DISSERTATION

Krylov techniques and approximations to the action of matrix exponentials

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften unter der Leitung von

Prof. Dr. Winfried Auzinger
E101 – Institut für Analysis und Scientific Computing, TU Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Tobias Jawecki
Matrikelnummer: 01026658

Diese Dissertation haben begutachtet:

1. **Prof. Dr. Winfried Auzinger**
   Institute of Analysis and Scientific Computing, TU Wien
2. **Prof. Dr. Stefan Güttel**
   Department of Mathematics, University of Manchester
3. **Prof. Dr. Mechthild Thalhammer**
   Institut für Mathematik, Universität Innsbruck

Wien, am 07.11.2022

Kurzfassung

Die Matrixexponentialfunktion stellt eine Fundamentallösung für autonome Systemen linearer Differentialgleichungen dar. Angewendet auf einen Startvektor lässt sich damit die Zeitentwicklung dieses Vektors für einen gegebenen Zeitschritt berechnen. In der aktuellen Arbeit betrachten wir schwachbesetzte Systeme von großer Dimension, welche sich häufig aus der numerischen Behandlung von partiellen Differentialgleichungen (Evolutionsgleichungen) ergeben. Für solche Systeme ist eine direkte Auswertung der Matrixexponentialfunktion nicht praxistauglich. Die Anwendung der Matrixexponentialfunktion auf einem Startvektor lässt sich aber effizient durch Krylov-Unterraum Methoden annähern.

Solche Methoden sind Thema dieser Arbeit. Unter anderem betrachten wir Abschätzungen für den Fehler von Krylov-Unterraum Verfahren zur numerischen Zeitintegration. Damit lässt sich die Dimension des Unterraums bzw. die Größe von Zeitschritten des Verfahrens steuern, um die numerische Zeitenwicklung mit hinreichender Genauigkeit zu berechnen. Für verschiedene Arten von Systemen leiten wir auch neue obere Schranken für die Fehlernorm her, welche eine besonders zuverlässige Fehlerschätzung erlauben.

Ein weiteres Thema dieser Arbeit ist der Zusammenhang zwischen der Spektralzerlegung des Startvektors und Spektralzerlegungen, welche sich aus der Projektion auf einen Krylov-Unterraum ergeben. Der Fehler der betrachteten Verfahren zur numerischen Zeitintegration lässt sich über eine Spektralzerlegung des Startvektors darstellen, welche aber in der Praxis nicht zur Verfügung steht. Abschätzungen zur Spektralzerlegung des Startvektors basierend auf Krylov-Unterräume können aber hilfreiche Informationen zum Problem liefern, z.B. um rationale Matrixfunktionen zur numerischen Zeitintegration zu generieren, die relevante Charakteristiken der Lösung besonders gut imitieren.

In dieser Arbeit betrachten wir auch das Konvergenzverhalten von Methoden zur numerischen Zeitintegration von schief-Hermiteschen Systemen basierend auf rationalen Krylov-Unterraum Verfahren. Die Struktur der Spektralzerlegung des Startvektors bleibt im rationalen Krylov-Unterraum teilweise erhalten. Bestimmte Charakteristiken solcher Systeme, welche sich vorteilhaft auf rationale Annäherungen auswirken, lassen sich daher auch mit Verfahren basierend auf rationalen Krylov-Unterräumen nutzen. Damit lässt sich teilweise ein von einer zugrundeliegenden Gitterdiskretisierung unabhängiges Konvergenzverhalten solcher Methoden erklären.

Abstract

The matrix exponential represents a time evolution operator for a linear autonomous system of ordinary differential equations. The action of the matrix exponential on a starting vector yields its time evolution for a given time step. In the present thesis, we consider sparse and large systems, which appear frequently in the context of discretized partial differential equations of evolutionary type. In this setting, a direct computation of the matrix exponential is not practicable. However, the action of the matrix exponential can be efficiently approximated using Krylov techniques which are the topic of the present thesis.

The first part of this thesis is mainly dedicated to error estimates for the Krylov approximation to the action of matrix exponentials. Error estimates provide a proper dimension for the underlying Krylov subspace, or a proper choice of time steps such that the constructed time propagator is sufficiently accurate. For various types of systems, we introduce upper bounds on the error norm, which constitute most reliable error estimates.

Another topic of the present thesis is the relation between the spectral distribution of the starting vector and spectral distributions which result from projection on Krylov subspaces. The error of the discussed time integration methods can be represented by the spectral distribution of the starting vectors, which is not accessible in practice. Estimates on this spectral distribution based on Krylov subspace techniques can be of some use for numerical time integration, e.g., to design rational approximants to the matrix exponential which imitate specific characteristics of the exact time evolution.

In the present thesis, we also study the convergence behavior of rational Krylov approximations to the action of the exponential of skew-Hermitian matrices. The structure of the spectral distribution of the starting vector is partly preserved in the rational Krylov subspace. Thus, specific characteristics of such systemts which are desirable for rational approximation are also of some use for rational Krylov approximations. Such ideas yield some insight on convergence behavior independent of a refinement of an underlying grid discretization for such methods.

Acknowledgment

This work was supported by the Doctoral College TU-D, Technische Universität Wien.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 07.11.2022

Tobias Jawecki

Contents

1 Introduction

1.1 Disclaimer

This work consists of four main chapters, Chapter 2–5. Each of these chapters is written in a self-contained manner including an individual introductory section.

Some of these materials have been published in parallel as a full paper or a preprint:

- The content of Chapter 2 has been published in [JAK20] and is cited literally in the present thesis up to some editorial changes. The paper [JAK20], representing the start of my scientific activity on Krylov methods in the context of the TU-D doctoral program, was composed in collaboration with my advisor Winfried Auzinger (TU Wien) and Othmar Koch (Universität Wien). This deserves some explanation: The ideas about error estimates considered in this work originate from the specification of the topic when co-applying for the TU-D program by my advisor. Moreover, Othmar Koch was in parallel working on a related project on quantum dynamics granted by the Austrian Science Fund (FWF), where adaptive Krylov techniques play also an important role.

  In this context I wish to emphasize that essential results in this work are based on my personal ideas. This includes a new rigorous a posteriori error estimator, extension to φ-functions, and the concept of effective order. I also have proposed and implemented a practical stepsize selection strategy. The results were checked and verified by my co-authors, including some technical modifications.

- Chapter 3 has been published in [Jaw22b] and is cited literally in the present thesis up to some editorial changes plus Section 3.D which contains additional unpublished material.

- Chapter 4 is also available online as an individual preprint at [Jaw22a]; these versions are identical up to some editorial changes.

1.2 Overview and setting of Chapter 2

In this chapter we consider polynomial Krylov approximations to the action of matrix φ-functions, i.e., the vector $\varphi_p(\sigma\, tA)v \in \mathbb{C}^n$ for $p \geq 0$, a time step t, a complex phase $\sigma \in \mathbb{C}$ with $|\sigma| = 1$, a given matrix $A \in \mathbb{C}^{n\times n}$ and a given vector $v \in \mathbb{C}^n$. This includes the action of the matrix exponential $\mathrm{e}^{\sigma\, tA}v$, which corresponds to the case $p = 0$ with $\varphi_0 \equiv \exp$, and yields a solution to linear systems of ordinary differential equations. For the case $p > 0$, the φ-functions yield solutions to systems of inhomogenous differential equations and have some relevance in exponential integrators. The polynomial Krylov approximation of $\varphi_p(\sigma tA)v$

is based on the Krylov subspace $\mathcal{K}_m(A, v) = \operatorname{span}\{v, Av, \ldots, A^{m-1}v\}$ where m denotes the dimension of the subspace. By explicitly denoting the complex phase σ, we aim to simplify the notation in skew-Hermitian cases, e.g., the approximation of $\mathrm{e}^{\mathrm{i}tH}v$ in the Krylov subspace $\mathcal{K}_m(H, v)$ which corresponds to the case $\sigma = \mathrm{i}$ and $A = H$ for an Hermitian matrix H. In the following we assume a *non-expansive* case, i.e., the numerical range of σA is located in the left complex plane including the imaginary axis.

Throughout this chapter, our main focus is on a posteriori error estimates and upper bounds on the error norm. We first recall a *defect*-based error representation for the Krylov approximation of $\mathrm{e}^{\sigma tA}v$. In the non-expansive case, this yields an upper bound on the error norm via the *defect integral*, i.e., the integral of the absolute value of the *scalar defect*. The scalar defect is closely related to divided differences of the exponential function over the *Ritz values*, i.e., the eigenvalues of the representation of A in the Krylov basis. Making use of properties of divided differences for the scalar defect, we derive a new upper bound on the error norm which corresponds to the leading order term of the error norm for a time step $t \to 0$. We also recall an error representation based on an expansion in φ-functions. The first term of this expansion is an asymptotically correct error estimate in terms of the time step. We show that, for the Hermitian case, this term also yields an upper bound on the error norm. Similar results are also derived for the Krylov approximation of $\varphi_p(\sigma tA)v$ with $p > 0$ and the so called corrected Krylov approximation for the case $p \geq 0$. However, a defect formulation is not directly given for the case $p > 0$ in Chapter 2 (see Chapter 3 for such results). For the approximation of the matrix exponential, we study further a posteriori error estimates which can be understood as quadrature estimates of the defect integral. In this concern, we derive a new error estimate which is based on the *effective order* of the defect. The Krylov approximation of the matrix exponential can be understood as a time integrator, and restarted accordingly. In this context, we apply error estimates to adaptively choose proper sub steps in time or as a stopping criteria to avoid using an unnecessary large dimension for the underlying Krylov subspace. We compare the performance of different error estimates for these tasks using numerical experiments.

1.3 Overview and setting of Chapter 3

Similar to Chapter 2, we consider a polynomial Krylov approximation of $\varphi_p(tA)v$ using the Krylov subspace $\mathcal{K}_m(A, v)$, where, in contrast to Chapter 2, the complex phase σ is included in A to simplify the notation. In Chapter 3, we derive a formulation for the defect of the Krylov approximation of $\varphi_p(tA)v$ with $p > 0$, and a defect-based error representation in this setting. A similar error representation was restricted to the case $p = 0$ previously. Furthermore, we discuss effects of floating point arithmetic to the Krylov approximation and we include round-off errors in the error representation.

Based on the defect formulation, we derive an upper bound on the error norm for the non-expansive case with $p \geq 0$, particularly, using the defect integral. With further treatment of the defect integral, this results in computable upper bounds. We consider different cases: For the case that Ritz values are real, e.g., when A is Hermitian, the defect integral can be evaluated exactly in practice. For complex Ritz values this is not the case in general. However, we derive a computable upper bound on the defect integral by neglecting the

imaginary parts of the Ritz values for the scalar defect. This result is based on properties of divided differences of the exponential function. We also discuss the accuracy of this upper bound on the defect integral using an asymptotic expansion of the effective order of the defect.

Further error estimates can be ascribed to quadrature formulae applied to the defect integral. In particular, the error estimate based on the effective order of the defect which we derive for $p \geq 0$ in this chapter, the case $p = 0$ is already covered in Chapter 2. We also show that the effective order error estimate constitutes an upper bound on the error norm when considering time steps close to an asymptotic time regime.

Furthermore, we study effects of clustered Ritz values on the scalar defect for time steps close to an asymptotic time regime. This describes oscillatory behaviour of the scalar defect, and has some relevance for quadrature estimates on the defect integral. Clustered Ritz values can be ascribed to regularity of an underlying continuous problem.

We also derive practical stopping criteria for lucky breakdown which are based on a posteriori error bounds. Various numerical examples illustrate the performance of the different a posteriori error estimates discussed in this chapter.

1.4 Overview and setting of Chapter 4

In this chapter we consider a vector $u \in \mathbb{C}^n$, and a matrix $A \in \mathbb{C}^{n\times n}$ which is Hermitian w.r.t. a given inner product (referred to as M-inner product). The linear functional $f \mapsto (u, f(A)u)_{\mathsf{M}}$ can be understood as a Riemann-Stieltjes integral associated with a step function α_n. This step function is defined by eigenvalues of A and *spectral coefficients* of u, i.e., the coefficients in the M-orthonormal eigenbasis of A. The zeros and the *Christoffel numbers* corresponding to orthogonal polynomials associated with the distribution $\mathrm{d}\alpha_n$ yield quadrature nodes and weights, respectively, of Gaussian quadrature formulae for the respective Riemann-Stieltjes integral. The accumulated quadrature weights of Gaussian quadrature formulae constitute bounds on the integral over the intervals between the quadrature nodes. Classical results in this concern date back to works of Chebyshev, Markov and Stieltjes and are referred to as Separation Theorem of Chebyshev-Markov-Stieltjes (CMS Theorem).

We recall the relation between polynomial Krylov subspaces $\mathcal{K}_m(A, u)$, where m denotes the dimension of the Krylov subspace, and orthogonal polynomials associated with the distribution $\mathrm{d}\alpha_n$. In the context of the Krylov subspace, the Gaussian quadrature nodes and weights correspond to the Ritz values and spectral coefficients of u projected onto the Krylov subspace, respectively. In this chapter, we review the classical CMS Theorem and corresponding intertwining properties in the context of Krylov subspaces. E.g., as a consequence of the CMS Theorem, the Krylov subspace yields computable bounds on the sums of spectral coefficients of u related to eigenvalues located between Ritz values. Such results hold independently of the convergence of Ritz values.

Similar results hold true for rational Krylov subspaces $\mathcal{Q}_m(A, u) = \{r(A)u\}$ with $r = p/q$, where m denotes the dimension of the Krylov subspace, p is a polynomial of degree $\leq m-1$, and q is a preassigned denominator polynomial of degree $\leq m - 1$. The basis of a rational Krylov subspace is closely related to orthogonal rational functions associated with the

distribution $\mathrm{d}\alpha_n$, and rational Gaussian quadrature formulae for the respective Riemann-Stieltjes integral. We derive new CMS type results for some classes of rational Gaussian quadrature formulae which are related to rational Krylov subspaces with a single pole of higher multiplicity. These results and other known CMS type results are applied in the context of rational Krylov subspaces with a single pole and some extended Krylov subspaces.

In this chapter, we also consider polynomial and rational *quasi-orthogonal residual (qor-)* Krylov representations which are closely related to Gauss-Radau quadrature formulae where one of the quadrature nodes is preassigned. CMS type results also hold in this context. Our results are illustrated by numerical examples.

Chapter 4 and the action of the matrix exponential. This chapter has some relevance for the present thesis, particularly, for the approximation of $\mathrm{e}^{\mathrm{i}tA}u$ where A is Hermitian. The error of polynomial or rational approximation of $\mathrm{e}^{\mathrm{i}tA}u$ can be represented by a scalar approximation error at the eigenvalues of A weighted by the respective spectral coefficients of u, see Chapter 5 for more details. The eigenvalues of A and spectral coefficients of u are not available in practice. However, piecewise bounds on spectral coefficients of u, as provided in Chapter 4, together with some understanding of polynomial and rational approximation of the scalar imaginary exponential provide some practical information on how to efficiently approximate $\mathrm{e}^{\mathrm{i}tA}u$. Especially for rational approximations, some knowledge on the problem is required to choose proper poles of the rational approximant. Furthermore, the results of Chapter 4 give some motivation for Chapter 5, and the qor-Krylov approximation is practical for the approximation of the matrix exponential.

1.5 Overview and setting of Chapter 5

In this chapter we a consider rational Krylov approximation of $\mathrm{e}^{-\mathrm{i}tA}u$ for a time step t, a matrix $A \in \mathbb{C}^{n\times n}$ which is Hermitian w.r.t. a given inner product, and an initial vector $u \in \mathbb{C}^n$. The rational Krylov approximation is based on a rational Krylov subspace $\mathcal{Q}_m(A,u)$ with preassigned poles which define the denominator of the rational approximant, and previous results show a near-best approximation property comparing with other rational approximants with the same denominator. Such results rely on the error of a scalar best approximation over the full range of the matrix spectrum and do not consider properties of the initial vector. In the present work, a localized near-best approximation property is formulated. This property is based on assumptions on the spectral distribution of the initial vector, and the assumption that the problem projected onto the Krylov subspaces satisfies a similar distribution. Whether the latter holds true is not known a priori without further consideration. However, it is reasonable to assume that desired properties of the spectral distribution of the initial vector carry over to the distribution of the projected problem due to an intertwining property which goes back to the Separation Theorem of Chebyshev-Markov-Stieltjes, see Chapter 4. The required assumption is tested for practical numerical examples. A localized near-best approximation property potentially yields grid-indepentent convergence rates if the matrix is based on an underlying spatial discretization of a continuous operator. In that concern, assumptions on the spectral distribution of the

initial vector are related to regularity properties of an underlying initial state.

2 Computable upper error bounds for Krylov approximations to matrix exponentials and associated φ-functions

2.1 Introduction

We consider Krylov approximations to the matrix exponential function for the purpose of the solution of a linear, homogeneous system of differential equations

$$\psi'(t) = M\psi(t), \quad \psi(0) = \psi_0, \qquad \psi(t) = \mathrm{e}^{tM}\psi_0.$$

The complex-valued matrix M commonly results from the discretization of a partial differential equation. In this work we present new results for precise a posteriori error estimation, which also extend to the evaluation of so-called φ-functions. The application of these estimates for the purpose of time propagation is also discussed and illustrated. Theoretical results are verified by numerical experiments, which are classified into Hermitian (dissipative), skew-Hermitian (Schrödinger-type) and general non-normal problems.

Overview of existing approaches and results. The approximate evaluation of large matrix exponential functions is a topic which has been extensively treated in the numerical analysis literature, for basic reference see e.g. [GVL89, MVL03]. A standard approach is to project the given matrix M to a low-dimensional Krylov space via Arnoldi or Lanczos iteration, and to directly exponentiate the projected small matrix. A first mention of the Lanczos approach can be found in [PL86], where it is also recognized that for the method to perform satisfactorily, the time-steps have to be controlled. However, the control mechanism from [PL86] is not very elaborate and is based on a series expansion of the error, which is only valid in the asymptotic regime, see for instance [NW12]. For discretizations of parabolic problems, [GS92] uses an error estimator to choose the step-size, this approach is improved in [SL96] and has been generalized in [MC10]. Notably, in the latter reference a strict error bound is used to estimate the time-step instead of asymptotic techniques. It is argued in [MC10] that the strategy from [MC10] performs better than [MA06] and better in turn than [PL86].

A first systematic study of Krylov-based methods for the matrix exponential function was given in [Saa92]. The error is analyzed theoretically, yielding both a priori and computable a posteriori estimates. The analysis there relies on approximation theory and yields a priori error bounds which are asymptotically optimal in the dimension of the Krylov subspace in important situations. The analysis moreover implies correction schemes to lift the convergence order which are cheap to compute based on the already available information. The error expansion also suggests a posteriori error estimators resorting to the leading

error term. This approach relies on the assumption of the sufficiently rapid decay of the series representation of the error. A recent generalization of this work together with a more rigorous justification is given in [JL15]. For early studies of a priori error estimates see also [DK89, DK95].

A thorough theoretical analysis of the error of Krylov methods for the exponential of a Hermitian or skew- (anti-) Hermitian matrix was given in [HL97]. The analysis derives an asymptotic error expansion and shows superlinear error decay in the dimension m of the approximation subspace for sufficiently large m. These results are further improved in [BR09]. In [HL97], a posteriori error estimation is also discussed. This topic is furthermore addressed in [Lub08]. There, the Krylov approximation method is interpreted as a Galerkin method, whence an error bound can be obtained from an error representation for this variational approximation. This yields a computable estimate via a quadrature approximation of the error integral involving the defect of the numerical approximation. The a priori error analysis reveals a step-size restriction for the convergence of the method, which is less stringent when the subspace dimension is larger.

Further work in the direction of controlling the Lanczos process through information gained from the defect is given in [BGH13]. The defect is a scalar multiple of the successive Krylov vector arising in the iteration and can be evaluated efficiently. If the error is approximated by a Galerkin approach, the resulting estimator corresponds to the difference of two Lanczos iterates. For the purpose of practical error estimation, in [BGH13] it is seen as preferable to continue the original Krylov process. Some other defect-based upper bounds for the error of the matrix exponential are given in [JL15], including a closer analysis of the error estimate of [Saa92]. These results still require some a priori information on the matrix spectrum.

Various improved methods for computing the matrix exponential function are given in the literature, for example restarted methods, deflated restarting methods or quadrature based restarting methods, see [AEEG08], [EEG11], and [FGS14].

It has also been advocated in [vdEH06] to use preconditioning in the Lanczos method by a shifted inverse in order to get a good approximation of the leading invariant subspaces. The shift-and-invert approach (a specific choice to construct a rational Krylov subspace) for the matrix exponential function was introduced earlier in [MN04]. However, the choice of the shift is critical for the success of this procedure. This strategy amounts to a transformation of the spectrum which grants a convergence speed which is independent of the norm of the given matrix. In [vdEH06], a posteriori error estimation based on the asymptotical expansion of the error is advocated as well. We note that our results do not immediately carry over to the shift-and-invert approach, see Remark 2.3.4.

Overview of present work. In Section 2.2 we introduce the Krylov approximation and the integral representation of the approximation error in terms of its defect. In Section 2.3 we derive a new computable upper bound for the error by using data available from the Krylov process with negligible additional computational effort (Theorem 2.3.2). This upper bound is cheap to evaluate and update on the fly during the Lanczos iteration. It is also asymptotically correct, i.e., for $t \to 0$ the error of the error estimator tends to zero faster asymptotically than the error itself. In Section 2.4 these results are extended to

the case where the Krylov approach is employed to approximate the φ-functions of matrices (generalizing the exponential function), see Theorem 2.4.1. In Section 2.5, improved approximations derived from a corrected Krylov process [Saa92] are discussed, and corresponding error estimators are analyzed, including an asymptotically correct true upper bound on the error (Theorem 2.5.4). This approach can be used to increase the order, but it has the drawback of violating mass conservation. In Proposition 2.5.5 error estimates are particularized to the Hermitian case. Another view on defect-based error estimation is presented in Section 2.6.

Section 2.7 is devoted to practical application of the various error estimators for the control of the time steps t including smaller substeps Δt if it appears indicated. In Section 2.8 we present numerical results for a finite difference discretization of the free Schrödinger equation, a Hubbard model of solar cells, the heat equation, and a convection-diffusion problem, illustrating our theoretical results. Additional practical aspects are also investigated: A priori estimates and the role of restarting are discussed in particular in the context of practical step-size adaptation. Finally, we demonstrate the computational efficiency of our adaptive strategy.

2.2 Problem setting, Krylov approximation, and defect-based representation of the approximation error

We discuss the approximation of the matrix exponential,

$$E(t)v = \mathrm{e}^{\sigma tA}v, \quad A \in \mathbb{C}^{n\times n}, \quad \sigma \in \mathbb{C}, \tag{2.2.1}$$

with step size t, applied to an initial vector $v \in \mathbb{C}^n$. To simplify the notation we assume $|\sigma| = 1$ and $\|v\|_2 = 1$ without loss of generality. In many relevant applications (Schrödinger-type problems) a complex prefactor is applied to the matrix A. The parameter σ is introduced here to separate the prefactor of the matrix A. The standard notation for Schrödinger-type problems is obtained in (2.2.1) with $\sigma = -\mathrm{i}$ and a Hermitian matrix A. For such problems our notation is helpful to simplify the construction of the Krylov subspace.

The exponential $E(t) = \mathrm{e}^{\sigma tA}$ satisfies the matrix differential equation

$$E'(t) = \sigma AE(t), \quad E(0) = I.$$

We assume that $\mu_2(\sigma A) \le 0$, where $\mu_2(\sigma A)$ denotes the logarithmic norm of σA, or equivalently, $W(\sigma A) \subseteq \mathbb{C}_-$, where $W(\sigma A)$ denotes the field of values of σA and we will refer to this assumption as the *nonexpansive case*. Following [Hig08, Theorem 10.11] and associated references we conclude that $\mu_2(\sigma A) \le 0$ implies $\|E(t)\|_2 \le 1$ for $t \ge 0$. This is essentially a technical assumption, and most of our theoretical results carry over to a more general setting, in particular if a priori information about $\mu_2(\sigma A)$ is available, such that $E(t)$ can be estimated as $\|E(t)\|_2 \le \mathrm{e}^{\mu_2(\sigma A)}$.

For the skew-Hermitian case with $\sigma = -\mathrm{i}$ we write[1]

$$E(t)v = \mathrm{e}^{-\mathrm{i}tH}v, \quad H \in \mathbb{C}^{n\times n} \text{ Hermitian.}$$

[1] In this case the matrix A is usually named H (Hamiltonian).

In this case, $E(t)$ represents a unitary evolution, i.e., $\|E(t)\|_2 = 1$.

