$\Longrightarrow$b): Let
   $$\mathsf{A} = S \begin{bmatrix} \mathbf{J} & \\ & \mathbf{N} \end{bmatrix} S^{-1}$$
   be a Jordan decomposition where $\mathbf{J}$ is nonsingular and $\mathbf{N}$ is nilpotent, i.e., $\mathbf{N}^m = 0$. If $v \in \mathcal{R}(\mathsf{A}^m)$, then $v = S \begin{bmatrix} \mathbf{J} & \\ & 0 \end{bmatrix} S^{-1}s$ for a nonzero $s \in \mathcal{H}$. Assume that b) does not hold, i.e., there exists a nonzero polynomial $p \in \mathbb{P}_{d-1}$ such that $\mathsf{A}p(\mathsf{A})v = 0$. Then
   $$0 = S \begin{bmatrix} \mathbf{J}p(\mathbf{J}) & \\ & 0 \end{bmatrix} S^{-1}v$$

which is equivalent to

$$0 = S \begin{bmatrix} p(\mathbf{J}) & \\ & 0 \end{bmatrix} S^{-1} v = p(\mathsf{A})v$$

because $\mathbf{J}$ is nonsingular. However, this means that there exists a nonzero polynomial of degree less than d with $p(\mathsf{A})v = 0$, which is a contradiction to the fact that d is the degree of the minimal polynomial of v with respect to A.

3. Follows directly from 2.b).

□

The condition 2.d) in the above proposition that involves the index of the zero eigenvalue appeared in a work of Ipsen and Meyer [82], see also the discussion following corollary 2.41.

The remaining part of this section shows how Krylov subspaces naturally fit into the setting of the (Petrov–)Galerkin method for solving the linear system (2.3), cf. section 2.4. In the projection framework of section 2.4, two subspaces of $\mathcal{H}$ of equal dimension $n < \infty$ can be chosen: the search space $\mathcal{S}$ and the constraint space $\mathcal{C}$. It was shown that the projection process described by equations (2.4) and (2.5) is well defined if and only if $\mathsf{A}\mathcal{S} \oplus \mathcal{C}^\perp = \mathcal{H}$ and a solution of the linear system (2.3) is found if and only if $r_0 \in \mathsf{A}\mathcal{S}$, cf. theorem 2.25 and corollary 2.26.

Because the Krylov subspace built with the initial residual r_0, i.e., $\mathcal{K}_n(\mathsf{A}, r_0)$, eventually becomes invariant at index $d = d(\mathsf{A}, r_0)$, the idea is to use the sequence of Krylov subspaces $\mathcal{K}_1(\mathsf{A}, r_0) \subseteq \ldots \subseteq \mathcal{K}_d(\mathsf{A}, r_0)$ as search spaces in a (Petrov–) Galerkin method.

Definition 2.40. A Krylov subspace method as described above is called *well defined* if the (Petrov–)Galerkin method is well defined for each $n \in \{1, \ldots, d\}$ and it terminates with $x_d \in \mathcal{H}$ satisfying $\mathsf{A}x_d = b$.

Lemma 2.28 states conditions under which the choices $\mathcal{C} = \mathcal{S}$ or $\mathcal{C} = \mathsf{A}\mathcal{S}$ of the constraint space lead to a well-defined (Petrov–)Galerkin method. The following corollary summarizes the results.

Corollary 2.41. *Consider a consistent linear system* $\mathsf{A}x = b$ *with* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *and* $b \in \mathcal{H}$. *Furthermore, let* $x_0 \in \mathcal{H}$ *be an initial guess with corresponding initial residual* $r_0 = b - \mathsf{A}x_0$ *such that* $d = d(\mathsf{A}, r_0) < \infty$ *and let* $\mathcal{K}_d(\mathsf{A}, r_0) \cap \mathcal{N}(\mathsf{A}) = \{0\}$. *The sequence of iterates* $(x_n)_{n \in \{1,\ldots,d\}}$ *that satisfy*

$$x_n = x_0 + s_n \quad \textit{with} \quad s_n \in \mathcal{K}_n(\mathsf{A}, r_0)$$

and

$$r_n := b - \mathsf{A}x_n \perp \mathcal{C}_n$$

is well defined and x_d *is a solution of the linear system* (2.3), *i.e.,* $\mathsf{A}x_d = b$, *if one of the following conditions holds:*

1. *$\mathcal{C}_n = \mathcal{K}_n(\mathsf{A}, r_0)$ and A is self-adjoint and positive semidefinite. Then the iterates x_n satisfy the optimality property*

$$\|x - x_n\|_{\mathsf{A}} = \min_{z \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0)} \|x - z\|_{\mathsf{A}} .$$

2. *$\mathcal{C}_n = \mathsf{A}\mathcal{K}_n(\mathsf{A}, r_0)$. Then the iterates x_n satisfy the optimality property*

$$\|b - \mathsf{A}x_n\| = \min_{z \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0)} \|b - \mathsf{A}z\| .$$

Proof. In both cases, the well-definedness and optimality property of the approximate solutions follows from lemma 2.28. $\mathsf{A}x_d = b$ follows from corollary 2.26 and the equivalence of 2.b) and 2.c) in proposition 2.39. □

Corollary 2.41 shows that a well-defined Krylov subspace method finds the exact solution in a finite number of steps if $d(\mathsf{A}, r_0) < \infty$ and $\mathcal{K}_d(\mathsf{A}, r_0) \cap \mathcal{N}(\mathsf{A}) = \{0\}$. The first condition is met, e.g., if $\mathcal{H}$ is finite-dimensional. In the finite-dimensional case, the second condition is equivalent to $r_0 \in \mathcal{R}(\mathsf{A}^m)$, where m is the index of the zero eigenvalue of A, cf. item 2.d) in proposition 2.39. The observation that a solution is contained in a Krylov subspace if and only if $r_0 \in \mathcal{R}(\mathsf{A}^m)$ has already been made by Ipsen and Meyer in [82]. Furthermore, they showed that the unique solution in the Krylov subspace is the Drazin inverse solution, see [35, 82].

Definition 2.42 (Drazin inverse)**.** Let $N := \dim \mathcal{H} < \infty$ and let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$. The *Drazin inverse* of A is defined as the unique $\mathsf{A}^D \in \mathcal{L}(\mathcal{H})$ that satisfies

$$\mathsf{A}^D\mathsf{A}\mathsf{A}^D = \mathsf{A}^D, \qquad \mathsf{A}^D\mathsf{A} = \mathsf{A}\mathsf{A}^D \qquad \text{and} \qquad \mathsf{A}^{m+1}\mathsf{A}^D = \mathsf{A}^m,$$

where m is the index of the zero eigenvalue of A.

Proposition 2.43. *Let $\dim \mathcal{H} < \infty$ and let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$. If*

$$\mathsf{A} = S \begin{bmatrix} \mathbf{J} & \\ & \mathbf{N} \end{bmatrix} S^{-1}$$

is a Jordan decomposition of A with nonsingular $\mathbf{J}$ and nilpotent $\mathbf{N}$, then the Drazin inverse of A is given by

$$\mathsf{A}^D = S \begin{bmatrix} \mathbf{J}^{-1} & \\ & 0 \end{bmatrix} S^{-1}.$$

Theorem 2.44. *Let $\dim \mathcal{H} < \infty$ and let $\mathsf{A}x = b$ be a consistent linear system with $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $b \in \mathcal{R}(\mathsf{A})$. Furthermore, let $x_0 \in \mathcal{H}$ be an initial guess with corresponding initial residual $r_0 = b - \mathsf{A}x_0$ of grade $d = d(\mathsf{A}, r_0)$.*

1. *There exists a $x_d \in \mathcal{K}_d(\mathsf{A}, r_0)$ with $\mathsf{A}x_d = b$ if and only if $r_0 \in \mathcal{R}(\mathsf{A}^m)$, where m is the index of the zero eigenvalue of A.*

2. *If a solution $x_d \in \mathcal{K}_d(\mathsf{A}, r_0)$ with $\mathsf{A}x_d = b$ exists, then it is unique and $x_d = \mathsf{A}^D b$ is the Drazin inverse solution.*

Proof. See theorems 2–3 in [82]. □

2.7. Arnoldi and Lanczos relations

For $n \leq d(\mathsf{A}, v)$, the Krylov basis $v, \mathsf{A}v, \mathsf{A}^2v, \ldots, \mathsf{A}^{n-1}v$ of $\mathcal{K}_n(\mathsf{A}, v)$ is ill-conditioned in practice even for a moderate order of n and should be avoided in the presence of round-off errors. Instead, an orthonormal basis can be used which can be obtained with one of the variants of the Arnoldi algorithm [6], e.g., the Gram–Schmidt variant in algorithm 2.1.

Algorithm 2.1 Arnoldi algorithm (modified Gram–Schmidt). Implemented in [60] as `krypy.utils.Arnoldi`, also with iterated Gram–Schmidt (`ortho='dmgs'`) and Householder (`ortho='house'`) orthogonalization.

```
Input: $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$ and $n \in \mathbb{N}_+$.
$v_1 = \frac{v}{\|v\|}$
for $k = 1, \ldots, n$ do
    $z \leftarrow \mathsf{A}v_k$
    for $m = 1, \ldots, k$ do
        $h_{m,k} = \langle v_m, z \rangle$
        $z \leftarrow z - v_m h_{m,k}$
    end for
    $h_{k+1,k} = \|z\|$
    if $h_{k+1,k} = 0$ then
        return $V_k = [v_1, \ldots, v_k] \in \mathcal{H}^k$ and $\mathbf{H}_k = [h_{i,j}]_{i,j=1,\ldots,k} \in \mathbb{C}^{k,k}$
    end if
    $v_{k+1} = \frac{z}{h_{k+1,k}}$
end for
return $V_{n+1} = [v_1, \ldots, v_{n+1}] \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}_n = [h_{i,j}]_{\substack{i=1,\ldots,n+1\\ j=1,\ldots,n}} \in \mathbb{C}^{n+1,n}$
```

In order to facilitate the characterization of the results of the Arnoldi algorithm it is helpful to extend the definition of an upper Hessenberg matrix to the non-square case:

Definition 2.45 (extended upper Hessenberg matrix)**.** $\underline{\mathbf{H}} = [h_{i,j}] \in \mathbb{C}^{n+1,n}$ is called an *extended upper Hessenberg matrix* if $h_{i,j} = 0$ for all $i > j + 1$. An extended upper Hessenberg matrix $\underline{\mathbf{H}}$ is said to be *unreduced* if $h_{i+1,i} \neq 0$ for $i \in \{1, \ldots, n\}$.

In the following, an extended upper Hessenberg matrix $\underline{\mathbf{H}} = [h_{i,j}] \in \mathbb{C}^{n+1,n}$ is denoted by an underline and its upper square part $\mathbf{H}$ without, i.e.,

$$\underline{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ h_{n+1,n} e_n^\mathsf{T} \end{bmatrix}.$$

In order to describe the computed quantities of algorithm 2.1, the following definition is helpful.

Definition 2.46 (Arnoldi relation)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$ and $n \in \mathbb{N}_+$.

1. An *Arnoldi relation* for $\mathcal{K}_n(\mathsf{A},v)$ is defined by $V_{n+1}=[v_1,\ldots,v_{n+1}]\in\mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}\in\mathbb{C}^{n+1,n}$ if $\underline{\mathbf{H}}$ is an extended unreduced upper Hessenberg matrix and if the following conditions are met:

$$v_1\in[\![v]\!],\qquad \langle V_{n+1},V_{n+1}\rangle=\mathbf{I}_{n+1}\quad\text{and}\quad \mathsf{A}V_n=V_{n+1}\underline{\mathbf{H}},$$

   where $V_n=[v_1,\ldots,v_n]$.

2. An *invariant Arnoldi relation* for $\mathcal{K}_n(\mathsf{A},v)$ is defined by $V_n=[v_1,\ldots,v_n]\in\mathcal{H}^n$ and $\mathbf{H}\in\mathbb{C}^{n,n}$ if $\mathbf{H}$ is an unreduced upper Hessenberg matrix and if the following conditions are met:

$$v_1\in[\![v]\!],\qquad \langle V_n,V_n\rangle=\mathbf{I}_n\quad\text{and}\quad \mathsf{A}V_n=V_n\mathbf{H}.$$

The vectors $v_1,v_2,\ldots$ are called an *Arnoldi basis* and $\underline{\mathbf{H}}$ and $\mathbf{H}$ are called an *Arnoldi matrix.*

In the above definition, it is not yet clear how an Arnoldi relation is linked to the Krylov subspace $\mathcal{K}_n(\mathsf{A},v)$. The next lemma shows that an Arnoldi relation consists of a nested orthonormal basis of Krylov subspaces and a projected version of the operator A.

Lemma 2.47. *Let $\mathsf{A}\in\mathcal{L}(\mathcal{H})$, $0\neq v\in\mathcal{H}$ and $n\in\mathbb{N}_+$. The following statements are equivalent for $V_{n+1}=[v_1,\ldots,v_{n+1}]\in\mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}\in\mathbb{C}^{n+1,n}$:*

1. *V_{n+1} and $\underline{\mathbf{H}}$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A},v)$.*

2. *$v_1,\ldots,v_k$ is an orthonormal basis of $\mathcal{K}_k(\mathsf{A},v)$ for $k\in\{1,\ldots,n+1\}$ and $\underline{\mathbf{H}}=\langle V_{n+1},\mathsf{A}V_n\rangle$.*

Proof. 1.$\Longrightarrow$2.: The orthonormality of V_{n+1} is clear. Because $\mathsf{A}V_n=V_{n+1}\underline{\mathbf{H}}$ holds by definition also $\langle V_{n+1},\mathsf{A}V_n\rangle=\langle V_{n+1},V_{n+1}\rangle\underline{\mathbf{H}}=\underline{\mathbf{H}}$ holds. It remains to show that $[\![v_1,\ldots,v_k]\!]=\mathcal{K}_k(\mathsf{A},v)$ for $k\in\{1,\ldots,n+1\}$. For $k=1$ this is true by definition. Assume that $[\![v_1,\ldots,v_k]\!]=\mathcal{K}_k(\mathsf{A},v)$ holds for a fixed $k\in\{1,\ldots,n\}$. Then $v_{k+1}=\frac{1}{h_{k+1,k}}(\mathsf{A}v_k-\sum_{i=1}^k h_{i,k}v_i)\in\mathcal{K}_{k+1}(\mathsf{A},v)$. Because $v_1,\ldots,v_{k+1}\in\mathcal{K}_{n+1}(\mathsf{A},v)$ are linearly independent and $\dim\mathcal{K}_{k+1}\leq k+1$ it follows that $[\![v_1,\ldots,v_{k+1}]\!]=\mathcal{K}_{k+1}(\mathsf{A},v)$. By induction the first part of the proof is complete.

2.$\Longrightarrow$1.: Again, the orthonormality of V_{n+1} is clear and $v_1\in[\![v]\!]$ follows from $[\![v_1]\!]=\mathcal{K}_1(\mathsf{A},v)=[\![v]\!]$. Furthermore, $[\![\mathsf{A}V_n]\!]\subseteq\mathcal{K}_{n+1}(\mathsf{A},v)=[\![V_{n+1}]\!]$ holds and thus

$$\mathsf{A}V_n=\mathsf{P}_{\mathcal{K}_{n+1}(\mathsf{A},v)}\mathsf{A}V_n=V_{n+1}\langle V_{n+1},\mathsf{A}V_n\rangle=V_{n+1}\underline{\mathbf{H}}.$$

Because of $\mathsf{A}v_j\in\mathcal{K}_{j+1}(\mathsf{A},v)=[\![V_{j+1}]\!]$ and the orthonormality of V_{n+1}, the entries of $\underline{\mathbf{H}}$ are $h_{i,j}=\langle v_i,\mathsf{A}v_j\rangle=0$ for $2\leq j+1<i\leq n+1$ and thus $\underline{\mathbf{H}}$ is an extended upper Hessenberg matrix. In order to show that it is also unreduced, it is assumed that

$h_{i+1,i} = 0$ for a $i \in \{1,\dots,n\}$. For $j \in \{1,\dots,i\}$ there exist polynomials $q_j \in \mathbb{P}_{j-1}$ of degree $j-1$ such that $v_j = q_j(\mathsf{A})v$. Then it follows from $\mathsf{A}V_n = V_{n+1}\underline{\mathbf{H}}$ that

$$0 = \mathsf{A}v_i - \sum_{j=1}^{i} h_{j,i}v_j = \mathsf{A}q_i(\mathsf{A})v - \sum_{j=1}^{i} h_{j,i}q_j(\mathsf{A})v = p(\mathsf{A})v,$$

where $p \in \mathbb{P}_i$ is of degree i. Thus $d(\mathsf{A},v) \le i$ which is a contradiction to $d(\mathsf{A},v) \ge \dim \mathcal{K}_{n+1}(\mathsf{A},v) = n+1$. □

With the above definition, the two possible results of algorithm 2.1 are characterized in the following lemma:

Lemma 2.48. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$ and $n \in \mathbb{N}_+$.*

1. *If algorithm 2.1 terminates in line 14, then $n < d(\mathsf{A},v)$ and V_{n+1} and $\underline{\mathbf{H}}_n$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A},v)$.*
2. *If algorithm 2.1 terminates in line 10 at iteration k then $k = d(\mathsf{A},v)$ and V_k and $\mathbf{H}_k$ define an invariant Arnoldi relation for $\mathcal{K}_k(\mathsf{A},v)$.*

Proof. The proof is analogous to the proofs of propositions 6.4–6.6 in the book of Saad [147]. □

The Arnoldi vectors can be represented as a polynomial in A whose zeros are the Ritz values of A with respect to a Krylov subspace:

Lemma 2.49. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$ and let $V_{n+1} = [v_1,\dots,v_{n+1}] \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}_n \in \mathbb{C}^{n+1,n}$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A},v)$.*

Then

$$v_{i+1} = \alpha_{i+1}\frac{p_i(\mathsf{A})v}{\|p_i(\mathsf{A})v\|}$$

for all $i \in \{1,\dots,n\}$, where $|\alpha_{i+1}| = 1$ and $p_i \in \mathbb{P}_i$ is the characteristic polynomial of $\mathbf{H}_i$, i.e., $p_i(\lambda) = \prod_{k=1}^{i}(\lambda - \theta_k^{(i)})$ with the Ritz values $\theta_1^{(i)},\dots,\theta_i^{(i)}$ of A with respect to $\mathcal{K}_i(\mathsf{A},v)$.

Proof. It has been shown by Saad in [144, Theorem 5.1] that the characteristic polynomial of $\mathbf{H}_i$ minimizes the norm $\|p(\mathsf{A})v\|$ over all $p \in \mathbb{P}_{i,\infty}$. Thus

$$\|p_i(\mathsf{A})v\| = \min_{p\in\mathbb{P}_{i,\infty}} \|p(\mathsf{A})v\| = \min_{w\in\mathcal{K}_i(\mathsf{A},v)} \left\|\mathsf{A}^i v - w\right\| = \left\|\mathsf{P}_{\mathcal{K}_i(\mathsf{A},v)^\perp}\mathsf{A}^i v\right\|$$

and because the minimizer is unique $p_i(\mathsf{A})v = \mathsf{P}_{\mathcal{K}_i(\mathsf{A},v)^\perp}\mathsf{A}^i v$ holds. The proof is complete by recognizing that $v_{i+1} \in [\![\mathsf{P}_{\mathcal{K}_i(\mathsf{A},v)^\perp}\mathsf{A}^i v]\!]$. □

Arnoldi relations as defined in definition 2.46 are not unique and different variants of the Arnoldi algorithm may generate different Arnoldi relations for a Krylov subspace. Though the results of the Gram–Schmidt Arnoldi algorithm (algorithm 2.1) are uniquely determined, this is not true for the Householder Arnoldi algorithm [180] where some freedom is left in the construction of the involved Householder transformations which results in possibly differing Arnoldi relations. Furthermore, non-unique Arnoldi relations are constructed by other means in section 3.4.3 for Krylov subspaces with perturbed operators and initial vectors. The following lemma characterizes all possible Arnoldi relations for a Krylov subspace.

Lemma 2.50. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$ and let $V_{n+1} \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}} \in \mathbb{C}^{n+1,n}$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A}, v)$. The following statements are equivalent for $W_{n+1} \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{G}} \in \mathbb{C}^{n+1,n}$:*

1. *W_{n+1} and $\underline{\mathbf{G}}$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A}, v)$.*
2. *$W_{n+1} = V_{n+1}\mathbf{D}_{n+1}$ and $\mathbf{G} = \overline{\mathbf{D}}_{n+1}\underline{\mathbf{H}}\mathbf{D}_n$, where $\mathbf{D}_k = \operatorname{diag}(d_1, \ldots, d_k)$ is a diagonal matrix with diagonal entries $d_i \in \mathbb{C}$ such that $|d_i| = 1$ for $i \in \{1, \ldots, n+1\}$.*

Proof. Let $V_{n+1} = [v_1, \ldots, v_{n+1}]$ and $W_{n+1} = [w_1, \ldots, w_{n+1}]$.

1.$\Longrightarrow$2.: By lemma 2.47 $\mathcal{K}_k(\mathsf{A}, v) = [\![v_1, \ldots, v_k]\!] = [\![w_1, \ldots, w_k]\!]$ holds for all $k \in \{1, \ldots, n+1\}$. It follows that $[\![v_k]\!] = [\![w_k]\!]$ because V_{n+1} and W_{n+1} are orthogonal. Then $v_k = w_k d_k$ with $|d_k| = 1$ follows from $\|v_k\| = \|w_k\| = 1$.

2.$\Longrightarrow$1.: It is clear that $w_1 = v_1 d_1 \in [\![v]\!]$, $\langle W_{n+1}, W_{n+1}\rangle = \mathbf{I}_{n+1}$ and that $\underline{\mathbf{G}}$ is an extended unreduced upper Hessenberg matrix. The proof is complete by noticing that

$$\mathsf{A}W_n = \mathsf{A}V_n\mathbf{D}_n = V_{n+1}\underline{\mathbf{H}}\mathbf{D}_n = V_{n+1}\mathbf{D}_{n+1}\overline{\mathbf{D}}_{n+1}\underline{\mathbf{H}}\mathbf{D}_n = W_{n+1}\underline{\mathbf{G}}.$$

□

Remark 2.51. Lemma 2.47 and lemma 2.50 also hold in the case of invariant Arnoldi relations and the proofs are analogous. The statements are omitted here for brevity.

The following proposition is a standard result concerning the evaluation of a polynomial in A if an Arnoldi relation is known.

Proposition 2.52. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$ and let $V_{n+1} = [v_1, \ldots, v_{n+1}] \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}} \in \mathbb{C}^{n+1,n}$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A}, v)$. Then the following holds for any polynomial $p \in \mathbb{P}_i$ with $i < n$:*

$$p(\mathsf{A})v_1 = V_n p(\mathbf{H}_n)e_1 = V_{i+1}p(\mathbf{H}_{i+1})e_1.$$

Proof. The proof can be found in [148, proposition 6.4]. □

If the linear operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ is self-adjoint and V_{n+1} and $\underline{\mathbf{H}}_n$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A}, v)$, then trivially

$$\mathbf{H}_n = \langle V_n, \mathsf{A}V_n\rangle = \langle \mathsf{A}V_n, V_n\rangle = \mathbf{H}_n^{\mathsf{H}}$$

and thus $\mathbf{H}_n$ is Hermitian and tridiagonal. This represents a remarkable special case of an Arnoldi relation and is referred to as a *Lanczos relation*:

Definition 2.53 (Lanczos relation)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$, $n \in \mathbb{N}_+$ and let $V_{n+1} \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}} = \begin{bmatrix} \mathbf{H} \\ h_{n+1,n} e_n^\mathsf{T} \end{bmatrix}$ define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A}, v)$. If $\mathbf{H} = \mathbf{H}^\mathsf{H}$, then V_{n+1} and $\underline{\mathbf{H}}$ define a *Lanczos relation.* Because the (extended) Hessenberg matrix is tridiagonal, it is denoted by

$$\underline{\mathbf{T}} := \underline{\mathbf{H}} = \begin{bmatrix} \mathbf{T} \\ \delta_{n+1} e_n^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \gamma_1 & \overline{\delta}_2 & & & \\ \delta_2 & \gamma_2 & \overline{\delta}_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \delta_{n-1} & \gamma_{n-1} & \overline{\delta}_n \\ & & & \delta_n & \gamma_n \\ & & & & \delta_{n+1} \end{bmatrix}. \tag{2.13}$$

Analogous to the invariant Arnoldi relation from definition 2.46, the invariant Lanczos relation can be defined with a basis tuple $V_n \in \mathcal{H}^n$ and a Hermitian tridiagonal matrix $\mathbf{T} \in \mathbb{C}^{n,n}$. Note that according to lemma 2.50, an equivalent Lanczos relation can be constructed where the tridiagonal matrix is real and symmetric.

The importance of Lanczos relations is primarily of algorithmic nature: if $\mathbf{H}_n$ is tridiagonal, then $h_{m,k} = 0$ for $m < k-1$ and thus all but two orthogonalizations in the Arnoldi algorithm can be omitted, i.e., the *for*-loop in line 4 of algorithm 2.1 only runs from $k-1$ to k. The resulting algorithm is given in algorithm 2.2 and is called the *Lanczos algorithm*, attributed to the works of Lanczos [102, 103] in 1950 and 1952. However, there is a one-to-one correspondence between the Lanczos algorithm and the Stieltjes recurrence for the computation of orthonormal polynomials whose history goes far beyond the 20th century. An extensive description of this link and profound historical remarks can be found in chapter 3 of [105].

In comparison with the *full recurrence* in the Arnoldi algorithm, the *short recurrence* in the Lanczos algorithm reduces the computation time and memory requirements to a constant per iteration. However, in order to carry over the mathematical and computational appeal to practical applications, counter measures have to be taken against the Lanczos algorithm's high sensitivity to round-off errors. Some implications of round-off errors and possible remedies are discussed briefly in section 2.10.

2.8. CG method

In order to solve a linear system $\mathsf{A}x = b$ with a self-adjoint and positive-definite operator A, Hestenes and Stiefel introduced the *method of conjugate gradients* (CG method) in their seminal paper [80] in 1952. In contrast to the original publication, the presentation in this section makes use of the Galerkin framework (cf. section 2.4).

Throughout this section the operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ of the linear system (2.3) is assumed to be self-adjoint and positive semidefinite and the linear system is assumed

Algorithm 2.2 Lanczos algorithm. Implemented in [60] as `krypy.utils.Arnoldi (ortho='lanczos')`.

```
Input: Self-adjoint $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $0 \neq v \in \mathcal{H}$ and $n \in \mathbb{N}$.
$v_0 = 0$, $\delta_1 = 0$
$v_1 = \frac{v}{\|v\|}$
for $k = 1, \ldots, n$ do
    $z \leftarrow \mathsf{A} v_k$
    $z \leftarrow z - \delta_k v_{k-1}$
    $\gamma_k = \langle v_k, z \rangle$
    $z \leftarrow z - \gamma_k v_k$
    $\delta_{k+1} = \|z\|$
    if $\delta_{k+1} = 0$ then
        return $V_k = [v_1, \ldots, v_k] \in \mathcal{H}^k$ and $\mathbf{T}_k \in \mathbb{R}^{k,k}$ ($k \times k$-submatrix of (2.13))
    end if
    $v_{k+1} = \frac{z}{\delta_{k+1}}$
end for
return $V_{n+1} = [v_1, \ldots, v_{n+1}] \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{T}}_n \in \mathbb{R}^{n+1,n}$ as in (2.13)
```

to be consistent, i.e., $b \in \mathcal{R}(\mathsf{A})$. Furthermore, an initial guess $x_0 \in \mathcal{H}$ is assumed to be given such that the corresponding initial residual $r_0 = b - \mathsf{A}x_0$ satisfies $d = d(\mathsf{A}, r_0) < \infty$, which holds, e.g., if $\dim \mathcal{H} < \infty$.

Corollary 2.41 gives a sufficient condition for a Galerkin method with Krylov subspaces as search and constraint spaces to be well defined in the case where A is self-adjoint and positive semidefinite. The assumption $\mathcal{K}_d(\mathsf{A}, r_0) \cap \mathcal{N}(\mathsf{A}) = \{0\}$ is satisfied since $\mathcal{K}_d(\mathsf{A}, r_0) \subseteq \mathcal{R}(\mathsf{A}) \perp \mathcal{N}(\mathsf{A})$ holds because A is self-adjoint. Item 1. in corollary 2.41 states that the Galerkin method is well defined if the search and constraint spaces are chosen as Krylov subspaces, i.e., $\mathcal{S}_n = \mathcal{C}_n = \mathcal{K}_n(\mathsf{A}, r_0)$. Furthermore, the iterates x_n minimize the A-norm of the error, i.e.,

$$\|x - x_n\|_{\mathsf{A}} = \min_{z \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0)} \|x - z\|_{\mathsf{A}} ,$$

and x_d is a solution of the linear system.

Because A is self-adjoint, the Lanczos algorithm (cf. section 2.7) can be applied with A and the initial vector r_0. For $n < d = d(\mathsf{A}, r_0)$, the Lanczos algorithm generates a $V_{n+1} \in \mathcal{H}^{n+1}$ and a $\underline{\mathbf{T}}_n = \begin{bmatrix} \mathbf{T}_n \\ \delta_{n+1} e_n^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{n+1,n}$ that define a Lanczos relation for $\mathcal{K}_n(\mathsf{A}, r_0)$, cf. definition 2.53. In the d-th step, the generated $V_d \in \mathcal{H}^d$ and $\mathbf{T}_d \in \mathbb{R}^{d,d}$ define an invariant Lanczos relation for $\mathcal{K}_d(\mathsf{A}, r_0)$.

For $n \leq d$, the Galerkin method with $\mathcal{S}_n = \mathcal{C}_n = \mathcal{K}_n(\mathsf{A}, r_0) = [\![V_n]\!]$ is well defined and by theorem 2.25 the matrix $\mathbf{T}_n = \langle V_n, \mathsf{A}V_n \rangle$ is nonsingular and thus Hermitian and positive definite. Note that although A is allowed to be singular, the matrix $\mathbf{T}_n$ is nonsingular. By corollary 2.26, the iterates x_n are given by

$$x_n = x_0 + V_n \langle V_n, \mathsf{A}V_n \rangle^{-1} \langle V_n, r_0 \rangle = x_0 + V_n \mathbf{T}_n^{-1} e_1 \|r_0\| .$$

A rather technical derivation, which is omitted here, then leads to the CG algorithm 2.3, cf. [105, section 2.5.1] for the full derivation.

Algorithm 2.3 CG algorithm.

```
Input: Self-adjoint and positive-semidefinite $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, right hand side $b \in \mathcal{R}(\mathsf{A})$,
    initial guess $x_0 \in \mathcal{H}$ and maximal number of iterations $n_{\max} \in \mathbb{N}$.
$r_0 = b - \mathsf{A}x_0$, $p_0 = r_0$
for $n = 1, \ldots, n_{\max}$ do
    $\alpha_{n-1} = \frac{\|r_{n-1}\|^2}{\langle p_{n-1}, \mathsf{A}p_{n-1}\rangle}$
    $x_n = x_{n-1} + \alpha_{n-1}p_{n-1}$
    $r_n = r_{n-1} - \alpha_{n-1}\mathsf{A}p_{n-1}$
    if stopping criterion is reached then
        return $x_n$
    end if
    $\omega_n = \frac{\|r_n\|^2}{\|r_{n-1}\|^2}$
    $p_n = r_n + \omega_n p_{n-1}$
end for
return $x_{n_{\max}}$
```

Convergence of CG

The most important properties of the CG method (algorithm 2.3) are gathered in the following theorem.

Theorem 2.54 (Convergence of CG). *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be self-adjoint and positive semidefinite, $b \in \mathcal{R}(\mathsf{A})$ and $x \in \mathcal{H}$ such that $\mathsf{A}x = b$. Furthermore, let $x_0 \in \mathcal{H}$ be an initial guess with corresponding initial residual $r_0 = b - \mathsf{A}x_0$ such that $d = d(\mathsf{A}, r_0) < \infty$.*

Then the CG method (see algorithm 2.3) is well defined and the following holds:

1. *The error norm in iteration $n \leq d$ satisfies*

$$\|x - x_n\|_{\mathsf{A}} = \min_{z \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0)} \|x - z\|_{\mathsf{A}} = \min_{p \in \mathbb{P}_{n,0}} \|p(\mathsf{A})(x - x_0)\|_{\mathsf{A}}.$$

2. *The error in iteration $n \leq d$ is given by*

$$x - x_n = p_n^{CG}(\mathsf{A})(x - x_0)$$

with $p_n^{CG} \in \mathbb{P}_{n,0}$ and

$$p_n^{CG}(\lambda) = \prod_{i=1}^{n} \left(1 - \frac{\lambda}{\theta_i^{(n)}}\right),$$

where $\theta_1^{(n)}, \ldots, \theta_n^{(n)} \in \mathbb{R}_+$ are the Ritz values of A with respect to $\mathcal{K}_n(\mathsf{A}, r_0)$, i.e., the eigenvalues of $\mathbf{T}_n = \langle V_n, \mathsf{A}V_n\rangle$ (cf. section 2.5).

Proof. The well-definedness of the method and the first equality in item 1. follow from corollary 2.41. The last equality in item 1. follows from the fact that for any $y \in \mathcal{K}_n(\mathsf{A}, r_0)$, there exists a $q \in \mathbb{P}_{n-1}$ such that $y = q(\mathsf{A})r_0 = q(\mathsf{A})\mathsf{A}(x - x_0)$. A proof of item 2. can be found in chapter 5.6 in [105]. □

Item 1 in theorem 2.54 states that the iterates of the CG method minimize the error in the A-norm in exact arithmetic. This allows the construction of a priori bounds on the error in the A-norm, i.e., bounds that are based on certain properties of the operator A.

In the finite-dimensional case, i.e., $N := \dim \mathcal{H} < \infty$, a self-adjoint operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ has an eigen-decomposition of the form

$$\mathsf{A} = U\mathbf{D}U^\star,$$

where $U = [u_1, \ldots, u_N] \in \mathcal{H}^N$ is orthonormal, i.e., $U^\star U = \langle U, U \rangle = \mathbf{I}_N$ with $U^\star x := \langle U, x \rangle$, and $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ is a diagonal matrix with A's eigenvalues. Let the eigenvalues be sorted and let $j+1$ be the index of the first nonzero eigenvalue, i.e., $0 = \lambda_1 = \ldots = \lambda_j < \lambda_{j+1} \le \ldots \le \lambda_N$. Furthermore, let $U_2 := [u_{j+1}, \ldots, u_N]$ and $\mathbf{D}_2 := \operatorname{diag}(\lambda_{j+1}, \ldots, \lambda_N)$. Then, in the setting of theorem 2.54, the n-th iterate of the CG method x_n satisfies

$$\begin{aligned}
\|x - x_n\|_{\mathsf{A}} &= \min_{p \in \mathbb{P}_{n,0}} \|p(\mathsf{A})(x - x_0)\|_{\mathsf{A}} = \min_{p \in \mathbb{P}_{n,0}} \left\| \mathsf{A}^{\frac{1}{2}} p(\mathsf{A})(x - x_0) \right\| \\
&= \min_{p \in \mathbb{P}_{n,0}} \left\| p(\mathsf{A})\mathsf{A}^{\frac{1}{2}}(x - x_0) \right\| = \min_{p \in \mathbb{P}_{n,0}} \left\| U p(\mathbf{D}) U^\star U \mathbf{D}^{\frac{1}{2}} U^\star (x - x_0) \right\| \\
&= \min_{p \in \mathbb{P}_{n,0}} \left\| p(\mathbf{D}_2)\mathbf{D}_2^{\frac{1}{2}} U_2^\star (x - x_0) \right\|_2 \le \min_{p \in \mathbb{P}_{n,0}} \|p(\mathbf{D}_2)\|_2 \left\| \mathbf{D}_2^{\frac{1}{2}} U_2^\star (x - x_0) \right\|_2 \qquad (2.14) \\
&= \min_{p \in \mathbb{P}_{n,0}} \max_{\lambda \in \Lambda(\mathsf{A}) \setminus \{0\}} |p(\lambda)| \, \|x - x_0\|_{\mathsf{A}} \, . \qquad (2.15)
\end{aligned}$$

The discrete set $\Lambda(\mathsf{A}) \setminus \{0\}$ in the min-max problem (2.15) can be replaced by the interval $[\lambda_{j+1}, \lambda_N]$ and the bound becomes

$$\|x - x_n\|_{\mathsf{A}} \le \min_{p \in \mathbb{P}_{n,0}} \max_{\lambda \in [\lambda_{j+1}, \lambda_N]} |p(\lambda)| \, \|x - x_0\|_{\mathsf{A}} \, . \qquad (2.16)$$

The probably best known convergence bound for CG results from (2.16) if the polynomial minimization is replaced by an appropriately scaled and shifted Chebyshev polynomial. The resulting bound makes use of the following definition of the *effective condition number* of an operator:

Definition 2.55 (Effective condition number)**.** Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be self-adjoint and let $\alpha(\mathsf{A}) := \inf_{\lambda \in \Lambda(\mathsf{A}) \setminus \{0\}} |\lambda| > 0$. Then $\kappa_{\text{eff}}(\mathsf{A}) := \frac{\|\mathsf{A}\|}{\alpha(\mathsf{A})}$ is the *effective condition number* of A.

Theorem 2.56 (CG κ-bound)**.** *Let the assumptions of theorem 2.54 hold and let* $\mathsf{A} \neq 0$. *Then*

$$\|x - x_n\|_{\mathsf{A}} \leq 2\left(\frac{\sqrt{\kappa_{\mathit{eff}}(\mathsf{A})} - 1}{\sqrt{\kappa_{\mathit{eff}}(\mathsf{A})} + 1}\right)^n \|x - x_0\|_{\mathsf{A}}\,. \tag{2.17}$$

Proof. First note that the linear system $\mathsf{A}x = b$ with a possibly singular operator A can be reduced to the linear system $\mathsf{B}y = b$ where $\mathsf{B} := \mathsf{A}|_{\mathcal{N}(\mathsf{A})^\perp}$ is nonsingular.

The proof for the finite-dimensional case can be found in many books on Krylov subspace methods, e.g., in [105]. The infinite-dimensional case has been handled by Daniel in [24]. □

Note that the influence of the solution x and the initial approximation x_0 has been separated from the polynomial minimization in the inequality (2.14) and all bounds derived from inequality (2.15), e.g., the κ-bound (2.17), may severely overestimate the error.

Other well-known approaches for convergence bounds try to capture the often observed "superlinear" convergence behavior of the CG method. One such approach is to choose p in (2.15) as a *composite polynomial* $p = q_l \cdot \tilde{p}$ where $q_l \in \mathbb{P}_{l,0}$ is a fixed polynomial for $l \leq n$ and to carry out the minimization for $\tilde{p} \in \mathbb{P}_{n-l,0}$, i.e.,

$$\frac{\|x - x_n\|_{\mathsf{A}}}{\|x - x_0\|_{\mathsf{A}}} \leq \min_{\tilde{p}\in\mathbb{P}_{n-l,0}} \max_{\lambda\in\{\lambda_{j+1},\ldots,\lambda_N\}} |q_l(\lambda)\tilde{p}(\lambda)|.$$

Picking $q_l(\lambda) = \prod_{i=N-l+1}^{N}\left(1 - \frac{\lambda}{\lambda_i}\right)$ yields $|q_l(\lambda_i)| \leq 1$ for $i \in \{j+1, \ldots, N-l\}$ and thus

$$\begin{aligned}\frac{\|x - x_n\|_{\mathsf{A}}}{\|x - x_0\|_{\mathsf{A}}} &\leq \max_{\lambda\in\{\lambda_{j+1},\ldots,\lambda_{N-l}\}} |q_l(\lambda)| \min_{\tilde{p}\in\mathbb{P}_{n-l,0}} \max_{\lambda\in\{\lambda_{j+1},\ldots,\lambda_{N-l}\}} |\tilde{p}(\lambda)| \\ &\leq \min_{\tilde{p}\in\mathbb{P}_{n-l,0}} \max_{\lambda\in\{\lambda_{j+1},\ldots,\lambda_{N-l}\}} |\tilde{p}(\lambda)| \leq 2\left(\frac{\sqrt{\kappa_l(\mathsf{A})} - 1}{\sqrt{\kappa_l(\mathsf{A})} + 1}\right)^{n-l}, \end{aligned} \tag{2.18}$$

where $\kappa_l(\mathsf{A}) = \frac{\lambda_{N-l}}{\lambda_{j+1}} \leq \kappa_{\mathrm{eff}}(\mathsf{A})$. Bounds of this form have been proposed by Axelsson [7], Jennings [83] and others. The bound (2.18) predicts a faster convergence rate of the CG method (in infinite precision) after l iterations if there are l eigenvalues at the upper end of A's spectrum that are well-separated from the remaining spectrum. However, as illustrated in the book by Liesen and Strakoš [105] and the article by Gergelits and Strakoš [66], bounds based on composite polynomials are questionable in the presence of round-off errors.

Preconditioned CG

In order to speed up the computation in practical applications, CG is not directly applied to the linear system $\mathsf{A}x = b$, but instead applied to a *preconditioned* linear system

$$\mathsf{MA}x = \mathsf{M}b, \tag{2.19}$$

Algorithm 2.4 Preconditioned CG algorithm, cf. [80]. Implemented as `krypy.linsys.Cg` in [60].

Input: Self-adjoint and positive-semidefinite $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, self-adjoint and positive-definite preconditioner $\mathsf{M} \in \mathcal{L}(\mathcal{H})$, right hand side $b \in \mathcal{R}(\mathsf{A})$, initial guess $x_0 \in \mathcal{H}$ and maximal number of iterations $n_{\max} \in \mathbb{N}$.

```
$r_0 = b - \mathsf{A}x_0$, $p_0 = z_0 = \mathsf{M}r_0$
for $n = 1, \dots, n_{\max}$ do
    $\alpha_{n-1} = \frac{\langle z_{n-1}, r_{n-1} \rangle}{\langle p_{n-1}, \mathsf{A}p_{n-1} \rangle}$
    $x_n = x_{n-1} + \alpha_{n-1} p_{n-1}$
    $r_n = r_{n-1} - \alpha_{n-1} \mathsf{A} p_{n-1}$
    if stopping criterion is reached then
        return $x_n$
    end if
    $z_k = \mathsf{M}r_n$
    $\omega_n = \frac{\langle z_n, r_n \rangle}{\langle z_{n-1}, r_{n-1} \rangle}$
    $p_n = z_n + \omega_n p_{n-1}$
end for
return $x_{n_{\max}}$
```

with the inner product $\langle \cdot, \cdot \rangle_{\mathsf{M}^{-1}}$ defined by $\langle x, y \rangle_{\mathsf{M}^{-1}} = \langle x, \mathsf{M}^{-1} y \rangle$. Here, $\mathsf{M} \in \mathcal{L}(\mathcal{H})$ is a suitably chosen self-adjoint and positive-definite operator, the *preconditioner*. The actual choice of the preconditioner highly depends on the problem that has to be solved.

Note that the CG method is in fact well defined when applied to the linear system (2.19), because

$$\langle x, \mathsf{MA}y \rangle_{\mathsf{M}^{-1}} = \langle x, \mathsf{A}y \rangle = \langle \mathsf{A}x, y \rangle = \langle \mathsf{MA}x, y \rangle_{\mathsf{M}^{-1}}$$

and $\langle x, \mathsf{MA}x \rangle_{\mathsf{M}^{-1}} = \langle x, \mathsf{A}x \rangle \geq 0$ hold for all $x, y \in \mathcal{H}$. Thus the linear operator MA is self-adjoint and positive (semi-)definite with respect to $\langle \cdot, \cdot \rangle_{\mathsf{M}^{-1}}$. If CG is applied to the preconditioned linear system (2.19) with the inner product $\langle \cdot, \cdot \rangle_{\mathsf{M}^{-1}}$, then

$$\langle x, y \rangle_{\mathsf{MA}} = \langle x, \mathsf{MA}y \rangle_{\mathsf{M}^{-1}} = \langle x, \mathsf{A}y \rangle$$

shows that the error is still minimized in the A-norm, i.e., the iterates x_n satisfy

$$\|x - x_n\|_{\mathsf{A}} = \min_{p \in \mathbb{P}_{n,0}} \|p(\mathsf{MA})(x - x_0)\|_{\mathsf{A}}.$$

For the preconditioned CG method, the κ-bound (2.17) becomes

$$\|x - x_n\|_{\mathsf{A}} \leq 2 \left(\frac{\sqrt{\kappa_{\text{eff}}(\mathsf{MA})} - 1}{\sqrt{\kappa_{\text{eff}}(\mathsf{MA})} + 1} \right)^n \|x - x_0\|_{\mathsf{A}}.$$

Comparing this bound with its unpreconditioned counterpart (2.17), the term *preconditioning* seems reasonable. However, it should be pointed out, that the κ-bound is not descriptive in many cases and that the condition number is not enough to estimate the convergence behavior of the CG method. The actual convergence behavior of the CG method is often superlinear while the κ-bound only predicts linear convergence, i.e., a constant factor for the reduction of the error norm.

By storing an additional vector, the preconditioned CG method can be implemented with only one application of A and one application of M per iteration, see algorithm 2.4. A derivation of this algorithm can be found in the book of Elman, Silvester and Wathen [46]. An algorithm for the CG method in a Hilbert space setting can also be found in the work of Günnel, Herzog and Sachs [77].

2.9. Minimal residual methods

In 1975, Paige and Saunders [131] introduced the minimal residual method (MINRES method) for the solution of the linear system (2.3) with a self-adjoint but possibly indefinite operator A. The method is derived from a Lanczos relation and the constructed iterates $x_n \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0)$ are chosen such that the residual norm is minimal. A decade later, Saad and Schultz [145] proposed the generalized minimal residual method (GMRES method) for a general (possibly non-self-adjoint) operator A by using an Arnoldi relation instead of a Lanczos relation. For a self-adjoint operator A, GMRES is *mathematically* equivalent to MINRES but the results of both methods can differ significantly in the presence of round-off errors, see section 2.10. The presentation in this section proceeds in reverse chronological order by first introducing the GMRES method and afterwards the special case MINRES.

2.9.1. GMRES method

In this subsection, the linear system (2.3) is again assumed to be consistent, i.e., $b \in \mathcal{R}(\mathsf{A})$. The operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, the initial guess $x_0 \in \mathcal{H}$ and the corresponding initial residual $r_0 = b - \mathsf{A}x_0$ are assumed to fulfill $d = d(\mathsf{A}, r_0) \leq \infty$ and

$$\mathcal{K}_d(\mathsf{A}, r_0) \cap \mathcal{N}(\mathsf{A}) = \{0\}. \tag{2.20}$$

The first condition holds true, e.g., if $\dim \mathcal{H} < \infty$. Unlike the situation of the CG method in section 2.8, the latter condition is not always fulfilled. However, it is obviously fulfilled if A is nonsingular. Condition (2.20) also holds if $\mathcal{R}(\mathsf{A}) \cap \mathcal{N}(\mathsf{A}) = \{0\}$ because $b \in \mathcal{R}(\mathsf{A})$. Condition (2.20) is the subject of further discussion in section 3.2.4 in the context of the deflated GMRES method.

With the above assumptions, item 2. in corollary 2.41 shows that the Petrov–Galerkin method with search space $\mathcal{S}_n = \mathcal{K}_n(\mathsf{A}, r_0)$ and constraint space $\mathcal{C}_n = \mathsf{A}\mathcal{S}_n$ is well defined. Furthermore, the iterates x_n minimize the residual norm, i.e.

$$\|b - \mathsf{A}x_n\| = \min_{z \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0)} \|b - \mathsf{A}z\|, \tag{2.21}$$

and x_d is a solution of the linear system (2.3).

For a general $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, the Arnoldi algorithm can be used with the initial vector r_0. For $n < d$, a $V_n + 1 \in \mathcal{H}^{n+1}$ and a $\underline{\mathbf{H}}_n = \begin{bmatrix} \mathbf{T}_n \\ h_{n+1,n} e_n^\mathsf{T} \end{bmatrix} \in \mathbb{C}^{n+1,n}$ are generated that define an Arnoldi relation for $\mathcal{K}_n(\mathsf{A}, r_0)$. In the d-th step, the generated $V_d \in \mathcal{H}^d$ and $\mathcal{H}_d \in \mathbb{C}^{d,d}$ define an invariant Arnoldi relation for $\mathcal{K}_d(\mathsf{A}, r_0)$.

The Petrov–Galerkin method with search space $\mathcal{S}_n = \mathcal{K}_n(\mathsf{A}, r_0)$ and constraint space $\mathcal{C}_n = \mathsf{A}\mathcal{S}_n$ is well defined for $n \leq d$ and the above assumptions. Because $x_n = x_0 + V_n y_n$ for a $y_n \in \mathbb{C}^n$, the residual norm becomes for $n < d$:

$$\|b - \mathsf{A}x_n\| = \|V_{n+1}(e_1 \|r_0\| - \underline{\mathbf{H}}_n y_n)\| = \|e_1 \|r_0\| - \underline{\mathbf{H}}_n y_n\|_2 .$$

The optimality property (2.21) then reads

$$\|e_1 \|r_0\| - \underline{\mathbf{H}}_n y_n\|_2 = \min_{z \in \mathbb{C}^n} \|e_1 \|r_0\| - \underline{\mathbf{H}}_n z\|_2 . \tag{2.22}$$

In the GMRES algorithm, this least squares problem is solved by maintaining a QR-factorization of $\underline{\mathbf{H}}_n = \mathbf{Q}_n \begin{bmatrix} \mathbf{R}_n \\ 0 \end{bmatrix}$ with $\mathbf{Q}_n^\mathsf{H} \mathbf{Q}_n = \mathbf{I}_{n+1}$ and $\mathbf{R}_n \in \mathbb{C}^{n,n}$ upper triangular. The factors $\mathbf{Q}_n$ and $\mathbf{R}_n$ can be updated from the factors $\mathbf{Q}_{n-1}$ and $\mathbf{R}_{n-1}$ from the last Arnoldi step by using Givens rotations. By theorem 2.25, the matrix

$$\begin{aligned} \langle \mathsf{A}V_n, \mathsf{A}V_n \rangle = \langle V_{n+1}\underline{\mathbf{H}}_n, V_{n+1}\underline{\mathbf{H}}_n \rangle = \underline{\mathbf{H}}_n^\mathsf{H} \underline{\mathbf{H}}_n &= \begin{bmatrix} \mathbf{R}_n \\ 0 \end{bmatrix}^\mathsf{H} \mathbf{Q}_n^\mathsf{H} \mathbf{Q}_n \begin{bmatrix} \mathbf{R}_n \\ 0 \end{bmatrix} \\ &= \mathbf{R}_n^\mathsf{H} \mathbf{R}_n \end{aligned}$$

is nonsingular and thus also $\mathbf{R}_n$ is nonsingular. The residual norm

$$\left\| e_1 \|r_0\| - \mathbf{Q}_n \begin{bmatrix} \mathbf{R}_n \\ 0 \end{bmatrix} y_n \right\|_2 = \left\| \mathbf{Q}_n^\mathsf{H} e_1 \|r_0\| - \begin{bmatrix} \mathbf{R}_n \\ 0 \end{bmatrix} y_n \right\|_2$$

is thus minimized by $y_n = \mathbf{R}^{-1} u_n$, where $[u_n, \eta_n]^\mathsf{T} = \mathbf{Q}_n^\mathsf{H} e_1 \|r_0\|$ with $u_n \in \mathbb{C}^n$ and $\eta_n \in \mathbb{C}$. The n-th GMRES approximation then is given

$$x_n = x_0 + V_n \mathbf{R}_n^{-1} u_n$$

and the residual norm is

$$\|b - \mathsf{A}x_n\| = |\eta_n|.$$

A generic version of the GMRES algorithm is given in algorithm 2.5. In line 3 of this algorithm, the Arnoldi relation can be constructed with the modified Gram–Schmidt or Householder Arnoldi algorithms, cf. section 2.7. The computation of the QR factorization in line 7 can be computed iteratively by applying n Givens rotations in the n-th iteration, cf. [145].

Algorithm 2.5 GMRES algorithm (generic version; cf. [145]). Implemented as `krypy.linsys.Gmres` in [60].

```
Input: $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, right hand side $b \in \mathcal{R}(\mathsf{A})$ and initial guess $x_0 \in \mathcal{H}$ such that
    condition (2.20) is fulfilled. Maximal number of iterations $n_{\max} \in \mathbb{N}$.
$r_0 = b - \mathsf{A}x_0$
for $n = 1, \ldots, n_{\max}$ do
    Generate Arnoldi relation for $\mathcal{K}_n(\mathsf{A}, r_0)$: either
      • $V_{n+1} \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}_n \in \mathbb{C}^{n+1,n}$ or
      • $V_n \in \mathcal{H}^n$ and $\underline{\mathbf{H}}_n \in \mathbb{C}^{n,n}$ if $\mathcal{K}_n(\mathsf{A}, r_0)$ is $\mathsf{A}$-invariant.
    if $\mathcal{K}_n(\mathsf{A}, r_0)$ is $\mathsf{A}$-invariant then
        return $x_n = x_0 + V_n \mathbf{H}_n^{-1} e_1 \|r_0\|$
    else
        Compute QR factorization of $\underline{\mathbf{H}}_n = \mathbf{Q}_n \begin{bmatrix} \mathbf{R}_n \\ 0 \end{bmatrix}$, $\mathbf{Q}_n \in \mathbb{C}^{n+1,n+1}$, $\mathbf{R}_n \in \mathbb{C}^{n,n}$.
        $[u_n, \eta_n]^\mathsf{T} = \mathbf{Q}_n^\mathsf{H} e_1 \|r_0\|$ with $u_n \in \mathbb{C}^n$, $\eta_n \in \mathbb{C}$.
        Form $x_n = x_0 + \mathbf{R}_n^{-1} u_n$ if necessary (note that $\|b - \mathsf{A}x_n\| = |\eta_n|$).
        if stopping criterion is reached then
            return $x_n$
        end if
    end if
end for
return $x_{n_{\max}}$
```

Convergence of GMRES

In this subsection, some well-known convergence properties of the GMRES method are stated for the finite-dimensional case, i.e., $N := \dim \mathcal{H} < \infty$. Compared to the analysis of the convergence behavior of the CG method for a self-adjoint and positive-definite operator A, the analysis becomes more intricate in the situation of the GMRES method with a general linear operator A. The most important properties of the GMRES method are gathered in the following theorem.

Theorem 2.57 (Convergence of GMRES)**.** *Let $N := \dim \mathcal{H} < \infty$ and let $\mathsf{A}x = b$ be a consistent linear system with $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $b \in \mathcal{H}$. Furthermore, let $x_0 \in \mathcal{H}$ be an initial guess such that condition* (2.20) *is fulfilled with $d = d(\mathsf{A}, r_0)$.*

Then the GMRES method (see algorithm 2.5) is well defined and the following holds:

1. *The residual norm in iteration $n \leq d$ satisfies*

$$\|b - \mathsf{A}x_n\| = \min_{z \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0)} \|b - \mathsf{A}z\| = \min_{p \in \mathbb{P}_{n,0}} \|p(\mathsf{A}) r_0\| . \tag{2.23}$$

2. *The residual in iteration $n \leq d$ is given by*

$$b - \mathsf{A}x_n = p_n^{MR}(\mathsf{A}) r_0$$

with $p_n^{MR} \in \mathbb{P}_{n,0}$ *and*

$$p_n^{MR}(\lambda) = \prod_{i \in J_n} \left(1 - \frac{\lambda}{\theta_i^{(n)}}\right),$$

where $\theta_1^{(n)}, \ldots, \theta_n^{(n)} \in \mathbb{C} \cup \{\infty\}$ *are the harmonic Ritz values of* A *with respect to* $\mathcal{K}_n(\mathsf{A}, r_0)$ *and* $J_n = \{i \in \{1, \ldots, n\} \mid \theta_i^{(n)} \neq \infty\}$ *(cf. section 2.5).*

3. *The following statements are equivalent for an iteration* $1 \leq n \leq d$*:*
   a) *GMRES stagnates at iteration n, i.e.,* $x_n = x_{n-1}$*.*
   b) $\mathbf{H}_n = \langle V_n, \mathsf{A} V_n \rangle$ *is singular.*
   c) *There exists an infinite harmonic Ritz value of* A *with respect to the subspace* $\mathcal{K}_n(\mathsf{A}, r_0)$*, i.e.,* $\infty \in \{\theta_1^{(n)}, \ldots, \theta_n^{(n)}\}$*.*
   d) *The minimal residual polynomial* $p_n^{MR} \in \mathbb{P}_{n,0}$ *does not have full degree, i.e.,* $\deg p_n^{MR} < n$*.*

Proof. The well-definedness of the method and item 1. are analogous to theorem 2.54. A proof of item 2. was given by Freund in [57]. The equivalence of 3.a) and 3.b) was shown by Brown in [17]. The equivalence of 3.b) and 3.c) follows from the fact that the existence of an infinite harmonic Ritz value is equivalent to the existence of a zero eigenvalue of $(\underline{\mathbf{H}}_n^{\mathsf{H}} \underline{\mathbf{H}}_n)^{-1} \mathbf{H}_n^{\mathsf{H}}$, cf. section 2.5. The last equivalence in item 3. immediately follows from item 2. □

A meaningful a priori characterization of the convergence behavior of the GMRES method for general linear operators is even harder than in the case of the CG method. The remaining part of this subsection gives a brief overview of common approaches for describing the convergence behavior of the GMRES method.

Theorem 2.58 (GMRES spectral bound)**.** *Let* $N := \dim \mathcal{H} < \infty$ *and let* $\mathsf{A}x = b$ *be a consistent linear system with* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *and* $b \in \mathcal{H}$*. Furthermore, let* $\mathsf{A} = S\mathbf{J}S^{-1}$ *be a Jordan decomposition with* $S \in \mathcal{H}^N$ *and* $\mathbf{J} = \operatorname{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_k)$*, where* $\mathbf{J}_i = \mathbf{J}_i(\lambda_i) \in \mathbb{C}^{N_i, N_i}$ *are the Jordan blocks for* $i \in \{1, \ldots, k\}$*. The eigenvalues are assumed to be ordered such that* $\lambda_1, \ldots, \lambda_l \neq 0$ *and* $\lambda_{l+1} = \ldots = \lambda_k = 0$ *and* $m := \max\{N_{l+1}, \ldots, N_k\}$ *denotes the index of the zero eigenvalue. Furthermore, assume that* $x_0 \in \mathcal{H}$ *is an initial guess such that* $0 \neq r_0 = b - \mathsf{A}x_0 \in \mathcal{R}(\mathsf{A}^m)$*.*

Then for n $\leq d$*, the residuals* r_n *of the GMRES method satisfy*

$$\frac{\|r_n\|}{\|r_0\|} \leq \kappa(S) \min_{p \in \mathbb{P}_{n,0}} \max_{i \in \{1, \ldots, l\}} \|p(\mathbf{J}_i)\|_2 . \tag{2.24}$$

Proof. The proof for the case where A is nonsingular has been given, e.g., by Jia in [84, theorem 1]. Here, the proof is given with a minor modification that takes into account the possibility of a singular operator A. If $r_0 \in \mathcal{R}(\mathsf{A}^m)$, then $r_0 = S \begin{bmatrix} \mathbf{I}_M & \\ & 0 \end{bmatrix} S^{-1} r_0$, where $M = \sum_{i=1}^{l} N_i$. Thus the following holds for a $p \in \mathbb{P}_n$

$$p(\mathsf{A}) r_0 = S p(\mathbf{J}) S^{-1} r_0 = S \operatorname{diag}(p(\mathbf{J}_1), \ldots, p(\mathbf{J}_l), 0_{N-M}) S^{-1} r_0$$

and with equation (2.23) therefore

$$\|r_n\| = \min_{p\in\mathbb{P}_{n,0}} \|p(\mathsf{A})r_0\| \leq \|S\| \left\|S^{-1}\right\| \min_{p\in\mathbb{P}_{n,0}} \max_{i\in\{1,\ldots,l\}} \|p(\mathbf{J}_i)\|_2 \|r_0\| .$$

□

For a diagonalizable matrix $\mathsf{A} = S\mathbf{D}S^{-1} \in \mathbb{C}^{N,N}$ with $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, equation (2.24) reads

$$\frac{\|r_n\|}{\|r_0\|} \leq \kappa(S) \min_{p\in\mathbb{P}_{n,0}} \max_{i\in\{1,\ldots,l\}} |p(\lambda_i)|, \tag{2.25}$$

where the eigenvalues are ordered and the index l is defined as in theorem 2.58. Because of this bound it is tempting to say that the convergence behavior of the GMRES method is determined by the eigenvalues of the operator A. However, this would be as misleading as calling the κ-bound (2.17) a descriptive bound for the CG method. Note that the bound (2.25) exhibits the condition number of the eigenvector matrix S which may render the bound useless in applications.

Greenbaum, Pták and Strakoš showed in [74] that the spectrum alone cannot describe the convergence behavior for the GMRES method for a general linear operator A. Their analysis shows that if any set $L \subset \mathbb{C} \setminus \{0\}$ with $|L| \leq N$ and any N numbers $\eta_0 \geq \eta_1 \geq \cdots \geq \eta_{N-1} > \eta_N = 0$ are given, then there exists a matrix $\mathsf{A} \in \mathbb{C}^{N,N}$ with $\Lambda(\mathsf{A}) = L$ and a right hand side $b \in \mathcal{H}$ such that GMRES applied to $\mathsf{A}x = b$ with $x_0 = 0$ constructs residuals r_n that satisfy $\|r_n\| = \eta_n$ for $n \in \{0, \ldots, N\}$. A parametrization of *all* matrices $\mathsf{A} \in \mathbb{C}^{N,N}$ and right hand sides $b \in \mathcal{H}$ with the above property was provided by Arioli, Pták and Strakoš in [5].

Trefethen took a different path in [175] by using the *ϵ-pseudospectrum* $\Lambda_\epsilon(\mathsf{A})$ (which was called the *set of ϵ-approximate eigenvalues of* A in [175]).

Definition 2.59 (Pseudospectrum)**.** For $\dim\mathcal{H} < \infty$, $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $\epsilon > 0$, the *ϵ-pseudospectrum* of A is defined by

$$\begin{aligned}\Lambda_\epsilon(\mathsf{A}) &= \bigcup_{\substack{\mathsf{E}\in\mathcal{L}(\mathcal{H})\\ \|\mathsf{E}\|<\epsilon}} \Lambda(\mathsf{A}+\mathsf{E})\\ &= \left\{\lambda \in \mathbb{C} \;\middle|\; \left\|(\lambda\mathsf{I} - \mathsf{A})^{-1}\right\| > \epsilon^{-1}\right\}\\ &= \left\{\lambda \in \mathbb{C} \;\middle|\; \|(\lambda\mathsf{I} - \mathsf{A})v\| < \epsilon \text{ for a } v \in \mathcal{H}\right\},\end{aligned}$$

where $\left\|(\lambda\mathsf{I} - \mathsf{A})^{-1}\right\| := \infty$ if the resolvent $(\lambda\mathsf{I} - \mathsf{A})^{-1}$ does not exist, i.e., if $\lambda \in \Lambda(\mathsf{A})$.

A detailed treatment of pseudospectra including a proof of the equivalence of the above definitions can be found in the monograph of Trefethen and Embree [177]. For $\epsilon > 0$, the spectrum is contained in the ϵ-pseudospectrum, i.e., $\Lambda(\mathsf{A}) \subset \Lambda_\epsilon(\mathsf{A})$. Note that the pseudospectrum is defined as an open set in definition 2.59. If the boundary of $\Lambda_\epsilon(\mathsf{A})$ is denoted by $\partial\Lambda_\epsilon(\mathsf{A})$ and $p \in \mathbb{P}_n$ is a polynomial, then $p(\mathsf{A})$ can be expressed as a Cauchy integral along the curve $\partial\Lambda_\epsilon(\mathsf{A})$:

$$p(\mathsf{A}) = \frac{1}{2\pi\mathbb{i}} \int_{\partial\Lambda_\epsilon(\mathsf{A})} p(\lambda)(\lambda\mathsf{I} - \mathsf{A})^{-1} d\lambda.$$

This equation is used to prove the following theorem, see Nachtigal, Reddy and Trefethen [126].

Theorem 2.60 (GMRES pseudospectral bound)**.** *Let the assumptions of theorem 2.58 hold and let $\epsilon > 0$.*

Then for $n \leq d$, the residuals of the GMRES method satisfy

$$\frac{\|r_n\|}{\|r_0\|} \leq \frac{|\partial\Lambda_\epsilon(\mathsf{A}_1)|}{2\pi\epsilon} \min_{p\in\mathbb{P}_{n,0}} \sup_{\lambda\in\Lambda_\epsilon(\mathsf{A}_1)} |p(\lambda)|,$$

where $\mathsf{A}_1 := \mathsf{A}|_{\mathcal{R}(\mathsf{A}^m)}$, i.e. $\mathsf{A}_1 = S\operatorname{diag}(\mathbf{J}_1,\ldots,\mathbf{J}_l,0,\ldots,0)S^{-1}|_{\mathcal{R}(\mathsf{A}^m)}$, and $|\partial\Lambda_\epsilon(\mathsf{A}_1)|$ denotes the curve's arc length.

Proof. If $r_0 \in \mathcal{R}(\mathsf{A}^m)$, then

$$p(\mathsf{A})r_0 = p(\mathsf{A}_1)r_0 = \frac{1}{2\pi\mathbb{i}} \int_{\partial\Lambda_\epsilon(\mathsf{A}_1)} p(\lambda)(\lambda\mathsf{I} - \mathsf{A}_1)^{-1} d\lambda\ r_0$$

holds for any $p \in \mathbb{P}_n$. Estimating the norm yields

$$\begin{aligned}\|p(\mathsf{A})r_0\| &\leq \frac{1}{2\pi} \int_{\partial\Lambda_\epsilon(\mathsf{A}_1)} |p(\lambda)| \left\|(\lambda\mathsf{I} - \mathsf{A}_1)^{-1}\right\| d\lambda\, \|r_0\| = \frac{1}{2\pi\epsilon} \int_{\partial\Lambda_\epsilon(\mathsf{A}_1)} |p(\lambda)| d\lambda\, \|r_0\| \\ &\leq \frac{|\partial\Lambda_\epsilon(\mathsf{A}_1)|}{2\pi\epsilon} \max_{\lambda\in\partial\Lambda_\epsilon(\mathsf{A}_1)} |p(\lambda)|\, \|r_0\| = \frac{|\partial\Lambda_\epsilon(\mathsf{A}_1)|}{2\pi\epsilon} \sup_{\lambda\in\Lambda_\epsilon(\mathsf{A}_1)} |p(\lambda)|\, \|r_0\|\,,\end{aligned}$$

where the last equality is due to the fact that p is holomorphic and its maximum is attained on the boundary of the pseudospectrum $\Lambda_\epsilon(\mathsf{A}_1)$. The proof is complete with the minimization property of GMRES, i.e., equation (2.23). □

Theorem 2.60 does not answer the question of an appropriate choice of ϵ and this question turns out to be tricky in practice. On the one hand, ϵ should be chosen large to make the factor $\frac{1}{\epsilon}$ small, but on the other hand it has to be chosen small enough such that the pseudospectrum $\Lambda_\epsilon(\mathsf{A}_1)$ and its boundary are not too large. In order to obtain a relevant bound, the pseudospectrum $\Lambda_\epsilon(\mathsf{A}_1)$ has to exclude the origin because $p(0) = 1$ for $p \in \mathbb{P}_{n,0}$.

Note that theorem 2.60 is usually stated without taking special care of the zero eigenvalues [175, 126, 177]. For a singular operator A, the naive way of taking the boundary of $\Lambda_\epsilon(\mathsf{A})$ as the curve for the Cauchy integral does not provide a usable bound because the origin is included. The restriction of the operator to the invariant subspace $\mathcal{R}(\mathsf{A}^m)$ that corresponds to the nonzero eigenvalues can be used because r_0 is also contained in this subspace by assumption. Note that $r_0 \in \mathcal{R}(\mathsf{A}^m)$ if and only if a solution can be found in a Krylov subspace, see proposition 2.39. Another detail in theorem 2.60 should be made clear here because it might be puzzling at first sight. If the linear operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ is singular, then the operator A_1 in theorem 2.60 has to be treated as the *nonsingular* operator $\mathsf{A}_1 \in \mathcal{L}(\mathcal{G})$, where

$\mathcal{G} = \mathcal{R}(\mathsf{A}^m)$. Thus, the perturbations E in the first equation of the definition of the pseudospectrum (cf. definition 2.59) also have to be elements of $\mathcal{L}(\mathcal{G})$, i.e.,

$$\Lambda_\epsilon(\mathsf{A}_1) = \bigcup_{\substack{\mathsf{E}\in\mathcal{L}(\mathcal{G}) \\ \|\mathsf{E}\|<\epsilon}} \Lambda(\mathsf{A}+\mathsf{E}).$$

The pseudospectral approach reappears in the analysis of perturbed Krylov subspace methods in section 3.3.3.

In [161, 162], Starke generalized and improved a convergence bound for the GCR method of Elman [45] for the GMRES method which is based on the field of values:

Definition 2.61 (Field of values)**.** The *field of values* of $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ is defined by

$$W(\mathsf{A}) \coloneqq \{\langle v, \mathsf{A}v\rangle \mid v \in \mathcal{H},\ \|v\| = 1\}.$$

The distance of $W(\mathsf{A})$ to the origin is denoted by

$$\nu(\mathsf{A}) \coloneqq \inf\left\{|w| \;\middle|\; w \in W(\mathsf{A})\right\}.$$

The field of value bound of Starke in [161, 162] is restricted to matrices with positive-definite symmetric part but it has been shown by Eiermann and Ernst [42] that the bound actually holds for any nonsingular matrix. Here, their result is stated for a possibly singular operator:

Theorem 2.62 (GMRES field of value bound)**.** *Let the assumptions of theorem 2.58 hold.*

Then for $n \leq d$, the residuals of the GMRES method satisfy

$$\frac{\|r_n\|}{\|r_0\|} \leq \left(1 - \nu(\mathsf{A}_1)\nu(\mathsf{A}_1^{-1})\right)^{\frac{n}{2}}, \tag{2.26}$$

where $\mathsf{A}_1 \coloneqq \mathsf{A}|_{\mathcal{R}(\mathsf{A}^m)}$ as in theorem 2.60.