Krylov subspaces and associated identities. The numerical approximation of (2.2.1) considered here (see (2.2.6) below) is based on the conventional Krylov subspace

$$\mathcal{K}_m(A, v) = \operatorname{span}\{v, Av, \ldots, A^{m-1}v\} \subseteq \mathbb{C}^n.$$

First, an orthonormal basis of $\mathcal{K}_m(A, v)$ is obtained by the well-known Arnoldi iteration, see [Saa03]. This produces a basis matrix $V_m \in \mathbb{C}^{n\times m}$ satisfying $V_m^* V_m = I_{m\times m}$, and an upper Hessenberg matrix $T_m \in \mathbb{C}^{m\times m}$ such that the Krylov identity[2]

$$A\,V_m = V_m T_m + \tau_{m+1,m}\, v_{m+1} e_m^* \tag{2.2.2}$$

is valid, with $\tau_{m+1,m} \in \mathbb{R}_+$ and $v_{m+1} \in \mathbb{C}^n$ with $\|v_{m+1}\|_2 = 1$.

Remark 2.2.1. *We are assuming that the Arnoldi iteration is executed until the desired dimension m. Then, by construction, all lower diagonal entries of T_m are positive [Saa03]. If this is not the case, i.e., if a breakdown occurs, it is known that this breakdown is* lucky,, *i.e., the approximation* (2.2.6) *below obtained in the step before breakdown is already exact, see [Saa92].*

For the case of a Hermitian matrix A the Krylov subspace can be constructed using the Lanczos iteration, which is a special case of the Arnoldi iteration, resulting in a tridiagonal matrix $T_m \in \mathbb{R}^{m\times m}$. In the following we discuss the general case and comment on the case of a Hermitian matrix A whenever appropriate.

The following identities hold true due to the upper Hessenberg [tridiagonal] structure of T_m together with (2.2.2):

$$e_m^*\, T_m^j\, e_1 = 0 \quad \text{for} \quad j = 0, \ldots, m-2, \tag{2.2.3}$$

and

$$A^j v = V_m T_m^j e_1, \quad 0 \le j \le m-1, \tag{2.2.4}$$

see for instance [DK89, Theorem 2] or [Saa92]. Furthermore, let

$$\gamma_m = e_m^*\, T_m^{m-1} e_1 = \prod_{j=1}^{m-1} (T_m)_{j+1,j}\,, \tag{2.2.5}$$

where the claimed identity also follows from the upper Hessenberg [tridiagonal] structure of T_m.

Krylov approximation. The standard Krylov approximation to $E(t)v$ is

$$S_m(t)v = V_m\, \mathrm{e}^{\sigma\, tT_m}\, V_m^*\, v = V_m\, \mathrm{e}^{\sigma\, tT_m} e_1. \tag{2.2.6}$$

We denote the corresponding error operator by $L_m(t)$, with

$$L_m(t) = E(t) - S_m(t) \in \mathbb{C}^{n\times n}. \tag{2.2.7}$$

[2] Here, $e_m = (0, \ldots, 0, 1)^* \in \mathbb{C}^m$, and in the sequel we also denote $e_1 = (1, 0, \ldots, 0)^* \in \mathbb{C}^m$.

Defect-based integral representation of the approximation error. We define the *defect* (or residual) operator $D_m(t)$ of $S_m(t)$ by

$$D_m(t) = \sigma\, A\, S_m(t) - S_m'(t) \in \mathbb{C}^{n\times n}.$$

Then, $L_m(t)v$ and $D_m(t)v$ are related via the differential equation

$$L_m'(t)v = \sigma\, A\, L_m(t)v + D_m(t)v, \quad L_m(0)v = 0,$$

whence

$$L_m(t)v = \int_0^t E(t-s)\, D_m(s)v \,\mathrm{d}s. \tag{2.2.8}$$

An explicit representation for $D_m(s)v$ is obtained from (2.2.2),

$$\begin{aligned} D_m(s)v &= \sigma\, A\, V_m\, \mathrm{e}^{\sigma\, s\, T_m} e_1 - \sigma\, V_m T_m\, \mathrm{e}^{\sigma\, s\, T_m} e_1 = \sigma\,(A\, V_m - V_m T_m)\, \mathrm{e}^{\sigma\, s\, T_m} e_1 \\ &= \sigma \tau_{m+1,m}\, \left(e_m^*\, \mathrm{e}^{\sigma\, s\, T_m} e_1\right) v_{m+1}. \end{aligned} \tag{2.2.9}$$

Asymptotically for $t \to 0$,

$$D_m(t)v = \sigma \tau_{m+1,m}\, \gamma_m \frac{(\sigma t)^{m-1}}{(m-1)!}\, v_{m+1} + \mathcal{O}(t^m), \tag{2.2.10}$$

which follows from the Taylor series representation for $\mathrm{e}^{\sigma\, t\, T_m}$ together with (2.2.3) and (2.2.5). Thus, by (2.2.8) and (2.2.10) we obtain

$$\|D_m(t)v\| = \mathcal{O}(t^{m-1}), \quad \text{and} \quad \|L_m(t)v\| = \mathcal{O}(t^m). \tag{2.2.11}$$

We can also characterize the asymptotically leading term of the error:

Proposition 2.2.2. *For any $A \in \mathbb{C}^{n\times n}$ the error $L_m(t)v$ satisfies the asymptotic relation*

$$L_m(t)v = \tau_{m+1,m}\gamma_m \frac{(\sigma\, t)^m}{m!} v_{m+1} + R_{m+1}(t), \quad R_{m+1}(t) = \mathcal{O}(t^{m+1}), \tag{2.2.12}$$

for $t \to 0$.

Proof. Taylor expansion. Due to $L_m(t)v = \mathcal{O}(t^m)$, see (2.2.11),

$$\begin{aligned} &L_m(t)v = E(t)v - S_m(t)v = \frac{(\sigma\, t)^m}{m!}(A^m v - V_m T_m^m\, e_1) + R_{m+1}(t), \\ &\text{with Taylor remainder } \; R_{m+1}(t) = \mathcal{O}(t^{m+1}). \end{aligned} \tag{2.2.13}$$

Multiplying the identity (2.2.4) (with $j = m-1$) by A and using (2.2.2) gives

$$\begin{aligned} A^m\, v &= A\, V_m T_m^{m-1} e_1 = (V_m T_m + \tau_{m+1,m}\, v_{m+1} e_m^*) T_m^{m-1} e_1 \\ &= V_m T_m^m\, e_1 + \tau_{m+1,m}\, (e_m^*\, T_m^{m-1} e_1)\, v_{m+1} = V_m T_m^m\, e_1 + \tau_{m+1,m}\gamma_m v_{m+1}, \end{aligned}$$

whence (2.2.13) simplifies to (2.2.12). □

Remark 2.2.3. *The Taylor remainder R_{m+1} in (2.2.12) can be specified in a more explicit way showing its dependence on m,*

$$R_{m+1}(t) = \frac{(\sigma\, t)^{m+1}}{m!} \int_0^1 \left(A^{m+1} \mathrm{e}^{\sigma\, \theta\, t\, A} v - V_m T_m^{m+1} \mathrm{e}^{\sigma\, \theta\, t\, T_m} e_1\right)(1-\theta)^m \,\mathrm{d}\theta.$$

2.3 An upper error bound for the nonexpansive case in (2.2.1)

For the nonexpansive case we have $\|E(t-s)\|_2 \le 1$ for $0 \le s \le t$, and (2.2.8) implies

$$\|L_m(t)v\|_2 = \Big\| \int_0^t E(t-s)\, D_m(s)v \,\mathrm{d}s \Big\|_2 \le \int_0^t \|D_m(s)v\|_2 \,\mathrm{d}s.$$

With $\|v_{m+1}\|_2 = 1$, and

$$\delta_m(s) = e_m^*\, \mathrm{e}^{\sigma\, s\, T_m} e_1 = \big(\mathrm{e}^{\sigma\, s\, T_m}\big)_{m,1}, \tag{2.3.1a}$$

together with (2.2.9) we obtain

$$\|L_m(t)v\|_2 \le \tau_{m+1,m} \int_0^t |\delta_m(s)| \,\mathrm{d}s. \tag{2.3.1b}$$

This estimate is also given in [Lub08, Section III.2] and appeared earlier in [DGK98, Subsection 2.2]. Of course, the integral in (2.3.1b) cannot be computed exactly. In [Lub08] it is proposed to use numerical quadrature[3] to approximate the integral in (2.3.1b). In contrast, our aim here is to derive a computable upper bound. We proceed in two steps.[4]

Analytic matrix function via interpolation on the spectrum. To approximate the error integral in (2.3.1b) we use the representation of matrix exponentials via Hermite interpolation of the scalar exponential function on the spectrum of the matrix T_m, see [Hig08, Chap. 1]: If $\mu_1, \dots, \mu_r$ ($r \le m$) denote the distinct eigenvalues of T_m and n_j is the dimension of the largest Jordan block associated with μ_j, then

$$\mathrm{e}^{\sigma\, t T_m} = p_t(T_m), \tag{2.3.2}$$

where $p_t(\lambda)$ is the Hermite interpolant of degree $\le m-1$ of the function

$$f_t(\lambda) = \mathrm{e}^{\sigma\, t\lambda} \tag{2.3.3}$$

over the nodes $\mu_1, \dots, \mu_m$ in the sense of [Hig08, (1.7)],

$$p_t^{(\ell)}(\mu_j) = f_t^{(\ell)}(\mu_j), \quad j = 1, \dots, r, \quad \ell = 0, \dots, n_j - 1.$$

For a general matrix, the degree of p_t may be smaller than $m-1$. However, in our context a special case occurs: Since the lower diagonal entries of T_m do not vanish, T_m is nonderogatory, i.e., for each eigenvalue μ_j the associated eigenspace is one-dimensional, see [HJ85, Section 3.1]. Then, $\sum_{j=1}^r n_j = m$, which implies that the degree of p_t is exactly $m-1$.

In the following we denote the full sequence of the m eigenvalues of T_m by $\lambda_1, \dots, \lambda_m$. By applying basic properties of the Krylov decomposition and imposed conditions on the numerical range of A we obtain

$$\operatorname{spec}(\sigma\, T_m) \subseteq W(\sigma\, T_m) \subseteq W(\sigma\, A) \subseteq \mathbb{C}_-. \tag{2.3.4}$$

[3] See also Section 2.6 below.

[4] In the sequel, the argument of $\delta_m(\cdot)$ is again denoted by t instead of s.

The following proposition is partially related to [CM97, Sec. 3] or [JL15]. Here, divided differences have to be understood in the general sense, i.e., in the confluent sense if multiple eigenvalues occur; for the detailed definition and properties see [Hig08, Section B.16].

Proposition 2.3.1. *Let $T_m \in \mathbb{C}^{m\times m}$ be an upper Hessenberg matrix with eigenvalues $\lambda_1, \ldots, \lambda_m$ and spec$(\sigma T_m) \subseteq \mathbb{C}_-$. Then the function $\delta_m(t)$ defined as in (2.3.1a), i.e.,*

$$\delta_m(t) = e_m^*\, \mathrm{e}^{\sigma\, t\, T_m} e_1 = \left(\mathrm{e}^{\sigma\, t\, T_m}\right)_{m,1},$$

satisfies

$$\delta_m(t) = f_t[\lambda_1, \ldots, \lambda_m]\gamma_m \le \frac{t^{m-1}}{(m-1)!}\,\gamma_m, \tag{2.3.5}$$

with γ_m from (2.2.5) and where $f_t[\lambda_1, \ldots, \lambda_m]$ is the $(m-1)$-th divided difference over spec(T_m) of the function f_t defined in (2.3.3).

Proof. We proceed from the Newton representation of the interpolant $p_t(\lambda)$ from (2.3.2),

$$p_t(\lambda) = \sum_{j=0}^{m-1} f_t[\lambda_1, \ldots, \lambda_{j+1}]\, \omega_j(\lambda),$$

with $\omega_j(\lambda) = (\lambda - \lambda_1)\cdots(\lambda - \lambda_j)$. From (2.2.3) and by definition of γ_m, it is obvious that the ω_j satisfy

$$e_m^*\, \omega_j(T_m)\, e_1 = \begin{cases} 0, & j = 0, \ldots, m-2, \\ \gamma_m, & j = m-1. \end{cases}$$

Together with (2.3.2) this shows that the identity claimed in (2.3.5) is valid:

$$\delta_m(t) = e_m^*\, \mathrm{e}^{\sigma\, t T_m}\, e_1 = e_m^*\, p_t(T_m) e_1 = \sum_{j=0}^{m-1} f_t[\lambda_1, \ldots, \lambda_{j+1}] e_m^*\, \omega_j(T_m) e_1 = f_t[\lambda_1, \ldots, \lambda_m]\gamma_m.$$

According to [Hig08, (B.28)] the divided difference can be estimated by

$$|f_t[\lambda_1, \ldots, \lambda_m]| \le \frac{\max_{z\in\Omega} D^{(m-1)} f_t(z)}{(m-1)!}$$
for convex $\Omega \subseteq \mathbb{C}$ which contains all eigenvalues λ_j.

With $D^{(m-1)} f_t(\lambda) = (\sigma\, t)^{m-1}\mathrm{e}^{\sigma\, t\lambda}$, $|\sigma| = 1$ and $\mathrm{Re}(\sigma\lambda_j) \le 0$ we obtain

$$|f_t[\lambda_1, \ldots, \lambda_m]| \le \frac{t^{m-1}}{(m-1)!},$$

which implies the estimate (2.3.5) for $\delta_m(t)$. □

Error estimate and asymptotical correctness. Now we apply Proposition 2.3.1 in the context of our Krylov approximation.

Theorem 2.3.2 (Computable upper bound)**.** *For the nonexpansive case the error $L_m(t)v$ of the Krylov approximation* (2.2.6) *to $E(t)v$ satisfies*

$$\|L_m(t)v\|_2 \leq \tau_{m+1,m}\gamma_m \frac{t^m}{m!} \tag{2.3.6}$$

with $\tau_{m+1,m}$ from (2.2.2) *and γ_m from* (2.2.5)*.*

Proof. We proceed from (2.3.1). For δ_m defined in (2.3.1a), Proposition 2.3.1 implies

$$|\delta_m(s)| \leq \frac{s^{m-1}}{(m-1)!}\gamma_m,$$

and this gives an upper bound for the error integral (2.3.1b):

$$\|L_m(t)v\|_2 \leq \tau_{m+1,m}\gamma_m \int_0^t \frac{s^{m-1}}{(m-1)!}\,\mathrm{d}s = \tau_{m+1,m}\gamma_m \frac{t^m}{m!},$$

which completes the proof.

Proposition 2.3.1 is applied here in the nonexpansive case ($W(\sigma A) \subseteq \mathbb{C}_-$) which implies the requirement $\operatorname{spec}(\sigma T_m) \subseteq \mathbb{C}_-$, see (2.3.4).

□

The upper bound (2.3.6) corresponds to the 2-norm of the leading error term (2.2.12) according to Proposition 2.2.2. It is easily computable from the Krylov decomposition (2.2.2). We denote the error estimate given by (2.3.6) as

$$\mathrm{Err}_a = \tau_{m+1,m}\gamma_m \frac{t^m}{m!}. \tag{Err_a}$$

Proposition 2.3.3 (Asymptotical correctness)**.** *The upper bound* (2.3.6) *is asymptotically correct for $t \to 0$, i.e.,*

$$\|L_m(t)v\|_2 = \tau_{m+1,m}\gamma_m \frac{t^m}{m!} + \mathcal{O}(t^{m+1}). \tag{2.3.7}$$

Proof. The asymptotic estimate

$$\begin{aligned}\left|\|L_m(t)v\|_2 - \tau_{m+1,m}\gamma_m \frac{t^m}{m!}\right| &= \left|\|L_m(t)v\|_2 - \left\|\tau_{m+1,m}\gamma_m \frac{(\sigma\,t)^m}{m!}v_{m+1}\right\|_2\right| \\ &\leq \left\|L_m(t)v - \tau_{m+1,m}\gamma_m \frac{(\sigma\,t)^m}{m!}v_{m+1}\right\|_2 = \mathcal{O}(t^{m+1})\end{aligned}$$

is valid due to Proposition 2.2.2, and this proves (2.3.7). □

Remark 2.3.4. *In [vdEH06, Section 4] a defect-based error formulation is given for the shift-and-invert Krylov approximation of the matrix exponential function. In contrast to the standard Krylov method, the defect is not of order $m-1$ for $t \to 0$ there. Hence, our new results do not directly apply to shift-and-invert Krylov approximations. A study of a posteriori error estimates for the shift-and-invert approach is a topic of future investigations.*

2.4 Krylov approximation to φ-functions.

As another application we consider the so-called φ-functions, with power series representation

$$\varphi_p(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+p)!}, \quad p \geq 0. \tag{2.4.1a}$$

We have $\varphi_0(z) = \mathrm{e}^z$, and

$$\varphi_p(z) = \frac{1}{(p-1)!} \int_0^1 (1-\theta)^{p-1} \mathrm{e}^{\theta z} \, \mathrm{d}\theta, \quad p \geq 1. \tag{2.4.1b}$$

As the matrix exponential, φ-functions of matrices also appear in a wide range of applications, such as exponential integrators, see for instance [AMH11, HLS98, HO10, NW12, Sid98]. Krylov approximation is a common technique to evaluate φ-functions of matrices applied to a starting vector,

$$\varphi_p(\sigma\, tA)v \approx V_m \varphi_p(\sigma\, tT_m)e_1, \quad p \geq 0. \tag{2.4.2}$$

Since φ-functions are closely related to the matrix exponential, our ideas can be applied to these as well. We use the following notation for the error in the φ-functions:

$$L_m^p(t)v = \varphi_p(\sigma\, t\, A)v - V_m \varphi_p(\sigma\, t\, T_m)e_1. \tag{2.4.3}$$

With (2.4.3) we generalize the previously used notation: $L_m(t) = L_m^0(t)$.

Theorem 2.4.1. *The error of the Krylov approximation* (2.4.2) *to* $\varphi_p(\sigma\, tA)v$ *with* $p \geq 0$ *satisfies*

$$L_m^p(t)v = \tau_{m+1,m}\, \gamma_m \frac{(\sigma\, t)^m}{(m+p)!}\, v_{m+1} + \mathcal{O}(t^{m+1}). \tag{2.4.4a}$$

Furthermore, in the nonexpansive case its norm is bounded by

$$\|L_m^p(t)v\|_2 \leq \tau_{m+1,m}\, \gamma_m \, \frac{t^m}{(m+p)!}, \tag{2.4.4b}$$

and this bound is asymptotically correct for $t \to 0$.

Proof. For $p = 0$ the result directly follows from Propositions 2.2.2, 2.3.3 and Theorem 2.3.2. We now assume $p \geq 1$. Via the series representation (2.4.1a) of φ_p we can determine the leading term of the error in an analogous way as in Proposition 2.2.2:

$$\begin{aligned} \varphi_p(\sigma\, tA)v - V_m \varphi_p(\sigma\, tT_m)e_1 &= \frac{(\sigma\, t)^m\, (A^m v - V_m T_m^m e_1)}{(m+p)!} + \mathcal{O}(t^{m+1}) \\ &= \tau_{m+1,m}\, \gamma_m \frac{(\sigma\, t)^m}{(m+p)!}\, v_{m+1} + \mathcal{O}(t^{m+1}), \end{aligned}$$

which proves (2.4.4a).

Furthermore, proceeding from (2.4.1b) we obtain

$$\begin{aligned}\varphi_p(\sigma\, tA)v - V_m\varphi_p(\sigma\, tT_m)e_1 &= \frac{1}{(p-1)!}\int_0^1 (1-\theta)^{p-1}\big(\mathrm{e}^{\sigma\,\theta\, tA}v - V_m\,\mathrm{e}^{\sigma\,\theta\, tT_m}e_1\big)\,\mathrm{d}\theta\\ &= \frac{1}{(p-1)!}\int_0^1 (1-\theta)^{p-1}L_m(\theta\, t)v\,\mathrm{d}\theta,\end{aligned}$$

with the error $L_m(t)v$ for the matrix exponential case. Now we apply Theorem 2.3.2 to obtain

$$\begin{aligned}\|\varphi_p(\sigma\, tA)v - V_m\varphi_p(\sigma\, tT_m)e_1\|_2 &\le \frac{1}{(p-1)!}\int_0^1 (1-\theta)^{p-1}\|L_m(\theta\, t)v\|_2\,\mathrm{d}\theta\\ &\le \tau_{m+1,m}\,\gamma_m\,\frac{t^m}{(p-1)!\,m!}\int_0^1 (1-\theta)^{p-1}\theta^m\,\mathrm{d}\theta\\ &= \tau_{m+1,m}\,\gamma_m\,\frac{t^m}{(m+p)!},\end{aligned}$$

which proves (2.4.4b). □

2.5 Corrected Krylov approximation to the exponential and φ-functions.

Let us recall the well-known error representation given in [Saa92].

Proposition 2.5.1. *see [Saa92, Theorem 5.1] With the φ-functions defined in* (2.4.1)*, the error* (2.2.7) *can be represented in the form*

$$L_m(t)v = \tau_{m+1,m}\,\sigma\, t\sum_{j=1}^{\infty} e_m^*\varphi_j(\sigma\, tT_m)\,e_1(\sigma\, tA)^{j-1}v_{m+1}. \tag{2.5.1}$$

In [Saa92] it is stated that, typically, the first term of the sum given in Proposition 2.5.1, formula (Err_1), is already a good approximation to $L_m(t)v$. Analogously to [Saa92, Section 5.2] we use the notation Err_1 for the norm of this term,

$$\mathrm{Err}_1 = \tau_{m+1,m}\,t\,|e_m^*\varphi_1(\sigma\, tT_m)e_1|. \tag{Err_1}$$

In [JL15] it is even shown that Err_1 is an upper bound up to a factor depending on spectral properties of the matrix A. For the case of Hermitian σA we show $\|L_m(t)v\|_2 \le \mathrm{Err}_1$ in Proposition 2.5.5 below.

In Remark 2.5.3 below we show that Err_1 is also an asymptotically correct approximation to the error norm (in the sense of Proposition 2.3.3). Furthermore, the error estimate Err_1 is computable at nearly no extra cost, see [Saa92, Proposition 2.1].

According to [Saa92, Proposition 2.1], $\varphi_1(\sigma\, tT_m)$ can be computed from the extended matrix

$$T_m^+ = \begin{bmatrix} T_m & 0\\ \tau_{m+1,m}\,e_m^* & 0\end{bmatrix} \in \mathbb{C}^{(m+1)\times(m+1)} \tag{2.5.2a}$$

as

$$\mathrm{e}^{\sigma\, t T_m^+} e_1 = \begin{bmatrix} \mathrm{e}^{\sigma\, t T_m} e_1 \\ \tau_{m+1,m}\, \sigma\, t\, (e_m^*\, \varphi_1(\sigma\, t T_m)\, e_1) \end{bmatrix} \in \mathbb{C}^{m+1}. \tag{2.5.2b}$$

Equation (2.5.2b) can be used to evaluate the error estimate Err_1 or a corrected Krylov approximation in the form

$$S_m^+(t)v = V_m^+\, \mathrm{e}^{\sigma\, t T_m^+} e_1 \quad \text{with} \quad V_m^+ = \left[V_m \,\middle|\, v_{m+1} \right] \in \mathbb{C}^{n\times(m+1)}, \tag{2.5.3}$$

for which the first term of the error expansion according to Proposition 2.5.1 vanishes, see [Saa92]. For the error of the corrected Krylov approximation we use the notation

$$L_m^+(t)v = E(t)v - S_m^+(t)v.$$

For general φ-functions we obtain an error representation similar to Proposition 2.5.1 and a corrected Krylov approximation to φ-functions. The corrected Krylov approximation to $\varphi_p(\sigma\, t\, A)v$ is given in [Sid98, Theorem 2]:

$$\varphi_p(\sigma\, t\, A)v \approx V_m^+ \varphi_p(\sigma\, t\, T_m^+) e_1$$

with T_m^+ and V_m^+ given in (2.5.2a) and (2.5.3). The error of the corrected Krylov approximation is denoted by

$$L_m^{p,+}(t)v = \varphi_p(\sigma\, t\, A)v - V_m^+ \varphi_p(\sigma\, t\, T_m^+) e_1. \tag{2.5.4}$$

Proposition 2.5.2 (see [Sid98, Theorem 2])**.** *The error of the Krylov approximation $L_m^p(t)v$, see (2.4.3), satisfies*

$$L_m^p(t)v = \tau_{m+1,m}\sigma\, t \sum_{j=p+1}^{\infty} (e_m^* \varphi_j(\sigma\, t\, T_m) e_1)\, (\sigma\, t\, A)^{j-p-1} v_{m+1}. \tag{2.5.5a}$$

The error of the corrected Krylov approximation $L_m^{p,+}(t)v$, see (2.5.4), is given by

$$L_m^{p,+}(t)v = \tau_{m+1,m}\sigma\, t \sum_{j=p+2}^{\infty} (e_m^* \varphi_j(\sigma\, t\, T_m) e_1)\, (\sigma\, t\, A)^{j-p-1} v_{m+1}. \tag{2.5.5b}$$

The following remark will be used later on.

Remark 2.5.3. *From the representation (2.4.1a) for the φ_j together with (2.2.3) and (2.2.5) we observe*

$$e_m^* \varphi_j(\sigma t T_m) e_1 = \frac{(\sigma t)^{m-1} e_m^* T_m^{m-1}\, e_1}{(m-1+j)!} + \mathcal{O}(t^m) = \gamma_m \frac{(\sigma t)^{m-1}}{(m-1+j)!} + \mathcal{O}(t^m). \tag{2.5.6}$$

By (2.5.6) we observe $e_m^* \varphi_j(\sigma\, t T_m)\, e_1 = \mathcal{O}(t^{m-1})$ for $j \geq 0$ and we conclude that the asymptotically leading order term of $L_m^p(t)v$ for $t \to 0$ is obtained by the leading term ($j = p+1$) of the series (2.5.5a):

$$L_m^p(t)v = \tau_{m+1,m}\sigma\, t (e_m^* \varphi_{p+1}(\sigma\, t\, T_m) e_1) v_{m+1} + \mathcal{O}(t^m). \tag{2.5.7a}$$

Analogously we obtain the asymptotically leading order term of $L_m^{p,+}(t)v$ *for* $t \to 0$ *by the leading term* ($j = p+2$) *of the series* (2.5.5b)*:*

$$L_m^{p,+}(t)v = \tau_{m+1,m}(\sigma\, t)^2(e_m^*\varphi_{p+2}(\sigma\, t\, T_m)e_1)\, Av_{m+1} + \mathcal{O}(t^{m+1}). \tag{2.5.7b}$$

The asymptotically leading terms in (2.5.7a) *and* (2.5.7b) *can be used as error estimators:*

$$\|L_m^p(t)v\|_2 \approx \tau_{m+1,m}\, t|e_m^*\varphi_{p+1}(\sigma\, t\, T_m)e_1| \tag{2.5.8a}$$

and

$$\|L_m^{p,+}(t)v\|_2 \approx \|Av_{m+1}\|_2\tau_{m+1,m}\, t^2|e_m^*\varphi_{p+2}(\sigma\, t\, T_m)e_1|. \tag{2.5.8b}$$

The error estimators (2.5.8a) *and* (2.5.8b) *are already suggested in [Sid98, NW12]. We will refer to them as* Err_1 *in the context of the* φ*-functions with standard and corrected Krylov approximation, generalizing the corresponding quantities for the exponential case* $p = 0$*.*

We also obtain true upper bounds for the matrix exponential ($p = 0$) and general φ-functions with $p \geq 1$.

Theorem 2.5.4. *The error of the corrected Krylov approximation* (2.5.4) *to* $\varphi_p(\sigma\, tA)v$ *with* $p \geq 0$ *satisfies*

$$L_m^{p,+}(t)v = \tau_{m+1,m}\,\gamma_m\frac{(\sigma\, t)^{m+1}}{(m+p+1)!}\, Av_{m+1} + \mathcal{O}(t^{m+2}). \tag{2.5.9a}$$

Furthermore, in the nonexpansive case its norm is bounded by

$$\|L_m^{p,+}(t)v\|_2 \leq \|Av_{m+1}\|_2\,\tau_{m+1,m}\,\gamma_m\frac{t^{m+1}}{(m+p+1)!}, \tag{2.5.9b}$$

and this bound is asymptotically correct for $t \to 0$*.*

Proof. Applying (2.5.6) (with $j = p+2$) to (2.5.7b) shows (2.5.9a):

$$L_m^{p,+}(t)v = \tau_{m+1,m}\,\gamma_m\,\frac{(\sigma\, t)^{m+1}}{(m+p+1)!}\, Av_{m+1} + \mathcal{O}(t^{m+2}).$$

From Proposition 2.5.2 we observe

$$L_m^{p,+}(t)v = \sigma\, t\, A\; L_m^{p+1}(t)v.$$

Using the integral representation analogously as in the proof of Theorem 2.4.1 for $L_m^{p+1}(t)v$ and formula (2.2.8) for $L_m(t)v$, we obtain

$$\begin{aligned} L_m^{p,+}(t)v = \sigma\, t\, A\; L_m^{p+1}(t)v &= \tau_{m+1,m}\sigma\, t\frac{1}{p!}\int_0^1 (1-\theta)^p A\, L_m(\theta\, t)v\,\mathrm{d}\theta \\ &= \tau_{m+1,m}\sigma\, t\frac{1}{p!}\int_0^1 (1-\theta)^p \int_0^{\theta t} \mathrm{e}^{\sigma(\theta t - s)A} Av_{m+1}\delta_m(s)\,\mathrm{d}s\,\mathrm{d}\theta. \end{aligned}$$

With norm inequalities (note the nonexpansive case) and Proposition 2.3.1 we obtain

$$\begin{aligned}\|L_m^{p,+}(t)v\|_2 &\le \tau_{m+1,m}\, t\, \|Av_{m+1}\|_2 \frac{1}{p!}\int_0^1 (1-\theta)^p \int_0^{\theta t} |\delta_m(s)|\,\mathrm{d}s\,\mathrm{d}\theta \\ &\le \|Av_{m+1}\|_2 \tau_{m+1,m}\gamma_m\, t^{m+1}\,\frac{1}{p!\,m!}\int_0^1 (1-\theta)^p\theta^m\,\mathrm{d}s\,\mathrm{d}\theta \\ &= \|Av_{m+1}\|_2 \tau_{m+1,m}\gamma_m \frac{t^{m+1}}{(m+p+1)!},\end{aligned}$$

which proves (2.5.9b). Proposition 2.3.1 is applied here in the nonexpansive case, see also the proof of Theorem 2.3.2. □

If the error estimate (2.5.9b) is to be evaluated, the effort of the computation of $\|Av_{m+1}\|_2$ is comparable to one additional step of the Krylov iteration.

As mentioned before, we also can show that for Hermitian σA the estimate Err_1 gives a true upper bound:

Proposition 2.5.5. *For the nonexpansive case with $\sigma = 1$ and a Hermitian matrix A we obtain*

$$|\delta_m(t)| = \delta_m(t) > 0 \quad \textit{for} \quad t > 0.$$

This leads to the following upper bounds for the errors L_m^p and $L_m^{p,+}$ with $p \ge 0$:

$$\|L_m^p(t)v\|_2 \le \tau_{m+1,m}\, t\, \underbrace{e_m^*\,\varphi_{p+1}(tT_m)\,e_1}_{\ge 0} \tag{2.5.10a}$$

and

$$\|L_m^{p,+}(t)v\|_2 \le \|Av_{m+1}\|_2\, \tau_{m+1,m}\, t^2\, \underbrace{e_m^*\,\varphi_{p+2}(tT_m)\,e_1}_{\ge 0}\,. \tag{2.5.10b}$$

Proof. For a Hermitian matrix A we obtain a symmetric, tridiagonal matrix T_m with distinct, real eigenvalues via Lanczos approximation, see [HJ85, Chap. 3.1]. By Proposition 2.3.1 we observe

$$\delta_m(t) = f_t[\lambda_1,\ldots,\lambda_m]\gamma_m$$

with $f_t(\lambda) = \mathrm{e}^{t\,\lambda}$ for the case $\sigma = 1$. For divided differences of real-valued functions over real nodes we obtain $f_t[\lambda_1,\ldots,\lambda_m] \in \mathbb{R}$ and

$$f_t[\lambda_1,\ldots,\lambda_m] = \frac{D^{(m-1)}f_t(\xi)}{(m-1)!} = \frac{t^{m-1}\mathrm{e}^{t\xi}}{(m-1)!} \quad \text{for} \quad \xi \in [\lambda_1,\lambda_m]. \tag{2.5.11}$$

Equation (2.5.11) shows $f_t[\lambda_1,\ldots,\lambda_m] > 0$ and with $\gamma_m > 0$ we conclude

$$\delta_m(t) > 0, \quad \text{and} \quad |\delta_m(t)| = \delta_m(t).$$

We continue with (2.5.10a) in the case $p = 0$:

$$\|L_m(t)v\|_2 \le \tau_{m+1,m}\int_0^t |\delta_m(s)|\,\mathrm{d}s = \tau_{m+1,m}\int_0^t e_m^*\,\mathrm{e}^{sT_m}\,e_1\,\mathrm{d}s = \tau_{m+1,m}t\,e_m^*\,\varphi_1(tT_m)\,e_1.$$

For the case $p \geq 1$ we start analogously to Theorem 2.4.1. Using definition (2.4.1b) for the φ-functions and resorting to the case $p = 0$ we find

$$\begin{aligned}\|L_m^p(t)v\|_2 &\leq \frac{1}{(p-1)!}\int_0^1 (1-\theta)^{p-1}\|L_m(\theta\, t)v\|_2\,\mathrm{d}\theta \\ &\leq \tau_{m+1,m}\, t\,\frac{1}{(p-1)!}\int_0^1 (1-\theta)^{p-1}\theta\, e_m^*\,\varphi_1(\theta\, t\, T_m)\, e_1\,\mathrm{d}\theta.\end{aligned}$$

Evaluation of the integral yields

$$\begin{aligned}\|L_m^p(t)v\|_2 &\leq \tau_{m+1,m}\, t\,\frac{1}{(p-1)!}\int_0^1 (1-\theta)^{p-1}\theta\, e_m^*\,\varphi_1(\theta\, t\, T_m)\, e_1\,\mathrm{d}\theta \\ &= \tau_{m+1,m}\, t\sum_{k=0}^{\infty}\frac{e_m^*\,(t\,T_m)^k\, e_1}{(p-1)!\,(k+1)!}\int_0^1 (1-\theta)^{p-1}\theta^{k+1}\,\mathrm{d}\theta \\ &= \tau_{m+1,m}\, t\sum_{k=0}^{\infty}\frac{e_m^*\,(t\,T_m)^k\, e_1}{(p+k+1)!} \\ &= \tau_{m+1,m}\, t\, e_m^*\varphi_{p+1}(t\,T_m)\, e_1.\end{aligned} \tag{2.5.12}$$

This shows (2.5.10a). To show (2.5.10b) we start analogously to Theorem 2.5.4:

$$\|L_m^{p,+}(t)v\|_2 = \|t\,A\,L_m^{p+1}(t)v\|_2 \leq \|Av_{m+1}\|_2\,\tau_{m+1,m}\,t\,\frac{1}{p!}\int_0^1 (1-\theta)^p\int_0^{\theta t}|\delta_m(s)|\,\mathrm{d}s\,\mathrm{d}\theta.$$

Using $|\delta_m(s)| = \delta_m(s)$ and evaluating the inner integral by the φ_1 function, we obtain

$$\|L_m^{p,+}(t)v\|_2 \leq \|Av_{m+1}\|_2\,\tau_{m+1,m}\,t^2\,\frac{1}{p!}\int_0^1 (1-\theta)^p\theta\, e_m^*\,\varphi_1(\theta\, t\, T_m)\, e_1\,\mathrm{d}\theta.$$

Evaluation of the integral analogously to (2.5.12),

$$\|L_m^{p,+}(t)v\|_2 \leq \|Av_{m+1}\|_2\,\tau_{m+1,m}\,t^2\, e_m^*\varphi_{p+2}(t\,T_m)\, e_1,$$

completes the proof. □

2.6 Defect-based quadrature error estimates revisited

The term on the right-hand side of (2.2.12) is a computable error estimate, which has been investigated more closely in Section 2.3. It can also be interpreted in an alternative way. To this end we again proceed from the integral representation (2.2.8),

$$L_m(t)v = \int_0^t \underbrace{E(t-s)\,D_m(s)}_{=:\,\Theta_m(s,t)}\,v\,\mathrm{d}s. \tag{2.6.1}$$

Due to $\|D_m(t)v\| = \mathcal{O}(t^{m-1})$,

$$\tfrac{\mathrm{d}^j}{\mathrm{d}\,s^j}\,D_m(s)v\big|_{s=0} = 0, \quad j = 0,\ldots,m-2,$$

and the same is true for the integrand in (2.6.1),

$$\tfrac{\partial^j}{\partial s^j}\,\Theta_m(s,t)v\big|_{s=0} = 0, \quad j = 0,\dots,m-2.$$

Analogously as in [AKT14], this allows us to approximate (2.6.1) by a Hermite quadrature formula in the form

$$\int_0^t \Theta_m(s,t)v\,\mathrm{d}s \approx \frac{t}{m}\,\Theta_m(t,t)v = \frac{t}{m}\,D_m(t)v. \tag{2.6.2}$$

From (2.2.10),

$$\frac{t}{m}\,D_m(t)v = \tau_{m+1,m}\,\gamma_m\frac{(\sigma t)^m}{m!}\,v_{m+1} + \mathcal{O}(t^{m+1}),$$

which is the same as (2.2.12). This means that the quadrature approximation (2.6.2) approximates the leading error term in an asymptotically correct way.

From (2.6.2), (2.2.9) and (2.3.1a) we obtain

$$\|L_m(t)v\|_2 \approx \tau_{m+1,m}\frac{t}{m}|\delta_m(t)|. \tag{2.6.3}$$

The quadrature error in (2.6.2) is $\mathcal{O}(t^{m+1})$. It is useful to argue this also in a direct way: By construction, the Hermite quadrature formula underlying (2.6.2) is of order m, and its error has the Peano representation (cf. also [AKT14])

$$\frac{t}{m}\,\Theta_m(t,t) - \int_0^t \Theta_m(s,t)v\,\mathrm{d}s = \int_0^t \frac{s\,(t-s)^{m-1}}{m!}\,\tfrac{\partial^m}{\partial s^m}\,\Theta_m(s,t)v\,\mathrm{d}s. \tag{2.6.4}$$

Here, $\frac{\partial^m}{\partial s^m}\,\Theta_m(s,t)v = \mathcal{O}(1)$, because $\frac{\mathrm{d}^m}{\mathrm{d}s^m}\,D_m(s)v = \mathcal{O}(1)$ which follows from $D_m(s)v = \mathcal{O}(s^{m-1})$. This shows that, indeed, the quadrature error (2.6.4) is $\mathcal{O}(t^{m+1})$. Furthermore, a quadrature formula of order $m+1$ can be constructed by including an additional evaluation of

$$\tfrac{\partial}{\partial s}\,\Theta_m(s,t)v\big|_{s=t} = D_m^{[2]}(t)v, \quad \text{with} \quad D_m^{[2]}(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}D_m(t) - \sigma A\,D_m(t).$$

A routine calculation shows

$$\int_0^t \Theta_m(s,t)v\,\mathrm{d}s = \frac{2\,t}{m+1}\,D_m(t)v - \frac{t^2}{m(m+1)}\,D_m^{[2]}(t)v + \mathcal{O}(t^{m+2}), \tag{2.6.5}$$

where the error depends on $\frac{\mathrm{d}^{m+1}}{\mathrm{d}s^{m+1}}\,D_m(s)v = \mathcal{O}(1)$. This may be considered as an improved error estimate[5] which can be evaluated using

$$\tfrac{\mathrm{d}}{\mathrm{d}t}D_m(t)v = \sigma^2\tau_{m+1,m}\,e_m^*(T_m\,\mathrm{e}^{\sigma\,t\,T_m})e_1 v_{m+1}.$$

With the solution in the Krylov subspace, $\mathrm{e}^{\sigma\,t\,T_m}e_1$ with $e_m^*\mathrm{e}^{\sigma\,t\,T_m}e_1 = (\mathrm{e}^{\sigma\,t\,T_m}e_1)_m$, we can compute the derivative of the defect at $\mathcal{O}(1)$ cost,

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}D_m(t)v &= \sigma^2\tau_{m+1,m}\,e_m^*(T_m\,\mathrm{e}^{\sigma\,t\,T_m})e_1 v_{m+1}\\ &= \sigma^2\tau_{m+1,m}\,\big((T_m)_{m,m}\,(\mathrm{e}^{\sigma\,t\,T_m}e_1)_m + (T_m)_{m,m-1}\,(\mathrm{e}^{\sigma\,t\,T_m}e_1)_{m-1}\big)v_{m+1}.\end{aligned}$$

[5] In the setting of [AKT14] (higher-order splitting methods) such an improved error estimate was not taken into account since it cannot be evaluated with reasonable effort in that context.

Also longer expansions may be considered, for instance

$$\int_0^t \Theta_m(s,t)v\,\mathrm{d}s = \frac{3\,t}{m+2}\,D_m(t)v - \frac{3\,t^2}{(m+1)(m+2)}\,D_m^{[2]}(t)v + \frac{t^3}{m(m+1)(m+2)}\,D_m^{[3]}(t)v + \mathcal{O}(t^{m+3}), \quad \text{with} \quad D_m^{[3]}(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}D_m^{[2]}(t) - A\,D_m^{[2]}(t),$$

etc.

This alternative way of computing improved error estimates is worth investigating but will not be pursued further here.

Quadrature estimate for (2.3.1b) **revisited.** We proceed from (2.3.1b) which is valid for the nonexpansive case. In [Lub08] it is suggested to use the right-endpoint rectangle rule as a practical approximation to the integral (2.3.1b),

$$\|L_m(t)v\|_2 \le \tau_{m+1,m}\int_0^t |\delta_m(s)|\,\mathrm{d}s \approx \tau_{m+1,m}\,t|\delta_m(t)|, \tag{2.6.6}$$

or alternatively the Simpson rule, which is also suggested in [WY17]. The error estimate in (2.6.6) is also referred to as *generalized residual estimate* in [HLS98, BGH13] and similar error estimates also appeared earlier in [Saa92]. In [DGK98, eq. (32)] an a priori upper bound on the integral in (2.6.6) is obtained by $t\max_{s\in[0,t]}|\delta_m(s)|$. Applying Hermite quadrature to (2.3.1b) also directly leads to the error estimate (2.6.3).

For a better understanding of the approximation (2.6.6) we consider the effective order of $|\delta_m(t)|$ as a function of t: Let us denote $f(t) := |\delta_m(t)|$ and assume $f(t) > 0$ in a sufficiently small interval $(0,T]$. For the Hermitian case this assumption is fulfilled for all $t > 0$, see Proposition 2.5.5. We define the *effective order* $\rho : (0,T] \to \mathbb{R}$ of f by

$$\rho(t) = \frac{f'(t)\,t}{f(t)}. \tag{2.6.7}$$

This definition is motivated by the slope of f in a double-logarithmic graph, i.e., the graph of

$$\xi(\tau) = \ln(f(\mathrm{e}^\tau)) \quad \text{for} \quad \tau = \ln t,$$

which, for example, corrsponds to the Matlab `loglog` plot of $f(t)$ over t. The slope in the double-logarithmic graph, i.e., the derivative of the auxiliary function ξ, satisfies

$$\xi'(\tau) = \frac{f'(\mathrm{e}^\tau)\,\mathrm{e}^\tau}{f(\mathrm{e}^\tau)},$$

and substituting $\tau = \ln t$ therein yields the effective order (2.6.7). The concept of the effective order generalizes the asymptotic order, which can be determined by the slope of $f(t)$ for $t \to 0$ in a double-logarithmic graph. In particular, we have $\rho(0+) = m-1$ for the effective order of the defect, which satisfies $\delta(t) = \mathcal{O}(t^{m-1})$ for $t \to 0$.

Assuming $\rho(t) > 0$ for $t \in (0,T]$, the effective order satisfies

$$f(t) = \frac{f'(t)\,t}{\rho(t)}.$$

Integration and application of the mean value theorem shows the existence of $t^* \in [0,t]$ with

$$\int_0^t f(s)\,\mathrm{d}s = \frac{1}{\rho(t^*)}\int_0^t f'(s)\,s\,\mathrm{d}s,$$

and integration by parts gives

$$\int_0^t |\delta_m(s)|\,\mathrm{d}s = \frac{t\,|\delta_m(t)|}{1+\rho(t^*)}.$$

With the plausible assumption that the order is bounded by $0 \le \tilde{m} \le \rho(t) \le m-1 = \rho(0+)$ for $t \in [0,T]$, we obtain

$$\tfrac{t}{m}\,|\delta_m(t)| \le \int_0^t |\delta_m(s)|\,\mathrm{d}s \le \tfrac{t}{\tilde{m}+1}\,|\delta_m(t)| \le t\,|\delta_m(t)|. \tag{2.6.8}$$

This shows that under such an assumption the generalized residual estimate (2.6.6) gives an upper bound on the error. With the assumption $0 \le \rho(t)$, the defect $|\delta_m(t)|$ (also called the residual in the literature) is monotonically increasing and the upper bound suggested in [DGK98, eq. (32)] is equivalent to the generalized residual estimate. However, in contrast to (2.6.3), the error estimate (2.6.6) is not asymptotically correct for $t \to 0$. In the following remark we suggest a practical approach to tighten the generalized residual estimate retaining the property of an upper bound in (2.6.8).

Remark 2.6.1. *With $\rho(0+) = m-1$ and the assumptions that the effective order is slowly decreasing locally at $t=0$ and sufficiently smooth, we suggest choosing $\tilde{m} = \rho(t)$ for a step of size t to improve the quadrature based estimate.*

$$\|L_m(t)v\|_2 \le \tau_{m+1,m}\int_0^t |\delta_m(s)|\,\mathrm{d}s \approx \tau_{m+1,m}\tfrac{t}{\rho(t)+1}\,|\delta_m(t)|. \tag{2.6.9}$$

We will refer to this as effective order quadrature *estimate.*

Substitute $f(t) = |\delta_m(t)| = \left((\mathrm{e}^{\sigma\,t\,T_m}e_1)_m(\mathrm{e}^{\bar{\sigma}\,t\,\bar{T}_m}e_1)_m\right)^{1/2}$ in the ansatz (2.6.7) to obtain a computable formula for the effective order $\rho(t)$,

$$\rho(t) = \frac{t\left(|\delta_m(t)|\right)'}{|\delta_m(t)|} = t\,Re\left(\sigma(T_m)_{m,m} + \sigma(T_m)_{m,m-1}\frac{(\mathrm{e}^{\sigma\,t\,T_m}e_1)_{m-1}}{(\mathrm{e}^{\sigma\,t\,T_m}e_1)_m}\right). \tag{2.6.10}$$

The computation of $(\mathrm{e}^{\sigma\,t\,T_m}e_1)_{m-1}$ usually comes hand in hand with the computation of $\delta_m(t) = (\mathrm{e}^{\sigma\,t\,T_m}e_1)_m$. For a numerical implementation of (2.6.10) we suggest computing $\mathrm{e}^{\sigma\,t\,T_m}e_1$ by a Taylor or Páde approximation.

In the limit $t \to 0$ this choice of quadrature is equivalent to the Hermite quadrature and, therefore, asymptotically correct.

Up to now we did refer to the effective order of the defect $|\delta_m(t)|$. For $t \to 0$ the effective order of the error is given by $\rho(t)+1$.

2.7 The matrix exponential as a time integrator.

For simplicity we assume the nonexpansive case of (2.2.1) in this section.

We recall from [HL97] that superlinear convergence as a function of m, the dimension of the underlying Krylov space, sets in for

$$t\,\|A\|_2 \lessapprox m. \tag{2.7.1}$$

This relation also affects the error considered as a function of time t. Equation (2.7.1) can be seen as a very rough estimate for a choice of t which leads to a systematic error and convergence behavior. Only for special classes of problems as for instance symmetric negative definite matrices, the relation (2.7.1) can be weakened, see [HL97, BR09] for details.

In general a large time step t would necessitate large m or a restart of the Krylov method. For larger dimensional problems memory issues can limit the choice of m and make a restart necessary. Considering global computational cost it may also be favorable to use a moderate value of m in combination with restarts. Even if increasing m results in a larger time step t, the increase in computational cost can lead to a decrease of total performance in some cases. This issue is relevant particularly for rather large choices of m, especially if computational cost scaling with m^2 or worse gets noticeable. We further discuss effects of computer arithmetic on the Krylov approximation of the matrix exponential in Section 2.8 without going into details.

For the matrix exponential seen as a time propagator, a simple restart is possible. The following procedure has been introduced in [Sid98] and is recapitulated here to fix the notation.

We split the time range $[0,t]$ into N subintervals,

$$\begin{aligned} 0 = t_0 < t_1 < \ldots < t_N = t, \\ \text{with step sizes} \quad \Delta t_j = t_j - t_{j-1}, \quad j = 1,\ldots,N. \end{aligned}$$

The exact solution at time t_j is denoted by $v^{[j]}$, whence

$$v^{[j]} = E(\Delta t_j)v^{[j-1]} = E(t_j)v, \quad \text{with} \quad v^{[0]} = v.$$

For simplicity we assume that the dimension m of the Krylov subspace is fixed over the substeps. We obtain approximations $w^{[j]}$ to $v^{[j]}$ by applying multiple restarted Krylov steps, with orthonormal bases $V_m^{[j]}$ and upper Hessenberg matrices $T_m^{[j]}$. Starting from $w^{[0]} = v$, for $j = 1,\ldots,N$,

$$w^{[j]} := S_m^{[j]}(\Delta t_j)w^{[j-1]} = V_m^{[j]}\mathrm{e}^{\sigma\,\Delta t_j T_m^{[j]}}\big(V_m^{[j]}\big)^* w^{[j-1]} = V_m^{[j]}\mathrm{e}^{\sigma\,\Delta t_j T_m^{[j]}}e_1.$$

The error matrix in the j-th step is denoted by

$$L_m^{[j]}(\Delta t_j) := E(\Delta t_j) - S_m^{[j]}(\Delta t_j),$$

and the accumulated error by

$$L_m^\star(t)v = v^{[N]} - w^{[N]}. \tag{2.7.2}$$

With

$$\begin{aligned} v^{[j]} - w^{[j]} &= E(\Delta t_j) v^{[j-1]} - S_m^{[j]}(\Delta t_j) w^{[j-1]} \\ &= E(\Delta t_j)(v^{[j-1]} - w^{[j-1]}) + L_m^{[j]}(\Delta t_j) w^{[j-1]} \end{aligned}$$

we obtain

$$L_m^\star(t)v = \sum_{j=1}^{N} E(\Delta t_N) \cdots E(\Delta t_{j+1})\, L_m^{[j]}(\Delta t_j) w^{[j-1]}.$$

Recall our premise that $E(\cdot)$ is nonexpansive and assume that the local error is bounded by

$$\|L_m^{[j]}(\Delta t_j) w^{[j]}\|_2 \leq \mathrm{tol} \cdot \Delta t_j. \tag{2.7.3}$$

Then,

$$\|L_m^\star(t)v\|_2 \leq \sum_{j=1}^{N} \|L_m^{[j]}(\Delta t_j) w^{[j-1]}\|_2 \leq \mathrm{tol} \sum_{j=1}^{N} \Delta t_j = \mathrm{tol} \cdot t.$$

The term $\|L_m^{[j]}(\Delta t_j) w^{[j]}\|_2$ denotes the truncation error of a single substep and is studied in the first part of this paper. We now apply local error estimates to predict acceptable time steps.