Proof. The proof is given in [42, theorem 6.1 and corollary 6.2] for a nonsingular operator A. Analogous to the proof of theorem 2.60, the result can be generalized to a singular operator because $r_0 \in \mathcal{R}(\mathsf{A}^m)$. □

Note that the bound (2.26) is useless if $0 \in W(\mathsf{A}_1)$, e.g., if A is indefinite. Starke showed in [162] and with Klawonn in [96], that the field of values can be used to obtain useful GMRES convergence bounds for preconditioned non-symmetric elliptic problems and saddle point problems. Benzi and Olshanskii [12] were able to mathematically justify the already observed convergence behavior of a preconditioned GMRES method for the Navier–Stokes problem. Furthermore, Liesen and Tichý showed in [106] that the bound in theorem 2.62 not only bounds the worst-case GMRES residual norm but also the ideal GMRES approximation, i.e.,

$$\frac{\|r_n\|}{\|r_0\|} \leq \underbrace{\max_{0\neq v\in\mathcal{H}} \min_{p\in\mathbb{P}_{n,0}} \frac{\|p(\mathsf{A})v\|}{\|v\|}}_{\text{worst-case GMRES}} \leq \underbrace{\min_{p\in\mathbb{P}_{n,0}} \|p(\mathsf{A})\|}_{\text{ideal GMRES}} \leq \left(1 - \nu(\mathsf{A})\nu(\mathsf{A}^{-1})\right)^{\frac{n}{2}}.$$

While the worst-case GMRES bound is attainable for at least one right hand side and initial guess, this does not need to be true for the ideal GMRES bound; see also the discussion in section 5.7.3 in the book of Liesen and Strakoš [105]. The term *ideal GMRES* has been introduced by Greenbaum and Trefethen in [75]. More results on the field of values in the context of Krylov subspace methods can be found in [41, 162, 96, 52, 42, 12].

Preconditioned GMRES

Instead of applying GMRES to the linear system $\mathsf{A}x = b$, it is in practice applied to the linear system

$$\mathsf{MA}x = \mathsf{M}b \tag{2.27}$$

$$\text{or} \qquad \mathsf{AM}y = b \quad \text{with} \quad x = \mathsf{M}y, \tag{2.28}$$

where $\mathsf{M} \in \mathcal{L}(\mathcal{H})$ is invertible and chosen such that convergence is faster than when applied to the original linear system. Probably because the κ-bound (2.17) for the CG method is so wide-spread, the term *preconditioner* was established for the operator M also in the case of the GMRES method for a general operator A. Note that the term *preconditioner* is highly misleading because the behavior of the GMRES method is usually not governed by the condition number of the operator. However, because of the wide-spread use, M is also called a preconditioner here and equations (2.27)–(2.28) are called the left and right preconditioned linear system. Unlike the CG method, the preconditioner can be any invertible operator in the GMRES method and the inner product is not required to be adapted to the preconditioner.

2.9.2. MINRES method

Let the operator A of the linear system (2.3) be self-adjoint in this subsection. In this case, the Arnoldi algorithm in line 3 of the GMRES algorithm 2.5 can be replaced by the Lanczos algorithm. Paige and Saunders showed in [131], that not only the Lanczos algorithm benefits from a three-term recurrence, but also that the least squares problem (2.22) can be solved by a three-term recurrence. The resulting MINRES algorithm 2.6 makes use of Givens rotations in order to maintain a QR factorization of the Hermitian tridiagonal matrix that is produced by the Lanczos algorithm, cf. Elman, Silvester and Wathen [46] and Fischer [54].

Because MINRES is mathematically equivalent to GMRES, all statements from theorem 2.57 and the convergence bounds for GMRES are also valid for MINRES. Note that condition (2.20) is always fulfilled for a self-adjoint operator A, see section 2.9.1. Furthermore, if A is self-adjoint and $N := \dim \mathcal{H} < \infty$, then a bound of the form of the κ-bound (2.17) for the CG method can be derived.

Algorithm 2.6 MINRES algorithm (based on algorithm 2.4 in [46], see also the preconditioned version in algorithm 2.7)

```
Input: Self-adjoint $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, right hand side $b \in \mathcal{R}(\mathsf{A})$ and initial guess $x_0 \in \mathcal{H}$.
    Maximal number of iterations $n_{\max} \in \mathbb{N}$.
$v_0 = w_0 = w_1 = 0$
$s_0 = s_1 = 0$, $c_0 = c_1 = 1$
$r_0 = b - \mathsf{A}x_0$, $\eta_0 = \|r_0\|$, $v_1 = \frac{r_0}{\eta_0}$
for $n = 1, \ldots, n_{\max}$ do
    $z \leftarrow \mathsf{A}v_n$
    $z \leftarrow z - \delta_n v_{n-1}$
    $\gamma_n = \langle v_n, z \rangle$
    $z \leftarrow z - \gamma_n v_n$
    $\delta_{n+1} = \|z\|$
    $\alpha_0 = c_n \gamma_n - c_{n-1} s_n \delta_n$
    $\alpha_1 = \sqrt{\alpha_0^2 + \delta_{n+1}^2}$
    $\alpha_2 = s_n \gamma_n + c_{n-1} c_n \delta_n$
    $\alpha_3 = s_{n-1} \delta_n$
    $c_{n+1} = \alpha_0 / \alpha_1$
    $s_{n+1} = \delta_{n+1} / \alpha_1$
    $w_{n+1} = \frac{1}{\alpha_1}(v_n - \alpha_3 w_{n-1} - \alpha_2 w_n)$
    $x_n = x_{n-1} + c_{n+1} \eta_{n-1} w_{n+1}$
    $\eta_n = -s_{n+1} \eta_{n-1}$
    if stopping criterion is reached (note that $\|b - \mathsf{A}x_n\| = |\eta_n|$) then
        return $x_n$
    end if
    $v_{n+1} = \frac{z}{\delta_{n+1}}$
end for
return $x_{n_{\max}}$
```

Theorem 2.63. *Let $N := \dim \mathcal{H} < \infty$ and let $\mathsf{A}x = b$ be a consistent linear system with a self-adjoint and indefinite operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $b \in \mathcal{H}$. Furthermore, let the eigenvalues of A be sorted such that $\lambda_1 \leq \ldots \leq \lambda_s < 0 = \lambda_{s+1} = \ldots = \lambda_{t-1} < \lambda_t \leq \ldots \leq \lambda_N$ holds. Assume that an initial guess $x_0 \in \mathcal{H}$ is given and the corresponding initial residual $r_0 = b - \mathsf{A}x_0$ is of grade $d = d(\mathsf{A}, r_0)$.*

Then for $n \leq d$, the residuals r_n of the MINRES/GMRES method satisfy

$$\frac{\|r_n\|}{\|r_0\|} \leq 2 \left(\frac{\sqrt{|\lambda_1 \lambda_N|} - \sqrt{|\lambda_s \lambda_t|}}{\sqrt{|\lambda_1 \lambda_N|} + \sqrt{|\lambda_s \lambda_t|}} \right)^{\left[\frac{n}{2}\right]}, \tag{2.29}$$

where $\left[\frac{n}{2}\right]$ is the integer part of $\frac{n}{2}$.

Proof. The proof was given in the monograph of Greenbaum [73] for a nonsingular operator. Analogous to the previous proofs for the GMRES method, it can be generalized to the singular case by using the restriction $\mathsf{A}|_{\mathcal{R}(\mathsf{A})}$. Note that the index m of the zero eigenvalue of A is at the maximum 1 because A is self-adjoint. Therefore, the condition $b \in \mathcal{R}(\mathsf{A})$ implies $r_0 \in \mathcal{R}(\mathsf{A})$ and guarantees that a solution can be found in a Krylov subspace. □

In order to speed up the convergence of the MINRES method, a preconditioner can be applied. However, the preconditioner is subject to the same restrictions as in the preconditioned CG method: it has to be self-adjoint and positive definite and the inner product is also changed by the choice of the preconditioner. A variant of the preconditioned MINRES algorithm is given in algorithm 2.7, cf. [46].

Algorithm 2.7 Preconditioned MINRES algorithm (based on algorithm 6.1 in [46]). Implemented as `krypy.linsys.Minres` in [60].

```
Input: Self-adjoint $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, self-adjoint and positive-definite preconditioner $\mathsf{M} \in \mathcal{L}(\mathcal{H})$, right hand side $b \in \mathcal{R}(\mathsf{A})$ and initial guess $x_0 \in \mathcal{H}$. Maximal number of iterations $n_{\max} \in \mathbb{N}$.
$v_0 = w_0 = w_1 = 0$
$s_0 = s_1 = 0$, $c_0 = c_1 = 1$
$z \leftarrow b - \mathsf{A}x_0$, $r_0 = \mathsf{M}z$, $\eta_0 = \sqrt{\langle z, r_0 \rangle}$, $z_1 = \frac{z}{\eta_0}$, $v_1 = \frac{r_0}{\eta_0}$
for $n = 1, \ldots, n_{\max}$ do
    $z \leftarrow \mathsf{A}v_n$
    $z \leftarrow z - \delta_n z_{n-1}$
    $\gamma_n = \langle v_n, z \rangle$
    $z \leftarrow z - \gamma_n z_n$
    $v \leftarrow \mathsf{M}z$
    $\delta_{n+1} = \sqrt{\langle v, z \rangle}$
    $\alpha_0 = c_n\gamma_n - c_{n-1}s_n\delta_n$
    $\alpha_1 = \sqrt{\alpha_0^2 + \delta_{n+1}^2}$
    $\alpha_2 = s_n\gamma_n + c_{n-1}c_n\delta_n$
    $\alpha_3 = s_{n-1}\delta_n$
    $c_{n+1} = \alpha_0/\alpha_1$
    $s_{n+1} = \delta_{n+1}/\alpha_1$
    $w_{n+1} = \frac{1}{\alpha_1}(v_n - \alpha_3 w_{n-1} - \alpha_2 w_n)$
    $x_n = x_{n-1} + c_{n+1}\eta_{n-1}w_{n+1}$
    $\eta_n = -s_{n+1}\eta_{n-1}$
    if stopping criterion is reached (note that $\|b - \mathsf{A}x_n\| = |\eta_n|$) then
        return $x_n$
    end if
    $z_{n+1} = \frac{z}{\delta_{n+1}}$
    $v_{n+1} = \frac{v}{\delta_{n+1}}$
end for
return $x_{n_{\max}}$
```

2.10. Round-off errors

Round-off errors unavoidably affect the results of all variants of the Arnoldi and Lanczos algorithms and consequently affect all methods that rely on these algorithms. Observable effects of round-off errors in Krylov subspace methods for linear systems are, e.g., a limited maximal attainable accuracy way above machine precision, a delay of convergence or even misconvergence. In general, the round-off properties heavily depend on the actual algorithm that is used. This section gives a brief overview on the difficulties that round-off errors can pose and states some well-known results for $N := \dim \mathcal{H} < \infty$ and the Euclidean case, i.e., $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_2$. Throughout this section, the unit round-off is denoted by ε and C denotes a positive constant that is independent of the operator A, the initial vector v, the dimension N and the Arnoldi/Lanczos step n. Two questions regarding the *computed* Arnoldi or Lanczos basis V_{n+1} and Hessenberg matrix $\underline{\mathbf{H}}$ are of major interest:

- To what extent is the Arnoldi recurrence fulfilled, e.g., how large is the Arnoldi residual $\eta := \|\mathsf{A}V_n - V_{n+1}\underline{\mathbf{H}}\|$?
- How far is V_n from orthonormality, e.g., how large is $\zeta := \|\mathbf{I}_n - \langle V_n, V_n \rangle\|_2$?

Two variants of the Arnoldi algorithm are commonly used: the modified Gram–Schmidt variant (cf. algorithm 2.1) and the Householder variant [180, 147]. In the Euclidean case, i.e., $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_2$, both variants fulfill [36, inequality (2.3)]

$$\eta \le CnN^{\frac{3}{2}}\varepsilon \|\mathsf{A}\| .$$

According to inequality (2.5) in [36], the loss of orthogonality in the modified Gram–Schmidt Arnoldi algorithm can be bounded by

$$\zeta \le Cn^2 N\varepsilon\kappa([v_1, \mathsf{A}V_n])$$

if $nN^2\varepsilon\kappa([v_1, \mathsf{A}V_n]) \ll 1$. Here, $\kappa(X)$ is the condition number of a full rank $X \in \mathcal{H}^l$, which is defined as the ratio of its largest singular value $\sigma_1(X)$ and smallest singular value $\sigma_l(X)$, i.e. $\kappa(X) := \frac{\sigma_1(X)}{\sigma_l(X)}$.

The loss of orthogonality in the Householder Arnoldi algorithm can be bounded independently of the operator A (cf. inequality 2.4 in [36]):

$$\zeta \le Cn^{\frac{3}{2}}N\varepsilon.$$

The operator-independent orthogonality makes the Householder variant very attractive. A forward error analysis of the GMRES algorithm with the Householder Arnoldi algorithm can be found in the work of Arioli and Fassino [2].

For GMRES with modified Gram–Schmidt orthogonalization, it was shown by Greenbaum, Rozložník and Strakoš in [72] that a severe loss of orthogonality occurs only after the residual norm has almost reached the maximum attainable accuracy. In [134], Paige, Rozložník and Strakoš showed that GMRES with modified Gram–Schmidt orthogonalization computes a backward stable solution for linear systems

$\mathsf{A}x = b$ with "sufficiently nonsingular A". These works justified the widespread use of the modified Gram–Schmidt orthogonalization in GMRES for a wide range of problems.

In cases where the modified Gram–Schmidt Arnoldi algorithm suffers from severe loss of orthogonality and the Householder method is not applicable, the orthogonalization can be iterated, i.e., the block of the *for*-loop in lines 4–7 of algorithm 2.1 is repeated. Regarding the number of required reorthogonalizations, Kahan coined the phrase "twice is enough", which was made precise in later articles, see [69, 70]. It should be noted, that one additional iteration of orthogonalization also cures the poor round-off properties of the classical Gram–Schmidt variant if the initial set of vectors are numerically nonsingular, see [70]. The classical Gram–Schmidt variant is not considered here but can be beneficial in parallel algorithms. The (iterated) Gram–Schmidt and Householder variants are implemented in [60] as `krypy.utils.Arnoldi`.

In the Lanczos algorithm, the sensitivity to round-off errors is even worse due to the fact that the orthogonalization is only performed against the last two Lanczos vectors. The consequences of round-off errors in the Lanczos algorithm have already been observed and studied in early works on Krylov subspace methods, e.g., by Lanczos [102, 103], Hestenes and Stiefel [80] and Wilkinson [183]. By this time, in order to overcome the gradual loss of orthogonality, a common strategy was to reorthogonalize a new Lanczos vector against all previously computed Lanczos vectors, thus essentially carrying out a full Gram–Schmidt orthogonalization. With regard to this strategy, Parlett writes from the perspective of 1994 in [137]:

> It is not disrespectful to say that Lanczos himself, and J. H. Wilkinson, the leading expert in matrix computations from 1960–1984, both panicked at this phenomenon. Each of them insisted on doing what we now call full reorthogonalization.

In the meantime, Paige showed in his PhD thesis [133] from 1971, that the loss of orthogonality of the Lanczos basis coincides with the convergence of certain Ritz values and Ritz vectors (see section 2.5 for the definition of Ritz values and Ritz vectors). This observation led to more efficient orthogonalization strategies. In 1979, Parlett and Scott [136] proposed the *selective orthogonalization* which consists of the orthogonalization of each new Lanczos vector against the last two Lanczos vectors and all previously converged Ritz vectors and those that are about to converge. In 1984, Simon [154] introduced the *partial orthogonalization* where the orthogonalization is performed against the last two and some of the previous Lanczos vectors, depending on a recurrence describing the loss of orthogonality.

In [158] the impact of certain round-off errors on the relative residual was analyzed for an unpreconditioned MINRES variant. An upper bound on the difference between the exact arithmetic residual r_n and the finite precision residual $\widehat{r}_n$ was given [158, formula (26)]

$$\frac{\|r_n - \widehat{r}_n\|}{\|b\|} \leq \varepsilon \left(3\sqrt{3n}\kappa_2(\mathsf{A})^2 + n\sqrt{n}\kappa_2(\mathsf{A})\right).$$

The corresponding bound for GMRES [158, formula (17)] only involves a factor of $\kappa_2(\mathsf{A})$ instead of its square. The numerical results in [158] also indicate that the maximal attainable accuracy of MINRES is worse than the one of GMRES. Thus, if very high accuracy is required, the GMRES algorithm should be used, preferably with Householder orthogonalization.

The analysis of round-off errors in Krylov subspace methods is a wide field of research with an overwhelming amount of articles. The monograph of Meurant [116] provides an in-depth analysis of the Lanczos and CG methods and their algorithmic realizations in exact and finite precision arithmetic. A discussion of the finite precision properties of the CG and GMRES algorithms and pointers to further literature can be found in the monograph of Liesen and Strakoš [105].

In the next chapter, projections are used extensively in the context of deflated Krylov subspace methods. For orthogonal projections, the implementation can be realized with (iterated) Gram–Schmidt variants whose properties with respect to round-off errors are well understood. In 2011, Stewart [165] complemented the literature with a thorough round-off error analysis of oblique projections. For the Euclidean inner product on $\mathcal{H} = \mathbb{C}^N$ and two n-dimensional subspaces $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$, the analysis in [165] suggests the XQRY-form of the projection $\mathsf{P} = \mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}$, where the application to a vector $z \in \mathbb{C}^N$ is carried out by a right-to-left evaluation of

$$\mathbf{V}\mathbf{R}^{-1}\mathbf{Q}^{\mathsf{H}}\mathbf{W}^{\mathsf{H}}z. \tag{2.30}$$

Here, $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{N,n}$ are such that $\mathcal{V} = [\![\mathbf{V}]\!]$, $\mathcal{W} = [\![\mathbf{W}]\!]$, $\mathbf{V}^{\mathsf{H}}\mathbf{V} = \mathbf{W}^{\mathsf{H}}\mathbf{W} = \mathbf{I}_n$ and $\mathbf{QR} = \mathbf{W}^{\mathsf{H}}\mathbf{V}$ is a QR-decomposition, i.e., $\mathbf{Q}, \mathbf{R} \in \mathbb{C}^{n,n}$ with $\mathbf{Q}^{\mathsf{H}}\mathbf{Q} = \mathbf{I}_n$ and $\mathbf{R}$ upper triangular. If $\tilde{\mathsf{P}}z$ denotes the computed application of the projection (2.30) in finite precision, then [165, corollary 5.2] shows that

$$\frac{\|\tilde{\mathsf{P}}z - \mathsf{P}z\|_2}{\|\mathsf{P}z\|_2} \leq \left(1 + \|\mathsf{P}\|_2 + \frac{\|\mathsf{P}\|_2 \, \|z\|_2}{\|\mathsf{P}z\|_2}\right) C\varepsilon + O(\varepsilon^2).$$

As Stewart points out, the appearance of $\|\mathsf{P}\|_2$ is disturbing but on the other hand it is not surprising since $\|\mathsf{P}\|_2 = \frac{1}{\cos\theta_{\max}(\mathcal{V},\mathcal{W})}$ can be seen as a measure for the departure from orthogonality (or conditioning) of the range $\mathcal{V}$ and null space $\mathcal{W}^\perp$ of P. Another interesting result from [165] is that care has to be taken when applying the complementary projection $\mathsf{P}_{\mathcal{W}^\perp,\mathcal{V}} = \mathsf{id} - \mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}$. In [165, section 7], it is shown that the rule of thumb “twice is enough” also holds for oblique projections, i.e., that the application of $\mathsf{P}_{\mathcal{W}^\perp,\mathcal{V}} = \mathsf{id} - \mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}$ to a vector z should be carried out twice with the XQRY-form of $\mathsf{P}_{\mathcal{V},\mathcal{W}^\perp}$ in equation (2.30). The proposed algorithms from [165] are implemented in [60] as `krypy.utils.Projection`.

Also the computation of small angles between subspaces turns out to be affected severely by round-off errors. The naive algorithm via the arccos of the singular values of $\langle V, W \rangle$ for orthonormal bases $V \in \mathcal{H}^n$ and $W \in \mathcal{H}^m$ is not able to accurately compute angles smaller than $\sqrt{\varepsilon}$. In [98], Knyazev and Argentati showed how small angles can be computed accurately. Their algorithm is implemented in [60] as `krypy.utils.angles`.

3. Recycling for sequences of linear systems

The goal of this thesis is to explore possibilities to improve the convergence behavior of Krylov subspace methods in situations where a sequence of linear systems has to be solved. Let

$$\mathsf{A}^{(i)}x^{(i)} = b^{(i)}, \quad i \in \{1, \dots, M\}, \tag{3.1}$$

be a sequence of consistent linear systems with linear operators $\mathsf{A}^{(i)} \in \mathcal{L}(\mathcal{H})$ and right hand sides $b^{(i)} \in \mathcal{H}$. It is assumed that the linear systems can only be solved subsequently and not all at once in parallel. This is, e.g., the case in many practical applications where the operator $\mathsf{A}^{(i)}$ and the right hand side $b^{(i)}$ depend on the solution $x^{(i-1)}$ of the last linear system. Often, the operators and right hand sides in the sequence are not random but subsequent linear operators and right hand sides are "close" to each other, i.e., $\mathsf{A}^{(i-1)} \approx \mathsf{A}^{(i)}$ and $b^{(i-1)} \approx b^{(i)}$. Each application has its own meaning of the symbol "≈" and the influence of differences between subsequent operators and right hand sides is treated in more detail in sections 3.3 and 3.4 of this chapter.

Such a sequence of linear systems arises in a wide range of applications, e.g., in

- implicit time-stepping schemes for time-dependent partial differential equations, cf. [115, 14, 16],
- optimization algorithms, where a problem has to be solved for a range of predefined or iteratively computed parameters, cf. [181, 114],
- nonlinear equations, where several iterations of Newton's method have to be performed, cf. [92, 32, 63, 64] and chapter 4.

The basic idea of *recycling* in the context of Krylov subspace methods for sequences of linear systems is to re-use information that has been computed in the solution process of the linear systems $1, \dots, i-1$ in order to speed up the solution process for the i-th linear system.

An effective recycling strategy for solving the sequence (3.1) addresses the following two questions:

1. How can external data be incorporated into a Krylov subspace method in order to influence its convergence behavior?
2. Which data from previously solved linear systems results in the best overall performance when it is incorporated into a Krylov subspace method for the next linear system?

An overview of some well-known approaches that address the first question is provided in section 3.1 for the CG, GMRES and MINRES methods, see sections 2.8 and 2.9 for basic properties and algorithms of these methods. In section 3.2, deflated Krylov subspace methods are discussed and analyzed for the purpose of incorporating external data. Furthermore, the close relationship between deflated and augmented Krylov subspace methods is analyzed.

The second question is dealt with in section 3.4 which relies on the perturbation results from section 3.3.

Since this chapter deals with practical mathematical strategies for actual computations, it is assumed that $\mathcal{H}$ is finite-dimensional, i.e., $N := \dim \mathcal{H} < \infty$.

3.1. Strategies

Of course, the question of how to accelerate Krylov subspace methods with external data is not new. This section aims at giving a brief overview of strategies that have been proposed and used in the literature.

In the context of sequences of linear systems, the Krylov subspaces that were constructed for previous linear systems potentially contain valuable information for the solution of the next linear system. A key question thus is: how can subspaces be used to influence the convergence behavior of a Krylov subspace method? This section omits the notational overhead of sequences and it is assumed that a single consistent linear system

$$\mathsf{A}x = b \tag{3.2}$$

is given with a linear operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and right hand side $b \in \mathcal{H}$.

In order to further simplify the notation in this chapter, the following projections are defined for a given $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $X \in \mathcal{H}^n$ such that $\mathcal{X} = [\![X]\!]$ and $\mathsf{A}\mathcal{X}$ are n-dimensional.

1. If $\theta_{\max}(\mathcal{X}, \mathsf{A}\mathcal{X}) < \frac{\pi}{2}$, then define

$$\mathsf{P}_{\mathcal{X}}^{\mathrm{CG}} := \mathsf{P}_{\mathcal{X},(\mathsf{A}^\star\mathcal{X})^\perp}, \qquad \text{i.e.} \quad \mathsf{P}_{\mathcal{X}}^{\mathrm{CG}} x = X\langle X, \mathsf{A}X\rangle^{-1}\langle X, \mathsf{A}x\rangle, \tag{3.3}$$

$$\text{and} \quad \mathsf{Q}_{\mathcal{X}}^{\mathrm{CG}} := \mathsf{P}_{\mathcal{X}^\perp,\mathsf{A}\mathcal{X}}, \qquad \text{i.e.} \quad \mathsf{Q}_{\mathcal{X}}^{\mathrm{CG}} x = x - \mathsf{A}X\langle X, \mathsf{A}X\rangle^{-1}\langle X, x\rangle. \tag{3.4}$$

If A is self-adjoint and positive semidefinite, then $\mathsf{P}_{\mathcal{X}}^{\mathrm{CG}}$ is the orthogonal projection on $\mathcal{X}$ with respect to the (possibly semidefinite) inner product $\langle\cdot,\cdot\rangle_{\mathsf{A}}$.

2. For a general A

$$\mathsf{P}_{\mathcal{X}}^{\mathrm{MR}} := \mathsf{P}_{\mathcal{X},(\mathsf{A}^\star\mathsf{A}\mathcal{X})^\perp}, \qquad \text{i.e.} \quad \mathsf{P}_{\mathcal{X}}^{\mathrm{MR}} x = X\langle \mathsf{A}X, \mathsf{A}X\rangle^{-1}\langle \mathsf{A}X, \mathsf{A}x\rangle \tag{3.5}$$

is the orthogonal projection on $\mathcal{X}$ with respect to the (possibly semidefinite) inner product $\langle\cdot,\cdot\rangle_{\mathsf{A}^\star\mathsf{A}}$. Furthermore, the following projection is defined:

$$\mathsf{Q}_{\mathcal{X}}^{\mathrm{MR}} := \mathsf{P}_{(\mathsf{A}\mathcal{X})^\perp}, \qquad \text{i.e.} \quad \mathsf{Q}_{\mathcal{X}}^{\mathrm{MR}} x = x - \mathsf{A}X\langle \mathsf{A}X, \mathsf{A}X\rangle^{-1}\langle \mathsf{A}X, x\rangle. \tag{3.6}$$

The complementary projections are denoted by $\mathsf{P}^{\mathrm{CG}}_{\mathcal{X}^\perp} = \mathsf{id} - \mathsf{P}^{\mathrm{CG}}_{\mathcal{X}}$ and $\mathsf{P}^{\mathrm{MR}}_{\mathcal{X}^\perp} = \mathsf{id} - \mathsf{P}^{\mathrm{MR}}_{\mathcal{X}}$.

Lemma 3.1. *If the above projections are well defined then the following statements hold:*

1. $\mathsf{A}\mathsf{P}^{CG}_{\mathcal{X}^\perp} = \mathsf{Q}^{CG}_{\mathcal{X}}\mathsf{A}$.
2. $\mathsf{A}\mathsf{P}^{MR}_{\mathcal{X}^\perp} = \mathsf{Q}^{MR}_{\mathcal{X}}\mathsf{A}$.

Proof. See lemma 2.12. □

For $1 \le n \le d(\mathsf{A}, r_0)$, the projections $\mathsf{P}^{\mathrm{CG}}_{\mathcal{X}}$ and $\mathsf{P}^{\mathrm{MR}}_{\mathcal{X}}$ can be used to represent the iterates x_n^{CG} of the CG method and the iterates x_n^{MR} of the GMRES and MINRES methods, cf. sections 2.4, 2.8 and 2.9:

$$\begin{aligned} x_n^{\mathrm{CG}} &= x_0 + \mathsf{P}^{\mathrm{CG}}_{\mathcal{K}_n(\mathsf{A},r_0)}(x - x_0) \\ \text{and} \qquad x_n^{\mathrm{MR}} &= x_0 + \mathsf{P}^{\mathrm{MR}}_{\mathcal{K}_n(\mathsf{A},r_0)}(x - x_0). \end{aligned}$$

The corresponding residuals are

$$\begin{aligned} r_n^{\mathrm{CG}} &= b - \mathsf{A}x_n^{\mathrm{CG}} = \mathsf{A}\left(\mathsf{id} - \mathsf{P}^{\mathrm{CG}}_{\mathcal{K}_n(\mathsf{A},r_0)}\right)(x - x_0) = \mathsf{Q}^{\mathrm{CG}}_{\mathcal{K}_n(\mathsf{A},r_0)} r_0 \\ \text{and} \qquad r_n^{\mathrm{MR}} &= b - \mathsf{A}x_n^{\mathrm{MR}} = \mathsf{A}\left(\mathsf{id} - \mathsf{P}^{\mathrm{MR}}_{\mathcal{K}_n(\mathsf{A},r_0)}\right)(x - x_0) = \mathsf{Q}^{\mathrm{MR}}_{\mathcal{K}_n(\mathsf{A},r_0)} r_0. \end{aligned}$$

The following subsections give a brief overview on well-known techniques that allow to incorporate external data in Krylov subspace methods by adapting the initial guess, modifying the preconditioner or augmenting the search space. Section 3.2 and subsequent sections concentrate on deflated Krylov subspace methods.

3.1.1. Initial guess

The most obvious approach is to construct a "good" initial guess x_0. Given an m-dimensional subspace $\mathcal{U}$ that potentially contains a good approximation of an exact solution x, the following two approaches appear to be sensible in the light of the minimization properties of the CG, GMRES and MINRES methods:

1. Choose $x_0^{\mathrm{CG}} \in \mathcal{U}$ such that the A-norm of the error is minimized if A is self-adjoint and positive (semi-)definite:

$$\left\| x - x_0^{\mathrm{CG}} \right\|_{\mathsf{A}} = \min_{u \in \mathcal{U}} \| x - u \|_{\mathsf{A}} . \tag{3.7}$$

2. Choose $x_0^{\mathrm{MR}} \in \mathcal{U}$ such that the residual norm is minimized:

$$\left\| b - \mathsf{A}x_0^{\mathrm{MR}} \right\| = \min_{u \in \mathcal{U}} \| b - \mathsf{A}u \| = \min_{z \in \mathsf{A}\mathcal{U}} \| b - z \| . \tag{3.8}$$

The first option matches the minimization property of the CG method while the latter one fits the minimization property of the GMRES and MINRES methods. If $U \in \mathcal{H}^m$ is given with $\mathcal{U} = [\![U]\!]$ and $\mathcal{U} \cap \mathcal{N}(\mathsf{A}) = \{0\}$, then the optimal initial guesses with respect to $\mathcal{U}$ in (3.7) and (3.8) can be represented with the help of corollary 2.26 by

$$x_0^{\text{CG}} = \mathsf{P}_{\mathcal{U}}^{\text{CG}} x = U\langle U, \mathsf{A}U\rangle^{-1}\langle U, b\rangle \tag{3.9}$$

$$\text{and} \qquad x_0^{\text{MR}} = \mathsf{P}_{\mathcal{U}}^{\text{MR}} x = U\langle \mathsf{A}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{A}U, b\rangle. \tag{3.10}$$

The initial guess x_0^{CG} has been used by Fischer in [55] for a sequence of linear systems with a fixed self-adjoint and positive-definite operator $\mathsf{A}^{(1)} = \ldots = \mathsf{A}^{(M)}$. In [55], the used recycle data are approximate solutions of the previously solved linear systems.

If a nonzero initial guess $x_0 \in \mathcal{H}$ is given, then the computation of the initial guess x_0^{MR} with $U = x_0$ is known as the *Hegedüs trick* [105].

If an m-dimensional subspace $\mathcal{U} = [\![u_1, \ldots, u_m]\!]$ and a nonzero initial guess $x_0 \notin \mathcal{U}$ is given, then the optimal initial guesses (3.9)–(3.10) with respect to $\mathcal{U}$ can be constructed with $\widehat{U} := [x_0, u_1, \ldots, u_m]$ instead of U.

In order to implement the computation of x_0^{CG}, a QR decomposition of $U = Q\mathbf{R}$ with $Q \in \mathcal{H}^m$, $\langle Q, Q\rangle_{\mathsf{A}} = \mathbf{I}_m$ and upper triangular and invertible $\mathbf{R} \in \mathbb{C}^{m,m}$ can be computed. A straightforward computation shows that the initial guess (3.9) can then be obtained by

$$x_0^{\text{CG}} = Q\langle Q, b\rangle.$$

The computational cost of this implementation is the cost of orthogonalizing U in the inner product $\langle\cdot,\cdot\rangle_{\mathsf{A}}$ (which includes m applications of the operator A), and m inner products and vector updates.

For the computation of x_0^{MR}, a QR decomposition of $\mathsf{A}U = Q\mathbf{R}$ with $Q \in \mathcal{H}^m$, $\langle Q, Q\rangle = \mathbf{I}_m$ and upper triangular and invertible $\mathbf{R} \in \mathbb{C}^{m,m}$ can be performed such that the optimal initial guess with respect to $\mathcal{U}$ becomes

$$x_0^{\text{MR}} = U\mathbf{R}^{-1}\langle Q, b\rangle.$$

The computational cost for this implementation is the cost of m applications of the operator A, the cost of the QR decomposition of $\mathsf{A}U$, the cost of m inner products and vector updates and the application of $\mathbf{R}^{-1}$ (backward substitution).

For $m \ll N$, the complexity for the computation of x_0^{CG} and x_0^{MR} is thus dominated by the m applications of the operator A in practice. Note that if GMRES is used with a singular operator A, the solvability condition $r_0 = b - \mathsf{A}x_0^{\text{MR}} \in \mathcal{R}(\mathsf{A}^m)$ (with m being the index of the zero eigenvalue of A) may be violated by using the initial guesses (3.9)–(3.10), cf. theorem 2.57 and proposition 2.39.

The use of a nonzero initial guess can reduce the initial residual in some cases where a problem-related initial guess is at hand. However, it will not change the overall convergence behavior of the Krylov subspace method in general – unless the initial guess is of a very special form, e.g., such that the corresponding initial residual has only a few nonzero entries in a representation in terms of an eigenvector basis.

In most cases, however, the adaption of the initial guess alone does not change the qualitative convergence behavior. In practice, the adaption is roughly as costly as m iterations of CG, MINRES or GMRES. Therefore, it should only be used if there is evidence that it reduces the number of iterations to reach a given tolerance by m or more.

3.1.2. Preconditioning

Another strategy concerns the use of preconditioners. In practice, Krylov subspace methods are usually only feasible in combination with a preconditioner. The choice of the preconditioner is highly problem-dependent and thus there is no general recipe for the incorporation of external data into preconditioners.

If the operators in a sequence of linear systems are explicitly given as a matrices, one strategy is to use a (possibly incomplete) factorization of the matrix of one linear system in the sequence as a preconditioner for subsequent linear systems. When the performance of the Krylov subspace method degrades due to the accumulated changes in the matrices, a new factorization can be computed.

Knoll and Keyes [97] used this approach in the context of Newton-Krylov methods (or inexact Newton methods), i.e., Newton's method where the linear systems with the Jacobian operator are solved up to a specified tolerance with a Krylov subspace method.

A similar strategy was used by Mehrmann and Schröder in [114] for a sequence of large scale linear systems in an industrial application. The linear systems represent a frequency response problem stemming from the optimization of acoustic fields and are of the form

$$(-\omega^2\mathbf{M} + \mathbb{i}\omega\mathbf{D} + \mathbf{K}_{\mathbb{R}}^{(\omega)} + \mathbb{i}\mathbf{K}_{\mathbb{C}})x^{(\omega)} = b^{(\omega)} \tag{3.11}$$

with real symmetric matrices $\mathbf{M}, \mathbf{K}_{\mathbb{R}}^{(\omega)}, \mathbf{K}_{\mathbb{C}} \in \mathbb{R}^{N,N}$ and a complex symmetric matrix $\mathbf{D} \in \mathbb{C}^{N,N}$. The parameter ω represents the frequency and the linear system (3.11) has to be solved for the frequency range $\omega \in \{1, 2, \dots, 10^3\}$. For small values of ω, the real symmetric matrix

$$\mathbf{P} := -\omega^2\mathbf{M} + \mathbf{K}_{\mathbb{R}}^{(\omega)}$$

is factorized with the MUMPS software package [1] into $\mathbf{P} = \mathbf{LDL}^\mathsf{T}$ and used as a preconditioner for a GMRES variant. A once computed factorization is reused as a preconditioner for subsequent linear systems until the performance of the Krylov subspace method deteriorates. The factorization is then recomputed for the matrix of the current linear system. For large frequencies close to 10^3, Mehrmann and Schröder observed that the used Krylov subspace method hardly converged. They thus switched from the Krylov subspace method to the direct solver MUMPS which is used to perform a $\mathbf{LDL}^\mathsf{T}$ factorization of the entire complex symmetric matrix in equation (3.11) for these frequencies.

For specific classes of preconditioners, update formulas for the preconditioner are known. Meurant showed in [115] how a Cholesky-like factorization of the matrix $\epsilon\mathbf{I}+\mathbf{A}$ with $\epsilon \in \mathbb{R}_+$ can be obtained as an update of an already computed incomplete Cholesky factorization of a symmetric M-matrix $\mathbf{A}$. The proposed update schemes have been applied to a finite difference discretization of the heat equation in [115]. Benzi and Bertaccini [11] showed how an update can be performed for the perturbed matrix $\epsilon\mathbf{I}+\mathbf{A}$ with a general symmetric and positive-definite matrix A. In [14], Bertaccini allows the perturbed matrix to take the form $\mathbf{D}+\mathbf{A}$, where $\mathbf{D}\in\mathbb{C}^{N,N}$ is a diagonal matrix and $\mathbf{A}\in\mathbb{C}^{N,N}$ is Hermitian and positive semidefinite. In a series of articles, Duintjer Tebbens and Tůma [38, 39, 40] and their coauthors Birken and Meister [16] generalized the approaches from [11, 14] and showed how (incomplete) LU factorizations can be updated. In these works, no assumptions are made on the structure of the perturbation. In [40], the updates for an incomplete LU factorization can be used in a matrix-free setting, i.e., when the matrix $\mathbf{A}$ is not explicitly formed but only the matrix-vector multiplication can be evaluated. Note that these updating schemes do not carry over data from the solution process from one linear system to the next but try to construct a preconditioner update by only analyzing the changes in the matrix.

In the case of geometric multigrid [78] or algebraic multigrid [141] preconditioners, data can often be reused easily in a sequence of linear systems. If the linear systems result from the discretization of a partial differential equation and the geometry is fixed throughout the sequence, then the restriction and prolongation operators in the multigrid method remain constant. Similarly, the sparsity pattern of the matrix does not change in the sequence in many applications of algebraic multigrid preconditioners and thus the costly initial setup of the restriction and prolongation operators only has to be carried out once for the first linear system.

All of the above strategies aim at reducing the costs for the construction or application of a preconditioner. However, they do not incorporate external data into the solution process and do not benefit from the fact that previous linear systems have already been solved. Nevertheless, it should be emphasized strongly, that the application of a problem-dependent preconditioner is crucial in almost all applications and the approaches in subsequent sections should be seen as complementary to preconditioning. An overview of preconditioning techniques can be found in the survey article of Benzi [10] and the books of Saad [147] and Greenbaum [73].

3.1.3. Augmentation

In augmented Krylov subspace methods, the search space is enlarged by an m-dimensional subspace $\mathcal{U}$. With a Krylov subspace method that satisfies an optimality condition like CG, GMRES or MINRES, the underlying minimization can be carried out over the sum of the Krylov subspace and the augmentation subspace $\mathcal{U}$. Ideally, the augmentation subspace $\mathcal{U}$ contains information about the problem that is only slowly revealed in the Krylov subspace itself, e.g., eigenvectors corresponding to eigenvalues of small magnitude. This approach was used by Morgan [118] and also by Chapman and Saad [23] in order to overcome recurring convergence slowdowns after restarting the GMRES method. For $n < d(\mathsf{A}, r_0)$, the method described by Morgan computes the minimal residual approximation with respect to an initial guess x_0 and the subspace $\mathcal{K}_n(\mathsf{A}, r_0) + \mathcal{U}$, i.e., the approximate solution $x_n^{\mathrm{MR}} \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0) + \mathcal{U}$ that satisfies

$$\left\|b - \mathsf{A}x_n^{\mathrm{MR}}\right\| = \min_{z \in x_0 + \mathcal{K}_n(\mathsf{A}, r_0) + \mathcal{U}} \|b - \mathsf{A}z\| .$$

In the restarted GMRES setup in [118] and [23], the subspace $\mathcal{U}$ is chosen as the span of a few harmonic Ritz vectors which correspond to the harmonic Ritz values of smallest magnitude. In [121], Morgan showed how Ritz vectors can be included implicitly as an augmentation space upon restarts.

If a nonzero initial guess x_0 is used with augmentation, the initial guess can also be included in the augmentation subspace, i.e., the augmentation space $\widehat{\mathcal{U}} = \mathcal{U} + [\![x_0]\!]$ can be used instead of $\mathcal{U}$. Without loss of generality, it is now assumed that the initial guess is already included in $\mathcal{U}$. Then the optimal initial guess x_0^{MR} with minimal residual r_0^{MR} with respect to $\mathcal{U}$ can be used, see subsection 3.1.1. For $0 \leq n \leq d(\mathsf{A}, r_0^{\mathrm{MR}})$, the iterates x_n^{MR} then can be represented by

$$x_n^{\mathrm{MR}} = \mathsf{P}^{\mathrm{MR}}_{\mathcal{K}_n(\mathsf{A}, r_0^{\mathrm{MR}}) + \mathcal{U}} x$$

and satisfy the optimality property

$$\left\|b - \mathsf{A}x_n^{\mathrm{MR}}\right\| = \min_{z \in \mathcal{K}_n(\mathsf{A}, r_0^{\mathrm{MR}}) + \mathcal{U}} \|b - \mathsf{A}z\| . \tag{3.12}$$

For self-adjoint and positive-semidefinite operators A, an analogous strategy can be carried out for the CG method by minimizing the A-norm of the error over the augmented Krylov subspace. For $0 \leq n \leq d(\mathsf{A}, r_0^{\mathrm{CG}})$, the iterates x_n^{CG} are given by

$$x_n^{\mathrm{CG}} = \mathsf{P}^{\mathrm{CG}}_{\mathcal{K}_n(\mathsf{A}, r_0^{\mathrm{CG}}) + \mathcal{U}} x$$

and satisfy the optimality property

$$\left\|x - x_n^{\mathrm{CG}}\right\|_{\mathsf{A}} = \min_{z \in \mathcal{K}_n(\mathsf{A}, r_0^{\mathrm{CG}}) + \mathcal{U}} \|x - z\|_{\mathsf{A}} . \tag{3.13}$$

Note that the minimization properties in equations (3.13)–(3.12) coincide for $n = 0$ with the respective minimization properties for the initial guess in equations (3.7)–(3.8). However, the minimization in equations (3.7)–(3.8) is only used to obtain an initial guess with minimal residual before applying the GMRES or MINRES method and the subspace $\mathcal{U}$ is ignored in the actual GMRES or MINRES method. In contrast, the subspace $\mathcal{U}$ is explicitly included in the minimization in equations (3.13)–(3.12) while the Krylov subspace method proceeds. Thus, the above augmentation approach can be seen as a straightforward extension of the adaption of the initial guess to the full solution process in a Krylov subspace method.

Because an augmented Krylov subspace method of the above type just includes the subspace $\mathcal{U}$ in the minimization, it is obvious that the errors in the A-norm or residual norms are bounded by their non-augmented counterparts, i.e.

$$\left\|x - x_n^{\mathrm{CG}}\right\|_{\mathsf{A}} \le \min_{z \in \mathcal{K}_n(\mathsf{A}, r_0^{\mathrm{CG}})} \|x - z\|_{\mathsf{A}}$$

$$\text{and} \qquad \left\|b - \mathsf{A}x_n^{\mathrm{MR}}\right\| \le \min_{z \in \mathcal{K}_n(\mathsf{A}, r_0^{\mathrm{MR}})} \|b - \mathsf{A}z\| .$$

It is important to note that there is no guarantee that the sum $\mathcal{K}_n(\mathsf{A}, r_0) + \mathcal{U}$ is a direct sum. In finite precision, the sum will often be direct but the minimal angle $\theta_{\min}(\mathcal{K}_n(\mathsf{A}, r_0), \mathcal{U})$ may become small and thus care has to be taken if separate bases for $\mathcal{K}_n(\mathsf{A}, r_0)$ and $\mathcal{U}$ are used.

Erhel and Guyomarc'h [49, 48] introduced an augmented variant of the CG algorithm for sequences of linear systems where the operators $\mathsf{A}^{(1)} = \ldots = \mathsf{A}^{(M)}$ are constant and only the right hand sides change. In their method, the subspace $\mathcal{U}$ is the full Krylov subspace that was constructed in the solution process of the previous linear system. The method minimizes the A-norm of the error over the subspace $\mathcal{K}_n(\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}, r_0) + \mathcal{U}$, i.e., the Krylov subspace is built with the projected operator $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}$. This also makes this method a *deflated* method which is the subject of the next subsection.

Saad [146] complemented the literature with convergence estimates of augmented minimal residual Krylov subspace methods in the case of nearly A-invariant augmentation spaces $\mathcal{U}$. Let

$$\mathsf{A} = [S_1, S_2] \begin{bmatrix} \mathbf{J}_1 & \\ & \mathbf{J}_2 \end{bmatrix} [S_1, S_2]^{-1}$$

be a partitioned Jordan decomposition and let $\mathcal{S}_1 = [\![S_1]\!], \mathcal{S}_2 = [\![S_2]\!]$. Then Saad showed in [146, corollary 3.6] that the iterates x_n^{MR} from the augmented minimal residual method, see equation (3.12), satisfy

$$\left\|b - \mathsf{A}x_n^{\mathrm{MR}}\right\| \le \frac{\|\bar{r}_n\|}{\sin\left(\frac{\theta_{\min}(\mathcal{S}_1, \mathcal{S}_2)}{2}\right)},$$

where $\bar{r}_n$ is the residual of n steps GMRES applied to the linear system

$$\mathsf{A}\delta = \theta_{\max}(\mathsf{A}\mathcal{U}, \mathcal{S}_1)\mathsf{P}_{\mathcal{S}_1, \mathcal{S}_2} r_0^{\mathrm{MR}} + \mathsf{P}_{\mathcal{S}_2, \mathcal{S}_1} r_0^{\mathrm{MR}}$$

with initial guess zero. The analysis suggests that the convergence behavior of the augmented minimal residual method is similar to GMRES applied to the linear system

$$\mathsf{A}\delta = \mathsf{P}_{\mathcal{S}_2, \mathcal{S}_1} r_0^{\mathrm{MR}} \tag{3.14}$$

if the sine of the maximal angle $\sin\theta_{\max}(\mathsf{A}\mathcal{U}, \mathcal{S}_1)$ between $\mathsf{A}\mathcal{U}$ and an invariant subspace $\mathcal{S}_1$ is small. The right hand side in equation (3.14) has no components in the invariant subspace $\mathcal{S}_1$ but only in $\mathcal{S}_2$ which may significantly improve the convergence behavior compared to the original right hand side. However, as it is shown in section 3.3.3, Krylov subspace methods are very sensitive to perturbations of the operator or the right hand side without further assumptions. Krylov subspace methods with perturbations are discussed in detail in the context of deflation in sections 3.3.3 and 3.4.3.

Recently, Wakam and Erhel [179] proposed a restarted GMRES variant that uses an augmented non-orthogonal basis in order to reduce the communication time in the context of parallel computations on distributed memory machines.

In [43], Eiermann, Ernst and Schneider provided a clear abstract framework for analyzing augmented Krylov subspace methods in the setting of restarted minimal residual methods.

3.2. Deflation

In *deflated* Krylov subspace methods, the operator of the linear system is explicitly modified in order to improve the convergence behavior. This is, e.g., achieved by multiplying the linear system with a projection P and then applying a Krylov subspace method to the projected linear system

$$\mathsf{P}\mathsf{A}x_\star = \mathsf{P}b. \tag{3.15}$$

The projected operator PA exhibits the eigenvalue zero with multiplicity at least $m = \dim \mathcal{N}(\mathsf{P})$. Of course, the projection P has to fulfill certain criteria in order to obtain a well-defined Krylov subspace method because PA is singular if P is not the identity projection, i.e., if $m > 0$. Furthermore, a solution $x_\star$ of the deflated linear system (3.15) is not necessarily a solution of the original linear system (3.2) and some "correction" scheme is required in order to recover a solution of (3.2) from $x_\star$.

The next subsection further motivates deflated Krylov subspace methods and gives a brief overview of the literature.

3.2.1. Motivation

A popular choice in the literature are projections that are based on approximations of invariant subspaces. In order to motivate this idea, let

$$\mathsf{A} = [S_1, S_2] \begin{bmatrix} \mathbf{J}_1 & \\ & \mathbf{J}_2 \end{bmatrix} [S_1, S_2]^{-1}$$

be a partitioned Jordan decomposition and $\mathcal{S}_1 = [\![S_1]\!], \mathcal{S}_2 = [\![S_2]\!]$. Then the projection $\mathsf{P}_{\mathcal{S}_2,\mathcal{S}_1}$ is well defined and a Jordan decomposition of the projected operator $\mathsf{P}_{\mathcal{S}_2,\mathcal{S}_1}\mathsf{A}$ is

$$\mathsf{P}_{\mathcal{S}_2,\mathcal{S}_1}\mathsf{A} = [S_1, S_2] \begin{bmatrix} 0 & \\ & \mathbf{J}_2 \end{bmatrix} [S_1, S_2]^{-1}.$$

If a Krylov subspace method is applied to the deflated linear system (3.15) with $\mathsf{P} = \mathsf{P}_{\mathcal{S}_2,\mathcal{S}_1}$, then the invariant subspace $\mathcal{S}_1$ and its associated spectrum $\Lambda(\mathbf{J}_1)$ are "hidden" from the method and the impact on the convergence behavior can be substantial. Also the spectral convergence bounds for CG, GMRES and MINRES may improve significantly due to the fact that only the subset $\Lambda(\mathbf{J}_2) \subseteq \Lambda(\mathsf{A})$ has to be taken into account, see theorems 2.56, 2.58 and 2.63. An approximate solution x_n of the deflated linear system 3.15 can be turned into an approximate solution of the original linear system (3.2) by either using an adapted initial guess or by a simple correction scheme. The details are addressed in section 3.2.2.

In practice, only an approximation $\mathcal{U}$ to an invariant subspace $\mathcal{S}_1$ is at hand. Still $m = \dim \mathcal{N}(\mathsf{P})$ eigenvalues are sent to zero but the remaining part of the spectrum will differ from $\Lambda(\mathbf{J}_2)$ in general. However, it is to be expected that small perturbations of an invariant subspace lead to small perturbations in the remaining spectrum if the minimal angle $\theta_{\min}(\mathcal{S}_1, \mathcal{S}_2)$ is not too small, i.e., if the invariant subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ are well-conditioned. The effects of perturbed subspaces on projections and spectral properties of projected operators are analyzed in sections 3.3.1 and 3.3.2.

Furthermore, it is shown in section 3.2.6, that there is a close relationship between deflated and augmented Krylov subspace methods. Because deflated methods essentially replace the operator with a projected operator, the impact on the convergence behavior can be studied by analyzing Krylov subspace methods for singular linear systems. In the remaining part of this subsection, a brief overview of deflation strategies in the literature is given. A comprehensive presentation with a broad overview on the literature can be found in the survey article by Simoncini and Szyld [156]. Subsequent sections present details for the deflated CG, MINRES and GMRES methods along with new results that emerge from the analysis in this thesis.

The first deflated Krylov subspace methods have been introduced in 1987 by Nicolaides [128] and in 1988 by Dostál [34]. Both showed how the convergence behavior can be improved by using a deflated CG method for symmetric and positive-definite operators A. Since these early works, the idea of deflation and augmentation has evolved and several authors applied them to a wide range of Krylov subspace methods. Saad, Yeung, Erhel and Guyomarc'h [143] presented a deflated CG algorithm that is mathematically equivalent to the one that Nicolaides introduced in [128]. The authors of [143] also seem to have made the first link between augmented and deflated Krylov subspace methods by showing that the deflated CG algorithm is a generalization of the augmented CG algorithm of Erhel and Guyomarc'h [49] to arbitrary augmentation spaces.

The determination of a deflation subspace varies across the literature but there is a strong focus on using eigenvector approximations, e.g., Ritz or harmonic Ritz vectors. However, other choices have been studied by several authors. In the context of discretized elliptic partial differential equations, already the early work of Nicolaides [128] considers the use of deflation subspaces that are based on piecewise constant interpolation from a set of subdomains. Mansfield [112] showed that Schur complement-type domain decomposition methods can be seen as a series of deflated methods. Similar to the work of Nicolaides [128], Mansfield [111] used a deflation subspace that is based on a coarse grid interpolation and combined this approach with a damped Jacobi smoother as a preconditioner that is related to the two-grid method. Nabben and Vuik pointed out similarities between the deflated CG method and domain decomposition methods in [124, 125, 123] and extended the comparison to multigrid methods with Tang and Erlangga in [173] and with Tang and MacLachlan in [172]. In [88], Kahl and Rittich presented convergence estimates for the deflated CG method with arbitrary deflation subspaces based on the effective condition number and techniques from algebraic multigrid methods.

For non-self-adjoint operators, de Sturler introduced the GCRO method in [169] which is a nested Krylov subspace method involving an outer and an inner iteration. The outer method is the GCR method [45, 44] and in each iteration of GCR several steps of GMRES are applied to the projected linear system (3.15) with $\mathsf{P} = \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}$, where the subspace $\mathcal{U}$ is determined by the outer GCR iteration. In [170], de Sturler enhanced the GCRO method with an optimal truncation variant for restarts (GCROT). Furthermore, the GCRO method was generalized to an arbitrary subspace $\mathcal{U}$ by Kilmer and de Sturler in [94].

The eigenvalue translation preconditioner by Kharchenko and Yeremin [93] also aims at moving some eigenvalues but instead of moving them to zero they are moved to a vicinity of 1. However, the technique does not fit into the setting of a deflated linear system 3.15 but rather is a classical preconditioner. The same holds for the approach that was taken by Erhel, Burrage and Pohl in [47] where the presented preconditioner moves some eigenvalues to the spectral radius $\rho(\mathsf{A})$ of the operator A. For the CG method, Nabben and Vuik [125] showed that a pure deflation approach yields iterates with smaller or equal A-norm of the error compared to a deflation-based preconditioner that shifts eigenvalues to 1. Deflation-based preconditioners have also been used and analyzed by Erlangga and Nabben [50, 51] and Tang, Nabben, Vuik and Erlangga [173].

3.2.2. Projections and corrections

As mentioned in the previous section 3.2.1, A-invariant subspaces are usually not known in practice. Thus, analogous to the strategies in section 3.1, it is now just assumed that an m-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$ is given in order to construct a projection P for the deflated linear system (3.15). Furthermore, let $U \in \mathcal{H}^m$ be given such that $\mathcal{U} = [\![U]\!]$. In principle, one could apply CG, GMRES or MINRES to the projected linear system (3.15) with any projection, e.g., $\mathsf{P} = \mathsf{P}_{\mathcal{U}^\perp}$. However, care has to be taken such that the following conditions are met:

1. The Krylov subspace method is well defined when it is applied to the projected linear system (3.15) which implies that it terminates with a solution $x_\star$ of (3.15), see definition 2.40.

2. A solution x of the original linear system (3.2) can be recovered from $x_\star$.

This subsection discusses existing approaches from the literature and studies when they satisfy the above conditions.

Two projections are dominant in the literature for deflated Krylov subspace methods, i.e., Krylov subspace methods that are applied to the projected linear system (3.15):

1. $\mathsf{P} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$, cf. equation (3.4), is primarily used with the CG method, e.g., in [128, 34, 100, 143, 56, 124, 125, 123, 173, 172]. The projection $\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$ is well defined if the linear operator A is self-adjoint and positive semidefinite but may cease to exist for indefinite or non-self-adjoint linear operators A.

2. $\mathsf{P} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}$, cf. equation (3.6), appears in the context of the GMRES, MINRES and GCRO methods, e.g., in [169, 94, 135, 181, 160].

The use of the two projections $\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$ and $\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}$ can be motivated as follows. Consider the projections $\mathsf{P}_{\mathcal{U}}^{\mathrm{CG}}$ and $\mathsf{P}_{\mathcal{U}}^{\mathrm{MR}}$ that already showed up in the computation of an optimal initial guess with respect to a subspace, cf. section 3.1.1, and in augmented methods, cf. section 3.1.3. Both projections can be used to decompose a solution x of the original linear system (3.2) into two parts:

$$\begin{aligned} x &= \mathsf{P}_{\mathcal{U}}^{\mathrm{CG}} x + \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}} x \\ &= \mathsf{P}_{\mathcal{U}}^{\mathrm{MR}} x + \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}} x. \end{aligned}$$

The first terms of both decompositions are the optimal initial guesses x_0^{CG} and x_0^{MR} with respect to the subspace $\mathcal{U}$ that have been derived in section 3.1.1 and can be computed from the right hand side b:

$$\begin{aligned} & \mathsf{P}_{\mathcal{U}}^{\mathrm{CG}} x = x_0^{\mathrm{CG}} = U\langle U, \mathsf{A}U\rangle^{-1}\langle U, b\rangle \\ \text{and} \qquad & \mathsf{P}_{\mathcal{U}}^{\mathrm{MR}} x = x_0^{\mathrm{MR}} = U\langle \mathsf{A}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{A}U, b\rangle. \end{aligned}$$

For the second terms, let $x_\star^{\mathrm{CG}}, x_\star^{\mathrm{MR}} \in \mathcal{H}$ such that $\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}} x_\star^{\mathrm{CG}} = \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}} x$ and $\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}} x_\star^{\mathrm{MR}} = \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}} x$. Inserting the representations of the solution

$$\begin{aligned} x &= \mathsf{P}_{\mathcal{U}}^{\mathrm{CG}} x + \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}} x_\star^{\mathrm{CG}} \\ &= \mathsf{P}_{\mathcal{U}}^{\mathrm{MR}} x + \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}} x_\star^{\mathrm{MR}} \end{aligned} \tag{3.16}$$

into the original linear system (3.2) yields with lemma 3.1 the following equivalences:

$$\begin{aligned} & \mathsf{A}x = b \iff \mathsf{A}\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}} x_\star^{\mathrm{CG}} = \mathsf{A}(\mathsf{id} - \mathsf{P}_{\mathcal{U}}^{\mathrm{CG}})x \iff \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}} \mathsf{A} x_\star^{\mathrm{CG}} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}} b \\ \text{and} \qquad & \mathsf{A}x = b \iff \mathsf{A}\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}} x_\star^{\mathrm{MR}} = \mathsf{A}(\mathsf{id} - \mathsf{P}_{\mathcal{U}}^{\mathrm{MR}})x \iff \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}} \mathsf{A} x_\star^{\mathrm{MR}} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}} b. \end{aligned}$$

Thus, a solution x of the original linear system (3.2) can be obtained in two steps:

1. Obtain a solution $x_\star$ of the projected linear system (3.15) with a projection $\mathsf{P} \in \{\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}, \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\}$.

2. Update the obtained solution $x_\star$ with the corresponding correction from equation (3.16).

In the literature, the correction in step 2 is often carried out in the beginning by adapting the initial guess. Both approaches are known to be equivalent and thus only make an algorithmic difference, see section 3.2.7.

The remaining question is: under which conditions are the CG, GMRES and MINRES methods well defined when they are applied to the projected linear system (3.15) with $\mathsf{P} \in \{\mathsf{Q}_\mathcal{U}^{\mathrm{CG}}, \mathsf{Q}_\mathcal{U}^{\mathrm{MR}}\}$.

3.2.3. CG method

The following theorem gathers well-known results about the CG method for the deflated linear system 3.15 with $\mathsf{P} = \mathsf{Q}_\mathcal{U}^{\mathrm{CG}}$.

Theorem 3.2 (Deflated CG)**.** *Let $\mathsf{A}x = b$ be a consistent linear system with a self-adjoint and positive-semidefinite operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and right hand side $b \in \mathcal{H}$. Let $U \in \mathcal{H}^m$ be given such that $\mathcal{U} = [\![U]\!]$ and $\mathsf{A}\mathcal{U}$ are m-dimensional.*

Then the following holds:

1. *The projection $\mathsf{Q}_\mathcal{U}^{CG}$ is well defined.*

2. *For all initial guesses x_0, the CG method applied to the linear system*

$$\widehat{\mathsf{A}}\widehat{x} = \widehat{b} \tag{3.17}$$

with $\widehat{\mathsf{A}} := \mathsf{Q}_\mathcal{U}^{CG}\mathsf{A}$ and $\widehat{b} = \mathsf{Q}_\mathcal{U}^{CG} b$ is well defined.

3. *If $\widehat{x}_n$ is the constructed iterate in the n-th step in 2., then the corrected iterate*

$$x_n := \mathsf{P}_{\mathcal{U}^\perp}^{CG}\widehat{x}_n + U\langle U, \mathsf{A}U\rangle^{-1}\langle U, b\rangle \tag{3.18}$$

satisfies

$$\begin{aligned} &\|x - x_n\|_\mathsf{A} = \|x - \widehat{x}_n\|_{\widehat{\mathsf{A}}} \\ \textit{and} \qquad &r_n = b - \mathsf{A}x_n = \widehat{b} - \widehat{\mathsf{A}}\widehat{x}_n = \widehat{r}_n, \end{aligned}$$

where $x \in \mathcal{H}$ is a solution of $\mathsf{A}x = b$.

Proof. 1. The projection $\mathsf{Q}_\mathcal{U}^{\mathrm{CG}}$ is well defined if and only if $\langle U, \mathsf{A}U\rangle$ is nonsingular, cf. theorem 2.15. If A is nonsingular the statement is trivial because $\langle U, \mathsf{A}U\rangle = \langle U, U\rangle_\mathsf{A}$. If A is singular, the statement can be shown analogously to the first part of the proof of theorem 2.28.

2. The projected linear operator $\widehat{\mathsf{A}}$ is self-adjoint because by lemma 3.1:

$$\widehat{\mathsf{A}}^\star = (\mathsf{Q}_\mathcal{U}^{\mathrm{CG}}\mathsf{A})^\star = (\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A})^\star = \mathsf{A}\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}} = \mathsf{A}\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}} = \mathsf{Q}_\mathcal{U}^{\mathrm{CG}}\mathsf{A} = \widehat{\mathsf{A}}. \tag{3.19}$$

Furthermore, the operator $\widehat{\mathsf{A}}$ is positive semidefinite because for all $v \in \mathcal{H}$:

$$\begin{aligned}\langle v, \widehat{\mathsf{A}}v\rangle &= \langle v, \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}\mathsf{A}v\rangle = \langle v, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}v\rangle = \langle v, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}v\rangle \\ &= \langle \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}v, \mathsf{A}\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}v\rangle = \left\|\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}v\right\|_{\mathsf{A}}^2 \geq 0.\end{aligned}$$

Since the projected linear system (3.17) results from the original linear system by applying the projection $\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$, it is consistent, i.e., $\widehat{b} \in \mathcal{R}(\widehat{\mathsf{A}})$. The well-definedness thus follows from theorem 2.54.

3. The error norm equality follows from

$$\begin{aligned}\|x - x_n\|_{\mathsf{A}}^2 &= \left\|x - \mathsf{P}_{\mathcal{U}}^{\mathrm{CG}}x - \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}}\widehat{x}_n\right\|_{\mathsf{A}}^2 = \left\|\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}}(x - \widehat{x}_n)\right\|_{\mathsf{A}}^2 = \left\|\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}(x - \widehat{x}_n)\right\|_{\mathsf{A}}^2 \\ &= \langle \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}(x - \widehat{x}_n), \mathsf{A}\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}(x - \widehat{x}_n)\rangle = \langle x - \widehat{x}_n, \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}\mathsf{A}(x - \widehat{x}_n)\rangle \\ &= \|x - \widehat{x}_n\|_{\widehat{\mathsf{A}}}^2\end{aligned}$$

and the residual equality follows from

$$b - \mathsf{A}x_n = b - \mathsf{A}U\langle U, \mathsf{A}U\rangle^{-1}\langle U, b\rangle - \mathsf{A}\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}}\widehat{x}_n = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}b - \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}\mathsf{A}\widehat{x}_n = \widehat{b} - \widehat{\mathsf{A}}\widehat{x}_n.$$

□

The conditions under which the deflated CG method is well defined and the basic properties in the above theorem are well-known in the literature. This is not unconditionally true for all aspects of the deflated GMRES and MINRES methods which are covered in the next section. Details concerning the implementation of the deflated CG method are discussed together with GMRES and MINRES in section 3.2.7. New statements about the spectral properties of the deflated operator $\widehat{\mathsf{A}}$ as well as convergence bounds and selection strategies for deflation vectors in the setting of a sequence of linear systems are provided in sections 3.3 and 3.4.

3.2.4. GMRES method and breakdowns

The question of the conditions under which GMRES is well defined for a singular but consistent linear system $\mathsf{A}x = b$ has been answered in parts in section 2.9.1 and theorem 2.57. There, only a particular initial guess was assumed to be given such that equation (2.20) holds:

$$\mathcal{K}_d(\mathsf{A}, r_0) \cap \mathcal{N}(\mathsf{A}) = \{0\}.$$

Here, the question is: under which conditions is the GMRES method well defined for every initial guess x_0. As already noticed in section 2.9.1, a sufficient condition is

$$\mathcal{R}(\mathsf{A}) \cap \mathcal{N}(\mathsf{A}) = \{0\} \tag{3.20}$$

because $\mathcal{K}_d(\mathsf{A}, r_0) \subseteq \mathcal{R}(\mathsf{A})$ holds for a consistent linear system, i.e., if $b \in \mathcal{R}(\mathsf{A})$. This has also been shown by Brown and Walker in [18, theorem 2.6]. In [62, theorem 5.1], the author, Gutknecht, Liesen and Nabben extended the result of Brown and Walker [18, theorem 2.6] by showing that condition (3.20) is also a necessary condition.

Theorem 3.3. *Let* $\mathsf{A}x = b$ *be a consistent linear system with* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *(possibly singular) and* $b \in \mathcal{H}$*. Then the following two conditions are equivalent:*

1. *For every initial guess* $x_0 \in \mathcal{H}$*, the GMRES method applied to the linear system* $\mathsf{A}x = b$ *is well defined.*
2. $\mathcal{R}(\mathsf{A}) \cap \mathcal{N}(\mathsf{A}) = \{0\}$.

Proof. It has been shown in [18, theorem 2.6] and in section 2.9.1, that condition 2 implies condition 1. The reverse is proved by contradiction. If $0 \neq v \in \mathcal{N}(\mathsf{A}) \cap \mathcal{R}(\mathsf{A})$, then an initial guess can be constructed such that GMRES does not terminate with the solution. Because $v \in \mathcal{R}(\mathsf{A})$, there exists a nonzero $w \in \mathcal{H}$ such that $v = \mathsf{A}w$. Then the initial guess $x_0 := x - w$ yields $r_0 = b - \mathsf{A}x_0 = b - \mathsf{A}x + \mathsf{A}w = \mathsf{A}w = v$. But then $\mathsf{A}r_0 = 0$ because $v \in \mathcal{N}(\mathsf{A})$, such that the GMRES method terminates at the first iteration with the approximation x_0 which has a nonzero residual $r_0 = v$. □

If a Krylov subspace method is not well defined, i.e., it terminates without finding the solution, this situation is also referred to as a *breakdown* of the method. The above proof of theorem 3.3 leads to the following characterization of all initial guesses that lead to a breakdown of GMRES at the first iteration (see [62, corollary 5.2]).

Corollary 3.4. *Let* $\mathsf{A}x = b$ *be a consistent linear system with* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *(possibly singular), right hand side* $b \in \mathcal{H}$ *and a solution* $x \in \mathcal{H}$*. Then the GMRES method breaks down at the first iteration for all initial guesses*

$$x_0 \in \mathcal{X}_0 := \{x - w \mid \mathsf{A}w \in \mathcal{N}(\mathsf{A}) \setminus \{0\}\}.$$

As already mentioned in section 2.9.2, a self-adjoint operator A always fulfills condition (3.20). However, care has to be taken in the non-self-adjoint case.

Consider a nonsingular operator A and the deflated linear system (3.15) with $\mathsf{P} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}$, i.e,

$$\widehat{\mathsf{A}}\widehat{x} = \widehat{b}, \tag{3.21}$$

where $\widehat{\mathsf{A}} := \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A}$ and $\widehat{b} := \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}b$. The well-definedness of the GMRES method applied to the deflated linear system 3.21 can now be answered with theorem 3.3 (see also [62, corollary 5.3]).

Corollary 3.5. *Let* $\mathsf{A}x = b$ *with a nonsingular operator* A *and* $b \in \mathcal{H}$*. For a subspace* $\mathcal{U} \subseteq \mathcal{H}$*, the following statements are equivalent:*

1. *For every initial guess* x_0*, the GMRES method applied to the linear system* (3.21) *is well defined.*

2. $\mathcal{U} \cap (\mathsf{A}\mathcal{U})^\perp = \{0\}$.

3. $\mathcal{H} = \mathcal{U} \oplus (\mathsf{A}\mathcal{U})^\perp$.

In particular, the above conditions are fulfilled if $\mathcal{U}$ *is an* A*-invariant subspace, i.e., if* $\mathsf{A}\mathcal{U} = \mathcal{U}$*.*

Proof. The equivalence of condition 2 in theorem 3.3 with condition 2 of corollary 3.5 becomes apparent by calculating the range and null space of the operator $\widehat{\mathsf{A}}$:

$$\begin{aligned} &\mathcal{R}(\widehat{\mathsf{A}}) = \mathcal{R}(\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}\mathsf{A}) = \mathcal{R}(\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}) = (\mathsf{A}\mathcal{U})^\perp \\ \text{and}\quad &\mathcal{N}(\widehat{\mathsf{A}}) = \mathcal{N}(\mathsf{A}\mathsf{P}_{(\mathsf{A}^\star\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}) = \mathcal{N}(\mathsf{P}_{(\mathsf{A}^\star\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}) = \mathcal{U}. \end{aligned}$$

For the equivalence of condition 2 and 3, let $U \in \mathcal{H}^m$ with $m = \dim\mathcal{U}$. Then condition 2 is equivalent to

$$0 \neq \langle Uu, \mathsf{A}U\rangle = u^\mathsf{H}\langle U, \mathsf{A}U\rangle$$

for all nonzero $u \in \mathbb{C}^m$. This is equivalent to the nonsingularity of $\langle U, \mathsf{A}U\rangle$ which is in turn equivalent to $\mathcal{H} = \mathcal{U} \oplus (\mathsf{A}\mathcal{U})^\perp$ by theorem 2.15. □

To illustrate the possibility of breakdowns of the GMRES method, two examples from [62] are given below.