Step size control. For a single substep, the error estimate (Err_a) suggests a step size to satisfy a given error tolerance tol as

$$\Delta t_j = \left(\frac{\mathrm{tol}\ m!}{\tau_{m+1,m}^{[j]} \gamma_m^{[j]}} \right)^{1/m}, \quad j = 1, \ldots, N. \tag{2.7.4}$$

For a local error as in (2.7.3), we replace tol by (Δt_j tol) in (2.7.4) and obtain

$$\Delta t_j = \left(\frac{\mathrm{tol}\ m!}{\tau_{m+1,m}^{[j]} \gamma_m^{[j]}} \right)^{1/(m-1)}, \quad j = 1, \ldots, N. \tag{2.7.5}$$

We remark that Δt_j can be computed together with the construction of the Krylov subspace, therefore, $\tau_{m+1,m}^{[j]}$ and $\gamma_m^{[j]}$ are known values at this point. For the corrected Krylov approximation $S_m^+(t)v^{[j]}$, see (2.5.3), the error estimate given in (2.5.9b) ($p = 0$) suggests a local step size of

$$\Delta t_j = \left(\frac{\mathrm{tol}\ (m+1)!}{\left\| A v_{m+1}^{[j]} \right\|_2 \tau_{m+1,m}^{[j]} \gamma_m^{[j]}} \right)^{1/m}, \quad j = 1, \ldots, N. \tag{2.7.6}$$

The error estimator Err_1 and estimates given in Section 2.6 cannot be inverted directly to predict the step size. Computing a feasible step size is still possible via heuristic step size control. This approach will be formulated for a general error estimate Err. Ideas of heuristic step size control are given in [Gus91] in general and [Sid98] or [NW12] for a Krylov

approximation of the matrix exponential. For a step with step size Δt_{j-1} and estimated error $\mathrm{Err}^{[j-1]}$, $j = 2, \ldots, N$, a reasonable size for the subsequent step can be chosen as the solution of

$$\Delta t_j = \left(\frac{\Delta t_j \ \mathrm{tol}}{\mathrm{Err}^{[j-1]}}\right)^{1/m} \Delta t_{j-1} \quad \text{resulting in} \quad \Delta t_j = \left(\frac{\mathrm{tol}}{\mathrm{Err}^{[j-1]}}\right)^{1/(m-1)} \Delta t_{j-1}^{m/(m-1)}. \tag{2.7.7}$$

In (2.7.7) we only need the evaluation of the error estimate for the previously computed step with step size Δt_{j-1}. In the Expokit package [Sid98] the heuristic step size control is used similarly to (2.7.7) and an initial step size Δt_1 is chosen by an a priori choice, which we recall for comparison on numerical examples,

$$\Delta t_1 = \frac{1}{\|H\|_\infty} \left(\frac{\mathrm{tol}\ ((m+1)/\mathrm{e})^{m+1} \sqrt{2\,\pi\,(m+1)}}{4\,\|H\|_\infty}\right)^{1/m}. \tag{2.7.8}$$

In many cases the construction of the Krylov subspace, which is independent of the step size, contributes the largest part to the computational cost. In this case we can improve the choice of Δt_j relatively cheaply in an iterative manner before continuing to time step $j+1$:

$$\begin{aligned} \Delta t_{j,1} &:= \Delta t_{j-1} \ \text{ or result of (2.7.5)}, \\ \Delta t_{j,l} &:= \left(\frac{\mathrm{tol}}{\mathrm{Err}^{[j,l-1]}}\right)^{1/(m-1)} \Delta t_{j,l-1}^{m/(m-1)}, \qquad l = 2, \ldots, N_j, \\ \Delta t_j &:= \Delta t_{j,N_j}. \end{aligned} \tag{2.7.9}$$

By choosing $\mathrm{Err}^{[j,l-1]}$ as an error estimate for the Krylov approximation of the j-th step with time step $\Delta t_{j,l-1}$.

The aim of the iteration (2.7.9) is to determine a step size $\Delta t_{j,\infty}$ with $\mathrm{Err}^{[j,\infty]} = \Delta t_{j,\infty}\mathrm{tol}$, see (2.7.3). The convergence behavior of iteration (2.7.9) depends on the structure of the corresponding error estimate. The idea of the heuristic step size control is based on the asymptotic order of the error for $\Delta t \to 0$, which in (2.7.7) and (2.7.9) is assumed to be m, see (2.2.11). By substituting the asymptotic order m by the effective order $\rho(\Delta t)+1$, which is introduced in (2.6.10), the iteration (2.7.9) could be improved for a step size Δt away from the asymptotic regime. In our practical examples this iteration does not seem to be sensitive with respect to the effective order of the error and converges in a small number of steps using the asymptotic order m.

For the following remarks on T_m we neglect the index j in $T_m^{[j]}$ to simplify the notation. In the case of a Hermitian matrix A the matrix T_m is symmetric, tridiagonal and real-valued which allows cheap and robust computation of its eigenvalue decomposition. The eigenvalue decomposition of T_m is independent of the step size Δt and allows cheap evaluation of $\mathrm{e}^{\sigma\,\Delta t\,T_m} e_1$ or $\varphi_1(\sigma\,\Delta t\,T_m)e_1$ and corresponding error estimates for multiple choices of Δt.

For a non-Hermitian matrix A computing $\mathrm{Err}^{[j,l]}$ for multiple choices of l, hence different step sizes $\Delta t_{j,l}$, only leads to slightly larger computational cost, which is usually negligible.

2.8 Numerical considerations and examples

We give an illustration of our theoretical results for two different skew-Hermitian problems in Subsection 2.8.1, a Hermitian problem in Subsection 2.8.2, and a non-normal problem in Subsection 2.8.3. We also compare the performance of different error estimates for practical step size control (Section 2.7) in Subsection 2.8.1. To show that our error estimate (2.3.6) is efficient in practice we also compare it with results delivered by the standard package Expokit [Sid98] and a priori error estimates.

2.8.1 The skew-Hermitian case

For our tests we use different types of matrices.

Free Schrödinger equation. We consider

$$H = \tfrac{1}{4}\,\mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}, \tag{2.8.1}$$

with dimension $n = 10\,000$. The matrix H is associated with a finite difference or finite element discretization of the one-dimensional negative Laplacian. With $A = H$ and $\sigma = -\mathrm{i}$, in (2.2.1) we obtain the free Schrödinger equation. The eigenvalue decomposition of H is well known, and we can use the discrete sine transform with high precision arithmetic in Matlab to compute the exact solution $E(t)v$, see (2.2.1). The starting vector v is chosen randomly. To compute the Krylov subspace approximation $S_m(t)v$, see (2.2.6), we use the eigenvalue decomposition of the tridiagonal matrix T_m.

Discrete Hubbard model. For the description of the Hubbard model we employ a self-contained notation. The Hubbard model first appears in [Hub63] and was further used in many papers and books, e.g. [Mah93, PKvdBS16]. The Hubbard model is used to describe electron density on a given number of sites, which correspond to Wannier discretization of orbitals, and spin up or down. We consider the following Hubbard Hamiltonian, in second quantization and without chemical potential:

$$H = \frac{1}{2}\sum_{i,j,\sigma} v_{ij} c_{j\sigma}^{\dagger} c_{i\sigma} + \sum_{j,\sigma} U \hat{n}_{j\sigma}\hat{n}_{j\sigma'}, \tag{2.8.2}$$

where i, j sum over the number of sites n_{sites} and the spins $\sigma, \sigma' \in \{\uparrow, \downarrow\}$ where σ' is the opposite spin to σ. The entries v_{ij} with $i, j = 1, \ldots, n_{\text{sites}}$ describe electron hopping from site i to j. In (2.8.2), the notation $c_{j\sigma}^{\dagger} c_{i\sigma}$ describes the 2nd quantization operator and $\hat{n}_{j\sigma} = c_{j\sigma}^{\dagger} c_{j\sigma}$ the occupation number operator. For details on the notation in (2.8.2) we can recommend several references, e.g. [Hub63, Jaf08, Mah93, PKvdBS16].

For our tests we model 8 electrons at 8 sites ($n_{\text{sites}} = 8$) with spin up and down for each site, this leads to 16 possible states for electrons. Such an electron distribution is also referred to as half-filled in the literature. We also restrict our model by considering the number of electrons with spin up and down to be fixed as $n_{\text{sites}}/2$. This leads to $n = (\text{binomial}(8, 4))^2 = 4900$ considered occupation states which create a discrete basis. For the numerical implementation of the basis we consider 16-bit integers for which each

bit describes a position which is occupied in case the bit is equal to 1 or empty otherwise. The set of occupation states can be ordered by the value of the integers which leads to a unique representation of the Hubbard Hamiltonian (2.8.2) by a matrix $H \in \mathbb{C}^{n\times n}$. Such an implementation of the Hubbard Hamiltonian is also described in [Jaf08, Section 3].

In our test setting we use $U = 5$ and parameter-dependent values for electron hopping $v_{ij} = v_{ij}(\omega) \in \mathbb{C}$ with $\omega \in (0, 2\pi]$:

$$v_{11} = v_{88} = -1.75, \quad v_{jj} = -2 \quad \text{for} \quad j = 2, \dots, 7,$$
$$v_{j,j+1} = \bar{v}_{j+1,j} = -\cos\omega + \mathrm{i}\sin\omega \quad \text{for} \quad j = 1, \dots, 7 \quad \text{and} \quad v_{ij} = 0 \quad \text{otherwise.}$$

For this choice of $v_{ij}(\omega)$ we obtain a Hermitian matrix $H_\omega \in \mathbb{C}^{n\times n}$ with 43980 nonzero entries (for a general choice of ω) and $\operatorname{spec}(H_\omega) \subseteq (-19.1, 8.3)$. The spectrum of H_ω is independent of ω.

A relevant application where the Hubbard Hamiltonian (2.8.2) is of importance is the simulation of oxide solar cells with the goal of finding candidates for new materials promising a gain in the efficiency of the solar cell, see [Hel07]. The study of solar cells considers time-dependent electron hoppings $v_{ij} = v_{ij}(t)$ to model time-dependent potentials which lead to Hamiltonian matrices $H(t)$. The time-dependent Hamiltonian can be parameterized via ω. Time propagation of a linear, non-autonomous ODE system can be approximated by Magnus-type integrators which are based on one or more evaluations of matrix exponentials applied to different starting vectors at several times t, see for instance [BCOR09, BM05]. Our test setting for the Hubbard Hamiltonian with arbitrary ω is then obtained by (2.2.1) with the matrix $A = H_\omega$ as described above and $\sigma = -\mathrm{i}$.

In the following Subsection 2.8.1 we focus on the skew-Hermitian case. For tests on the Hermitian case see Subsection 2.8.2 below.

Verification of upper error bound. In the following Figures 2.1 and 2.2 we compare the error $\|L_m(t)v\|_2$ with the error estimates Err_1 and Err_a. Figure 2.1 refers to the matrix (2.8.1) of the free Schrödinger problem and Figure 2.2 to the Hubbard Hamiltonian (2.8.2) with $\omega = 0.123$. For both cases we show results with Krylov subspace dimensions $m = 10$ and $m = 30$, respectively.

We observe that the error estimate Err_1 is a good approximation to the error, but it is not an upper bound in general. In contrast, Err_a is a proven upper error bound. Up to round-off error, for $m = 10$ we observe the correct asymptotic behavior of Err_a and Err_1. For larger choices of m the asymptotic regime starts at time steps for which the error is already close to round-off precision. Therefore, for larger choices of m, the Krylov approximation, as a time integrator, cannot achieve its full order for typical time steps in double precision.

The matrix (2.8.1) has been scaled such that $\operatorname{spec}(H) \subseteq (0, 1)$ and $\|H\|_2 \approx 1$. In accordance with (2.7.1) stagnation of the error is observed for times $t \lessapprox m$, see Figure 2.1.

We verify the error estimates in the skew-Hermitian setting of the free Schrödinger equation (2.8.1) for the standard Krylov approximation of the φ_1 function in Figure 2.3 and the corrected Krylov approximation of the matrix exponential function in Figure 2.4. In Figure 2.3 the error estimator Err_1 refers to formula (2.5.8a) and Err_a shows the upper error bound (2.4.4b) from Theorem 2.4.1, both for the case $p = 1$. In Figure 2.4, Err_1 is

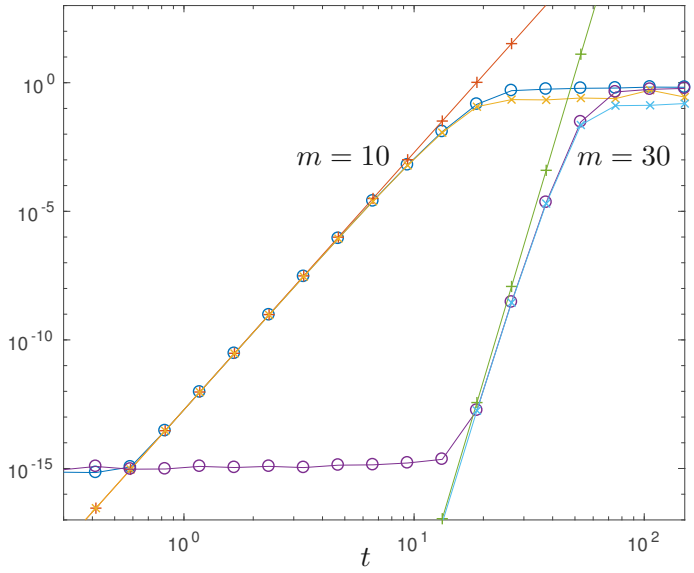


Figure 2.1: Error $\|L_m(t)v\|_2$ ($\circ$) and the error estimates Err_1 ($\times$) and Err_a ($+$) for the free Schrödinger problem and Krylov subspace dimensions $m = 10$ and $m = 30$. Err_a is an upper bound for the error, and both estimates show the correct asymptotical behavior. Due to round-off error, for $m = 30$ the observed effective order is less clear than for $m = 10$.

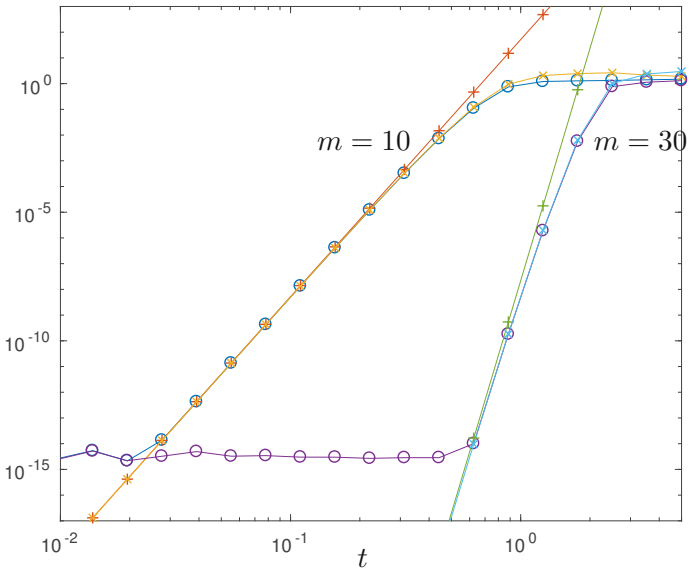


Figure 2.2: Error $\|L_m(t)v\|_2$ ($\circ$) and the error estimates Err_1 ($\times$) and Err_a ($+$) for the Hubbard Hamiltonian with $\omega = 0.123$ and Krylov subspace dimensions $m = 10$ and $m = 30$. This shows the same behavior as in Figure 2.1.

from formula (2.5.8b) and Err_a denotes the upper error bound (2.5.9b) from Theorem 2.5.4, both for the case $p = 0$.

Illustration of defect-based quadrature error estimates from Section 2.6. We first illustrate the performance of the estimates based on Hermite quadrature according to (2.6.3) and improved Hermite quadrature according to (2.6.5) for the Hubbard model, see Figure 2.5. Both estimates are asymptotically correct, whereas the improved quadrature (2.6.5) is slightly better for larger time steps t, with the drawback of one additional matrix-vector multiplication. (See Remark 2.8.2 below for cost efficiency of more expensive error estimates.)

Figure 2.6 refers to the generalized residual estimate (2.6.6), and estimates based on the effective order quadrature according to Remark 2.6.1, and the Hermite quadrature (2.6.3). For our test problems the assumptions from Section 2.6 on the defect and its effective order are satisfied for a significant range of values of t. We also observe that the inequalities (2.6.8) are satisfied. The effective order and Hermite quadrature estimates behave in an asymptotically correct way, while the generalized residual estimate leads to an upper error bound which is, however, not sharp for $t \to 0$.

For the skew-Hermitian case use $\sigma = -\mathrm{i}$ and $T_m \in \mathbb{R}^{m\times m}$ in (2.6.10) to obtain

$$\rho(t) = t\,(T_m)_{m-1,m}\,\mathrm{Re}\left(\frac{-\mathrm{i}\,(\mathrm{e}^{-\mathrm{i}\,t\,T_m}e_1)_{m-1}}{(\mathrm{e}^{-\mathrm{i}\,t\,T_m}e_1)_m}\right).$$

For computing the effective order we only consider time steps $\rho(t) > 0$, and where ρ

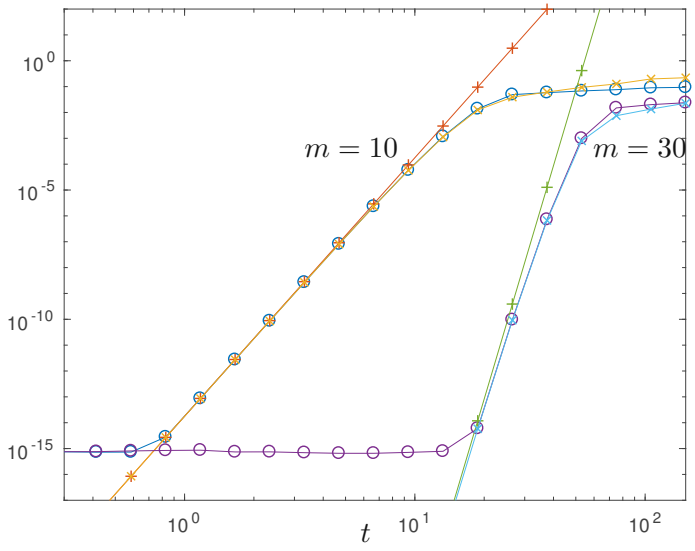


Figure 2.3: Error $\|L_m^1(t)v\|_2$ ($\circ$) and the error estimates Err_1 ($\times$) and Err_a ($+$) for the free Schrödinger problem and Krylov subspace dimension $m = 10$ and $m = 30$.

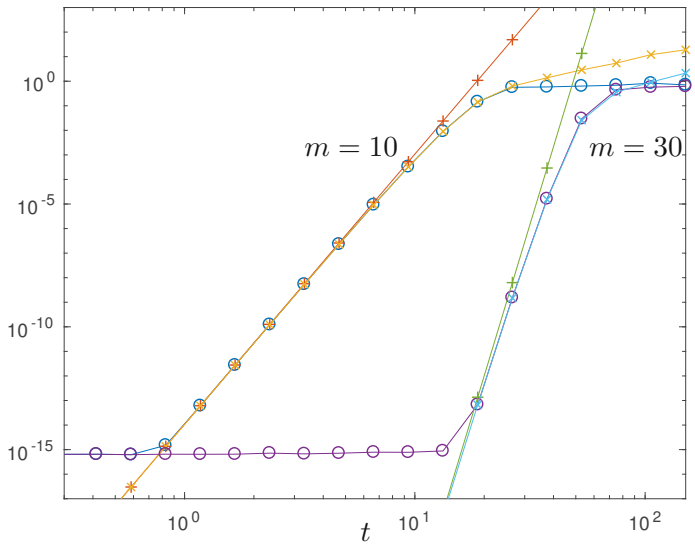


Figure 2.4: Error $\|L_m^+(t)v\|_2$ ($\circ$) and the error estimates Err_1 ($\times$) and Err_a ($+$) for the free Schrödinger problem and Krylov subspace dimension $m = 10$ and $m = 30$.

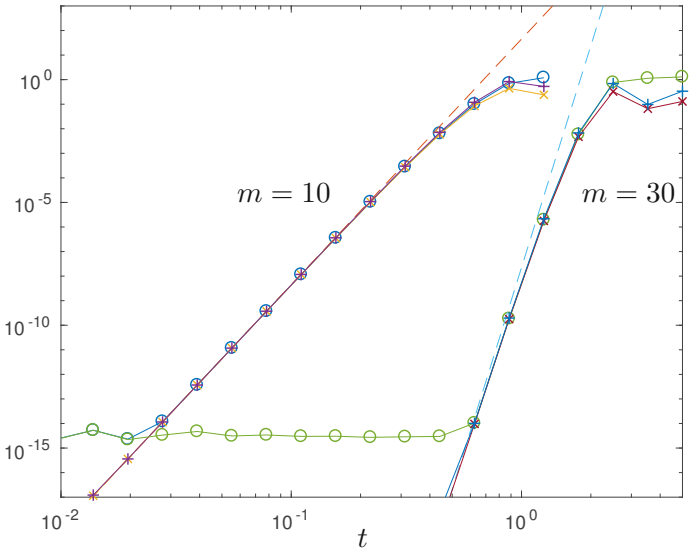


Figure 2.5: Error $\|L_m(t)v\|_2$ ($\circ$) and the error estimates based on the Hermite quadrature ($\times$) and improved Hermite quadrature ($+$), see (2.6.3) and (2.6.5), for the Hubbard Hamiltonian with $m = 10$ and $m = 30$. The dashed lines show the error estimate Err_a.

appears indeed to be monotonically decreasing over the computed discrete time steps. This restriction is compatible with our assumptions in Section 2.6.

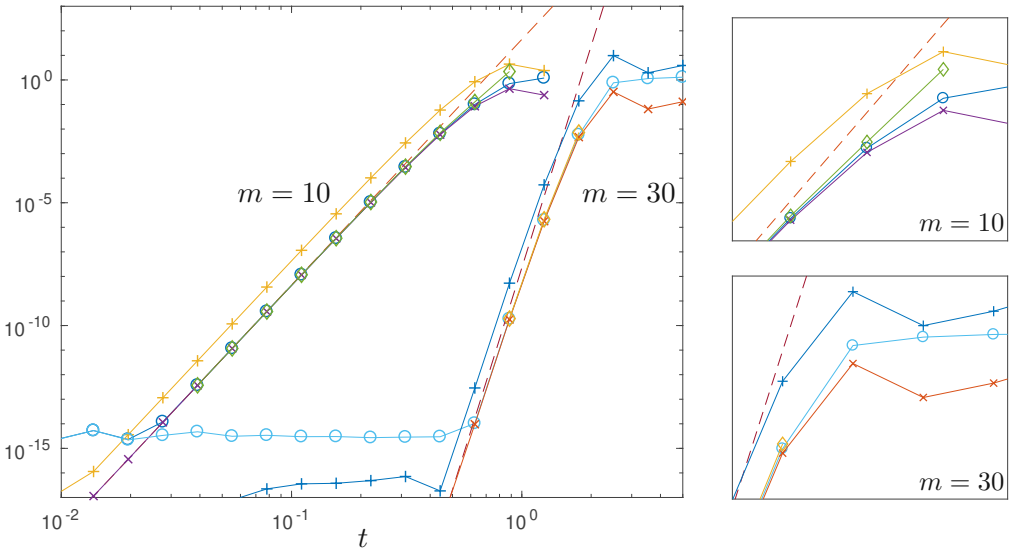
Corrected Krylov approximation and mass conservation. We remark that error estimates for the corrected Krylov approximation usually require one additional matrix-vector multiplication, and applying a standard Krylov approximation of dimension $m + 1$ seems to be a more favorable choice in our approach to error estimation.

The Krylov approximation of the matrix exponential conserves the mass for the skew-Hermitian case in contrast to the corrected Krylov approximation. Whether this is a real drawback of the corrected Krylov approximation depends on the emphasis placed on mass conservation. In the following examples we focus on the standard Krylov approximation, with some exceptions which serve for comparisons with the original Expokit code, which is based on the corrected Krylov approximation.

In exact arithmetic we obtain mass conservation for the skew-Hermitian case: For the case $\|v\|_2 = 1$ and the standard Krylov approximation $S_m(t)v$ we have

$$\|S_m(t)v\|_2 = \|V_m \mathrm{e}^{-\mathrm{i}tT_m} e_1\|_2 = e_1^* \mathrm{e}^{\mathrm{i}tT_m} V_m^* V_m \mathrm{e}^{-\mathrm{i}tT_m} e_1 = 1. \tag{2.8.3}$$

The requirement $V_m^* V_m = I$ is essential to obtain mass conservation in (2.8.3). In computer arithmetic the loss of orthogonality of the Krylov basis V_m has been studied earlier, see also [Pai76]. To preserve the property of mass conservation a reorthogonalization, see [Par98], may be advisable in this case.



Computed effective order of the defect for $m = 10$:

t	$3.9 \cdot 10^{-2}$	$5.5 \cdot 10^{-2}$	$7.8 \cdot 10^{-2}$	$1.1 \cdot 10^{-1}$	$1.5 \cdot 10^{-1}$	$2.2 \cdot 10^{-1}$	$3.1 \cdot 10^{-1}$	$4.4 \cdot 10^{-1}$	$6.3 \cdot 10^{-1}$	$8.8 \cdot 10^{-1}$
$\rho(t)$	8.99	8.98	8.95	8.90	8.81	8.63	8.24	7.44	5.66	1.00

and $m = 30$:

t	$8.8 \cdot 10^{-1}$	$1.2 \cdot 10^{0}$	$1.8 \cdot 10^{0}$
$\rho(t)$	26.68	24.18	18.42

Figure 2.6: The upper left plot shows the error $\|L_m(t)v\|_2$ ($\circ$), the generalized residual estimate (2.6.6) ($+$) and the error estimates based on the Hermite quadrature (2.6.3) ($\times$), and the effective order quadrature (2.6.9) ($\Diamond$) for the Hubbard Hamiltonian with $m = 10$ and $m = 30$. The dashed lines show the error estimate Err_a. On the right-hand side the graphics show a detail from the error plots to illustrate the inequalities (2.6.8). The table on the bottom shows the computed effective order of the defect for $m = 10$ and $m = 30$ which is used for the effective order quadrature.

Krylov approximation of the matrix exponential in computer arithmetic. It has been shown in [DGK98, DK92, Gre89] that a priori error estimates for the Krylov approximation of the matrix exponential remain valid also taking account of affects of arithmetic. Such results imply that in general in computer arithmetic the convergence of the Krylov approximation is not precluded and round-off errors are not critical. In practice round-off errors may in some cases lead to a delay of convergence which can make a reorthogonalization relevant. Stability of the Krylov approximation has been discussed by many authors, see also [MMS18], but is not further discussed here in detail. In the next paragraph we will give an argument, following [DGK98], that the a posteriori error estimates which are the topic of this work are robust with respect to round-off errors.

We recall that the Krylov subspace constructed in computer arithmetic satisfies the Krylov identity (2.2.2) with a small perturbation, see also [Pai76] for the Lanczos case and [BLR00, Zem03] for the Arnoldi case, which can both be extended to complex-valued problems using results from [Hig02]. Following results from [DGK98] we conclude that a small perturbation of the Krylov identity leads to a small perturbation of the defect (residual) $\delta_m(t)$ in (2.3.1a) and the integral representation of the error in (2.3.1b). Thus the error estimates given in Section 2.6 remain stable with respect to round-off.

We further use that by construction the computed T_m is still upper Hessenberg with a positive lower diagonal and in the Lanczos case also real-valued and symmetric. Then following Proposition 2.5.5 in the Hermitian (Lanczos) case, the integral representation of the error in (2.3.1b) results in the upper error bound Err_1, which is not critically affected by round-off errors. For the upper bound Err_a we further assume that spectral properties of (2.3.4) still hold mutatis mutandis under a small perturbation, see [Pai80] for such results for the Lanczos case, to obtain stability of this upper error bound also with round-off.

Numerical tests for step size control. The idea of choosing discrete time steps for the Krylov approximation is described in Section 2.7. The following tests are applied to the matrix exponential of the Hubbard Hamiltonian. We first clarify the notation used for our test setting.

Expokit and Expokit⋆. The original Expokit code uses the corrected Krylov approximation with heuristic step size control and an error estimator which is based on the error expansion (2.5.1), see [Sid98, Algorithm 3.2] for details. Since the standard Krylov approximation is not part of the Expokit package, we have slightly adapted the code and its error estimate such that the standard Krylov approximation is used. We refer to the adapted package as Expokit⋆. With Expokit⋆ our comparison can be drawn with the standard Krylov approximation which may in some cases be the method of choice as discussed above.

Step size based on Err_a*.* In another test code the upper error bound Err_a from Theorem 2.3.2 is used. With Err_a we obtain proven upper bounds on the error and reliable step sizes (2.7.5).

By *gen.res,* eff.o.quad*, and* Err_1 *we refer to the generalized residual estimate* (2.6.6)*, the effective order quadrature* (2.6.9)*, and* (Err_1)*, respectively. Because these error estimates cannot be inverted directly we need to apply heuristic ideas for the step*

size control, see (2.7.7). *In addition, we use the iteration* (2.7.9) *to improve step sizes. For the test problems we have solved, iteration* (2.7.9) *converges in less than* 2 *iterations for* $m = 10$ *or less than* 5 *iterations for* $m = 30$. *We simply choose* $N_j = 5$ *for our tests.*

The a priori estimates (2.7.8), [HL97, Theorem 4] and [MC10, eq. (20)] *are given in the corresponding references. Formula* (2.7.8) *taken from the Expokit code directly provides a step size. In [MC10, eq. (20)] the computation of the step size is described. For the error estimate given in [HL97, Theorem 4] we apply Newton iteration to determine an appropriate step size. For tests on the Hubbard model we use* $(\lambda_{max} - \lambda_{min}) = 27.4$ *as suggested in the description of the Hubbard Hamiltonian.*

In Remark 2.8.2 below we also investigate the following variants:

Step size based on Err_a^+. By Err_a^+ we denote the upper error bound for the corrected Krylov approximation as given in Theorem 2.5.4 with $p = 0$. The corresponding step size is given by (2.7.6).

By *i.H.quad* we refer to the improved Hermite quadrature (2.6.5). Similarly to other quadrature error estimates we use heuristic step size control and iteration (2.7.9) to determine adequate step sizes.

Remark 2.8.1. *In the Expokit code the step sizes are rounded to* 2 *digits in every step. Rounding the step size can give too large errors in some steps. This makes it necessary to include safety parameters in Expokit which on the other hand slow down the performance of the code. It seems advisable to avoid any kind of rounding of step sizes.*

In Table 2.1 we compare the total time step t for the Krylov approximation with $m = 10$ and $m = 30$ after $N = 10$ steps obtained with the different step size control strategies. For the local error we choose the tolerance $\mathrm{tol} = 10^{-8}$. The original Expokit code seems to give larger step sizes, but it also uses a larger number of matrix-vector multiplications, see Remark 2.8.2. The error estimate Err_a leads to optimal step sizes for $m = 10$ and close to optimal step sizes for $m = 30$. For any choice of m the error estimate Err_a gives reliable step sizes. The generalized residual estimate overestimates the error and, therefore, step sizes are smaller. The effective order quadrature and Err_1 give optimal step sizes. With the assumptions from Section 2.6 (which apply to our test examples), the generalized residual estimate and effective order quadrature give reliable step sizes. For the error estimate Err_1 we do not have results on the reliability of the step sizes since the error estimate Err_1 does not lead to an upper bound of the error in general. The tested a priori estimates (2.7.8), [HL97, Th. 4], and [MC10, (20)] overestimate the error and lead to precautious step size choices. For all the tested versions the accumulated error $L_m^\star$ (see (2.7.2)) satisfies $\|L_m^\star v\|_2/t \leq \mathrm{tol}$.

Apart from step size control, the upper error bound Err_a can be used on the fly to test if the dimension of the Krylov subspace is already sufficiently large to solve the problem in a single time step with the required accuracy. For our test problems this stopping criterion is applied to the Err_a estimate.We refer to Table 2.2, in which we observe the Krylov method with error estimate Err_a to stop after 17 steps instead of computing the

Table 2.1: The displayed step size t is the sum of $N = 10$ substeps computed by different versions of step size control, as described above. In the top table we show the results for $m = 10$, in the bottom table for $m = 30$, both for tolerance $\text{tol} = 10^{-8}$, for the Hubbard Hamiltonian.

$m = 10$	Expokit	Expokit*	Err_a	gen.res	eff.o.quad	Err_1	(2.7.8)	[HL97, Th. 4]	[MC10, (20)]
t	0.9020	0.6850	0.8468	0.6568	0.8488	0.8489	0.1918	0.4918	0.6879
N	10	10	10	10	10	10	10	10	10
# m-v	110	100	100	100	100	100	100	100	100
$\|L_m^\star v\|_2/t$	$3.5{\cdot}10^{-09}$	$2.9{\cdot}10^{-09}$	$9.8{\cdot}10^{-09}$	$1.0{\cdot}10^{-09}$	$1.0{\cdot}10^{-08}$	$1.0{\cdot}10^{-08}$	$3.0{\cdot}10^{-14}$	$7.5{\cdot}10^{-11}$	$1.5{\cdot}10^{-09}$

m=30	Expokit	Expokit*	Err_a	gen.res	eff.o.quad	Err_1	(2.7.8)	[HL97, Th. 4]	[MC10, (20)]
t	8.5700	8.2500	9.7248	9.0091	10.2127	10.2222	2.1131	8.2642	8.8111
N	10	10	10	10	10	10	10	10	10
# m-v	310	300	300	300	300	300	300	300	300
$\|L_m^\star v\|_2/t$	$2.6{\cdot}10^{-10}$	$2.9{\cdot}10^{-10}$	$2.6{\cdot}10^{-09}$	$3.5{\cdot}10^{-10}$	$9.5{\cdot}10^{-09}$	$9.7{\cdot}10^{-09}$	$2.9{\cdot}10^{-15}$	$3.4{\cdot}10^{-11}$	$1.9{\cdot}10^{-10}$

Table 2.2: With a test setting similar to Table 2.1, we now compute up to a fixed time $t = 0.3$ and choose the number N of steps according to the step size control. We use a tolerance $\text{tol} = 10^{-8}$ and $m = 30$. For this problem we see a significant reduction in the number of matrix-vector multiplications used for the estimate Err_a by the stopping criteria described in the text.

$m = 10$	Expokit	Expokit*	Err_a	gen.res	Err_1	(2.7.8)	[HL97, Th. 4]	[MC10, (20)]
t	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
N	2	2	1	1	1	2	1	1
# m-v	62	60	17	30	30	60	30	30
$\|L_m^\star v\|_2/t$	$8.4{\cdot}10^{-15}$	$8.4{\cdot}10^{-15}$	$1.0{\cdot}10^{-09}$	$9.7{\cdot}10^{-15}$	$9.7{\cdot}10^{-15}$	$1.0{\cdot}10^{-14}$	$9.7{\cdot}10^{-15}$	$9.7{\cdot}10^{-15}$

full Krylov subspace of dimension 30. In comparison, the original Expokit package needs a total of 62 matrix-vector multiplications.

Remark 2.8.2. *Error estimates for the corrected Krylov approximation or improved error estimates such as the improved Hermite quadrature* (2.6.5) *require additional matrix-vector multiplications. Instead of investing computational effort in improving the error estimate, one may as well increase the dimension of the standard Krylov subspace. For comparison we test the original Expokit code, the corrected Krylov approximation with error estimate* Err_a^+ *and the improved Hermite quadrature* (2.6.5) *with Krylov subspace* $m - 1$*. Table 2.3 shows that a standard Krylov approximation with dimension* m *leads to better results, although all considered versions use the same number of matrix-vector multiplications. Since the reliability of error estimates such as* Err_a *has been demonstrated earlier, it appears that additional cost to improve the error estimate is not justified.*

2.8.2 The Hermitian case

To obtain a more complete picture, we also briefly consider the case of a Hermitian matrix $A = H$ with $\sigma = 1$ in (2.2.1). Such a model is typical of the discretization of a parabolic PDE. Thus, the result may depend on the regularity of the initial data, which is chosen to be random in our experiments.

Table 2.3: All variants shown use exactly m matrix-vector multiplications. Whereas Expokit, improved Hermite quadrature (i.H.quad) and Err_a^+ imply higher cost for the error estimate, the other codes Err_a, effective order quadrature (eff.o.quad) and Err_1 use standard Krylov subspaces and do not spend additional matrix-vector multiplications on error estimates.

$m = 10$	Expokit	Err_a^+	i.H.quad	Err_a	eff.o.quad	Err_1
t	0.6620	0.7828	0.5863	0.8346	0.8366	0.8368
N	10	10	10	10	10	10
# m-v	100	100	100	100	100	100
$\|L_m^\star v\|_2/t$	$4.1 \cdot 10^{-09}$	$8.8 \cdot 10^{-09}$	$1.0 \cdot 10^{-08}$	$9.8 \cdot 10^{-09}$	$1.0 \cdot 10^{-08}$	$1.0 \cdot 10^{-08}$

$m = 30$	Expokit	Err_a^+	i.H.quad	Err_a	eff.o.quad	Err_1
t	8.1900	9.5763	9.6591	9.7482	10.2378	10.2473
N	10	10	10	10	10	10
# m-v	100	100	100	100	100	100
$\|L_m^\star v\|_2/t$	$3.6 \cdot 10^{-10}$	$2.7 \cdot 10^{-09}$	$9.2 \cdot 10^{-09}$	$2.6 \cdot 10^{-09}$	$9.5 \cdot 10^{-09}$	$9.7 \cdot 10^{-09}$

Heat equation. To obtain the heat equation in (2.2.1) we choose $A = H$ in (2.8.1) and $\sigma = -1$. Details on the test setting are already given in Subsection 2.8.1.

For the heat equation, H given in (2.8.1), we can also verify the error estimates, see Figure 2.7. In comparison to the skew-Hermitian case we do not observe a large time regime for which the error is of the asymptotic order m. As shown in Proposition 2.5.5 we do obtain an upper error bound using Err_1 for the heat equation.

Similarly to the skew-Hermitian case, we can also apply the effective order quadrature according to Remark 2.6.1 to the Hermitian case. Use $\sigma = -1$ and $T_m \in \mathbb{R}^{m\times m}$ in (2.6.10) to obtain

$$\rho(t) = -t\left((T_m)_{m,m} + (T_m)_{m,m-1}\frac{(\mathrm{e}^{-tT_m}e_1)_{m-1}}{(\mathrm{e}^{-tT_m}e_1)_m}\right).$$

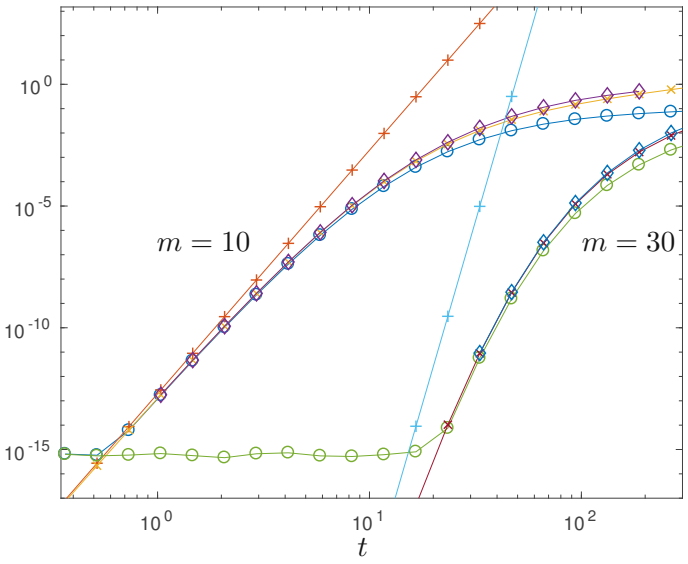
For computing the effective order we only consider time steps $\rho(t) > 0$, and where ρ appears indeed to be monotonically decreasing over the computed discrete time steps. This restriction is compatible with our assumptions in Section 2.6.

2.8.3 A non-normal problem

For a more general case we consider a convection-diffusion equation (see [MN01, EE06]).

$$\begin{aligned}&\partial_t u = \Delta u - \tau_1\partial_{x_1}u - \tau_2\partial_{x_2}u, \quad \tau_1,\tau_2 \in \mathbb{R}, \quad u = u(t,x), \quad t \geq 0, \quad x \in \Omega = [0,1]^3,\\ &u(0,x) = v(x) \quad \text{for } x \in \Omega, \quad u(t,x) = 0 \quad \text{for } x \in \partial\Omega.\end{aligned} \tag{2.8.4}$$

Following [MN01, EE06] we use a central finite difference scheme to discretize the partial differential operator in (2.8.4). The grid is chosen uniformly with $(n+2)^3$ points and mesh



Computed effective order of the defect for $m = 10$ (partly):

t	$1.0 \cdot 10^0$	$1.5 \cdot 10^0$	$2.1 \cdot 10^0$	$2.9 \cdot 10^0$	$4.1 \cdot 10^0$	$5.9 \cdot 10^0$	$8.3 \cdot 10^0$	$1.2 \cdot 10^1$	$1.7 \cdot 10^1$	$2.3 \cdot 10^1$	$3.3 \cdot 10^1$
$\rho(t)$	8.50	8.30	8.02	7.64	7.14	6.48	5.65	4.66	3.58	2.52	1.60

and $m = 30$:

t	$3.3 \cdot 10^1$	$4.7 \cdot 10^1$	$6.6 \cdot 10^1$	$9.4 \cdot 10^1$	$1.3 \cdot 10^2$	$1.9 \cdot 10^2$	$2.7 \cdot 10^2$
$\rho(t)$	16.60	13.47	10.33	7.48	5.15	3.36	2.07

Figure 2.7: Error $\|L_m(t)v\|_2$ ($\circ$), the error estimates Err_1 ($\times$) and Err_a ($+$) and the error estimate based on the effective order quadrature (2.6.9) ($\Diamond$) for the heat equation with $m = 10$ and $m = 30$. The tabular on the bottom shows some of the computed values for the effective order.

width $h = 1/(n+1)$. The dimension N of the discrete operator is $N = n^3$. Choosing $n = 15$ we obtain $N = 3357$. The discretized operator is given by

$$A = (I_{n\times n} \otimes (I_{n\times n} \otimes C_1)) + (B \otimes I_{n\times n} + I_{n\times n} \otimes C_2) \otimes I_{n\times n}) \in \mathbb{R}^{N\times N}, \quad \text{with} \quad (2.8.5)$$
$$B = \tfrac{1}{h^2}\text{tridiag}(1, -2, 1) \in \mathbb{R}^n, \quad C_i = \tfrac{1}{h^2}\text{tridiag}(1+\mu_i, -2, 1-\mu_i) \in \mathbb{R}^n, \quad i = 1, 2,$$

and $\mu_i = \tau_i\,(h/2)$. The spectrum of the non-normal matrix A in (2.8.5) (see [MN01]) satisfies

$$\text{spec}(A) \subseteq \tfrac{1}{h^2}\,[-6 - 2\cos(\pi\,h)\text{Re}(\theta), -6 + 2\cos(\pi\,h)\text{Re}(\theta)]$$
$$\times\ \tfrac{1}{h^2}\,\text{i}\,[-2\cos(\pi\,h)\text{Im}(\theta), 2\cos(\pi\,h)\text{Im}(\theta)].$$

with $\theta = 1 + \sqrt{1-\mu_1^2} + \sqrt{1-\mu_2^2}$. Therefore, the eigenvalues are complex-valued if at least one $\mu_i > 1$. The matrix A depends on the parameters μ_i, correspondingly τ_i, for which we consider two different cases,

$$\mu_1 = 0.9,\ \mu_2 = 1.1, \quad \text{with} \quad \text{spec}(h^2\,A) \subseteq [-9, -3] \times \text{i}[-1, 1], \qquad (2.8.6)$$

and

$$\mu_1 = \mu_2 = 10, \quad \text{with} \quad \text{spec}(h^2\,A) \subseteq [-8, -4] \times \text{i}[-39, 39]. \qquad (2.8.7)$$

In the following numerical experiments we apply the Krylov approximation to $\text{e}^{t\,A}v$ ($\sigma = 1$ in (2.2.1)) for different time steps t and starting vector $v = (1, \ldots, 1)^* \in \mathbb{R}^N$ as in [MN01]. For non-normal A we use the Arnoldi method based on a modified Gram-Schmidt procedure (see [Saa03, Algorithm 6.2]) to generate the Krylov subspace.

The error estimates Err_a and Err_1 are compared to the exact error norm $\|L_m(t)v\|_2$ in Figure 2.8 for the case (2.8.6) and in Figure 2.9 for the case (2.8.7). As shown in Theorem 2.3.2 the error estimate Err_a constitutes an upper error bound. The error estimate Err_1 gives a good approximation of the error but has not been proven to give an upper bound in general.

Compared to (2.8.7), the spectrum for (2.8.6) is closer to the Hermitian case. The spectrum for (2.8.7), on the other hand, is dominated by large imaginary parts similarly as in the skew-Hermitian case.

In Figure 2.8 we observe effects similar to the Hermitian case. The asymptotic order m of the error does not hold for a large time regime, and the error estimate Err_a is not as sharp as in the skew-Hermitian case. On the other hand, in Figure 2.9, we observe that the performance of the error estimates is closer to the skew-Hermitian case. Therefore, the upper error bound Err_a is sharp for a larger range of time steps. As already observed for the Hermitian and skew-Hermitian cases, the error of the Krylov approximation is closer to its asymptotic order m for smaller choices of m.

2.9 Summary and outlook.

We have studied a new reliable error estimate Err_a for Krylov approximations to the matrix exponential and φ-functions. This error estimate constitutes an upper bound on the error, and it can be computed on the fly at nearly no additional cost. The Krylov process can

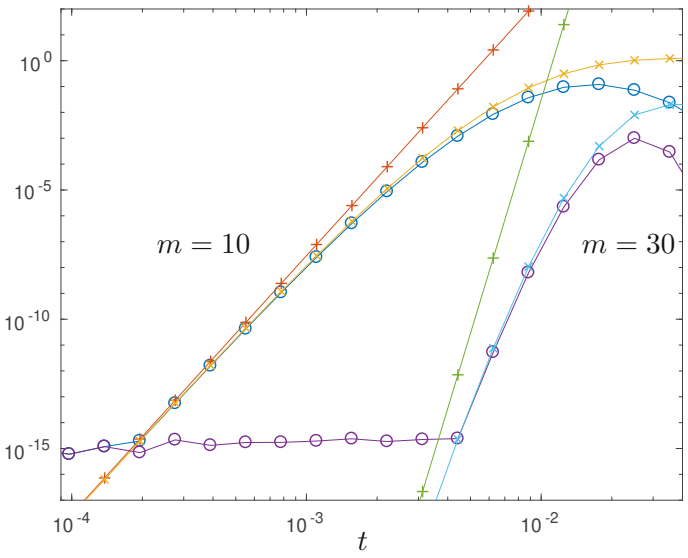


Figure 2.8: Error $\|L_m(t)v\|_2$ ($\circ$) and the error estimates Err_1 ($\times$) and Err_a ($+$) for the convection-diffusion problem (2.8.5) with $\mu_1 = 0.9$ and $\mu_2 = 1.1$ and Krylov subspace dimensions $m = 10$ and $m = 30$.

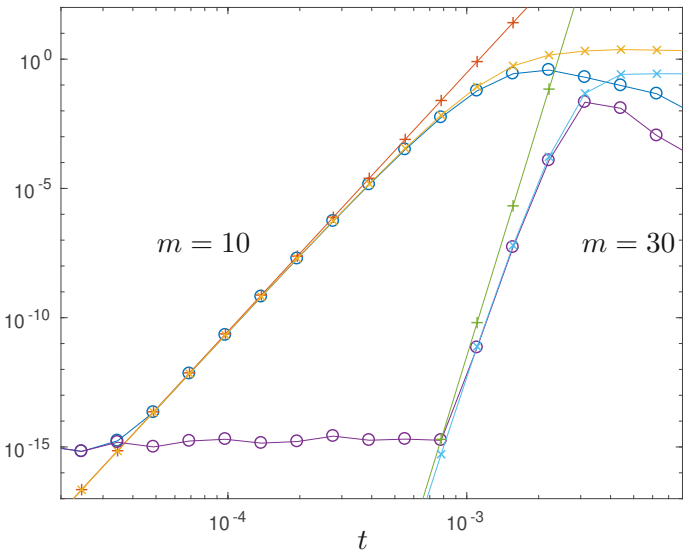


Figure 2.9: Error $\|L_m(t)v\|_2$ ($\circ$) and the error estimates Err_1 ($\times$) and Err_a ($+$) for the convection-diffusion problem (2.8.5) with $\mu_1 = \mu_2 = 10$ and Krylov subspace dimensions $m = 10$ and $m = 30$.

be stopped as soon as the error estimate satisfies a given tolerance. Err_a is asymptotically correct for $t \to 0$ and very tight in the asymptotic regime. Our numerical experiments illustrate that the asymptotic regime is more relevant for the skew-Hermitian case (compared to the Hermitian case) and for a smaller choice of m and tolerances. The non-normal examples seem to be in between the skew-Hermitian and Hermitian cases.