Example 3.6. Consider the linear system $\mathsf{A}x = b$ with

$$\mathsf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and the Euclidean inner product $\langle\cdot,\cdot\rangle_2$. Let the deflation basis be chosen as $U = \begin{bmatrix} 1, 0 \end{bmatrix}^\mathsf{T}$, i.e., the deflation space is $\mathcal{U} = [\![U]\!]$. Then

$$\mathsf{Q}_\mathcal{U}^{\mathrm{MR}} = \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp} = \mathsf{P}_{[\![x]\!]^\perp} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \widehat{\mathsf{A}} = \mathsf{Q}_\mathcal{U}^{\mathrm{MR}}\mathsf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \widehat{b} = \mathsf{Q}_\mathcal{U}^{\mathrm{MR}} b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

If $x_0 = 0$, then $\widehat{r}_0 = \widehat{b}$ and $\widehat{\mathsf{A}}\widehat{r}_0 = 0$ and the GMRES method applied to the deflated linear system (3.21) terminates at the first iteration with the approximation x_0. Since $\widehat{\mathsf{A}}x_0 \neq \widehat{b}$, this is a breakdown of GMRES. Furthermore, the correction (3.16) does not recover the solution of the original linear system $\mathsf{A}x = b$ from x_0 because

$$\mathsf{P}_\mathcal{U}^{\mathrm{MR}} x + \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}} x_0 = U\langle \mathsf{A}U, \mathsf{A}U\rangle_2^{-1}\langle \mathsf{A}U, b\rangle_2 = U\langle \mathsf{A}U, \mathsf{A}U\rangle_2^{-1}\left\langle \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\rangle_2 = 0 \neq x.$$

In corollary 3.5, it was stated, that breakdowns cannot occur in the GMRES method applied to the deflated linear system (3.21) if $\mathcal{U}$ is A-invariant. However, the following example shows that care has also to be taken with approximate A-invariant subspaces if A is non-normal.

Example 3.7. Let $\alpha > 0$ be a small positive number. Then $v = \begin{bmatrix}0, 1, \alpha\end{bmatrix}^{\mathsf{T}}$ is an eigenvector of the matrix

$$\mathsf{A} = \begin{bmatrix} 0 & 1 & -\alpha^{-1} \\ 1 & 0 & \alpha^{-1} \\ 0 & 0 & 1 \end{bmatrix}$$

corresponding to the eigenvalue 1. Instead of v, the perturbed vector $U = \begin{bmatrix}0, 1, 0\end{bmatrix}^{\mathsf{T}}$ is chosen as a basis for the deflation space $\mathcal{U} = [\![U]\!]$. Then

$$\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}} = \mathsf{P}_{[\![e_1]\!]^\perp} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & \alpha^{-1} \\ 0 & 0 & 1 \end{bmatrix}$$

and for $x, b \in \mathbb{C}^3$ with $\mathsf{A}x = b$, the GMRES method breaks down in the first step for all $x_0 \in \{x + \beta\begin{bmatrix}1, 0, 0\end{bmatrix}^{\mathsf{T}} \mid \beta \neq 0\}$. Note that $\|U - v\|_2 = \alpha$ can be chosen arbitrarily small.

The preceding example shows that measuring the departure of $\mathcal{U}$ from being A-invariant by the norm of the vector differences is inappropriate. A more appropriate measure is the maximum angle $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})$ between the subspaces $\mathcal{U}$ and $\mathsf{A}\mathcal{U}$. Note that the condition $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ is equivalent to the conditions in corollary 3.5 (see theorem 2.15) and thus also to the well-definedness of the GMRES method applied to the deflated linear system (3.21), cf. theorem 3.3. The following theorem gathers the observations and states further properties of the deflated GMRES method.

Theorem 3.8 (Deflated GMRES, variant 1)**.** *Let $\mathsf{A}x = b$ be a linear system with a nonsingular operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and right hand side $b \in \mathcal{H}$. Furthermore, let $U \in \mathcal{H}^m$ be given such that $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ for $\mathcal{U} = [\![U]\!]$.*

Then the following holds:

1. *For all initial guesses x_0, the GMRES method applied to the linear system*

$$\widehat{\mathsf{A}}\widehat{x} = \widehat{b} \tag{3.22}$$

with $\widehat{\mathsf{A}} = \mathsf{Q}_{\mathcal{U}}^{MR}\mathsf{A}$ and $\widehat{b} = \mathsf{Q}_{\mathcal{U}}^{MR}b$ is well defined.

2. *If $\widehat{x}_n$ is the constructed iterate in the n-th step in 1., then the corrected iterate*

$$x_n := \mathsf{P}_{\mathcal{U}^\perp}^{MR}\widehat{x}_n + U\langle \mathsf{A}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{A}U, b\rangle \tag{3.23}$$

satisfies

$$r_n = b - \mathsf{A}x_n = \widehat{b} - \widehat{\mathsf{A}}\widehat{x}_n = \widehat{r}_n. \tag{3.24}$$

Proof. Item 1. is just a summary of theorem 3.3, corollary 3.5 and theorem 2.15.

The residual equality in item 2. follows with lemma 3.1 by noticing that

$$b - \mathsf{A}x_n = b - \mathsf{A}U\langle \mathsf{A}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{A}U, b\rangle - \mathsf{A}\mathsf{P}_{U^\perp}^{\mathrm{MR}}\widehat{x}_n = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}b - \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A}\widehat{x}_n = \widehat{b} - \widehat{\mathsf{A}}\widehat{x}_n.$$

□

The key observation, that $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ is a necessary and sufficient condition for GMRES to be well defined for all initial guesses, opens up another possibility for a deflated GMRES method. The condition $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})$ is also necessary and sufficient for the existence of the projection $\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$ which is used in the deflated CG method, see equation (3.3). As it turns out, the GMRES method is also well defined when it is applied to the deflated linear system (3.15) with $\mathsf{P} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$.

Theorem 3.9 (Deflated GMRES, variant 2)**.** *Let the assumptions of theorem 3.8 hold. Then the following holds:*

1. *The projection $\mathsf{Q}_{\mathcal{U}}^{CG}$ is well defined.*
2. *For all initial guesses x_0, the GMRES method applied to the linear system*
$$\widehat{\mathsf{A}}\widehat{x} = \widehat{b} \tag{3.25}$$
with $\widehat{\mathsf{A}} := \mathsf{Q}_{\mathcal{U}}^{CG}\mathsf{A}$ and $\widehat{b} = \mathsf{Q}_{\mathcal{U}}^{CG} b$ is well defined.
3. *The corrected iterates*
$$x_n := \mathsf{P}_{\mathcal{U}^\perp}^{CG}\widehat{x}_n + U\langle U, \mathsf{A}U\rangle^{-1}\langle U, b\rangle$$
satisfy
$$r_n = b - \mathsf{A}x_n = \widehat{b} - \widehat{\mathsf{A}}\widehat{x}_n = \widehat{r}_n.$$

Proof. 1. The well-definedness of the projection directly follows from the assumption $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, cf. theorem 2.14.

2. The well-definedness of GMRES follows from theorem 3.3 by calculating the range and null space of the deflated operator:
$$\begin{aligned} &\mathcal{R}(\widehat{\mathsf{A}}) = \mathcal{R}(\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}) = \mathcal{R}(\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}) = \mathcal{U}^\perp \\ \text{and} \quad &\mathcal{N}(\widehat{\mathsf{A}}) = \mathcal{N}(\mathsf{A}\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp, \mathcal{U}}) = \mathcal{N}(\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp, \mathcal{U}}) = \mathcal{U}. \end{aligned}$$

3. The proof of the residual equality is analogous to the proof for the CG method in theorem 3.2.

□

The second variant of the deflated GMRES variant is rarely used in the literature. Erlangga and Nabben [50] used a more general projection P for the deflated system (3.15) which is defined by $\mathsf{P}x = x - \mathsf{A}Z\langle Y, \mathsf{A}Z\rangle^{-1}\langle Y, x\rangle$. For $Z = Y$, the method coincides with the second deflated GMRES variant in theorem 3.9. Yeung, Tang and Vuik [188] also used this deflated GMRES variant and presented an analysis for the case where $\mathcal{U}$ is an exact A-invariant subspace.

3.2.5. MINRES method and breakdowns

In cases where the original operator A is self-adjoint but possibly indefinite, the MINRES method is attractive, see section 2.9.2. Since GMRES is mathematically equivalent to MINRES for a self-adjoint operator, all results for GMRES carry over to MINRES if the deflated operator is self-adjoint. However, as outlined in this subsection, also a non-self-adjoint deflated operator is feasible due to the special structure of the involved Krylov subspace.

Kilmer and de Sturler [94] applied the MINRES method to the deflated linear system (3.22), i.e., with the projected operator $\widehat{\mathsf{A}} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A}$. In this case, the operator $\widehat{\mathsf{A}}$ is in general *not* self-adjoint, even if A is self-adjoint. However, as pointed out in [94, footnote on p. 2153], it turns out that

$$\mathcal{K}_n(\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A}, \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}v) = \mathcal{K}_n(\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A}\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}, \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}v)$$

holds for every vector $v \in \mathcal{H}$. Now the operator $\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A}\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}$ is self-adjoint because both A and $\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}$ are self-adjoint. Thus the Krylov subspace that is constructed in the MINRES method when it is applied to the linear system (3.22) is implicitly generated by a self-adjoint operator and the MINRES method is well defined for this linear system under the same condition as the GMRES method, i.e., $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$.

Corollary 3.10 (Deflated MINRES variants)**.** *Let the assumptions of theorem 3.8 hold and let* A *be self-adjoint.*

Then the statements in theorems 3.8 and 3.9 hold if GMRES is replaced by MINRES.

Proof. That the MINRES method is well defined when it is applied to the deflated linear system (3.22) has been pointed out in the discussion preceding this theorem. For the second variant, analogous to the deflated CG method, it can be shown that the deflated operator $\widehat{\mathsf{A}} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}\mathsf{A}$ is self-adjoint, see equation (3.19). □

Because the matrix in example 3.6 is Hermitian, this example also serves as an example of a breakdown of the first deflated MINRES variant if the condition $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ is violated.

In [130], Olshanskii and Simoncini used an augmented and deflated MINRES method where deflation is achieved with the second deflated MINRES variant, i.e., the MINRES method applied to the deflated linear system (3.25). However, they included an additional explicit augmentation in the MINRES algorithm. The author and Schlömer used the second deflated MINRES variant in [63] in order to solve a sequence of linear systems that stems from nonlinear Schrödinger equations. This application is discussed in detail in chapter 4.

3.2.6. Equivalence of deflation and augmentation

In the literature, many deflated Krylov subspace methods also incorporate some sort of explicit augmentation. One example is the deflated and augmented version of the

CG algorithm by Saad, Yeung, Erhel and Guyomarc'h [143] where an approximate solution x_n is sought such that the Galerkin conditions

$$x_n \in x_0 + \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0) + \mathcal{U}$$
$$r_n = b - \mathsf{A}x_n \perp \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0) + \mathcal{U}$$

hold with $\widehat{\mathsf{A}} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}\mathsf{A}$ and $\widehat{r}_0 = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}(b - \mathsf{A}x_0)$. However, they showed that the proposed deflated and augmented method is mathematically equivalent to the plain CG method applied to the deflated linear system (3.17) *without* augmentation but with an adapted initial guess, cf. section 3.2.7. A natural question is whether also other deflated methods implicitly augment the search space.

In [62], the author, Gutknecht, Liesen and Nabben showed that a (Petrov–) Galerkin method can in general be augmented by either explicitly enlarging the search space as in in equations (3.12)–(3.13), or implicitly by applying an appropriate projection to the residuals and correcting the thus computed approximate solutions.

The result is given here in its original form [62, theorem 3.2] before it is recast in a more accessible form for the two special cases of the deflated CG method (see theorem 3.2) and the first variant of the deflated GMRES and MINRES method (see theorem 3.8 and corollary 3.10).

Theorem 3.11. *Let* $\mathsf{A}x = b$ *be a consistent linear system with linear operator* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *and right hand side and initial guess* $b, x_0 \in \mathcal{H}$. *Furthermore, let* $\mathsf{B} \in \mathcal{L}(\mathcal{H})$ *and* $U \in \mathcal{H}^m$ *be such that* $\langle \mathsf{B}U, \mathsf{A}U\rangle$ *is nonsingular and let* $\widehat{\mathsf{A}} \in \mathcal{L}(\mathcal{H})$, $\widehat{v} \in \mathcal{H}$ *and* $n \le d(\widehat{\mathsf{A}}, \widehat{v})$.

Then, with $\mathcal{U} = [\![U]\!]$, *the following two pairs of conditions*

$$\left.\begin{aligned} x_n &\in x_0 + \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v}) + \mathcal{U}, \\ r_n &= b - \mathsf{A}x_n \perp \mathsf{B}\mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v}) + \mathsf{B}\mathcal{U} \end{aligned}\right\} \tag{3.26}$$

and

$$\left.\begin{aligned} \widehat{x}_n &\in x_0 + \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v}), \\ \widehat{r}_n &= \mathsf{P}_{(\mathsf{B}\mathcal{U})^\perp, \mathsf{A}\mathcal{U}}(b - \mathsf{A}\widehat{x}_n) \perp \mathsf{B}\mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v}) \end{aligned}\right\} \tag{3.27}$$

are equivalent in the sense that

$$x_n = \mathsf{P}_{(\mathsf{A}^\star\mathsf{B}\mathcal{U})^\perp, \mathcal{U}}\widehat{x}_n + U\langle \mathsf{B}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{B}U, b\rangle \qquad \textit{and} \qquad r_n = \widehat{r}_n. \tag{3.28}$$

Proof. The proof starts with the first pair of conditions (3.26) and shows the equivalence to the second pair (3.27). Let the approximate solution be

$$x_n = x_0 + V_n v_n + U u_n \tag{3.29}$$

for a $V_n \in \mathcal{H}^n$ with $[\![V_n]\!] = \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v})$, $v_n \in \mathbb{C}^n$ and $u_n \in \mathbb{C}^m$. In order to satisfy the residual constraint in (3.26), the residual $r_n = b - \mathsf{A}x_n$ must be orthogonal to both $\mathsf{B}\mathcal{K}_n(\widehat{\mathsf{A}},\widehat{v})$ and $\mathsf{B}\mathcal{U}$. Thus the residual constraint is equivalent to the pair of orthogonality conditions

$$r_n \perp \mathsf{B}\mathcal{K}_n(\widehat{\mathsf{A}},\widehat{v}) \quad \text{and} \quad r_n \perp \mathsf{B}\mathcal{U}. \tag{3.30}$$

With $r_0 = b - \mathsf{A}x_0$, the second condition is equivalent to

$$0 = \langle \mathsf{B}U, r_n\rangle = \langle \mathsf{B}U, r_0 - \mathsf{A}V_n v_n - \mathsf{A}U u_n\rangle = \langle \mathsf{B}U, r_0 - \mathsf{A}V_n v_n\rangle - \langle \mathsf{B}U, \mathsf{A}U\rangle u_n.$$

By assumption, $\langle \mathsf{B}U, \mathsf{A}U\rangle$ is nonsingular and thus

$$u_n = \langle \mathsf{B}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{B}U, r_0 - \mathsf{A}V_n v_n\rangle.$$

Substituting this into equation (3.29) yields

$$\begin{aligned} x_n &= x_0 + V_n v_n + U\langle \mathsf{B}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{B}U, r_0 - \mathsf{A}V_n v_n\rangle \\ &= x_0 + V_n v_n - U\langle \mathsf{B}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{B}U, \mathsf{A}(x_0 + V_n v_n)\rangle + U\langle \mathsf{B}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{B}U, b\rangle \\ &= \mathsf{P}_{(\mathsf{A}^\star\mathsf{A}\mathcal{U})^\perp,\mathcal{U}}(x_0 + V_n v_n) + U\langle \mathsf{B}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{B}U, b\rangle \qquad (3.31) \\ \text{and} \qquad r_n &= b - \mathsf{A}U\langle \mathsf{B}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{B}U, b\rangle - \mathsf{A}\mathsf{P}_{(\mathsf{A}^\star\mathsf{B}\mathcal{U})^\perp,\mathcal{U}}(x_0 + V_n v_n) \\ &= \mathsf{P}_{(\mathsf{B}\mathcal{U})^\perp,\mathsf{A}\mathcal{U}}(b - \mathsf{A}(x_0 + V_n v_n)). \qquad (3.32) \end{aligned}$$

After defining

$$\widehat{x}_n := x_0 + V_n v_n \in x_0 + \mathcal{K}_n(\widehat{\mathsf{A}},\widehat{v})$$

the first orthogonality condition in (3.30) can be imposed on the residual (3.32):

$$\mathsf{P}_{(\mathsf{B}\mathcal{U})^\perp,\mathsf{A}\mathcal{U}}(b - \mathsf{A}\widehat{x}_n) \perp \mathsf{B}\mathcal{K}_n(\widehat{\mathsf{A}},\widehat{v}).$$

The last two conditions now yield the second pair of conditions (3.27) and the correction (3.28) follows from equation (3.31). □

For $\mathsf{B} = \mathsf{id}$, $\widehat{\mathsf{A}} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}\mathsf{A}$ and $\widehat{v} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}(b - \mathsf{A}x_0)$, the second pair of conditions (3.27) represents the Galerkin conditions for the deflated CG method in theorem 3.2 and the correction (3.28) is the corresponding correction (3.18). This observation is registered in the following corollary.

Corollary 3.12. *With the assumptions and notation from theorem 3.2 and a solution $x \in \mathcal{H}$ of $\mathsf{A}x = b$, the corrected iterates x_n generated by the deflated CG method satisfy*

$$x_n = x_0 + \mathsf{P}^{CG}_{\mathcal{K}_n(\widehat{\mathsf{A}},\widehat{r}_0)+\mathcal{U}}(x - x_0)$$

or, equivalently,

$$\|x - x_n\|_{\mathsf{A}} = \min_{z \in x_0 + \mathcal{K}_n(\widehat{\mathsf{A}},\widehat{r}_0)+\mathcal{U}} \|x - z\|_{\mathsf{A}}.$$

As mentioned before, the fact that the deflated CG method implicitly performs augmentation was already recognized by Saad, Yeung, Erhel and Guyomarc'h [143].

An analogous result can be stated for the deflated GMRES and MINRES methods. For $\mathsf{B} = \mathsf{A}$, the orthogonality constraint in (3.27) is equivalent to

$$\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}(b - \mathsf{A}\widehat{x}_n) \perp \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{B}\mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v})$$

because $\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}} = \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}$ is an orthogonal projection and thus self-adjoint. With $\widehat{\mathsf{A}} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\mathsf{A}$ and $\widehat{v} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}(b - \mathsf{A}x_0)$, the conditions (3.27) thus represent the Petrov–Galerkin conditions for the first variant of the deflated GMRES method in theorem 3.8 and the correction (3.28) equals the correction (3.23). The following corollary gathers the made observations for the GMRES method.

Corollary 3.13. *With the assumptions and notation from theorem 3.8 and a solution $x \in \mathcal{H}$ of $\mathsf{A}x = b$, the corrected iterates x_n generated by the first variant of the deflated GMRES method satisfy*

$$x_n = x_0 + \mathsf{P}^{MR}_{\mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0) + \mathcal{U}}(x - x_0) \tag{3.33}$$

or, equivalently,

$$\|b - \mathsf{A}x_n\| = \min_{z \in x_0 + \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0) + \mathcal{U}} \|b - \mathsf{A}z\| .$$

The statement also holds for the first variant of the deflated MINRES method in the case of a self-adjoint operator A.

In [169], de Sturler introduced the GCRO method which is a nested Krylov subspace method with an outer GCR method [45, 44] and an inner GMRES method. Inside each iteration of the outer GCR method, a special subspace $\mathcal{U}$ is determined which is then used as a deflation space in the deflated GMRES method (first variant). In the setting of GCRO, the implicit augmentation of the deflated GMRES method has already been shown in [169, theorem 2.2]. That the augmentation is also carried out in general has been shown in [62] thanks to the equivalence theorem 3.11.

Wang, de Sturler and Paulino introduced the RMINRES method [181] in order to solve sequences of linear systems with self-adjoint operators. Essentially, the RMINRES method consists of two main parts that can be analyzed separately: an augmented and deflated MINRES method which is based on the GCRO method and an extraction procedure for harmonic Ritz vectors. Here, the augmented and deflated MINRES method is called the *solver part* of the RMINRES method. The solver part of the RMINRES method is presented as a rather intricate algorithm [181, algorithm 2] and essentially seeks to find the approximate solution x_n with minimal residual norm from an augmented and deflated Krylov subspace, i.e., the approximate solution (3.33). Corollary 3.13 states that this can be achieved by simply applying the standard MINRES algorithm to a deflated linear system, cf. the first variant of deflated MINRES in corollary 3.10.

Remark 3.14. Note that the second variant of the deflated GMRES and MINRES method (see theorem 3.8 and corollary 3.10) does not fit into the equivalence theorem because the orthogonality condition then reads

$$\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}(b - \mathsf{A}\widehat{x}_n) \perp \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}\mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v}),$$

which is in general incompatible with the orthogonality constraint in the second pair of conditions (3.27).

3.2.7. Implementation

In the previous subsection, it was shown that augmentation is implicitly achieved by the deflated CG method and the first variant of the deflated GMRES and MINRES methods. In these methods, the standard CG, GMRES or MINRES algorithms are applied to a deflated linear system and the approximate solutions are corrected afterwards. Regarding the implementation, such a pure deflated Krylov subspace method has clear advantages because the algorithm of the underlying Krylov subspace method (e.g., CG, GMRES or MINRES) do not have to be modified and one can draw on the most robust already existing algorithms and their implementations. In terms of programming, deflation can thus be seen as a wrapper around Krylov subspace methods.

For all deflated methods that were discussed above, it is also possible to start with a corrected initial guess and use a projection as a right "preconditioner" which makes the post-correction superfluous. The difference is only of algorithmic nature and is described briefly for the deflated methods discussed in this thesis. First, the concept of right preconditioning is recapitulated.

For a linear system $\mathsf{A}x = b$ and an invertible linear operator $\mathsf{M} \in \mathcal{L}(\mathcal{H})$, the right preconditioned system $\mathsf{AM}y = b$ can be solved for y and then the original solution can be obtained from $x = \mathsf{M}y$. Instead of x_0, the initial guess $y_0 := \mathsf{M}^{-1}x_0$ is used implicitly and the initial residual $r_0 = b - \mathsf{AM}y_0 = b - \mathsf{A}x_0$ equals the residual of the unpreconditioned system. For the CG, GMRES and MINRES methods, iterates of the form

$$y_n = y_0 + z_n \quad \text{with} \quad z_n \in \mathcal{K}_n(\mathsf{AM}, r_0)$$

and $x_n := \mathsf{M}y_n = x_0 + \mathsf{M}z_n$ are constructed such that $\|x - x_n\|_\mathsf{A}$ or $\|b - \mathsf{A}x_n\|$ are minimal. For well-definedness, the preconditioned operator AM has to be self-adjoint and positive definite for the CG method and self-adjoint for the MINRES method. Note that y_0 is not needed and will never be computed explicitly.

Assume that two projections $\mathsf{Q}, \mathsf{P} \in \mathcal{L}(\mathcal{H})$ are given such that $\mathsf{QA} = \mathsf{AP}$ and let CG, GMRES or MINRES be well defined when applied to the projected linear system $\widehat{\mathsf{A}}\widehat{x} = \widehat{b}$ with $\widehat{\mathsf{A}} = \mathsf{QA}$ and $\widehat{b} = \mathsf{Q}b$. The cases that are of interest here are:

1. $\mathsf{Q} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$ and $\mathsf{P} = \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}}$ (with the CG, GMRES or MINRES method).
2. $\mathsf{Q} = \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}$ and $\mathsf{P} = \mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}}$ (with the GMRES or MINRES method).

For an initial guess x_0 with corresponding initial residual $\widehat{r}_0 = \widehat{b} - \widehat{\mathsf{A}}x_0$ and $n \leq d(\widehat{\mathsf{A}}, \widehat{r}_0)$, iterates $\widehat{x}_n \in x_0 + \widehat{\mathcal{K}}_n$ are constructed where $\widehat{\mathcal{K}}_n := \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0)$. If $V_n \in \mathcal{H}^n$ with $[\![V_n]\!] = \widehat{\mathcal{K}}_n$, then the iterates satisfy

$$\widehat{x}_n = x_0 + V_n \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}V_n \right\rangle^{-1} \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}(x - x_0) \right\rangle$$

with $\mathsf{B} = \mathsf{id}$ for the CG method and $\mathsf{B} = \widehat{\mathsf{A}}$ for the GMRES and MINRES methods. The iterates then have to be corrected according to

$$x_n = \mathsf{P}\widehat{x}_n + (\mathsf{id} - \mathsf{P})x.$$

Now a closer look is taken at right preconditioning with $\mathsf{M} = \mathsf{P}$ which differs from the above description because P is singular in general. However, even if the right preconditioned system

$$\mathsf{AP}\tilde{y} = b \tag{3.34}$$

is not consistent (i.e., $b \notin \mathcal{R}(\mathsf{AP})$), the right preconditioning strategy can be used to solve the original linear system if the initial guess

$$\tilde{x}_0 := \mathsf{P}x_0 + (\mathsf{id} - \mathsf{P})x \tag{3.35}$$

is used. The key issues are that $\mathsf{AP} = \widehat{\mathsf{A}}$ and that the initial residual for the right preconditioned system is computed as

$$\tilde{r}_0 = b - \mathsf{A}\tilde{x}_0 = b - \mathsf{AP}x_0 - \mathsf{A}(\mathsf{id} - \mathsf{P})x = b - \mathsf{QA}x_0 - (\mathsf{id} - \mathsf{Q})\mathsf{A}x = \mathsf{Q}(b - \mathsf{A}x_0) = \widehat{b} - \widehat{\mathsf{A}}x_0.$$

Thus, the Krylov subspace that is generated by the CG, GMRES or MINRES method when it is applied to the right preconditioned system (3.34) with the initial guess (3.35) actually equals the Krylov subspace $\widehat{\mathcal{K}}_n$ which is constructed in the corresponding method when it is applied to the linear system $\widehat{\mathsf{A}}\widehat{x} = \widehat{b}$ with initial guess x_0. The methods thus are also well defined and the generated iterates are given by

$$\begin{aligned}
\tilde{x}_n &= \tilde{x}_0 + \mathsf{P}V_n \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}V_n \right\rangle^{-1} \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}(x - \tilde{x}_0) \right\rangle \\
&= \tilde{x}_0 + \mathsf{P}V_n \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}V_n \right\rangle^{-1} \langle \mathsf{B}V_n, \mathsf{AP}(x - \mathsf{P}x_0 - (\mathsf{id} - \mathsf{P})x) \rangle \\
&= \tilde{x}_0 + \mathsf{P}V_n \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}V_n \right\rangle^{-1} \langle \mathsf{B}V_n, \mathsf{AP}(x - x_0) \rangle \\
&= \tilde{x}_0 + \mathsf{P}V_n \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}V_n \right\rangle^{-1} \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}(x - x_0) \right\rangle \\
&= \mathsf{P}x_0 + (\mathsf{id} - \mathsf{P})x + \mathsf{P}V_n \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}V_n \right\rangle^{-1} \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}(x - x_0) \right\rangle \\
&= \mathsf{P}\left(x_0 + V_n \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}V_n \right\rangle^{-1} \left\langle \mathsf{B}V_n, \widehat{\mathsf{A}}(x - x_0) \right\rangle \right) + (\mathsf{id} - \mathsf{P})x \\
&= \mathsf{P}\widehat{x}_n + (\mathsf{id} - \mathsf{P})x.
\end{aligned}$$

The equivalent methods are gathered in table 3.1 and the following corollary.

Method	Projections		Initial guess	Final iterate
	P_l	P_r	$\overline{x}_0$	x_n
CG (theorem 3.2)	$\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$	–	x_0	$c^{\mathrm{CG}}(\overline{x}_n)$
	–	$\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}}$	$c^{\mathrm{CG}}(x_0)$	$\overline{x}_n$
GMRES/MINRES variant 1	$\mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}$	–	x_0	$c^{\mathrm{MR}}(\overline{x}_n)$
(theorem 3.8, corollary 3.10)	–	$\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{MR}}$	$c^{\mathrm{MR}}(x_0)$	$\overline{x}_n$
GMRES/MINRES variant 2	$\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}$	–	x_0	$c^{\mathrm{CG}}(\overline{x}_n)$
(theorem 3.9, corollary 3.10)	–	$\mathsf{P}_{\mathcal{U}^\perp}^{\mathrm{CG}}$	$c^{\mathrm{CG}}(x_0)$	$\overline{x}_n$

Table 3.1.: Overview of well-defined deflated Krylov subspace methods, see corollary 3.15. Implementations can be found in [60] under `krypy.deflation.Deflated{Cg,Minres,Gmres}`.

Corollary 3.15 (Overview of well-defined deflated Krylov subspace methods)**.** *Let a linear system $\mathsf{A}x = b$ with a nonsingular operator A and right hand side and initial guess $b, x_0 \in \mathcal{H}$ be given. Furthermore, let a deflation space basis $U \in \mathcal{H}^m$ be given such that $\mathcal{U} = [\![U]\!]$ is m-dimensional with $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ and let $c^{CG}, c^{MR} : \mathcal{H} \longrightarrow \mathcal{H}$ be defined by*

$$
\begin{aligned}
& c^{CG}(z) = \mathsf{P}_{\mathcal{U}^\perp}^{CG} z + U\langle U, \mathsf{A}U\rangle^{-1}\langle U, b\rangle \\
\textit{and} \qquad & c^{MR}(z) = \mathsf{P}_{\mathcal{U}^\perp}^{MR} z + U\langle \mathsf{A}U, \mathsf{A}U\rangle^{-1}\langle \mathsf{A}U, b\rangle .
\end{aligned}
$$

Considering table 3.1, the listed Krylov subspace methods are well defined when they are applied to the deflated linear system

$$\mathsf{P}_l \mathsf{A} \mathsf{P}_r y = \mathsf{P}_l b$$

with initial guess $\overline{x}_0$ and yield iterates $\overline{x}_n$ from which the final approximation x_n can be obtained. The operator A has to be self-adjoint and positive definite for the CG method and self-adjoint for the MINRES method. The right preconditioner P_r is treated as described in section 3.2.7.

In a brief algorithmic interlude, the situation of a sequence of linear systems (3.1) is considered here again. In some cases, good candidates for deflation subspaces are known for the problem at hand (see chapter 4). However, in general, the goal is to determine and improve suitable deflation subspaces on the fly while solving the linear systems one after another. A prototype algorithm for the solution of the sequence of linear systems is given in algorithm 3.1. In each iteration, the deflation vectors $U^{(i)}$ are chosen as the auxiliary vectors $Z^{(i)}$ and a set of vectors from the previous Krylov subspace $\mathcal{K}^{(i-1)}$ and the span of the previous deflation vectors $U^{(i-1)}$. The algorithm deliberately leaves open the question of how to exactly choose the deflation subspace for the next linear system in line 4. Selection strategies for

Algorithm 3.1 Prototype of recycling Krylov subspace methods for sequences of linear systems. Implemented as `krypy.recycling.Recycling{Cg,Minres,Gmres}` in [60].

```
Input: For $i \in \{1,\ldots,M\}$, the following is assumed to be given for solving the
    sequence of linear systems (3.1):
      • $\mathsf{A}^{(i)} \in \mathcal{L}(\mathcal{H})$                                ▷ operator
      • $b^{(i)}, x_0^{(i)} \in \mathcal{H}$                     ▷ right hand side and initial guess
      • $Z^{(i)} \in \mathcal{H}^{l^{(i)}}$ for $l^{(i)} \in \mathbb{N}$       ▷ auxiliary deflation vectors (may be empty)
$W^{(1)} = [\ ] \in \mathcal{H}^0$         ▷ no deflation vectors can be determined in the first step
for $i = 1,\ldots,M$ do
    if $i > 1$ then
        Choose $W^{(i)} \in \left(\mathcal{K}^{(i-1)} + [\![U^{(i-1)}]\!]\right)^{k^{(i)}}$ for a $k^{(i)} \in \mathbb{N}$.
        | Only information from the previous linear system is used, e.g., the
        | Arnoldi relation for the Krylov subspace $\mathcal{K}^{(i-1)}$, $U^{(i-1)}$ and character-
        | izations of the differences $\mathsf{A}^{(i)} - \mathsf{A}^{(i-1)}$, $b^{(i)} - b^{(i-1)}$ and $x_0^{(i)} - x_0^{(i-1)}$.
        | The selection of such vectors is discussed in section 3.4.
    end if
    $U^{(i)} = [W^{(i)}, Z^{(i)}]$                                ▷ new deflation subspace
    if $\langle U^{(i)}, \mathsf{A}^{(i)}U^{(i)} \rangle$ is singular then  ▷ solvability condition $\theta(\mathcal{U}, \mathsf{A}^{(i)}\mathcal{U}) < \frac{\pi}{2}$
        $U^{(i)} \leftarrow U^{(i)}\mathbf{X}$
        | Choose $\mathbf{X} \in \mathbb{C}^{k^{(i)}+l^{(i)},m}$ such that $\langle U^{(i)}, \mathsf{A}^{(i)}U^{(i)} \rangle$ is nonsingular.
    end if
    Solve deflated linear system to given tolerance with one of the methods in
    table 3.1 and the deflation space $\mathcal{U} = [\![U^{(i)}]\!]$.
    | The underlying Krylov subspace method is applied to $\mathsf{P}_l\mathsf{A}^{(i)}x^{(i)} = \mathsf{P}_l b^{(i)}$
    | or $\mathsf{A}^{(i)}\mathsf{P}_r y^{(i)} = b^{(i)}$ with the appropriate correction of the initial guess or
    | approximate solution. In the course of the method, an Arnoldi relation
    | for the Krylov subspace $\mathcal{K}^{(i)} = \mathcal{K}_{n^{(i)}}(\mathsf{P}_l\mathsf{A}^{(i)}, \mathsf{P}_l r_0^{(i)})$ is constructed with
    | $r_0^{(i)} = b^{(i)} - \mathsf{A}^{(i)}x_0^{(i)}$.
end for
```

the deflation subspace are the subject of section 3.4. Algorithm 3.1 can be seen as a straightforward building block for the strategies that are about to be outlined later in this thesis. Algorithms along the lines of algorithm 3.1 have been used without the safety check in line 7, e.g., by Kilmer and de Sturler [94], Parks et al. [135] and the author and Schlömer [63].

3.3. Perturbation theory

The goal of this subsection is to investigate the convergence behavior of the deflated Krylov subspace methods that have been presented in section 3.2. For a given deflation subspace $\mathcal{U} \subseteq \mathcal{H}$ with $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, all of the presented deflated methods essentially consist of the application of the CG, GMRES or MINRES method to a projected linear system

$$\mathsf{PA}\widehat{x} = \mathsf{P}b$$

with $\mathsf{P} \in \{\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}, \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}\}$. For a self-adjoint and positive-definite operator A and the projection $\mathsf{P} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$, it was shown by Vuik, Nabben and Tang [178] that the effective condition number $\kappa_{\text{eff}}(\mathsf{PA})$ (see definition 2.55) satisfies[1]

$$\kappa_{\text{eff}}(\mathsf{PA}) \leq \kappa(\mathsf{A}) \tag{3.36}$$

for all deflation spaces $\mathcal{U} \neq \mathcal{H}$. This implies that the asymptotic κ-bound for the CG method (cf. theorem 2.56) cannot grow for any deflation space $\mathcal{U}$. The effective condition number $\kappa_{\text{eff}}(\mathsf{PA})$ can be quantified easily if $\mathcal{U}$ is A-invariant. Assume that the spectrum of A is $\Lambda(\mathsf{A}) = \{\lambda_1, \ldots, \lambda_N\}$ and that $\mathcal{U}$ is an m-dimensional invariant subspace associated with the eigenvalues $\{\lambda_1, \ldots, \lambda_m\}$. Then the spectrum of PA is

$$\Lambda(\mathsf{PA}) = \{0, \lambda_{m+1}, \ldots, \lambda_N\}$$

and the effective condition number of PA thus is

$$\kappa_{\text{eff}}(\mathsf{PA}) = \frac{\max_{i \in \{m+1,\ldots,N\}} \lambda_i}{\min_{i \in \{m+1,\ldots,N\}} \lambda_i}.$$

Of course, an exact A-invariant subspace is not available in practice but only approximations, e.g., the span of Ritz or harmonic Ritz vectors, cf. section 2.5. In the deflation and augmentation literature [43, 143], it is sometimes argued that an almost A-invariant subspace will not result in substantial differences of the spectrum or the convergence behavior.

To the knowledge of the author, no meaningful quantifications of the impact of deflation space perturbations on the spectrum of the deflated operator or the convergence behavior of deflated methods have been established in the literature.

[1] It was stated in [178, 173] that $\kappa_{\text{eff}}(\mathsf{PA}) < \kappa(\mathsf{A})$. The example $\mathsf{A} = \mathbf{I}_2$ and $\mathcal{U} = [\![e_1]\!]$, where $\kappa_{\text{eff}}(\mathsf{PA}) = \kappa(\mathsf{A}) = 1$, shows that the strict inequality does not hold in general. However, inequality (3.36) is true which can be seen easily from corollary 3.25 and the Cauchy interlacing property.

When studying perturbations for deflated Krylov subspace methods, several topics that are interesting and partly unexplored themselves lie along the way. These topics include perturbation theory for projections, deflated operators and Krylov subspace methods. The theory of this section is also of importance in section 3.4 which deals with the determination of deflation spaces in the setting of a sequence of linear systems.

3.3.1. Projections

This subsection complements the results on projections and angles between subspaces from sections 2.2 and 2.3 with a new perturbation result for projections. In this subsection, the Hilbert space $\mathcal{H}$ may be infinite-dimensional. Given two projections $\mathsf{P}, \mathsf{Q} \in \mathcal{L}(\mathcal{H})$, the question is: how can the norm

$$\|\mathsf{P} - \mathsf{Q}\|$$

be characterized? For special cases, this problem has been addressed in the literature and a brief overview is given here. Since a projection is uniquely defined by its range and null space, it is assumed that four closed and nonzero subspaces $\mathcal{V}, \mathcal{W}, \mathcal{X}, \mathcal{Y} \subseteq \mathcal{H}$ are given such that $\mathcal{V} \oplus \mathcal{W} = \mathcal{X} \oplus \mathcal{Y} = \mathcal{H}$. Then with $\mathsf{P} = \mathsf{P}_{\mathcal{V},\mathcal{W}}$ and $\mathsf{Q} = \mathsf{P}_{\mathcal{X},\mathcal{Y}}$ the task is to express or bound

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \tag{3.37}$$

in terms of angles between the involved subspaces. Note that this subsection makes heavy use of the statements on projections in section 2.2 and on angles and gaps between subspaces in section 2.3.

If both projections are orthogonal, i.e., $\mathcal{W} = \mathcal{V}^\perp$ and $\mathcal{Y} = \mathcal{X}^\perp$, then the norm of the difference is by definition 2.21 the sine of the maximal canonical angle between $\mathcal{V}$ and $\mathcal{X}$:

$$\|\mathsf{P}_{\mathcal{V}} - \mathsf{P}_{\mathcal{X}}\| = \sin\theta_{\max}(\mathcal{V}, \mathcal{X}) = \Theta(\mathcal{V}, \mathcal{X}).$$

With $\mathcal{W} = \mathcal{Y}$, a more interesting situation was considered by Berkson [13] in the more general setting of a Banach space. In the situation of a Hilbert space and with the notation used in this thesis, he showed in [13, theorem 5.2] that

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{W}}\| \le \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \frac{\mu}{1-\mu} \quad \text{with} \quad \mu = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|\, \Theta(\mathcal{V}, \mathcal{X}) \tag{3.38}$$

holds if $\mu < 1$. The terms in the right hand side of inequality (3.38) can be expressed in terms of angles between the involved subspaces by using the identities from lemma 2.23 and definition 2.21:

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{W}}\| \le \frac{1}{\cos\theta_{\max}(\mathcal{V}, \mathcal{W}^\perp)} \frac{\mu}{1-\mu} \quad \text{with} \quad \mu = \frac{\sin\theta_{\max}(\mathcal{V}, \mathcal{X})}{\cos\theta_{\max}(\mathcal{V}, \mathcal{W}^\perp)}.$$

Dirr, Rakočević and Wimmer sharpened the bound in [33, theorem 3.1] by showing that

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{W}}\| \le \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \frac{\eta}{1-\eta} \quad \text{with} \quad \eta = \|\mathsf{P}_{\mathcal{X}^\perp}\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \tag{3.39}$$

holds if $\eta < 1$. In general, η cannot be directly represented in terms of angles and is often not available in practice. However, η can be bounded by

$$\eta = \|\mathsf{P}_{\mathcal{X}^\perp}\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \|\mathsf{P}_{\mathcal{X}^\perp}\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \le \|\mathsf{P}_{\mathcal{X}^\perp}\mathsf{P}_{\mathcal{V}}\|\,\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|\,\Theta(\mathcal{V},\mathcal{X}) = \mu$$

which leads to the bound (3.38) of Berkson. This relation between the bounds has already been established in [33]. Though sharper, the bound (3.39) is not considered in the following for the above reason.

In this subsection, a bound for the general problem (3.37) with four independent subspaces $\mathcal{V},\mathcal{W},\mathcal{X},\mathcal{Y} \subseteq \mathcal{H}$ satisfying $\mathcal{V} \oplus \mathcal{W} = \mathcal{X} \oplus \mathcal{Y} = \mathcal{H}$ is presented. The bound appears to be new and it is shown that the new bound is sharper than Berkson's bound (3.38) in the special case $\mathcal{W} = \mathcal{Y}$. As a preparation for the proof of the bound in the upcoming theorem 3.17, the following lemma is stated.

Lemma 3.16. *Let $\mathcal{V},\mathcal{W},\mathcal{X},\mathcal{Y} \subseteq \mathcal{H}$ be closed and nonzero subspaces with $\mathcal{V} \oplus \mathcal{W} = \mathcal{X} \oplus \mathcal{Y} = \mathcal{H}$. Then*

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \le \frac{\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)}.$$

Proof. The lemma follows from statements 3, 4 and 6 of lemma 2.23 and statement 3 of lemma 2.18:

$$\begin{aligned}\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| &= \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{X}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \le \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|\,\|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{X}}\|\,\|\mathsf{P}_{\mathcal{X},\mathcal{Y}}\|\\ &= \frac{\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)}.\end{aligned}$$

□

The bound in lemma 3.16 is sharp. For an arbitrary Hilbert space $\mathcal{H}$ with $\dim\mathcal{H} \ge 2$ (for $\dim\mathcal{H} < 2$ one of the involved subspaces has to be zero) this can be seen by choosing four vectors $v,w,x,y \in \mathcal{H}$ of norm 1 with $\langle v,w\rangle \ne 0$ and $\langle x,y\rangle \ne 0$. Let the four subspaces be defined by $\mathcal{V} = [\![v]\!]$, $\mathcal{W} = [\![w]\!]^\perp$, $\mathcal{X} = [\![x]\!]$ and $\mathcal{Y} = [\![y]\!]^\perp$. With $z \in \mathcal{H}$ and theorem 2.15, the projections can then be represented by

$$\mathsf{P}_{\mathcal{V},\mathcal{W}}z = v\frac{\langle w,z\rangle}{\langle w,v\rangle} \qquad \text{and} \qquad \mathsf{P}_{\mathcal{X},\mathcal{Y}}z = x\frac{\langle y,z\rangle}{\langle y,x\rangle}.$$

Then

$$\begin{aligned}\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| &= \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}y\| = \frac{1}{|\langle y,x\rangle|}\|\mathsf{P}_{\mathcal{V},\mathcal{W}}x\| = \frac{|\langle w,x\rangle|}{|\langle w,v\rangle|\,|\langle y,x\rangle|}\\ &= \frac{\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{X})}{\cos\theta_{\min}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\min}(\mathcal{X},\mathcal{Y}^\perp)}.\end{aligned}$$

At this point, statement 5 of lemma 2.18 (because $\dim\mathcal{W}^\perp = 1$) and statement 5 of lemma 2.23 (because $\mathcal{H} = \mathcal{V} \oplus \mathcal{W}$) can be applied in order to obtain

$$\cos\theta_{\min}(\mathcal{V},\mathcal{W}^\perp) = \sin\theta_{\min}(\mathcal{V},\mathcal{W}) = \cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp).$$

An analogous argument shows that $\cos\theta_{\min}(\mathcal{X},\mathcal{Y}^\perp) = \cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)$ and the bound is attained because

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| = \frac{\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)}.$$

Now this section's main result can be shown. In order to serve the applications in later sections, the theorem bounds $\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\|$, where $\mathsf{L} \in \mathcal{L}(\mathcal{H})$ is an arbitrary linear operator. Obviously, a bound for the original problem $\|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}}\|$ results with $\mathsf{L} = \mathsf{id}$.

Theorem 3.17. *Let $\mathcal{V},\mathcal{W},\mathcal{X},\mathcal{Y} \subseteq \mathcal{H}$ be closed and nonzero subspaces with $\mathcal{V}\oplus\mathcal{W} = \mathcal{X}\oplus\mathcal{Y} = \mathcal{H}$ and let $\mathsf{L} \in \mathcal{L}(\mathcal{H})$. Then*

$$\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| \le \frac{\|\mathsf{L}|_{\mathcal{V}}\|\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{Y}) + \|\mathsf{L}|_{\mathcal{W}}\|\cos\theta_{\min}(\mathcal{V}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)}.$$

Proof. The theorem is proved by decomposing the identity operator as $\mathsf{id} = \mathsf{P}_{\mathcal{V},\mathcal{W}} + \mathsf{P}_{\mathcal{W},\mathcal{V}}$:

$$\begin{aligned}
\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| &= \|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}}) + \mathsf{L}\mathsf{P}_{\mathcal{W},\mathcal{V}}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| \\
&= \|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}(\mathsf{id} - \mathsf{P}_{\mathcal{X},\mathcal{Y}}) - \mathsf{L}\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&= \|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}} - \mathsf{L}\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&\le \|\mathsf{L}\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| + \|\mathsf{L}\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&= \|\mathsf{L}\mathsf{P}_{\mathcal{V}}\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| + \|\mathsf{L}\mathsf{P}_{\mathcal{W}}\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&\le \|\mathsf{L}\mathsf{P}_{\mathcal{V}}\|\,\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| + \|\mathsf{L}\mathsf{P}_{\mathcal{W}}\|\,\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&= \|\mathsf{L}|_{\mathcal{V}}\|\,\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| + \|\mathsf{L}|_{\mathcal{W}}\|\,\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&\le \frac{\|\mathsf{L}|_{\mathcal{V}}\|\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{Y})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{Y},\mathcal{X}^\perp)} \\
&\quad + \frac{\|\mathsf{L}|_{\mathcal{W}}\|\cos\theta_{\min}(\mathcal{V}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{W},\mathcal{V}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)} \\
&= \frac{\|\mathsf{L}|_{\mathcal{V}}\|\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{Y}) + \|\mathsf{L}|_{\mathcal{W}}\|\cos\theta_{\min}(\mathcal{V}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)}.
\end{aligned}$$

□

The bound in theorem 3.17 is attained for special choices of the subspaces where some of the terms disappear. However, it is shown that the bound can be attained up to a factor for a rather technical construction of subspaces.

Let $\mathsf{L} \in \mathcal{L}(\mathcal{H})$ have a bounded inverse and let $\mathcal{Z}_1,\mathcal{Z}_2,\mathcal{Z}_3 \subseteq \mathcal{H}$ be closed subspaces such that $\mathcal{H} = \mathcal{Z}_1\oplus\mathcal{Z}_2\oplus\mathcal{Z}_3$, $\dim\mathcal{Z}_3 = 2$ and $\mathcal{Z}_i \perp \mathcal{Z}_j$ for $i,j \in \{1,2,3\}$ and $i \ne j$. Furthermore, let $v,w,x,y \in \mathcal{Z}_3$ such that $\mathcal{Z}_3 = [\![v]\!]\oplus[\![x]\!] = [\![w]\!]\oplus[\![y]\!]$ and $\langle v,w\rangle, \langle x,y\rangle \ne 0$. Then the following subspaces are defined:

$$\begin{aligned}
\mathcal{V} &:= \mathcal{Z}_1 \oplus [\![v]\!], & \mathcal{W} &:= (\mathcal{Z}_1 \oplus w)^\perp, \\
\mathcal{X} &:= (\mathcal{Z}_2 \oplus [\![x]\!])^\perp \quad\text{and} & \mathcal{Y} &:= \mathcal{Z}_2 \oplus y.
\end{aligned} \tag{3.40}$$

For this constellation of subspaces, it is shown that

$$\begin{aligned}\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| &\le \frac{\|\mathsf{L}|_{\mathcal{V}}\| \cos\theta_{\min}(\mathcal{W}^\perp, \mathcal{Y}) + \|\mathsf{L}|_{\mathcal{W}}\| \cos\theta_{\min}(\mathcal{V}^\perp, \mathcal{X})}{\cos\theta_{\max}(\mathcal{V}, \mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X}, \mathcal{Y}^\perp)} \\ &\le 2\kappa(\mathsf{L})\,\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\|\,\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\|\,.\end{aligned} \tag{3.41}$$

The two summands in the right hand side of (3.41) are treated separately. With lemma 2.8 it can be seen that $\mathsf{P}_{\mathcal{V},\mathcal{W}}$ and $\mathsf{P}_{\mathcal{Y},\mathcal{X}}$ are well defined and are given by

$$\mathsf{P}_{\mathcal{V},\mathcal{W}} = \mathsf{P}_{\mathcal{Z}_1} + \mathsf{P}_{[\![v]\!],[\![w]\!]^\perp} \qquad \text{and} \qquad \mathsf{P}_{\mathcal{Y},\mathcal{X}} = \mathsf{P}_{\mathcal{Z}_2} + \mathsf{P}_{[\![y]\!],[\![x]\!]^\perp}.$$

Because of the orthogonality conditions a direct computation shows that

$$\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}} = \mathsf{P}_{[\![v]\!],[\![w]\!]^\perp}\mathsf{P}_{[\![y]\!],[\![x]\!]^\perp}.$$

In analogy to the discussion following lemma 3.16, the following equality holds:

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| = \frac{\cos\theta_{\min}([\![w]\!],[\![y]\!])}{\cos\theta_{\max}([\![v]\!],[\![w]\!])\cos\theta_{\max}([\![y]\!],[\![x]\!])}.$$

Lemma 2.8, lemma 2.18 and lemma 2.20 now yield

$$\begin{aligned}\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{Y}) &= \|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{Y}}\| = \left\|(\mathsf{P}_{\mathcal{Z}_1} + \mathsf{P}_{[\![w]\!]})(\mathsf{P}_{\mathcal{Z}_2} + \mathsf{P}_{[\![y]\!]})\right\| = \left\|\mathsf{P}_{[\![w]\!]}\mathsf{P}_{[\![y]\!]}\right\| \\ &= \cos\theta_{\min}([\![w]\!],[\![y]\!]), \\ \sin\theta_{\max}(\mathcal{V},\mathcal{W}^\perp) &= \|\mathsf{P}_{\mathcal{V}} - \mathsf{P}_{\mathcal{W}^\perp}\| = \left\|\mathsf{P}_{\mathcal{Z}_1} + \mathsf{P}_{[\![v]\!]} - \mathsf{P}_{\mathcal{Z}_1} - \mathsf{P}_{[\![w]\!]}\right\| \\ &= \left\|\mathsf{P}_{[\![v]\!]} - \mathsf{P}_{[\![w]\!]}\right\| = \sin\theta_{\max}([\![v]\!],[\![w]\!]), \\ \sin\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp) &= \sin\theta_{\max}(\mathcal{X}^\perp,\mathcal{Y}) = \|\mathsf{P}_{\mathcal{X}^\perp} - \mathsf{P}_{\mathcal{Y}}\| = \left\|\mathsf{P}_{\mathcal{Z}_2} + \mathsf{P}_{[\![x]\!]} - \mathsf{P}_{\mathcal{Z}_2} - \mathsf{P}_{[\![y]\!]}\right\| \\ &= \left\|\mathsf{P}_{[\![x]\!]} - \mathsf{P}_{[\![y]\!]}\right\| = \sin\theta_{\max}([\![x]\!],[\![y]\!])\end{aligned}$$

and thus

$$\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| = \frac{\cos\theta_{\min}(\mathcal{W}^\perp,\mathcal{Y})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{Y}^\perp)}. \tag{3.42}$$

The right hand side of equation (3.42) is the first summand in (3.41) and now the second summand is considered. Therefore, let $\widehat{v},\widehat{w},\widehat{x},\widehat{y} \in \mathcal{Z}_3$ be nonzero such that $\widehat{v} \perp v$, $\widehat{w} \perp w$, $\widehat{x} \perp x$ and $\widehat{y} \perp y$. The orthogonal complements of the subspaces defined in (3.40) then are

$$\begin{aligned}\mathcal{V}^\perp &= \mathcal{Z}_2 \oplus [\![\widehat{v}]\!], & \mathcal{W}^\perp &= (\mathcal{Z}_2 \oplus [\![\widehat{w}]\!])^\perp, \\ \mathcal{X}^\perp &= (\mathcal{Z}_1 \oplus [\![\widehat{x}]\!])^\perp \quad \text{and} & \mathcal{Y}^\perp &= \mathcal{Z}_1 \oplus [\![\widehat{y}]\!].\end{aligned}$$

Similar arguments as above lead to

$$\begin{aligned}\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| &= \left\|\mathsf{P}^\star_{\mathcal{X},\mathcal{Y}}\mathsf{P}^\star_{\mathcal{W},\mathcal{V}}\right\| = \left\|\mathsf{P}_{\mathcal{Y}^\perp,\mathcal{X}^\perp}\mathsf{P}_{\mathcal{V}^\perp,\mathcal{W}^\perp}\right\| = \left\|\mathsf{P}_{[\![\widehat{y}]\!],[\![\widehat{x}]\!]^\perp}\mathsf{P}_{[\![\widehat{v}]\!],[\![\widehat{w}]\!]^\perp}\right\| \\ &= \frac{\cos\theta_{\min}([\![\widehat{x}]\!],[\![\widehat{v}]\!])}{\cos\theta_{\max}([\![\widehat{y}]\!],[\![\widehat{x}]\!])\cos\theta_{\max}([\![\widehat{v}]\!],[\![\widehat{w}]\!])}\end{aligned}$$

$$= \frac{\cos\theta_{\min}(\mathcal{V}^\perp, \mathcal{X})}{\cos\theta_{\max}(\mathcal{V}, \mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X}, \mathcal{Y}^\perp)}.$$

The last line of the preceding equation is the second term in (3.41) and the following estimate results:

$$\begin{aligned}
\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| &\le \frac{\|\mathsf{L}|_{\mathcal{V}}\| \cos\theta_{\min}(\mathcal{W}^\perp, \mathcal{Y}) + \|\mathsf{L}|_{\mathcal{W}}\| \cos\theta_{\min}(\mathcal{V}^\perp, \mathcal{X})}{\cos\theta_{\max}(\mathcal{V}, \mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X}, \mathcal{Y}^\perp)} \\
&= \|\mathsf{L}|_{\mathcal{V}}\| \, \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| + \|\mathsf{L}|_{\mathcal{W}}\| \, \|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&\le \|\mathsf{L}\| \left(\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\mathsf{P}_{\mathcal{Y},\mathcal{X}}\| + \|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X},\mathcal{Y}}\|\right) \\
&= \|\mathsf{L}\| \left(\|\mathsf{P}_{\mathcal{V},\mathcal{W}}(\mathsf{id} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| + \|\mathsf{P}_{\mathcal{W},\mathcal{V}}(\mathsf{id} - \mathsf{P}_{\mathcal{Y},\mathcal{X}})\|\right) \\
&= \|\mathsf{L}\| \left(\|\mathsf{P}_{\mathcal{V},\mathcal{W}}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| + \|\mathsf{P}_{\mathcal{W},\mathcal{V}}(\mathsf{P}_{\mathcal{W},\mathcal{V}} - \mathsf{P}_{\mathcal{Y},\mathcal{X}})\|\right) \\
&\le \|\mathsf{L}\| \left(\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \, \|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}}\| + \|\mathsf{P}_{\mathcal{W},\mathcal{V}}\| \, \|\mathsf{P}_{\mathcal{W},\mathcal{V}} - \mathsf{P}_{\mathcal{Y},\mathcal{X}}\|\right) \\
&= 2\,\|\mathsf{L}\| \, \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \, \|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \\
&\le 2\,\|\mathsf{L}\| \, \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \, \left\|\mathsf{L}^{-1}\right\| \, \|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| \\
&= 2\kappa(\mathsf{L}) \, \|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| \, \|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| \, .
\end{aligned}$$

In an application of theorem 3.17 (cf. lemma 3.30) the first projection is orthogonal, i.e., $\mathcal{W} = \mathcal{V}^\perp$, and thus $\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\| = 1$.

In practice, the cosines in the numerator of the bound in theorem 3.17 are often not available. In the following corollary, the cosines are replaced with the sines of the maximal angle between the ranges and the maximal angle between the null spaces of the projections.

Corollary 3.18. *Let the assumptions of theorem 3.17 hold. Then*

$$\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| \le \frac{\|\mathsf{L}|_{\mathcal{V}}\| \sin\theta_{\max}(\mathcal{W}, \mathcal{Y}) + \|\mathsf{L}|_{\mathcal{W}}\| \sin\theta_{\max}(\mathcal{V}, \mathcal{X})}{\cos\theta_{\max}(\mathcal{V}, \mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X}, \mathcal{Y}^\perp)}.$$

Proof. By item 3 of lemma 2.18, item 4 of lemma 2.20 and definition 2.21 the statement follows from:

$$\begin{aligned}
\cos\theta_{\min}(\mathcal{W}^\perp, \mathcal{Y}) &= \|\mathsf{P}_{\mathcal{W}^\perp}\mathsf{P}_{\mathcal{Y}}\| \le \sin\theta_{\max}(\mathcal{W}, \mathcal{Y}) \\
\cos\theta_{\min}(\mathcal{V}^\perp, \mathcal{X}) &= \|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{X}}\| \le \sin\theta_{\max}(\mathcal{V}, \mathcal{X}).
\end{aligned}$$

Note that equality holds if $\mathcal{W}^\perp \oplus \mathcal{Y} = \mathcal{V}^\perp \oplus \mathcal{X} = \mathcal{H}$ by item 5 of lemma 2.23. □

The bound in corollary 3.18 is visualized in figure 3.1 with $\mathcal{H} = \mathbb{R}^2$ and the Euclidean inner product. An interactive version of this figure is available online[2].

Remark 3.19. Note that if the roles of $\mathcal{V}$ and $\mathcal{X}$ as well as $\mathcal{W}$ and $\mathcal{Y}$ are interchanged in theorem 3.17, then the bound can be improved to

$$\|\mathsf{L}(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})\| \le \frac{\min\{\eta_1, \eta_2\}}{\cos\theta_{\max}(\mathcal{V}, \mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X}, \mathcal{Y}^\perp)},$$

[2] `http://andrenarchy.github.io/talk-2013-01-projections/`

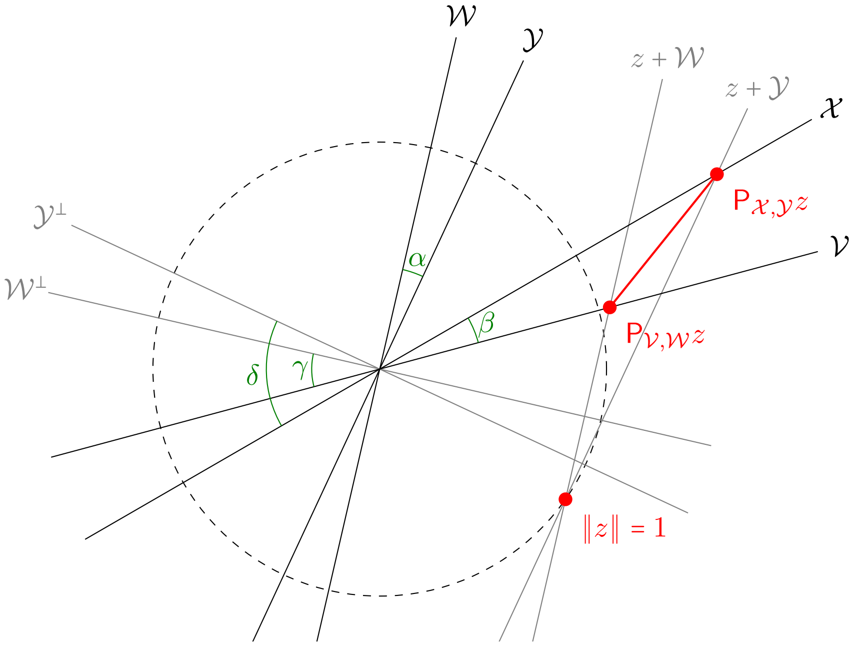


Figure 3.1.: Illustration of corollary 3.18 with $\mathcal{H} = \mathbb{R}^2$, the Euclidean inner product, $\mathsf{L} = \mathsf{id}$ and one-dimensional subspaces $\mathcal{V}, \mathcal{W}, \mathcal{X}$ and $\mathcal{Y}$. The projections $\mathsf{P}_{\mathcal{V},\mathcal{W}}$ and $\mathsf{P}_{\mathcal{X},\mathcal{Y}}$ are applied to a normalized vector z. The theorem guarantees that $\|(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})z\| \le \frac{\sin\alpha + \sin\beta}{\cos\gamma\cos\delta}$. The angle α measures how close the null spaces $\mathcal{W}$ and $\mathcal{Y}$ are and β measures how close the ranges $\mathcal{V}$ and $\mathcal{X}$ are. The angles γ and δ indicate the departure of the projections from orthogonal projections. In this example $\|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}}\| \approx 0.7559$, $\|(\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{Y}})z\| \approx 0.7534$ and $\frac{\sin\alpha + \sin\beta}{\cos\gamma\cos\delta} \approx 0.9216$.

$$\begin{aligned} \text{where} \qquad & \eta_1 = \|\mathsf{L}|_{\mathcal{V}}\| \cos\theta_{\min}(\mathcal{W}^\perp, \mathcal{Y}) + \|\mathsf{L}|_{\mathcal{W}}\| \cos\theta_{\min}(\mathcal{V}^\perp, \mathcal{X}) \\ \text{and} \qquad & \eta_2 = \|\mathsf{L}|_{\mathcal{X}}\| \cos\theta_{\min}(\mathcal{W}, \mathcal{Y}^\perp) + \left\|\mathsf{L}|_{\mathcal{Y}}\right\| \cos\theta_{\min}(\mathcal{V}, \mathcal{X}^\perp). \end{aligned}$$

Analogously, the same holds true for the bound in corollary 3.18 where

$$\begin{aligned} & \eta_1 = \|\mathsf{L}|_{\mathcal{V}}\| \sin\theta_{\max}(\mathcal{W}, \mathcal{Y}) + \|\mathsf{L}|_{\mathcal{W}}\| \sin\theta_{\max}(\mathcal{V}, \mathcal{X}) \\ \text{and} \qquad & \eta_2 = \|\mathsf{L}|_{\mathcal{X}}\| \sin\theta_{\max}(\mathcal{W}, \mathcal{Y}) + \left\|\mathsf{L}|_{\mathcal{Y}}\right\| \sin\theta_{\max}(\mathcal{V}, \mathcal{X}). \end{aligned}$$

A remaining question is how the bound in theorem 3.17 or corollary 3.18 relates to the bound (3.38) by Berkson in the special case $\mathcal{W} = \mathcal{Y}$. The following lemma shows that the bound (3.38) by Berkson is weaker than the new bounds in theorem 3.17 or corollary 3.18.

Lemma 3.20. *Let $\mathcal{V}, \mathcal{W}, \mathcal{X} \subseteq \mathcal{H}$ be closed and nonzero subspaces with $\mathcal{V} \oplus \mathcal{W} =$*

$\mathcal{X} \oplus \mathcal{W} = \mathcal{H}$. If

$$\mu = \frac{\sin\theta_{\max}(\mathcal{V},\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)} < 1,$$

then

$$\begin{aligned}\left\|\mathsf{P}_{\mathcal{V},\mathcal{W}} - \mathsf{P}_{\mathcal{X},\mathcal{W}}\right\| &\le \frac{\cos\theta_{\min}(\mathcal{V}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{W}^\perp)}\\ &\le \frac{1}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)}\frac{\mu}{1-\mu}.\end{aligned}$$

Proof. The first inequality directly follows from theorem 3.17 and only the second inequality remains to show. Because of item 3 of lemma 2.18 and item 6 of lemma 2.23, the following holds:

$$\begin{aligned}\frac{\cos\theta_{\min}(\mathcal{V}^\perp,\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)\cos\theta_{\max}(\mathcal{X},\mathcal{W}^\perp)} &= \frac{\left\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{X}}\right\|\left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}\right\|}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)} \le \frac{\sin\theta_{\max}(\mathcal{V},\mathcal{X})}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)}\left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}\right\|\\ &= \mu\left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}\right\|. \end{aligned} \tag{3.43}$$

The projection $\mathsf{P}_{\mathcal{X},\mathcal{W}}$ can be represented by

$$\mathsf{P}_{\mathcal{X},\mathcal{W}} = \mathsf{P}_{\mathcal{X},\mathcal{W}}(\mathsf{P}_{\mathcal{V},\mathcal{W}} + \mathsf{P}_{\mathcal{W},\mathcal{V}}) = \mathsf{P}_{\mathcal{X},\mathcal{W}}\mathsf{P}_{\mathcal{V},\mathcal{W}} = \mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}}\,\mathsf{P}_{\mathcal{V},\mathcal{W}}. \tag{3.44}$$

With item 1 of lemma 2.23 the restricted operator $\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}} : \mathcal{V} \longrightarrow \mathcal{X}$ can be expressed as

$$\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}} = \mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{W}^\perp}\,\mathsf{P}_{\mathcal{W}^\perp}|_{\mathcal{V}} = \mathsf{Q}_{\mathcal{X},\mathcal{W}}^{-1}\mathsf{Q}_{\mathcal{V},\mathcal{W}}.$$

Thus, it is invertible and the inverse is given by

$$\left(\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}}\right)^{-1} = \mathsf{Q}_{\mathcal{V},\mathcal{W}}^{-1}\mathsf{Q}_{\mathcal{X},\mathcal{W}} = \mathsf{P}_{\mathcal{V},\mathcal{W}}\,\mathsf{P}_{\mathcal{W}^\perp}|_{\mathcal{X}} = \mathsf{P}_{\mathcal{V},\mathcal{W}}|_{\mathcal{X}}.$$

Then the norm of $\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}}$ can be estimated as follows:

$$\begin{aligned}\left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}}\right\| &= \sup_{0\ne v\in\mathcal{V}}\frac{\left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}v\right\|}{\|v\|} = \sup_{0\ne x\in\mathcal{X}}\frac{\left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}}\,\mathsf{P}_{\mathcal{V},\mathcal{W}}|_{\mathcal{X}}\,x\right\|}{\left\|\mathsf{P}_{\mathcal{V},\mathcal{W}}x\right\|} = \sup_{0\ne x\in\mathcal{X}}\frac{\|x\|}{\left\|\mathsf{P}_{\mathcal{V},\mathcal{W}}x\right\|}\\ &= \sup_{0\ne x\in\mathcal{X}}\frac{\|x\|}{\left\|x - \mathsf{P}_{\mathcal{W},\mathcal{V}}x\right\|} \le \sup_{0\ne x\in\mathcal{X}}\frac{\|x\|}{\|x\| - \left\|\mathsf{P}_{\mathcal{W},\mathcal{V}}x\right\|} \le \sup_{0\ne x\in\mathcal{X}}\frac{\|x\|}{\|x\| - \left\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X}}\right\|\|x\|}\\ &= \frac{1}{1 - \left\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{X}}\right\|} \le \frac{1}{1 - \left\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{X}}\right\|} \le \frac{1}{1 - \left\|\mathsf{P}_{\mathcal{W},\mathcal{V}}\right\|\left\|\mathsf{P}_{\mathcal{V}^\perp}\mathsf{P}_{\mathcal{X}}\right\|} \le \frac{1}{1-\mu}.\end{aligned}$$

This bound together with equation (3.44) results in

$$\begin{aligned}\left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}\right\| = \left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}}\,\mathsf{P}_{\mathcal{V},\mathcal{W}}\right\| &\le \left\|\mathsf{P}_{\mathcal{X},\mathcal{W}}|_{\mathcal{V}}\right\|\left\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\right\| \le \frac{1}{1-\mu}\left\|\mathsf{P}_{\mathcal{V},\mathcal{W}}\right\|\\ &\le \frac{1}{1-\mu}\frac{1}{\cos\theta_{\max}(\mathcal{V},\mathcal{W}^\perp)}.\end{aligned}$$

The statement of the lemma then follows with equation (3.43). □

3.3.2. Spectrum of deflated operators

In this subsection, the spectrum of deflated operators is analyzed for a finite-dimensional Hilbert space $\mathcal{H}$, i.e., $N := \dim \mathcal{H} < \infty$. If an exact A-invariant subspace $\mathcal{V}$ is used as a deflation subspace, then the spectrum and the invariant subspaces of the deflated operator PA are trivial to compute for the choices of P that have been discussed in the previous sections. The following theorem has been presented in [62, theorem 3.3] and characterizes the spectrum and invariant subspaces of a deflated operator with an invariant deflation subspace.

Theorem 3.21. *Let* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *have the Jordan decomposition*

$$\mathsf{A} = \begin{bmatrix} S_1 & S_2 \end{bmatrix} \begin{bmatrix} \mathbf{J}_1 & 0 \\ 0 & \mathbf{J}_2 \end{bmatrix} \begin{bmatrix} \widehat{S}_1^\star \\ \widehat{S}_2^\star \end{bmatrix},$$

where $S_1, \widehat{S}_1 \in \mathcal{H}^m$, $S_2, \widehat{S}_2 \in \mathcal{H}^{N-m}$, $\mathbf{J}_1 \in \mathbb{C}^{m,m}$, $\mathbf{J}_2 \in \mathbb{C}^{N-m,N-m}$ *for* $m > 0$ *and* $\begin{bmatrix} \widehat{S}_1 & \widehat{S}_2 \end{bmatrix}^\star = \begin{bmatrix} S_1 & S_2 \end{bmatrix}^{-1}$. *If* $\mathcal{V} = [\![S_1]\!]$ *and* $\mathbf{J}_1$ *is nonsingular, then*

1. $\mathsf{Q}_{\mathcal{V}}^{CG} = \mathsf{Q}_{\mathcal{V}}^{MR} = \mathsf{P}_{\mathcal{V}^\perp}$.

2. *The deflated operator* $\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$ *has the Jordan decomposition*

$$\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A} = \begin{bmatrix} S_1 & \mathsf{P}_{\mathcal{V}^\perp} S_2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{J}_2 \end{bmatrix} \begin{bmatrix} S_1 & \mathsf{P}_{\mathcal{V}^\perp} S_2 \end{bmatrix}^{-1}$$

$$\textit{with} \quad \begin{bmatrix} S_1 & \mathsf{P}_{\mathcal{V}^\perp} S_2 \end{bmatrix}^{-1} = \begin{bmatrix} S_1 \langle S_1, S_1 \rangle^{-1} & \widehat{S}_2 \end{bmatrix}^\star.$$

3. $\Lambda(\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}) = \{0\} \cup \Lambda(\mathbf{J}_2)$.

Proof. 1. Because $\mathsf{A}\mathcal{V} = \mathcal{V}$ the projections $\mathsf{Q}_{\mathcal{V}}^{\mathrm{CG}} = \mathsf{P}_{\mathcal{V}^\perp, \mathsf{A}\mathcal{V}}$ and $\mathsf{Q}_{\mathcal{V}}^{\mathrm{MR}} = \mathsf{P}_{(\mathsf{A}\mathcal{V})^\perp}$ obviously equal $\mathsf{P}_{\mathcal{V}^\perp}$.

2. A direct computation shows that

$$\begin{aligned} \mathsf{P}_{\mathcal{V}^\perp}\mathsf{A} &= \begin{bmatrix} 0 & \mathsf{P}_{\mathcal{V}^\perp} S_2 \end{bmatrix} \begin{bmatrix} \mathbf{J}_1 & 0 \\ 0 & \mathbf{J}_2 \end{bmatrix} \begin{bmatrix} \widehat{S}_1^\star \\ \widehat{S}_2^\star \end{bmatrix} = \mathsf{P}_{\mathcal{V}^\perp} S_2 \mathbf{J}_2 \widehat{S}_2^\star \\ &= \begin{bmatrix} S_1 & \mathsf{P}_{\mathcal{V}^\perp} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{J}_2 \end{bmatrix} \begin{bmatrix} S_1 \langle S_1, S_1 \rangle^{-1} & \widehat{S}_2 \end{bmatrix}^\star \end{aligned}$$

and because $[\![\widehat{S}_2]\!] \perp \mathcal{V}$ also

$$\begin{aligned} \begin{bmatrix} S_1 \langle S_1, S_1 \rangle^{-1} & \widehat{S}_2 \end{bmatrix}^\star \begin{bmatrix} S_1 & \mathsf{P}_{\mathcal{V}^\perp} S_2 \end{bmatrix} &= \begin{bmatrix} \left\langle S_1 \langle S_1, S_1 \rangle^{-1}, S_1 \right\rangle & \left\langle S_1 \langle S_1, S_1 \rangle^{-1}, \mathsf{P}_{\mathcal{V}^\perp} S_2 \right\rangle \\ \left\langle \widehat{S}_2, S_1 \right\rangle & \left\langle \widehat{S}_2, \mathsf{P}_{\mathcal{V}^\perp} S_2 \right\rangle \end{bmatrix} \\ &= \begin{bmatrix} \langle S_1, S_1 \rangle^{-1} \langle S_1, S_1 \rangle & \left\langle \mathsf{P}_{\mathcal{V}^\perp} S_1 \langle S_1, S_1 \rangle^{-1}, S_2 \right\rangle \\ 0 & \left\langle \mathsf{P}_{\mathcal{V}^\perp} \widehat{S}_2, S_2 \right\rangle \end{bmatrix} \\ &= \mathbf{I}_N. \end{aligned}$$

3. Immediately follows from item 2.

□

Unlike the situation in the preceding theorem, usually only approximations to invariant subspaces can be used in practice. However, things become more complicated if only an approximation $\mathcal{U}$ to an A-invariant subspace $\mathcal{V}$ is used and not much is known in the literature about the spectrum of the deflated operator PA for $\mathsf{P} \in \{\mathsf{Q}_{\mathcal{U}}^{\mathrm{CG}}, \mathsf{Q}_{\mathcal{U}}^{\mathrm{MR}}\}$. In the introduction of section 3.3, it was stated for a self-adjoint and positive operator A, that the effective condition number of $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ is always less than or equal to the condition number of the original operator A – independent of the choice of the subspace $\mathcal{U}$. However, the next example shows that this does not hold for self-adjoint but indefinite operators.

Example 3.22. Consider the operator A and the deflation basis U defined by

$$\mathsf{A} = \begin{bmatrix} -1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \qquad \text{and} \qquad U = \begin{bmatrix} \sqrt{1+\varepsilon} \\ 1 \\ 0 \end{bmatrix}$$

with $0 < \varepsilon \ll 1$ and the Euclidean inner product. Then $\langle U, \mathsf{A}U\rangle = -\varepsilon < 0$ and with $\mathcal{U} = [\![U]\!]$ a direct computation shows that the projected operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ has the spectrum $\Lambda(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) = \{0, 1, 1 + \frac{2}{\varepsilon}\}$. The effective condition number thus is $\kappa_{\mathrm{eff}}(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) = 1 + \frac{2}{\varepsilon}$ which can become arbitrarily large.

In practice, it is often observed that an approximation $\mathcal{U}$ to an A-invariant subspace only leads to a slight difference between $\Lambda(\mathsf{P}_{\mathcal{V}^\perp,\mathsf{A}\mathcal{V}}\mathsf{A})$ and $\Lambda(\mathsf{P}_{\mathcal{U},\mathsf{A}\mathcal{U}}\mathsf{A})$ if $\mathcal{U}$ is a “good enough” approximation. Likewise, the convergence of Krylov subspace methods for linear systems is often observed to behave very similarly for the exact and the approximate invariant subspaces. The following example demonstrates this effect for the MINRES method.

Example 3.23. Let $\mathsf{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{104})$ with $\lambda_1 = -10^{-3}$, $\lambda_2 = -10^{-4}$, $\lambda_3 = -10^{-5}$ and

$$\lambda_{i+4} = 1 + \frac{i}{100} \quad \text{for} \quad i \in \{0, \ldots, 100\}$$

and let $b = [1, 1, 1, 0.1, \ldots, 0.1]^\mathsf{T}$. The convergence of MINRES applied to $\mathsf{A}x = b$ with $x_0 = 0$ and the corresponding bound (2.29) from theorem 2.63 are visualized in figure 3.2. It is observed that MINRES almost stagnates for the initial 20 steps, and that the bound (2.29) is quite descriptive in this phase. Deflating the three negative eigenvalues using $\mathcal{V} = [\![e_1, e_2, e_3]\!]$ reduces the effective condition number to $\kappa_{\mathrm{eff}}(\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}) = 2$ (note that $\mathsf{P}_{\mathcal{V}^\perp,\mathsf{A}\mathcal{V}} = \mathsf{P}_{\mathcal{V}^\perp}$) and the initial phase of slow convergence is completely removed. The faster convergence is also described by the bound (2.29). Furthermore, a similar behavior is observed for a slightly perturbed deflation space given by $\mathcal{U} = [\![[e_1, e_2, e_3] + 10^{-5}\mathbf{E}]\!]$ with a random $\mathbf{E} \in \mathbb{R}^{N,3}$ of norm 1.

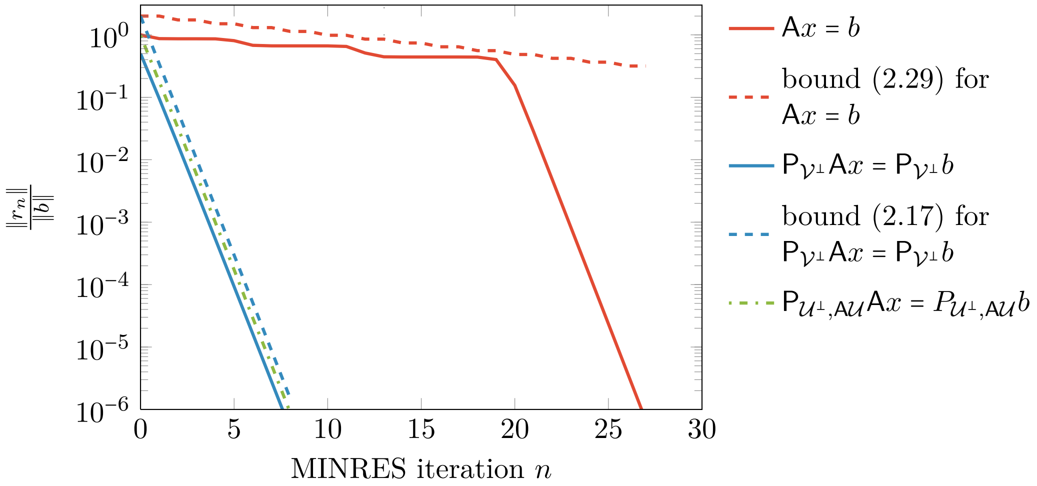


Figure 3.2.: Convergence history of MINRES with the linear system from example 3.23: without deflation, with deflation of an exact invariant subspace $\mathcal{V}$ and with deflation of an approximate invariant subspace $\mathcal{U}$. The curve for the approximate invariant subspace only differs slightly from the curve for the exact invariant subspace.

Example 3.23 clearly shows that it can be worthwhile to use an approximate invariant subspace to get rid of certain eigenvalues in a deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$. The experiment seems to indicate that a small perturbation of an invariant subspace $\mathcal{V}$ only leads to a small change in the convergence behavior of MINRES. In order to understand the underlying mechanisms mathematically, a better understanding of the spectrum of deflated operators is needed.

Already in the first article on deflation, Nicolaides [128] gave a characterization of the smallest and largest positive eigenvalues of the deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ in terms of a Rayleigh quotient in the case where A is positive definite. In [128, lemma 3.1], Nicolaides states that

$$\min \Lambda(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) \setminus \{0\} = \min_{u\in\mathcal{U}^\perp} \frac{\langle u,u\rangle}{\langle u,\mathsf{A}^{-1}u\rangle} \tag{3.45}$$

$$\text{and} \qquad \max \Lambda(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) = \max_{u\in\mathcal{U}^\perp} \frac{\langle u,u\rangle}{\langle u,\mathsf{A}^{-1}u\rangle}. \tag{3.46}$$

Inspired by this result, the following theorem shows how the *full* spectrum of the deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ can be characterized via A's inverse for *any* invertible operator A – even if A is indefinite, non-self-adjoint or non-diagonalizable. To the knowledge of the author, no similar statements exist in the literature.

Theorem 3.24. *Let* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *be nonsingular and let* $\mathcal{U} \subseteq \mathcal{H}$ *be a subspace of dimension* $m > 0$ *such that* $\sin\theta_{\max}(\mathcal{U},\mathsf{A}\mathcal{U}) < \frac{\pi}{2}$.

Then $\mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}^{-1}|_{\mathcal{U}^\perp} : \mathcal{U}^\perp \longrightarrow \mathcal{U}^\perp$ *is nonsingular and*

$$\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A} = \left(\mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}^{-1}|_{\mathcal{U}^\perp}\right)^{-1}\mathsf{P}_{\mathcal{U}^\perp}.$$

In particular, the spectrum satisfies

$$\Lambda\left(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}\right) = \{0\} \cup \Lambda\left(\left(\mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}^{-1}|_{\mathcal{U}^\perp}\right)^{-1}\right).$$

Proof. First note that

$$\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A} = \mathsf{A}\mathsf{P}_{(\mathsf{A}^\star\mathcal{U})^\perp,\mathcal{U}} = \mathsf{A}\mathsf{P}_{(\mathsf{A}^\star\mathcal{U})^\perp,\mathcal{U}}\mathsf{P}_{\mathcal{U}^\perp} = \mathsf{A}\mathsf{P}_{(\mathsf{A}^\star\mathcal{U})^\perp,\mathcal{U}}|_{\mathcal{U}^\perp}\mathsf{P}_{\mathcal{U}^\perp}.$$

Then item 1 of lemma 2.23 can be applied by noticing that

$$\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}|_{\mathcal{U}^\perp} = \mathsf{A}\mathsf{P}_{(\mathsf{A}^\star\mathcal{U})^\perp,\mathcal{U}}|_{\mathcal{U}^\perp} = \mathsf{A}\mathsf{Q}^{-1}_{(\mathsf{A}^\star\mathcal{U})^\perp,\mathcal{U}} : \mathcal{U}^\perp \longrightarrow \mathcal{U}^\perp$$

is nonsingular:

$$\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A} = \mathsf{A}\mathsf{Q}^{-1}_{(\mathsf{A}^\star\mathcal{U})^\perp,\mathcal{U}}\mathsf{P}_{\mathcal{U}^\perp} = \left(\mathsf{Q}_{(\mathsf{A}^\star\mathcal{U})^\perp,\mathcal{U}}\mathsf{A}^{-1}|_{\mathcal{U}^\perp}\right)^{-1}\mathsf{P}_{\mathcal{U}^\perp} = \left(\mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}^{-1}|_{\mathcal{U}^\perp}\right)^{-1}\mathsf{P}_{\mathcal{U}^\perp}.$$

The statement regarding the spectrum follows from the fact that both $\mathcal{U}$ and $\mathcal{U}^\perp$ are $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$-invariant subspaces. □

The following corollary recasts the above result on the spectrum of the deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ in a more accessible form for the case where an orthonormal basis of $\mathcal{U}^\perp$ is at hand.

Corollary 3.25. *Let the assumptions of theorem 3.24 hold. Furthermore, assume that* $U_\perp \in \mathcal{H}^{N-m}$ *is given with* $\mathcal{U}^\perp = [\![U_\perp]\!]$ *and* $\langle U_\perp, U_\perp\rangle = \mathbf{I}_{N-m}$.

Then

$$\Lambda\left(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}\right) = \{0\} \cup \Lambda\left(\left\langle U_\perp, \mathsf{A}^{-1}U_\perp\right\rangle^{-1}\right). \tag{3.47}$$

Proof. The statement immediately follows from theorem 3.24 by representing the projection $\mathsf{P}_{\mathcal{U}^\perp}$ in terms of $U_\perp$:

$$\mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}^{-1}|_{\mathcal{U}^\perp}U_\perp = \mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}^{-1}U_\perp = U_\perp\left\langle U_\perp, \mathsf{A}^{-1}U_\perp\right\rangle.$$

□

In the case of a positive-definite operator A, it becomes apparent from equation (3.47) and an application of Rayleigh quotients, that Nicolaides' characterizations (3.45)–(3.46) of the smallest and largest positive eigenvalue of $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ are special cases of the above results.

For a self-adjoint operator A, it is interesting to know how the number of positive and negative eigenvalues of A relate to the ones in $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$.

Definition 3.26 (Inertia)**.** For a self-adjoint operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, the *inertia* of A is defined as

$$\operatorname{In}(\mathsf{A}) = (n_-(\mathsf{A}), n_0(\mathsf{A}), n_+(\mathsf{A}))\,,$$

where $n_-(\mathsf{A})$, $n_0(\mathsf{A})$ and $n_+(\mathsf{A})$ denotes the number of negative, zero and positive eigenvalues of A.

The following theorem gives a precise characterization of $\operatorname{In}(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A})$, which appears to be new.

Theorem 3.27. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be self-adjoint and let $U \in \mathcal{H}^m$ be such that $\mathcal{U} = [\![U]\!]$ is an m-dimensional subspace with $\sin\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$.*

Then

$$\operatorname{In}(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) = \operatorname{In}(\mathsf{A}) - \operatorname{In}(\langle U, \mathsf{A}U\rangle) + (0, m, 0).$$

Proof. The proof's central element is the application of the Haynsworth inertia identity [79]. For a partitioned Hermitian matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12}^{\mathsf{H}} & \mathbf{H}_{22} \end{bmatrix} \in \mathbb{C}^{n,n}$$

with a nonsingular submatrix $\mathbf{H}_{11}$, the Haynsworth inertia identity states that

$$\operatorname{In}(\mathbf{H}) = \operatorname{In}(\mathbf{H}_{11}) + \operatorname{In}\left(\mathbf{H}_{22} - \mathbf{H}_{12}^{\mathsf{H}}\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\right).$$

Assume that $U_\perp \in \mathcal{H}^{N-m}$ is given such that $\mathcal{U}^\perp = [\![U_\perp]\!]$ and $\langle U_\perp, U_\perp\rangle = \mathbf{I}_{N-m}$. Because $\mathcal{U}$ is an invariant subspace of $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$, it follows that

$$\operatorname{In}(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) = (0, m, 0) + \operatorname{In}\left(\langle U_\perp, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}U_\perp\rangle\right). \tag{3.48}$$

By applying the Haynsworth identity to the matrix

$$\mathbf{H} = \left\langle \begin{bmatrix} U & U_\perp \end{bmatrix}, \mathsf{A}\begin{bmatrix} U & U_\perp \end{bmatrix} \right\rangle = \begin{bmatrix} \langle U, \mathsf{A}U, \rangle & \langle U, \mathsf{A}U_\perp\rangle \\ \langle U_\perp, \mathsf{A}U, \rangle & \langle U_\perp, \mathsf{A}U_\perp\rangle \end{bmatrix},$$

the following equation can be obtained:

$$\begin{aligned} \operatorname{In}(\mathsf{A}) = \operatorname{In}(\mathsf{H}) &= \operatorname{In}(\langle U, \mathsf{A}U\rangle) + \operatorname{In}\left(\langle U_\perp, \mathsf{A}U_\perp\rangle - \langle U_\perp, \mathsf{A}U\rangle\langle U, \mathsf{A}U\rangle^{-1}\langle U, \mathsf{A}U_\perp\rangle\right) \\ &= \operatorname{In}(\langle U, \mathsf{A}U\rangle) + \operatorname{In}\left(\langle U_\perp, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}U_\perp\rangle\right). \end{aligned}$$

Note that the first equality also holds for non-orthonormal U because of Sylvester's law of inertia. The statement then follows with equation (3.48). □

Theorem 3.27 can be helpful in practice, e.g., if a self-adjoint and indefinite operator A is given that exhibits $0 < m \ll N$ negative eigenvalues and $N-m$ positive eigenvalues. If an $U \in \mathcal{H}^m$ is known such that $\langle U, \mathsf{A}U\rangle$ has m negative eigenvalues, then with $\mathcal{U} = [\![U]\!]$ the deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ has the inertia $(0, m, N-m)$ and thus is positive semidefinite. This property is independent of the closeness of $\mathcal{U}$ to an A-invariant subspace. In such a case, the CG method can be used with the deflated linear system (cf. section 3.2) instead of the MINRES method in order to solve a given linear system with the indefinite operator A. Although no positive eigenvalue of A can turn into a negative eigenvalue of $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ in such a situation, care has to be taken because nothing can be said about how close a positive eigenvalue can come to zero or how large they can grow. Recall that it was demonstrated in example 3.22 that – unlike in the positive-definite case – the effective condition number $\kappa_{\text{eff}}(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A})$ is not bounded by $\kappa(\mathsf{A})$ in the self-adjoint and indefinite case and can grow arbitrarily if no further restrictions are imposed on U.

In the remaining part of this subsection, bounds for the spectrum of a deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ are developed in the case of a self-adjoint operator A. In order to prove the apparently new bounds, some well-known auxiliary perturbation results are needed and are recalled in the course of this subsection. If knowledge about the spectrum $\Lambda(\mathsf{A})$ of the original operator A is at hand, the perturbation bounds can be used in order to estimate the spectrum $\Lambda(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A})$ of a deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ if an approximate invariant subspace $\mathcal{U}$ is used as the deflation subspace. For the deflated CG and MINRES methods, these spectral estimations then may give insight into the convergence behavior of these methods, cf. the convergence bounds in sections 2.8 and 2.9.2.

For the non-self-adjoint case, the perturbation theory is more intricate and the effects of non-normality can be disastrous. The breakdown in example 3.7 may serve as a warning sign for the sometimes counter-intuitive behavior of non-normal operators. Furthermore, the spectrum may have no influence on the convergence behavior of Krylov subspace methods like the GMRES method, cf. the discussion following theorem 2.58.

In the following, only the minimal required subset of spectral perturbation results for self-adjoint operators is presented. An extensive treatment of spectral perturbation theory can, e.g., be found in the books by Bhatia [15], Stewart and Sun [167], Demmel [31] and Parlett [138].

Given two self-adjoint linear operators $\mathsf{L}, \mathsf{E} \in \mathcal{L}(\mathcal{H})$, the question of how the eigenvalues of $\widehat{\mathsf{L}} = \mathsf{L} + \mathsf{E}$ relate to the eigenvalues of L and E is, e.g., answered by Weyl's theorem:

Theorem 3.28 (Weyl)**.** *Let $\mathsf{L}, \mathsf{E} \in \mathcal{L}(\mathcal{H})$ be two self-adjoint linear operators and let $\widehat{\mathsf{L}} := \mathsf{L} + \mathsf{E}$. The (real) eigenvalues of L, $\widehat{\mathsf{L}}$ and E are denoted by $\lambda_1 \leq \ldots \leq \lambda_N$, $\widehat{\lambda}_1 \leq \ldots \leq \widehat{\lambda}_N$ and $\epsilon_1 \leq \ldots \leq \epsilon_N$.*

Then for $i \in \{1, \ldots, N\}$

$$\widehat{\lambda}_i \in [\lambda_i + \epsilon_1, \lambda_i + \epsilon_N].$$

Proof. See [167, corollary 4.9]. □

Weyl's theorem is often stated in the following weaker form:

Corollary 3.29. *With the assumptions and notation of theorem 3.28 the following holds:*

$$\max_{i\in\{1,\ldots,N\}} |\widehat{\lambda}_i - \lambda_i| \leq \|\mathsf{E}\| \,.$$

Proof. The statement directly follows from theorem 3.28 since $\|\mathsf{E}\| = \max\{|\epsilon_1|, |\epsilon_N|\}$. □

Assume that the spectrum of A is known and that an m-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}$ is given such that $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, i.e., such that the projection $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}$ is well defined. The question now is: how can bounds on the spectrum of $\widehat{\mathsf{A}} = \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ be established?

A naive approach would be to choose $\mathsf{L} = \mathsf{A}$ and $\mathsf{E} = \mathsf{P}_{\mathsf{A}\mathcal{U},\mathcal{U}^\perp}\mathsf{A}$ in Weyl's theorem. However, m eigenvalues will be moved to zero in the deflated operator $\widehat{\mathsf{A}}$ and thus the perturbation's norm $\|\mathsf{E}\|$ can grow exceedingly.

A more reasonable idea is to choose $\mathsf{L} = \mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$ and $\mathsf{E} = \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A} - \mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$, where $\mathcal{V}$ is an A-invariant subspace that is supposed to be "close" to the subspace $\mathcal{U}$. The "closeness" is quantified in terms of an angle or a residual in the course of this subsection. The next lemma is an intermediate result that makes use of Weyl's theorem and the new bound on the norm of the difference of two projections, cf. corollary 3.18.

Lemma 3.30. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be self-adjoint and $\mathcal{V} \subseteq \mathcal{H}$ an A-invariant subspace with*

$$\rho_{\mathcal{V}} := \rho\left(\mathsf{A}|_{\mathcal{V}}\right) = \max_{\lambda\in\Lambda(\mathsf{A}|_{\mathcal{V}})} |\lambda| \quad \text{and} \quad \rho_{\mathcal{V}^\perp} := \rho\left(\mathsf{A}|_{\mathcal{V}^\perp}\right) = \max_{\lambda\in\Lambda(\mathsf{A}|_{\mathcal{V}^\perp})} |\lambda|.$$

Furthermore, let $\mathcal{U} \subseteq \mathcal{H}$ be a subspace such that $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ and let the eigenvalues of $\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$ and $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ be sorted as

$$\overline{\lambda}_1 \leq \ldots \leq \overline{\lambda}_N \quad \text{and} \quad \widehat{\lambda}_1 \leq \ldots \leq \widehat{\lambda}_N.$$

Then:

$$\max_{i\in\{1,\ldots,N\}} |\widehat{\lambda}_i - \overline{\lambda}_i| \leq \frac{\rho_{\mathcal{V}} \sin\theta_{\max}(\mathcal{V}, \mathsf{A}\mathcal{U}) + \rho_{\mathcal{V}^\perp} \sin\theta_{\max}(\mathcal{V},\mathcal{U})}{\cos\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})}. \tag{3.49}$$

Proof. Let $\mathsf{L} = \mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$ and $\mathsf{E} = \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A} - \mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$. Then with corollary 3.18 the following bound holds:

$$\begin{aligned}\|\mathsf{E}\| &= \left\|\mathsf{A}(\mathsf{P}_{\mathcal{V}^\perp} - \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp,\mathcal{U}})\right\| \\ &\leq \frac{\|\mathsf{A}|_{\mathcal{V}}\| \sin\theta_{\max}(\mathcal{V}^\perp, (\mathsf{A}\mathcal{U})^\perp) + \|\mathsf{A}|_{\mathcal{V}^\perp}\| \sin\theta_{\max}(\mathcal{V},\mathcal{U})}{\cos\theta_{\max}(\mathcal{V},\mathcal{V}) \cos\theta_{\max}(\mathcal{U},\mathsf{A}\mathcal{U})} \\ &= \frac{\rho_{\mathcal{V}} \sin\theta_{\max}(\mathcal{V}, \mathsf{A}\mathcal{U}) + \rho_{\mathcal{V}^\perp} \sin\theta_{\max}(\mathcal{V},\mathcal{U})}{\cos\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})}.\end{aligned}$$

The lemma's statement then follows with Weyl's theorem, cf. corollary 3.29. □

The bound (3.49) in lemma 3.30 exhibits an invariant subspace $\mathcal{V}$ which is not known in general. In practice, the deflation space $\mathcal{U}$ is often chosen as an approximation to an invariant subspace, e.g., as the subspace spanned by Ritz or harmonic Ritz vectors from a larger subspace. Assume that an orthonormal basis of approximate eigenvectors $U \in \mathcal{H}^m$, i.e., $\langle U, U\rangle = \mathbf{I}_m$ and $\mathcal{U} = [\![U]\!]$, and approximations to eigenvalues $\mu_1, \ldots, \mu_m \in \mathbb{R}$ are given. With $\mathbf{D} = \operatorname{diag}(\mu_1, \ldots, \mu_m)$, the norm of the residual $R = \mathsf{A}U - U\mathbf{D}$ can be used as an indicator for the quality of the approximate eigenvectors and eigenvalues. In the Arnoldi and Lanczos methods, the residual norm of Ritz and harmonic Ritz pairs is available as a cheap byproduct of the iteration.

An interesting question is if a similar statement to lemma 3.30 can be formulated in terms of the residual norm $\|R\|$. In fact, such a bound can be established if additional knowledge about the spectrum of A is available. In the course of this subsection, the following definitions of the spectral gap and the spectral interval gap are of importance.

Definition 3.31 (Spectral gap)**.** Let two bounded sets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$ be given.

1. The *spectral gap* between A and B is defined as

$$\delta(A, B) = \min_{\substack{\alpha \in A \\ \beta \in B}} |\alpha - \beta|.$$

2. The *spectral interval gap* between A and B is defined as

$$\underline{\delta}(A, B) = \begin{cases} \delta(A, B) & \text{if } B \cap [\min(A), \max(A)] = \varnothing \\ & \text{or } A \cap [\min(B), \max(B)] = \varnothing, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the spectral (interval) gap should not be confused with the gap between subspaces which also appears in the following theorems in form of the sine of the maximal canonical angle, cf. definition 2.19. The next theorem can be found in the seminal paper of Davis and Kahan [26] and relates the maximal angle between an invariant subspace and an approximation to the residual of the given eigenpair approximations and a spectral interval gap.

Theorem 3.32 ($\sin\theta$-theorem)**.** *Let L be a self-adjoint linear operator and $\mathcal{V} \subseteq \mathcal{H}$ an L-invariant subspace. Furthermore, let $Z \in \mathcal{H}^m$ be with $\langle Z, Z\rangle = \mathbf{I}_m$, $\mathcal{Z} = [\![Z]\!]$ and $\mathbf{M} = \mathbf{M}^\mathsf{H} \in \mathbb{C}^{m,m}$. If $\underline{\delta} = \underline{\delta}\,(\Lambda(\mathbf{M}), \Lambda(\,\mathsf{L}|_{\mathcal{V}^\perp}))>0$, then*

$$\sin\theta_{\max}(\mathcal{V}, \mathcal{Z}) \le \frac{\|\mathsf{L}Z - Z\mathbf{M}\|}{\underline{\delta}}.$$

Proof. See [167, theorem 3.6]. The original proof by Davis and Kahan can be found in [26]. Their theorem does not require the Hilbert space $\mathcal{H}$ to be finite-dimensional and even allows unbounded operators. □

In the proof of the following theorem, the $\sin\theta$-theorem and lemma 3.30 are used in order to obtain a bound that depends on computable quantities in terms of the given eigenpair approximations and the spectral interval gap between the given eigenvalue approximations and a part of A's spectrum.

Theorem 3.33. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be self-adjoint and let $U \in \mathcal{H}^m$ such that $\langle U, U\rangle = \mathbf{I}_m$, $\mathcal{U} = [\![U]\!]$ and $\sin\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$. Furthermore, let $\mathbf{D} = \operatorname{diag}(\mu_1, \dots, \mu_m) \in \mathbb{R}^{m,m}$ be given and $R = \mathsf{A}U - U\mathbf{D}$. Let the eigenvalues of A be $\lambda_1, \dots, \lambda_N$.*

If there exists a permutation $k_1, \dots, k_N$ of $1, \dots, N$ such that

$$\underline{\delta} = \underline{\delta}\left(\Lambda(\mathbf{D}), \{\lambda_{k_{m+1}}, \dots, \lambda_{k_N}\}\right) > 0,$$

then the eigenvalues $\widehat{\lambda}_1 \le \dots \le \widehat{\lambda}_N$ of $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}$ satisfy

$$\max_{i\in\{1,\dots,N\}} |\widehat{\lambda}_i - \underline{\lambda}_i| \le \frac{(\rho + \underline{\rho})\frac{\|R\|}{\underline{\delta}} + \rho \sin\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})}{\cos\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})}, \tag{3.50}$$

where $\underline{\lambda}_1 \le \dots \le \underline{\lambda}_N$ are the ordered values $\lambda_{k_{m+1}}, \dots, \lambda_{k_N}, 0, \dots, 0$ (zero is added m times), $\rho = \max_{j\in\{1,\dots,m\}} |\lambda_{k_j}|$ and $\underline{\rho} = \max_{j\in\{m+1,\dots,N\}} |\lambda_{k_j}|$.

Proof. Let $\mathcal{V} \subseteq \mathcal{H}$ be an A-invariant subspace that is associated with the eigenvalues $\lambda_{k_{m+1}}, \dots, \lambda_{k_N}$. Note that $\Lambda(\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}) = \{\underline{\lambda}_1 \dots, \underline{\lambda}_N\}$. With $\mathsf{L} = \mathsf{A}$, $Z = U$ and $\mathbf{M} = \mathbf{D}$, the requirements of the $\sin\theta$-theorem 3.32 are satisfied and thus

$$\sin\theta_{\max}(\mathcal{V}, \mathcal{U}) \le \frac{\|R\|}{\underline{\delta}}. \tag{3.51}$$

With $\rho_{\mathcal{V}} = \rho$ and $\rho_{\mathcal{V}^\perp} = \underline{\rho}$, the following inequality holds according to lemma 3.30:

$$\max_{i\in\{1,\dots,N\}} |\widehat{\lambda}_i - \underline{\lambda}_i| \le \frac{\rho \sin\theta_{\max}(\mathcal{V}, \mathsf{A}\mathcal{U}) + \underline{\rho} \sin\theta_{\max}(\mathcal{V}, \mathcal{U})}{\cos\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})}.$$

With item 3 of lemma 2.20, the inequality

$$\sin\theta_{\max}(\mathcal{V}, \mathsf{A}\mathcal{U}) \le \sin\theta_{\max}(\mathcal{V}, \mathcal{U}) + \sin\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})$$

holds and thus

$$\max_{i\in\{1,\dots,N\}} |\widehat{\lambda}_i - \underline{\lambda}_i| \le \frac{(\rho + \underline{\rho}) \sin\theta_{\max}(\mathcal{V}, \mathcal{U}) + \rho \sin\theta_{\max}(\mathcal{V}, \mathsf{A}\mathcal{U})}{\cos\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})}.$$

The proof is complete after inserting the inequality (3.51) into the last inequality. □

The quantities ρ and $\underline{\rho}$ in theorem 3.33 can both be bounded by $\|\mathsf{A}\|$ which can often be approximated cheaply. As mentioned before, the residual norm $\|R\|$ can be obtained as a byproduct in the Arnoldi or Lanczos algorithms, cf. section 2.7. The maximal angle $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U})$ can be computed by

$$\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) = \arccos\left(\sigma_{\min}(\langle U, Q\rangle)\right), \tag{3.52}$$

where $\sigma_{\min}(\langle U, Q\rangle)$ denotes the minimal singular value of $\langle U, Q\rangle$ and $Q \in \mathcal{H}^m$ is such that $\langle Q, Q\rangle = \mathbf{I}_m$ and $\mathsf{A}\mathcal{U} = [\![Q]\!]$. However, due to round-off errors, care has to be taken because small angles cannot be found accurately with naive implementations, e.g., with equation (3.52). The issue is discussed briefly in section 2.10.

The only quantity that is hard to get in practical applications is the spectral interval gap $\underline{\delta}$ which describes how well the approximate eigenvalues $\mu_1, \ldots, \mu_m$ are separated from $N - m$ eigenvalues $\lambda_{k_{m+1}}, \ldots, \lambda_{k_N}$ of A. The exact value of this quantity is rarely available in practice but it can sometimes be estimated from properties of the originating problem or from eigenvalue approximations of a similar operator. For example, an operator A may be known to have only a few negative eigenvalues $\lambda_1 \leq \ldots \leq \lambda_m < 0$ while the other eigenvalues $0 < \lambda_{m+1} \leq \ldots \leq \lambda_N$ are positive. If then eigenvalue approximations $\mu_1 \leq \ldots \leq \mu_m < 0$ are given, then $\underline{\delta}(\{\mu_1, \ldots, \mu_m\}, \{\lambda_{m+1}, \ldots, \lambda_N\}) = \lambda_{m+1} - \mu_m > |\mu_m|$. Furthermore, the spectral interval gap is rather demanding since it requires that one part of the spectrum lies in an interval that does not contain any of the complementary eigenvalues. The spectral interval gap requirement is visualized in figure 3.3. However, it is not unusual to remove a contiguous set of eigenvalues by deflation, e.g., only a few eigenvalues that are closest to the origin.

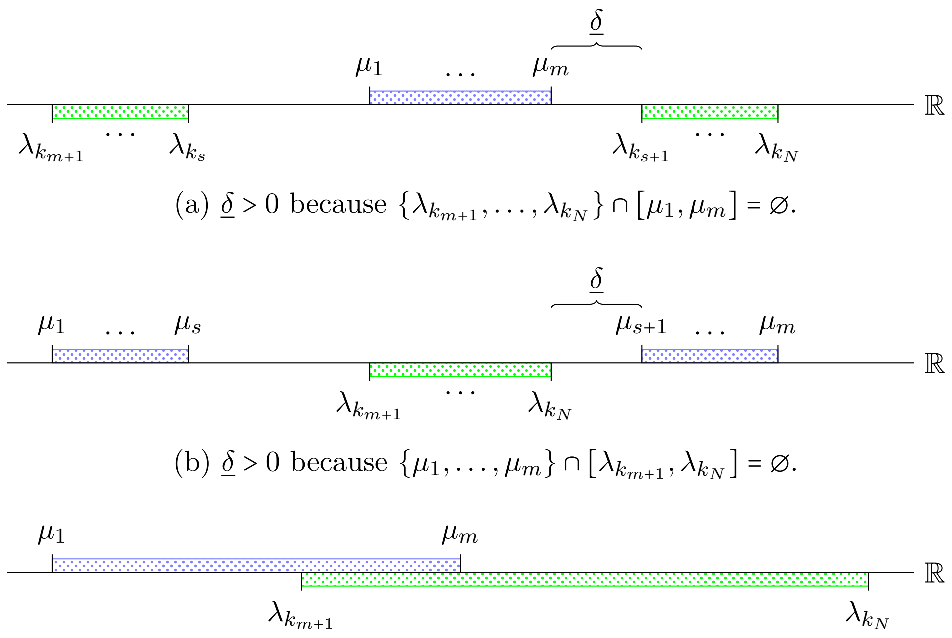
In the next example, it is analyzed to what extent the eigenvalue bound in theorem 3.33 can capture the actual eigenvalue error.

Example 3.34. Let the matrix A be defined as in example 3.23. This example consists of the following two experiments:

1. For the first experiment, let the subspace $\mathcal{V} = [\![\mathbf{I}_{N,3}]\!]$ be also given as in example 3.23. Recall that the eigenvalues of $\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$ are $0, 0, 0 < \lambda_4 \leq \ldots \leq \lambda_N$. For $\varepsilon > 0$, let $\mathcal{U}_\varepsilon = [\![\mathbf{I}_{N,3} + \varepsilon\mathbf{E}]\!]$ with a random $\mathbf{E} \in \mathbb{R}^{N,3}$ of norm 1 and let $U_\varepsilon \in \mathbb{R}^{N,3}$ an orthonormal Ritz vector basis of $\mathcal{U}_\varepsilon$, i.e., $\mathcal{U}_\varepsilon = [\![U_\varepsilon]\!]$, $\langle U_\varepsilon, U_\varepsilon\rangle = \mathbf{I}_3$ and $\langle U_\varepsilon, \mathsf{A}U_\varepsilon\rangle = \mathbf{D}_\varepsilon = \operatorname{diag}(\mu_1^{(\varepsilon)}, \mu_2^{(\varepsilon)}, \mu_3^{(\varepsilon)}) \in \mathbb{R}^{3,3}$, where $\mu_1^{(\varepsilon)} \leq \mu_2^{(\varepsilon)} \leq \mu_3^{(\varepsilon)}$ are the Ritz values. The Ritz residual then is $R_\varepsilon = \mathsf{A}U_\varepsilon - U_\varepsilon\mathbf{D}_\varepsilon$.

   With $U = U_\varepsilon$, both sides of the inequality (3.50) in theorem 3.33 are plotted in figure 3.4a versus the perturbation size ε. The figure clearly shows that the bound in theorem 3.33 lies above the actual eigenvalue error by several orders of magnitude if the perturbation ε is small. The overestimation is not that severe for larger Ritz residual norms. For small perturbations, the plot suggests that the actual eigenvalue error depends quadratically on the Ritz residual norm while the bound in theorem 3.33 only provides a linear bound.

2. The second experiment only differs from experiment 1 in the choice of the subspace $\mathcal{V}$. Here, $\mathcal{V} = [\![\mathbf{I}_{N,2}]\!]$ is chosen and thus the eigenvalues of $\mathsf{P}_{\mathcal{V}^\perp}\mathsf{A}$



(a) $\underline{\delta} > 0$ because $\{\lambda_{k_{m+1}}, \dots, \lambda_{k_N}\} \cap [\mu_1, \mu_m] = \varnothing$.

(b) $\underline{\delta} > 0$ because $\{\mu_1, \dots, \mu_m\} \cap [\lambda_{k_{m+1}}, \lambda_{k_N}] = \varnothing$.

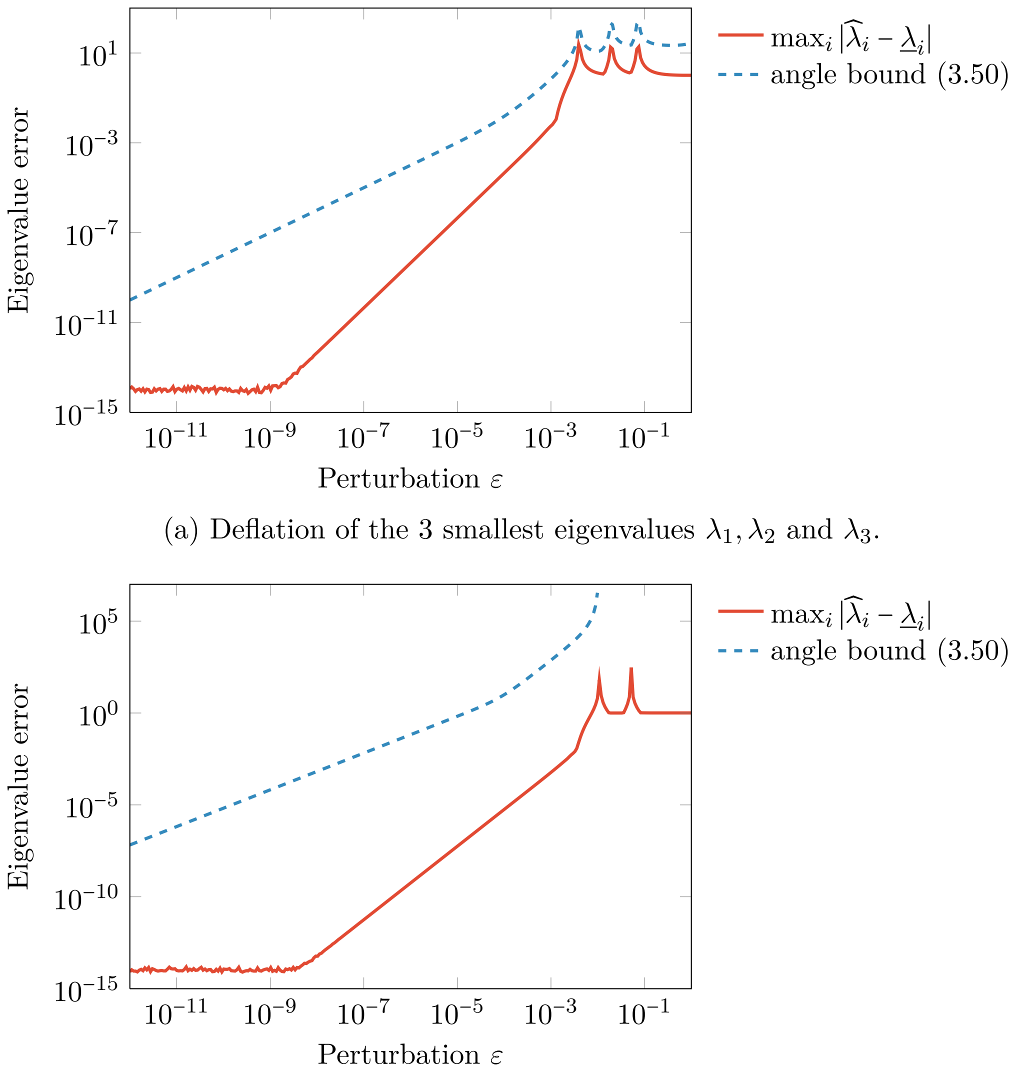
(c) $\underline{\delta} = 0$. See figure 3.5 for a visualization of the spectral gap in this situation.

Figure 3.3.: Visualization of the assumption on the spectral interval gap in theorem 3.33. See definition 3.31 for the definition of the spectral interval gap.

> are $0, 0, \lambda_3, \dots, \lambda_N$. Let the Ritz vector basis U_ε, the Ritz value matrix $\mathbf{D}_\varepsilon = \operatorname{diag}(\mu_1^{(\varepsilon)}, \mu_2^{(\varepsilon)})$ and the Ritz residual R_ε be defined analogously to experiment 1.
>
> Analogous to experiment 1, figure 3.4b shows both sides of the inequality theorem 3.33. Now the difference between the bound the actual error is even around 10^3 for large perturbations. The reason for this severe overestimation is the presence of the spectral interval gap $\underline{\delta}(\{\mu_1^{(\varepsilon)}, \mu_2^{(\varepsilon)}\}, \{\lambda_3, \dots, \lambda_N\})$ which becomes very small. For $\varepsilon \approx 10^{-2}$, the spectral interval gap even is zero because none of the two spectra $\{\mu_1^{(\varepsilon)}, \mu_2^{(\varepsilon)}\}$ and $\{\lambda_3, \dots, \lambda_N\}$ can be placed into an interval such that it does not contain any of the eigenvalues from the other one.

Note that the “exact” eigenvalue error $\max_{i \in \{1,\dots,N\}} |\widehat{\lambda}_i - \underline{\lambda}_i|$ gets stuck around machine precision since the involved eigenvalues are computed numerically.

It could be observed in the previous experiment, that the bound (3.50) in theorem 3.33 deviates from the actual eigenvalue error by several orders of magnitude and it was conjectured that the eigenvalue errors depend quadratically on the perturbation or the residual norm. In [166], Stewart showed that the eigenvalue error for self-adjoint (undeflated) operators can be bounded quadratically in



(a) Deflation of the 3 smallest eigenvalues λ_1, λ_2 and λ_3.

(b) Deflation of the 2 smallest eigenvalues λ_1 and λ_2.

Figure 3.4.: Spectral bound from theorem 3.33 for deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ with approximate invariant subspaces. The setup is described and discussed in example 3.34.

terms of the residual. However, since his eigenvalue error bound is also based on the $\sin\theta$-theorem 3.32 by Davis and Kahan, it still has the same limitation: the spectrum has to be separated such that the spectral interval gap is positive, cf. definition 3.31 and figure 3.3. Mathias [113] improved the eigenvalue bounds significantly by allowing the eigenvalues and eigenvalue approximations to be scattered throughout the spectrum. A subset of the results from [113] is stated here in terms of Ritz residuals rather than in the original form of a perturbed block matrix.

Theorem 3.35. *Let $\mathsf{L} \in \mathcal{L}(\mathcal{H})$ be self-adjoint with eigenvalues $\lambda_1 \leq \ldots \leq \lambda_N$. Assume that $X \in \mathcal{H}^m$ and $Y \in \mathcal{H}^{N-m}$ are given such that $\langle [X,Y],[X,Y] \rangle = \mathbf{I}_N$ and $\tilde{\lambda}_1 \leq \ldots \leq \tilde{\lambda}_N$ are the eigenvalues of* $\operatorname{diag}(\mathbf{M},\mathbf{N})$*, where $\mathbf{M} = \langle X, \mathsf{L}X \rangle$ and $\mathbf{N} = \langle Y, \mathsf{L}Y \rangle$. Furthermore, let $i_1 < \ldots < i_m$ be such that $\tilde{\lambda}_{i_1} \leq \ldots \leq \tilde{\lambda}_{i_m}$ are the eigenvalues of $\mathbf{M}$. Then with $R = \mathsf{L}X - X\mathbf{M}$ the following holds:*

1. *If $\delta_i = \delta(\{\lambda_i\}, \Lambda(\mathbf{N})) > 0$ for a $i \in \{1,\ldots,N\}$, then*
$$|\lambda_i - \tilde{\lambda}_i| \leq \frac{\|R\|^2}{\delta_i}.$$

2. *If $\delta = \delta(\{\lambda_{i_1},\ldots,\lambda_{i_m}\}, \Lambda(\mathbf{N})) > 0$ then*
$$\max_{k\in\{1,\ldots,m\}} |\lambda_{i_k} - \tilde{\lambda}_{i_k}| \leq \frac{\|R\|^2}{\delta}.$$

Proof.

1. The statement immediately follows by applying the first statement of theorem 1 from [113] to the matrices $\mathbf{A} = \langle [X,Y], \mathsf{L}[X,Y] \rangle$ and $\tilde{\mathbf{A}} = \operatorname{diag}(\mathbf{M},\mathbf{N})$. Note that $\delta_i = \delta(\{\lambda_i\}, \Lambda(\mathbf{N})) > 0$ is equivalent to the condition $\lambda_i \notin \Lambda(\mathbf{N})$ in [113].

2. The result follows from
$$\max_{k\in\{1,\ldots,m\}} |\lambda_{i_k} - \tilde{\lambda}_{i_k}| \leq \frac{\|R\|^2}{\min_{k\in\{1,\ldots,m\}} \delta_{i_k}} \leq \frac{\|R\|^2}{\delta}.$$

□

With Mathias' quadratic residual norm bound for the eigenvalues of a self-adjoint operator A, an apparently new quadratic bound can be derived for the deflated operator $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}$.

Theorem 3.36. *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be self-adjoint with eigenvalues $\lambda_1 \leq \ldots \leq \lambda_N$ and let $U \in \mathcal{H}^m$ and $U_\perp \in \mathcal{H}^{N-m}$ be such that $\langle [U,U_\perp],[U,U_\perp] \rangle = \mathbf{I}_N$ and $\sin\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ with $\mathcal{U} = [\![U]\!]$. Furthermore, assume that $\tilde{\lambda}_1 \leq \ldots \leq \tilde{\lambda}_N$ are the eigenvalues of* $\operatorname{diag}(\mathbf{M},\mathbf{N})$*, where $\mathbf{M} = \langle U_\perp, \mathsf{A}U_\perp \rangle$ and $\mathbf{N} = \langle U, \mathsf{A}U \rangle$ and let $1 \leq i_1 < \ldots < i_{N-m} \leq N$ be given such that $\tilde{\lambda}_{i_1} \leq \ldots \leq \tilde{\lambda}_{i_{N-m}}$ are the eigenvalues of $\mathbf{M}$.*

Then with $R = \mathsf{A}U - U\mathbf{N}$ and $\mu_{\min} := \min_{\mu\in\Lambda(\mathbf{N})} |\mu|$ the following holds for the spectrum $\Lambda(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) = \{0\} \cup \{\widehat{\lambda}_1 \leq \ldots \leq \widehat{\lambda}_{N-m}\}$:

1. *If $\delta_k = \delta(\{\lambda_{i_k}\}, \Lambda(\mathbf{N})) > 0$ for a $k \in \{1,\ldots,N-m\}$, then*
$$|\lambda_{i_k} - \widehat{\lambda}_k| \leq \|R\|^2 \left(\frac{1}{\delta_k} + \frac{1}{\mu_{\min}} \right). \tag{3.53}$$

2. *If $\delta = \delta(\{\lambda_{i_1},\ldots,\lambda_{i_{N-m}}\}, \Lambda(\mathbf{N})) > 0$, then*
$$\max_{k\in\{1,\ldots,N-m\}} |\lambda_{i_k} - \widehat{\lambda}_k| \leq \|R\|^2 \left(\frac{1}{\delta} + \frac{1}{\mu_{\min}} \right). \tag{3.54}$$

Proof. Because $\mathcal{U} \subseteq \mathcal{N}(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A})$, the spectrum of $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ can be characterized as

$$\Lambda(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}) = \{0\} \cup \Lambda\left(\left\langle U_\perp, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}U_\perp\right\rangle\right),$$

where the eigenvalues of $\left\langle U_\perp, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}U_\perp\right\rangle$ are $\widehat{\lambda}_1 \leq \ldots \leq \widehat{\lambda}_{N-m}$. Note that

$$\left\langle U_\perp, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}U_\perp\right\rangle = \langle U_\perp, \mathsf{A}U_\perp\rangle - \langle U_\perp, \mathsf{A}U\rangle\langle U, \mathsf{A}U\rangle^{-1}\langle \mathsf{A}U, U_\perp\rangle = \mathbf{M} - \widehat{R}\mathbf{N}^{-1}\widehat{R}^{\mathsf{H}},$$

where $\widehat{R} = \langle U_\perp, \mathsf{A}U\rangle$. Since $\tilde{\lambda}_{i_1} \leq \ldots \leq \tilde{\lambda}_{i_{N-m}}$ are the eigenvalues of $\mathbf{M}$, Weyl's theorem (cf. corollary 3.29) and the equality

$$\begin{aligned}\left\|\widehat{R}\right\|_2 &= \|\langle U_\perp, \mathsf{A}U\rangle\|_2 = \|U_\perp\langle U_\perp, \mathsf{A}U\rangle\| = \|\mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}U\| = \|(\mathsf{id} - \mathsf{P}_{\mathcal{U}})\mathsf{A}U\| \\ &= \|\mathsf{A}U - U\langle U, \mathsf{A}U\rangle\| = \|R\|\end{aligned}$$

yield for $k \in \{1, \ldots, N-m\}$

$$|\tilde{\lambda}_{i_k} - \widehat{\lambda}_k| \leq \left\|\widehat{R}\mathbf{N}^{-1}\widehat{R}^{\mathsf{H}}\right\|_2 \leq \left\|\widehat{R}\right\|_2^2 \left\|\mathbf{N}^{-1}\right\|_2 = \frac{\|R\|^2}{\mu_{\min}}. \tag{3.55}$$

With theorem 3.35, the eigenvalues $\tilde{\lambda}_{i_1} \leq \ldots \leq \tilde{\lambda}_{i_{N-m}}$ of $\mathbf{M}$ can now be related with the eigenvalues $\lambda_{i_1} \leq \ldots \leq \lambda_{i_{N-m}}$ of A. For $k \in \{1, \ldots, N-m\}$, the theorem states that if $\delta_k > 0$, then

$$|\lambda_{i_k} - \tilde{\lambda}_{i_k}| \leq \frac{\|R\|^2}{\delta_k}. \tag{3.56}$$

The inequalities (3.55) and (3.56) now yield

$$|\lambda_{i_k} - \widehat{\lambda}_k| \leq |\lambda_{i_k} - \tilde{\lambda}_{i_k}| + |\tilde{\lambda}_{i_k} - \widehat{\lambda}_k| \leq \|R\|^2\left(\frac{1}{\delta_k} + \frac{1}{\mu_{\min}}\right)$$

which proves the first inequality of the theorem. Analogous to the second inequality of theorem 3.35, the second inequality of this theorem follows by maximizing $\frac{1}{\delta_k}$. □

Compared to the angle-based bound in theorem 3.33, the new bound in theorem 3.36 is more attractive in several ways. First, the most obvious feature is it's quadratic dependency on the residual norm $\|R\|$. Second, the bound uses the spectral gap instead of the spectral interval gap and thus also allows the spectra to be scattered. The only requirement is that the subset $\{\lambda_{i_1}, \ldots, \lambda_{i_{N-m}}\}$ of the eigenvalues of A is disjoint from the spectrum of $\mathbf{N} = \langle U, \mathsf{A}U\rangle$. The spectral gap requirement is visualized in figure 3.5. Another benefit is that the first inequality (3.53) of theorem 3.36 gives an individual bound for each eigenvalue of $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ by taking into account the location of the corresponding eigenvalue of the original operator A.

In order to assess the bound in theorem 3.36, the following example demonstrates the behavior for the setup that was described in example 3.34.

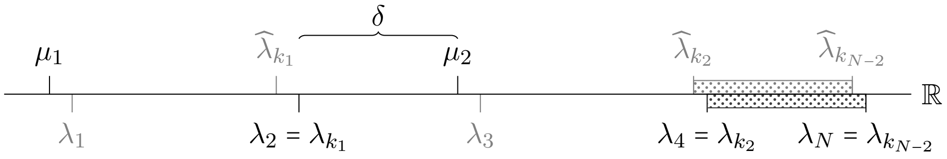

δ
μ_1
$\widehat{\lambda}_{k_1}$
μ_2
$\widehat{\lambda}_{k_2}$
$\widehat{\lambda}_{k_{N-2}}$
$\mathbb{R}$
λ_1
$\lambda_2 = \lambda_{k_1}$
λ_3
$\lambda_4 = \lambda_{k_2}$
$\lambda_N = \lambda_{k_{N-2}}$

Figure 3.5.: Visualization of the assumption on the spectral gap in theorem 3.36 for the case of two eigenvalues μ_1 and μ_2 of $\langle U, \mathsf{A}U\rangle$ that are scattered through the eigenvalues $\lambda_{k_1} \leq \ldots \leq \lambda_{k_{N-2}}$ of A. The spectral gap δ is positive, whereas the spectral interval gap $\underline{\delta}$ is zero, cf. figure 3.3c. See definition 3.31 for the definition of the spectral gap.

Example 3.37. Both experiments in example 3.34 are repeated with the new bound. In figure 3.6 the bound (3.54) is plotted versus the perturbation ε. For comparison, the angle-based bound (3.50) is also included. In both cases, i.e., when deflating λ_1, λ_2 and λ_3 or when deflating λ_1 and λ_2, the quadratic bound is clearly favorable for the full range of the perturbation ε. Remarkably, the quadratic bound does not show the offset of $\approx 10^3$ for large perturbations and stays rather close to the actual eigenvalue error.

The derived bounds are used in section 3.4 in order to select a deflation subspace when a sequence of linear systems has to be solved.

3.3.3. Krylov subspace methods

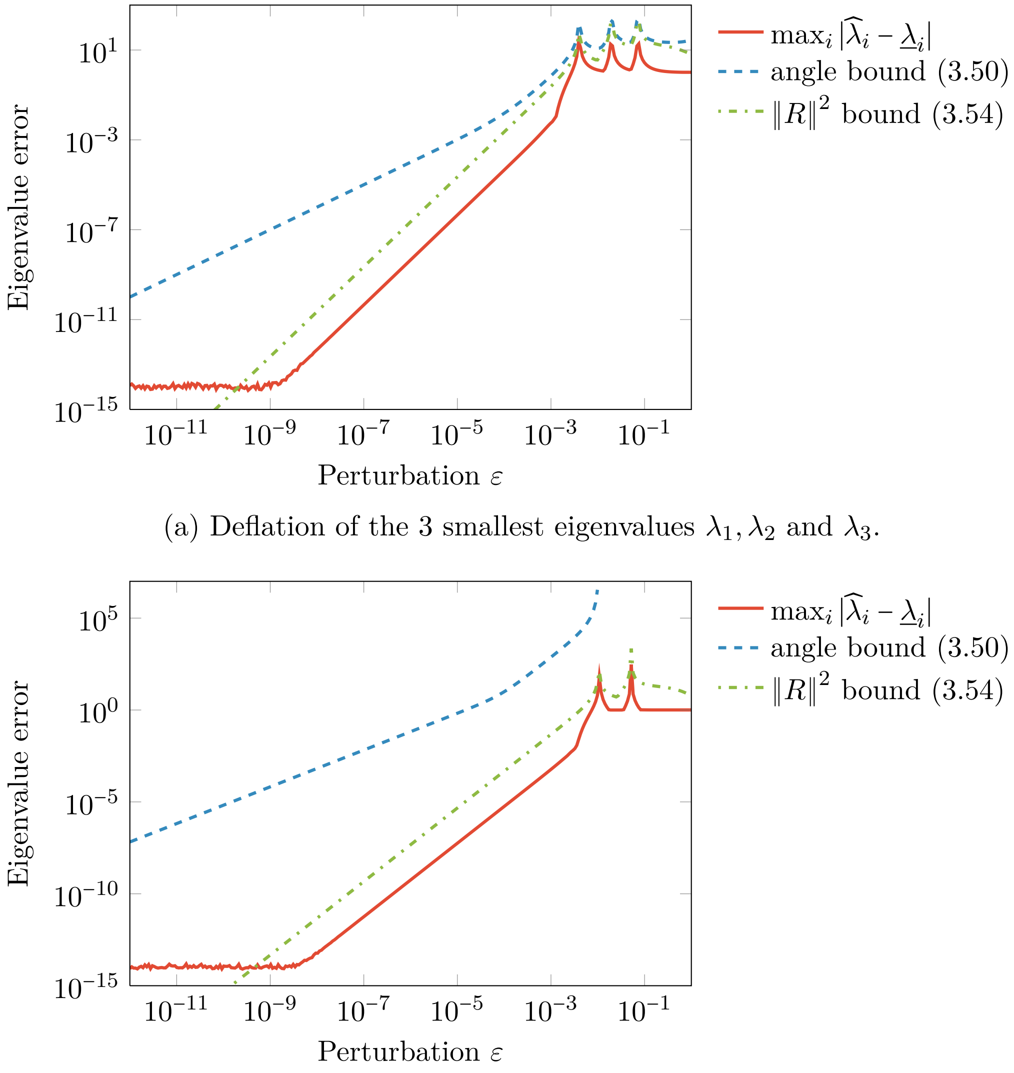
In section 3.4, it is analyzed how a suitable deflation subspace can be extracted from a Krylov subspace that was generated with an operator and initial vector that correspond to a previous linear system in a sequence of linear systems. The analysis leads to the following question: how does the already known behavior of a Krylov subspace method that is applied to $\mathsf{A}x = b$ relate to the behavior of the Krylov subspace method when it is applied to a perturbed linear system $(\mathsf{A} + \mathsf{F})\widehat{x} = b + f$?

The following statement describes a widespread opinion:

> A small perturbation F of the operator and a small perturbation f of the initial vector only leads to a small perturbation of the convergence behavior of Krylov subspace methods for solving linear systems.

The following example shows that this statement is not true in general.

Example 3.38. The example shows results of the GMRES and MINRES methods with academic examples where the operator or the right hand side is perturbed.



(a) Deflation of the 3 smallest eigenvalues λ_1, λ_2 and λ_3.

(b) Deflation of the 2 smallest eigenvalues λ_1 and λ_2.

Figure 3.6.: Spectral bound from theorem 3.36 for a deflated operator $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}$ with approximate invariant subspaces. The setup is described and discussed in example 3.37.

1. For $n \in \mathbb{N}$ and $\alpha \gg \varepsilon > 0$, let the matrices

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{C} = \alpha \begin{bmatrix} 0 & & & & 1 \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix} \in \mathbb{R}^{n,n}, \quad \mathbf{F} = \varepsilon e_4 e_1^\mathsf{T} \in \mathbb{R}^{n+3,n+3}$$

and the vectors

$$b = e_1 \in \mathbb{R}^{n+3} \quad \text{and} \quad f = \varepsilon e_4$$

be given. With $n = 20$, $\alpha = 100$, $\varepsilon = 10^{-8}$ and $\mathbf{A} = \operatorname{diag}(\mathbf{B}, \mathbf{C})$, the GMRES method is now applied to the linear systems $\mathbf{A}x = b$, $(\mathbf{A} + \mathbf{F})x_{\mathbf{F}} = b$ and $\mathbf{A}x_f = b + f$. The resulting convergence histories are shown in figure 3.7a. The original linear system is solved up to machine precision after 3 iterations while both perturbed versions exhibit $n = 20$ iterations where the residual norm stagnates. In fact, increasing n, e.g., to 10^3 or larger, still leads to n iterations of stagnation. Here, $n = 20$ is simply chosen because the resulting convergence histories begin to look ambiguous for large values of n. Varying ε or α results in different levels where the stagnation takes place.

Clearly, the example is constructed such that the Krylov subspace $\mathcal{K}_3(\mathbf{A}, b)$ is A-invariant and the exact solution is thus found after 3 iterations (in fact, the exact solution is found numerically without round-off errors in this special case because only integer operations are performed). Adding the perturbation $\mathbf{F}$ or f drastically changes the behavior because the right hand side is no longer an element of a 3-dimensional invariant subspace and due to the dominance of the submatrix $\mathbf{C}$, a significant residual norm reduction just happens once the Krylov subspace has full dimension $n + 3$.

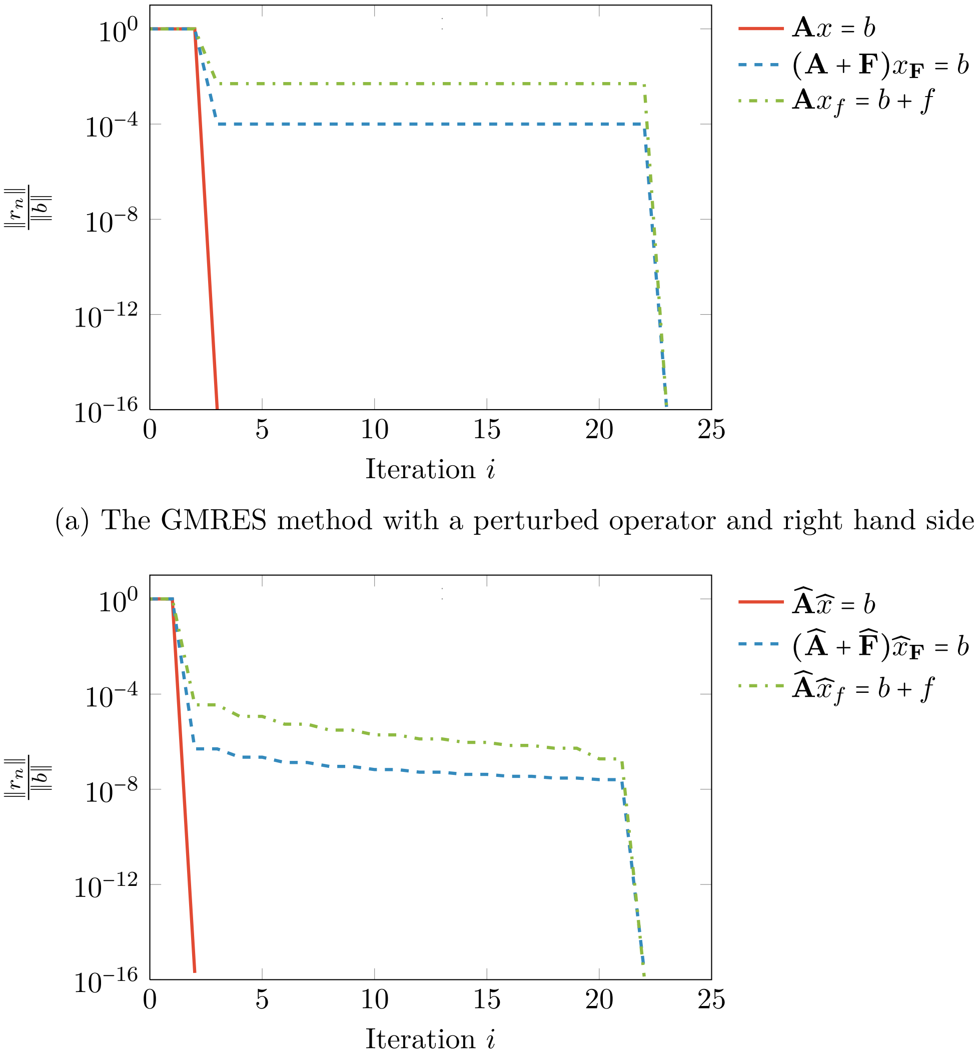
2. An interesting question is if such a behavior can also be observed with self-adjoint operators. Therefore, let

$$\widehat{\mathbf{B}} = \mathbf{B} + \mathbf{B}^{\mathsf{T}}, \quad \widehat{\mathbf{C}} = \alpha \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix} \quad \text{and} \quad \widehat{\mathbf{F}} = \mathbf{F} + \mathbf{F}^{\mathsf{H}}.$$

   Again with $n = 20$, $\alpha = 100$, $\varepsilon = 10^{-8}$ and $\widehat{\mathbf{A}} = \operatorname{diag}(\widehat{\mathbf{B}}, \widehat{\mathbf{C}})$, the MINRES method is applied to the three linear systems $\widehat{\mathbf{A}}\widehat{x} = b$, $(\widehat{\mathbf{A}} + \widehat{\mathbf{F}})\widehat{x}_{\mathbf{F}} = b$ and $\widehat{\mathbf{A}}\widehat{x}_f = b + f$. The resulting convergence histories are presented in figure 3.7b. There is no phase of persistent complete stagnation like in the first experiment but the reduction of the residual norm is severely impeded by adding the perturbation $\widehat{\mathbf{F}}$ or f.

Although this example is contrived, it stresses the fact that the GMRES residual can stagnate at any time during the iteration and that it can by no means be concluded that the residual norm reduction by a factor of 10^{-4} from iteration 2 to iteration 3 is continued in subsequent iterations.

Although the previous example draws a dark picture of perturbation theory for Krylov subspace methods, useful bounds can be derived that characterize the behavior of Krylov subspaces under perturbations. Similar to the perturbation theory



(a) The GMRES method with a perturbed operator and right hand side.

(b) The MINRES method with a perturbed operator and right hand side.

Figure 3.7.: Krylov subspace methods with a perturbed operator and right hand side. The setup is described and discussed in example 3.38.

for deflated operators, the bounds often exhibit some information about the spectrum that is usually not available in practice. However, the bounds lead to a better understanding of the behavior of Krylov subspace methods and can explain under which circumstances the statement in the fictional quote at the beginning of this subsection is true.

Consider the GMRES method with initial guess $x_0 = 0$ for the linear systems $\mathsf{A}x = b$ and $\widehat{\mathsf{A}}\widehat{x} = \widehat{b}$, where $\widehat{\mathsf{A}} = \mathsf{A} + \mathsf{F}$ and $\widehat{b} = b + f$. Let the constructed iterates and residuals be denoted by x_n, r_n, $\widehat{x}_n$ and $\widehat{r}_n$. Assume that the residual r_n or its norm $\|r_n\|$ is known and the task is to bound the deviation of $\widehat{r}_n$ from r_n, e.g., measured by $\|\widehat{r}_n - r_n\|$ or $\|\widehat{r}_n\| - \|r_n\| \leq \|\widehat{r}_n - r_n\|$.

Several approaches to perturbations of Krylov subspaces and Krylov subspace methods are known in the literature and serve very different purposes. Here, a brief overview is presented before extending one of the approaches for the situation described above.

In [22], Carpraux, Godunov and Kuznetsov studied the sensitivity of a Krylov subspace to a perturbation of the operator A and present an analysis of the condition number of Arnoldi bases[3] and Krylov subspaces. Unfortunately, the presented algorithm for the estimation of the condition numbers involves the in practice infeasible solution of a large linear system and does not lead to a deeper analytic understanding of the behavior of Krylov subspaces. Kuznetsov extended the treatment in [101] to perturbations of the initial vector v and gave further bounds for the Arnoldi basis and the corresponding Hessenberg matrix in terms of the condition numbers derived in [22].

Given any subspace $\mathcal{U} \subset \mathcal{H}$, Stewart showed in [163], how a backward perturbation F with minimal norm $\|\mathsf{F}\|$ can be constructed such that $\mathcal{U}$ is a Krylov subspace of $\mathsf{A} + \mathsf{F}$. A similar strategy is applied in section 3.4.3 in order to construct a perturbation F such that $\mathcal{U}$ is a Krylov subspace of $\mathsf{A} + \mathsf{F}$ with a *specific* initial vector $u \in \mathcal{U}$ and such that the influence of the perturbation is minimal *in each iteration.*

For the case of dual Krylov subspace methods like BiCG or QMR, Wu, Wei, Jia and Ling [187] considered the case of two subspaces $\mathcal{U}, \mathcal{V} \subset \mathcal{H}$ of same dimension and determined a backward perturbation F such that $\mathcal{U}$ and $\mathcal{V}$ are Krylov subspaces of $\mathsf{A} + \mathsf{F}$ and its adjoint, respectively.

The theory of inexact Krylov subspace methods by Simoncini and Szyld in [157] focuses on Krylov subspace methods where the application of the operator A is not carried out exactly but instead a perturbed operator $\mathsf{A} + \mathsf{F}_i$ is applied in each iteration i. The goal is to determine an admissible perturbation norm $\|\mathsf{F}_i\|$ for each iteration such that the method still constructs iterates whose residual norm eventually drops below a given tolerance. The perturbations F_i of the original operator A are allowed to change in each iteration but can be seen as the application of $\mathsf{A}+\mathsf{F}$ with a constant operator $\mathsf{F} = \sum_{i=1}^{n} \mathsf{F}_i \mathsf{P}_{[\![v_i]\!]}$, where $v_1, \ldots, v_n$ is the constructed orthonormal basis. Thus, the iterates and residuals in [157] are $\widehat{x}_n$ and $\widehat{r}_n$ in the notation that is used here. However, their analysis differs from the situation here because in [157], the residual norm $\|b - \mathsf{A}\widehat{x}_n\|$ is bounded and the residual r_n that corresponds to the Krylov subspace method applied to the unperturbed linear system is not considered. The representation of the perturbation F as a sum where each summand acts on one Arnoldi vector reappears in section 3.4.3.