In our numerical experiments the defect (residual) is seen to behave nicely close to the asymptotic regime and the generalized residual estimate is observed to constitute an upper bound. The generalized residual estimate can be tightened by applying an effective order quadrature.

For the Hermitian case we have shown that the error estimate Err_1 constitutes an upper bound and, compared to other error estimates, seems to be the most appropriate choice for Hermitian problems.

Step size control for a simple restarted scheme is an important application. The upper error bound Err_a is an appropriate tool for this task, since the optimal step size for a given tolerance can be computed directly. This is not the case for other error estimates for the Krylov approximation, which usually employ heuristic schemes to compute optimal step sizes in the restarting approach. We have shown that the step size can be cheaply improved by using a heuristic step size approach in an iterative manner. Also the use of a priori bounds is not optimal in most cases.

3 A study of defect-based error estimates for the Krylov approximation of φ-functions

3.1 Introduction

Overview on prior work. The matrix exponential and associated φ-functions play a crucial role in some numerical methods for solving systems of differential equations. In practice this means that the vector $\mathrm{e}^{tA}v$, for a time step t, a given matrix A and a given vector v, representing the time propagation for a linear initial value problem, is to be approximated. Similarly, the associated φ-functions (see (3.2.2) below) conform to solutions of certain inhomogeneous differential equations. In particular, evaluation of φ-functions is used in exponential integrators [HO10].

If the matrix A is sparse and large, approximation of the action of these matrix functions in the class of Krylov subspaces is a general and well-established technique. For the matrix exponential and φ-functions this goes back to early works in the field of chemical physics [NW83, PL86], parabolic problems [GS92], some nonlinear problems [FTDR89], etc. The case of a symmetric or skew-Hermitian matrix A is the most prominent one. Krylov approximations of the matrix exponential were early studied for the symmetric case in [DK89, DK92, Saa92], and together with φ-functions in a more general setting [HLS98, HL97].

Concerning different approaches for the numerical approximation of the matrix exponential see [MVL03]. In [Saa92] it is shown for the symmetric case that the Krylov approximation is equivalent to interpolation of the exponential function at associated Ritz values. This automatically results in a near-best approximation among other choices of interpolation nodes, see also [DK89, SL96] and further works [BR09] with similar results for the non-symmetric case and general functions including φ-functions. For other polynomial approaches approximating the matrix exponential we mention truncated Taylor series [AMH11] (and many works well in advance), Chebyshev interpolation [TEK84], or the Leja method [CKOR16], where [AMH11] also covers φ-functions.

In general, Krylov approximations (or other polynomial approximations) result in an accurate approximation if the time step t in $\mathrm{e}^{tA}v$ is sufficiently small or the dimension of the Krylov subspace (i.e., the degree of the approximating matrix polynomial) is sufficiently large, see for instance [HL97]. The dimension of the Krylov subspace is limited in practice, and large time steps require a restart of the iteration generating the Krylov basis. A larger time step t can be split into smaller substeps for which the Krylov approximation can be applied in a nested way. Such a restarting strategy in the sense of a time integrator was already exploited in [PL86]. In particular we refer to the EXPOKIT package [Sid98]. Similar ideas can be applied for the evaluation of φ-functions [HLS98, NW12, Sid98].

In practice, a posteriori error estimates are used to choose a proper Krylov dimension

or proper (adaptive) substeps if the method is restarted as a time integrator. Different approaches for a posteriori error estimation concerning the exponential function make use of a series expansion for the error given [Saa92] or use a formulation via the defect (also called residual) of the Krylov approximation [DGK98, HLS98, CM97, BGH13]. A prominent error estimate concerning φ-functions is the generalized residual estimate introduced in [HLS98], which is based on the residual of a matrix inverse. Furthermore, a series expansion of the error concerning φ-functions is given in [Sid98] (similar to the series expansion concerning the exponential in [Saa92]) and leading terms of this series are used for a posteriori error estimation in [Sid98, NW12]. Further a priori as well as a posteriori error estimates for the exponential function are are given in [MN01, Lub08, DMR09, BR09, JL15, WY17, JAK20], where [DMR09, JAK20] also consider φ-functions. Restarting via substeps based on different choices of error estimates is further discussed in [JAK20]. A restart with substeps together with a strategy to choose the Krylov dimension in terms of computational cost was presented in [NW12, BK19]. For various other approaches for restarting (without adapting the time step) we refer to [CM97, EE06, TE07, Nie07, AEEG08, EEG11, BGH13, Sch15].

The influence of round-off errors on the construction of the Krylov basis in floating point arithmetic was early studied for the symmetric case in [Pai76, Par98]. The orthogonalization procedure can behave numerically unstable, typically due to a loss of orthogonality. Nevertheless, the near-best approximation property and related a priori convergence results are not critically affected [DK92, DGK98]. Following [DGK98], in the symmetric case the defect obtained in floating point arithmetic results in numerically stable error estimates.

Beside the polynomial Krylov method, further studies are devoted to the approximation of matrix functions using so called extended Krylov subspaces [DK98, KS10, GG13], rational Krylov subspaces [MN04, vdEH06, Güt10], or polynomial Krylov subspaces with a harmonic Ritz approach [HH05, Sch15, WZX16].

Overview on results presented here. In Section 3.2 we introduce the problem setting and recapitulate basic properties of Krylov subspaces.

In Section 3.3 we introduce the *defect* associated with Krylov approximations to φ-functions, including the exponential function as the basic case. Our approach for the defect is different from [WZX16] and is based on an inhomogeneous differential equation for the approximation error. This is used in Theorem 3.3.1 to obtain an integral representation of the error, also taking effects of floating point arithmetic into account. [1] In contrast to previous works ([DGK98, JAK20]), this result is extended to φ-functions here.

This upper bound is further analyzed in Section 3.4 to obtain computable a posteriori bounds, in particular a new a posteriori bound (Theorem 3.4.3). We also study the accuracy of our and other existing defect-based bounds [JAK20] with respect to spectral properties of the Krylov Hessenberg matrix (the representation of A in the orthogonal Krylov basis). To this end we use properties of divided differences including a new asymptotic expansion for these given in Appendix 3.C. In Subsection 3.4.1 we consider error estimates based on a quadrature estimate of the defect norm integral: The generalized residual estimate [HLS98] for the approximation of φ-functions which conforms to a quadrature of the defect norm

[1] Cf. [DGK98] for the case of the matrix exponential.

integral (namely, the right-endpoint rectangle rule), and the effective order estimate, which was introduced for the approximation of the matrix exponential in [JAK20] and is extended to φ-functions in the present work. We also discuss cases for which the defect norm behaves oscillatory and reliable quadrature estimates may be difficult to obtain. In Subsection 3.4.2 we specify a stopping criterion for the so-called lucky breakdown in floating point arithmetic which is justified by our a posteriori error bounds.

In Section 3.5 we illustrate our results via numerical experiments. This includes further remarks on previously known error estimates for the Krylov approximation of φ-functions.

3.2 Problem statement and Krylov approximation

We discuss the approximation via Krylov techniques for evaluation of the matrix exponential, and in particular of the associated φ-functions, for a step size $t > 0$ and matrix $A \in \mathbb{C}^{n\times n}$ applied to an initial vector $v \in \mathbb{C}^n$. Here,

$$\mathrm{e}^{tA}v = \sum_{k=0}^{\infty} \frac{(tA)^k}{k!}v. \tag{3.2.1}$$

The matrix exponential $u(t) = \mathrm{e}^{tA}v$ is the solution of the differential equation

$$u'(t) = Au(t), \quad u(0) = v.$$

The associated φ-functions are given by

$$\varphi_p(tA)v = \sum_{k=0}^{\infty} \frac{(tA)^k}{(k+p)!}v, \quad p \in \mathbb{N}_0. \tag{3.2.2}$$

This includes the case $\varphi_0 = \exp$. The matrix functions (3.2.1) and (3.2.2) are defined according to their scalar counterparts. The following definitions of φ_p are equivalent to (3.2.2): For $z \in \mathbb{C}$ we have $\varphi_0(z) = \mathrm{e}^z$, and

$$\varphi_p(z) = \frac{1}{(p-1)!}\int_0^1 \mathrm{e}^{(1-\theta)z}\theta^{p-1}\,\mathrm{d}\theta, \quad p \in \mathbb{N}. \tag{3.2.3}$$

(See also [Hig08, Subsection 10.7.4].) The function $w_p(t) = t^p\varphi_p(tA)v$ ($p \in \mathbb{N}$) is the solution of an inhomogeneous differential equation of the form

$$w_p'(t) = Aw_p(t) + \frac{t^{p-1}}{(p-1)!}v, \quad w_p(0) = 0, \tag{3.2.4}$$

see for instance [NW12]. This follows from (3.2.2),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big(t^p\varphi_p(tA)v\big) &= \frac{\mathrm{d}}{\mathrm{d}t}\Big(\sum_{k=0}^{\infty}\frac{t^{k+p}A^kv}{(k+p)!}\Big) = A\sum_{k=0}^{\infty}\frac{t^{k+p}A^kv}{(k+p)!} + \frac{t^{p-1}v}{(p-1)!}\\ &= A(t^p\varphi_p(tA)v) + \frac{t^{p-1}v}{(p-1)!}.\end{aligned}$$

The φ-functions appear for instance in the field of exponential integrators, see for instance [HO10].

For the case of A being a large and sparse matrix, e.g., the spatial discretization of a partial differential operator using a localized basis, Krylov subspace techniques are commonly used to approximate (3.2.2) in an efficient way.

Notation and properties of Krylov subspaces. [2] We briefly recapitulate the usual notation and properties of standard Krylov subspaces, see for instance [Saa03]. For a given matrix $A \in \mathbb{C}^{n\times n}$, a starting vector $v \in \mathbb{C}^n$ and Krylov dimension $0 < m \leq n$, the Krylov subspace is given by

$$\mathcal{K}_m(A, v) = \operatorname{span}(v, Av, \ldots, A^{m-1}v).$$

Let $V_m \in \mathbb{C}^{n\times m}$ represent the orthonormal basis of $\mathcal{K}_m(A, v)$ with respect to the Hermitian inner product, constructed by the Arnoldi method and satisfying $V_m^* V_m = I_{m\times m}$. Its first column is given by $V_m^* v = \beta e_1$ with $\beta = \|v\|_2$. Here, the matrix

$$H_m = V_m^* A V_m \in \mathbb{C}^{m\times m}$$

is upper Hessenberg. We further use the notation $h_{m+1,m} = (H_{m+1})_{m+1,m} \in \mathbb{R}$, and $v_{m+1} \in \mathbb{C}^n$ for the $(m+1)$-th column of V_{m+1}, with $V_m^* v_{m+1} = 0$ and $\|v_{m+1}\|_2 = 1$.

The Arnoldi decomposition (in exact arithmetic) can be expressed in matrix form,

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^*. \tag{3.2.5}$$

Remark 3.2.1. *The numerical range* $\mathrm{W}(A) = \{y^*Ay/y^*y,\ 0 \neq y \in \mathbb{C}^n\}$ *plays a role in our analysis. Note that* $\mathrm{W}(H_m) \subseteq \mathrm{W}(A)$ *(see* (3.A.1)*).*

Remark 3.2.2. *The case* $(H_m)_{k+1,k} = 0$ *occurs if* $\mathcal{K}_k(A, v)$ *is an invariant subspace of* A*, whence the Krylov approximation given in* (3.2.9) *below is exact. This exceptional case is referred to as a* lucky breakdown. *In general we assume that no lucky breakdown occurs, whence the lower subdiagonal entries of* H_m *are real and positive,* $0 < (H_m)_{j+1,j}$ *for* $j = 1, \ldots, m-1$*, and* $0 < h_{m+1,m} \in \mathbb{R}$.

For the special case of a Hermitian or skew-Hermitian matrix A the Arnoldi iteration simplifies to a three-term recurrence, the so-called Lanczos iteration. This case will be addressed in Remark 3.2.4 below.

Krylov subspaces in floating point arithmetic. We proceed with some results for the Arnoldi decomposition in computer arithmetic, assuming complex floating point arithmetic with a relative machine precision ε, see also [Hig02]. For practical implementation different variants of the Arnoldi procedure exist, using different ways for the orthogonalization of the Krylov basis. These are based on classical Gram-Schmidt, modified Gram-Schmidt, the Householder algorithm, the Givens algorithm, or variants of Gram-Schmidt with reorthogonalization (see also [Saa03, Algorithm 6.1–6.3] and others). We refer to [BLR00] and references therein for an overview on the stability properties of these different variants.

[2] In the sequel, e_j denotes the j-th unit vector in $\mathbb{C}^m$ or $\mathbb{C}^n$, respectively.

In the sequel the notation V_m, H_m, etc., will again be used for the result of the Arnoldi method in floating point arithmetic. We now accordingly adapt some statements formulated in the previous paragraph. By construction, H_m remains to be upper Hessenberg with positive lower subdiagonal entries. Assuming floating point arithmetic we use the notation $U_m \in \mathbb{C}^{n\times m}$ for a perturbation of the Arnoldi decomposition (3.2.5) caused by round-off, i.e.,

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^* + U_m. \tag{3.2.6}$$

An upper norm bound for U_m was first introduced in [Pai76] for the Lanczos iteration in real arithmetic. For different variants of the Arnoldi or Lanczos iteration this is discussed in [Zem03] and others. We assume $\|U_m\|_2$ is bounded by a constant C_1 which can depend on m and n in a moderate way and is sufficiently small in a typical setting,

$$\|U_m\|_2 \leq C_1 \varepsilon \|A\|_2. \tag{3.2.7a}$$

We further assume that the normalization of the columns of V_m is accurate, in particular that the $(m+1)$-th basis vector v_{m+1} is normalized correctly up round-off with a sufficiently small constant C_2 (see e.g., [Pai76, eq. (14)]),

$$|\|v_{m+1}\|_2 - 1| \leq C_2 \varepsilon. \tag{3.2.7b}$$

Concerning V_{m+1} which represents an orthogonal basis in exact arithmetic, numerical loss of orthogonality has been well-studied. Loss of orthogonality can be significant (see for instance [Par98, BLR00] and others), depending on the starting vector v. Reorthogonalization schemes or orthogonalization via Householder or Givens algorithm can be used to obtain orthogonality of V_{m+1} on a sufficiently accurate level.

The numerical range of H_m obtained in floating point arithmetic (see (3.2.6)) can be characterized as

$$\mathrm{W}(H_m) \subseteq U_{C_3\varepsilon}(\mathrm{W}(A)), \tag{3.2.7c}$$

with $U_{C_3\varepsilon}(\mathrm{W}(A))$ being the neighborhood of $\mathrm{W}(A)$ in $\mathbb{C}$ with a distance $C_3\varepsilon$. With the assumption that V_{m+1} is sufficiently close to orthogonal (e.g., semiorthogonal [Sim84]), the constant C_3 in (3.2.7c) (which also depends on C_1 and problem sizes) can be shown to be moderate-sized. Further details on this aspect are given in Appendix 3.A.

Krylov approximation of φ-functions. [3] Let $V_m \in \mathbb{C}^{n\times m}$, $H_m \in \mathbb{C}^{m\times m}$ and $\beta \in \mathbb{R}$ be the result of the Arnoldi method in floating point arithmetic for $\mathcal{K}_m(A, v)$ as described above. For a time-step $0 < t \in \mathbb{R}$ and $p \geq 0$ the vector $\varphi_p(tA)v$ can be approximated in the Krylov subspace $\mathcal{K}_m(A, v)$ by the *Krylov propagator*

$$u_{p,m}(t) := V_m \varphi_p(tV_m^* A V_m) V_m^* v = \beta V_m \varphi_p(tH_m) e_1, \quad p \in \mathbb{N}. \tag{3.2.8a}$$

The special case $p = 0$ reads

$$u_{0,m}(t) = \beta V_m \mathrm{e}^{tH_m} e_1. \tag{3.2.8b}$$

[3] Remark concerning notation: 'u' objects live in $\mathbb{C}^n$, and 'y' objects live in $\mathbb{C}^m$.

We remark that the small-dimensional problem $\varphi_p(tH_m)e_1 \in \mathbb{C}^m$, typically with $m \ll n$, can be evaluated cheaply by standard methods. In the sequel we denote

$$y_{p,m}(t) = \beta\varphi_p(tH_m)e_1 \in \mathbb{C}^m, \quad \text{i.e.,} \quad u_{p,m}(t) = V_m y_{p,m}(t). \tag{3.2.9}$$

For $p = 0$ the small dimensional problem $y_{0,m}(t) = \beta \mathrm{e}^{tH_m} e_1$ solves the differential equation

$$y'_{0,m}(t) = H_m y_{0,m}(t), \quad y_{0,m}(0) = \beta e_1, \tag{3.2.10}$$

For later use we introduce the notation

$$\widehat{y}_{p,m}(t) = t^p y_{p,m}(t), \tag{3.2.11a}$$

which for $p \in \mathbb{N}$ and according to (3.2.4) satisfies the differential equation

$$\widehat{y}'_{p,m}(t) = H_m \widehat{y}_{p,m}(t) + \tfrac{t^{p-1}}{(p-1)!}\beta e_1, \quad \widehat{y}_{p,m}(0) = 0. \tag{3.2.11b}$$

Remark 3.2.3. *Although we take rounding effects in the Arnoldi decomposition into account, we do not give a full study of round-off errors at this point. Round-off errors in substeps such as the evaluation of $y_{p,m}(t)$ or the matrix-vector multiplication $V_m y_{p,m}(t)$ will be ignored. We refer to [Hig02] for a more general study of these effects.*

Remark 3.2.4. *In the special cases $A = B$ or $A = \mathrm{i}B$ for a Hermitian matrix $B \in \mathbb{C}^{n\times n}$ (with A being skew-Hermitian in the latter case) the orthogonalization of the Krylov basis of $\mathcal{K}_m(B, v)$ simplifies to a three-term recursion, the so-called Lanczos method. In the skew-Hermitian case ($A = \mathrm{i}B$) the Krylov propagator* (3.2.8a) *can be evaluated by $\beta V_m \varphi_p(\mathrm{i}tH_m)e_1$, i.e., we approximate the function $\lambda \mapsto \varphi_p(\mathrm{i}t\lambda)$ in the Krylov subspace $\mathcal{K}_m(B, v)$. The advantage is a cheaper computation of the Krylov subspace in terms of computational cost and better conservation of geometric properties. For details we refer to the notation $\mathrm{e}^{\sigma tB}$ as introduced in Chapter 2, with $\sigma = \pm\mathrm{i}$ and a Hermitian matrix B for the skew-Hermitian case.*

The error of the Krylov propagator. We denote the error of the Krylov propagator given in (3.2.9) by

$$l_{p,m}(t) = \beta V_m \varphi_p(tH_m)e_1 - \varphi_p(tA)v, \quad p \in \mathbb{N}_0. \tag{3.2.12}$$

We are further interested in computable a posteriori estimates for the error norm, $\zeta_{p,m}(t) \approx \|l_{p,m}(t)\|_2$, which in the best case can be proven to be upper bounds on the error norm $\|l_{p,m}(t)\|_2 \le \zeta_{p,m}(t)$. Norm estimates of the error (3.2.12) can be used in practice to stop the Krylov iteration after k steps if $\|l_{p,k}(t)\|_2$ satisfies (3.2.13) below, or to restrict the time-step t to obtain an accurate approximation and restart the method with the remaining time. For details on the total error with this restarting approach see also [Sid98, JAK20].

A prominent task is to test if the error norm per unit step is bounded by a tolerance tol,

$$\zeta_{p,m}(t) \le t \cdot \mathrm{tol}, \quad \text{which should entail} \quad \|l_{p,m}(t)\|_2 \le t \cdot \mathrm{tol}. \tag{3.2.13}$$

In case of $\zeta_{p,m}(t)$ being an upper bound on the error norm, this results in a reliable bound.

3.3 An integral representation for the error of the Krylov propagator

We proceed with discussing the error $l_{p,m}$ of the Krylov propagator. To this end we first define its scalar *defect* by

$$\delta_{p,m}(t) = \beta e_m^* t^p \varphi_p(tH_m)e_1 = t^p \big(y_{p,m}(t)\big)_m \in \mathbb{C}, \tag{3.3.1a}$$

and the *defect integral* by[4]

$$L_{p,m}(t) = \frac{h_{m+1,m}}{t^p} \int_0^t |\delta_{p,m}(s)|\,\mathrm{d}s \in \mathbb{R}. \tag{3.3.1b}$$

Theorem 3.3.1. *Let $\delta_{p,m}(t) \in \mathbb{C}$ be the defect defined in* (3.3.1a)*. For $y_{p,m}(t) \in \mathbb{C}^m$ defined in* (3.2.9) *and a numerical perturbation $U_m \in \mathbb{C}^{n\times m}$ of the Arnoldi decomposition (see* (3.2.6)*), we have:*

(a) The error $l_{p,m}(t)$ of the Krylov propagator (see (3.2.12)*) enjoys the integral representation*

$$l_{p,m}(t) = -\frac{h_{m+1,m}}{t^p} \int_0^t \mathrm{e}^{(t-s)A} v_{m+1}\delta_{p,m}(s)\,\mathrm{d}s - \frac{1}{t^p}\int_0^t \mathrm{e}^{(t-s)A} U_m s^p y_{p,m}(s)\,\mathrm{d}s. \tag{3.3.2a}$$

(b) For given machine precision ε and constants C_1, C_2 representing round-off effects (see (3.2.7a),(3.2.7b)*), and with $\kappa_1 = \max_{s\in[0,t]} \|\mathrm{e}^{sA}\|_2$ and $\kappa_2 = \max_{s\in[0,t]} \|\mathrm{e}^{sH_m}\|_2$ the error norm is bounded by*

$$\|l_{p,m}(t)\|_2 \le (1 + C_2\varepsilon)\kappa_1 L_{p,m}(t) + C_1\varepsilon\|A\|_2 \frac{\beta\kappa_1\kappa_2 t}{(p+1)!}, \tag{3.3.2b}$$

with the defect integral $L_{p,m}(t)$ defined in (3.3.1b)*.*

Proof.

(a) For the exact matrix function we use the notation

$$u_p(t) = \varphi_p(tA)v, \quad \text{and} \quad w_p(t) = t^p u_p(t).$$

For the Krylov propagator we denote

$$u_{p,m}(t) = V_m y_{p,m}(t) \quad \text{with} \quad y_{p,m}(t) = \beta\varphi_p(tH_m)e_1$$

(see (3.2.9)), and we also define

$$w_{p,m}(t) = t^p u_{p,m}(t) = V_m \widehat{y}_{p,m}(t), \quad \text{with} \quad \widehat{y}_{p,m}(t) = t^p y_{p,m}(t) \ \text{ defined in (3.2.11a).}$$

[4] This and the result of Theorem 3.3.1 remain valid for the case $t = 0$.

- For $p \in \mathbb{N}$, the functions $w_p(t)$ and $w_{p,m}(t)$ satisfy the differential equations (see (3.2.4), (3.2.11b))

$$\begin{aligned} w'_{p,m}(t) &= V_m \widehat{y}'_{p,m}(t) = V_m\big(H_m \widehat{y}_{p,m}(t) + \tfrac{t^{p-1}}{(p-1)!}\beta e_1\big), \\ w'_p(t) &= A w_p(t) + \tfrac{t^{p-1}}{(p-1)!} v, \quad \text{and} \quad w_p(0) = w_{p,m}(0) = 0. \end{aligned} \tag{3.3.3}$$

- For $p = 0$, i.e., $w_0(t) = u_0(t)$ and $w_{0,m}(t) = V_m y_{0,m}(t)$, according to (3.2.10) we have

$$\begin{aligned} &w'_0(t) = A w_0(t), \quad w'_{0,m}(t) = V_m H_m y_{0,m}(t), \\ &\text{and} \quad w_0(0) = v, \quad w_{0,m}(0) = \beta V_m e_1 = v. \end{aligned}$$

Local error representation in terms of the defect. We defined the re-scaled error

$$\widehat{l}_{p,m}(t) = w_{p,m}(t) - w_p(t) = t^p l_{p,m}(t).$$

- For $p \in \mathbb{N}$ this satisfies

$$\widehat{l}'_{p,m}(t) = w'_{p,m}(t) - w'_p(t) = A\,\widehat{l}_{p,m}(t) + d_{p,m}(t), \quad \widehat{l}_{p,m}(0) = 0, \tag{3.3.4}$$

  with the *defect* of $w_{p,m}(t)$ with respect to the differential equation (3.3.3),

$$\begin{aligned} d_{p,m}(t) &= w'_{p,m}(t) - A w_{p,m}(t) - \tfrac{t^{p-1}}{(p-1)!} v \\ &= V_m\big(H_m \widehat{y}_{p,m}(t) + \tfrac{t^{p-1}}{(p-1)!}\beta e_1\big) - A V_m \widehat{y}_{p,m}(t) - \tfrac{t^{p-1}}{(p-1)!} v \\ &= \big(V_m H_m - A V_m\big)\widehat{y}_{p,m}(t) + \tfrac{t^{p-1}}{(p-1)!}(\beta V_m e_1 - v). \end{aligned}$$

  Together with (3.2.6) and using of $\beta V_m e_1 = v$ the defect can be written as

$$d_{p,m}(t) = -h_{m+1,m}(e_m^* \widehat{y}_{p,m}(t)) v_{m+1} - U_m \widehat{y}_{p,m}(t).$$

- For $p = 0$, in an analogous way we obtain

$$d_{0,m}(t) = -h_{m+1,m}(e_m^* y_{0,m}(t)) v_{m+1} - U_m y_{0,m}(t).$$

We conclude

$$d_{p,m}(t) = -h_{m+1,m}\delta_{p,m}(t) v_{m+1} - t^p U_m y_{p,m}(t), \quad p \in \mathbb{N}_0, \tag{3.3.5}$$

with the scalar defect defined in (3.3.1a). Due to (3.3.4) we have

$$\widehat{l}_{p,m}(t) = \int_0^t \mathrm{e}^{(t-s)A} d_{p,m}(s)\mathrm{d}s, \quad p \in \mathbb{N}_0,$$

and for $l_{p,m}(t) = t^{-p}\widehat{l}_{p,m}(t)$ together with (3.3.5) this implies (3.3.2a).