A recent approach that almost fits the situation of this subsection has been presented by Sifuentes, Embree and Morgan in [153]. They provide bounds for the quantity $\|\widehat{r}_n\| - \|r_n\|$ for the case of a perturbed operator but an unperturbed

[3] The Arnoldi basis is called *Krylov basis* in [22] which is used here for the basis $v, \mathsf{A}v, \ldots, \mathsf{A}^{n-1}v$, cf. section 2.7.

right hand side, i.e., $f = 0$. Their analysis is based on resolvent estimates and the pseudospectrum of the operator A and is closely related to the pseudospectral residual norm bound for GMRES, cf. theorem 2.60. The approach of Sifuentes, Embree and Morgan is presented here in more detail and is slightly generalized in order to also allow arbitrary perturbations f of the right hand side and arbitrary initial guesses. Also, the same approach can be used for the CG method, where the theory becomes much simpler due to the favorable spectral properties. First, the strategy for the proof of the main result in the work of Sifuentes, Embree and Morgan [153] is outlined.

Let $\mathsf{A}x = b$ and $\widehat{\mathsf{A}}\widehat{x} = b$ with nonsingular $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ and $\widehat{\mathsf{A}} = \mathsf{A} + \mathsf{F}$ be two linear systems and let $d_1 = d(\mathsf{A}, b)$ and $d_2 = d(\widehat{\mathsf{A}}, b)$. For $n \leq \min\{d_1, d_2\}$, the residuals r_n and $\widehat{r}_n$ of the GMRES method applied to the two linear systems with the zero initial guess can be expressed as $r_n = p_n(\mathsf{A})b$ and $\widehat{r}_n = \widehat{p}_n(\widehat{\mathsf{A}})b$ with polynomials $p_n, \widehat{p}_n \in \mathbb{P}_{n,0}$. Due to the minimization property of the GMRES polynomials p_n and $\widehat{p}_n$ and the triangle inequality, the following inequality can be obtained:

$$\begin{aligned}\|\widehat{r}_n\| - \|r\| = \left\|\widehat{p}_n(\widehat{\mathsf{A}})b\right\| - \|p_n(\mathsf{A})b\| \leq \left\|p_n(\widehat{\mathsf{A}})b\right\| - \|p_n(\mathsf{A})b\| \leq \left\|p_n(\widehat{\mathsf{A}})b - p_n(\mathsf{A})b\right\| \\ \leq \left\|p_n(\widehat{\mathsf{A}}) - p_n(\mathsf{A})\right\| \|b\| .\end{aligned}$$

If now $\Gamma \subset \mathbb{C}$ is a finite union of Jordan curves that enclose the spectra $\Lambda(\mathsf{A})$ and $\Lambda(\widehat{\mathsf{A}})$, then the polynomial expressions can be written as a Cauchy integral similar to the proof of theorem 2.60:

$$p_n(\mathsf{A}) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma p_n(\lambda)(\lambda \mathsf{id} - \mathsf{A})^{-1} d\lambda \quad \text{and} \quad p_n(\widehat{\mathsf{A}}) = \frac{1}{2\pi \mathbb{i}} \int_\Gamma p_n(\lambda)(\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} d\lambda.$$

Thus

$$\begin{aligned}\frac{\|\widehat{r}_n\| - \|r_n\|}{\|b\|} &\leq \frac{1}{2\pi} \left\| \int_\Gamma p_n(\lambda) \left((\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} - (\lambda \mathsf{id} - \mathsf{A})^{-1} \right) d\lambda \right\| \\ &\leq \frac{1}{2\pi} \int_\Gamma |p_n(\lambda)| \left\| (\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} - (\lambda \mathsf{id} - \mathsf{A})^{-1} \right\| d\lambda. \qquad (3.57)\end{aligned}$$

The difference of the resolvents can be tackled with the second resolvent identity, cf. Nevanlinna [127]:

$$(\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} - (\lambda \mathsf{id} - \mathsf{A})^{-1} = (\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} \mathsf{F} (\lambda \mathsf{id} - \mathsf{A})^{-1}. \qquad (3.58)$$

The first resolvent can be expressed as

$$\begin{aligned}(\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} = (\lambda \mathsf{id} - \mathsf{A} - \mathsf{F})^{-1} &= \left((\lambda \mathsf{id} - \mathsf{A}) \left(\mathsf{id} - (\lambda \mathsf{id} - \mathsf{A})^{-1} \mathsf{F} \right)^{-1} \right)^{-1} \\ &= \left(\mathsf{id} - (\lambda \mathsf{id} - \mathsf{A})^{-1} \mathsf{F} \right)^{-1} (\lambda \mathsf{id} - \mathsf{A})^{-1}.\end{aligned}$$

If $\varepsilon := \|\mathsf{F}\| < \frac{1}{\|(\lambda \mathsf{id} - \mathsf{A})^{-1}\|} =: \delta_\lambda$, then $\left\|(\lambda \mathsf{id} - \mathsf{A})^{-1}\mathsf{F}\right\| \leq \frac{\varepsilon}{\delta_\lambda} < 1$ and $\left(\mathsf{id} - (\lambda \mathsf{id} - \mathsf{A})^{-1} \mathsf{F} \right)^{-1}$ can be expressed as a Neumann series, cf. Kato [90]:

$$\left(\mathsf{id} - (\lambda \mathsf{id} - \mathsf{A})^{-1} \mathsf{F} \right)^{-1} = \sum_{j=0}^{\infty} \left((\lambda \mathsf{id} - \mathsf{A})^{-1} \mathsf{F} \right)^j .$$

Its norm can be bounded by the geometric series

$$\left\|\left(\mathsf{id} - (\lambda\mathsf{id} - \mathsf{A})^{-1}\mathsf{F}\right)^{-1}\right\| \leq \sum_{j=0}^{\infty}\left(\frac{\varepsilon}{\delta_\lambda}\right)^j = \frac{1}{1 - \frac{\varepsilon}{\delta_\lambda}}$$

and the norm of the resolvent difference (3.58) can thus be bounded by

$$\left\|(\lambda\mathsf{id} - \widehat{\mathsf{A}})^{-1} - (\lambda\mathsf{id} - \mathsf{A})^{-1}\right\| \leq \frac{\varepsilon}{\delta_\lambda\left(\delta_\lambda - \varepsilon\right)}.$$

For any $\delta > \varepsilon$, the boundary of the δ-pseudospectrum $\Lambda_\delta(\mathsf{A})$ of A can be used as the contour of integration in (3.57) because it encloses by definition 2.59 the spectra $\Lambda(\mathsf{A})$ and $\Lambda(\widehat{\mathsf{A}})$. Note that $\delta = \delta_\lambda = \frac{1}{\|(\lambda\mathsf{id}-\mathsf{A})^{-1}\|}$ holds for any $\lambda \in \partial\Lambda_\delta(\mathsf{A})$ and the following bound follows directly:

$$\frac{\|\widehat{r}_n\| - \|r_n\|}{\|b\|} \leq \frac{\varepsilon}{\delta - \varepsilon}\,\frac{|\partial\Lambda_\delta(\mathsf{A})|}{2\pi\delta}\max_{\lambda\in\partial\Lambda_\delta(\mathsf{A})}|p_n(\lambda)| = \frac{\varepsilon}{\delta - \varepsilon}\,\frac{|\partial\Lambda_\delta(\mathsf{A})|}{2\pi\delta}\sup_{\lambda\in\Lambda_\delta(\mathsf{A})}|p_n(\lambda)|.$$

As in theorem 2.60, the last equality holds because p_n is holomorphic and thus its maximum is attained on the boundary. Note that $|\partial\Lambda_\delta(\mathsf{A})|$ again denotes the curve's arc length. The above inequality is the main result of of Sifuentes, Embree and Morgan [153] and is gathered in the following theorem for later reference.

Theorem 3.39. *Assume that* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ *is nonsingular,* $\mathsf{F} \in \mathcal{L}(\mathcal{H})$*,* $b \in \mathcal{H}$ *and* $n \leq d(\mathsf{A}, b)$*. Let the n-th residual of the GMRES method applied to* $\mathsf{A}x = b$ *with initial guess* $x_0 = 0$ *be denoted by* $r_n = p_n(\mathsf{A})b$ *with* $p_n \in \mathbb{P}_{n,0}$*.*

Then for all $\delta > \varepsilon := \|\mathsf{F}\|$*, the n-th residual* $\widehat{r}_n$ *of the GMRES method applied to* $(\mathsf{A} + \mathsf{F})\widehat{x} = b$ *with initial guess* $x_0 = 0$ *satisfies*

$$\frac{\|\widehat{r}_n\| - \|r_n\|}{\|b\|} \leq \frac{\varepsilon}{\delta - \varepsilon}\,\frac{|\partial\Lambda_\delta(\mathsf{A})|}{2\pi\delta}\sup_{\lambda\in\Lambda_\delta(\mathsf{A})}|p_n(\lambda)|. \tag{3.59}$$

Proof. See above or the original proof in [153]. □

Sifuentes, Embree and Morgan noted in [153], that the bound in theorem 3.39 has major parts in common with the pseudospectral GMRES bound for an unperturbed linear system in theorem 2.60. Furthermore, they provided another bound that resembles the bound in theorem 2.60 even more closely. However, this bound is not of relevance here and it is referred to [153] for details.

Similar to the pseudospectral GMRES bound in theorem 2.60, theorem 3.39 does not provide an answer to the question of how $\delta > \varepsilon$ should be chosen. Again, on the one hand, δ should be chosen larger than ε such that the factor $\frac{1}{\delta-\varepsilon}$ becomes small. On the other hand δ should be close to ε such that the pseudospectrum $\Lambda_\delta(\mathsf{A})$ and with it its boundary length $|\partial\Lambda_\delta(\mathsf{A})|$ and the supremum of the polynomial on the pseudospectrum does not grow too large.

Before moving on, theorem 3.39 is discussed with regard to example 3.38.

Example 3.40. In example 3.38, the convergence of the GMRES method was severely delayed by a perturbation of the size $\varepsilon = \|\mathsf{F}\| = 10^{-8}$. The inequality in theorem 3.39 can only bound the residual for GMRES applied to the perturbed linear system for the first 3 iterations because then the exact solution of the unperturbed linear system is found. In the third iteration, the GMRES polynomial for the unperturbed linear system is given by

$$p_n(\lambda) = 1 - \lambda^3 = \prod_{j=1}^{3}(1 - \frac{\lambda}{\zeta_3^j}),$$

where $\zeta_k = e^{\frac{2\pi \mathrm{i}}{k}}$ is an k-th root of unity for $k \in \mathbb{N}$. Note that ζ_3, ζ_3^2 and ζ_3^3 are the eigenvalues of $\mathbf{B}$. However, the polynomial has no roots close to the eigenvalues $\alpha\zeta_n, \alpha\zeta_n^2, \ldots, \alpha\zeta_n^n$ of $\mathbf{C}$ and the supremum in bound (3.59) is always greater than $\alpha^3 - 1 = 10^6 - 1$. In figure 3.8, the bound is evaluated numerically for a wide range of δ. The minimal value of the bound is $1.601 \cdot 10^{-3}$ and thus differs by one order of magnitude from the level of the residual norm stagnation in figure 3.7a which is 10^{-4}. However, beyond the third iteration, the bound can give no insight into the convergence behavior of the perturbed linear system. This example is constructed such that the GMRES method for the unperturbed problem has gathered no "knowledge" about the spectrum of the submatrix $\mathbf{C}$ and the supremum in bound (3.59) cannot be small. The bound can only give more insightful answers to the question of how perturbations affect the GMRES convergence if the GMRES polynomial p_n has zeros distributed across the full spectrum of A and not only a small part of it that is well-separated from the rest.

The construction of the boundaries of pseudospectra is achieved with the free software Python package PseudoPy [61] which was developed by the author. PseudoPy is a Python version of the original EigTool [184] by Wright. Details on the algorithms in PseudoPy and EigTool can be found in the works by Trefethen [175, 176] and his coauthors Toh [174] and Wright [185, 186]. An extensive treatment of the computation of pseudospectra can be found in the book of Trefethen and Embree [177]. In this example, the matrix A is normal and hence the pseudospectrum is the union of circles around the eigenvalues of A:

$$\Lambda_\delta(\mathsf{A}) = \bigcup_{\lambda \in \Lambda(\mathsf{A})} \{\mu \in \mathbb{C} \mid |\lambda - \mu| < \delta\}.$$

The following theorem generalizes the result of Sifuentes, Embree and Morgan to the case where not only the linear operator A but also the right hand side b and the initial guess x_0 can be perturbed.

Theorem 3.41. *Assume that $\mathsf{A}x = b$ is a linear system with an operator $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, right hand side $b \in \mathcal{H}$ and let $x_0 \in \mathcal{H}$ be an initial guess with corresponding initial residual $r_0 = b - \mathsf{A}x_0$. For $n \leq d(\mathsf{A}, r_0)$, let the n-th residual of the GMRES method applied to $\mathsf{A}x = b$ with initial guess x_0 be denoted by $r_n = p_n(\mathsf{A})r_0$ with $p_n \in \mathbb{P}_{n,0}$.*

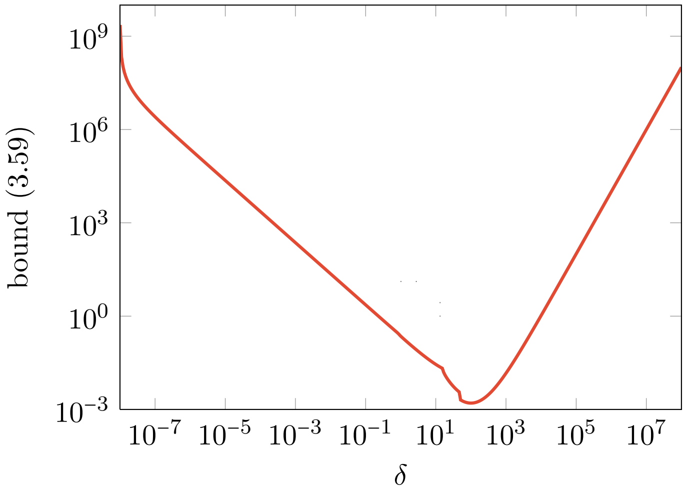


Figure 3.8.: Evaluation of the bound (3.59) for example 3.40 with $\delta \in]10^{-8}, 10^8]$. The minimum is attained for $\delta \approx 96.96$ where the bound evaluates to $1.601 \cdot 10^{-3}$. Note that the level of the residual norm stagnation in figure 3.7a is 10^{-4}.

Let $\widehat{\mathsf{A}}\widehat{x} = \widehat{b}$ be a perturbed linear system with $\widehat{\mathsf{A}} = \mathsf{A} + \mathsf{F}$ and $\widehat{b} = b + f$, where $\mathsf{F} \in \mathcal{L}(\mathcal{H})$ and $f \in \mathcal{H}$, and let $\widehat{x}_0 \in \mathcal{H}$ be an initial guess with corresponding residual $\widehat{r}_0 = \widehat{b} - \widehat{\mathsf{A}}\widehat{x}_0$.

Then for all $\delta > \varepsilon := \|\mathsf{F}\|$, the n-th residual $\widehat{r}_n$ of the GMRES method applied to $(\mathsf{A} + \mathsf{F})\widehat{x} = b + f$ with initial guess $\widehat{x}_0$ satisfies

$$\|\widehat{r}_n\| - \|r_n\| \le \frac{|\partial\Lambda_\delta(\mathsf{A})|}{2\pi\delta}\left(\frac{\varepsilon}{\delta - \varepsilon}\|\widehat{r}_0\| + \|\widehat{r}_0 - r_0\|\right) \sup_{\lambda\in\Lambda_\delta(\mathsf{A})} |p_n(\lambda)| \tag{3.60}$$

and also

$$\|\widehat{r}_n\| - \|r_n\| \le \frac{|\partial\Lambda_\delta(\widehat{\mathsf{A}})|}{2\pi\delta}\left(\frac{\varepsilon}{\delta - \varepsilon}\|r_0\| + \|\widehat{r}_0 - r_0\|\right) \sup_{\lambda\in\Lambda_\delta(\widehat{\mathsf{A}})} |p_n(\lambda)|.$$

Proof. The proof is very similar to the proof of theorem 3.39 and its major difference is that the separation of the operator polynomials and the initial residuals is deferred until the end. In order to prove inequality (3.60), let $\widehat{r}_n = \widehat{p}_n(\widehat{\mathsf{A}})\widehat{r}_0$ with $\widehat{p}_n \in \mathbb{P}_{n,0}$. The minimization property of $\widehat{p}_n$ and the triangle inequality yield

$$\|\widehat{r}_n\| - \|r_n\| = \left\|\widehat{p}_n(\widehat{\mathsf{A}})\widehat{r}_0\right\| - \|p_n(\mathsf{A})r_0\| \le \left\|p_n(\widehat{\mathsf{A}})\widehat{r}_0\right\| - \|p_n(\mathsf{A})r_0\| \le \left\|p_n(\widehat{\mathsf{A}})\widehat{r}_0 - p_n(\mathsf{A})r_0\right\|.$$

Because $\delta > \epsilon = \|\mathsf{F}\|$, the pseudospectrum $\Lambda_\delta(\mathsf{A})$ contains the spectra $\Lambda(\mathsf{A})$ and $\Lambda(\widehat{\mathsf{A}})$ and thus the operator polynomials can be expressed as Cauchy integrals over the pseudospectrum's boundary:

$$\|\widehat{r}_n\| - \|r_n\| \le \frac{1}{2\pi}\left\|\int_{\partial\Lambda_\delta(\mathsf{A})} p_n(\lambda)\left((\lambda\mathsf{id} - \widehat{\mathsf{A}})^{-1}\widehat{r}_0 - (\lambda\mathsf{id} - \mathsf{A})^{-1}r_0\right)d\lambda\right\|$$

$$\leq \frac{1}{2\pi} \int_{\partial\Lambda_\delta(\mathsf{A})} |p_n(\lambda)| \left\| (\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} \widehat{r}_0 - (\lambda \mathsf{id} - \mathsf{A})^{-1} r_0 \right\| d\lambda. \qquad (3.61)$$

Analogous to the proof of theorem 3.39, the first resolvent can be expressed as

$$(\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} = (\mathsf{id} - (\lambda \mathsf{id} - \mathsf{A})^{-1}\mathsf{F})^{-1}(\lambda \mathsf{id} - \mathsf{A})^{-1}.$$

By the definition of the pseudospectrum, $\delta = \frac{1}{\|(\lambda \mathsf{id} - \mathsf{A})^{-1}\|}$ for all $\lambda \in \partial\Lambda_\delta(\mathsf{A})$ and the above resolvent expression can be formulated with a Neumann series:

$$(\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} = \sum_{j=0}^{\infty} \left((\lambda \mathsf{id} - \mathsf{A})^{-1}\mathsf{F} \right)^j (\lambda \mathsf{id} - \mathsf{A})^{-1}.$$

This equation yields with the geometric series

$$\begin{aligned}
&\left\| (\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1} \widehat{r}_0 - (\lambda \mathsf{id} - \mathsf{A})^{-1} r_0 \right\| \\
&\quad = \left\| \sum_{j=1}^{\infty} \left((\lambda \mathsf{id} - \mathsf{A})^{-1}\mathsf{F} \right)^j (\lambda \mathsf{id} - \mathsf{A})^{-1} \widehat{r}_0 + (\lambda \mathsf{id} - \mathsf{A})^{-1} (\widehat{r}_0 - r_0) \right\| \\
&\quad \leq \frac{1}{\delta} \left(\sum_{j=1}^{\infty} \left(\frac{\varepsilon}{\delta} \right)^j \|\widehat{r}_0\| + \|\widehat{r}_0 - r_0\| \right) = \frac{1}{\delta} \left(\left(\frac{1}{1 - \frac{\varepsilon}{\delta}} - 1 \right) \|\widehat{r}_0\| + \|\widehat{r}_0 - r_0\| \right) \\
&\quad = \frac{1}{\delta} \left(\frac{\varepsilon}{\delta - \varepsilon} \|\widehat{r}_0\| + \|\widehat{r}_0 - r_0\| \right).
\end{aligned}$$

Inserting this inequality into (3.61) and bounding the polynomial yields the inequality (3.60).

The second inequality can be proved analogously by using the pseudospectrum $\Lambda_\delta(\widehat{\mathsf{A}})$ and expressing the resolvent $(\lambda \mathsf{id} - \mathsf{A})^{-1}$ with a Neumann series in terms of the resolvent $(\lambda \mathsf{id} - \widehat{\mathsf{A}})^{-1}$ and the perturbation F. □

Remark 3.42. Theorem 3.41 gives a bound on the absolute residual norm. Relative residual norms can be achieved trivially by pre-multiplying the linear systems with $\frac{1}{\|b\|}$ and $\frac{1}{\|\widehat{b}\|}$, respectively.

The operator A is not required to be nonsingular in theorem 3.41 but the statement becomes trivial if $0 \in \Lambda(\mathsf{A}) \subseteq \Lambda_\delta(\mathsf{A})$. Nevertheless, allowing the operator A to be nonsingular in the above formulation eases the application of the theorem in section 3.4.3.

Also note that if the right hand sides coincide and the zero initial guess is used for both linear systems in theorem 3.41, i.e., $b = \widehat{b}$ and $x_0 = \widehat{x}_0 = 0$, then the bound (3.60) reduces to the bound by Sifuentes, Embree and Morgan in theorem 3.39.

Clearly, theorem 3.41 also holds for MINRES if the operators A and F are self-adjoint. In this case, the pseudospectrum in inequality (3.60) can be replaced by the union of δ-intervals around the eigenvalues of A, i.e.,

$$\Lambda_\delta(\mathsf{A}) = \bigcup_{\lambda \in \Lambda(\mathsf{A})} [\lambda - \delta, \lambda + \delta].$$

Similarly, the approach of theorem 3.41 can be applied to the A-norm of the error in the CG method. The derivation is analogous and is not presented here.

Another perturbation result for Krylov subspaces is based on theorem 3.41 in section 3.4.3 in the context of selection strategies of deflation vectors for sequences of linear systems.

3.4. Selection of recycling vectors

In the introduction of chapter 3, the question of recycling data in Krylov subspace methods was split into two separate questions. The first one was the question of how to incorporate external data into a Krylov subspace method in order to influence its convergence behavior. This question was addressed in sections 3.1 and 3.2 with a focus on deflated Krylov subspace methods. The previous section 3.3 paved the way to attack the second question, which is: which data from a dataset with possible recycling data results in the best overall performance? Since this thesis concentrates on deflation, the question boils down to the following: given a subspace $\mathcal{W} \subseteq \mathcal{H}$ with candidates for recycling data, select a deflation subspace $\mathcal{U} \subseteq \mathcal{W}$ such that the deflated Krylov subspace method performs best.

Before moving on, it has to be made clear what is asked for by stating more precisely what "best performance" means here. When analyzing deflated Krylov subspace methods, it appears to be mathematically sound to measure the number of iterations that are needed in order to reduce the residual below a prescribed tolerance. Of course, the goal is in practice to reduce the time that is needed to solve the linear system up to a given tolerance with the available memory. Thus, instead of counting the raw number of iterations, the *computational cost* can be measured in terms of computation time and memory requirements. With respect to the computational cost, some thoughts have to be kept in mind. The first addresses the tolerance itself. In many applications, the linear systems are solved up to a very high accuracy which is often unnecessary because other errors like discretization errors for discretized partial differential equations or errors in parameters or measurements dominate the overall error. Thus, the tolerance should be adapted to the properties of the underlying problem, see, e.g., the articles by Arioli, Noulard and Russo [4] and Arioli, Loghin and Wathen [3] for linear systems and the PhD thesis of Międlar [117] for eigenvalue problems in the context of partial differential equations. The second issue is that the finite precision behavior of Krylov subspace methods may deviate significantly from their exact arithmetic counterparts, see section 2.10. Another issue is that counting the overall number of floating point operations (*flops*) is often impossible because of complex preconditioners, e.g., (algebraic) multigrid preconditioners that again constitute an iterative method with possibly varying runtime performance inside the Krylov subspace method.

Unfortunately, the convergence behavior of Krylov subspace methods is hard to predict accurately a priori (cf. sections 2.8 and 2.9) and thus – even if the above problems were non-existent – the determination of an *optimal* deflation subspace seems to be far from possible with today's knowledge. It comes as no real surprise that simple heuristics are customary in the literature for the selection of deflation subspaces. As already mentioned in section 3.2.1, there is a strong focus on using eigenvector approximations for the deflation subspace, e.g., Ritz or harmonic Ritz vectors from a given subspace $\mathcal{W}$.

For a prescribed integer m, Parks et al. [135] as well as Wang, de Sturler and Paulino [181] chose m harmonic Ritz vectors from previously computed Krylov subspaces in the GCRO-DR and the RMINRES method (cf. Morgan [120] and section 3.2.6). An analogous strategy is performed in the context of sequences of shifted linear systems in the works of Darnell, Morgan and Wilcox [25] and Soodhalter, Szyld and Xue [160]. In [63], the author and Schlömer used m Ritz vectors in the deflated MINRES method (cf. the second variant of the deflated MINRES method in section 3.2.5). The heuristic of picking the m (harmonic) Ritz vectors that correspond to the (harmonic) Ritz values of smallest magnitude works well in the above cases, but a major drawback of all methods is that the number of chosen deflation vectors m has to be defined in advance. In each iteration, the application of the projections $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}$ and $\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}$ with an m-dimensional subspace $\mathcal{U}$ to a vector needs at least m inner products and m vector updates. Furthermore, adding a "bad" Ritz vector to the deflation subspace may result in no decrease but instead in an increase of the number of iterations. For example, in the context of Newton's method for a nonlinear Schrödinger equation, it was noticed in [63] that the strategy of always choosing m deflation vectors does not pay off in the early phase of Newton's method because of large changes in the operators and because the Ritz vectors are poor eigenvector approximations in the first steps of Newton's method.

As mentioned above, the determination of an *optimal* deflation subspace without any heuristics seems to be impossible today. However, the perturbation theory that was developed in section 3.3 can be used to assess a given deflation subspace and to drive the existing heuristics forward. A careful evaluation of given candidates for a deflation subspace can lead to deflated Krylov subspace methods that automatically determine the number of needed deflation vectors in order to reduce the overall computation time. In order to do so, such methods may need to keep track of timings of potentially costly operations inside the iterations, such as the application of the operator A and a preconditioner. The construction of such estimations is the subject of this section.

In order to keep the notational overhead at a minimum, only two subsequent linear systems of a sequence (3.1) are considered in this section. The two linear systems are

$$\mathsf{A}x = b \qquad \text{and} \qquad \mathsf{B}y = c \tag{3.62}$$

with nonsingular operators $\mathsf{A}, \mathsf{B} \in \mathcal{L}(\mathcal{H})$ and right hand sides $b, c \in \mathcal{H}$. Furthermore, it is assumed that the first linear system $\mathsf{A}x = b$ has been solved approximately with a well-defined deflated Krylov subspace method from section 3.2, cf. table 3.1 for an overview. Thus, for an initial guess x_0 and a deflation subspace $\mathcal{U} \subseteq \mathcal{H}$ with

$\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, the underlying Krylov subspace method was not applied to $\mathsf{A}x = b$ but to the deflated linear system

$$\widehat{\mathsf{A}}\widehat{x} = \widehat{b}, \tag{3.63}$$

where $\widehat{\mathsf{A}} = \mathsf{P}\mathsf{A}$ and $\widehat{b} = \mathsf{P}b$ for a $\mathsf{P} \in \{\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}, \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}\}$. If no deflation space is used, e.g., for the first linear system, then all results hold with $\mathcal{U} = \{0\}$ and $\mathsf{P} = \mathsf{id}$. Let $n \in \mathbb{N}$ be the number of required iterations and x_n the approximate solution of the original linear system (3.62), i.e., $x_n = c(\overline{x}_n)$ is the corrected version of the approximate solution $\overline{x}_n \in x_0 + \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0)$ of the deflated linear system (3.63) with $\widehat{r}_0 = \widehat{b} - \widehat{\mathsf{A}}x_0$.

The next subsection deals with the extraction of Ritz and harmonic Ritz pairs for the "undeflated" operator A from the data that was generated while solving the first deflated linear system (3.63), i.e., with the deflated operator $\mathsf{P}\mathsf{A}$. The two subsequent subsections describe new strategies to assess a given set of deflation vectors. The first strategy is to use the perturbation theory for the spectrum of deflated operators from section 3.3.2 in order to evaluate a priori bounds for the convergence of the deflated Krylov subspace method for the next linear system, e.g., the CG or MINRES bounds in sections 2.8 and 2.9.2. The second strategy is to extract an Arnoldi relation for a deflated linear system that is "close" to the deflated linear system that is about to be solved by only using data from the linear system that has already been solved. Then the perturbation theory for Krylov subspace methods from section 3.3.3 can be applied in order to estimate a few steps of the convergence behavior of the next linear system with the given set of deflation vectors.

3.4.1. Computation of Ritz and harmonic Ritz pairs

Before discussing the actual selection of recycling vectors, this subsection briefly describes how Ritz and harmonic Ritz pairs can be obtained in the setting of a sequence of linear systems. The computation of Ritz and harmonic Ritz pairs is a fairly straightforward application of the results in section 2.5 and is only slightly complicated by the fact that Ritz or harmonic Ritz pairs have to be computed for the original operator A but the Krylov subspace has been built with a projected operator $\mathsf{P}\mathsf{A}$. This constellation makes the presentation in this subsection look very technical but the underlying ideas are simple and the extraction procedures for Ritz and harmonic Ritz pairs are easy to implement.

In this section, it is assumed that

$$V_{n+1} = [v_1, \dots, v_{n+1}] \in \mathcal{H}^{n+1} \quad \text{and} \quad \underline{\mathbf{H}}_n = \begin{bmatrix} \mathbf{H}_n \\ h_{n+1,n} e_n^\mathsf{T} \end{bmatrix} \in \mathbb{C}^{n+1,n} \tag{3.64}$$

define an Arnoldi relation for the Krylov subspace $\mathcal{K}_n = \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0)$, cf. section 2.7. If the Krylov subspace $\mathcal{K}_n$ is A-invariant, it is assumed that $V_n = [v_1, \dots, v_n] \in \mathcal{H}^n$ and $\mathbf{H}_n \in \mathbb{C}^{n,n}$ define an invariant Arnoldi relation for $\mathcal{K}_n$. Because the invariant Arnoldi relation is a rare case in practice, only the regular Arnoldi relation is considered for reasons of better readability. The invariant Arnoldi relation can always be treated analogously and only leads to minor modifications of the statements and their proofs. The data for the Arnoldi relation is generated explicitly or implicitly in all Krylov subspace methods, e.g., it is directly available after n steps of the GMRES and MINRES algorithms. If A is self-adjoint, the Arnoldi relation reduces to the Lanczos relation, i.e., $\mathbf{H}_n = \mathbf{H}_n^{\mathsf{H}}$, and the tridiagonality of $\mathbf{H}_n$ leads to short recurrences in the CG and MINRES algorithms. Although these methods only require to store 3 vectors, all Arnoldi/Lanczos vectors are assumed to be available for deflation purposes. In the CG algorithm, the Lanczos basis V_{n+1} and Lanczos matrix $\underline{\mathbf{H}}_n$ can be recovered efficiently from the quantities that are computed in algorithm 2.4. The extraction of the Lanczos relation from the CG method is, e.g., implemented in `krypy.linsys.Cg` in [60].

The subspaces that are considered here for the extraction of Ritz and harmonic Ritz pairs are the Krylov subspace $\mathcal{K}_n = \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{r}_0)$ and the deflation subspace $\mathcal{U}$ that defines the used projection P. From these two subspaces the Ritz pairs can be computed as follows in the case $\mathsf{P} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$.

Lemma 3.43 (Ritz pairs with $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$)**.** *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $v \in \mathcal{H}$ and $U \in \mathcal{H}^m$ with $\langle U, U\rangle = \mathbf{I}_m$ and $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, where $\mathcal{U} = [\![U]\!]$. Assume that $V_{n+1} \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}_n \in \mathbb{C}^{n+1,n}$ define an Arnoldi relation for $\mathcal{K}_n = \mathcal{K}_n(\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}, \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}v)$ with the usual decomposition* (3.64)*. Let $\mathbf{Y} = [y_1, \ldots, y_{n+m}] \in \mathbb{C}^{n+m,n+m}$ and $\mu_1, \ldots, \mu_{n+m} \in \mathbb{C}$ solve the eigenvalue problem*

$$\begin{bmatrix} \mathbf{H}_n + \mathbf{B}\mathbf{E}^{-1}\mathbf{C} & \mathbf{B} \\ \mathbf{C} & \mathbf{E} \end{bmatrix} \mathbf{Y} = \mathbf{Y} \operatorname{diag}(\mu_1, \ldots, \mu_{n+m}) \tag{3.65}$$

with $\mathbf{B} = \langle V_n, \mathsf{A}U\rangle$, $\mathbf{E} = \langle U, \mathsf{A}U\rangle$ and $\mathbf{C} = \langle U, \mathsf{A}V_n\rangle$.

Then $(w_1, \mu_1), \ldots, (w_{n+m}, \mu_{n+m})$ are the Ritz pairs of A with respect to $\mathcal{K}_n + \mathcal{U}$, where $w_i = [V_n, U]y_i$ for $i \in \{1, \ldots, n+m\}$. The Ritz residuals satisfy

$$\|\mathsf{A}w_i - \mu_i w_i\| = \sqrt{(\mathbf{G}_i y_i)^{\mathsf{H}} \mathbf{S} \mathbf{G}_i y_i},$$

where

$$\mathbf{G}_i = \begin{bmatrix} \underline{\mathbf{H}}_n - \mu_i \underline{\mathbf{I}}_n & 0 \\ \mathbf{E}^{-1}\mathbf{C} & \mathbf{I}_m \\ 0 & -\mu_i \mathbf{I}_m \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \mathbf{I}_{n+1} & \underline{\mathbf{B}} & 0 \\ \underline{\mathbf{B}}^{\mathsf{H}} & \mathbf{F} & \mathbf{E} \\ 0 & \mathbf{E}^{\mathsf{H}} & \mathbf{I}_m \end{bmatrix},$$

$$\underline{\mathbf{I}}_n = \begin{bmatrix} \mathbf{I}_n \\ 0 \end{bmatrix}, \quad \underline{\mathbf{B}} = \langle V_{n+1}, \mathsf{A}U\rangle = \begin{bmatrix} \mathbf{B} \\ \langle v_{n+1}, \mathsf{A}U\rangle \end{bmatrix} \quad \text{and} \quad \mathbf{F} = \langle \mathsf{A}U, \mathsf{A}U\rangle.$$

Proof. First note that $[V_n, U]$ is orthonormal and that

$$\begin{aligned} & \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}} \mathsf{A} V_n = V_{n+1} \underline{\mathbf{H}}_n \\ \Longleftrightarrow \quad & \mathsf{A}V_n = V_{n+1}\underline{\mathbf{H}}_n + \mathsf{P}_{\mathsf{A}\mathcal{U}, \mathcal{U}^\perp} \mathsf{A}V_n = V_{n+1}\underline{\mathbf{H}}_n + \mathsf{A}U \langle U, \mathsf{A}U\rangle^{-1} \langle U, \mathsf{A}V_n\rangle \\ & \qquad = V_{n+1}\underline{\mathbf{H}}_n + \mathsf{A}U\mathbf{E}^{-1}\mathbf{C}. \end{aligned}$$

That $(w_i, \mu_i)_{i\in\{1,\dots,n+m\}}$ are the Ritz pairs of A with respect to $\mathcal{K}_n + \mathcal{U}$ follows directly from remark 2.30 by noticing that

$$\langle [V_n, U], \mathsf{A}[V_n, U]\rangle = \begin{bmatrix} \langle V_n, \mathsf{A}V_n\rangle & \langle V_n, \mathsf{A}U\rangle \\ \langle U, \mathsf{A}V_n\rangle & \langle U, \mathsf{A}U\rangle \end{bmatrix} = \begin{bmatrix} \mathbf{H}_n + \mathbf{B}\mathbf{E}^{-1}\mathbf{C} & \mathbf{B} \\ \mathbf{C} & \mathbf{E} \end{bmatrix}.$$

It remains to show the residual norm identity. Therefore observe that

$$\begin{aligned} \mathsf{A}w_i - \mu_i w_i = \mathsf{A}[V_n, U]y_i - \mu_i[V_n, U]y_i &= [V_{n+1}\underline{\mathbf{H}}_n + \mathsf{A}U\mathbf{E}^{-1}\mathbf{C}, \mathsf{A}U]y_i - \mu_i[V_n, U]y_i \\ &= [V_{n+1}, \mathsf{A}U, U] \begin{bmatrix} \underline{\mathbf{H}}_n - \mu_i \underline{\mathbf{I}}_n & 0 \\ \mathbf{E}^{-1}\mathbf{C} & \mathbf{I}_m \\ 0 & -\mu_i \mathbf{I}_m \end{bmatrix} y_i = [V_{n+1}, \mathsf{A}U, U]\mathbf{G}_i y_i. \end{aligned}$$

The proof is complete by computing

$$\|\mathsf{A}w_i - \mu_i w_i\|^2 = (\mathbf{G}_i y_i)^{\mathsf{H}} \langle [V_{n+1}, \mathsf{A}U, U], [V_{n+1}, \mathsf{A}U, U]\rangle \mathbf{G}_i y_i = (\mathbf{G}_i y_i)^{\mathsf{H}} \mathbf{S}\mathbf{G}_i y_i.$$

Note that $\mathbf{S} = \langle [V_{n+1}, \mathsf{A}U, U], [V_{n+1}, \mathsf{A}U, U]\rangle$ is Hermitian and positive semidefinite. □

The following remark gives details on the impact of the Ritz pair computation on the computational cost.

Remark 3.44. The matrix $\mathbf{E} = \langle U, \mathsf{A}U\rangle$ has to be formed anyway in order to construct the projection $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$. The matrix $\mathbf{C}$ can be retrieved from the Arnoldi or Lanczos algorithm because in each iteration $1 \le i \le n$, the operator $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}$ has to be applied to the i-th Arnoldi/Lanczos vector v_i and from

$$\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}v_i = \mathsf{A}v_i - \mathsf{A}U\mathbf{E}^{-1}\langle U, \mathsf{A}v_i\rangle$$

it becomes clear that the columns of the matrix $\mathbf{C}$ are formed one after another during the iteration. In the CG algorithm, the matrix $\mathbf{C}$ can also be formed efficiently with a three-term recurrence which is implemented in [60] in `krypy.deflation.DeflatedCg`. Furthermore, note that if A is self-adjoint, then $\mathbf{C} = \mathbf{B}^{\mathsf{H}}$ and $\mathbf{E} = \mathbf{E}^{\mathsf{H}}$. Only if A is non-self-adjoint, $n \cdot m$ additional inner products have to be performed in order to compute the matrix $\mathbf{B} = \langle V_n, \mathsf{A}U\rangle$. However, no additional applications of A are necessary because $\mathsf{A}U$ can and should be kept in memory. For the computation of the Ritz residual norms, $\frac{m(m+1)}{2}$ additional inner products have to be computed for $\langle v_{n+1}, \mathsf{A}U\rangle$ and the matrix $\mathbf{F} = \langle \mathsf{A}U, \mathsf{A}U\rangle$ but again no applications of A are necessary.

The harmonic Ritz pairs of A with respect to $\mathcal{K}_n + \mathcal{U}$ can be computed as follows in the case $\mathsf{P} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$.

Lemma 3.45 (Harmonic Ritz pairs with $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$)**.** *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be nonsingular, $v \in \mathcal{H}$ and $U \in \mathcal{H}^m$ with $\langle U, U\rangle = \mathbf{I}_m$ and $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, where $\mathcal{U} = [\![U]\!]$. Assume that $V_{n+1} \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}_n \in \mathbb{C}^{n+1,n}$ define an Arnoldi relation for $\mathcal{K}_n =$*

$\mathcal{K}_n(\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}, \mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}v)$ *with the usual decomposition* (3.64). *Let* $\mathbf{Y} = [y_1, \ldots, y_{n+m}] \in \mathbb{C}^{n+m,n+m}$ *and* $\sigma_1, \ldots, \sigma_{n+m} \in \mathbb{C}$ *solve the generalized eigenvalue problem*

$$\begin{bmatrix} \mathbf{H}_n + \mathbf{B}\mathbf{E}^{-1}\mathbf{C} & \mathbf{B} \\ \mathbf{C} & \mathbf{E} \end{bmatrix}^{\mathsf{H}} \mathbf{Y} = (\mathbf{L}^{\mathsf{H}}\mathbf{K}\mathbf{L})\mathbf{Y}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_{n+m})$$

with $\mathbf{L} = \begin{bmatrix} \underline{\mathbf{H}}_n & 0 \\ \mathbf{E}^{-1}\mathbf{C} & \mathbf{I}_m \end{bmatrix}$, $\mathbf{K} = \begin{bmatrix} \mathbf{I}_{n+1} & \underline{\mathbf{B}} \\ \underline{\mathbf{B}}^{\mathsf{H}} & \mathbf{F} \end{bmatrix}$, *where the matrices* $\mathbf{B}, \underline{\mathbf{B}}, \mathbf{E}, \mathbf{C}, \mathbf{F}$ *are defined as in lemma 3.43.*

Then with

$$\mu_i := \begin{cases} \frac{1}{\sigma_i} & \text{if } \sigma_i \neq 0 \\ \infty & \text{else} \end{cases} \qquad \text{and} \quad w_i = [V_n, U]y_i \quad \text{for } i \in \{1, \ldots, n+m\},$$

the pairs $(w_1, \mu_1), \ldots, (w_{n+m}, \mu_{n+m})$ *are the harmonic Ritz pairs of* A *with respect to* $\mathcal{K}_n + \mathcal{U}$. *If* $\mu_i \neq \infty$ *for a* $i \in \{1, \ldots, n+m\}$, *then the Ritz pair* (w_i, μ_i) *satisfies*

$$\|\mathsf{A}w_i - \mu_i w_i\| = \sqrt{(\mathbf{G}_i y_i)^{\mathsf{H}}\mathbf{S}\mathbf{G}_i y_i} = \sqrt{|\mu_i(\mu_i - \rho_i)|} \leq |\mu_i|,$$

where the matrices $\mathbf{G}_i, \mathbf{S}$ *are defined as in lemma 3.43 and*

$$\rho_i = y_i^{\mathsf{H}} \begin{bmatrix} \mathbf{H}_n + \mathbf{B}\mathbf{E}^{-1}\mathbf{C} & \mathbf{B} \\ \mathbf{C} & \mathbf{E} \end{bmatrix} y_i.$$

Proof. According to definition (2.32) and equation (2.11), the generalized eigenvalue problem

$$\langle \mathsf{A}S, S\rangle \mathbf{Y} = \langle \mathsf{A}S, \mathsf{A}S\rangle \mathbf{Y}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_{n+m})$$

can be solved with $S = [V_n, U]$ in order to obtain the desired harmonic Ritz pairs. The matrix on the left hand side is just the Hermitian transpose of the matrix in (3.65). For the matrix on the right hand side, observe that

$$\mathsf{A}[V_n, U] = [V_{n+1}, \mathsf{A}U] \begin{bmatrix} \underline{\mathbf{H}}_n & 0 \\ \mathbf{E}^{-1}\mathbf{C} & \mathbf{I}_m \end{bmatrix} = [V_{n+1}, \mathsf{A}U]\mathbf{L}$$

and thus

$$\begin{aligned} \langle \mathsf{A}[V_n, U], \mathsf{A}[V_n, U]\rangle &= \mathbf{L}^{\mathsf{H}}\langle [V_{n+1}, \mathsf{A}U], [V_{n+1}, \mathsf{A}U]\rangle \mathbf{L} \\ &= \mathbf{L}^{\mathsf{H}} \begin{bmatrix} \langle V_{n+1}, V_{n+1}\rangle & \langle V_{n+1}, \mathsf{A}U\rangle \\ \langle \mathsf{A}U, V_{n+1}\rangle & \langle \mathsf{A}U, \mathsf{A}U\rangle \end{bmatrix} \mathbf{L} = \mathbf{L}^{\mathsf{H}}\mathbf{K}\mathbf{L}. \end{aligned}$$

The first residual norm equality is analogous to the corresponding part of the proof of lemma 3.43 while the second equality and the inequality follow from lemma 2.33. □

If the projection $\mathsf{P} = \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}$ is used, the Ritz pairs can be obtained as described in the following lemma:

Lemma 3.46 (Ritz pairs with $\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}$)**.** *Let* $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $v \in \mathcal{H}$ *and* $U \in \mathcal{H}^m$ *with* $\langle U, U\rangle = \mathbf{I}_m$ *and* $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, *where* $\mathcal{U} = \llbracket U \rrbracket$*. Furthermore, let* $\mathbf{R} \in \mathbb{C}^{m,m}$ *be such that* $Z = \mathsf{A}U\mathbf{R}$ *fulfills* $\langle Z, Z\rangle = \mathbf{I}_m$*. Assume that* $V_{n+1} \in \mathcal{H}^{n+1}$ *and* $\underline{\mathbf{H}}_n \in \mathbb{C}^{n+1,n}$ *define an Arnoldi relation for* $\mathcal{K}_n = \mathcal{K}_n(\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}\mathsf{A}, \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}v)$ *with the usual decomposition* (3.64)*. Let* $\mathbf{Y} = [y_1, \ldots, y_{n+m}] \in \mathbb{C}^{n+m,n+m}$ *and* $\mu_1, \ldots, \mu_{n+m} \in \mathbb{C}$ *solve the generalized eigenvalue problem*

$$\begin{bmatrix} \mathbf{H}_n & 0 \\ \underline{\widehat{\mathbf{B}}}^{\mathsf{H}} \underline{\mathbf{H}}_n + \mathbf{E}\mathbf{R}\widehat{\mathbf{C}} & \mathbf{E} \end{bmatrix} \mathbf{Y} = \begin{bmatrix} \mathbf{I}_n & \widehat{\mathbf{B}} \\ \widehat{\mathbf{B}}^{\mathsf{H}} & \mathbf{I}_m \end{bmatrix} \mathbf{Y} \operatorname{diag}(\mu_1, \ldots, \mu_{n+m}) \tag{3.66}$$

with $\underline{\widehat{\mathbf{B}}} = \langle V_{n+1}, U\rangle = \begin{bmatrix} \widehat{\mathbf{B}} \\ \langle v_{n+1}, U\rangle \end{bmatrix}$, $\widehat{\mathbf{C}} = \langle Z, \mathsf{A}V_n\rangle$, $\mathsf{E} = \langle U, \mathsf{A}U\rangle$ *and* $\underline{\mathbf{I}}_n = \begin{bmatrix} \mathbf{I}_n \\ 0 \end{bmatrix}$.

Then $(w_1, \mu_1), \ldots, (w_{n+m}, \mu_{n+m})$ *are the Ritz pairs of* A *with respect to* $\mathcal{K}_n + \mathcal{U}$, *where* $w_i = [V_n, U]y_i$ *for* $i \in \{1, \ldots, n+m\}$*. The Ritz residuals satisfy*

$$\|\mathsf{A}w_i - \mu_i w_i\| = \sqrt{(\widehat{\mathbf{G}}_i y_i)^{\mathsf{H}} \widehat{\mathbf{S}} \widehat{\mathbf{G}}_i y_i},$$

where

$$\widehat{\mathbf{G}}_i = \begin{bmatrix} \underline{\mathbf{H}}_n - \mu_i \underline{\mathbf{I}}_n & 0 \\ \widehat{\mathbf{C}} & \mathbf{R}^{-1} \\ 0 & -\mu_i \mathbf{I}_m \end{bmatrix}, \quad \text{and} \quad \widehat{\mathbf{S}} = \begin{bmatrix} \mathbf{I}_{n+1} & 0 & \underline{\widehat{\mathbf{B}}} \\ 0 & \mathbf{I}_m & (\mathbf{E}\mathbf{R})^{\mathsf{H}} \\ \underline{\widehat{\mathbf{B}}}^{\mathsf{H}} & \mathbf{E}\mathbf{R} & \mathbf{I}_m \end{bmatrix}.$$

Proof. First note that $\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp} x = x - Z\langle Z, x\rangle$ holds for $x \in \mathcal{H}$. Then $\mathcal{K}_n \subseteq \llbracket Z \rrbracket^\perp$ shows that $[V_n, Z]$ is orthogonal. From the Arnoldi relation and the definition of Z it can be seen that

$$\mathsf{A}[V_n, U] = [V_{n+1}\underline{\mathbf{H}}_n + Z\langle Z, \mathsf{A}V_n\rangle, \mathsf{A}U] = [V_{n+1}, Z] \underbrace{\begin{bmatrix} \underline{\mathbf{H}}_n & 0 \\ \widehat{\mathbf{C}} & \mathbf{R}^{-1} \end{bmatrix}}_{=:\widehat{\mathbf{L}}}. \tag{3.67}$$

With $S = [V_n, U]$, the Ritz pairs can be obtained by solving the generalized eigenvalue problem

$$\langle S, \mathsf{A}S\rangle \mathbf{Y} = \langle S, S\rangle \mathbf{Y} \operatorname{diag}(\mu_1, \ldots, \mu_{n+m}).$$

The inner product matrix on the left hand side is

$$\langle S, \mathsf{A}S\rangle = \langle [V_n, U], [V_{n+1}, Z]\rangle \widehat{\mathbf{L}} = \begin{bmatrix} \underline{\mathbf{I}}_n^{\mathsf{H}} & 0 \\ \underline{\widehat{\mathbf{B}}}^{\mathsf{H}} & \mathbf{E}\mathbf{R} \end{bmatrix} \widehat{\mathbf{L}} = \begin{bmatrix} \mathbf{H}_n & 0 \\ \underline{\widehat{\mathbf{B}}}^{\mathsf{H}} \underline{\mathbf{H}}_n + \mathbf{E}\mathbf{R}\widehat{\mathbf{C}} & \mathbf{E} \end{bmatrix}$$

while the matrix on the right hand side is

$$\langle [V_n, U], [V_n, U]\rangle = \begin{bmatrix} \mathbf{I}_n & \langle V_n, U\rangle \\ \langle U, V_n\rangle & \mathbf{I}_m \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \widehat{\mathbf{B}} \\ \widehat{\mathbf{B}}^{\mathsf{H}} & \mathbf{I}_m \end{bmatrix}.$$

It remains to show the residual norm equality. Note that

$$\mathsf{A}w_i - \mu_i w_i = [V_{n+1}, Z]\widehat{\mathbf{L}} y_i - \mu_i [V_n, U] y_i = [V_{n+1}, Z, U] \begin{bmatrix} \underline{\mathbf{H}}_n - \mu_i \underline{\mathbf{I}}_n & 0 \\ \widehat{\mathbf{C}} & \mathbf{R}^{-1} \\ 0 & -\mu_i \mathbf{I}_m \end{bmatrix} y_i$$

$$= [V_{n+1}, Z, U]\widehat{\mathbf{G}}_i y_i$$

and thus

$$\begin{aligned}\|\mathsf{A}w_i - \mu_i w_i\|^2 &= (\widehat{\mathbf{G}}_i y_i)^{\mathsf{H}}\langle [V_{n+1}, Z, U], [V_{n+1}, Z, U]\rangle \widehat{\mathbf{G}}_i y_i \\ &= (\widehat{\mathbf{G}}_i y_i)^{\mathsf{H}} \begin{bmatrix} \mathbf{I}_{n+1} & 0 & \langle V_{n+1}, U\rangle \\ 0 & \mathbf{I}_m & \langle Z, U\rangle \\ \langle U, V_{n+1}\rangle & \langle U, Z\rangle & \mathbf{I}_m \end{bmatrix} \widehat{\mathbf{G}}_i y_i = (\widehat{\mathbf{G}}_i y_i)^{\mathsf{H}} \widehat{\mathbf{S}} \widehat{\mathbf{G}}_i y_i.\end{aligned}$$

□

Remark 3.47. As noted in remark 3.44 for the case of $\mathsf{P} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}$, the matrix $\widehat{\mathbf{C}}$ is implicitly constructed in the Arnoldi/Lanczos algorithm. The matrices $\mathbf{R}$, $\widehat{\underline{\mathbf{B}}}$ and E have to be computed but no applications of A are involved.

If A is self-adjoint, then the matrix on the left hand side of equation (3.66) is Hermitian, i.e., $\widehat{\underline{\mathbf{B}}}^{\mathsf{H}} \underline{\mathbf{H}}_n + \mathbf{E}\mathbf{R}\widehat{\mathbf{C}} = 0$.

Lemma 3.48 (Harmonic Ritz pairs with $\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}$)**.** *Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$ be nonsingular, $v \in \mathcal{H}$ and $U \in \mathcal{H}^m$ with $\langle U, U\rangle = \mathbf{I}_m$ and $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$, where $\mathcal{U} = [\![U]\!]$. Furthermore, let $\mathbf{R} \in \mathbb{C}^{m,m}$ be such that $Z = \mathsf{A}U\mathbf{R}$ fulfills $\langle Z, Z\rangle = \mathbf{I}_m$. Assume that $V_{n+1} \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}_n \in \mathbb{C}^{n+1,n}$ define an Arnoldi relation for $\mathcal{K}_n = \mathcal{K}_n(\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}\mathsf{A}, \mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp} v)$ with the usual decomposition* (3.64)*. Let $\mathbf{Y} = [y_1, \ldots, y_{n+m}] \in \mathbb{C}^{n+m,n+m}$ and $\sigma_1, \ldots, \sigma_{n+m} \in \mathbb{C}$ solve the generalized eigenvalue problem*

$$\begin{bmatrix} \mathbf{H}_n & 0 \\ \widehat{\underline{\mathbf{B}}}^{\mathsf{H}} \underline{\mathbf{H}}_n + \mathbf{E}\mathbf{R}\widehat{\mathbf{C}} & \mathbf{E} \end{bmatrix}^{\mathsf{H}} \mathbf{Y} = \widehat{\mathbf{L}}^{\mathsf{H}} \widehat{\mathbf{L}} \mathbf{Y} \operatorname{diag}(\sigma_1, \ldots, \sigma_{n+m}),$$

where $\widehat{\mathbf{L}} = \begin{bmatrix} \underline{\mathbf{H}}_n & 0 \\ \widehat{\mathbf{C}} & \mathbf{R}^{-1} \end{bmatrix}$ and $\widehat{\mathbf{B}}$, $\widehat{\mathbf{C}}$ and $\mathbf{E}$ are defined as in lemma 3.46.

Then with

$$\mu_i := \begin{cases} \frac{1}{\sigma_i} & \text{if } \sigma_i \neq 0 \\ \infty & \text{else} \end{cases} \quad \text{and} \quad w_i = [V_n, U] y_i \quad \text{for } i \in \{1, \ldots, n+m\},$$

the pairs $(w_1, \mu_1), \ldots, (w_{n+m}, \mu_{n+m})$ are the harmonic Ritz pairs of A with respect to $\mathcal{K}_n + \mathcal{U}$. If $\mu_i \neq \infty$ for a $i \in \{1, \ldots, n+m\}$, then the Ritz pair (w_i, μ_i) satisfies

$$\|\mathsf{A}w_i - \mu_i w_i\| = \sqrt{(\widehat{\mathbf{G}}_i y_i)^{\mathsf{H}} \widehat{\mathbf{S}} \widehat{\mathbf{G}}_i y_i} = \sqrt{|\mu_i(\mu_i - \rho_i)|} \leq |\mu_i|,$$

where the matrices $\widehat{\mathbf{G}}_i, \widehat{\mathbf{S}}$ are defined as in lemma 3.46 and

$$\rho_i = y_i^{\mathsf{H}} \begin{bmatrix} \mathbf{H}_n & 0 \\ \widehat{\underline{\mathbf{B}}}^{\mathsf{H}} \underline{\mathbf{H}}_n + \mathbf{E}\mathbf{R}\widehat{\mathbf{C}} & \mathbf{E} \end{bmatrix} y_i.$$

Proof. With $S = [V_n, U]$, the Ritz pairs can again be computed by solving the generalized eigenvalue problem

$$\langle \mathsf{A}S, S\rangle \mathbf{Y} = \langle \mathsf{A}S, \mathsf{A}S\rangle \mathbf{Y} \operatorname{diag}(\sigma_1, \dots, \sigma_{n+m}).$$

The inner product matrix on the left hand side is the Hermitian transpose of the matrix on the left hand side of equation (3.66). For the inner product matrix on the right hand side, note that equation (3.67) also holds with the assumptions of this lemma and the inner product becomes

$$\langle \mathsf{A}S, \mathsf{A}S\rangle = \widehat{\mathbf{L}}^{\mathsf{H}}\widehat{\mathbf{L}}.$$

The residual norm statement can be shown analogously to the corresponding part of the proof of lemma 3.45. □

In the literature, the computation of Ritz pairs or harmonic Ritz pairs has been carried out along the lines of the (generalized) eigenvalue problems in the above lemmas. Parks et al. [135] and Wang, de Sturler and Paulino [181] computed harmonic Ritz pairs similarly to lemma 3.48. It should be noted, that the possibility of an infinite harmonic Ritz value is largely overlooked in the literature when it comes to the actual computation of harmonic Ritz pairs, see also section 2.5. Regular Ritz pairs have been computed in the manner of lemma 3.43 by the author and Schlömer in [63].

3.4.2. Estimation with a priori bounds

In this subsection, it is assumed that both operators A and B of the linear systems in (3.62) are self-adjoint and that the first linear system has been solved approximately via the deflated linear system (3.63). Furthermore, it is assumed that Ritz or harmonic Ritz pairs of the original operator A have been computed as presented in the previous subsection. The aim of this subsection is to derive a strategy for selecting some of the computed Ritz or harmonic Ritz vectors in order to use them as a basis for the deflation subspace for the second linear system in (3.62). The selection strategy takes into account the a priori bounds for the CG and MINRES methods that have been presented in sections 2.8 and 2.9.2. For this subsection let $(w_1, \mu_1), \dots, (w_t, \mu_t) \in \mathcal{H} \times \mathbb{R}$ be the computed Ritz or harmonic Ritz pairs with corresponding residual norms $\|\mathsf{A}w_i - \mu_i w_i\|$ for $i \in \{1, \dots, t\}$. As usual, it is assumed that the Ritz or harmonic Ritz values are sorted such that $\mu_1 \leq \dots \leq \mu_t$ and that the corresponding vectors $w_1, \dots, w_t$ are normalized.

In section 3.3.2 it was shown for a self-adjoint operator A, that if an approximation $\mathcal{U} \subseteq \mathcal{H}$ to an A-invariant subspace with associated spectrum $\Lambda_1 \subset \mathbb{R}$ is used as the deflation subspace, then the spectrum of the deflated operator $\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}$ consists of zero and a perturbed version of the spectrum Λ_2 of A that is complementary to Λ_1. Theorem 3.36 shows that the difference of the eigenvalues can be bounded quadratically by the Ritz residual norm $\|R\| = \|\mathsf{P}_{\mathcal{U}^\perp}\mathsf{A}\mathsf{P}_{\mathcal{U}}\|$ of the subspace $\mathcal{U}$ and a the spectral gap between the Ritz values that correspond to the deflation subspace $\mathcal{U}$ and the spectrum Λ_2. The results of theorem 3.36 help to theoretically understand the spectrum of the deflated operator $\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}$ in the case where the spectrum of A is fully disclosed. In practice, this is usually not the case but some insight into the spectrum of A can be gained from the computed Ritz or harmonic Ritz values and their residuals. The following lemma is due to Cao, Xie and Li [21] and is an improved version of a theorem by Kahan [87]; see also Davis, Kahan and Weinberger [27].

Lemma 3.49. *Let $\mathsf{A} \in \mathcal{H}$ be self-adjoint, $\mathbf{M} = \mathbf{M}^\mathsf{H} \in \mathbb{C}^{m,m}$ and $Z \in \mathcal{H}^m$ of rank m. Let $\lambda_1 \leq \cdots \leq \lambda_N$ be the eigenvalues of A and let $\mu_1 \leq \cdots \leq \mu_m$ be the eigenvalues of $\mathbf{M}$.*

Then there exist indices $1 \leq j_1 < \cdots < j_m \leq N$ such that

$$\max_{i\in\{1,\ldots,m\}} |\lambda_{j_i} - \mu_i| \leq \frac{\|\mathsf{A}Z - Z\mathbf{M}\|}{\sigma_{\min}(Z)},$$

where $\sigma_{\min}(Z)$ is the smallest singular value of Z.

Proof. See Cao, Xie and Li [21]. □

Let $I = \{i_1 < \cdots < i_m\} \subset \{1,\ldots,t\}$, $J = \{j_1 < \cdots < j_{t-m}\} = \{1,\ldots,t\} \setminus I$ and define $W_I = [w_{i_1},\ldots,w_{i_m}]$, $\mathbf{M}_I = \operatorname{diag}(\mu_{i_1},\ldots,\mu_{i_m})$, $W_J = [w_{j_1},\ldots,w_{j_{t-m}}]$ and $\mathbf{M}_J = \operatorname{diag}(\mu_{j_1},\ldots,\mu_{j_{t-m}})$. Lemma 3.49 now guarantees that there exist indices $1 \leq k_1 < \cdots < k_m \leq N$ such that

$$\max_{s\in\{1,\ldots,m\}} |\lambda_{k_s} - \mu_{i_s}| \leq \frac{\|\mathsf{A}W_I - W_I\mathbf{M}_I\|}{\sigma_{\min}(W_I)} =: \delta_I,$$

where $\lambda_1 \leq \cdots \leq \lambda_N$ are the eigenvalues of A. Likewise, the lemma guarantees that there exist indices $1 \leq l_1 < \cdots < l_{t-m} \leq N$ such that

$$\max_{s\in\{1,\ldots,t-m\}} |\lambda_{l_i} - \mu_{l_i}| \leq \frac{\|\mathsf{A}W_J - W_J\mathbf{M}_J\|}{\sigma_{\min}(W_J)} =: \delta_J.$$

Assume that the Ritz or harmonic Ritz pairs satisfy the following spectral gap condition (cf. definition 3.31):

$$\delta_I + \delta_J < \delta(\{\mu_{i_1},\ldots,\mu_{i_m}\},\{\mu_{j_1},\ldots,\mu_{j_{t-m}}\}) =: \delta. \tag{3.68}$$

Then it can be concluded that there exist indices $1 \leq p_1 < \cdots < p_t \leq N$ such that

$$\max_{i\in I} |\lambda_{p_i} - \mu_i| \leq \delta_I \quad \text{and} \quad \max_{j\in J} |\lambda_{p_j} - \mu_j| \leq \delta_J.$$

The above condition (3.68) and a possible distribution of A's eigenvalues is visualized in figure 3.9. Lemma 3.49 can give valuable insight in the distribution of t eigenvalues of A but nothing can be said about the remaining $N - t$ eigenvalues in general.

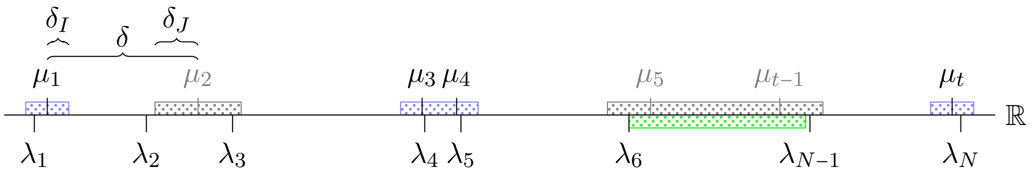


Figure 3.9.: Visualization of the spectral gap condition $\delta_I + \delta_J < \delta$ for Ritz or harmonic Ritz pairs $(w_1, \mu_1), \ldots, (w_t, \mu_t)$ and a possible distribution of the eigenvalues $\lambda_1, \ldots, \lambda_N$ of A, cf. equation (3.68). Lemma 3.49 states that the intervals around μ_1, μ_2 and μ_t each contain at least one eigenvalue of A while the union of the intervals around μ_3 and μ_4 contains at least two eigenvalues. The union of the intervals around $\mu_5, \ldots, \mu_{t-1}$ has to contain at least $t-5$ eigenvalues. Note that lemma 3.49 cannot provide information about the remaining $N-m$ eigenvalues and that eigenvalues like λ_2 can lie outside the intervals.

This is the point where the heuristics enter the analysis because the actual spectrum of A is unknown in general. If more is known about A's spectrum, e.g., containment intervals for the eigenvalues and information about the number of eigenvalues in these intervals, then this information can be used to improve the heuristics of this subsection or even construct actual a priori bounds. If nothing more is known about A's eigenvalues, then one strategy is to assume that the set $\bigcup_{i\in I}[\mu_i - \delta_I, \mu_i + \delta_I]$ contains exactly m eigenvalues of A and $\bigcup_{j\in J}[\mu_j - \delta_J, \mu_j + \delta_J]$ contains *all* remaining $N-m$ eigenvalues and not only the $t-m$ eigenvalues that are guaranteed to be contained by lemma 3.49.

In practice, it can often be observed that some Ritz or harmonic Ritz pairs approximate certain eigenpairs of A very well once the Krylov subspace method has found an approximate solution. For example, after the last iteration of MINRES with the undeflated linear system in example 3.23, the three Ritz pairs with smallest Ritz value magnitude approximate the three eigenpairs that correspond to the eigenvalues $\lambda_1 = -10^{-3}$, $\lambda_2 = -10^{-4}$ and $\lambda_3 = -10^{-5}$. The residual norms of these Ritz pairs lie around 10^{-11} while all other Ritz pairs exhibit large residual norms between 10^{-2} and 10^{-1}. It was shown by Simoncini and Szyld in [155], that MINRES enters the phase of superlinear convergence once some harmonic Ritz values approximate well the eigenvalues of smallest or largest magnitude. Often, the corresponding eigenvalues are well-separated from the remaining spectrum and the above heuristic can be considered as justified.

In the described situation, the following theorem narrows down the spectrum of the deflated operator $\mathsf{P}_{\mathcal{W}_I^\perp, \mathsf{B}\mathcal{W}_I}\mathsf{B}$ with the deflation subspace $\mathcal{W}_I = [\![W_I]\!]$ based on properties of the difference $\mathsf{F} = \mathsf{B} - \mathsf{A}$ and Ritz residuals. The statement appears to be complicated at first sight because of quite technical assumptions but its underlying arguments are simple and the train of thought is the following: with proper assumptions on A's spectrum and the available Ritz pairs, the spectrum of B is characterized with the difference F via Weyl's theorem, cf. theorem 3.28. Then the new quadratic residual bound from theorem 3.36 can be used in order to localize the eigenvalues of $\mathsf{P}_{\mathcal{W}_I^\perp,\mathsf{B}\mathcal{W}_I}\mathsf{B}$. To the knowledge of the author, no comparable result is known in the literature.

Theorem 3.50. *Let* $\mathsf{A},\mathsf{B} \in \mathcal{L}(\mathcal{H})$ *be self-adjoint with* $\Lambda(\mathsf{A}) = \{\lambda_1 \le \dots \le \lambda_N\}$, $\mathsf{F} = \mathsf{B} - \mathsf{A}$, $\Lambda(\mathsf{F}) = \{\epsilon_1 \le \dots \le \epsilon_N\}$ *and let* $(w_1,\mu_1),\dots,(w_t,\mu_t) \in \mathcal{H}\times\mathbb{R}$ *be* t *Ritz pairs of* A *with* $\|w_1\| = \dots = \|w_t\| = 1$. *For* $m \in \mathbb{N}$ *and an index set* $I = \{i_1 < \dots < i_m\} \subset \{1,\dots,t\}$, *let* $J = \{j_1 < \dots < j_{t-m}\} = \{1,\dots,t\} \setminus I$ *and define*

$$\begin{aligned} W_I &= [w_{i_1},\dots,w_{i_m}], & \mathcal{W}_I &= [\![W_I]\!], & \mathbf{D}_I &= \operatorname{diag}(\mu_{i_1},\dots,\mu_{i_m}), & \Lambda_I &= \{\mu_i\}_{i\in I},\\ W_J &= [w_{j_1},\dots,w_{j_{t-m}}], & \mathcal{W}_J &= [\![W_J]\!], & \mathbf{D}_J &= \operatorname{diag}(\mu_{j_1},\dots,\mu_{j_{t-m}}), & \Lambda_J &= \{\mu_j\}_{j\in J}. \end{aligned}$$

Assume that

$$\delta_I + \delta_J + \epsilon_N - \epsilon_1 < \delta(\Lambda_I,\Lambda_J) =: \delta \tag{3.69}$$

for

$$\delta_I = \frac{\|\mathsf{A}W_I - W_I\mathbf{D}_I\|}{\sigma_{\min}(W_I)} \quad \textit{and} \quad \delta_J = \frac{\|\mathsf{A}W_J - W_J\mathbf{D}_J\|}{\sigma_{\min}(W_J)}.$$

Furthermore, assume that there exists an index set $K = \{k_1 < \dots < k_m\} \subset \{1,\dots,N\}$ *with* $\{l_1 < \dots < l_{N-m}\} = \{1,\dots,N\} \setminus K$ *such that*

$$\lambda_{k_1},\dots,\lambda_{k_m} \in \bigcup_{i\in I}[\mu_i - \delta_I, \mu_i + \delta_I] \quad \textit{and} \quad \lambda_{l_1},\dots,\lambda_{l_{N-m}} \in \bigcup_{j\in J}[\mu_j - \delta_J, \mu_j + \delta_J]. \tag{3.70}$$

If

$$\mu_{\min} := \min_{\mu\in\bigcup_{i\in I}[\mu_i+\epsilon_1,\mu_i+\epsilon_N]} |\mu| > 0,$$

then $\mathsf{P}_{\mathcal{W}_J^\perp,\mathsf{B}\mathcal{W}_J}$ *is well defined and* $\Lambda(\mathsf{P}_{\mathcal{W}_J^\perp,\mathsf{B}\mathcal{W}_J}\mathsf{B}) = \{0\} \cup \{\widehat{\lambda}_1,\dots,\widehat{\lambda}_{N-m}\}$ *satisfies*

$$\widehat{\lambda}_i \in \bigcup_{j\in J}[\mu_j - \delta_J + \epsilon_1 - \eta, \mu_j + \delta_J + \epsilon_N + \eta] \tag{3.71}$$

with

$$\eta = \left(\delta_I + \left\|\mathsf{P}_{\mathcal{W}_I^\perp}\mathsf{F}\mathsf{P}_{\mathcal{W}_I}\right\|\right)^2 \left(\frac{1}{\delta - \delta_J - \epsilon_N + \epsilon_1} + \frac{1}{\mu_{\min}}\right).$$

Proof. Let the eigenvalues of B be denoted by $\overline{\lambda}_1 \le \dots \le \overline{\lambda}_N$. Because of Weyl's theorem (see theorem 3.28 the eigenvalues of B satisfy for $k \in \{1,\dots,N\}$

$$\overline{\lambda}_k \in [\lambda_k + \epsilon_1, \lambda_k + \epsilon_N].$$

With assumption (3.69) on the Ritz pairs and assumption (3.70) on A's eigenvalues, it follows that

$$\overline{\lambda}_{k_1},\dots,\overline{\lambda}_{k_m} \in \bigcup_{i\in I}[\mu_i - \delta_I + \epsilon_1, \mu_i + \delta_I + \epsilon_N]$$

and

$$\overline{\lambda}_{l_1},\dots,\overline{\lambda}_{l_{N-m}} \in \bigcup_{j\in J}[\mu_j - \delta_J + \epsilon_1, \mu_j + \delta_J + \epsilon_N]. \tag{3.72}$$

It is now shown that the projection $\mathsf{P}_{\mathcal{W}_I^\perp,\mathsf{B}\mathcal{W}_I}$ is well defined by analyzing the eigenvalue ν of smallest magnitude of

$$\langle W_I, \mathsf{B}W_I\rangle = \langle W_I, \mathsf{A}W_I\rangle + \langle W_I, \mathsf{F}W_I\rangle = \mathbf{D}_I + \langle W_I, \mathsf{F}W_I\rangle.$$

It follows from Weyl's theorem that $\nu \in \bigcup_{i\in I}[\mu_i + \epsilon_1, \mu_i + \epsilon_N]$ and thus $|\nu| \geq \mu_{\min} > 0$ by assumption. This means that $\langle W_I, \mathsf{B}W_I\rangle$ is nonsingular and thus the projection $\mathsf{P}_{\mathcal{W}_I^\perp,\mathsf{B}\mathcal{W}_I}$ is well defined.

Now let $W_\perp \in \mathcal{H}^{N-m}$ such that $\left\langle W_I^\perp, W_I^\perp\right\rangle = \mathbf{I}_{N-m}$ and $[\![W_I^\perp]\!] = \mathcal{W}_I^\perp$ and let $\mathbf{M} = \left\langle W_I^\perp, \mathsf{A}W_I^\perp\right\rangle$. The eigenvalues of $\operatorname{diag}(\mathbf{M}, \mathbf{D}_I)$ are assumed to be $\tilde\lambda_1 \leq \cdots \leq \tilde\lambda_N$ and the indices $1 \leq n_1 < \cdots < n_{N-m} \leq N$ are assumed to be given such that $\tilde\lambda_{n_1} \leq \cdots \leq \tilde\lambda_{n_{N-m}}$ are the eigenvalues of $\mathbf{M}$ and $1 \leq p_1 < \cdots < p_m \leq N$ such that $\tilde\lambda_{p_s} = \mu_{i_s}$ for $s \in \{1,\ldots,m\}$. Because $\left\|\mathsf{A}W_I^\perp - W_I^\perp\mathbf{M}\right\| = \|\mathsf{A}W_I - W_I\mathbf{D}_I\| = \delta_I$, the eigenvalues of $\operatorname{diag}(\mathbf{M}, \mathbf{D}_I)$ satisfy with lemma 3.49

$$\max_{i\in\{1,\ldots,N\}} |\lambda_i - \tilde\lambda_i| \leq \delta_I.$$

With assumption (3.70) it follows that

$$\tilde\lambda_{n_1},\ldots,\tilde\lambda_{n_{N-m}} \in \bigcup_{j\in J}[\mu_j - \delta_I - \delta_J, \mu_j + \delta_I + \delta_J].$$

Assumption (3.69) thus guarantees that $n_s = l_s$ for $s \in \{1,\ldots,N-m\}$. Now theorem 3.36 can be applied in order to bound the difference of the nonzero eigenvalues of $\mathsf{P}_{\mathcal{W}_I^\perp,\mathsf{B}\mathcal{W}_I}\mathsf{B}$ and the eigenvalues $\overline\lambda_{l_1},\ldots,\overline\lambda_{l_{N-m}}$ of B. This results in

$$\max_{s\in\{1,\ldots,N-m\}} |\overline\lambda_{l_s} - \widehat\lambda_s| \leq \|R\|^2\left(\frac{1}{\alpha} + \frac{1}{\beta}\right), \tag{3.73}$$

where

$$\begin{aligned} R &= \mathsf{B}W_I - W_I\langle W_I, \mathsf{B}W_I\rangle, \\ \alpha &= \delta(\{\overline\lambda_{l_1},\ldots,\overline\lambda_{l_{N-m}}\}, \Lambda(\langle W_I, \mathsf{B}W_I\rangle)) \\ \text{and}\quad \beta &= \min_{\mu\in\Lambda(\langle W_I,\mathsf{B}W_I\rangle)} |\mu|. \end{aligned}$$

Now note that by definition

$$\begin{aligned} \|R\| &= \left\|\mathsf{P}_{\mathcal{W}_I^\perp}\mathsf{B}\mathsf{P}_{\mathcal{W}_I}\right\| = \left\|\mathsf{P}_{\mathcal{W}_I^\perp}(\mathsf{A}+\mathsf{F})\mathsf{P}_{\mathcal{W}_I}\right\| \leq \left\|\mathsf{P}_{\mathcal{W}_I^\perp}\mathsf{A}\mathsf{P}_{\mathcal{W}_I}\right\| + \left\|\mathsf{P}_{\mathcal{W}_I^\perp}\mathsf{F}\mathsf{P}_{\mathcal{W}_I}\right\| \\ &\leq \delta_I + \left\|\mathsf{P}_{\mathcal{W}_I^\perp}\mathsf{F}\mathsf{P}_{\mathcal{W}_I}\right\|. \end{aligned}$$

If the eigenvalues of $\langle W_I, \mathsf{B}W_I\rangle$ are $\overline\mu_1 \leq \cdots \leq \overline\mu_m$, then with Weyl's theorem and Cauchy interlacing it follows that

$$\overline\mu_i \in [\mu_i + \epsilon_1, \mu_i + \epsilon_N]$$

for $i \in \{1, \dots, m\}$. With this observation it can be seen with equation (3.72) and assumption (3.69) that

$$\alpha \geq \delta\Big(\bigcup_{j \in J}[\mu_j - \delta_J + \epsilon_1, \mu_j + \delta_J + \epsilon_N], \bigcup_{i \in I}[\mu_i + \epsilon_1, \mu_i + \epsilon_N]\Big) \geq \delta - \delta_J - \epsilon_N + \epsilon_1 > 0.$$

It has already been shown in the course of the proof that $\beta \geq \mu_{\min} > 0$ and thus

$$\|R\|^2 \left(\frac{1}{\alpha} + \frac{1}{\beta}\right) \leq \eta.$$

The theorem's inclusion statement (3.71) now follows from (3.72) and (3.73). □

Remark 3.51. Theorem 3.50 can be improved trivially such that an individual interval for each eigenvalue is provided if the location of the original eigenvalues is known. This generalization is not presented here because the notation becomes even more cumbersome and the above version is sufficient for the intended use.

A brief discussion of the theorem seems to be appropriate. Roughly speaking, the theorem's assumptions are satisfied if the eigenvalues of A can be split into two parts: the first part consists of m eigenvalues and lies in the neighborhood of m Ritz values while the remaining $N - m$ eigenvalues lie in a (potentially large) neighborhood of the remaining $t - m$ Ritz values which are required to be separated from the other m Ritz values. The condition on $\mu_{\min}$ is required to ensure that the projection $\mathsf{P}_{\mathcal{W}_I^\perp, \mathsf{B}\mathcal{W}_I}$ is well defined. Note that throughout the theorem, the values ϵ_1 and ϵ_N can be replaced by $-\epsilon$ and ϵ, where $\epsilon = \max\{|\epsilon_1|, |\epsilon_N|\} = \|\mathsf{F}\|$. However, in certain cases, the more general formulation in the theorem pays off. For example, if several shifted linear systems have to be solved, then $\mathsf{F} = \alpha \mathsf{id}$ for an $\alpha \in \mathbb{R}$ and thus also all Ritz values are simply shifted and $\mu_{\min}$ can be determined exactly because $\epsilon_1 = \epsilon_N = \alpha$. Likewise, the intervals around the Ritz values in (3.71) are not enlarged by the perturbation F but just shifted in this case. The term $\left\|\mathsf{P}_{\mathcal{W}_I^\perp} \mathsf{F} \mathsf{P}_{\mathcal{W}_I}\right\| = \|\mathsf{F}W_I - W_I\langle W_I, \mathsf{F}W_I\rangle\|$ can also be bounded by $\|\mathsf{F}\|$ but such a rough estimation may be penalized severely because it enters the bound quadratically. For example, in the shifted case, $\left\|\mathsf{P}_{\mathcal{W}_I^\perp} \mathsf{F} \mathsf{P}_{\mathcal{W}_I}\right\| = 0$ but $\|\mathsf{F}\| = |\alpha|$.

Theorem 3.50 does not reveal which Ritz vectors should be used for deflation but can help to assess a possible choice of vectors in combination with a priori bounds for the used Krylov subspace method. If theorem 3.50 can be applied for a choice of Ritz vectors, then the resulting inclusion intervals for eigenvalues of the deflated operator can be used with the κ-bound for the CG method (see theorem 2.56), the MINRES bound in theorem 2.63 or any other a priori bound that can be evaluated using inclusion intervals for the eigenvalues. Of course, the bounds can only be applied if the inclusion intervals do not contain zero. If t Ritz pairs $R = \{(w_1, \mu_1), \dots, (w_t, \mu_t)\}$ have been computed, then one could simply test all elements of the power set $\mathcal{P}(R)$. Clearly, this becomes unattractive for large numbers of t because $|\mathcal{P}(R)| = 2^t$. What follows is a description of a strategy for selecting a subset of Ritz vectors from an already computed set of Ritz vectors. The strategy incorporates theorem 3.50 and a simple heuristic. Both the CG κ-bound in theorem 2.56 and the MINRES bound in theorem 2.63 only take into account the extremal eigenvalues, where this means

$$\lambda_{\min} = \min_{\lambda\in\Lambda} \lambda, \quad \lambda^-_{\max} = \max_{\substack{\lambda\in\Lambda\\ \lambda<0}} \lambda, \quad \lambda^+_{\min} = \min_{\substack{\lambda\in\Lambda\\ \lambda<0}} \lambda \quad \text{and} \quad \lambda_{\max} = \max_{\lambda\in\Lambda} \lambda$$

in the indefinite case and $\lambda_{\min}$ and $\lambda_{\max}$ in the positive-definite case. Algorithm 3.2 is an $O(t)$-algorithm which successively considers Ritz vectors for deflation that correspond to the extremal negative and extremal positive Ritz values. The selections are then evaluated with a cost function $\omega : \mathcal{P}(\{1,\ldots,t\}) \longrightarrow \mathbb{R}_{\geq} \cup \{\infty\}$. The cost function is allowed to return ∞ if no estimation is possible, e.g., if assumptions are not satisfied. The index set with the smallest cost is returned if at least one index set could be evaluated. Otherwise the empty set is returned.

One possibility to construct a cost functional ω with the insight of theorem 3.50 is to first compute estimated inclusion intervals for the eigenvalues of the deflated operator $\widehat{\mathsf{B}} = \mathsf{P}_{\mathcal{U}^\perp,\mathsf{B}\mathcal{U}}\mathsf{B}$ by simply assuming that the requirement (3.70) on the eigenvalues of A is satisfied. If also the other assumptions of theorem 3.50 are satisfied, then the right hand side of (3.71) defines a set $\widehat{\Lambda}$ that contains all nonzero eigenvalues of $\widehat{\mathsf{B}}$. If $0 \notin \widehat{\Lambda}$, then asymptotic convergence bounds can be used to determine the first iteration where the bound falls below a prescribed tolerance ε for the relative B-norm of the error (CG) or the relative residual norm (MINRES). For example, if A and B are positive definite, then the κ-bound for the CG method leads to

$$\frac{\|y - y_n\|_{\widehat{\mathsf{B}}}}{\|y - y_0\|_{\widehat{\mathsf{B}}}} \leq 2\left(\frac{\sqrt{\kappa_{\text{eff}}(\widehat{\mathsf{B}})} - 1}{\sqrt{\kappa_{\text{eff}}(\widehat{\mathsf{B}})} + 1}\right)^n \leq 2\left(\frac{\sqrt{\widehat{\kappa}} - 1}{\sqrt{\widehat{\kappa}} + 1}\right)^n = 2\rho_{\text{CG}}^n,$$

where

$$\widehat{\kappa} = \frac{\max\widehat{\Lambda}}{\min\widehat{\Lambda}} \geq \kappa(\widehat{\mathsf{B}}) \quad \text{and} \quad \rho_{\text{CG}} = \frac{\sqrt{\widehat{\kappa}} - 1}{\sqrt{\widehat{\kappa}} + 1}.$$

Thus

$$\frac{\|y - y_n\|_{\widehat{\mathsf{B}}}}{\|y - y_0\|_{\widehat{\mathsf{B}}}} \leq \varepsilon$$

is satisfied for $n = \lceil \log_{\rho_{\text{CG}}} \frac{\varepsilon}{2} \rceil$. Analogously, for the indefinite case and the MINRES method, the inequality

$$\frac{\|c - \mathsf{B}y_n\|}{\|c - \mathsf{B}y_0\|} \leq \varepsilon$$

is satisfied for $n = 2\lceil \log_{\rho_{\text{MR}}} \frac{\varepsilon}{2} \rceil$, where

$$\rho_{\text{MR}} = \frac{\alpha - \beta}{\alpha + \beta} \quad \text{with} \quad \alpha = \sqrt{|\min\widehat{\Lambda} \cdot \max\widehat{\Lambda}|} \quad \text{and} \quad \alpha = \sqrt{|\max_{\lambda\in\widehat{\Lambda},\lambda<0} \lambda \cdot \min_{\lambda\in\widehat{\Lambda},\lambda>0} \lambda|}.$$

Note that the κ-bound can be used if $\widehat{\Lambda}$ only contains positive (or only negative) values.

Algorithm 3.2 Basic selection strategy of Ritz vectors for self-adjoint operators. Implemented as `krypy.recycling.generators.RitzExtremal` in [60].

Input: Assume that the following is given:

- $\mathsf{A}, \mathsf{B} \in \mathcal{L}(\mathcal{H})$ self-adjoint and $c \in \mathcal{L}(\mathcal{H})$.
- Ritz pairs of $\mathsf{A} \in \mathcal{L}(\mathcal{H})$: $(w_1, \mu_1), \ldots, (w_t, \mu_t) \in \mathcal{H} \times \mathbb{R}$ with $\mu_i \neq 0$ and Ritz residual norm $\|r_i\| = \|\mathsf{A}w_i - \mu_i w_i\|$ for $i \in M := \{1, \ldots, t\}$.
- A cost function $\omega : \mathcal{P}(M) \longrightarrow \mathbb{R}_{\geq} \cup \{\infty\}$. For $I = \{i_1, \ldots, i_k\} \subseteq M$, $\omega(I)$ is an estimation of the time that is required to solve the deflated linear system $\mathsf{P}_{\mathcal{U}_I, \mathsf{B}\mathcal{U}_I} \mathsf{B}\tilde{y} = \mathsf{P}_{\mathcal{U}_I, \mathsf{B}\mathcal{U}_I} c$ up to a prescribed tolerance ε with the deflation space $\mathcal{U}_I = [\![w_{i_1}, \ldots, w_{i_k}]\!]$. The value $\omega(I) = \infty$ is allowed to indicate that no estimation is possible for the given set I.

```
$C = \varnothing$.          ▷ gathers subsets of $\{1, \ldots, t\}$ with finite time estimations via $\omega$
$I = \varnothing$, $J = \{1, \ldots, t\}$.
while $J \neq \varnothing$ do
    $\Lambda_J \leftarrow \{\mu_j \mid j \in J\}$.
    $\Lambda_- \leftarrow \Lambda_J \cap ]-\infty, 0[$.          ▷ negative Ritz values in $\Lambda_J$
    $\Lambda_+ \leftarrow \Lambda_J \cap ]0, \infty[$.          ▷ positive Ritz values in $\Lambda_J$
    $\Lambda_S \leftarrow \{\min \Lambda_J, \max \Lambda_J\}$
    $\Lambda_S \leftarrow \Lambda_S \cup \{\max \Lambda_-\}$   if   $\Lambda_- \neq \varnothing$.
    $\Lambda_S \leftarrow \Lambda_S \cup \{\min \Lambda_+\}$   if   $\Lambda_+ \neq \varnothing$.
    $S \leftarrow \{j \in J \mid \mu_j \in \Lambda_S\}$.          ▷ indices of extremal negative and positive Ritz values
    if $\min_{s \in S} \omega(I \cup \{s\}) < \infty$ then
        $s_{\text{sel}} \leftarrow \operatorname{argmin}_{s \in S} \omega(I \cup \{s\})$.
        $C \leftarrow C \cup \{I \cup \{s_{\text{sel}}\}\}$.
    else
        $s_{\text{sel}} \leftarrow \operatorname{argmin}_{s \in S} \|r_s\|$.          ▷ fallback: pick Ritz vector with smallest Ritz residual
    end if
    $I \leftarrow I \cup \{s_{\text{sel}}\}$.
    $J \leftarrow J \setminus \{s_{\text{sel}}\}$.
end while
if $C \neq \varnothing$ then
    return $\operatorname{argmin}_{K \in C} \omega(K)$.          ▷ return index set with smallest time estimation
else
    return $\varnothing$.
end if
```

A cost function ω can now be constructed by using the iteration count n and the number of deflation vectors. As mentioned before, also timings for the most expensive parts of the iteration can be included in order to balance the number of deflation vectors with the number of iterations n.

Example 3.52. Consider again the situation of example 3.23. After 27 iterations, the original linear system $\mathsf{A}x = b$ has been solved with a relative residual norm below 10^{-6}. Now assume – just out of curiosity – that the same linear system has to be solved again and the goal is to find an optimal deflation subspace from the Krylov subspace that was built in the first solution process. If the Ritz pairs of A with respect to this Krylov subspace are computed and algorithm 3.2 is employed with the cost function ω described above, then actually the three Ritz vectors corresponding to the Ritz values of smallest magnitude are chosen automatically. Also, all assumptions of theorem 3.50 are satisfied such that the prediction is not just a heuristic but an actual bound in this case. The bound is plotted in figure 3.10 along with the actual convergence of MINRES if the deflation subspace $\mathcal{W}$ is chosen as the span of the suggested Ritz vectors.

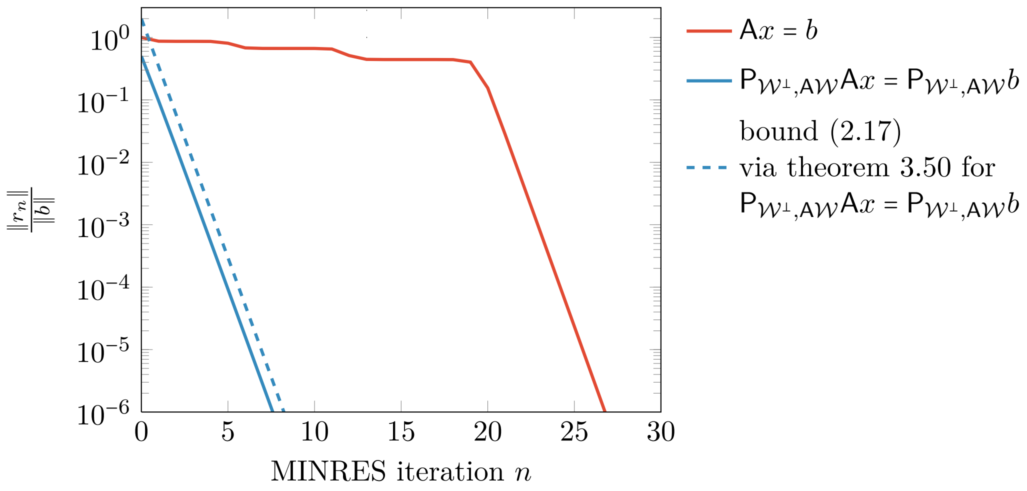


Figure 3.10.: Convergence bound for deflated MINRES via the a priori bound (2.17) based on the eigenvalue inclusion intervals (3.71). The setup is described in example 3.52.

Instead of the proposed strategy, a zoo of more sophisticated approaches can be used if more is known about the spectrum of A. Also, the Ritz values can be preprocessed by a clustering algorithm such that the Ritz pairs are not treated individually in algorithm 3.2 but Ritz values in a small neighborhood are treated as one set whose Ritz vectors are included for deflation or are left out. However, clustering algorithms need to be told how many clusters should be created or at which distance a value is included in a cluster.

It should be noted that linear asymptotic bounds like the CG or MINRES bounds that were used here discard a lot of information about the problem. First, only the extremal eigenvalues are considered and the actual distribution is not taken into account. Second, the bounds eliminate the effects of the right hand side and the initial guess. The often observed nonlinear behavior of Krylov subspace methods such as the transition from a phase of slow convergence to a "superlinear" phase (e.g., in figure 3.2) cannot be characterized with purely linear asymptotic bounds since they are based on the minimal reduction of the A-norm of the error or the residual norm from one iteration to the next. A discussion worth reading about the odd habit of interpreting the inherently nonlinear Krylov subspace methods by means of linear methods can be found in chapter 5.5.4 in the book of Liesen and Strakoš [105].

In principle, the a priori approach of this subsection could be extended to the GMRES method and non-normal operators. For example, the spectral GMRES bound in theorem 2.58 can be considered for this purpose. However, two obstacles render this approach useless in many cases. The first problem is that the spectral bound in equation (2.24) is even less descriptive than the a priori bounds for CG or MINRES due to the presence of the eigenvector basis condition number $\kappa(\mathsf{S})$. The second obstacle is that the perturbation theory for non-normal operators becomes much more intricate and general bounds lead to rough eigenvalue regions that are likely to include the origin and thus render the spectral bound irrelevant.

The difficulties with a priori bounds are avoided in the next subsection by using a very different approach that includes prior knowledge which is available from already computed quantities or can be approximated.

3.4.3. Estimation with approximate Krylov subspaces

A major drawback of the strategy in section 3.4.2 is that the employed a priori bounds for the CG and MINRES methods only take into account the extremal values of the eigenvalue inclusion intervals. The actual distribution of the eigenvalues as well as the right hand side are not considered and – as mentioned in the previous subsection – the linear asymptotic bounds cannot reproduce nonlinear convergence histories. Furthermore, the a priori bound approach in section 3.4.2 is unlikely to yield usable results for non-normal operators and the GMRES method.

In this subsection, a new strategy is proposed that is able to exploit more information from the already completed solution process for the previous linear system. In order to maintain readability, it is assumed in this subsection that all initial guesses are chosen as the zero vector. The general case is a straightforward extension and only adds notational complexity without leading to valuable insight. Apart from the initial guess, the setting in this subsection is identical to the one described in the introduction of section 3.4: the first of the two linear systems in (3.62) has been solved approximately via a Krylov subspace method applied to the deflated linear system (3.63) with the m-dimensional deflation space $\mathcal{U}$. For the first linear system, an Arnoldi relation (3.64) has been constructed for the Krylov subspace $\mathcal{K}_n = \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{b})$. The goal is again to assess a given k-dimensional subspace $\mathcal{W} \subseteq \mathcal{K}_n + \mathcal{U}$ that is considered as a deflation subspace for the next linear system, i.e.,

$$\widehat{\mathsf{B}}\widehat{y} = \widehat{c} \tag{3.74}$$

with $\widehat{\mathsf{B}} = \mathsf{P}_\mathsf{B}\mathsf{B}$ and $\widehat{c} = \mathsf{P}_\mathsf{B}c$ for $\mathsf{P}_\mathsf{B} \in \{\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}, \mathsf{P}_{(\mathsf{B}\mathcal{W})^\perp}\}$.

The idea here is to first figure out how the considered subspace $\mathcal{W}$ performs for the first linear system by constructing an approximate Krylov subspace for the deflated operator $\mathsf{P}_\mathsf{A}\mathsf{A}$ with initial vector $\mathsf{P}_\mathsf{A}b$, where $\mathsf{P}_\mathsf{A} \in \{\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}, \mathsf{P}_{(\mathsf{A}\mathcal{W})^\perp}\}$. Note that P_A is the projection that is built with the new subspace $\mathcal{W}$ and not the projection P that was used to construct the Krylov subspace $\mathcal{K}_n$. For each $i \leq n + m - k$, a self-adjoint perturbation $\mathsf{F}_i \in \mathcal{L}(\mathcal{H})$ is determined along with a perturbation $f \in \mathcal{H}$ of the right hand side such that an Arnoldi relation for $\mathcal{K}_i(\mathsf{P}_\mathsf{A}\mathsf{A} + \mathsf{F}_i, \mathsf{P}_\mathsf{A}b + f)$ is computable without any additional applications of the operator A or the projection P_A. The norm of the perturbations F_i and f are also available and are shown to be optimal for each step i. If A is self-adjoint, then the self-adjoint perturbation F_i shows that the Arnoldi relation in fact is a Lanczos relation. Once the Arnoldi relation for the approximate Krylov subspace is computed, the residual norms of the CG, MINRES or GMRES methods in the i-th iteration for the linear system

$$(\mathsf{P}_\mathsf{A}\mathsf{A} + \mathsf{F}_i)\widehat{x} = \mathsf{P}_\mathsf{A}b + f$$

can be extracted from the Arnoldi relation. In a last step, theorem 3.41 can be employed to bound the i-th residual norm of the used Krylov subspace method when it is applied to the second linear system (3.74) with the considered deflation subspace $\mathcal{W}$. The last step involves knowledge about the (pseudo-)spectrum of the operator A. Similar to the previous subsection, it is reasonable to draw on approximations of the (pseudo-)spectrum due to a lack of exact spectral information in practice.

The basic rationale behind the idea of approximate Krylov subspaces is the following: if it is known how the operator A acts on the subspace $\mathcal{K}_n + \mathcal{U}$, i.e., if $\mathsf{A}[V_n, U]$ is known for bases V_n and U of $\mathcal{K}_n$ and $\mathcal{U}$, then $\mathsf{P}_\mathsf{A}\mathsf{A}[V_n, U]$ can also be computed without additional applications of A. Thus, for any vector $v \in \mathcal{K}_n + \mathcal{U}$, the vector $\mathsf{P}_\mathsf{A}\mathsf{A}v$ can be computed but another application of $\mathsf{P}_\mathsf{A}\mathsf{A}$, i.e., $(\mathsf{P}_\mathsf{A}\mathsf{A})^2 v$ is usually not computable without further applications of A because the vector $\mathsf{P}_\mathsf{A}\mathsf{A}v$ will in general not lie in the subspace $\mathcal{K}_n + \mathcal{U}$ where the behavior of A is known. However, the closest vector in $\mathcal{K}_n + \mathcal{U}$ to $\mathsf{P}_\mathsf{A}\mathsf{A}v$ can be computed as $\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\mathsf{P}_\mathsf{A}\mathsf{A}v$ with the orthogonal projection $\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}$. The error that is made by projecting onto $\mathcal{K}_n + \mathcal{U}$ after the application of $\mathsf{P}_\mathsf{A}\mathsf{A}$ in the i-th iteration can be interpreted as a perturbation F_i of $\mathsf{P}_\mathsf{A}\mathsf{A}$.

The next theorem constitutes the starting point of the further analysis and states a basic observation about the construction of approximate Krylov subspaces. To the knowledge of the author, the result is new.

Theorem 3.53 (Backward error for approximate Krylov subspaces)**.** *Let $V \in \mathcal{H}^n$ and $W \in \mathcal{H}^m$ such that $\langle [V,W],[V,W]\rangle = \mathbf{I}_{n+m}$ and assume that*

$$\mathsf{A}V = V\mathbf{G} + W\mathbf{R} \tag{3.75}$$

for $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $\mathbf{G} \in \mathbb{C}^{n,n}$ and $\mathbf{R} \in \mathbb{C}^{m,n}$. Furthermore, let $\mathbf{G} = \mathbf{T}\mathbf{H}_n\mathbf{T}^\mathsf{H}$ with an upper Hessenberg matrix $\mathbf{H}_n \in \mathbb{C}^{n,n}$ and a unitary matrix $\mathbf{T} \in \mathbb{C}^{n,n}$.

Then for $i \in \{1,\ldots,n-1\}$

$$(\mathsf{A}+\mathsf{F}_i)\widehat{V}_i = \widehat{V}_{i+1}\underline{\mathbf{H}}_i \qquad \text{and} \qquad (\mathsf{A}+\mathsf{F}_n)\widehat{V}_n = \widehat{V}_n\mathbf{H}_n,$$

where $\widehat{V}_n = V\mathbf{T}$, $\widehat{V}_i = \widehat{V}_n\begin{bmatrix}\mathbf{I}_i\\0\end{bmatrix}$, $\underline{\mathbf{H}}_i = \begin{bmatrix}\mathbf{I}_{i+1} & 0\end{bmatrix}\mathbf{H}_n\begin{bmatrix}\mathbf{I}_i\\0\end{bmatrix}$ and

$$\mathsf{F}_i = \mathsf{F}_i^\star = -W\widehat{\mathbf{R}}_i\widehat{V}_i^\star - \widehat{V}_i\widehat{\mathbf{R}}_i^\mathsf{H}W^\star$$

with $\widehat{\mathbf{R}}_n = \mathbf{R}\mathbf{T}$ and $\widehat{\mathbf{R}}_i = \widehat{\mathbf{R}}\begin{bmatrix}\mathbf{I}_i\\0\end{bmatrix}$. Furthermore, $\|\mathsf{F}_i\| = \|\widehat{\mathbf{R}}_i\|_2$ and $\|\mathsf{F}_1\| \le \cdots \le \|\mathsf{F}_n\|$.

Proof. With $\langle W,\widehat{V}_n\rangle = 0$ and $\langle \widehat{V}_n,\widehat{V}_n\rangle = \mathbf{I}_n$, equation (3.75) yields

$$(\mathsf{A}+\mathsf{F}_n)\widehat{V}_n = \widehat{V}_n\mathbf{H}_n.$$

A multiplication with $\begin{bmatrix}\mathbf{I}_i\\0\end{bmatrix} \in \mathbb{R}^{n,i}$ from the right results in

$$(\mathsf{A}+\mathsf{F}_i)\widehat{V}_i = \widehat{V}_i\underline{\mathbf{H}}_i \quad \text{for } i \in \{1,\ldots,n-1\}.$$

The norm of the perturbation satisfies for $i \in \{1,\ldots,n\}$

$$\|\mathsf{F}_i\|^2 = \left\|[\widehat{V}_i,W]\begin{bmatrix}0 & \widehat{\mathbf{R}}_i^\mathsf{H}\\ \widehat{\mathbf{R}}_i & 0\end{bmatrix}[\widehat{V}_i,W]^\star\right\|^2 = \left\|\begin{bmatrix}0 & \widehat{\mathbf{R}}_i^\mathsf{H}\\ \widehat{\mathbf{R}}_i & 0\end{bmatrix}\right\|_2^2 = \|\widehat{\mathbf{R}}_i\|_2^2.$$

The norm inequality follows from the fact that

$$\|\mathsf{F}_i\| = \|\widehat{\mathbf{R}}_i\|_2 = \left\|\widehat{\mathbf{R}}_{i+1}\begin{bmatrix}\mathbf{I}_i\\0\end{bmatrix}\right\|_2 \le \|\widehat{\mathbf{R}}_{i+1}\|_2 = \|\mathsf{F}_{i+1}\| \quad \text{for } i \in \{1,\ldots,n-1\}.$$

□

Remark 3.54. If the matrix $\mathbf{T}$ in theorem 3.53 satisfies $\mathbf{T}e_1 = e_1$ then $\widehat{V}_ne_1 = Ve_1$ holds trivially. A Hessenberg decomposition $\mathbf{G} = \mathbf{T}\mathbf{H}_n\mathbf{T}^\mathsf{H}$ that satisfies $\mathbf{T}e_1 = e_1$ can be constructed with the Householder reduction to Hessenberg form, cf. algorithm 7.4.2 in the book of Golub and Van Loan [71].

The following corollary gives insight into the behavior of the constructed perturbation F_i in theorem 3.53.

Corollary 3.55. *Let the assumptions of theorem 3.53 hold with* $\mathbf{T}e_1 = e_1$ *and let* $\mathbf{H}_n = [h_{ij}]_{i,j\in\{1,\ldots,n\}}$ *and* $v = Ve_1$*. Furthermore, assume that* $j \in \{1,\ldots,n\}$ *is such that* $h_{i+1,i} \neq 0$ *for all* $i \in \{1,\ldots,j-1\}$*.*

Then

$$[\![\widehat{V}_j]\!] = \mathcal{K}_j(\mathsf{A}+\mathsf{F}_j, v) = \mathcal{K}_j(\mathsf{A}+\mathsf{F}_n, v) = \mathcal{K}_j(\mathsf{P}_{[\![V]\!]}\mathsf{A}, v). \tag{3.76}$$

If there exists a $j \in \{1,\ldots,n-1\}$ *with* $h_{j+1,j} = 0$*, then* $\mathcal{K}_j(\mathsf{A}+\mathsf{F}_j, v)$ *is invariant under* $\mathsf{A}+\mathsf{F}_j$*. In any case,* $\mathcal{K}_n(\mathsf{A}+\mathsf{F}_n, v)$ *is invariant under* $\mathsf{A}+\mathsf{F}_n$*.*

Proof. Almost all statements follow immediately from theorem 3.53 with the definitions in section 2.7 and lemma 2.47. The last equality in (3.76) follows from the assumption (3.75) by noticing that

$$\mathsf{P}_{[\![V]\!]}\mathsf{A}\widehat{V}_n = \mathsf{P}_{[\![V]\!]}\widehat{V}_n\mathbf{H}_n + \mathsf{P}_{[\![V]\!]}W\widehat{\mathbf{R}} = \widehat{V}_n\mathbf{H}_n.$$

□

From corollary 3.55, it can be seen that the addition of the perturbation F_j is equivalent to the application of the orthogonal projection $\mathsf{P}_{[\![V]\!]}$ after each application of the operator A. Thus, the perturbation can be seen as optimal in the sense that it yields the smallest perturbation in each step by projecting orthogonally on the subspace $[\![V]\!]$ where it is known how the operator A acts and discarding the part where nothing is known about A.

As mentioned in section 3.3.3, Stewart considered a similar problem in [163]. There, it was shown how to obtain a perturbation F of minimal norm such that a given subspace $\mathcal{U}$ is a Krylov subspace of the operator $\mathsf{A}+\mathsf{F}$. Here, the situation is that an orthonormal basis V is given and the task is to find a perturbation F such that $[\![V]\!]$ is a Krylov subspace of the operator $\mathsf{A}+\mathsf{F}$ with the initial guess Ve_1. The crucial point is that the perturbation is not fixed but is tailored to each iteration and that the perturbation's norm is allowed to grow. In fact, unless $[\![V]\!]$ is close to an A-invariant subspace, the perturbation's norm $\|\mathsf{F}_i\|$ can be expected to be very large for i close to n because the perturbation is forcing the Krylov subspace $\mathcal{K}_n(\mathsf{A}+\mathsf{F}_n, v)$ to be invariant. However, note that the Krylov subspace $\mathcal{K}_i(\mathsf{A}+\mathsf{F}_i, v)$ may already be invariant for $i < n$. The structure of the "mutable" perturbation in theorem 3.53 whose norm is allowed to grow resembles the situation of inexact Krylov subspace methods in the work of Simoncini and Szyld [157]. Their analysis already starts out with a perturbed Arnoldi relation and focuses on how to loosen the restrictions on the perturbation's norm. Here, only the abstract assumption (3.75) is made and the task is to construct an Arnoldi relation with the given data and initial vector Ve_1.

Remark 3.56. In some cases, an initial vector $v \neq 0$ is given that is an element of the given subspace $\mathcal{V} := [\![V]\!]$ where the action of the operator A is known but is not contained in the span of Ve_1 or is not even contained in the given subspace $\mathcal{V}$ at all. If $\langle V, V\rangle = \mathbf{I}_n$ and $q = \langle V, v\rangle \neq 0$, then a Householder transformation matrix

$\mathbf{Q}_q \in \mathbb{C}^{n,n}$ can be constructed such that $\mathbf{Q}_q q = \alpha \left\| q \right\|_2 e_1$ with $|\alpha| = 1$. Because $\mathbf{Q}_q$ is involutory, i.e., $\mathbf{Q}_q^2 = \mathbf{I}_n$, the following holds for $\tilde{V} := V\mathbf{Q}_q$:

$$\tilde{V}e_1 = V\mathbf{Q}_q e_1 = \overline{\alpha}\frac{V\langle V, v\rangle}{\left\| q \right\|_2} = \overline{\alpha}\frac{\mathsf{P}_{\mathcal{V}}v}{\left\| \mathsf{P}_{\mathcal{V}}v \right\|}.$$

Thus, $\tilde{V}e_1$ is a normalized version of $\mathsf{P}_{\mathcal{V}}v$, the vector in $\mathcal{V}$ with minimal distance to the initial vector v of interest. If a relation of the form (3.75) is given, the procedure in theorem 3.53 with $\tilde{V}$, $\tilde{\mathbf{G}} = \mathbf{Q}_q\mathbf{G}\mathbf{Q}_q$, $\tilde{\mathbf{R}} = \mathbf{R}\mathbf{Q}_q$ instead of V, $\mathbf{G}$ and $\mathbf{R}$ can thus be understood as the generation of an optimal approximate Krylov subspace for the initial vector v with respect to the available data. The optimality is meant with regard to each iteration as explained in the discussion following corollary 3.55.

In the next theorem, the following situation is analyzed: an Arnoldi relation for the Krylov subspace $\mathcal{K}_n = \mathcal{K}_n(\widehat{\mathsf{A}}, \widehat{v})$ with $\widehat{\mathsf{A}} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}$ and $\widehat{v} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}v$ for a given $v \in \mathcal{H}$ is known and the task is to construct an Arnoldi relation for an approximate Krylov subspace with the operator $\mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}}\mathsf{A}$ and the initial vector $\mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}}v$, where $\mathcal{W} \subseteq \mathcal{K}_n + \mathcal{U}$. The orthogonal projections $\mathsf{P}_{(\mathsf{A}\mathcal{U})^\perp}$ and $\mathsf{P}_{(\mathsf{A}\mathcal{W})^\perp}$ are not considered in the following theorem. The reason for this decision lies in the inapplicability of certain perturbation techniques which becomes apparent in the course of this section, cf. remark 3.66. The theorem is preceded by an assumption and a definition with quantities that are reused later. The definitions may seem technical at first sight but the computation of the defined quantities can be implemented efficiently in a few lines of code (implemented as `krypy.deflation.Arnoldifyer` in [60]).

Assumption 3.57. Let $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $v \in \mathcal{H}$, $U \in \mathcal{H}^m$ with $\langle U, U\rangle = \mathbf{I}_m$ such that $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ for $\mathcal{U} = [\![U]\!]$ and let $V_{n+1} = [V_n, v_{n+1}] \in \mathcal{H}^{n+1}$ and $\underline{\mathbf{H}}_n$ define an Arnoldi relation for $\mathcal{K}_n = \mathcal{K}_n(\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}\mathsf{A}, \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}v)$ with $V_{n+1}e_1 = \frac{\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}v}{\left\|\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}v\right\|}$. Furthermore, let $\tilde{\mathbf{W}} \in \mathbb{C}^{n+m,k}$ such that $\theta_{\max}(\mathcal{W}, \mathsf{A}\mathcal{W}) < \frac{\pi}{2}$ for $\mathcal{W} = [\![W]\!]$, where $W = [V_n, U]\tilde{\mathbf{W}}$.

Definition 3.58. Let assumption 3.57 hold and let $\tilde{\mathbf{W}}_\perp \in \mathbb{C}^{n+m,n+m-k}$ be such that

$$\tilde{\mathbf{W}}_\perp^{\mathsf{H}} \begin{bmatrix} \tilde{\mathbf{W}}_\perp \\ \tilde{\mathbf{W}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{n+m-k} \\ 0_{k,n+m-k} \end{bmatrix}.$$

With $\mathbf{B}$, $\underline{\mathbf{B}}$, $\mathbf{C}$, $\mathbf{E}$, $\mathbf{L}$ and $\underline{\mathbf{I}}_n$ as in lemmas 3.43–3.45, the following quantities are defined:

$$\mathbf{J} := \begin{bmatrix} \underline{\mathbf{I}}_n^{\mathsf{T}} & \mathbf{B} \\ 0_{m,n+1} & \mathbf{E} \end{bmatrix}, \qquad \tilde{\mathbf{P}} := \mathbf{I}_{n+m+1} - \mathbf{L}\tilde{\mathbf{W}}(\tilde{\mathbf{W}}^{\mathsf{H}}\mathbf{J}\mathbf{L}\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^{\mathsf{H}}\mathbf{J},$$

$$\tilde{q} := \tilde{\mathbf{P}} \begin{bmatrix} \left\|\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}}v\right\| \\ 0_{n,1} \\ \mathbf{E}^{-1}\langle U, v\rangle \end{bmatrix} \qquad \text{and} \qquad q := \tilde{\mathbf{W}}_\perp^{\mathsf{H}}\mathbf{J}\tilde{q}.$$

Furthermore, let the following quantities be given:

- A Householder transformation $\mathbf{Q}_q \in \mathbb{C}^{n+m-k,n+m-k}$ such that $\mathbf{Q}_q q = \alpha \left\| q \right\|_2 e_1$ with $|\alpha| = 1$.
- A Hessenberg decomposition of $\mathbf{Q}_q \tilde{\mathbf{W}}_\perp^\mathsf{H} \mathbf{J} \tilde{\mathbf{P}} \mathbf{L} \tilde{\mathbf{W}}_\perp \mathbf{Q}_q = \mathbf{T} \widehat{\mathbf{H}}_{n+m-k} \mathbf{T}^\mathsf{H}$ with an upper Hessenberg matrix $\widehat{\mathbf{H}}_{n+m-k} \in \mathbb{C}^{n+m-k,n+m-k}$ and a unitary matrix $\mathbf{T} \in \mathbb{C}^{n+m-k,n+m-k}$ such that $\mathbf{T}e_1 = e_1$.
- A QR decomposition (with column pivoting) of
$$(\mathsf{A}\mathcal{U} - U\mathbf{E} - V_{n+1}\underline{\mathbf{B}})\mathbf{P}_{QR} = [Q_1, Q_2] \begin{bmatrix} \mathbf{R}_1 & \mathbf{R}_2 \\ 0 & 0 \end{bmatrix},$$
where $\mathbf{P}_{QR} \in \mathbb{C}^{m,m}$ is a permutation matrix, $Q_1 \in \mathcal{H}^l$, $Q_2 \in \mathcal{H}^{m-l}$ are such that $\langle [Q_1, Q_2], [Q_1, Q_2] \rangle = \mathbf{I}_m$ and $\mathbf{R}_1 \in \mathbb{C}^{l,l}$ is nonsingular.

Then also define

$$\mathbf{N} := \begin{bmatrix} 1 & e_{n+1}^\mathsf{T} \underline{\mathbf{B}} \\ 0_{l,1} & [\mathbf{R}_1, \mathbf{R}_2]\mathbf{P}_{QR}^\mathsf{T} \end{bmatrix} \begin{bmatrix} 0_{m+1,n} & \mathbf{I}_{m+1} \end{bmatrix}, \qquad Z := [v_{n+1}, Q_1],$$

$$\widehat{V}_{n+m-k} := [V_n, U] \tilde{\mathbf{W}}_\perp \mathbf{Q}_q \mathbf{T}, \qquad \widehat{V}_i := \widehat{V}_{n+m-k} \begin{bmatrix} \mathbf{I}_i \\ 0 \end{bmatrix},$$

$$\widehat{\mathbf{R}}_{n+m-k} := \mathbf{N} \tilde{\mathbf{P}} \mathbf{L} \tilde{\mathbf{W}}_\perp \mathbf{Q}_q \mathbf{T}, \qquad \widehat{\mathbf{R}}_i := \widehat{\mathbf{R}}_{n+m-k} \begin{bmatrix} \mathbf{I}_i \\ 0 \end{bmatrix},$$

$$\mathsf{F}_i := -Z \widehat{R}_i \widehat{V}_i^\star - \widehat{V}_i \widehat{R}_i^\mathsf{H} Z^\star, \qquad \underline{\widehat{\mathbf{H}}}_j = \begin{bmatrix} \mathbf{I}_{j+1} & 0 \end{bmatrix} \widehat{\mathbf{H}}_{n+m-k} \begin{bmatrix} \mathbf{I}_j \\ 0 \end{bmatrix},$$

where $i \in \{1, \ldots, n+m-k\}$ and $j \in \{1, \ldots, n+m-k-1\}$.

Theorem 3.59. *Let assumption 3.57 hold and let the definitions from definition 3.58 be given.*

Then for $i \in \{1, \ldots, n+m-k-1\}$

$$(\tilde{\mathsf{A}} + \mathsf{F}_i)\widehat{V}_i = \widehat{V}_{i+1} \underline{\widehat{\mathbf{H}}}_i \tag{3.77}$$
$$\text{and} \qquad (\tilde{\mathsf{A}} + \mathsf{F}_{n+m-k})\widehat{V}_{n+m-k} = \widehat{V}_{n+m-k} \widehat{\mathbf{H}}_{n+m-k} \tag{3.78}$$

with $\tilde{\mathsf{A}} = \mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}} \mathsf{A}$. Furthermore, $\left\langle \widehat{V}_i, \widehat{V}_i \right\rangle = \mathbf{I}_i$ and $\left\| \mathsf{F}_i \right\| = \left\| \widehat{\mathbf{R}}_i \right\|_2$ hold for $i \in \{1, \ldots, n+m\}$ as well as $\left\| \mathsf{F}_1 \right\| \leq \cdots \leq \left\| \mathsf{F}_{n+m-k} \right\|$.

If there exists a $j \in \{1, \ldots, n+m-k-1\}$ such that $h_{i+1,i} \neq 0$ for all $i \in \{1, \ldots, j\}$, then $\widehat{V}_{j+1}$ and $\underline{\widehat{\mathbf{H}}}_j$ define an Arnoldi relation for

$$\mathcal{K}_j(\tilde{\mathsf{A}} + \mathsf{F}_j, \tilde{v} + f) = \mathcal{K}_j(\tilde{\mathsf{A}} + \mathsf{F}_{n+m-k}, \tilde{v} + f) = \mathcal{K}_j(\mathsf{P}_{\mathcal{K}_n + \mathcal{U}} \tilde{\mathsf{A}}, \mathsf{P}_{\mathcal{K}_n + \mathcal{U}} \tilde{v}), \tag{3.79}$$

where $\tilde{v} := \mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}} v$ and $f := -\mathsf{P}_{(\mathcal{K}_n + \mathcal{U})^\perp} \tilde{v}$ satisfy $\tilde{v} + f \in \mathcal{W}^\perp$ and $\left\| f \right\| = \left\| \mathbf{N} \tilde{q} \right\|_2$.

Proof. It is first shown that

$$\tilde{\mathsf{A}} \widehat{V}_{n+m-k} = \widehat{V}_{n+m-k} \widehat{\mathbf{H}}_{n+m-k} + Z \widehat{\mathbf{R}}_{n+m-k}.$$

Therefore, note the following equations which can be derived with simple calculations and the definitions in definition 3.58:

$$\begin{aligned}
\mathsf{A}[V_n, U] &= [V_{n+1}, \mathsf{A}U]\mathbf{L}, \\
\langle [V_n, U], [V_{n+1}, \mathsf{A}U] \rangle &= \mathbf{J}, \\
\mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}}[V_{n+1}, \mathsf{A}U] &= [V_{n+1}, \mathsf{A}U] - \mathsf{A}W \langle W, \mathsf{A}W \rangle^{-1} \langle W, [V_{n+1}, \mathsf{A}U] \rangle \\
&= [V_{n+1}, \mathsf{A}U](\mathbf{I}_{n+1+m} - \mathbf{L}\tilde{\mathbf{W}}(\tilde{\mathbf{W}}^\mathsf{H}\mathbf{J}\mathbf{L}\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^\mathsf{H}\mathbf{J}) \\
&= [V_{n+1}, \mathsf{A}U]\tilde{\mathbf{P}}, \\
\mathsf{P}_{\mathcal{K}_n + \mathcal{U}}[V_{n+1}, \mathsf{A}U] &= [V_n, U]\mathbf{J}, \\
\mathsf{A}U - V_n\mathbf{B} - U\mathbf{E} &= \mathsf{A}U - V_{n+1}\underline{\mathbf{B}} - U\mathbf{E} + v_{n+1}e_{n+1}^\mathsf{T}\underline{\mathbf{B}} \\
&= \mathsf{P}_{(\mathcal{K}_{n+1} + \mathcal{U})^\perp}\mathsf{A}U + v_{n+1}e_{n+1}^\mathsf{T}\underline{\mathbf{B}}, \\
\mathsf{P}_{(\mathcal{K}_n + \mathcal{U})^\perp}[V_{n+1}, \mathsf{A}U] &= [V_{n+1}, \mathsf{A}U] - [V_n, U]\mathbf{J} \\
&= [V_{n+1}, \mathsf{A}U] - [V_n, U]\begin{bmatrix} \underline{\mathbf{I}}_n^\mathsf{T} & \mathbf{B} \\ 0_{m,n+1} & \mathbf{E} \end{bmatrix} \\
&= [v_{n+1}, \mathsf{A}U - V_n\mathbf{B} - U\mathbf{E}][0_{m+1,n}, \mathbf{I}_{m+1}] \\
&= [v_{n+1}, \mathsf{A}U - V_{n+1}\underline{\mathbf{B}} - U\mathbf{E}]\begin{bmatrix} 1 & e_{n+1}^\mathsf{T}\underline{\mathbf{B}} \\ 0_{m,1} & \mathbf{I}_m \end{bmatrix}[0_{m+1,n}, \mathbf{I}_{m+1}] \\
&= [v_{n+1}, Q_1]\begin{bmatrix} 1 & e_{n+1}^\mathsf{T}\underline{\mathbf{B}} \\ 0_{l,1} & [\mathbf{R}_1, \mathbf{R}_2]\mathbf{P}_{QR}^\mathsf{T} \end{bmatrix}[0_{m+1,n}, \mathbf{I}_{m+1}] = Z\mathbf{N}.
\end{aligned}$$

The above equations yield with $\tilde{\mathbf{W}}_\perp\tilde{\mathbf{W}}_\perp^\mathsf{H}\mathbf{J}\tilde{\mathbf{P}} = \mathbf{J}\tilde{\mathbf{P}}$ the following equation:

$$\begin{aligned}
\tilde{\mathsf{A}}\widehat{V}_{n+m-k} &= \mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}}\mathsf{A}[V_n, U]\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} = \mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}}[V_{n+1}, \mathsf{A}U]\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} \\
&= [V_{n+1}, \mathsf{A}U]\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} \\
&= \mathsf{P}_{\mathcal{K}_n + \mathcal{U}}[V_{n+1}, \mathsf{A}U]\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} + \mathsf{P}_{(\mathcal{K}_n + \mathcal{U})^\perp}[V_{n+1}, \mathsf{A}U]\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} \\
&= [V_n, U]\mathbf{J}\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} + Z\mathbf{N}\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} \\
&= [V_n, U]\tilde{\mathbf{W}}_\perp\tilde{\mathbf{W}}_\perp^\mathsf{H}\mathbf{J}\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} + Z\mathbf{N}\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} \\
&= \widehat{V}_{n+m-k}\mathbf{T}^\mathsf{H}\mathbf{Q}_q\tilde{\mathbf{W}}_\perp^\mathsf{H}\mathbf{J}\tilde{\mathbf{P}}\mathbf{L}\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T} + Z\widehat{\mathbf{R}}_{n+m-k} \\
&= \widehat{V}_{n+m-k}\widehat{\mathbf{H}}_{n+m-k} + Z\widehat{\mathbf{R}}_{n+m-k}.
\end{aligned}$$

Now theorem 3.53 and corollary 3.55 apply and show that equations (3.77)–(3.78) hold and that $\widehat{V}_{j+1}$ and $\underline{\widehat{\mathbf{H}}}_n$ actually define Arnoldi relations hold for the Krylov subspaces in equation (3.79) under the made assumption on j.

It remains to show that $\tilde{v} + f \in \mathcal{W}^\perp$ and that $\|f\| = \|\mathbf{N}\tilde{q}\|$. Therefore, observe that

$$\begin{aligned}
\widehat{V}_{n+m-k}e_1 &= [V_n, U]\tilde{\mathbf{W}}_\perp\mathbf{Q}_q\mathbf{T}e_1 = [V_n, U]\tilde{\mathbf{W}}_\perp\mathbf{Q}_q e_1 = \frac{\overline{\alpha}}{\|q\|_2}[V_n, U]\tilde{\mathbf{W}}_\perp q \\
&= \frac{\overline{\alpha}}{\|q\|_2}[V_n, U]\tilde{\mathbf{W}}_\perp\tilde{\mathbf{W}}_\perp^\mathsf{H}\mathbf{J}\tilde{q} = \frac{\overline{\alpha}}{\|q\|_2}[V_n, U]\mathbf{J}\tilde{q}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\overline{\alpha}}{\|q\|_2}\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}[V_{n+1},\mathsf{A}U]\tilde{q} = \frac{\overline{\alpha}}{\|q\|_2}\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}[V_{n+1},\mathsf{A}U]\tilde{\mathbf{P}}\begin{bmatrix}\left\|\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}v\right\|\\ 0_{n,1}\\ \mathbf{E}^{-1}\langle U,v\rangle\end{bmatrix}\\
&= \frac{\overline{\alpha}}{\|q\|_2}\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}[V_{n+1},\mathsf{A}U]\begin{bmatrix}\left\|\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}v\right\|\\ 0_{n,1}\\ \mathbf{E}^{-1}\langle U,v\rangle\end{bmatrix}\\
&= \frac{\overline{\alpha}}{\|q\|_2}\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\left(v_1\left\|\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}v\right\| + \mathsf{A}U\mathbf{E}^{-1}\langle U,v\rangle\right)\\
&= \frac{\overline{\alpha}}{\|q\|_2}\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}v = \frac{\overline{\alpha}}{\|q\|_2}\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\tilde{v} = \overline{\alpha}\frac{\tilde{v}-\mathsf{P}_{(\mathcal{K}_n+\mathcal{U})^\perp}\tilde{v}}{\|q\|_2} = \overline{\alpha}\frac{\tilde{v}+f}{\|\tilde{v}+f\|}.
\end{aligned}
$$

Also, for any $w\in\mathcal{W}\subseteq\mathcal{K}_n+\mathcal{U}$, the perturbed initial vector $\tilde{v}+f=\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\tilde{v}$ satisfies

$$\langle w,\tilde{v}+f\rangle = \left\langle w,\mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}v\right\rangle = \left\langle \mathsf{P}_{\mathcal{K}_n+\mathcal{U}}w,\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}v\right\rangle = \left\langle w,\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}v\right\rangle = 0$$

and thus is orthogonal to $\mathcal{W}$. The norm equality directly follows by the above equations and the fact that $\langle Z,Z\rangle=\mathbf{I}_{l+1}$:

$$\|f\| = \left\|\mathsf{P}_{(\mathcal{K}_n+\mathcal{U})^\perp}\tilde{v}\right\| = \left\|\mathsf{P}_{(\mathcal{K}_n+\mathcal{U})^\perp}[V_{n+1},\mathsf{A}U]\tilde{q}\right\| = \|Z\mathbf{N}\tilde{q}\| = \|\mathbf{N}\tilde{q}\|_2\,.$$

□

Remark 3.60. Theorem 3.59 also holds if no deflation space was used for the construction of $\mathcal{K}_n$, i.e., $\mathcal{U}=\{0\}$, or if no deflation space is considered for the next linear system, i.e., $\mathcal{W}=\{0\}$. In assumption 3.57 and definition 3.58, the bases for the deflation spaces can simply be chosen as "empty" bases with $m=0$ or $k=0$, i.e., $U\in\mathcal{H}^0$ or $W=[V_n,U]\tilde{\mathbf{W}}\in\mathcal{H}^0$ with $\tilde{\mathbf{W}}\in\mathbb{C}^{n+m,0}$.

It remains to investigate if the proposed strategy can provide significant insight into the behavior of the deflated operator $\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}$. In the following example, the norms of the perturbation F_i are computed for the situation of examples 3.23 and 3.52.

Example 3.61. If MINRES is applied to the linear system $\mathsf{A}x=b$ from example 3.23 with $x_0=0$, the residual norm satisfies $\frac{\|r_n\|}{\|b\|}<10^{-6}$ after $n=27$ steps, see figure 3.2. Let V_{n+1} and $\underline{\mathbf{H}}_n$ define a Lanczos relation for the Krylov subspace $\mathcal{K}_n(\mathsf{A},b)$ that was constructed by MINRES at this step. Now let $\tilde{\mathbf{W}}\in\mathbb{C}^{n,3}$ be such that $W=V_n\tilde{\mathbf{W}}$ contains the 3 Ritz vectors of A with respect to $\mathcal{K}_n(\mathsf{A},b)$ that correspond to the Ritz values of smallest magnitude. With $\mathcal{W}=[\![W]\!]$ and $\mathcal{U}=\{0\}$, theorem 3.59 can be applied to construct a Lanczos relation for the Krylov subspace $\tilde{\mathcal{K}}_i=\mathcal{K}_i(\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}+\mathsf{F}_i,\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b+f)$.

Of interest are the norms of the perturbations of the right hand side $\|f\|$ and of the operator perturbations $\|\mathsf{F}_i\|$ in each step $i\in\{1,\ldots,l\}$, where $l=d(\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}+\mathsf{F}_{n-3},\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b+f)$ is the grade of $\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b+f$ with respect to $\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}+\mathsf{F}_{n-3}$. Figure 3.11 shows the development of the perturbation's norm for the mentioned choice of $\tilde{\mathbf{W}}$.

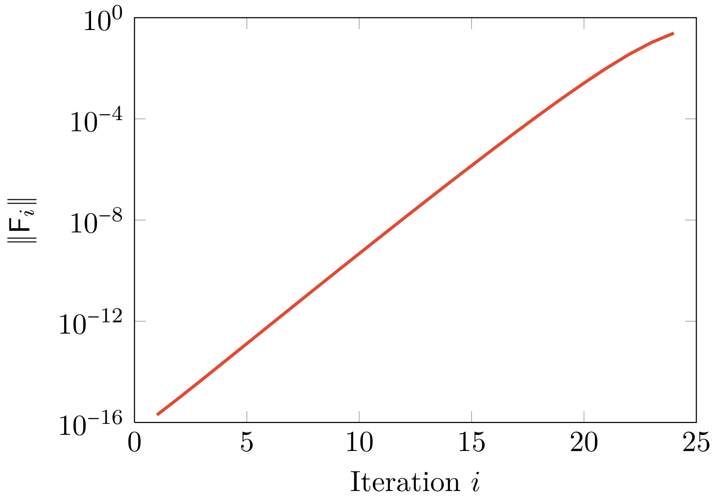


Figure 3.11.: Norm of the perturbation F_i in theorem 3.59 for example 3.61.

Clearly, the norm of the perturbation F_i is small in the first iterations but grows up to ≈ 0.1 at iteration 24 where the Krylov subspace $\tilde{\mathcal{K}}$ with the perturbed operator and initial vector becomes invariant. The perturbation of the right hand side satisfies $\|f\| \approx 1.359 \cdot 10^{-6}$.

After the above treatment of approximate Krylov subspaces, the focus is now on how to exploit the findings for the assessment of deflation vectors for the solution of linear systems. Let $U \in \mathcal{H}^m$ be such that $\langle U, U \rangle = \mathbf{I}_m$ and $\theta_{\max}(\mathcal{U}, \mathsf{A}\mathcal{U}) < \frac{\pi}{2}$ for $\mathcal{U} = [\![U]\!]$. If a Krylov subspace method is applied to the linear system

$$\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}} \mathsf{A} \widehat{x} = \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}} b,$$

then the method explicitly or implicitly computes a basis $V_{n+1} \in \mathcal{H}^{n+1}$ and an extended Hessenberg matrix $\underline{\mathbf{H}}_n$ which define an Arnoldi relation for the Krylov subspace

$$\mathcal{K}_n = \mathcal{K}_n(\mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}} \mathsf{A}, \mathsf{P}_{\mathcal{U}^\perp, \mathsf{A}\mathcal{U}} b).$$

Now assume that $\tilde{\mathbf{W}} \in \mathbb{C}^{n+m,k}$ is given such that $\theta_{\max}(\mathcal{W}, \mathsf{A}\mathcal{W}) < \frac{\pi}{2}$ for $\mathcal{W} = [\![W]\!]$, where $W = [V_n, U]\tilde{\mathbf{W}}$. Then assumption 3.57 is satisfied with $v = b$ and it can easily be verified that all quantities in definition 3.58 are available without any additional applications of the operator A (see also remark 3.44). Theorem 3.59 now provides perturbations F_i for $i \in \{1, \ldots, n+m-k\}$, a basis $\widehat{V}_{n+m-k}$ and a Hessenberg matrix $\widehat{\mathbf{H}}_{n+m-k} = [\widehat{h}_{i,j}]_{i,j \in \{1,\ldots,n+m-k\}}$. If j is the largest index $j \in \{1, \ldots, n+m-k-1\}$ such that $\widehat{h}_{i+1,i} \neq 0$ for all $i \in \{1, \ldots, j\}$, then $\widehat{V}_{i+1}$ and $\underline{\widehat{\mathbf{H}}}_i$ define an Arnoldi relation for the Krylov subspace

$$\tilde{\mathcal{K}}_i = \mathcal{K}_i(\mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}} \mathsf{A} + \mathsf{F}_i, \mathsf{P}_{\mathcal{W}^\perp, \mathsf{A}\mathcal{W}} b + f)$$

for all $i \in \{1,\ldots,j\}$, where the norm of $f = \mathsf{P}_{(\mathcal{K}_n+\mathcal{U})^\perp}\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b$ is available from theorem 3.59 as $\|f\| = \|\mathbf{N}\tilde{q}\|_2$. Now observe that $\tilde{\mathcal{K}}_i$ is the Krylov subspace that is constructed if the Krylov subspace method is applied to the linear system

$$(\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A} + \mathsf{F}_i)\tilde{x} = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b + f. \tag{3.80}$$

Note that quantities like the residual norm in the i-th step of the MINRES and GMRES algorithms for the linear system (3.80) can be directly retrieved from the available Hessenberg matrix $\widehat{\mathbf{H}}_{n+m-k}$ of the Arnoldi relation for $\tilde{\mathcal{K}}_i$ without any additional applications of the operator A, see sections 2.9.1 and 2.9.2.

Together with the perturbation theorem 3.41, the norm of the residual in the i-th step of the Krylov subspace method applied to the unperturbed linear system

$$\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}\overline{x} = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b$$

can be bounded[4]. A crucial observation is that the perturbation F_i is sufficient in order to obtain statements for the i-th step because for early iterations i, the norm $\|\mathsf{F}_i\|$ can be significantly smaller than the norm of F_{n+m-k} which can also be used for the generation of the same Krylov subspace, cf. example 3.61 and figure 3.11.

In fact, one can go further and directly obtain bounds for the Krylov subspace method when it is applied to the linear system that is of main interest in this section:

$$\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B}\widehat{y} = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c. \tag{3.81}$$

With $\mathsf{A}_{\mathsf{F}_i} = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A} + \mathsf{F}_i$ and $b_f = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b + f$, the operator and right hand side in equation (3.81) satisfy

$$\begin{aligned} &\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B} = \mathsf{A}_{\mathsf{F}_i} + \widehat{\mathsf{F}}_i && \text{with } \widehat{\mathsf{F}}_i = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B} - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A} - \mathsf{F}_i \\ \text{and } &\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c = b_f + \widehat{f} && \text{with } \widehat{f} = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b - f. \end{aligned}$$

The next theorem is required in order to bound the terms $\|\widehat{\mathsf{F}}_i\|$ and $\|\widehat{f}\|$ in the above equations. The result can be seen as a complement to theorem 3.17 by providing a computable bound on the norm of the difference $\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}} - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}$ in terms of the operator A and the difference $\mathsf{F} = \mathsf{B} - \mathsf{A}$.

Theorem 3.62. *Let* $\mathsf{A},\mathsf{B} \in \mathcal{L}(\mathcal{H})$ *and* $\mathsf{F} = \mathsf{B} - \mathsf{A}$. *Assume that* $W \in \mathcal{H}^m$ *is given such that* $\langle W, W\rangle = \mathbf{I}_m$ *and*

$$\sigma_{\min}(\langle W, \mathsf{A}W\rangle) > \|\langle W, \mathsf{F}W\rangle\|_2. \tag{3.82}$$

Then $\theta_{\max}(\mathcal{W}, \mathsf{A}\mathcal{W}) < \frac{\pi}{2}$ *and* $\theta_{\max}(\mathcal{W}, \mathsf{B}\mathcal{W}) < \frac{\pi}{2}$ *and the following holds:*

$$\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}} - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\| \le \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\| \frac{\|\mathsf{F}|_{\mathcal{W}}\|}{\sigma_{\min}(\langle W, \mathsf{A}W\rangle) - \|\langle W, \mathsf{F}W\rangle\|_2}.$$

[4] Theorem 3.41 only addresses GMRES (and thus also MINRES). However, an extension to the A-norm of the error in the CG method is straightforward.

Proof. First note that $\theta_{\max}(\mathcal{W},\mathsf{A}\mathcal{W}) < \frac{\pi}{2}$ follows immediately from theorem 2.15 since $\langle W,\mathsf{A}W\rangle$ has to be nonsingular by assumption (3.82). Now observe that the smallest singular value of $\langle W,\mathsf{B}W\rangle$ satisfies

$$\sigma_{\min}(\langle W,\mathsf{B}W\rangle) = \sigma_{\min}(\langle W,\mathsf{A}W\rangle + \langle W,\mathsf{F}W\rangle) \geq \sigma_{\min}(\langle W,\mathsf{A}W\rangle) - \|\langle W,\mathsf{F}W\rangle\|_2 > 0$$

also by assumption (3.82) and thus $\theta_{\max}(\mathcal{W},\mathsf{B}\mathcal{W}) < \frac{\pi}{2}$. Then the projections are well defined and

$$\begin{aligned}
\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}} - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\| = \left\|\mathsf{P}_{\mathsf{B}\mathcal{W},\mathcal{W}^\perp} - \mathsf{P}_{\mathsf{A}\mathcal{W},\mathcal{W}^\perp}\right\| &= \sup_{x\in\mathbb{C}^m,x\neq0} \frac{\left\|\mathsf{P}_{\mathsf{B}\mathcal{W},\mathcal{W}^\perp}Wx - \mathsf{P}_{\mathsf{A}\mathcal{W},\mathcal{W}^\perp}Wx\right\|}{\|x\|_2} \\
&= \sup_{x\in\mathbb{C}^m,x\neq0} \frac{\left\|\mathsf{B}W\langle W,\mathsf{B}W\rangle^{-1}x - \mathsf{A}W\langle W,\mathsf{A}W\rangle^{-1}x\right\|}{\|x\|_2} \\
&= \sup_{y\in\mathbb{C}^m,y\neq0} \frac{\left\|\mathsf{B}Wy - \mathsf{A}W\langle W,\mathsf{A}W\rangle^{-1}\langle W,\mathsf{B}W\rangle y\right\|}{\|\langle W,\mathsf{B}W\rangle y\|_2} \\
&= \sup_{y\in\mathbb{C}^m,y\neq0} \frac{\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{B}Wy\right\|}{\|\langle W,\mathsf{B}W\rangle y\|_2} \leq \sup_{y\in\mathbb{C}^m,y\neq0} \frac{\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{B}Wy\right\|}{\sigma_{\min}(\langle W,\mathsf{B}W\rangle)\,\|y\|_2} \\
&= \frac{\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{B}W\right\|}{\sigma_{\min}(\langle W,\mathsf{B}W\rangle)} \leq \frac{\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{B}W\right\|}{\sigma_{\min}(\langle W,\mathsf{A}W\rangle) - \|\langle W,\mathsf{F}W\rangle\|_2} \\
&= \frac{\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}(\mathsf{A}+\mathsf{F})W\right\|}{\sigma_{\min}(\langle W,\mathsf{A}W\rangle) - \|\langle W,\mathsf{F}W\rangle\|_2} \\
&= \frac{\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{F}W\right\|}{\sigma_{\min}(\langle W,\mathsf{A}W\rangle) - \|\langle W,\mathsf{F}W\rangle\|_2} \\
&\leq \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\| \frac{\|\,\mathsf{F}|_{\mathcal{W}}\|}{\sigma_{\min}(\langle W,\mathsf{A}W\rangle) - \|\langle W,\mathsf{F}W\rangle\|_2}.
\end{aligned}$$

□

The following corollary is a straightforward application of the above theorem.

Corollary 3.63. *Let the assumptions of theorem 3.62 hold and let*

$$\eta := \frac{\|\,\mathsf{F}|_{\mathcal{W}}\|}{\sigma_{\min}(\langle W,\mathsf{A}W\rangle) - \|\langle W,\mathsf{F}W\rangle\|_2}.$$

Then

$$\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B} - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}\right\| \leq \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\| \left(\eta(\|\mathsf{A}\| + \|\mathsf{F}\|) + \|\mathsf{F}\|\right).$$

If $b,f\in\mathcal{H}$ and $c := b+f$, then analogously

$$\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b\right\| \leq \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\| \left(\eta(\|b\| + \|f\|) + \|f\|\right).$$

Proof. Only the first inequality is shown because the proof of the second inequality is completely analogous.

$$\begin{aligned}\left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B}-\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}\right\| &= \left\|(\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}-\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}})\mathsf{B}+\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}(\mathsf{B}-\mathsf{A})\right\| \\ &\leq \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}-\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\,\|\mathsf{A}+\mathsf{F}\| + \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\,\|\mathsf{F}\| \\ &\leq \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\left(\eta(\|\mathsf{A}\|+\|\mathsf{F}\|)+\|\mathsf{F}\|\right).\end{aligned}$$

□

The following theorem finally wraps up the above results and shows how the convergence of a Krylov subspace method for the linear system

$$\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B}\widehat{y}=\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c$$

can be characterized for a given deflation subspace $\mathcal{W}\subseteq\mathcal{K}_n+\mathcal{U}$ after the Krylov subspace $\mathcal{K}_n$ has been generated by a Krylov subspace method applied to the linear system

$$\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}\mathsf{A}\widehat{x}=\mathsf{P}_{\mathcal{U}^\perp,\mathsf{A}\mathcal{U}}b.$$

To the knowledge of the author, no similar statements have been shown in the literature.