(b) With $\kappa_1 = \max_{s\in[0,t]} \|\mathrm{e}^{sA}\|_2$, $\|U_m\|_2 \le C_1\varepsilon\|A\|_2$ and $\|v_{m+1}\|_2 \le 1 + C_2\varepsilon$, the representation (3.3.2a) implies the upper bound

$$\begin{aligned}\|l_{p,m}(t)\|_2 \le (1 + C_2\varepsilon)\kappa_1 \frac{h_{m+1,m}}{t^p} \int_0^t |\delta_{p,m}(s)|\,\mathrm{d}s \\ + C_1\varepsilon\|A\|_2 \frac{\kappa_1}{t^p} \int_0^t s^p \|y_{p,m}(s)\|_2\,\mathrm{d}s.\end{aligned} \tag{3.3.6}$$

With the defect integral $L_{p,m}(t)$ defined in (3.3.1b) we obtain the first term in (3.3.2b). For the second integral term (with $y_{p,m}(t) = \beta\varphi_p(tH_m)e_1$) we use the upper bound

$$\int_0^t s^p \|\varphi_p(sH_m)e_1\|_2\,\mathrm{d}s \le \max_{s\in[0,t]} \|\varphi_p(sH_m)e_1\|_2 \frac{t^{p+1}}{p+1}. \tag{3.3.7}$$

- For $p \in \mathbb{N}$ we apply the integral representation due to (3.2.3) for $\varphi_p(tH_m)e_1$ to obtain the norm bound

$$\max_{s\in[0,t]} \|\varphi_p(sH_m)e_1\|_2 \le \frac{\max_{s\in[0,t]} \|\mathrm{e}^{sH_m}\|_2}{(p-1)!} \int_0^1 \theta^{p-1}\,\mathrm{d}\theta = \frac{\max_{s\in[0,t]} \|\mathrm{e}^{sH_m}\|_2}{p!}. \tag{3.3.8}$$

- For $p = 0$ we obtain (3.3.8) in a direct way.

Combining (3.3.7) with (3.3.8) and denoting $\kappa_2 = \max_{s\in[0,t]} \|\mathrm{e}^{sH_m}\|_2$ we obtain

$$\frac{\kappa_1}{t^p} \int_0^t s^p \|y_{p,m}(s)\|_2\,\mathrm{d}s \le \frac{\beta\kappa_1\kappa_2 t}{(p+1)!}.$$

Combining these estimates with (3.3.6) we conclude (3.3.2b).

□

Remark 3.3.2. *The error norm of the Krylov propagator scales with* $\kappa_1 = \max_{s\in[0,t]} \|\mathrm{e}^{sA}\|_2$ *and* $\kappa_2 = \max_{s\in[0,t]} \|\mathrm{e}^{sH_m}\|_2$ *in a natural way.* [5] *It is well known that*

$$\|\mathrm{e}^{tA}\|_2 \le \mathrm{e}^{t\mu_2(A)} \quad \textit{with the logarithmic norm}$$
$$\mu_2(A) = \max\{\mathrm{Re}(\mathrm{W}(A))\} = \max\{\mathrm{spec}(A + A^*)/2\},$$

see for instance [Hig08, Theorem 10.11]. Problems with $\mu_2(A) > 0$ *can be arbitrary ill-conditioned and difficult to solve with proper accuracy. (For further results on the stability of the matrix exponential see also [MVL03, VL77].) We will not further discuss problems with* $\mu_2(A) > 0$ *and assume* $\mu_2(A) \le 0$. *We refer to the case* $\mu_2(A) \le 0$ *as the dissipative case, with* $\kappa_1 = 1$.

[5] Taking the maximum $\max_{s\in[0,t]}$ in the definition of κ_1 and κ_2 is necessary to cover the case $p > 0$. For the special case $p = 0$ the upper norm bound given in Theorem 3.3.1 can be adapted to scale with $\mathrm{e}^{t\mu_2(A)}$.

For the dissipative case with $\mu_2(A) \leq 0$ the error bound (3.3.2b) from Theorem 3.3.1 reads

$$\|l_{p,m}(t)\|_2 \leq (1 + C_2\varepsilon)L_{p,m}(t) + C_1\varepsilon\|A\|_2\frac{\beta\kappa_2 t}{(p+1)!}. \tag{3.3.9}$$

The dissipative behavior of e^{tA} carries over to the Krylov propagator up to a perturbation which depends on round-off errors, including the loss of orthogonality of V_m. In terms of the numerical range $\mathrm{W}(H_m)$, with $\mathrm{W}(H_m) \subseteq U_{C_3\varepsilon}(\mathrm{W}(A))$ we have $\mu_2(H_m) \leq \mu_2(A) + C_3\varepsilon$, for a constant $C_3\varepsilon$ depending on round-off effects (3.2.7c). Thus, $\mu_2(H_m) \leq C_3\varepsilon$ and $\kappa_2 \leq \mathrm{e}^{tC_3\varepsilon}$.

Our aim is to construct an upper norm bound for the error per unit step (3.2.13) via (3.3.9). Let the tolerance tol be given and t be a respective time step for (3.2.13). Then the round-off error terms in (3.3.9) are negligible if

$$C_2\varepsilon \ll 1, \quad \text{and} \quad C_1\varepsilon\|A\|_2\beta\mathrm{e}^{tC_3\varepsilon}/(p+1)! \ll \mathrm{tol}. \tag{3.3.10}$$

Concerning the constants C_1, C_2 and C_3 see (3.2.7). We recapitulate that C_1 and C_2 given in (3.2.7a) and (3.2.7b) can be considered to be small enough in a standard Krylov setting. The constant C_3 can be larger in the case of a loss of orthogonality of the Krylov subspace, which can however be avoided at the cost of additional computational effort. The constant C_3 only appears as an exponential prefactor for the round-off term in (3.3.10) and is less critical compared to C_1 and C_2.

With the previous observation on the round-off errors taken into account in (3.3.9) we consider the following upper bound to be stable in computer arithmetic in accordance to a proper value of tol, see (3.3.10).

Corollary 3.3.3. *For the case $\mu_2(A) \leq 0$ and with the assumption that round-off error is negligible, the error of the Krylov propagator is bounded by the defect integral $L_{p,m}(t)$,*

$$\|l_{p,m}(t)\|_2 \leq \frac{h_{m+1,m}}{t^p}\int_0^t |\delta_{p,m}(s)|\,\mathrm{d}s = L_{p,m}(t), \quad p \in \mathbb{N}_0.$$

Note that the defect norm $|\delta_{p,m}(s)|$ cannot be integrated exactly in general. This point will further be studied in the sequel.

Representing the defect in terms of divided differences. Divided differences play an essential role in this work. We use the notation

$$f[\lambda_1, \ldots, \lambda_m]$$

for the divided differences of a function f over the nodes $\lambda_1, \ldots, \lambda_m$. (This is to be understood in the confluent sense for the case of multiple nodes λ_j, see for instance [Hig08, Section B.16].)

Theorem 3.3.4 (see for instance [CM97])**.** *Let $H_m \in \mathbb{C}^{m\times m}$ be an upper Hessenberg matrix with positive secondary diagonal entries, $0 < (H_m)_{j+1,j} \in \mathbb{R}$ for $j = 1, \ldots, m-1$, and eigenvalues $\lambda_1, \ldots, \lambda_m$. Let f be an analytic function for which $f(H_m)$ is well defined. Then,*

$$e_m^* f(H_m)e_1 = \gamma_m f[\lambda_1, \ldots, \lambda_m],$$

with $\gamma_m = \prod_{j=1}^{m-1}(H_m)_{j+1,j}$.

For $f = (\varphi_p)_t : \lambda \mapsto \varphi_p(t\lambda)$ we will also make use of the following result. [6]

Theorem 3.3.5 (Corollary 1 in [Sid98]; expressing φ-functions via dilated exp-functions)**.** *For $t \in \mathbb{R}$,*

$$t^p e_m^* \varphi_p(tH_m) e_1 = e_{m+p}^* \exp(t\widetilde{H}_{p,m}) e_1$$

with

$$\widetilde{H}_{p,m} = \begin{pmatrix} H_m & 0_{m\times p} \\ e_1 e_m^* & J_{p\times p} \end{pmatrix} \in \mathbb{C}^{(m+p)\times(m+p)} \quad \text{and} \quad J_{p\times p} = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} \in \mathbb{C}^{p\times p}.$$

The matrix $\widetilde{H}_{p,m}$ in Theorem 3.3.5 is block triangular with eigenvalues equal to those of H_m and $J_{p\times p}$. Therefore, $\operatorname{spec}(\widetilde{H}_m) = \{\lambda_1, \dots, \lambda_m, 0, \dots, 0\}$, with 0 as an eigenvalue of multiplicity p (at least). In our context, $\widetilde{H}_m$ is upper Hessenberg with a positive lower secondary diagonal and $\gamma_m = \prod_{j=1}^{m-1} (H_m)_{j+1,j} = \prod_{j=1}^{m+p-1} (\widetilde{H}_m)_{j+1,j}$. In accordance with Theorem 3.3.4 the result of Theorem 3.3.5 holds for divided differences in a similar manner,

$$t^p (\varphi_p)_t[\lambda_1, \dots, \lambda_m] = \exp_t[\lambda_1, \dots, \lambda_m, \underbrace{0, \dots, 0}_{p \text{ times}}].$$

With Theorem 3.3.4 and 3.3.5 the following equivalent formulations can be used the rewrite the scalar defect $\delta_{p,m}(t)$ defined in (3.3.1a).

Corollary 3.3.6. *Let $\delta_{p,m}(t)$ be the scalar defect given in* (3.3.1a) *for the upper Hessenberg matrix $H_m \in \mathbb{C}^{m\times m}$ with positive secondary diagonal entries. Denote $0 < \gamma_m = \prod_{j=1}^{m-1} (H_m)_{j+1,j}$. Let $\widetilde{H}_{p,m} \in \mathbb{C}^{m+p}$ be given as in Theorem 3.3.5. For the scalar defect we obtain the following equivalent formulations:*

(i) $\delta_{p,m}(t) = \beta e_m^* t^p \varphi_p(tH_m) e_1$

(ii) $= \beta \gamma_m t^p (\varphi_p)_t[\lambda_1, \dots, \lambda_m]$

(iii) $= \beta e_{m+p}^* \exp(t\widetilde{H}_{p,m}) e_1$

(iv) $= \beta \gamma_m \exp_t[\lambda_1, \dots, \lambda_m, 0_p]$ [7]

We remark that the eigenvalues $\lambda_1, \dots, \lambda_m$ of the Krylov Hessenberg matrix H_m are also referred to as Ritz values (of A) in the literature.

[6] Theorem 3.3.5 can be generalized to the case $t^p e_m^* \varphi_{k+p}(tH_m) e_1 = e_{m+p}^* \varphi_k(t\widetilde{H}_{p,m}) e_1$ with $k \in \mathbb{N}$, see [AMH11, Theorem 2.1]. The case $k = 0$ is sufficient for our purpose.

[7] Here we introduce the notation $(\lambda_1, \dots, \lambda_m, 0_p) = (\lambda_1, \dots, \lambda_m, 0, \dots, 0) \in \mathbb{C}^{m+p}$ for $p \in \mathbb{N}_0$.

3.4 Computable a posteriori error bounds for the Krylov propagator

The following two propositions are used for the proof of Theorem 3.4.3 below.[8]

Proposition 3.4.1. *For arbitrary nodes $\lambda_j \in \mathbb{C}$ and $p \in \mathbb{N}_0$,*

$$\int_0^t s^p(\varphi_p)_s[\lambda_1, \ldots, \lambda_k]\,\mathrm{d}s = t^{p+1}(\varphi_{p+1})_t[\lambda_1, \ldots, \lambda_k].$$

Proof. See Appendix 3.B. □

Proposition 3.4.2 (Lemma including eq. (5.1.1) in [MNP84])**.** *For arbitrary nodes $\lambda_j = \xi_j + \mathrm{i}\eta_j \in \mathbb{C}$,*

$$|\exp_t[\lambda_1, \ldots, \lambda_k]| \leq \exp_t[\xi_1, \ldots, \xi_k].$$

Proof. See Appendix 3.B. □

We now derive upper bounds for the error via its representation by the defect integral (3.3.1b).

Theorem 3.4.3. *Let $p \in \mathbb{N}_0$, $\mu_2(A) \leq 0$, and assume that round-off errors are sufficiently small (see Corollary 3.3.3). For the eigenvalues of H_m we write $\lambda_j = \xi_j + \mathrm{i}\eta_j$, $j = 1, \ldots, m$. An upper bound on the error norm is given by*

$$\|l_{p,m}(t)\|_2 \leq \beta h_{m+1,m}\gamma_m t(\varphi_{p+1})_t[\xi_1, \ldots, \xi_m]. \tag{3.4.1}$$

Proof. Due to Corollary 3.3.6, (iv),

$$\delta_{p,m}(t) = \beta\gamma_m \exp_t[\lambda_1, \ldots, \lambda_m, 0_p]. \tag{3.4.2a}$$

The divided differences in (3.4.2a) span over complex nodes $\lambda_1, \ldots, \lambda_m$ and $0_p \in \mathbb{C}^p$, with real parts $\xi_1, \ldots, \xi_m$. Propositions 3.4.2 and 3.4.1 imply

$$\int_0^t |\exp_s[\lambda_1, \ldots, \lambda_m, 0_p]|\,\mathrm{d}s \leq \int_0^t \exp_s[\xi_1, \ldots, \xi_m, 0_p]\,\mathrm{d}s = t(\varphi_1)_t[\xi_1, \ldots, \xi_m, 0_p]. \tag{3.4.2b}$$

From Corollary 3.3.6 we obtain

$$t(\varphi_1)_t[\xi_1, \ldots, \xi_m, 0_p] = \exp_t[\xi_1, \ldots, \xi_m, 0_{p+1}] = t^{p+1}(\varphi_{p+1})_t[\xi_1, \ldots, \xi_m]. \tag{3.4.2c}$$

Eqs. (3.4.2a)–(3.4.2c) together with Corollary 3.3.3 imply (3.4.1). □

For the case of H_m having real eigenvalues, the assertion of Theorem 3.4.3 can be reformulated in the following way (see Proposition 2.5.5 in Chapter 2).

[8] We use the notation introduced in the previous sections.

Corollary 3.4.4. *Assume $\mu_2(A) \leq 0$ and that round-off errors are sufficiently small (see Corollary 3.3.3). For the case of H_m having real eigenvalues $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$, the upper bound on the error norm in Theorem 3.4.3 yields an exact evaluation of the defect integral. Hence,*

$$\|l_{p,m}(t)\|_2 \leq L_{p,m}(t) = \beta h_{m+1,m} t\big(e_m^* \varphi_{p+1}(tH_m) e_1\big).$$

As a further corollary we formulate an upper bound on the error norm which is cheaper to evaluate compared to the bound from Theorem 3.4.3 but may be less tight. Using the Mean Value Theorem, [Hig08, eq. (B.26)] or [dB05, eq. (44)], for the divided differences in Theorem 3.4.3, eq. (3.4.1) we obtain the following result which corresponds to Theorem 2.3.2 and Theorem 2.4.1 in Chapter 2. For the exponential of a skew-Hermitian matrix a similar error estimate has been used in [KBC05] and is based on ideas of [PL86] with some lack of theory.

Corollary 3.4.5. *Let $p \in \mathbb{N}_0$, $\mu_2(A) \leq 0$, and assume that round-off errors are sufficiently small (see Corollary 3.3.3). Let $\xi_{\max} = 0$ for $p \in \mathbb{N}$ and $\xi_{\max} = \max_{j=1,\ldots,m} \xi_j \leq 0$ for $p = 0$ and eigenvalues $\lambda_j = \xi_j + \mathrm{i}\mu_j \in \mathbb{C}$ of H_m. An upper bound on the error norm is given by*

$$\|l_{p,m}(t)\|_2 \leq \beta h_{m+1,m} \frac{\gamma_m t^m \mathrm{e}^{t\xi_{\max}}}{(m+p)!} \leq \beta h_{m+1,m} \frac{\gamma_m t^m}{(m+p)!}.$$

For the case of H_m having purely imaginary eigenvalues, the divided differences in Theorem 3.4.3 (see (3.4.1)) can be evaluated directly via [Hig08, eq. (B.27)],

$$t(\varphi_{p+1})_t[0_m] = t^{-p} \exp_t[0_{m+p+1}] = \frac{t^m}{(m+p)!},$$

hence the assertions of Theorem 3.4.3 and Corollary 3.4.5 coincide in this case.

Accuracy of the previously specified upper bounds on the error norm. In the following we again denote $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$ for the eigenvalues of H_m, with $\lambda_j = \xi_j + \mathrm{i}\eta_j$. For the scalar defect $\delta_{p,m}(t)$ (see (3.3.1a)) we recapitulate Corollary 3.3.6, in particular

$$\delta_{p,m}(t) = \beta\gamma_m t^p (\varphi_p)_t[\lambda_1, \ldots, \lambda_m] = \beta\gamma_m \exp_t[\lambda_1, \ldots, \lambda_m, 0_p]. \tag{3.4.3}$$

Theorem 3.4.3 and its corollaries make use of the error bound given in Corollary 3.3.3 and computable upper bounds on the defect integral $L_{p,m}(t)$. A refinement of the upper bound from Corollary 3.3.3 would require further applications of the large-dimensional matrix-vector product with $A \in \mathbb{C}^{n\times n}$ and has been shown to be inefficient in terms of computational cost, see also Remark 2.8.2 in Chapter 2. The computable upper bounds on the defect integral $L_{p,m}(t)$ will be further discussed. We recapitulate the upper bound of the divided differences given in Proposition 3.4.2,

$$|\exp_t[\lambda_1, \ldots, \lambda_m, 0_p]| \leq \exp_t[\xi_1, \ldots, \xi_m, 0_p]. \tag{3.4.4}$$

Thus, in the case of H_m having eigenvalues with a sufficiently small imaginary part, the upper bound in Proposition 3.4.2, is tight. In the following proposition this statement is made more precise.

Proposition 3.4.6 (Part of a proof in [MNP84], eq. (5.2.3)). *For nodes* $\lambda_j = \xi_j + \mathrm{i}\eta_j \in \mathbb{C}$ *and* $t \geq 0$ *with* $\max_j t|\eta_j| \leq \widetilde{\eta}_t < \pi/2$,

$$0 < \cos(\widetilde{\eta}_t)\exp_t[\xi_1, \ldots, \xi_k] \leq |\exp_t[\lambda_1, \ldots, \lambda_k]|.$$

Proof. See Appendix 3.B. □

Under the assumptions of Proposition 3.4.6 we conclude

$$0 < \cos(\widetilde{\eta}_t)\exp_t[\xi_1, \ldots, \xi_m, 0_p] \leq |\exp_t[\lambda_1, \ldots, \lambda_m, 0_p]|. \tag{3.4.5}$$

With (3.4.3), (3.4.4), (3.4.5) and following the proof of Theorem 3.4.3 the defect integral in (3.3.1b) can be enclosed by

$$\begin{aligned} 0 < \cos(\widetilde{\eta}_t) \cdot \beta\gamma_m h_{m+1,m} t(\varphi_{p+1})_t[\xi_1, \ldots, \xi_m] \\ \leq L_{p,m}(t) \leq \beta\gamma_m h_{m+1,m} t(\varphi_{p+1})_t[\xi_1, \ldots, \xi_m]. \end{aligned} \tag{3.4.6}$$

Hence,

$$L_{p,m}(t) = \big(1 - \mathcal{O}(|t\eta|^2)\big)\beta\gamma_m h_{m+1,m} t(\varphi_{p+1})_t[\xi_1, \ldots, \xi_m], \tag{3.4.7}$$

using the notation $\mathcal{O}(|t\eta|^2)$ in the sense of $\mathcal{O}(|t\eta|) = \mathcal{O}(\max_j t|\eta_j|)$ for $t|\eta_j| \to 0$. Following Proposition 3.4.6 the choice of $\widetilde{\eta}_t$ is independent of $\xi_1, \ldots, \xi_m$, and this carries over to the constant in (3.4.7).

Summarizing, we see that the defect integral can be computed exactly for the case of H_m having real eigenvalues (Corollary 3.4.4), and a computable upper bound can be given which is tight for the case of H_m having eigenvalues sufficiently close to the real axis (Theorem 3.4.3 and eq. (3.4.7)).

The approach underlying Theorem 3.4.3 does not enable us to specify the asymptotic constant in (3.4.7). Therefore, we use the asymptotic expansion of the divided differences, $|\exp_t[\lambda_1, \ldots, \lambda_m, 0_p]|$ in (3.4.3), derived in Appendix 3.C, to discuss the asymptotic behavior of the defect norm $|\delta_{p,m}(t)|$ for $t \to 0$. Theorem 3.C.2 from Appendix 3.C implies

$$\begin{aligned} &|\exp_t[\lambda_1, \ldots, \lambda_m, 0_p]| = \frac{t^{m+p-1}}{(m+p-1)!}\exp\big(\rho_1 t + \rho_2 t^2/2 + \mathcal{O}(t^3)\big), \\ &\text{with} \quad \rho_1 = \mathrm{avg}_p(\xi) \quad \text{and} \quad \rho_2 = \frac{\mathrm{var}_p(\xi) - \mathrm{var}_p(\eta)}{m+p+1}. \end{aligned} \tag{3.4.8}$$

Here, the asymptotics holds for $t \to 0$, $\mathrm{avg}_p(\xi) = \sum_{j=1}^m \xi_j/(m+p)$ is the average, and $\mathrm{var}_p(\xi) = \big(\sum_{j=1}^m (\xi_j - \mathrm{avg}_p(\xi))^2 + p\,\mathrm{avg}_p(\xi)^2\big)/(m+p)$ is the variance of the sequence $\{\xi_1, \ldots, \xi_m, 0_p\}$ and $\mathrm{var}_p(\eta)$ is the variance of the sequence $\{\eta_1, \ldots, \eta_m, 0_p\}$.

Remark 3.4.7. *For* H_m *with purely imaginary eigenvalues* ($\lambda_j \in \mathrm{i}\mathbb{R}$), *e.g., in the skew-Hermitian case, the following asymptotic expansion for the defect is obtained from* (3.4.8), [9]

$$|\delta_{p,m}(t)| = \beta\gamma_m \frac{t^{m+p-1}}{(m+p-1)!}\exp\Big(-\frac{\mathrm{var}_p(\eta)}{2(m+p+1)}t^2 + \mathcal{O}(t^3)\Big) \quad \textit{for} \quad t \to 0. \tag{3.4.9}$$

[9] It can be shown that the remainder is of even order $\mathcal{O}(t^4)$ in this case.

We use the expansion from (3.4.8) for $|\exp_t[\lambda_1,\dots,\lambda_m,0_p]|$ and $\exp_t[\xi_1,\dots,\xi_m,0_p]$ to obtain

$$|\delta_{p,m}(t)| = \exp\Big(-\frac{\mathrm{var}_p(\eta)}{2(m+p+1)}t^2+\mathcal{O}(t^3)\Big)\cdot\beta\gamma_m t^p(\varphi_p)_t[\xi_1,\dots,\xi_m]. \tag{3.4.10}$$

Termwise integration of (3.4.10) and the proper prefactor gives an asymptotic expansion for the defect integral $L_{p,m}(t)$, similar to (3.4.7),

$$L_{p,m}(t) = \Big(1-\frac{\mathrm{var}_p(\eta)(m+p)t^2}{2(m+p+1)(m+p+2)}+\mathcal{O}(t^3)\Big)\cdot\beta h_{m+1,m}\gamma_m t(\varphi_{p+1})_t[\xi_1,\dots,\xi_m]. \tag{3.4.11}$$

Omitting further details we state that (3.4.11) is to be understood in an asymptotic sense with an remainder of $\mathcal{O}(t^3|\xi||\eta|^2+t^4|\eta|^4)$. In contrast to (3.4.7) the remainder is depending on ξ terms but (3.4.11) reveals further constants which can be relevant for practical applications.