Theorem 3.64. *Assume that* $\mathsf{A}x=b$ *and* $\mathsf{B}y=c$ *are linear systems with nonsingular operators* $\mathsf{A},\mathsf{B}\in\mathcal{L}(\mathbf{H})$*, right hand sides* $b,c\in\mathcal{H}$ *and assume that assumption 3.57 is fulfilled with* $v=b$*. Furthermore, let* $\widehat{\mathbf{H}}=\widehat{\mathbf{H}}_{n+m-k}$*,* F_i*,* $\widehat{\mathbf{R}}_i$ *for* $i\in\{1,\ldots,n+m-k\}$*,* $\mathbf{N}$*,* q *and* $\tilde{q}$ *be as in theorem 3.59 for* $i\in\{1,\ldots,n+m-k\}$*. Let* $l=d(\widehat{\mathbf{H}},e_1)$ *be the grade of* $\widehat{\mathbf{H}}$ *with respect to* e_1 *and let for* $i\in\{1,\ldots,l\}$ *the* i*-th residual of GMRES applied to the linear system*

$$\widehat{\mathbf{H}}z=\|q\|_2\,e_1 \tag{3.83}$$

be denoted by $\tilde{r}_i=p_i(\widehat{\mathbf{H}})\,\|q\|_2\,e_1$*, where* $p_i\in\mathbb{P}_{n,0}$*.*

Furthermore, let $\mathsf{G}=\mathsf{B}-\mathsf{A}$ *and* $g=c-b$*, assume that* $\sigma_{\min}(\langle W,\mathsf{A}W\rangle)>\|\langle W,\mathsf{G}W\rangle\|_2$ *holds and define*

$$\eta:=\frac{\|\,\mathsf{G}|_{\mathcal{W}}\|}{\sigma_{\min}(\langle W,\mathsf{A}W\rangle)-\|\langle W,\mathsf{G}W\rangle\|_2}.$$

Then the i*-th residual* $\widehat{r}_i$ *of the GMRES method applied to the linear system*

$$\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B}\widehat{y}=\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c \tag{3.84}$$

satisfies

$$\|\widehat{r}_i\|\leq\|\tilde{r}_i\|_2+\frac{|\partial\Lambda_\delta(\widehat{\mathsf{A}}_i)|}{2\pi\delta}\left(\frac{\varepsilon_i}{\delta-\varepsilon_i}(\|q\|_2+\beta)+\beta\right)\sup_{\lambda\in\Lambda_\delta(\widehat{\mathsf{A}}_i)}|p_i(\lambda)| \tag{3.85}$$

for all $\delta>\varepsilon_i$*, where*

$$\begin{aligned}\widehat{\mathsf{A}}_i &:= (\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A}+\mathsf{F}_i)\big|_{\mathcal{W}^\perp}, \\ \varepsilon_i &:= \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\left(\eta(\|\mathsf{A}\|+\|\mathsf{G}\|)+\|\mathsf{G}\|\right)+\left\|\widehat{\mathbf{R}}_i\right\|_2 \\ \text{and}\quad \beta &:= \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\left(\eta(\|b\|+\|g\|)+\|g\|\right)+\left\|\widehat{\mathbf{N}}\tilde{q}\right\|_2.\end{aligned}$$

Proof. Let $\widehat{V}_{n+m-k}$ be as in theorem 3.59. It then follows from proposition 2.52 that

$$\|\tilde{r}_i\|_2 = \left\|p_i(\widehat{\mathbf{H}})\,\|q\|_2\,e_1\right\|_2 = \left\|\widehat{V}_{n+m-k}p_i(\widehat{\mathbf{H}})\,\|q\|_2\,e_1\right\| = \left\|p_i(\tilde{\mathsf{A}}_i)\tilde{b}\right\|$$

with

$$\tilde{b} = \mathsf{P}_{\mathcal{K}_n+\mathcal{U}}\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b = \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b + f,$$

where = $\mathsf{P}_{(\mathcal{K}_n+\mathcal{U})^\perp}f\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b$ is as in theorem 3.59. Note that p_i is the GMRES residual polynomial for the i-th step of GMRES for the linear system $\tilde{\mathsf{A}}_i\tilde{x} = \tilde{b}$ and observe that

$$\begin{aligned} \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B} &= \tilde{\mathsf{A}}_i + \tilde{\mathsf{F}}_i && \text{with} \quad \tilde{\mathsf{F}}_i := \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B} - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\mathsf{A} - \mathsf{F}_i \\ \text{and} \quad \widehat{c} := \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c &= \tilde{b} + \tilde{f} && \text{with} \quad \tilde{f} := \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}c - \mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}b - f. \end{aligned}$$

Note that the application of corollary 3.63 yields with the norm equalities for F_i and f in theorem 3.59 the inequalities

$$\begin{aligned} \left\|\tilde{\mathsf{F}}_i\right\| &\leq \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\left(\eta(\|\mathsf{A}\| + \|\mathsf{G}\|) + \|\mathsf{G}\|\right) + \|\mathsf{F}_i\| \\ &= \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\left(\eta(\|\mathsf{A}\| + \|\mathsf{G}\|) + \|\mathsf{G}\|\right) + \left\|\widehat{\mathbf{R}}_i\right\|_2 = \varepsilon_i \\ \text{and} \quad \left\|\tilde{f}\right\| &\leq \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\left(\eta(\|b\| + \|g\|) + \|g\|\right) + \|f\| \\ &= \left\|\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}\right\|\left(\eta(\|b\| + \|g\|) + \|g\|\right) + \|\mathbf{N}\tilde{q}\|_2 = \beta. \end{aligned}$$

Because $\tilde{b},\widehat{c} \in \mathcal{W}^\perp$ and $\mathcal{R}(\tilde{\mathsf{A}}_i),\mathcal{R}(\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B}),\mathcal{R}(\tilde{\mathsf{F}}_i) \subseteq \mathcal{W}^\perp$, the operators can be restricted to the subspace $\mathcal{W}^\perp$, i.e.,

$$\begin{aligned} \widehat{\mathsf{A}}_i := \tilde{\mathsf{A}}_i\big|_{\mathcal{W}^\perp} : \mathcal{W}^\perp \longrightarrow \mathcal{W}^\perp, \qquad \widehat{\mathsf{F}}_i := \tilde{\mathsf{F}}_i\big|_{\mathcal{W}^\perp} : \mathcal{W}^\perp \longrightarrow \mathcal{W}^\perp \\ \text{and} \quad \widehat{\mathsf{B}} := \mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B}\big|_{\mathcal{W}^\perp} = \widehat{\mathsf{A}}_i + \widehat{\mathsf{F}}_i : \mathcal{W}^\perp \longrightarrow \mathcal{W}^\perp. \end{aligned}$$

The i-th residual $\overline{r}_i$ of the GMRES method applied to the linear system $\widehat{\mathsf{A}}_i\widehat{x} = \tilde{b}$ thus satisfies

$$\|\overline{r}_i\| = \left\|p_i(\widehat{\mathsf{A}}_i)\tilde{b}\right\| = \|\tilde{r}_i\|_2$$

while the i residual $\widehat{r}_i$ of the GMRES method applied to the linear system (3.84) equals the i-th residual of the GMRES method applied to the linear system $\widehat{\mathsf{B}}\widehat{y} = \widehat{c}$. Now theorem 3.41 can be applied in order to obtain

$$\begin{aligned} \|\widehat{r}_i\| &\leq \|\overline{r}_i\| + \frac{|\partial\Lambda_\delta(\widehat{\mathsf{A}}_i)|}{2\pi\delta}\left(\frac{\varepsilon_i}{\delta-\varepsilon_i}\|\widehat{c}\| + \left\|\widehat{c} - \tilde{b}\right\|\right)\sup_{\lambda\in\Lambda_\delta(\widehat{\mathsf{A}}_i)}|p_i(\lambda)| \\ &= \|\tilde{r}_i\|_2 + \frac{|\partial\Lambda_\delta(\widehat{\mathsf{A}}_i)|}{2\pi\delta}\left(\frac{\varepsilon_i}{\delta-\varepsilon_i}\left\|\tilde{b} + \tilde{f}\right\| + \left\|\tilde{f}\right\|\right)\sup_{\lambda\in\Lambda_\delta(\widehat{\mathsf{A}}_i)}|p_i(\lambda)| \\ &\leq \|\tilde{r}_i\|_2 + \frac{|\partial\Lambda_\delta(\widehat{\mathsf{A}}_i)|}{2\pi\delta}\left(\frac{\varepsilon_i}{\delta-\varepsilon_i}(\left\|\tilde{b}\right\| + \beta) + \beta\right)\sup_{\lambda\in\Lambda_\delta(\widehat{\mathsf{A}}_i)}|p_i(\lambda)| \\ &= \|\tilde{r}_i\|_2 + \frac{|\partial\Lambda_\delta(\widehat{\mathsf{A}}_i)|}{2\pi\delta}\left(\frac{\varepsilon_i}{\delta-\varepsilon_i}(\|q\|_2 + \beta) + \beta\right)\sup_{\lambda\in\Lambda_\delta(\widehat{\mathsf{A}}_i)}|p_i(\lambda)|. \end{aligned}$$

□

Remark 3.65. An algorithm for computing the right hand side of bound (3.85) is implemented in [60] as `krypy.deflation.bound_pseudo`. A representation of the polynomial p_i is obtained by computing harmonic Ritz values of $\widehat{\mathbf{H}}$, cf. theorem 2.57. Due to the fact that p_i is holomorphic, its maximum is attained on the boundary of the pseudospectrum, i.e., $\sup_{\lambda \in \Lambda_\delta(\widehat{\mathsf{A}}_i)} |p_i(\lambda)| = \sup_{\lambda \in \partial\Lambda_\delta(\widehat{\mathsf{A}}_i)} |p_i(\lambda)|$. The pseudospectrum package PseudoPy [61] is used to approximate $\partial\Lambda_\delta(\widehat{\mathsf{A}}_i)$ by $\partial\Lambda_\delta(\widehat{\mathbf{H}})$, see the discussion in the remaining part of this subsection. The boundary $\partial\Lambda_\delta(\widehat{\mathbf{H}})$ is approximated by polygons and the polynomial p_i is then evaluated on the polygons' vertices in order to obtain an approximation of the supremum in bound (3.85).

Remark 3.66. Skimming through the proof of theorem 3.64 reveals why only the projections $\mathsf{P}_{\mathcal{W}^\perp,\mathsf{A}\mathcal{W}}$ and $\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}$ are considered in this subsection: if the projections $\mathsf{P}_{(\mathsf{A}\mathcal{W})^\perp}$ and $\mathsf{P}_{(\mathsf{B}\mathcal{W})^\perp}$ are used, then the restrictions of the deflated operators $\mathsf{P}_{(\mathsf{A}\mathcal{W})^\perp}\mathsf{A}\big|_{\mathcal{W}^\perp}$ and $\mathsf{P}_{(\mathsf{B}\mathcal{W})^\perp}\mathsf{B}\big|_{\mathcal{W}^\perp}$ are operators from $\mathcal{W}^\perp$ to $(\mathsf{A}\mathcal{W})^\perp$ and $(\mathsf{B}\mathcal{W})^\perp$, respectively. Thus, the restricted operators cannot be used in the above setting because $\left(\mathsf{P}_{(\mathsf{A}\mathcal{W})^\perp}\mathsf{A}\big|_{\mathcal{W}^\perp}\right)^i$ and $\left(\mathsf{P}_{(\mathsf{B}\mathcal{W})^\perp}\mathsf{B}\big|_{\mathcal{W}^\perp}\right)^i$ are not well defined in general. Instead, the unrestricted operators could be used but then the spectra and thus also the pseudospectra in (3.85) contain zero if $\mathcal{W} \neq \{0\}$ which clearly renders the bound useless.

Remark 3.67. In finite precision arithmetic, the computed Arnoldi or Lanczos relations may deviate significantly from their exact arithmetic counterparts. As mentioned in section 2.10, this applies particularly to methods that are based on Lanczos recurrences. Theorem 3.53 and the descending theorems 3.59 and 3.64 rely on the orthogonality of the provided bases. In experiments, minor deviations from orthogonality do not seem to have an excessive effect on the computed quantities but the impact of the complete loss of orthogonality on the above quantities is unexplored.

Note that Sifuentes, Embree and Morgan [153] also applied the pseudospectral bound from their article to deflated methods, see theorem 4.5 in [153]. However, their approach is fundamentally different and serves a very different purpose. In [153], the goal is to analyze *restarted* GMRES with a deflation-based preconditioner that uses approximate invariant subspaces. In contrast, the goal is here to obtain estimates for deflated GMRES for a linear system with the data that has been computed while solving a *different* linear system. In [153], it is assumed that the GMRES convergence behavior is known in the case where the preconditioner is built with an *exact* invariant subspace. Clearly, this information is not available in practice and theorem 3.64 does not have this requirement. Instead, the new theorem shows how residual norms and polynomials can be computed cheaply from *available* data. Furthermore, a condition on the separation of certain eigenvalues has to be fulfilled in [153] while theorem 3.64 only depends on a condition that is required to guarantee a well-defined projection for the next linear system. In contrast to [153], the new theorem is not restricted to Ritz vectors as deflation vectors but holds for any vectors from the provided subspaces. Both approaches have in common that they require the pseudospectrum of a potentially large matrix or operator. The efficient approximation of this pseudospectrum is dealt with in the following.

With respect to the applicability of the bound in theorem 3.64, one puzzle still remains to be solved: the pseudospectrum $\Lambda_\delta(\widehat{\mathsf{A}}_i)$ is not available in practice. Nothing is known about the behavior of the operator A or $\widehat{\mathsf{A}}_i$ on the subspace $(\mathcal{K}_n + \mathcal{U})^\perp$ and as forecasted in the beginning of section 3.4, it is unlikely that a precise and entirely non-heuristic prediction of the convergence behavior is possible with the very limited data that is available in practice. Nevertheless, an approximation can fill the gap in applications because the action of $\widehat{\mathsf{A}}_i$ is known on the subspace $(\mathcal{K}_n + \mathcal{U}) \cap \mathcal{W}^\perp$.

If $\widehat{V}_{n+m-k} \in \mathcal{H}^{n+m-k}$ is as defined in theorem 3.59, then by construction $[\![\widehat{V}_{n+m-k}]\!] = (\mathcal{K}_n + \mathcal{U}) \cap \mathcal{W}^\perp$. Instead of asking for the full ε_i-pseudospectrum of $\widehat{\mathsf{A}}_i$, a projection technique can be employed as proposed by Toh and Trefethen in [174]; see also chapters 40 and 46 in the book of Trefethen and Embree [177]. In the following, the definition of the pseudospectrum of rectangular matrices as given in [177] is important.

Definition 3.68. Let $\mathbf{A} \in \mathbb{C}^{m,n}$ with $m \geq n$ and let $\varepsilon > 0$. The ε-pseudospectrum of $\mathbf{A}$ is defined by

$$\Lambda_\varepsilon(\mathbf{A}) = \{\lambda \in \mathbb{C} \mid \sigma_{\min}(\lambda \underline{\mathbf{I}}_n - \mathbf{A}) < \varepsilon\},$$

where $\underline{\mathbf{I}}_n = \begin{bmatrix} \mathbf{I}_n \\ 0_{m-n,n} \end{bmatrix}$.

The following theorem characterizes the pseudospectrum of rectangular matrices and can be found in [177].

Theorem 3.69. *Let $\mathbf{A} \in \mathbb{C}^{m,n}$ with $m \geq n$ and let $\varepsilon > 0$. Then*

1. $\Lambda_\varepsilon\left(\mathbf{A}\begin{bmatrix} \mathbf{I}_k \\ 0_{m-k,k} \end{bmatrix}\right) \subseteq \Lambda_\varepsilon(\mathbf{A})$ *for* $k \in \{1, \ldots, m-1\}$.
2. $\Lambda_\varepsilon(\mathbf{A}) \subseteq \Lambda_\varepsilon\left(\begin{bmatrix} \mathbf{I}_k & 0_{k,n-k} \end{bmatrix} \mathbf{A}\right)$ *for* $k \in \{1, \ldots, n-1\}$.

Proof. See theorem 46.2 in [177]. □

The following corollary applies the above theorem to the situation of a decomposition as given in theorem 3.53. The idea is not new and has been used with a regular Arnoldi relation in [174].

Corollary 3.70. *Let $V \in \mathcal{H}^n$ and $W \in \mathcal{H}^m$ such that $\langle [V, W], [V, W] \rangle = \mathbf{I}_{n+m}$ and assume that*

$$\mathsf{A}V = V\mathbf{G} + W\mathbf{R}$$

for $\mathsf{A} \in \mathcal{L}(\mathcal{H})$, $\mathbf{G} \in \mathbb{C}^{n,n}$ and $\mathbf{R} \in \mathbb{C}^{m,n}$.

Then for any $\varepsilon > 0$

$$\Lambda_\varepsilon\left(\begin{bmatrix}\mathbf{G}\\ \mathbf{R}\end{bmatrix}\right) \subseteq \Lambda_\varepsilon(\mathsf{A}).$$

Proof. Let $N = \dim(\mathsf{A})$ and $Z \in \mathcal{H}^{N-m-n}$ such that $\langle [V, W, Z], [V, W, Z]\rangle = \mathbf{I}_N$. Then

$$\Lambda_\varepsilon(\mathsf{A}) = \Lambda_\varepsilon\left(\langle [V, W, Z], \mathsf{A}[V, W, Z]\rangle\right)$$

and by theorem 3.69

$$\Lambda_\varepsilon(\langle [V, W, Z], \mathsf{A}V\rangle) = \Lambda_\varepsilon\left(\langle [V, W, Z], \mathsf{A}[V, W, Z]\rangle \begin{bmatrix}\mathbf{I}_n\\ 0_{N-n,n}\end{bmatrix}\right) \subseteq \Lambda_\varepsilon(\mathsf{A}).$$

Now notice that

$$\langle [V, W, Z], \mathsf{A}V\rangle = \begin{bmatrix}\mathbf{G}\\ \mathbf{R}\\ 0_{N-n-m,n}\end{bmatrix} \quad \text{and} \quad \sigma_{\min}\left(\begin{bmatrix}\mathbf{G}\\ \mathbf{R}\\ 0_{N-n-m,n}\end{bmatrix}\right) = \sigma_{\min}\left(\begin{bmatrix}\mathbf{G}\\ \mathbf{R}\end{bmatrix}\right).$$

Thus the statement follows with definition 3.68. □

With corollary 3.70, an approximation to the pseudospectrum of $\widehat{\mathsf{A}}_i$ can be computed in the setting of theorem 3.64 as demonstrated in the following theorem.

Theorem 3.71. *Let the assumptions of theorem 3.64 hold. Then*

$$\Lambda_\varepsilon\left(\begin{bmatrix}\widehat{\mathbf{H}}\\ \mathbf{S}_i\end{bmatrix}\right) \subseteq \Lambda_\varepsilon(\widehat{\mathsf{A}}_i)$$

holds for any $\varepsilon > 0$ with $\mathbf{S}_i := \widehat{\mathbf{R}}_{n+m-k}\operatorname{diag}(0_{i,i}, \mathbf{I}_{n+m-k-i})$.

Proof. It follows from theorem 3.64 that

$$\begin{aligned}\widehat{\mathsf{A}}_i\widehat{V}_{n+m-k} &= (\widehat{\mathsf{A}}_{n+m-k} + \mathsf{F}_i - \mathsf{F}_{n+m-k})\widehat{V}_{n+m-k}\\ &= \left(\widehat{\mathsf{A}}_{n+m-k} - Z\widehat{\mathbf{R}}_i\widehat{V}_i^\star + Z\widehat{\mathbf{R}}_{n+m-k}\widehat{V}_{n+m-k}^\star\right)\widehat{V}_{n+m-k}\\ &= \left(\widehat{\mathsf{A}}_{n+m-k} + Z(\widehat{\mathbf{R}}_{n+m-k} - [\widehat{\mathbf{R}}_i, 0_{n+m-k-i}])\widehat{V}_{n+m-k}^\star\right)\widehat{V}_{n+m-k}\\ &= \left(\widehat{\mathsf{A}}_{n+m-k} + Z\mathbf{S}_i\widehat{V}_{n+m-k}^\star\right)\widehat{V}_{n+m-k} = \widehat{V}_{n+m-k}\widehat{\mathbf{H}} + Z\mathbf{S}_i.\end{aligned}$$

Now corollary 3.70 applies with $V = \widehat{V}_{n+m-k}$ and $W = Z$ and the proof is complete. □

Theorem 3.71 shows a way to approximate the pseudospectrum of $\widehat{\mathsf{A}}_i$ from the inside. Although parts of the pseudospectrum may be missing, theorem 3.71 uses all information about $\widehat{\mathsf{A}}_i$ that is known, namely the behavior on the subspace $[\![\widehat{V}_{n+m-k}]\!] = (\mathcal{K}_n + \mathcal{U}) \cap \mathcal{W}^\perp$. In many cases, such a projection scheme onto a small subspace approximates the actual pseudospectrum very well. Examples and counter-examples with large-scale matrices and projections onto Krylov subspaces (without deflation) can be found in [177, chapter 40].

Another approximation to the pseudospectrum can be obtained by the pseudospectrum of $\widehat{\mathbf{H}} = \langle \widehat{V}_{n+m-k}, \mathsf{A}\widehat{V}_{n+m-k} \rangle$. However, $\Lambda_\varepsilon(\widehat{\mathbf{H}})$ does not satisfy the containment property in theorem 3.71 in general, i.e.,

$$\Lambda_\varepsilon(\widehat{\mathbf{H}}) \notin \Lambda_\varepsilon(\widehat{\mathsf{A}}_i).$$

Nevertheless, $\Lambda_\varepsilon(\widehat{\mathbf{H}})$ can provide a meaningful approximation to $\Lambda_\varepsilon(\widehat{\mathsf{A}}_i)$. If A is normal or self-adjoint, then $\widehat{\mathbf{H}}$ is normal or Hermitian, respectively, and $\Lambda_\varepsilon(\widehat{\mathbf{H}})$ becomes trivial to compute with the spectrum $\Lambda(\widehat{\mathbf{H}})$:

$$\begin{aligned} \Lambda_\varepsilon(\widehat{\mathbf{H}}) &= \bigcup_{\mu \in \Lambda(\mathbf{H})} \{\lambda \in \mathbb{C} \mid |\lambda - \mu| < \varepsilon\} \qquad \text{if } \widehat{\mathbf{H}} \text{ is normal} \\ \text{and} \quad \Lambda_\varepsilon(\widehat{\mathbf{H}}) &= \bigcup_{\mu \in \Lambda(\mathbf{H})} \{\lambda \in \mathbb{R} \mid |\lambda - \mu| < \varepsilon\} \qquad \text{if } \widehat{\mathbf{H}} \text{ is Hermitian.} \end{aligned}$$

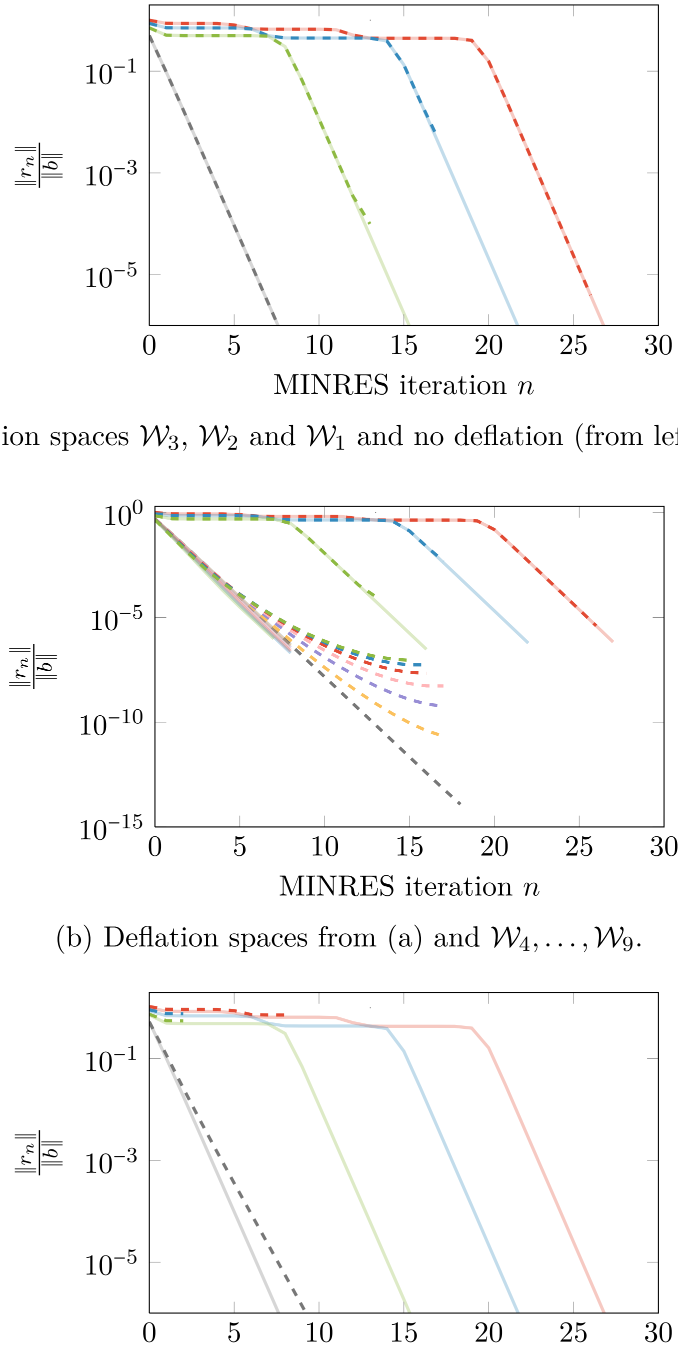
The above approximations of pseudospectra are implemented in PseudoPy [61]. For a non-normal matrix, the resolvent norm is sampled in a neighborhood of the spectrum which is guaranteed to contain the pseudospectrum's boundary that is of interest. The actual computation of bound (3.85) is described in remark 3.65.

Example 3.72. Similar to the experiment with a priori bounds in example 3.52, the approximate Krylov subspace bound (3.85) is applied to the linear system from example 3.23. It is again assumed that 27 MINRES iterations have been performed for the linear system $\mathsf{A}x = b$ and that the task is to find an optimal choice of Ritz vectors from the constructed Krylov subspace for deflation. Let $\mathcal{W}_i$ denote the subspace that is spanned by the i Ritz vectors that correspond to the i Ritz values of smallest magnitude.

In figure 3.12a, the bound (3.85) is plotted together with the actual convergence history of MINRES without deflation and with the deflation subspaces $\mathcal{W}_1$, $\mathcal{W}_2$ and $\mathcal{W}_3$. The bound is able to accurately capture the convergence behavior for all cases. For $\mathcal{W}_1$ and $\mathcal{W}_2$, the bound is not able to give a meaningful result beyond the steps 17 and 13 because the second summand in (3.85) grows too large such that the residual norm would start to increase. With $\mathcal{W}_3$ this happens only after the bound dropped below 10^{-14} (see figure 3.12b) which is surprising given that the original linear system was only solved up to the tolerance 10^{-6}.

Figure 3.12b shows that no significant improvement can be achieved by choosing more than 3 Ritz vectors. Note that the bound (3.85) is still able to capture the initial phase of convergence with 4 or more Ritz vectors before drifting apart from the actual convergence curve.

Besides the ability to bound the convergence behavior for the same linear system, theorem 3.64 allows to make statements for a perturbed linear system $\mathsf{B}y = c$ with $\mathsf{B} = \mathsf{A} + \mathsf{G}$ and $c = b + g$. The assumption $\sigma_{\min}(\langle W, \mathsf{A}W \rangle) > \|\langle W, \mathsf{G}W \rangle\|$ is, e.g., fulfilled by a random matrix G of norm 10^{-7}. Figure 3.12c shows the bound and actual convergence histories where g is chosen as a random vector of norm 0.1. Note that there is no visible difference of the actual convergence histories to the



(a) Deflation spaces $\mathcal{W}_3$, $\mathcal{W}_2$ and $\mathcal{W}_1$ and no deflation (from left to right).

(b) Deflation spaces from (a) and $\mathcal{W}_4, \dots, \mathcal{W}_9$.

(c) Same as (a) but for a perturbed matrix $\mathsf{A} + \mathsf{G}$ and right hand side $b + g$ with $\|\mathsf{G}\| = 10^{-7}$ and $\|g\| = 0.1$.

Figure 3.12.: Approximate Krylov subspace bound (3.85) (dashed) and actual convergence (light solid) of MINRES for example 3.72.

corresponding ones in figure 3.12a. The run with 3 Ritz vectors still provides a meaningful bound for the actual behavior but the bound does not give much insight for 2 or less Ritz vectors.

In the next section 3.5, bound (3.85) is evaluated for a less academic linear system with a non-symmetric matrix that stems from a convection-diffusion problem.

3.5. Numerical experiments with GMRES and a convection-diffusion problem

In this section, the results of the preceding section are applied to the GMRES method for the solution of a convection-diffusion model problem. The considered partial differential equation is

$$\begin{aligned} -\nu\Delta u + w \cdot \nabla u = 0 \quad &\text{in} \quad \Omega :=]-1,1[\times]-1,1[\\ \text{and} \quad u = q \quad &\text{on} \quad \partial\Omega, \end{aligned} \tag{3.86}$$

where $u : \overline{\Omega} \longrightarrow \mathbb{R}$ is the sought solution for the diffusion parameter $\nu > 0$, the velocity field $w = [w_1, w_2]^\mathsf{T} : \Omega \longrightarrow \mathbb{R}^2$ and the Dirichlet boundary function $q : \partial\Omega \longrightarrow \mathbb{R}$. Here, the reference problem 3.1.4 from the book of Elman, Silvester and Wathen [46] is used as a basis for numerical experiments. The velocity field is a recirculating flow defined by $w(x,y) := [2y(1-x^2), -2x(1-y^2)]^\mathsf{T}$ and the boundary conditions are given by

$$q(x,y) = \begin{cases} 1 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.87}$$

The diffusion parameter is chosen as $\nu = \frac{1}{200}$ which renders the equation convection-dominated. A finite element discretization is obtained for equation (3.86) with linear Lagrange elements on a regular triangulation exhibiting 1301 vertices. In order to avoid non-physical oscillations in the approximate solution, the discretization incorporates the streamline diffusion stabilization (SUPG) as described in [46]. The finite element implementation is based on the Python package DOLFIN from the FEniCS project [108, 109]. Figure 3.13 shows the discrete solution for the described configuration.

Besides requiring special attention in the discretization, the domination of the convection term also poses a challenge for Krylov subspace methods. The finite element discretization of (3.86) with streamline diffusion yields a non-symmetric and highly non-normal matrix $\mathbf{A}_h \in \mathbb{R}^{N,N}$ and a corresponding right hand side $b_h \in \mathbb{R}^N$ that incorporates the Dirichlet boundary conditions. The typical convergence behavior of GMRES for the linear system $\mathbf{A}_h x = b_h$ is an initial phase of slow convergence which is followed by a fast residual reduction. A characteristic quantity of convection-diffusion problems is the mesh Péclet number P_h which is defined by $P_h := \frac{h\|w\|}{2\nu}$ for a regular mesh with grid size h and constant velocity norm $\|w\|$. In

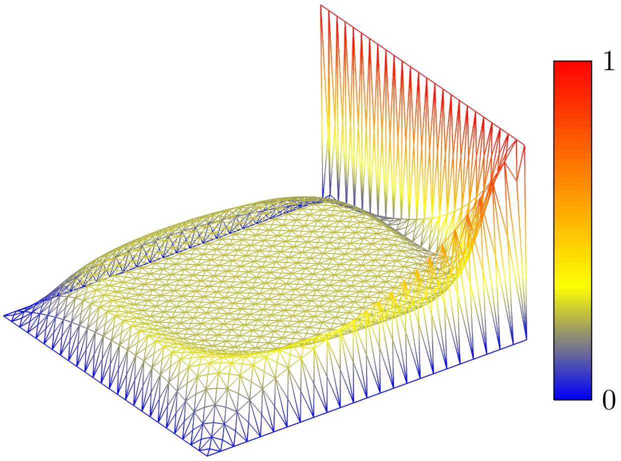


Figure 3.13.: Three-dimensional surface plot of the discrete solution of (3.86) on a mesh with 1301 vertices.

the model problem considered here, the maximal element Péclet number $P_{h_i} = \frac{h_i \|w\|}{2\nu}$ is ≈ 16.

In [52], Ernst analyzed the convergence behavior of GMRES for convection-dominated convection-diffusion model problems and showed that the condition number of the eigenvector basis of the resulting matrix grows exponentially with both the mesh Péclet number and grid size. Therefore, the spectral bound (2.24) gives no insight into the actual convergence behavior of GMRES. It was observed that the initial phase of slow convergence can be characterized by the field of values bound (2.26) while the second phase of faster convergence appears to be dominated by spectral properties. Furthermore, Ernst conjectured that the length of the initial phase of slow convergence is determined by the time the boundary values take to be transported along the longest streamline through the domain. In [104], Liesen and Strakoš illustrated and explained the drastic influence of the right hand side b_h on the initial period of slow convergence. Like in equation (3.86), the forcing term is zero in [104] and thus the right hand side b_h is determined by the boundary values.

For the experiments in this section, an incomplete LU (ILU) decomposition of $\mathbf{A}_h$ is used as a preconditioner $\mathbf{M} \approx \mathbf{A}_h^{-1}$. Instead of $\mathbf{A}_h x = b_h$, GMRES is applied to the linear system

$$\mathbf{A}x = b \tag{3.88}$$

with $\mathbf{A} = \mathbf{M}\mathbf{A}_h$, $b = \mathbf{M}b_h$ and the initial guess x_0 is zero except for the boundary unknowns where it satisfies the boundary conditions.

The convergence history of GMRES applied to (3.88) is plotted as the solid red line in figure 3.14 and clearly shows that the slow initial convergence is not remedied by the use of the ILU preconditioner. If Ritz pairs are computed from the Krylov subspace that has been generated by GMRES and a few Ritz vectors are used as deflation vectors for the *same* linear system, then the initial phase of near-stagnation is relieved or even completely removed, see the light colored curves in figure 3.14. The approximate Krylov subspace bound (3.85) is able to reproduce the convergence behavior for deflated GMRES accurately in the early phase until it diverges around 10^{-6}. However, the first summand $\|\tilde{r}_n\|_2$ (dotted) in (3.85) still lies on the actual convergence curve far beyond the point where the bound (3.85) diverges.

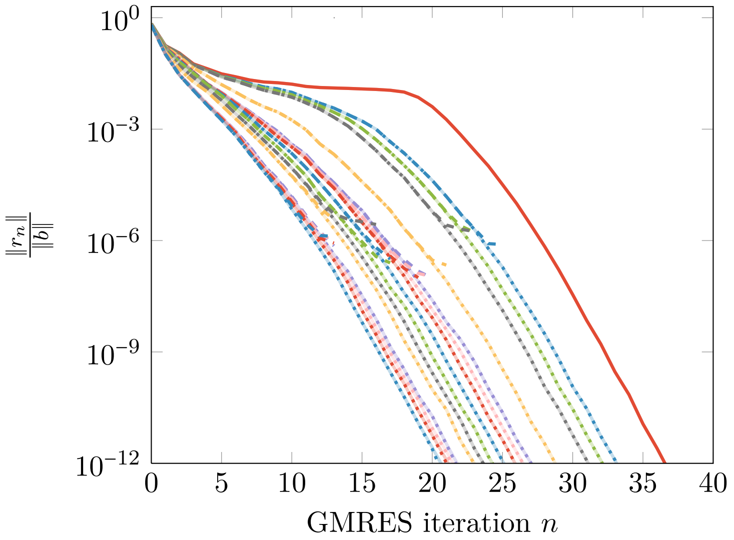


Figure 3.14.: Convergence histories for (deflated) GMRES applied to the linear system (3.88). The solid red curve represents the original run (without deflation) from which Ritz pairs are computed. The curves in light colors represent GMRES with deflation of one up to 14 Ritz vectors (from right to left) corresponding to the Ritz values of smallest magnitude. Bound (3.85) and its first summand $\|\tilde{r}_n\|_2$ are plotted as dashed and dotted lines, respectively. Note that the dotted curves coincide with the light curves.

An interesting question is how the bound (3.85) behaves in a sequence of linear systems. Similar to the previous sections, only two subsequent linear systems are considered here. First, it is analyzed how perturbations of the right hand side b affect the bound before considering perturbations in the matrix $\mathbf{A}$.

In order to perturb the right hand side, the boundary function (3.87) is modified with $\gamma \geq 0$ to

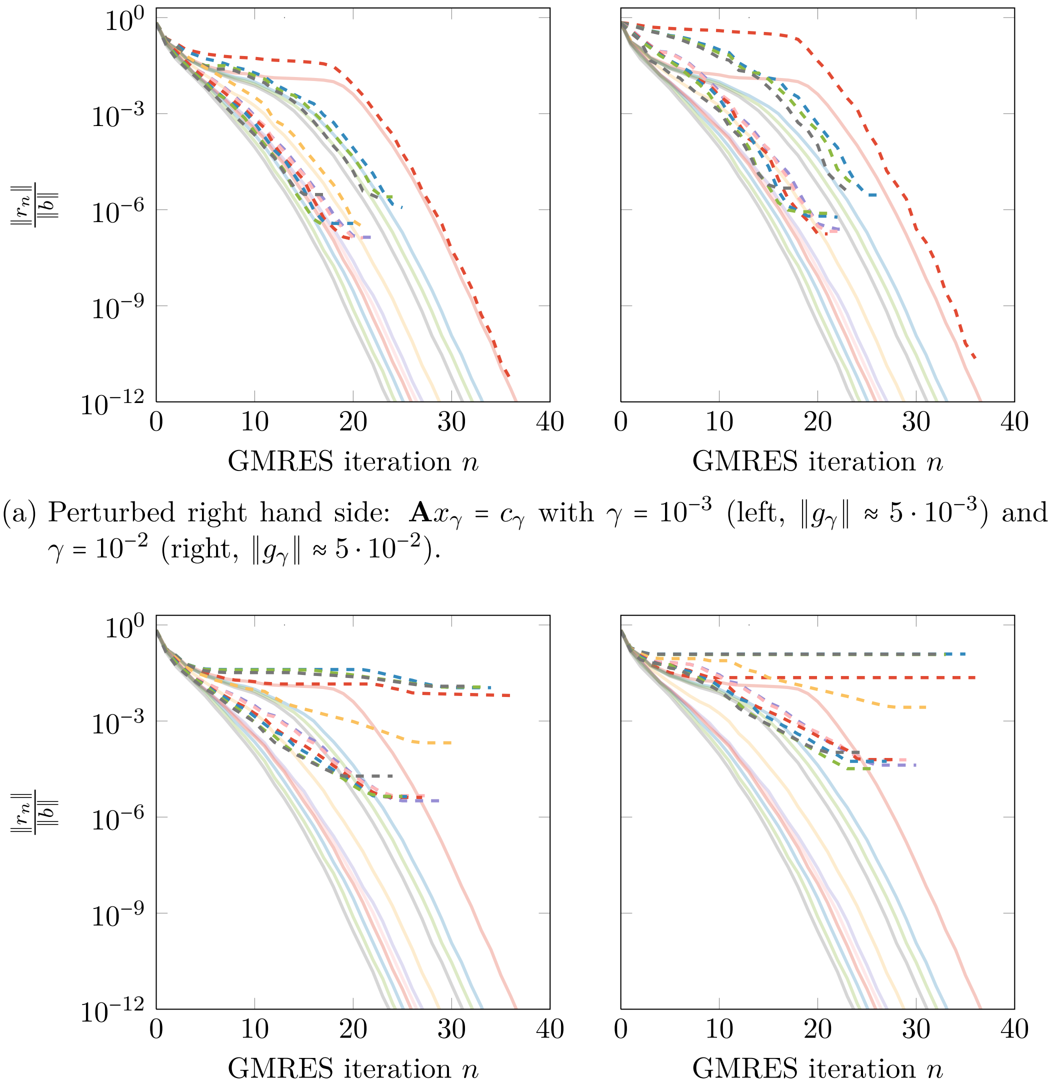
$$q_\gamma(x,y) = \begin{cases} 1+\gamma(1-y^2) & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The resulting (preconditioned) right hand side is denoted by c_γ and the difference by $g_\gamma = b - c_\gamma$. In figure 3.15a, the bound is evaluated for the linear system $\mathbf{A}x_\gamma = c_\gamma$, where γ takes the values 10^{-3} and 10^{-2} (dashed lines). The light colored curves represent the actual GMRES residual norms for the perturbed right hand side and do not differ significantly from the unperturbed counterparts in figure 3.14.

In the next experiment, the right hand side remains the same but the matrix is perturbed by replacing the velocity field w in (3.86) by $w_\omega := (1+\omega)w$ for a $\omega \in \mathbb{R}$. The resulting matrix $\mathbf{B}_{h,\omega}$ is preconditioned with the matrix $\mathbf{M}$ that is based on the unperturbed matrix $\mathbf{A}_h$. Thus, the linear system $\mathbf{B}_\omega x_\omega = b$ has to be solved, where $\mathbf{B}_\omega = \mathbf{M}\mathbf{B}_{h,\omega}$. The difference of the matrices is denoted by $\mathbf{G}_\omega = \mathbf{A} - \mathbf{B}_\omega$. Figure 3.15b shows that already very small perturbations in the matrix lead to crude estimations with bound (3.85).

Note that the second summand in inequality (3.85) depends linearly on β and that the range $]\epsilon_i, \infty[$ for the level δ of the pseudospectrum is thus not affected by the choice of the right hand side. In contrast, a perturbation of the matrix enlarges ϵ_i and thus restricts the range for δ. Usually, large values of δ minimize (3.85) in early iterations while small values close to ϵ_i are required for later iterations because the polynomial p_i is of higher degree and therefore grows faster away from its zeros. This observation explains why the bound is much more sensitive to perturbations of the matrix than of the right hand side.

The above experiments show that the bound (3.85) often severely overestimates the residual norms due to the presence of the pseudospectral term. Because the first summand $\|\tilde{r}_n\|_2$ captures the behavior in the above experiments far beyond the point where the pseudospectral term begins to dominate, it is reasonable to use only $\|\tilde{r}_n\|_2$ as an indicator for the effectiveness of a given set of deflation vectors. Furthermore, the computation of $\|\tilde{r}_n\|_2$ is cheap because the extraordinary expensive computation of the pseudospectrum and the construction and evaluation of the polynomials in (3.85) is not required. This strategy is combined with timings of the involved operations in chapter 4 for the solution of nonlinear Schrödinger equations. Clearly, dropping the pseudospectral term in inequality (3.85) renders the bound ignorant of perturbations in the matrix and right hand side. Recall that example 3.38 and figure 3.7 show that a small perturbation of the matrix or right hand side may result in a dramatic change of the convergence behavior of Krylov subspace methods.



(a) Perturbed right hand side: $\mathbf{A}x_\gamma = c_\gamma$ with $\gamma = 10^{-3}$ (left, $\|g_\gamma\| \approx 5 \cdot 10^{-3}$) and $\gamma = 10^{-2}$ (right, $\|g_\gamma\| \approx 5 \cdot 10^{-2}$).

(b) Perturbed matrix: $\mathbf{B}_\omega x_\omega = b$ with $\omega = 10^{-6}$ (left, $\|\mathbf{G}_\omega\| \approx 5 \cdot 10^{-6}$) and $\omega = 10^{-5}$ (right, $\|\mathbf{G}_\omega\| \approx 5 \cdot 10^{-5}$).

Figure 3.15.: Bound (3.85) (dashed) and actual residual norms (light solid) with perturbed right hand side c_γ (top) and perturbed matrix $\mathbf{B}_\omega$ (bottom). The deflation vectors are chosen as the $0, \ldots, 10$ Ritz vectors corresponding to the Ritz values of smallest magnitude.

4. Numerical solution of nonlinear Schrödinger equations

In [63], Schlömer and the author investigated a recycling MINRES variant for the numerical solution of nonlinear Schrödinger equations. Nonlinear Schrödinger equations are used to describe a wide variety of physical systems like superconductivity, quantum condensates, nonlinear acoustics [159], nonlinear optics [65], and hydrodynamics [129]. In the numerical solution of nonlinear Schrödinger equations with Newton's method, a linear system has to be solved with the Jacobian operator for each Newton update. Because the Jabobian operator is self-adjoint with respect to a non-Euclidean inner product and indefinite in general, the MINRES method can be used for the computation of the Newton updates. However, the spectrum of the Jacobian operators becomes unfavorable once the Newton iterates approach a nonzero solution and the MINRES residual norms stagnate for several iterations. The number of required MINRES iterations as well as the overall time consumption can be significantly reduced with a recycling strategy that uses Ritz vectors as deflation vectors as proposed in chapter 3. One important instance of nonlinear Schrödinger equations is the Ginzburg–Landau equation that models phenomena of certain superconductors. The nonlinear Schrödinger equations and the Ginzburg–Landau equation are only discussed briefly here. An in-depth description with further references can be found in [63] and the articles by Schlömer, Avitabile and Vanroose [151] and Schlömer and Vanroose [149]. The presented numerical results in this chapter and the used software package *PyNosh* [64] for the solution of nonlinear Schrödinger equations have been developed jointly by Schlömer and the author and have been published to a great extent in [63].

4.1. Nonlinear Schrödinger equations

For $d \in \{2,3\}$ and an open domain $\Omega \subseteq \mathbb{R}^d$, the general nonlinear Schrödinger equation can be derived by minimizing the Gibbs energy in a physical system. Let $\mathsf{S} : \mathcal{X} \longrightarrow \mathcal{Y}$ be defined by

$$\mathsf{S}(\psi) \coloneqq (\mathsf{K} + \mathsf{V} + g|\psi|^2)\psi \quad \text{in } \Omega, \tag{4.1}$$

where $\mathcal{X} \subseteq L^2(\Omega)$ is the natural energy space of the problem and $\mathcal{Y} \subseteq L^2(\Omega)$. The space $\mathcal{X}$ can incorporate appropriate boundary conditions in the setting of a bounded domain. The linear operator $\mathsf{K} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ is assumed to be self-adjoint and positive semidefinite with respect to the inner product $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$, $\mathsf{V} : \Omega \longrightarrow \mathbb{R}$

is a scalar potential and $g > 0$ is the nonlinearity parameter. A state $\psi^\star : \Omega \longrightarrow \mathbb{C}$ is called a solution of the nonlinear Schrödinger equation if

$$\mathsf{S}(\psi^\star) = 0 \tag{4.2}$$

and solutions of interest are nontrivial solutions, i.e., $\|\psi^\star\| \neq 0$. The magnitude $|\psi^\star|^2$ typically describes a particle density or, more generally, a probability distribution. It follows directly from

$$\mathsf{S}(\mathrm{e}^{\mathbb{i}\varphi}\psi) = \mathrm{e}^{\mathbb{i}\varphi}\mathsf{S}(\psi) \tag{4.3}$$

that a solution $\psi^\star$ defines an equivalence class $\{\mathrm{e}^{\mathbb{i}\varphi}\psi^\star \mid \varphi \in \mathbb{R}\}$ of physically equivalent solutions.

If a good-enough initial guess $\psi_0 \in \mathcal{X}$ is given, Newton's method can be used to generate a sequence of iterates $(\psi_i)_{i\in\mathbb{N}_+}$ which converges quadratically towards a solution $\psi^\star$ of the nonlinear Schrödinger equation (4.3). In each step of Newton's method, a linear system has to be solved with the Jacobian operator $\mathsf{J}_\psi : \mathcal{X} \longrightarrow \mathcal{Y}$ of S at ψ which is defined by

$$\mathsf{J}_\psi\phi := (\mathsf{K} + \mathsf{V} + 2g|\psi|^2)\phi + g\psi^2\overline{\phi}. \tag{4.4}$$

The functions ϕ are complex-valued in general and the Jacobian operator J_ψ is not linear if $\mathcal{X}$ and $\mathcal{Y}$ are treated as vector spaces over the field $\mathbb{C}$ because of the complex conjugation. However, the Jacobian J_ψ is linear if $\mathcal{X}$ and $\mathcal{Y}$ are defined as vector spaces over the field $\mathbb{R}$ with the corresponding inner product

$$\langle\cdot,\cdot\rangle_{\mathbb{R}} := \operatorname{Re}\langle\cdot,\cdot\rangle_{L^2(\Omega)}. \tag{4.5}$$

This definition is in accordance with the fact that the specific complex argument of a solution $\psi^\star$ is of no physical relevance, cf. equation (4.3). With an initial guess $\psi_0 \in \mathcal{X}$, the following linear systems have to be solved for $i \in \mathbb{N}_+$:

$$\mathsf{J}_{\psi_i}\delta_i = -\mathsf{S}(\psi_i), \quad \text{where} \quad \psi_{i+1} = \psi_i + \delta_i. \tag{4.6}$$

The adjoint operator of the Jacobian operator J_ψ can be characterized with a result by Schlömer, Avitabile and Vanroose [151] and is of importance for the numerical solution of the linear systems in Newton's method.

Corollary 4.1. *Let $\mathcal{X}, \mathcal{Y} \subseteq L^2(\Omega)$ and let $\psi \in \mathcal{X}$. Then $\mathsf{J}_\psi : \mathcal{X} \longrightarrow \mathcal{Y}$ as defined as above is self-adjoint with respect to the inner product $\langle\cdot,\cdot\rangle_{\mathbb{R}}$, cf. equation* (4.5).

Proof. The statement immediately follows from lemma 3.1 in [151]. □

The spectrum of J_ψ thus is real but J_ψ is indefinite, in general. Furthermore, it follows from (4.3) that for any $\psi \in \mathcal{X}$ the equation

$$\mathsf{J}_\psi(\mathbb{i}\psi) = (\mathsf{K} + \mathsf{V} + 2g|\psi|^2)\mathbb{i}\psi + g\psi^2\overline{\mathbb{i}\psi} = (\mathsf{K} + \mathsf{V} + 2g|\psi|^2)\mathbb{i}\psi - g|\psi|^2\mathbb{i}\psi = \mathbb{i}\mathsf{S}(\psi)$$

holds. For a non-trivial solution $\psi^\star$ of the nonlinear Schrödinger equation (4.2), the Jacobian operator $\mathsf{J}_{\psi^\star}$ thus is singular. The singularity of the Jacobian operator in a solution $\psi^\star$ is again a direct consequence of the fact that $\psi^\star$ defines an equivalence class of physically equivalent solutions.

If $\mathsf{J}_{\psi^\star}$ is positive semidefinite, the solution $\psi^\star$ is called a stable solution. In practice, solutions with low Gibbs energies tend to be stable whereas highly energetic solutions tend to be unstable, i.e., the Jacobian operator $\mathsf{J}_{\psi^\star}$ exhibits negative eigenvalues.

The singularity of the Jacobian operator imposes additional challenges to the numerical solution of nonlinear Schrödinger equations:

1. In the area of attraction of a solution $\psi^\star$, the quadratic convergence of Newton's method cannot be expected if the Jacobian operator $\mathsf{J}_{\psi^\star}$ is singular and only linear convergence is guaranteed, cf. Decker, Keller and Kelley [28], Deuflhard [32] and Kelley [92].

2. Although no linear system has to be solved with an exactly singular Jacobian operator, the spectrum of J_{ψ_i} will exhibit at least one eigenvalue of small magnitude when ψ_i approaches a solution $\psi^\star$. The presence of one or a few small magnitude eigenvalues may seriously impede the convergence of Krylov subspace methods, cf. sections 2.8 and 2.9.

In order to alleviate the effects of a singular Jacobian operator, several approaches have been proposed in the literature. Detailed treatments of this case can be found, e.g., in the works of Reddien [140], Decker and Kelley [29], Decker, Keller and Kelley [28] and Griewank [76]. A widely used approach is known as *bordering* and effectively modifies the original problem such that the singularity is eliminated, cf. Keller [91] and Griewank [76]. In the setting of nonlinear Schrödinger equations this amounts to prescribing the complex argument in a consistent way. The bordering approach has been applied to the Ginzburg–Landau equation in [151] and is naturally generalizable to nonlinear Schrödinger equations. Although bordering yields nonsingular Jacobian operators, the approach has the major disadvantage that the choice of an appropriate preconditioner is not clear for the bordered system even if a good preconditioner is available for the original problem. Thus bordering is impractical for discretizations of PDEs with a large number of unknowns. Furthermore, Schlömer and the author noticed in [63] that the singularity of $\mathsf{J}_{\psi^\star}$ in a solution $\psi^\star$ does not lead to a loss of the quadratic convergence in numerical solutions of the Ginzburg–Landau equation, see figure 4.1. Therefore, the bordering approach is not further pursued here.

The more severe consequence of the singularity of the Jacobian operator $\mathsf{J}_{\psi^\star}$ in a solution $\psi^\star$ is that the numerical solution of the linear systems (4.6) with Krylov subspace methods is complicated by the fact that at least one eigenvalue of the Jacobian operator J_{ψ_i} approaches the origin as the Newton iterates ψ_i approach a solution. In the numerical experiments in section 4.3, it can be observed that the MINRES residual norms stagnate for several iterations and it is shown that recycling strategies are able to remove these phases of stagnation and significantly reduce the number of MINRES iterations as well as the overall time consumption.

4.2. Ginzburg–Landau equation and discretization

An important instance of nonlinear Schrödinger equations (4.2) is the Ginzburg–Landau equation that models the physical phenomenon of superconductivity for extreme-type-II superconductors. Here, only the basic equations and their most important properties are presented. An extensive description with further pointers to literature can be found, e.g., in the article of Du, Gunzburger and Peterson [37]. For $d \in \{2,3\}$ and an open and bounded domain $\Omega \subseteq \mathbb{R}^d$ that describes the superconducting material, the Ginzburg–Landau equations are

$$0 = \begin{cases} \mathsf{K}\psi - \psi(1 - |\psi|^2) & \text{in } \Omega, \\ \mathrm{n} \cdot (-\mathbb{i}\nabla - A)\psi & \text{on } \partial\Omega, \end{cases} \tag{4.7}$$

where the linear operator $\mathsf{K} : \mathcal{X} \longrightarrow \mathcal{Y}$ between $\mathcal{X}, \mathcal{Y} \subset L^2(\Omega)$ is defined by

$$\mathsf{K}\phi := (-\mathbb{i}\nabla - A)^2\phi$$

with a given magnetic vector potential $A \in H^2_{\mathbb{R}^d}(\Omega)$. The operator K is often referred to as the kinetic energy operator and describes the energy of a charged particle when it is exposed to the magnetic field $B = \nabla \times A$. Solutions $\psi^\star$ of the Ginzburg–Landau equation (4.7) describe the density $|\psi^\star|^2$ of electric charge carriers (also referred to as *Cooper pair* density) and are known to fulfill $0 \le |\psi^\star| \le 1$ almost everywhere, cf. [37]. For two-dimensional domains, solutions $\psi^\star$ typically exhibit isolated zeros referred to as *vortices* while lines of zeros are the typical solution pattern for three-dimensional domains, see figure 4.2.

As outlined by Schlömer and Vanroose in [149], the operator K is self-adjoint and positive semidefinite with respect to the inner product $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$ on the space

$$\mathcal{X} = \{\psi \in H^2_{\mathbb{C}}(\Omega) \mid \mathrm{n} \cdot (-\mathbb{i}\nabla - A)\psi = 0 \text{ almost everywhere on } \partial\Omega\}.$$

Note that $\mathcal{X}$ is the subspace of $H^2_{\mathbb{C}}(\Omega)$ that satisfies the boundary conditions in (4.7). Furthermore, K was shown to be positive definite in [149] if and only if the magnetic field B does not vanish. Hence, the operator K and thus the Ginzburg–Landau equation fit in the setting of the nonlinear Schrödinger equations (4.1) with $\mathsf{V} \equiv -1$ and $g = 1$. In each iteration of Newton's method, a linear system (4.6) has to be solved with the Jacobian operator (4.4).

The discretization of the Ginzburg–Landau equation has to be carried out carefully in order to retain important properties of the infinite-dimensional problem such as self-adjointness of the Jacobian operator and gauge invariance of solutions, i.e., invariance with respect to the complex argument. In [149], a finite volume discretization of the Ginzburg–Landau equation is presented that preserves these properties. The discretized Jacobian operator $\mathsf{J}_v : \mathbb{C}^n \longrightarrow \mathbb{C}^n$ at a discrete state $v = [v_1, \ldots, v_n] \in \mathbb{C}^n$ is is defined with $z \in \mathbb{C}^n$ by

$$\mathsf{J}_v z := \left(\mathbf{D}^{-1}\mathbf{K} - \mathbf{I}_n + 2|\mathbf{D}_v|\right) z + \mathbf{D}_v \overline{z}, \tag{4.8}$$

where $\mathbf{K} \in \mathbb{C}^{n,n}$ is a Hermitian and positive-definite matrix, $\mathbf{D} = \operatorname{diag}(|\Omega_1|, \ldots, |\Omega_n|)$ is the diagonal matrix with the volumes of the Voronoi regions $\Omega_1, \ldots, \Omega_n$ that result from a Voronoi tesselation of a disctretization of the domain Ω. The remaining matrices that constitute the Jacobian operator (4.8) are defined by $\mathbf{D}_v := \operatorname{diag}(v_1^2, \ldots, v_n^2)$ and $|\mathbf{D}_v| := \operatorname{diag}(|v_1|^2, \ldots, |v_n|^2)$. The discrete Jacobian operator (4.8) is self-adjoint with respect to the discrete inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}}$ which is defined for $v, w \in \mathbb{C}^n$ by

$$\langle v, w \rangle_{\mathbb{R}} := \operatorname{Re}\left(v^{\mathsf{H}} \mathbf{D} w\right). \tag{4.9}$$

This inner product corresponds to the natural inner product (4.5) of the infinite-dimensional problem.

A remark seems to be appropriate here because the definition of a linear operator in the form (4.8) along with the inner product (4.9) may seem odd at first sight. As discussed in section 4.1, the vector space $\mathbb{C}^n$ is treated as a vector space over the field $\mathbb{R}$, thus effectively doubling the dimension to $2n$. An element $v \in \mathbb{C}^n$ can be represented in the basis $e_1, \ldots, e_n, \mathbb{i}e_1, \ldots, \mathbb{i}e_n$ as the vector $[\operatorname{Re} v, \operatorname{Im} v] \in \mathbb{R}^{2n}$ and the Jacobian operator J_v can therefore be represented as a matrix in $\mathbb{R}^{2n,2n}$ with respect to this basis. However, the equivalent real-valued formulation is not further considered here and instead the natural complex-valued formulation is used.

Summarizing, if an initial guess z_0 is given for Newton's method, then the linear systems

$$\mathsf{J}_{z_i} \delta_i = -\mathsf{S}(z_i) \quad \text{with} \quad z_{i+1} = z_i + \delta_i$$

have to be solved for $i \in \mathbb{N}$ until the Newton residual $\mathsf{S}(z_i)$ satisfies $\|\mathsf{S}(z_i)\| < \epsilon$ for a prescribed tolerance ϵ. Note that the norm $\|\cdot\|$ is the norm that is induced by the natural inner product (4.9).

4.3. Numerical experiments with recycling MINRES

In this section, numerical experiments are conducted with the Ginzburg–Landau equation from section 4.2 in two and in three dimensions.

Setup 4.2 (2D)**.** The domain is chosen as the circle $\Omega^{2\mathrm{D}} := \{x \in \mathbb{R}^2 \mid \|x\|_2 < 5\}$ and the magnetic vector potential is defined as

$$A^{2\mathrm{D}}([x_1, x_2]^{\mathsf{T}}) := m \times \frac{[x_1, x_2, 0]^{\mathsf{T}} - x_0}{\|[x_1, x_2, 0]^{\mathsf{T}} - x_0\|_2^3}$$

which corresponds to the magnetic field that is generated by a dipole at $x_0 := [0, 0, 5]^{\mathsf{T}}$ with orientation $m := [0, 0, 1]^{\mathsf{T}}$. A Delaunay triangulation for Ω^{2D} with $n = 3299$ points was created with *Triangle* [152]. The initial guess $z_0 \in \mathbb{C}^n$ is chosen as the interpolation of $\psi_0^{2\mathrm{D}}(x, y) = \cos(\pi x)$. With this setup, Newton's method takes 27 iterations until the Newton residual norm falls below 10^{-10}. The Newton residual norms and the found approximate solution $\psi^{2\mathrm{D}} := z_{27}$ are visualized in figures 4.1 and 4.2.

Setup 4.3 (3D)**.** The domain is chosen as the three-dimensional “L-shape” $\Omega^{3\mathrm{D}} := \{x \in \mathbb{R}^3 \mid \|x\|_\infty < 5\} \setminus \mathbb{R}^3_\geq$ and the magnetic field is chosen as the constant function $B^{3\mathrm{D}} \equiv \frac{1}{\sqrt{3}}[1,1,1]^\mathsf{T}$ which is represented by the magnetic vector potential $A^{3\mathrm{D}}(x) := \frac{1}{2}B^{3\mathrm{D}} \times x$. The domain was discretized using *Gmsh* [67] with 72166 points. The initial guess is chosen as the constant function $\psi_0^{3\mathrm{D}} \equiv 1$. After 22 iterations of Newton’s method, the residual norm drops below 10^{-10}. The history of the Newton residual norms and the final approximate solution $\psi^{3\mathrm{D}} := z_{22}$ are again visualized in figures 4.1 and 4.2.

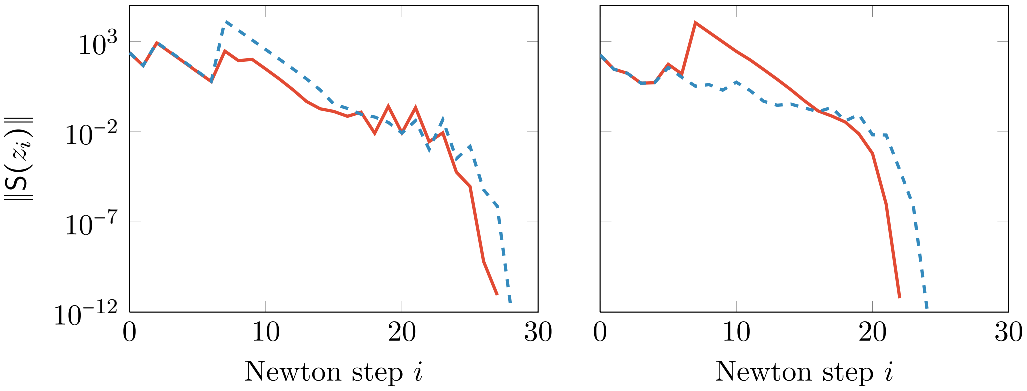
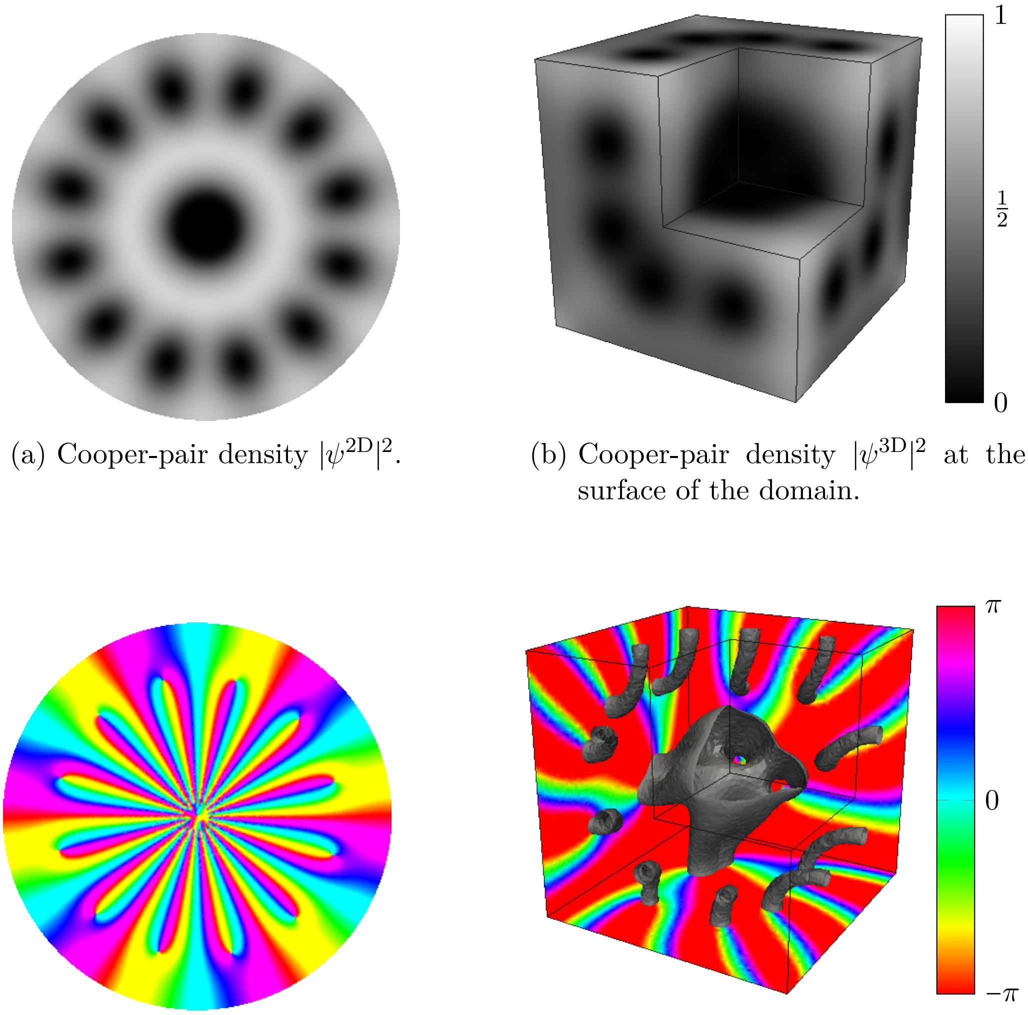


Figure 4.1.: Newton residual history for the two-dimensional setup 4.2 (left) and three-dimensional setup 4.3 (right), each with bordering (- - -) and without (——). With initial guesses $\psi_0^{2\mathrm{D}}(x,y) = \cos(\pi x)$ and $\psi_0^{3\mathrm{D}} \equiv 1$, respectively, the Newton process delivered the solutions as highlighted in figure 4.2 in 22 and 27 steps, respectively.

All experimental results presented in this section can be reproduced from the data published with the free and open source Python packages *PyNosh* [64] and *KryPy* [60]. *PyNosh* provides solvers for nonlinear Schrödinger equations as outlined in section 4.1 and the above test setups 4.2 and 4.3 for the Ginzburg–Landau equation.

Since the Jacobian operators (4.8) are self-adjoint with respect to the inner product (4.9), the MINRES method can be used. As proposed by Schlömer and Vanroose [149], a self-adjoint and positive-definite preconditioner is used that is based on an algebraic multigrid solver (AMG) and is able to reduce the number of MINRES iterations dramatically. Let $\mathrm{AMG}_k(\mathbf{B})$ denote the matrix that applies k cycles of an AMG scheme with a Hermitian and positive-definite matrix $\mathbf{B}$, i.e., $\mathrm{AMG}_k(\mathbf{B}) \approx \mathbf{B}^{-1}$. A review of applicable AMG methods can be found in the article of Stüben [168]. Here, the smoothed aggregation AMG solver with one pre- and one post-smoothing step with symmetric Gauss–Seidel from the Python package *PyAMG* [9] is used. The matrix $\mathrm{AMG}_k(\mathbf{B})$ then also is Hermitian and positive definite. In the setting of the Ginzburg–Landau equation, a preconditioner for the



(a) Cooper-pair density $|\psi^{2\text{D}}|^2$.

(b) Cooper-pair density $|\psi^{3\text{D}}|^2$ at the surface of the domain.

(c) $\arg \psi^{2\text{D}}$.

(d) Isosurface with $|\psi^{3\text{D}}|^2 = 0.1$ (see (b)), $\arg \psi^{3\text{D}}$ at the back sides of the cube.

Figure 4.2.: Solutions of the two-dimensional setup 4.2 (left) and three-dimensional setup 4.3 (right) as found in the Newton process illustrated in figure 4.1.

Jacobian operator (4.8) can be constructed by approximately inverting the matrix

$$\mathbf{C}_v = \mathbf{D}^{-1}\mathbf{K} + 2|\mathbf{D}_v| = \mathbf{D}^{-1}(\mathbf{K} + 2\mathbf{D}|\mathbf{D}_v|)$$

which is self-adjoint with respect to the inner product (4.9) and also positive definite if $A \not\equiv 0$ or $\|\psi\| \neq 0$. Because $\mathbf{K} + 2\mathbf{D}|\mathbf{D}_v|$ is Hermitian and positive definite with respect to the Euclidean inner product, the preconditioner can then be defined as

$$\mathsf{M}_v := \mathrm{AMG}_k(\mathbf{K} + 2\mathbf{D}|\mathbf{D}_v|)\mathbf{D} \approx \mathbf{C}_v^{-1}$$

and it can be verified that M_v is self-adjoint and positive definite with respect to the inner product (4.9). More details on this preconditioner can be found in [149].

For a Newton step $i \in \mathbb{N}$, the preconditioned MINRES algorithm 2.7 is applied to the preconditioned linear system

$$\mathsf{M}_i\mathsf{J}_i\delta_i = \mathsf{M}_i b_i,$$

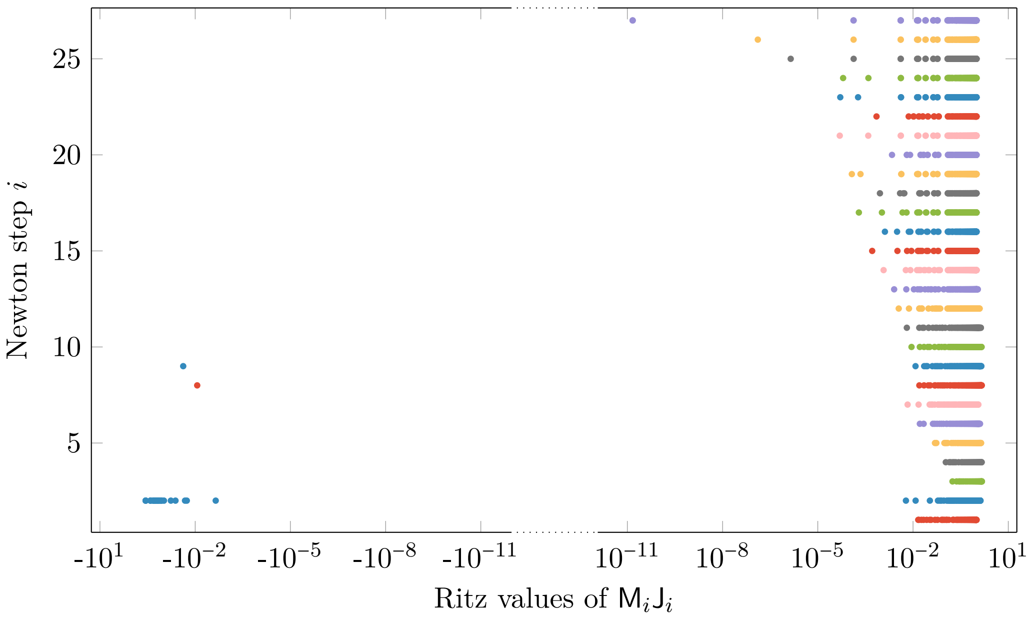
where $\mathsf{M}_i := \mathsf{M}_{z_i}$, $\mathsf{J}_i := \mathsf{J}_{z_i}$ and $b_i := -\mathsf{S}(z_i)$. Note that the preconditioner implicitly changes the inner product in the MINRES algorithm to $\langle\cdot,\cdot\rangle_{\mathsf{M}_i^{-1}}$ which is defined for $v, w \in \mathbb{C}^n$ by

$$\langle v, w\rangle_{\mathsf{M}_i^{-1}} = \left\langle v, \mathsf{M}_i^{-1}w\right\rangle_{\mathbb{R}}.$$

In figure 4.4, the convergence histories of MINRES with several recycling strategies are plotted for all Newton steps of the 2D setup (left) and the 3D setup (right). The tolerance for the relative residual norm is 10^{-10} throughout all Newton steps. In order to facilitate the notation, $\|r_n\|$ and $\|b\|$ denote the preconditioned residual and right hand side in the corresponding norm that is induced by the preconditioner. The underlying algorithm for all used recycling strategies in the experiments is algorithm 3.1.

No deflation. The convergence of standard MINRES is depicted in figure 4.4a. Towards the end of the Newton process, numerous phases of stagnation occur (dark red curves). In the last Newton step for the 2D setup, the residual norm ultimately stagnates above 10^{-9}. In order to get some insight into the spectrum of the preconditioned Jacobi operators $\mathsf{M}_i\mathsf{J}_i$, the Ritz values with respect to the constructed Krylov subspace at the end of each MINRES run are visualized in figure 4.3.

12 vectors. In a series of numerical experiments in [63, figure 3.4], it was analyzed how many Ritz vectors corresponding to the Ritz values of smallest magnitude lead to the minimal overall time consumption with deflated MINRES for each Newton step. In later Newton steps, the optimal number of deflation vectors was found to be around 12 in both the 2D and the 3D setup. Deflation does not pay off in early Newton steps due to the large changes of the Jacobian operator and the right hand side. Figure 4.4b shows the convergence history of MINRES if 12 deflation vectors are used throughout Newton's method. Except for the second Newton



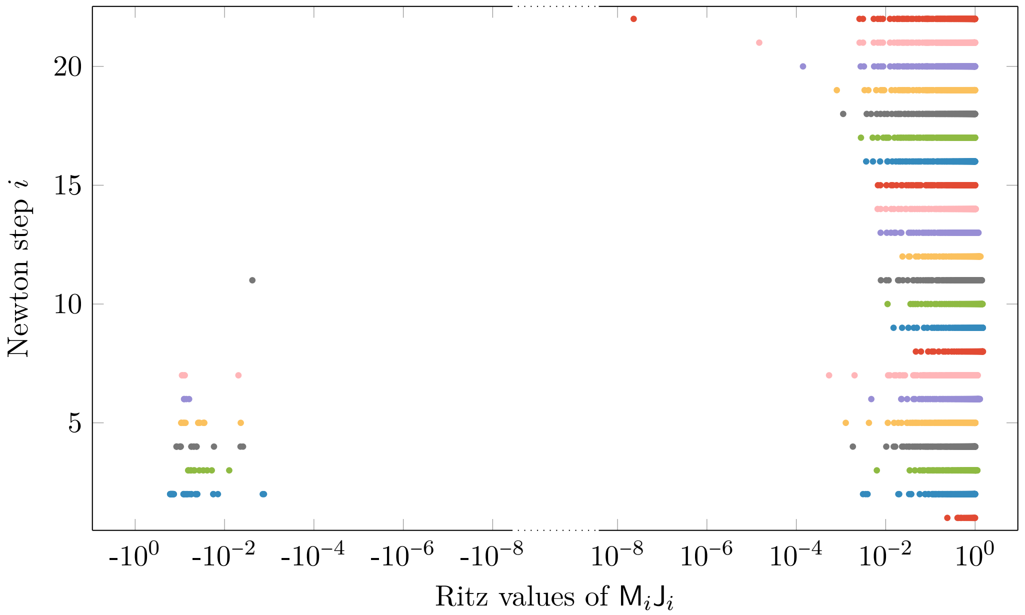
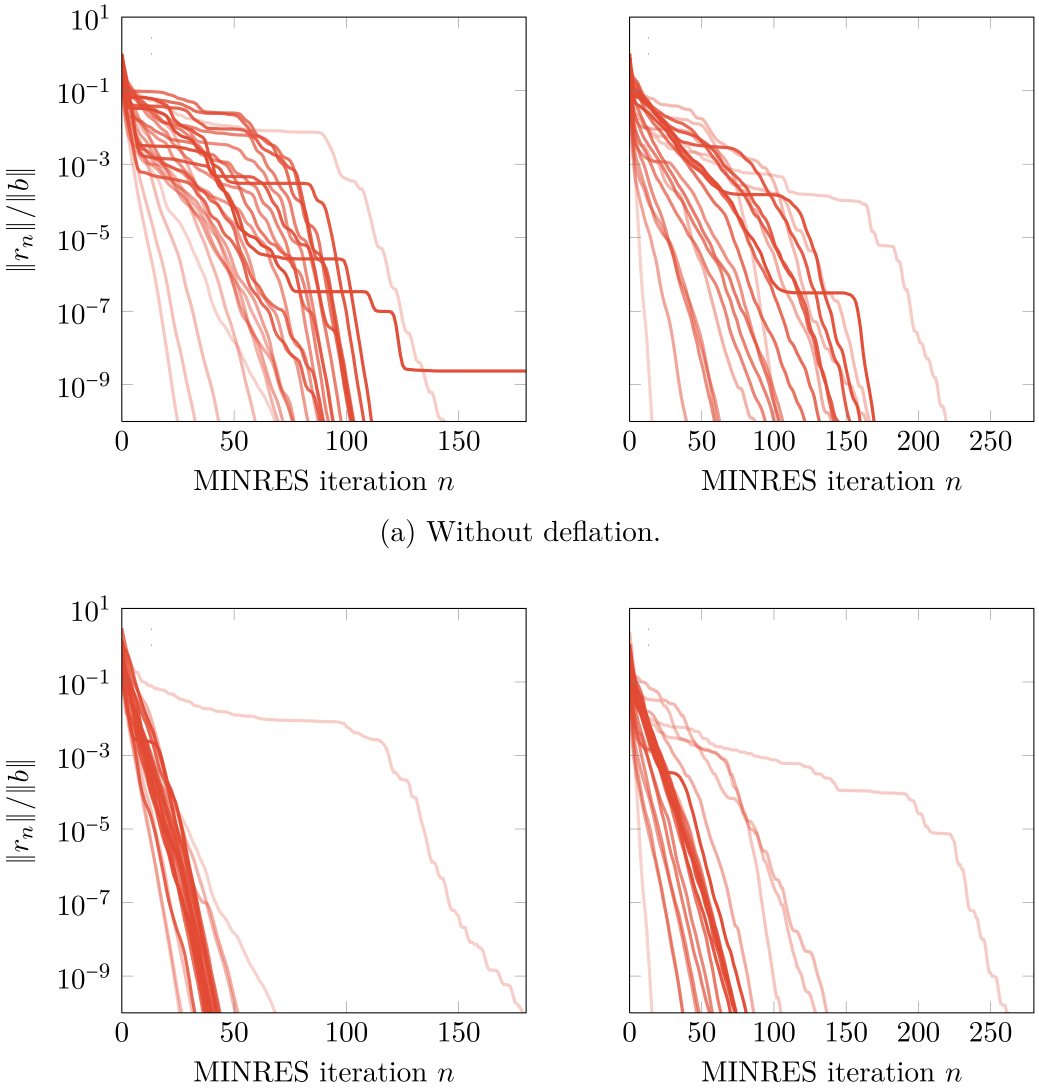
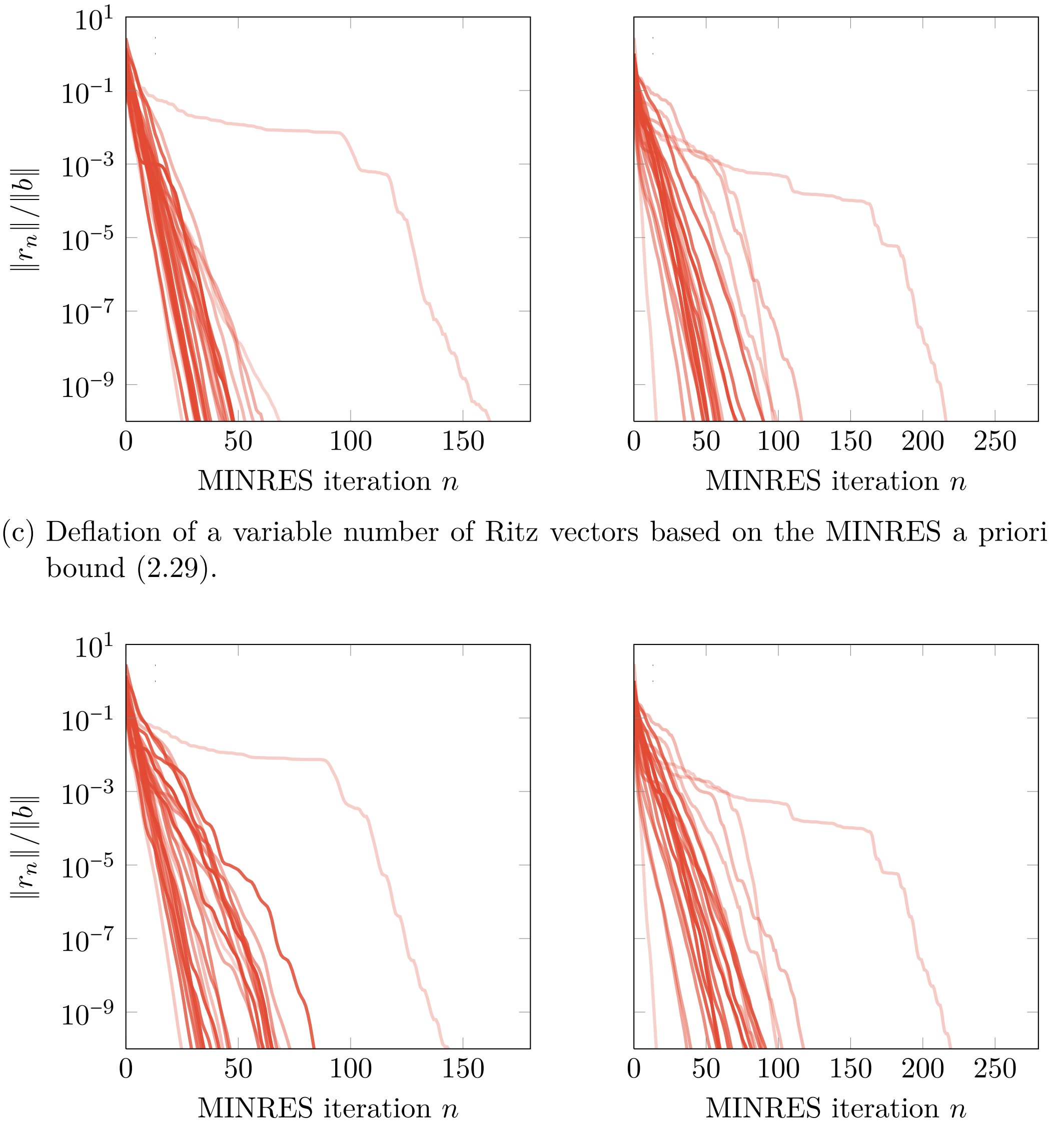


Figure 4.3.: Ritz values of $\mathsf{M}_i\mathsf{J}_i$ for the two-dimensional setup 4.2 (top) and the three-dimensional setup 4.3 (bottom) with respect to the constructed Krylov subspace after convergence of the MINRES method without deflation for all Newton steps.



(a) Without deflation.

(b) Deflation of 12 Ritz vectors corresponding to the Ritz values of smallest magnitude.



(c) Deflation of a variable number of Ritz vectors based on the MINRES a priori bound (2.29).

(d) Deflation of a variable number of Ritz vectors based on approximate Krylov subspaces.

Figure 4.4.: MINRES convergence histories for all Newton steps for the 2D setup 4.2 (left) and the 3D setup 4.3 (right). The shade of the curve corresponds to the Newton step: light red is the first Newton step while dark red is the last Newton step.

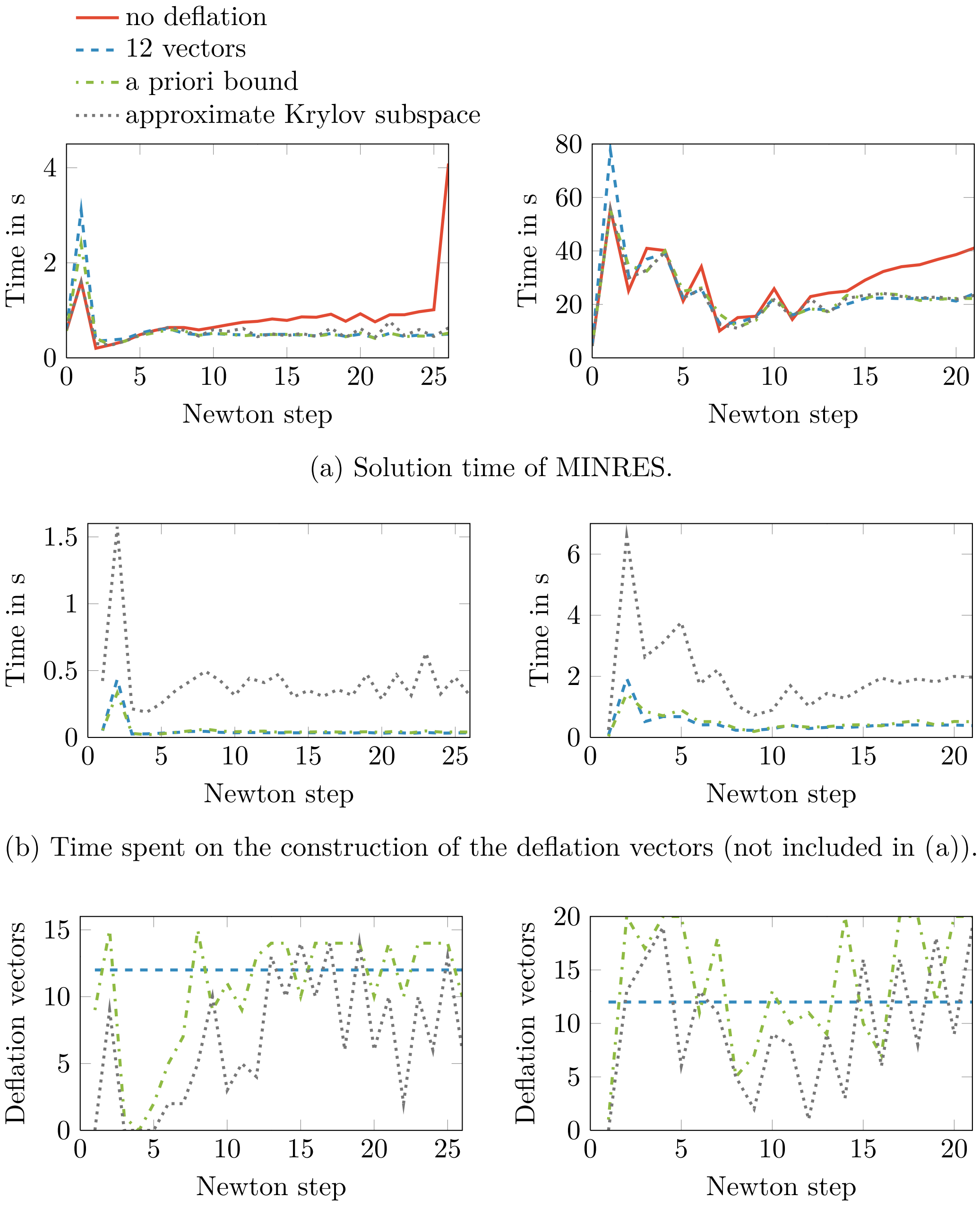
step, all phases of stagnation are removed. For most Newton steps, the number of MINRES iterations is reduced to 40 in the 2D setup and 80 or less in the 3D setup. Figure 4.5a compares the timings of the recycling MINRES strategies that are considered in this section. The fixed number of 12 deflation vectors leads to a significant reduction of the solution time beyond the 12th Newton step in both setups. In the last Newton step, the effective run time of MINRES is reduced by roughly 40%. However, no reduction is achieved for earlier Newton steps and the solution time even increases in the second Newton step. The second Newton step stands out in the 2D and 3D setup and also exhibits a special convergence behavior without deflation and the other recycling strategies that are discussed below. The distribution of Ritz values in figure 4.3 indicates a large number of negative and positive eigenvalues of the preconditioned Jacobian $\mathsf{M}_2\mathsf{J}_2$ in the second Newton step but this behavior is not fully understood.

The preceding strategy has two major drawbacks:

1. The determination of the number of deflation vectors requires user interaction.
2. The number of deflation vectors is fixed and is not dynamically adapted to the actual situation during the Newton process.

Instead of the fixed number of deflation vectors, the strategies from section 3.4 can be employed. Algorithm 3.2 provides a basic selection strategy for Ritz vectors that iteratively evaluates a cost function ω and picks the set of evaluated Ritz vectors with minimal cost. The algorithm first evaluates the empty set (no deflation vectors) and in each iteration, the Ritz vectors corresponding to the extremal Ritz values (smallest negative, largest negative, smallest negative and largest positive Ritz value) are evaluated and the one with minimal cost is added for all further evaluations. Here, the strategy is modified such that only the Ritz vectors corresponding to the Ritz value of smallest magnitude are considered and the maximum number of selected Ritz vectors is set to 15 for the 2D setup and 20 for the 3D setup. The strategy is implemented as `krypy.recycling.generators.RitzSmall` in [60].

In the following, two choices for the cost function ω are explored. Both cost functions first estimate the number m of MINRES iterations that are required until the relative residual norm falls below the prescribed tolerance 10^{-10}. Then the cost function uses timings in order to estimate the overall time that is required for m iterations of deflated MINRES. The timings are gathered on the fly in the preceding MINRES run and include the most expensive operations in the MINRES method, i.e., the application of the operator J_i, the preconditioner M_i, the evaluation of the inner product $\langle\cdot,\cdot\rangle_{\mathbb{R}}$ and the vector update (`daxpy`). It is important to note that the time estimation includes the time that is required to setup the projection for deflation as well as the time that is required to apply the projection in each step of the MINRES method. Also note that the use of timings may result in slightly different measurements and thus different decisions of the strategies in each run because the operating system's scheduler has to deal with varying constellations of



(a) Solution time of MINRES.

(b) Time spent on the construction of the deflation vectors (not included in (a)).

(c) Number of used deflation vectors.

Figure 4.5.: Timings and number of deflation vectors for the MINRES variants in figure 4.4.

concurrent processes in general. Although this renders the timing-based algorithms non-deterministic, the effect of different decisions has not been observed in experiments if the system load does not change dramatically throughout the different runs. When comparing results of the timing-based strategies, it should be kept in mind that the algorithms adapt to the performance of the specific machine and the implementations of underlying libraries such as NumPy, SciPy, LAPACK and BLAS.

A priori bound. In section 3.4.2, it was proposed to use an a priori bound like the MINRES bound (2.29) in order to assess a given set of Ritz vectors that are considered for deflation. Theorem 3.50 provides inclusion intervals for the nonzero eigenvalues of the deflated operator $\mathsf{P}_{\mathcal{W}^\perp,\mathsf{B}\mathcal{W}}\mathsf{B}$ based on Ritz pairs of the operator A and the difference $\mathsf{F} = \mathsf{B} - \mathsf{A}$. The inclusion intervals are defined as intervals around the Ritz values that correspond to the Ritz vectors that are not used for deflation. Unfortunately, in this particular application, either the stringent assumption (3.69) of theorem 3.50 on the spectral gap is not fulfilled (especially in the first Newton steps) or the resulting eigenvalue inclusion intervals contain zero which clearly renders the MINRES bound (2.29) useless. In practice, the Ritz values that correspond to the Ritz vectors that are not used for deflation can serve as a guideline for the spectrum of the deflated operator and can be used with the a priori MINRES bound (2.29) to estimate the number of required MINRES steps. This naive heuristic may provide too optimistic estimations. For example, if all but one Ritz vectors are chosen as deflation vectors, the remaining single Ritz value will indicate convergence in one step via the κ-bound (2.17) (which then also holds for the MINRES residual norm). Although the time estimation then includes the time that is needed for the application of J_i and M_i to all but one Ritz vectors, the resulting timings sometimes indicate to pick all Ritz vectors in experiments. For this reason, the use of deflation is penalized by multiplying the estimated time for the projection setup and application with a factor $\rho > 1$. Both the described simple strategy and the interval-based strategy with the bound from theorem 3.50 are implemented in [60] as `krypy.recycling.evaluators.RitzApriori`. Figure 4.4c shows the convergence histories of MINRES with the simple a priori bound strategy and $\rho = 2$. The convergence behavior is comparable to the one in figure 4.4b where 12 deflation vectors were used throughout the Newton process.

In figure 4.5c, the number of selected Ritz vectors for deflation is visualized for all Newton steps. In the 2D setup, the number grows to 15 after the second Newton step and then drops to zero before it starts to settle between 10 and 14 in the last 15 Newton steps. The alternating pattern between 10 and 14 is not fully understood but may be related to alternating locations of the smallest eigenvalues of the preconditioned Jacobian $\mathsf{M}_i\mathsf{J}_i$ which is indicated by the changes of the smallest Ritz values in figure 4.3.

Figure 4.5a reveals that the overall solution time with the a priori bound strategy is almost identical to the recycling strategy with the manually adjusted number of

12 deflation vectors. The computation time that is spent on the evaluation of the different selections of Ritz vectors is marginal and dominated by the explicit formation of the selected Ritz vectors from the Arnoldi basis and the previous deflation basis. Thus, the curves of the a priori strategy almost coincide with the static choice of 12 vectors in figure 4.5b.

Approximate Krylov subspaces. In section 3.4.3, an alternative to the recycling strategy based on a priori bounds was proposed. Let $\mathcal{K}_n$ denote the Krylov subspace that has been constructed by n iterations of deflated GMRES for the linear system $\mathsf{A}x = b$ with the m-dimensional deflation subspace $\mathcal{U}$. If a k-dimensional deflation subspace candidate $\mathcal{W} \subseteq \mathcal{K}_n + \mathcal{U}$ is given, then theorem 3.64 provides a bound on the residual norms $\|\widehat{r}_i\|$ of deflated GMRES for the linear system $\mathsf{B}y = c$ with deflation subspace $\mathcal{W}$ based on the data that has been computed by deflated GMRES for the linear system $\mathsf{A}x = b$. The bound (3.85) essentially consists of two terms. The first one is the residual norm $\|\tilde{r}_i\|_2$ that can be computed by GMRES for the small linear system (3.83) with a (square) upper Hessenberg matrix $\widehat{\mathbf{H}} \in \mathbb{C}^{n+m-k,n+m-k}$. The residual norms $\|\tilde{r}_i\|_2$ equal the residual norms of a perturbed linear system and can be cheaply computed from $\widehat{\mathbf{H}}$ because it already is an upper Hessenberg matrix and the right hand side is a multiple of e_1. The second term in (3.85) is based on the pseudospectrum of an operator $\widehat{\mathsf{A}}_i$ and the differences $\mathsf{G} = \mathsf{B} - \mathsf{A}$ and $g = c - b$. The pseudospectrum of $\widehat{\mathsf{A}}_i$ cannot be computed in practice but can be approximated by the pseudospectrum of $\widehat{\mathbf{H}}$, see the discussion following theorem 3.64. The matrix $\widehat{\mathbf{H}}$ is Hermitian in the situation of nonlinear Schrödinger equations because $\mathsf{M}_i\mathsf{J}_i$ is self-adjoint. Therefore, the δ-pseudospectrum $\Lambda_\delta(\widehat{\mathbf{H}})$ is easy to compute because it is the union of symmetric intervals of length 2δ around the eigenvalues of $\widehat{\mathbf{H}}$, i.e.,

$$\Lambda_\delta(\widehat{\mathbf{H}}) = \bigcup_{\lambda \in \Lambda(\widehat{\mathbf{H}})} [\lambda - \delta, \lambda + \delta].$$

Although the pseudospectrum can be obtained cheaply in this case, the second term in the bound (3.85) has to be evaluated for a wide range of the parameter δ because the bound holds for all $\delta > \epsilon_i$ and the optimal δ is usually not known. For each δ the maximum of a polynomial $p_i \in \mathbb{P}_{n,0}$ with known roots has to be computed on the pseudospectrum. Besides being exceedingly expensive, the second term in (3.85) often rises above the first term $\|\tilde{r}_i\|_2$ although the actual convergence behavior is still captured by $\|\tilde{r}_i\|_2$; see section 3.5 for a discussion. For the above reasons, the strategy for the experiments in this section only determines the first iteration i such that $\|\tilde{r}_i\|_2 < 10^{-10}$ and then estimates the expected time as described above. Note that the exact solution of the linear system (3.83) is found after at most $n + m - k$ steps because $\widehat{\mathbf{H}} \in \mathbb{C}^{n+m-k,n+m-k}$. Note that if all but j Ritz vectors are considered for deflation then $\tilde{r}_j = 0$. Therefore, the use of deflation is penalized as in the case of the a priori bound with the penalty factor $\rho = 2$. It can be seen in figure 4.5c that fewer vectors are chosen with this strategy than with the a priori strategy in almost all Newton steps. The convergence histories in figure 4.4d record slightly more MINRES iterations in some Newton steps but the timings in figure 4.5a show that the time consumption is similar to the a priori based strategy and the static choice of 12 vectors. The time that is spent on the construction of the deflation vectors is larger for the approximate Krylov subspace strategy, see figure 4.5b. However, it should be noted that the approximate Krylov subspace strategy only operates with "small" matrices and no applications of the operator $\mathsf{M}_i\mathsf{J}_i$ or evaluations of the inner product $\langle\cdot,\cdot\rangle_{\mathbb{R}}$ are necessary. Thus the time for the selection of Ritz vectors becomes negligible if the application of $\mathsf{M}_i\mathsf{J}_i$ or the evaluation of inner products consume more time while the number of MINRES iterations stays roughly the same. This can be observed for the 3D setup in figure 4.5a with figure 4.5b.

Summarizing, the proposed recycling strategies based on a priori bounds and approximate Krylov subspaces are able to achieve roughly the same performance in terms of the overall consumed time as the manually determined optimal choice of 12 vectors, cf. figure 4.5a. The bounds from sections 3.4.2 and 3.4.3 result in too pessimistic estimates and therefore trimmed variants have been used for the experiments in this section.

5. Discussion and outlook

In this thesis, several contributions have been made to the mathematical theory as well as the practical realization of recycling Krylov subspace methods that are based on deflation. This concluding discussion gives a very brief summary of the main contributions and points out possibilities for further investigations that arise from the analysis and experiments in this thesis.

By showing equivalences between several deflated and augmented methods in section 3.2 and characterizing the well-definedness of these methods, the author wishes to draw attention to the *mathematical* analysis of deflated methods – away from the urge to introduce "new" methods that often appear as complicated *algorithms* but are essentially equivalent to already existing ones. Of course, this does not apply to algorithmic contributions that are superior, e.g., with respect to their behavior in the presence of round-off errors. Furthermore, the results show that well-defined deflated methods can be constructed without hassle by using the most robust of the already existing implementations of Krylov subspace methods.

Due to their generality, the provided perturbation results for projections in section 3.3.1 may prove to be useful beyond deflated Krylov subspace methods. For example, the results can be applied in infinite-dimensional Hilbert spaces.

In the context of eigenvalue-based a priori bounds for CG and MINRES, the quadratic residual bound for all eigenvalues of the deflated operator in theorems 3.36 and 3.50 can provide significant insight, cf. example 3.52. In the particular application to the Ginzburg–Landau equation in section 4.3, either the tough assumptions of the theorem are not satisfied or the eigenvalue inclusion intervals contain zero which makes a meaningful automatic selection of Ritz vectors impossible. Therefore, a simplified selection strategy is employed in section 4.3 that uses a priori bounds in conjunction with Ritz values alone. For the Ginzburg–Landau equation, the strategy is able to reduce the overall computation time like the manually determined optimal number of deflation vectors. In future works, an unexplored path can be taken by using a priori bounds for Krylov subspace methods with the new characterization of the full spectrum of the deflated operator in theorem 3.24.

The novel approach from section 3.4.3 of using an approximate Krylov subspace in order to estimate the convergence behavior yields positive results for MINRES and, more interestingly, for GMRES with a non-normal operator. Unlike asymptotic bounds, the approximate Krylov subspace bound is able to capture an altering convergence behavior accurately in example 3.72 and the convection-diffusion problem in section 3.5. The experiments reveal that the first term of bound (3.85) describes the convergence behavior well beyond the point where the pseudospectral term causes the bound to grow. The pseudospectral approach seems to severely overestimate the effects of perturbations and it remains to be explored if another approach can do better. However, note that example 3.38 shows that a small perturbation of the operator or the right hand side may actually result in a drastic change of the convergence behavior of Krylov subspace methods. Because the pseudospectral term is exceedingly expensive and often overestimates the residual norms, only the cheaply computable residual norms from the approximate Krylov subspace are used in the experiments in section 4.3 for the automatic selection of Ritz vectors for recycling. Analogous to the strategy based on a priori bounds, also the approximate Krylov subspace strategy yields the same overall time consumption as with the manually determined optimal number of deflation vectors.

For practical applications, the automatic selection of deflation vectors based on the proposed bounds is attractive. Given candidates of deflation vectors, the proposed recycling strategies pick the choice that yields the best overall time estimation according to an a priori bound or the approximate Krylov subspace bound. In order to estimate the time accurately, the implementation keeps track of the timings for the most expensive operations. In prior algorithms in the literature, the number of used deflation vectors had to be specified manually. Besides requiring the user to specify the number of vectors, a fixed number of deflation vectors may also lead to an increased overall computation time for a poor choice of vectors. For the experiments in this thesis, the Ritz vectors corresponding to the Ritz values of smallest magnitude performed best as deflation vectors and were used with the proposed strategies. Harmonic Ritz vectors did not lead to a better performance. However, other choices may perform better and the challenging question of the optimal choice of deflation vectors from a given subspace remains open. General statements about the optimal choice of deflation vectors seem unrealistic without a better understanding of the convergence behavior of the underlying Krylov subspace methods. For a sequence of linear systems, the situation is even more complicated because the convergence behavior cannot be deduced with the help of simple measures such as the norm of the difference between subsequent operators and right hand sides. As explained in section 3.4, a heuristic should be expected at some point in order to obtain an efficient method. The results in section 4.3 show that the developed strategies drive the existing heuristics forward and result in efficient methods that require no manually tuned parameters.

Besides the mathematical results, the author wishes to make recycling Krylov subspace methods accessible to a broader audience by providing the free software package KryPy [60] which allows to use a recycling solver with a few lines of code, see appendix A.

A. Documentation of the KryPy software package

The KryPy [60] software package is free software and contains implementations of standard and deflated versions of CG, MINRES and GMRES as well as all recycling strategies that are discussed in this thesis. All methods are very flexible, e.g., by allowing to use non-Euclidean inner products and several orthogonalization strategies like Lanczos, (iterated) Gram–Schmidt or Householder orthogonalization. Furthermore, KryPy contains a rich toolbox for Krylov subspace methods such as methods for the computation of Ritz pairs, angles between subspaces, convergence bounds and stable implementations of projections that are of importance in deflated methods. This chapter provides minimalistic code snippets in order to get started with KryPy. The full documentation of KryPy can be found at `http://krypy.readthedocs.org`.

Requirements. Python (2.7 or ≥3.2) and the modules NumPy (≥1.7) and SciPy (≥0.12) are required to run KryPy. Optionally, PseudoPy (≥1.2.1) enables the computation of the pseudospectral bounds from this thesis.

Installation. KryPy can be installed easily with the Python package installer by issuing `pip install krypy`. Alternatively, it can be installed by downloading the source code from `https://github.com/andrenarchy/krypy` and then running `python setup.py install`.

First steps. The following code uses MINRES to solve a linear system with an indefinite diagonal matrix:

```
from numpy import diag, linspace, ones, eye
from krypy.linsys import LinearSystem, Minres

# construct the linear system
A = diag(linspace(1, 2, 20))
A[0, 0] = -1e-5
b = ones(20)
linear_system = LinearSystem(A, b, self_adjoint=True)

# solve the linear system (approximate solution is solver.xk)
solver = Minres(linear_system)
```

Deflation. The vector e_1 can be used as a deflation vector to get rid of the small negative eigenvalue -10^{-5}:

```
from krypy.deflation import DeflatedMinres
dsolver = DeflatedMinres(linear_system, U=eye(20, 1))
```

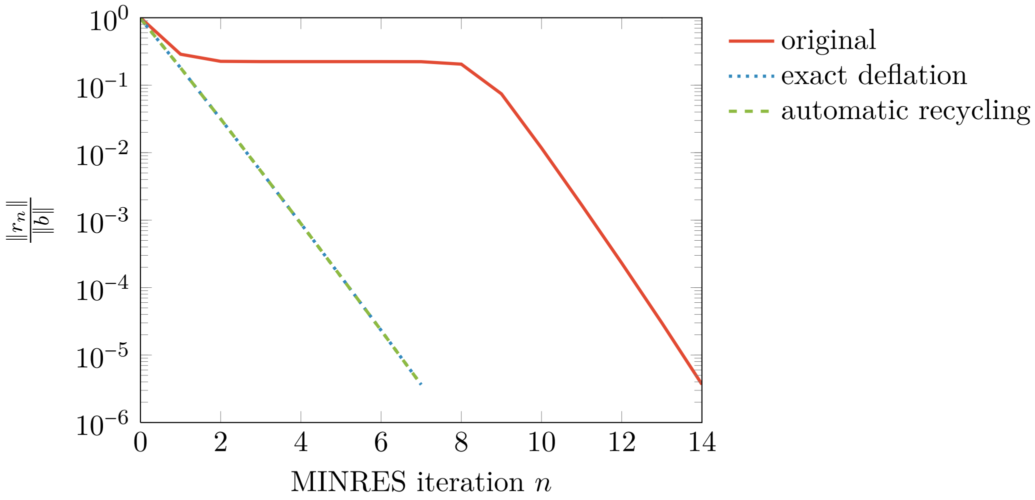
Recycling. The deflation subspace can also be determined automatically with a recycling strategy. Just for illustration, the same linear system is solved twice in the following code:

```
from krypy.recycling import RecyclingMinres

# get recycling solver with approximate Krylov subspace strategy
rminres = RecyclingMinres(vector_factory='RitzApproxKrylov')

# solve twice
rsolver1 = rminres.solve(linear_system)
rsolver2 = rminres.solve(linear_system)
```

The convergence histories can be plotted by

```
from matplotlib.pyplot import semilogy, show, legend
semilogy(solver.resnorms, label='original')
semilogy(dsolver.resnorms, label='exact␣deflation', ls='dotted')
semilogy(rsolver2.resnorms, label='automatic␣recycling',
         ls='dashed')
legend()
show()
```

which results in the following figure.



Bibliography

[1] P. Amestoy, A. Buttari, A. Guermouche, J.-Y. L'Excellent, and B. Ucar. *MUMPS: a MUltifrontal Massively Parallel sparse direct Solver*. URL: `http://mumps.enseeiht.fr/`.

[2] M. Arioli and C. Fassino. "Roundoff error analysis of algorithms based on Krylov subspace methods". In: *BIT* 36.2 (1996), pp. 189–205.

[3] M. Arioli, D. Loghin, and A. J. Wathen. "Stopping criteria for iterations in finite element methods". In: *Numer. Math.* 99.3 (2005), pp. 381–410.

[4] M. Arioli, E. Noulard, and A. Russo. "Stopping criteria for iterative methods: applications to PDE's". In: *Calcolo* 38.2 (2001), pp. 97–112.

[5] M. Arioli, V. Pták, and Z. Strakoš. "Krylov sequences of maximal length and convergence of GMRES". In: *BIT* 38.4 (1998), pp. 636–643.

[6] W. E. Arnoldi. "The principle of minimized iteration in the solution of the matrix eigenvalue problem". In: *Quart. Appl. Math.* 9 (1951), pp. 17–29.

[7] O. Axelsson. "A class of iterative methods for finite element equations". In: *Comput. Methods Appl. Mech. Engrg.* 9.2 (1976), pp. 123–127.

[8] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds. *Templates for the solution of algebraic eigenvalue problems*. Vol. 11. Software, Environments, and Tools. A practical guide. Philadelphia, PA: SIAM, 2000, pp. xxx+410.

[9] W. N. Bell, L. N. Olson, and J. B. Schroder. *PyAMG: Algebraic Multigrid Solvers in Python v2.0*. Release 2.0. 2011. URL: `http://www.pyamg.org`.

[10] M. Benzi. "Preconditioning techniques for large linear systems: a survey". In: *J. Comput. Phys.* 182.2 (2002), pp. 418–477.

[11] M. Benzi and D. Bertaccini. "Approximate inverse preconditioning for shifted linear systems". In: *BIT* 43.2 (2003), pp. 231–244.

[12] M. Benzi and M. A. Olshanskii. "Field-of-values convergence analysis of augmented Lagrangian preconditioners for the linearized Navier-Stokes problem". In: *SIAM J. Numer. Anal.* 49.2 (2011), pp. 770–788.

[13] E. Berkson. "Some metrics on the subspaces of a Banach space". In: *Pacific J. Math.* 13 (1963), pp. 7–22.

[14] D. Bertaccini. "Efficient preconditioning for sequences of parametric complex symmetric linear systems". In: *Electron. Trans. Numer. Anal.* 18 (2004), pp. 49–64.

[15] R. Bhatia. *Perturbation bounds for matrix eigenvalues.* Vol. 53. Classics in Applied Mathematics. Reprint of the 1987 original. Philadelphia, PA: SIAM, 2007, pp. xvi+191.

[16] P. Birken, J. Duintjer Tebbens, A. Meister, and M. Tůma. “Preconditioner updates applied to CFD model problems”. In: *Appl. Numer. Math.* 58.11 (2008), pp. 1628–1641.

[17] P. N. Brown. “A theoretical comparison of the Arnoldi and GMRES algorithms”. In: *SIAM J. Sci. Statist. Comput.* 12.1 (1991), pp. 58–78.

[18] P. N. Brown and H. F. Walker. “GMRES on (nearly) singular systems”. In: *SIAM J. Matrix Anal. Appl.* 18.1 (1997), pp. 37–51.

[19] D. Buckholtz. “Hilbert space idempotents and involutions”. In: *Proc. Amer. Math. Soc.* 128 (2000), pp. 1415–1418.

[20] D. Buckholtz. “Inverting the difference of Hilbert space projections”. In: *Amer. Math. Monthly* 104.1 (1997), pp. 60–61.

[21] Z.-h. Cao, J.-J. Xie, and R.-C. Li. “A sharp version of Kahan’s theorem on clustered eigenvalues”. In: *Linear Algebra Appl.* 245 (1996), pp. 147–155.

[22] J.-F. Carpraux, S. K. Godunov, and S. V. Kuznetsov. “Condition number of the Krylov bases and subspaces”. In: *Linear Algebra Appl.* 248 (1996), pp. 137–160.

[23] A. Chapman and Y. Saad. “Deflated and augmented Krylov subspace techniques”. In: *Numer. Linear Algebra Appl.* 4.1 (1997), pp. 43–66.

[24] J. W. Daniel. “The conjugate gradient method for linear and nonlinear operator equations”. In: *SIAM J. Numer. Anal.* 4 (1967), pp. 10–26.

[25] D. Darnell, R. B. Morgan, and W. Wilcox. “Deflated GMRES for systems with multiple shifts and multiple right-hand sides”. In: *Linear Algebra Appl.* 429.10 (2008), pp. 2415–2434.

[26] C. Davis and W. M. Kahan. “The rotation of eigenvectors by a perturbation. III”. In: *SIAM J. Numer. Anal.* 7 (1970), pp. 1–46.

[27] C. Davis, W. M. Kahan, and H. F. Weinberger. “Norm-preserving dilations and their applications to optimal error bounds”. In: *SIAM J. Numer. Anal.* 19.3 (1982), pp. 445–469.

[28] D. W. Decker, H. B. Keller, and C. T. Kelley. “Convergence rates for Newton’s method at singular points”. In: *SIAM J. Numer. Anal.* 20.2 (1983), pp. 296–314.

[29] D. W. Decker and C. T. Kelley. “Newton’s method at singular points. I”. In: *SIAM J. Numer. Anal.* 17.1 (1980), pp. 66–70.

[30] D. Del Pasqua. “Su una nozione di varietà lineari disgiunte di uno spazio di Banach”. In: *Rend. Mat. e Appl. (5)* 13 (1955), pp. 406–422.

[31] J. W. Demmel. *Applied numerical linear algebra.* Philadelphia, PA: SIAM, 1997, pp. xii+419.

[32] P. Deuflhard. *Newton methods for nonlinear problems.* Vol. 35. Springer Series in Computational Mathematics. Affine invariance and adaptive algorithms, First softcover printing of the 2006 corrected printing. Berlin: Springer-Verlag, 2011, pp. xii+424.

[33] G. Dirr, V. Rakočević, and H. K. Wimmer. "Estimates for projections in Banach spaces and existence of direct complements". In: *Studia Math.* 170.2 (2005), pp. 211–216.

[34] Z. Dostál. "Conjugate gradient method with preconditioning by projector". In: *Int. J. Comput. Math.* 23.3 (1988), pp. 315–323.

[35] M. P. Drazin. "Pseudo-inverses in associative rings and semigroups". In: *Amer. Math. Monthly* 65 (1958), pp. 506–514.

[36] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš. "Numerical stability of GMRES". In: *BIT* 35.3 (1995), pp. 309–330.

[37] Q. Du, M. D. Gunzburger, and J. S. Peterson. "Modeling and analysis of a periodic Ginzburg-Landau model for type-II superconductors". In: *SIAM J. Appl. Math.* 53.3 (1993), pp. 689–717.

[38] J. Duintjer Tebbens and M. Tůma. "Efficient preconditioning of sequences of nonsymmetric linear systems". In: *SIAM J. Sci. Comput.* 29.5 (2007), pp. 1918–1941.

[39] J. Duintjer Tebbens and M. Tůma. "Improving triangular preconditioner updates for nonsymmetric linear systems". In: *Large-scale scientific computing.* Vol. 4818. Lecture Notes in Comput. Sci. Berlin: Springer-Verlag, 2008, pp. 737–744.

[40] J. Duintjer Tebbens and M. Tůma. "Preconditioner updates for solving sequences of linear systems in matrix-free environment". In: *Numer. Linear Algebra Appl.* 17.6 (2010), pp. 997–1019.

[41] M. Eiermann. "Fields of values and iterative methods". In: *Linear Algebra Appl.* 180 (1993), pp. 167–197.

[42] M. Eiermann and O. G. Ernst. "Geometric aspects of the theory of Krylov subspace methods". In: *Acta Numer.* 10 (2001), pp. 251–312.

[43] M. Eiermann, O. G. Ernst, and O. Schneider. "Analysis of acceleration strategies for restarted minimal residual methods". In: *J. Comput. Appl. Math.* 123.1-2 (2000), pp. 261–292.

[44] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. "Variational iterative methods for nonsymmetric systems of linear equations". In: *SIAM J. Numer. Anal.* 20.2 (1983), pp. 345–357.

[45] H. C. Elman. "Iterative methods for large, sparse, nonsymmetric systems of linear equations". PhD thesis. Yale University, 1982.

[46] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics.* Numerical Mathematics and Scientific Computation. New York: Oxford University Press, 2005, pp. xiv+400.

[47] J. Erhel, K. Burrage, and B. Pohl. "Restarted GMRES preconditioned by deflation". In: *J. Comput. Appl. Math.* 69.2 (1996), pp. 303–318.

[48] J. Erhel and F. Guyomarc'h. "An augmented conjugate gradient method for solving consecutive symmetric positive definite linear systems". In: *SIAM J. Matrix Anal. Appl.* 21.4 (2000), pp. 1279–1299.

[49] J. Erhel and F. Guyomarc'h. *An Augmented Subspace Conjugate Gradient.* Research report RR-3278. INRIA, 1997.

[50] Y. A. Erlangga and R. Nabben. "Deflation and balancing preconditioners for Krylov subspace methods applied to nonsymmetric matrices". In: *SIAM J. Matrix Anal. Appl.* 30.2 (2008), pp. 684–699.

[51] Y. A. Erlangga and R. Nabben. "Multilevel projection-based nested Krylov iteration for boundary value problems". In: *SIAM J. Sci. Comput.* 30.3 (2008), pp. 1572–1595.

[52] O. G. Ernst. "Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations". In: *SIAM J. Matrix Anal. Appl.* 21.4 (2000), pp. 1079–1101.

[53] E. Feleqi. "Spectral stability estimates for the eigenfunctions of second order elliptic operators". PhD thesis. Università degli Studi di Padova, 2009.

[54] B. Fischer. *Polynomial based iteration methods for symmetric linear systems.* Vol. 68. Classics in Applied Mathematics. Reprint of the 1996 edition. Philadelphia, PA: SIAM, 2011. 283 pp.

[55] P. F. Fischer. "Projection techniques for iterative solution of $A\underline{x} = \underline{b}$ with successive right-hand sides". In: *Comput. Methods Appl. Mech. Engrg.* 163.1-4 (1998), pp. 193–204.

[56] J. Frank and C. Vuik. "On the construction of deflation-based preconditioners". In: *SIAM J. Sci. Comput.* 23.2 (2001), pp. 442–462.

[57] R. W. Freund. "Quasi-kernel polynomials and their use in non-Hermitian matrix iterations". In: *J. Comput. Appl. Math.* 43.1-2 (1992). Orthogonal polynomials and numerical methods, pp. 135–158.

[58] R. W. Freund, G. H. Golub, and N. M. Nachtigal. "Iterative solution of linear systems". In: *Acta numerica, 1992.* Acta Numer. Cambridge: Cambridge Univ. Press, 1992, pp. 57–100.

[59] A. Galántai. *Projectors and projection methods.* Vol. 6. Advances in Mathematics (Dordrecht). Boston, MA: Kluwer Academic Publishers, 2004, pp. x+287.

[60] A. Gaul. *KryPy: Krylov subspace methods package for Python.* 2013–. URL: `https://github.com/andrenarchy/krypy`.

[61] A. Gaul. *PseudoPy: Computation and visualization of pseudospectra in Python.* Jan. 2014. URL: `https://github.com/andrenarchy/pseudopy`.

[62] A. Gaul, M. H. Gutknecht, J. Liesen, and R. Nabben. "A framework for deflated and augmented Krylov subspace methods". In: *SIAM J. Matrix Anal. Appl.* 34.2 (2013), pp. 495–518.

[63] A. Gaul and N. Schlömer. *Preconditioned recycling Krylov subspace methods for self-adjoint problems.* Preprint. arXiv:1208.0264v2. 2013.

[64] A. Gaul and N. Schlömer. *pynosh: Python framework for nonlinear Schrödinger equations.* July 2013. URL: `https://bitbucket.org/nschloe/pynosh`.

[65] M. Gedalin, T. Scott, and Y. Band. "Optical solitary waves in the higher order nonlinear Schrödinger equation". In: *Phys. Rev. Lett.* 78.3 (1997), pp. 448–451.

[66] T. Gergelits and Z. Strakoš. "Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations". In: *Numer. Algorithms* 65.4 (2014), pp. 759–782.

[67] C. Geuzaine and J.-F. Remacle. "Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities". In: *Internat. J. Numer. Methods Engrg.* 79.11 (2009), pp. 1309–1331.

[68] S. Gillies et al. *Shapely: Manipulation and analysis of geometric objects in the Cartesian plane.* URL: `http://toblerity.org/shapely/`.

[69] L. Giraud and J. Langou. "When modified Gram-Schmidt generates a well-conditioned set of vectors". In: *IMA J. Numer. Anal.* 22.4 (2002), pp. 521–528.

[70] L. Giraud, J. Langou, and M. Rozloznik. "The loss of orthogonality in the Gram-Schmidt orthogonalization process". In: *Comput. Math. Appl.* 50.7 (2005), pp. 1069–1075.

[71] G. H. Golub and C. F. Van Loan. *Matrix computations.* Fourth. Johns Hopkins Studies in the Mathematical Sciences. Baltimore, MD: Johns Hopkins University Press, 2013, pp. xiv+756.

[72] A. Greenbaum, M. Rozložník, and Z. Strakoš. "Numerical behaviour of the modified Gram-Schmidt GMRES implementation". In: *BIT* 37.3 (1997). Direct methods, linear algebra in optimization, iterative methods (Toulouse, 1995/1996), pp. 706–719.

[73] A. Greenbaum. *Iterative methods for solving linear systems.* Vol. 17. Frontiers in Applied Mathematics. Philadelphia, PA: SIAM, 1997, pp. xiv+220.

[74] A. Greenbaum, V. Pták, and Z. Strakoš. “Any nonincreasing convergence curve is possible for GMRES”. In: *SIAM J. Matrix Anal. Appl.* 17.3 (1996), pp. 465–469.

[75] A. Greenbaum and L. N. Trefethen. “GMRES/CR and Arnoldi/Lanczos as matrix approximation problems”. In: *SIAM J. Sci. Comput.* 15.2 (1994). Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992), pp. 359–368.

[76] A. Griewank. “On solving nonlinear equations with simple singularities or nearly singular solutions”. In: *SIAM Rev.* 27.4 (1985), pp. 537–563.

[77] A. Günnel, R. Herzog, and E. Sachs. “A note on preconditioners and scalar products in Krylov Subspace methods for self-adjoint problems in Hilbert space”. In: *Electron. Trans. Numer. Anal.* 41 (2014), pages.

[78] W. Hackbusch. *Multigrid methods and applications.* Vol. 4. Springer Series in Computational Mathematics. Berlin: Springer-Verlag, 1985, pp. xiv+377.

[79] E. V. Haynsworth. “Determination of the inertia of a partitioned Hermitian matrix”. In: *Linear Algebra and Appl.* 1.1 (1968), pp. 73–81.

[80] M. R. Hestenes and E. Stiefel. “Methods of conjugate gradients for solving linear systems”. In: *J. Research Nat. Bur. Standards* 49 (1952), pp. 409–436.

[81] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing In Science and Engineering* 9.3 (2007), pp. 90–95.

[82] I. C. F. Ipsen and C. D. Meyer. “The idea behind Krylov methods”. In: *Amer. Math. Monthly* 105.10 (1998), pp. 889–899.

[83] A. Jennings. “Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method”. In: *J. Inst. Math. Appl.* 20.1 (1977), pp. 61–72.

[84] Z. Jia. “The convergence of Krylov subspace methods for large unsymmetric linear systems”. In: *Acta Math. Sinica (N.S.)* 14.4 (1998), pp. 507–518.

[85] E. Jones, T. Oliphant, P. Peterson, et al. *SciPy: Open source scientific tools for Python.* 2001–. URL: `http://www.scipy.org/`.

[86] A. Jujunashvili. “Angles between infinite dimensional subspaces”. PhD thesis. University of Colorado Denver, 2005.

[87] W. M. Kahan. *Inclusion theorems for clusters of eigenvalues of Hermitian matrices.* Univ. of Toronto: Department of Computer Science, 1967.

[88] K. Kahl and H. Rittich. *Analysis of the Deflated Conjugate Gradient Method Based on Symmetric Multigrid Theory.* Preprint. arXiv:1209.1963v1. 2012.

[89] T. Kato. “Estimation of iterated matrices, with application to the von Neumann condition”. In: *Numer. Math.* 2 (1960), pp. 22–29.

[90] T. Kato. *Perturbation theory for linear operators.* Classics in Mathematics. Reprint of the 1980 edition. Berlin: Springer-Verlag, 1995, pp. xxii+619.

[91] H. B. Keller. “The bordering algorithm and path following near singular points of higher nullity”. In: *SIAM J. Sci. Statist. Comput.* 4.4 (1983), pp. 573–582.

[92] C. T. Kelley. *Solving nonlinear equations with Newton’s method.* Vol. 1. Fundamentals of Algorithms. Philadelphia, PA: SIAM, 2003, pp. xiv+104.

[93] S. A. Kharchenko and A. Y. Yeremin. “Eigenvalue translation based preconditioners for the GMRES(k) method”. In: *Numer. Linear Algebra Appl.* 2.1 (1995), pp. 51–77.

[94] M. E. Kilmer and E. de Sturler. “Recycling subspace information for diffuse optical tomography”. In: *SIAM J. Sci. Comput.* 27.6 (2006), pp. 2140–2166.

[95] R. C. Kirby. “From functional analysis to iterative methods”. In: *SIAM Rev.* 52.2 (2010), pp. 269–293.

[96] A. Klawonn and G. Starke. “Block triangular preconditioners for nonsymmetric saddle point problems: field-of-values analysis”. In: *Numer. Math.* 81.4 (1999), pp. 577–594.

[97] D. A. Knoll and D. E. Keyes. “Jacobian-free Newton-Krylov methods: a survey of approaches and applications”. In: *J. Comput. Phys.* 193.2 (2004), pp. 357–397.

[98] A. V. Knyazev and M. E. Argentati. “Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates”. In: *SIAM J. Sci. Comput.* 23.6 (2002), pp. 2008–2040.

[99] A. Knyazev, A. Jujunashvili, and M. Argentati. “Angles between infinite dimensional subspaces with applications to the Rayleigh-Ritz and alternating projectors methods”. In: *J. Funct. Anal.* 259.6 (2010), pp. 1323–1345.

[100] L. Y. Kolotilina. “Twofold deflation preconditioning of linear algebraic systems. I. Theory”. In: *J. Math. Sci.* 89 (6 1998). Translation of Russian original from 1995, pp. 1652–1689.

[101] S. V. Kuznetsov. “Perturbation bounds of the Krylov bases and associated Hessenberg forms”. In: *Linear Algebra Appl.* 265 (1997), pp. 1–28.

[102] C. Lanczos. “An iteration method for the solution of the eigenvalue problem of linear differential and integral operators”. In: *J. Research Nat. Bur. Standards* 45 (1950), pp. 255–282.

[103] C. Lanczos. “Solution of systems of linear equations by minimized iterations”. In: *J. Research Nat. Bur. Standards* 49 (1952), pp. 33–53.

[104] J. Liesen and Z. Strakoš. “GMRES convergence analysis for a convection-diffusion model problem”. In: *SIAM J. Sci. Comput.* 26.6 (2005), pp. 1989–2009.

[105] J. Liesen and Z. Strakoš. *Krylov subspace methods. Principles and analysis.* Numerical Mathematics and Scientific Computation. Oxford: Oxford University Press, 2013, pp. xvi+391.

[106] J. Liesen and P. Tichý. *The field of values bound on ideal GMRES*. Tech. rep. arXiv:1211.5969v2. 2013.

[107] V. È. Ljance. “Certain properties of idempotent operators”. In: *Teoret. Prikl. Mat. Vyp.* 1 (1958). In Russian, pp. 16–22.

[108] A. Logg and G. N. Wells. “DOLFIN: automated finite element computing”. In: *ACM Trans. Math. Software* 37.2 (2010), Art. 20, 28.

[109] A. Logg, G. N. Wells, and J. Hake. “DOLFIN: a C++/Python Finite Element Library”. In: *Automated Solution of Differential Equations by the Finite Element Method, Volume 84 of Lecture Notes in Computational Science and Engineering*. Ed. by A. Logg, K.-A. Mardal, and G. N. Wells. Berlin: Springer-Verlag, 2012. Chap. 10.

[110] J. Málek and Z. Strakoš. *Preconditioning and the conjugate gradient method in the context of solving PDEs*. Submitted for publication. 2014.

[111] L. Mansfield. “Damped Jacobi preconditioning and coarse grid deflation for conjugate gradient iteration on parallel computers”. In: *SIAM J. Sci. Statist. Comput.* 12.6 (1991), pp. 1314–1323.

[112] L. Mansfield. “On the conjugate gradient solution of the Schur complement system obtained from domain decomposition”. In: *SIAM J. Numer. Anal.* 27.6 (1990), pp. 1612–1620.

[113] R. Mathias. “Quadratic residual bounds for the Hermitian eigenvalue problem”. In: *SIAM J. Matrix Anal. Appl.* 19.2 (1998), pp. 541–550.

[114] V. Mehrmann and C. Schröder. “Nonlinear eigenvalue and frequency response problems in industrial practice”. In: *J. Math. Ind.* 1 (2011), Art. 7, 18.

[115] G. Meurant. “On the incomplete Cholesky decomposition of a class of perturbed matrices”. In: *SIAM J. Sci. Comput.* 23.2 (2001). Copper Mountain Conference (2000), pp. 419–429.

[116] G. Meurant. *The Lanczos and conjugate gradient algorithms*. Vol. 19. Software, Environments, and Tools. From theory to finite precision computations. Philadelphia, PA: SIAM, 2006, pp. xvi+365.

[117] A. Międlar. “Inexact Adaptive Finite Element Methods for Elliptic PDE Eigenvalue Problems”. PhD thesis. Technische Universität Berlin, 2011.

[118] R. B. Morgan. “A restarted GMRES method augmented with eigenvectors”. In: *SIAM J. Matrix Anal. Appl.* 16.4 (1995), pp. 1154–1171.

[119] R. B. Morgan. “Computing interior eigenvalues of large matrices”. In: *Linear Algebra Appl.* 154/156 (1991), pp. 289–309.

[120] R. B. Morgan. “GMRES with deflated restarting”. In: *SIAM J. Sci. Comput.* 24.1 (2002), pp. 20–37.

[121] R. B. Morgan. "Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations". In: *SIAM J. Matrix Anal. Appl.* 21.4 (2000), pp. 1112–1135.

[122] R. B. Morgan. "On restarting the Arnoldi method for large nonsymmetric eigenvalue problems". In: *Math. Comp.* 65.215 (1996), pp. 1213–1230.

[123] R. Nabben and C. Vuik. "A comparison of abstract versions of deflation, balancing and additive coarse grid correction preconditioners". In: *Numer. Linear Algebra Appl.* 15.4 (2008), pp. 355–372.

[124] R. Nabben and C. Vuik. "A comparison of deflation and coarse grid correction applied to porous media flow". In: *SIAM J. Numer. Anal.* 42.4 (2004), pp. 1631–1647.

[125] R. Nabben and C. Vuik. "A comparison of deflation and the balancing preconditioner". In: *SIAM J. Sci. Comput.* 27.5 (2006), pp. 1742–1759.

[126] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen. "How fast are nonsymmetric matrix iterations?" In: *SIAM J. Matrix Anal. Appl.* 13.3 (1992). Iterative methods in numerical linear algebra (Copper Mountain, CO, 1990), pp. 778–795.

[127] O. Nevanlinna. *Convergence of iterations for linear equations.* Lectures in Mathematics ETH Zürich. Basel: Birkhäuser Verlag, 1993, pp. viii+177.

[128] R. A. Nicolaides. "Deflation of conjugate gradients with applications to boundary value problems". In: *SIAM J. Numer. Anal.* 24.2 (1987), pp. 355–365.

[129] C. Nore, M. E. Brachet, and S. Fauve. "Numerical study of hydrodynamics using the nonlinear Schrödinger equation". In: *Phys. D* 65.1-2 (1993), pp. 154–162.

[130] M. A. Olshanskii and V. Simoncini. "Acquired clustering properties and solution of certain saddle point systems". In: *SIAM J. Matrix Anal. Appl.* 31.5 (2010), pp. 2754–2768.

[131] C. C. Paige and M. A. Saunders. "Solutions of sparse indefinite systems of linear equations". In: *SIAM J. Numer. Anal.* 12.4 (1975), pp. 617–629.

[132] C. C. Paige, B. N. Parlett, and H. A. van der Vorst. "Approximate solutions and eigenvalue bounds from Krylov subspaces". In: *Numer. Linear Algebra Appl.* 2.2 (1995), pp. 115–133.

[133] C. C. Paige. "The computation of eigenvalues and eigenvectors of very large sparse matrices". PhD thesis. University of London, 1971.

[134] C. C. Paige, M. Rozložník, and Z. Strakoš. "Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES". In: *SIAM J. Matrix Anal. Appl.* 28.1 (2006), pp. 264–284.

[135] M. L. Parks, E. de Sturler, G. Mackey, D. D. Johnson, and S. Maiti. “Recycling Krylov subspaces for sequences of linear systems”. In: *SIAM J. Sci. Comput.* 28.5 (2006), pp. 1651–1674.

[136] B. N. Parlett and D. S. Scott. “The Lanczos algorithm with selective orthogonalization”. In: *Math. Comp.* 33.145 (1979), pp. 217–238.

[137] B. N. Parlett. “Do we fully understand the symmetric Lanczos algorithm yet?” In: *Proceedings of the Cornelius Lanczos International Centenary Conference (Raleigh, NC, 1993).* Philadelphia, PA: SIAM, 1994, pp. 93–107.

[138] B. N. Parlett. *The symmetric eigenvalue problem.* Vol. 20. Classics in Applied Mathematics. Corrected reprint of the 1980 original. Philadelphia, PA: SIAM, 1998, pp. xxiv+398.

[139] F. Pérez and B. E. Granger. “IPython: a System for Interactive Scientific Computing”. In: *Computing in Science and Engineering* 9.3 (May 2007), pp. 21–29.

[140] G. W. Reddien. “On Newton’s method for singular problems”. In: *SIAM J. Numer. Anal.* 15.5 (1978), pp. 993–996.

[141] J. Ruge and K. Stüben. “Efficient solution of finite difference and finite element equations”. In: *Multigrid methods for integral and differential equations (Bristol, 1983).* Vol. 3. Inst. Math. Appl. Conf. Ser. New Ser. New York: Oxford Univ. Press, 1985, pp. 169–212.

[142] Y. Saad. “Krylov subspace methods for solving large unsymmetric linear systems”. In: *Math. Comp.* 37.155 (1981), pp. 105–126.

[143] Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc’h. “A deflated version of the conjugate gradient algorithm”. In: *SIAM J. Sci. Comput.* 21.5 (2000), pp. 1909–1926.

[144] Y. Saad. “Projection methods for solving large sparse eigenvalue problems”. In: *Matrix Pencils.* Ed. by B. Kågström and A. Ruhe. Vol. 973. Lecture Notes in Mathematics. Springer, 1983, pp. 121–144.

[145] Y. Saad and M. H. Schultz. “GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems”. In: *SIAM J. Sci. Statist. Comput.* 7.3 (1986), pp. 856–869.

[146] Y. Saad. “Analysis of augmented Krylov subspace methods”. In: *SIAM J. Matrix Anal. Appl.* 18.2 (1997), pp. 435–449.

[147] Y. Saad. *Iterative methods for sparse linear systems.* Second. Philadelphia, PA: SIAM, 2003, pp. xviii+528.

[148] Y. Saad. *Numerical Methods for Large Eigenvalue Problems.* Philadephia, PA: SIAM, 2011, pp. xvi+276.

[149] N. Schlömer and W. Vanroose. “An optimal linear solver for the Jacobian system of the extreme type-II Ginzburg-Landau problem”. In: *J. Comput. Phys.* 234 (2013), pp. 560–572.

[150] N. Schlömer. *matplotlib2tikz*. 2010–. URL: `https://github.com/nschloe/matplotlib2tikz/`.

[151] N. Schlömer, D. Avitabile, and W. Vanroose. “Numerical bifurcation study of superconducting patterns on a square”. In: *SIAM J. Appl. Dyn. Syst.* 11.1 (2012), pp. 447–477.

[152] J. R. Shewchuk. “Delaunay refinement algorithms for triangular mesh generation”. In: *Comput. Geom.* 22.1-3 (2002). 16th ACM Symposium on Computational Geometry (Hong Kong, 2000), pp. 21–74.

[153] J. A. Sifuentes, M. Embree, and R. B. Morgan. “GMRES Convergence for Perturbed Coefficient Matrices, with Application to Approximate Deflation Preconditioning”. In: *SIAM J. Matrix Anal. Appl.* 34.3 (2013), pp. 1066–1088.

[154] H. D. Simon. “Analysis of the symmetric Lanczos algorithm with reorthogonalization methods”. In: *Linear Algebra Appl.* 61 (1984), pp. 101–131.

[155] V. Simoncini and D. B. Szyld. “On the Superlinear Convergence of MINRES”. English. In: *Numerical Mathematics and Advanced Applications 2011*. Ed. by A. Cangiani et al. Berlin: Springer, 2013, pp. 733–740.

[156] V. Simoncini and D. B. Szyld. “Recent computational developments in Krylov subspace methods for linear systems”. In: *Numer. Linear Algebra Appl.* 14.1 (2007), pp. 1–59.

[157] V. Simoncini and D. B. Szyld. “Theory of inexact Krylov subspace methods and applications to scientific computing”. In: *SIAM J. Sci. Comput.* 25.2 (2003), pp. 454–477.

[158] G. L. G. Sleijpen, H. A. van der Vorst, and J. Modersitzki. “Differences in the effects of rounding errors in Krylov solvers for symmetric indefinite linear systems”. In: *SIAM J. Matrix Anal. Appl.* 22.3 (2000), pp. 726–751.

[159] B. K. Som, M. R. Gupta, and B. Dasgupta. “Coupled nonlinear Schrödinger equation for Langmuir and dispersive ion acoustic waves”. In: *Phys. Lett. A* 72.2 (1979), pp. 111–114.

[160] K. M. Soodhalter, D. B. Szyld, and F. Xue. *Krylov Subspace Recycling for Sequences of Shifted Linear Systems*. Preprint. arXiv:1301.2650v3. 2013.

[161] G. Starke. “Iterative Methods and Decomposition-Based Preconditioners for Nonsymmetric Elliptic Boundary Value Problems”. Habilitationsschrift. Universität Karlsruhe, 1994.

[162] G. Starke. “Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems”. In: *Numer. Math.* 78.1 (1997), pp. 103–117.

[163] G. W. Stewart. “Backward error bounds for approximate Krylov subspaces”. In: *Linear Algebra Appl.* 340 (2002), pp. 81–86.

[164] G. W. Stewart. *Matrix algorithms. Vol. II.* Eigensystems. Philadelphia, PA: SIAM, 2001, pp. xx+469.

[165] G. W. Stewart. "On the numerical analysis of oblique projectors". In: *SIAM J. Matrix Anal. Appl.* 32.1 (2011), pp. 309–348.

[166] G. W. Stewart. "Two simple residual bounds for the eigenvalues of a Hermitian matrix". In: *SIAM J. Matrix Anal. Appl.* 12.2 (1991), pp. 205–208.

[167] G. W. Stewart and J. G. Sun. *Matrix perturbation theory.* Computer Science and Scientific Computing. Boston, MA: Academic Press Inc., 1990, pp. xvi+365.

[168] K. Stüben. "A review of algebraic multigrid". In: *J. Comput. Appl. Math.* 128.1-2 (2001). Numerical analysis 2000, Vol. VII, Partial differential equations, pp. 281–309.

[169] E. de Sturler. "Nested Krylov methods based on GCR". In: *J. Comput. Appl. Math.* 67.1 (1996), pp. 15–41.

[170] E. de Sturler. "Truncation strategies for optimal Krylov subspace methods". In: *SIAM J. Numer. Anal.* 36.3 (1999), pp. 864–889.

[171] D. B. Szyld. "The many proofs of an identity on the norm of oblique projections". In: *Numer. Algorithms* 42.3-4 (2006), pp. 309–323.

[172] J. M. Tang, S. P. MacLachlan, R. Nabben, and C. Vuik. "A Comparison of Two-Level Preconditioners Based on Multigrid and Deflation". In: *SIAM J. Matrix Anal. Appl.* 31.4 (2010), pp. 1715–1739.

[173] J. M. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga. "Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods". In: *J. Sci. Comput.* 39.3 (2009), pp. 340–370.

[174] K.-C. Toh and L. N. Trefethen. "Calculation of pseudospectra by the Arnoldi iteration". In: *SIAM J. Sci. Comput.* 17.1 (1996). Special issue on iterative methods in numerical linear algebra (Breckenridge, CO, 1994), pp. 1–15.

[175] L. N. Trefethen. "Approximation theory and numerical linear algebra". In: *Algorithms for approximation, II (Shrivenham, 1988).* London: Chapman and Hall, 1990, pp. 336–360.

[176] L. N. Trefethen. "Computation of pseudospectra". In: *Acta numerica, 1999.* Vol. 8. Acta Numer. Cambridge: Cambridge Univ. Press, 1999, pp. 247–295.

[177] L. N. Trefethen and M. Embree. *Spectra and pseudospectra.* The behavior of nonnormal matrices and operators. Princeton, NJ: Princeton University Press, 2005, pp. xviii+606.

[178] C. Vuik, R. Nabben, and J. Tang. "Deflation acceleration for domain decomposition preconditioners". In: *Proceedings of the 8th European Multigrid Conference September 27-30, 2005 Scheveningen The Hague, The Netherlands.* Ed. by P. Wesseling, C. Oosterlee, and P. Hemker. Delft: TU Delft, 2006.

[179] D. N. Wakam and J. Erhel. “Parallelism and robustness in GMRES with a Newton basis and deflated restarting”. In: *Electron. Trans. Numer. Anal.* 40 (2013), pp. 381–406.

[180] H. F. Walker. “Implementation of the GMRES method using Householder transformations”. In: *SIAM J. Sci. Statist. Comput.* 9.1 (1988), pp. 152–163.

[181] S. Wang, E. de Sturler, and G. H. Paulino. “Large-scale topology optimization using preconditioned Krylov subspace methods with recycling”. In: *Internat. J. Numer. Methods Engrg.* 69.12 (2007), pp. 2441–2468.

[182] D. Werner. *Funktionalanalysis.* Sechste Auflage. Berlin: Springer-Verlag, 2007, pp. xiii+532.

[183] J. H. Wilkinson. *The algebraic eigenvalue problem.* Clarendon Press, Oxford, 1965, pp. xviii+662.

[184] T. G. Wright. *EigTool.* 2002. URL: `http://www.comlab.ox.ac.uk/pseudospectra/eigtool/`.

[185] T. G. Wright and L. N. Trefethen. “Large-scale computation of pseudospectra using ARPACK and eigs”. In: *SIAM J. Sci. Comput.* 23.2 (2001). Copper Mountain Conference (2000), pp. 591–605.

[186] T. G. Wright and L. N. Trefethen. “Pseudospectra of rectangular matrices”. In: *IMA J. Numer. Anal.* 22.4 (2002), pp. 501–519.

[187] G. Wu, Y. Wei, Z.-g. Jia, S.-t. Ling, and L. Zhang. “Towards backward perturbation bounds for approximate dual Krylov subspaces”. In: *BIT* 53.1 (2013), pp. 225–239.

[188] M. Yeung, J. Tang, and C. Vuik. *On the Convergence of GMRES with Invariant-Subspace Deflation.* Report 10-14. Delft: Delft University of Technology, Delft Institute of Applied Mathematics, 2010.

[189] K. Yosida. *Functional analysis.* Classics in Mathematics. Reprint of the sixth (1980) edition. Berlin: Springer-Verlag, 1995, pp. xii+501.