Remark 3.4.8. *With* (3.4.11) *we obtain a computable estimate for the relative deviation from the defect integral to the upper bound in* (3.4.6)*. The criterion*

$$ac.est.1(t) := \frac{\mathrm{var}_p(\eta)(m+p)t^2}{2(m+p+1)(m+p+2)} > 0.1,$$

can indicate that a tighter estimate on the defect integral could improve the error bound given in Theorem 3.4.3 in terms of accuracy. A possible choice are quadrature estimates on the defect integral, see Subsection 3.4.1 below.

A similar criterion can be given for the accuracy of the upper bound,

$$L_{p,m}(t) \le \beta h_{m+1,m}\gamma_m\frac{t^m}{(m+p)!}, \tag{3.4.12}$$

which appears in Corollary 3.4.5 (with $\xi_{\max}=0$), and Theorem 2.3.2 and Theorem 2.4.1 in Chapter 2.
With (3.4.8), and ρ_1 and ρ_2 given therein, the defect integral can be written as

$$L_{p,m}(t) = \beta h_{m+1,m}\gamma_m\frac{t^m}{(m+p)!}\Big(1+\rho_1\frac{(m+p)t}{m+p+1}+(\rho_1^2+\rho_2)\frac{(m+p)t^2}{2(m+p+2)}+\mathcal{O}(t^3)\Big) \tag{3.4.13}$$

for $t\to 0$. In contrast to the error bound in Corollary 3.4.5, the formulas for ρ_1 and ρ_2 in (3.4.8) require the evaluation of the eigenvalues of H_m. The following Proposition gives a formula for ρ_1 and ρ_2 which does not require computation of the eigenvalues of H_m and can be evaluated on the fly.

Proposition 3.4.9 (Evaluation of ρ_1 and ρ_2 in terms of entries of H_m)**.** *The coefficients ρ_1 and ρ_2 in* (3.4.8) *can be rewritten as*

$$\rho_1 = \frac{\mathrm{Re}(S_1)}{m+p},\quad \rho_2 = \frac{\mathrm{Im}(S_1)^2-\mathrm{Re}(S_1)^2}{(m+p)^2}+\frac{\mathrm{Re}(S_1^2+S_2)}{(m+p)(m+p+1)},\quad \textit{with}$$

$$S_1 = \sum_{j=1}^{m}(H_m)_{j,j}\quad \textit{and}\quad S_2 = \sum_{j=1}^{m}(H_m)_{j,j}^2+2\sum_{j=1}^{m-1}(H_m)_{j+1,j}(H_m)_{j,j+1}.$$

Proof. For the coefficients ρ_1 and ρ_2 we use (3.C.17) with $m \leftarrow m+p$ and S_1 and S_2 from (3.C.3). For the nodes $\lambda_1, \ldots, \lambda_m, 0_p$ (with $\lambda_1, \ldots, \lambda_m$ eigenvalues of H_m) we obtain

$$\begin{aligned} S_1 &= \sum_{j=1}^{m} \lambda_j = \mathrm{Trace}(H_m) = \sum_{j=1}^{m} (H_m)_{j,j} \quad \text{and} \\ S_2 &= \sum_{j=1}^{m} \lambda_j^2 = \mathrm{Trace}(H_m^2) = \sum_{j=1}^{m} (H_m)_{j,j}^2 + 2\sum_{j=1}^{m-1} (H_m)_{j+1,j}(H_m)_{j,j+1}. \end{aligned} \tag{3.4.14}$$

The identity for $\mathrm{Trace}(H_m^2)$ in (3.4.14) holds true due to the upper Hessenberg structure of H_m. □

Following the proof of Theorem 3.C.2 we observe that the case $\rho_1 = 0$ is possible but results in $\rho_2 \neq 0$.

Remark 3.4.10. *With* (3.4.13) *and Proposition 3.4.9 we obtain a computable estimate for the relative deviation from the defect integral to the upper bound in* (3.4.12)*. The criterion*

$$ac.est.2(t) := \left| \rho_1 \frac{(m+p)t}{m+p+1} + (\rho_1^2 + \rho_2) \frac{(m+p)t^2}{2(m+p+2)} \right| > 0.1$$

can indicate that a tighter estimate on the defect integral could improve the error bound given in Corollary 3.4.5 in terms of accuracy. We refer to the error bound in Theorem 3.4.3 in case the eigenvalues of H_m have a significant real part (which can be observed via ρ_1).

3.4.1 Quadrature-based error estimates

First we recapitulate some prior results. In the dissipative case the integral formulation of the error from Theorem 3.3.1 can be bounded via the defect integral via Corollary 3.3.3 up to round-off. We conclude that the defect integral can be computed exactly for the case of H_m having real eigenvalues (Corollary 3.4.4), and a computable upper bound exists which is tight for the case of H_m having eigenvalues sufficiently close to the real axis (Theorem 3.4.3 and eq. (3.4.6)).

For the case of H_m having eigenvalues with a significant imaginary part, tight estimates are more difficult to obtain. It can be favorable to approximate the defect integral (3.3.1b) by quadrature to obtain an error estimate via Corollary 3.3.3. The aim of using quadrature is to obtain an error estimate which is tighter compared to previous upper norm bounds on the error. In contrast to the proven upper error bounds given in Theorem 3.4.3, Corollary 3.4.4 and 3.4.5 the following quadrature estimates do not result in upper error bounds in general. However, in many practical cases such quadrature estimates turn out to be still reliable.

Here, some remarks on the defect are in order to explain some subtleties with quadrature estimates for the defect integral $L_{p,m}(t)$. We discuss a test problem with a skew-Hermitian matrix $A \in \mathbb{C}^{n\times n}$. Following Remark 3.2.4 we choose $A = \mathrm{i}B$ with a Hermitian matrix B, in particularly, $B = \mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ with $n = 1000$. The matrix B is related to a finite difference discretization of the one-dimensional negative Laplacian operator and A

corresponds to a free Schrödinger type problem. The eigenvalues σ_j, for $j = 1, \dots, n$, of B are given by

$$\sigma_j = 4\sin(j\pi/(2(n+1)))^2 \quad \text{with respective eigenvector } \psi_j \in \mathbb{R}^n. \tag{3.4.15}$$

Here, $\mu_2(A) = 0$, and the conditions of Corollary 3.3.3 hold. For a given starting vector $v \in \mathbb{C}^n$ the time propagation for the discretized free Schrödinger equation is given by $\exp(tA)v$ and can be approximated by the Krylov propagator with $p = 0$. The following different cases for the starting vector v will be discussed.

(a) Choose a random starting vector $v \in \mathbb{R}^n$.

(b) Start close to a linear combination of eigenvectors, $v = 10^6 \sum_{j=1}^{25} \psi_j + \sum_{j=26}^{n} \psi_j$ for eigenvectors ψ_j of the discretized negative Laplacian operator, (3.4.15).

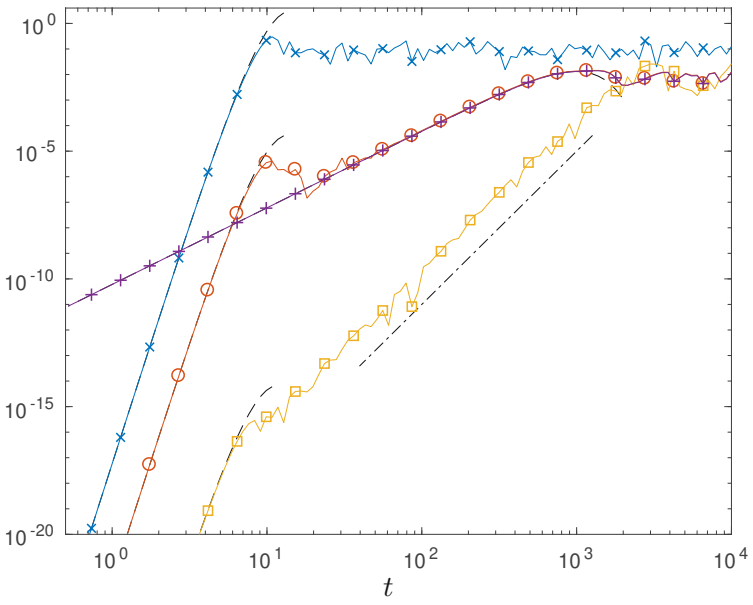
(c) Start close to a linear combination of eigenvectors which are more spread on the spectrum, $v = 10^5 \sum_{j=1}^{20} \psi_j + 10^5 \sum_{j=n-19}^{n} \psi_j$ for eigenvectors ψ_j of the discretized negative Laplacian operator, (3.4.15).

In addition to the setting from (a)–(c) we normalize v, $\|v\|_2 = 1$. The defect $\delta_{p,m}(t)$ for $p = 0$ is computed in MATLAB, using `expm` to evaluate the matrix exponential of H_m and divided differences for a fixed Krylov dimension $m = 20$.

In Figure 3.1 we observe $|\delta_{p,m}(t)| = \mathcal{O}(t^{m-1})$ (for $t \to 0$) up to $t \approx 10^1$ for the case (a)–(c). The values of $|\delta_{p,m}(t)|$ in this time regime vary strongly among these cases. We further remark that in the case (b) for $t \geq 4 \cdot 10^1$ the defect $|\delta_{p,m}(t)|$ behaves similar to the divided differences of the exponential over the first eigenvalues $\lambda_1^{(b)}, \dots, \lambda_4^{(b)}$ of H_m with a proper prefactor. This behavior occurs if eigenvalues of H_m are clustered, in this case $\lambda_1^{(b)}, \dots, \lambda_4^{(b)} \approx 0$, and will be further discussed below, see Figure 3.2. For the case (c) the eigenvalues of H_m are clustered at ≈ 0 and ≈ 4. Also in this case there is a time regime for which the defect behaves similar to a lower order function in t with some additional oscillations. (This may be explained by the existence of different eigenvalue clusters of the same size.)

As a conclusion from the example illustrated in Figure 3.1, we observe that quadrature of the defect can be relevant up to a time t for which the quadrature-based estimate of $\|l_{p,m}(t)\|_2$ (via the defect integral) is equal to a given tolerance, see (3.2.13). This regime of t would depend on the choice of tol and additional factors such as β, $h_{m+1,m}$ etc. which appear in the error bound from Corollary 3.3.3. Depending on parameters and the starting vector v the defect can be highly oscillatory for relevant times t and, respectively, a quadrature estimate of the defect integral can be difficult to obtain. Such effects seem to be relevant for special choices of starting vectors v, for example case (b) and (c). The effect of H_m having clustered eigenvalues and the prefactor used in Figure 3.1 (+) are explained in the following model problem, see Figure 3.2.

Divided differences with clustered nodes: an example. Choose $m = 3$ with nodes $a_1 = 1.123, a_2 = 1.231, a_3 = 5.43$. With this choice we obtain cluster of nodes, $a_1 \approx a_2$. For the given example we obtain $|\exp_t[ia_2, ia_3]| \ll |\exp_t[ia_1, ia_2]|$ for t large enough, hence, using



j	$\lambda_j^{(a)}$	$\lambda_j^{(b)}$	$\lambda_j^{(c)}$
1	0.0002	0.0003	0.0001
2	0.0422	0.0026	0.0005
3	0.1360	0.0054	0.0013
4	0.2712	0.0108	0.0023
5	0.4743	0.3378	0.0032
6	0.6921	0.5763	0.0039
7	0.9440	0.8428	0.0054
8	1.2105	1.1343	0.9160
9	1.5049	1.4444	1.3768
10	1.8318	1.7660	1.7847
11	2.1456	2.0913	2.2385
12	2.4621	2.4124	2.6623
13	2.7540	2.7216	3.1348
14	3.0393	3.0112	3.9938
15	3.2997	3.2741	3.9961
16	3.5088	3.5038	3.9968
17	3.7091	3.6948	3.9977
18	3.8402	3.8423	3.9987
19	3.9510	3.9427	3.9995
20	3.9945	3.9935	3.9999

Figure 3.1: The defect norm $|\delta_{p,m}(t)|$ ($p = 0$, $m = 20$) for the free Schrödinger example with different choices of starting vector case (a) ($\times$), case (b) ($\circ$) and case (c) ($\square$). The table on the right-hand side shows eigenvalues $\lambda_1^{(*)}, \dots, \lambda_m^{(*)}$ of H_m for the different starting vectors, case (a)–(c). For the case (b) the divided differences over the clustered eigenvalues $\gamma_m \big(\prod_{j=5}^{20} \lambda_j^{(b)} \big)^{-1} | \exp_t[\mathrm{i}\,\lambda_1^{(b)}, \dots, \mathrm{i}\,\lambda_4^{(b)}]|$ is illustrated by ($+$). The asymptotic expansion of the divided differences for $t \to 0$ given in (3.4.9) is illustrated using dashed lines. The dash-dotted line is $\mathcal{O}(t^6)$.

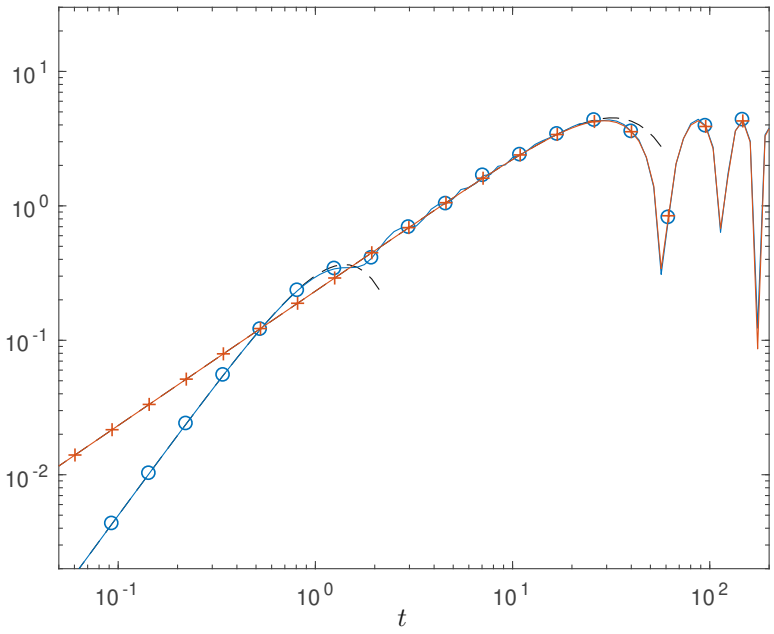


Figure 3.2: The divided differences $|\exp_t[\mathrm{i}a_1, \mathrm{i}a_2, \mathrm{i}a_3]|$ ($\circ$) and $|\exp_t[\mathrm{i}a_1, \mathrm{i}a_2]|/|a_3 - a_1|$ ($+$) for the choice of a_1, a_2, a_3 given in the text. The asymptotic expansion of the divided differences for $t \to 0$ given in (3.4.9) is illustrated using dashed lines.

the recursive definition of the divided differences (see [Hig08, eq. (B.24)] or others) we obtain

$$|\exp_t[\mathrm{i}a_1, \mathrm{i}a_2, \mathrm{i}a_3]| = \left|\frac{\exp_t[\mathrm{i}a_2, \mathrm{i}a_3] - \exp_t[\mathrm{i}a_1, \mathrm{i}a_2]}{a_3 - a_1}\right| \approx \left|\frac{\exp_t[\mathrm{i}a_1, \mathrm{i}a_2]}{a_3 - a_1}\right|, \quad \text{for larger } t.$$

This example is illustrated in Figure 3.2. This behavior can be generalized for a larger number of nodes and is also observed in Figure 3.1.

Quadrature estimates for the defect integral. With the previous observations on the defect we now discuss different quadrature-based estimates.

The generalized residual estimate, which was introduced in [HLS98] and appeared in a similar manner in [DGK98, Saa92, Lub08, BGH13], conforms to a quadrature on the defect norm integral which is related to the error norm via Corollary 3.3.3.

Remark 3.4.11 (Generalized residual estimate, see also [HLS98])**.** *Applying the right-endpoint rectangle rule we have*

$$\int_0^t |\delta_{p,m}(s)|\,\mathrm{d}s \approx t|\delta_{p,m}(t)|,$$

and with Corollary 3.3.3 (and $\delta_{p,m}(t)$ given in (3.3.1a)*) we obtain the error estimate*

$$\|l_{p,m}(t)\|_2 \approx h_{m+1,m}t^{1-p}|\delta_{p,m}(t)| = \beta h_{m+1,m}t|e_m^*\varphi_p(tH_m)e_1|.$$

Assume that $\max_{s\in[0,t]} |\delta_{p,m}(s)| = |\delta_{p,m}(t)|$, e.g., $|\delta_{p,m}(t)|$ is monotonically increasing in t. Then,

$$\int_0^t |\delta_{p,m}(s)|\,\mathrm{d}s \leq t \max_{s\in[0,t]} |\delta_{p,m}(s)| = t|\delta_{p,m}(t)|.$$

In this case the generalized residual estimate from Remark 3.4.11 results in an upper bound on the error norm.

In the most general case the defect is of high order for $t \to 0$ and in a relevant time regime, see also Figure 3.1 case (a) and previous remarks. Then the defect is a higher order function and the right-endpoint quadrature does result in an upper bound but is not tight. In this case we can improve the estimate by a prefactor depending on the *effective order* defined in Appendix 3.C, cf. (2.6.7) in Chapter 2. If the defect is sufficiently smooth in a relevant time regime this results in a tight upper bound on the error norm.

Remark 3.4.12 (Effective order estimate, see also Section 2.6 in Chapter 2)**.** *Denote $f(t) = |\exp_t[\lambda_1,\dots,\lambda_m,0_p]|$ for the time-dependent part of the defect with eigenvalues $\lambda_1,\dots,\lambda_m$ of H_m. Assume $f(t) > 0$ for a sufficiently small time regime $t > 0$. We consider the effective order $\rho(t)$ defined in* (3.C.4a)*. With the following estimate for the integral of the defect,*

$$\int_0^t |\delta_{p,m}(s)|\,\mathrm{d}s \approx \frac{t}{\rho(t)+1}|\delta_{p,m}(t)|,$$

and from Corollary 3.3.3 (with $\delta_{p,m}(t)$ given in (3.3.1a)*) we obtain*

$$\|l_{p,m}(t)\|_2 \approx h_{m+1,m}\frac{t^{1-p}}{\rho(t)+1}|\delta_{p,m}(t)| = \beta h_{m+1,m}\frac{t}{\rho(t)+1}|e_m^*\varphi_p(tH_m)e_1|.$$

In Section 2.6 (Chapter 2) the effective order is defined for $|e_m^*\mathrm{e}^{tH_m}e_1|$ ($p=0$) which is equivalent to the definition via the divided differences of $f(t)$. (This follows from Corollary 3.3.6 and the definition of the effective order which is independent of a constant prefactor.)

Some of the following observations already appeared in Section 2.6 (Chapter 2). The quadrature scheme in Remark 3.4.12 is motivated by the following relation of the effective order and the integral of the divided differences $f(t)$. From eq. (3.C.4a),

$$f(t) = \frac{f'(t)\,t}{\rho(t)}.$$

Integration and application of the mean value theorem shows the existence of $t^* \in [0,t]$ with

$$\int_0^t f(s)\,\mathrm{d}s = \frac{1}{\rho(t^*)}\int_0^t f'(s)\,s\,\mathrm{d}s,$$

and integration by parts gives

$$\int_0^t f(s)\,\mathrm{d}s = \frac{tf(t)}{1+\rho(t^*)}. \tag{3.4.16}$$

This result can passed over to the integral of the defect.

Assume the effective order is monotonically decreasing for t small enough, with $\min_{s\in(0,t]} \rho(s) = \rho(t) \geq 0$. This holds in an asymptotic regime for the dissipative case up to round-off, see also Theorem 3.C.2 with the real parts $\xi_1, \ldots, \xi_m$ of the eigenvalues of H_m being non-positive. With (3.4.16) and the assumption $0 \leq \rho(t) \leq \rho(s) \leq m+p-1 = \rho(0+)$ for $s \in [0,t]$, we inclose the integral of the defect by

$$\tfrac{t}{m}\,|\delta_{p,m}(t)| \leq \int_0^t |\delta_{p,m}(s)|\,\mathrm{d}s \leq \tfrac{t}{\rho(t)+1}\,|\delta_{p,m}(t)| \leq t\,|\delta_{p,m}(t)|. \tag{3.4.17}$$

Combining (3.4.17) and Corollary 3.3.3 we obtain the upper bound

$$\|l_{p,m}(t)\|_2 \leq \tfrac{h_{m+1,m}t^{1-p}}{\rho(t)+1} \cdot |\delta_{p,m}(t)| \leq h_{m+1,m}t^{1-p} \cdot |\delta_{p,m}(t)|.$$

A computable expression for the effective order was given in (2.6.10) (see Section 2.6 in Chapter 2). This result can be generalized to the case $p \in \mathbb{N}_0$,

$$\rho(t) = \begin{cases} t\,\mathrm{Re}\left((H_m)_{m,m} + (H_m)_{m,m-1}(y_{p,m}(t))_{m-1}/(y_{p,m}(t))_m\right) & \text{for } p = 0, \quad \text{and} \\ \mathrm{Re}((y_{p-1,m}(t))_m/(y_{p,m}(t))_m) & \text{for } p \in \mathbb{N}, \end{cases}$$

with $y_{p,m}(t) \in \mathbb{C}^m$ from (3.2.9). The expression for the case $p \in \mathbb{N}$ can be obtained by (2.6.10) (see Section 2.6 in Chapter 2) applied on the representation $|e_{m+p}^* \mathrm{e}^{t\widetilde{H}_m} e_1|$ for the defect ((iii) in Corollary 3.3.6) and making use of the special structure of $\widetilde{H}_m$, $\beta e_{m+p}^* \mathrm{e}^{t\widetilde{H}_m} e_1 = t^p (y_{p,m}(t))_m$ (see Corollary 3.3.6) and $\beta e_{m+p-1}^* \mathrm{e}^{t\widetilde{H}_m} e_1 = t^{p-1} (y_{p-1,m}(t))_m$ (see [Sid98, Corollary 1]).

As illustrated in Figure 3.1 the defect can be highly oscillatory in a relevant time regime, especially for specific starting vectors, and in this case the quadrature estimates should be handled with care.

3.4.2 A stopping criterion for the lucky breakdown

The special case $h_{k+1,k} = 0$ during the construction of the Krylov subspace is considered to be a *lucky breakdown,* a breakdown of the Arnoldi or Lanczos iteration with the benefit of an exact approximation of $\varphi_p(tA)v$ for any $t > 0$ via the Krylov subspace $\mathcal{K}_k(A, v)$. In floating point arithmetic the lucky breakdown results in $h_{k+1,k} \approx 0$ and can lead to stability issues if the Arnoldi or Lanczos method is not stopped properly. The condition that the Krylov propagator is exact is not exactly determinable in floating point arithmetic but can be weakened to the error condition in (3.2.13) for a given tolerance tol per unit step. With this approach we introduce a stopping criterion which can be applied on the fly to detect a lucky breakdown and satisfies an error bound. This does not depend on any a priori information as long the tolerance tol is chosen properly so that round-off errors can be neglected, see remarks before Corollary 3.3.3.

Proposition 3.4.13. *Let $\mu_2(A) \leq 0$ and assume that round-off errors are sufficiently small, see Corollary 3.3.3. Let tol be a given tolerance and*

$$\frac{\beta h_{k+1,k}}{(p+1)!} \leq tol \tag{3.4.18}$$

be satisfied at the k-th step of the Arnoldi or Lanczos iteration. Then the iteration can be stopped and the Krylov subspace $\mathcal{K}_k(A, v)$ can be used to approximate the vector $\varphi_p(tA)v$ with a respective error per unit step $\|l_{p,k}(t)\|_2 \leq t \cdot tol$.

Proof. We use the upper bound on the error norm from Corollary 3.3.3,

$$\|l_{p,k}(t)\|_2 \leq \frac{h_{k+1,k}}{t^p} \int_0^t |\delta_{p,k}(s)| \, \mathrm{d}s. \tag{3.4.19}$$

To obtain a uniform bound on the defect integral we use

$$|\delta_{p,k}(t)| \leq \beta t^p \|e_k\|_2 \|\varphi_p(tH_k)e_1\|_2 = \beta t^p \|\varphi_p(tH_k)e_1\|_2. \tag{3.4.20}$$

- For $p > 0$ we apply the integral representation (3.2.3) on $\varphi_p(tH_m)e_1$ to obtain the upper bound

$$\|\varphi_p(tH_m)e_1\|_2 \leq \frac{\max_{s\in[0,t]} \|\mathrm{e}^{sH_m}\|_2}{(p-1)!} \int_0^1 \theta^{p-1} \, \mathrm{d}\theta = \frac{\max_{s\in[0,t]} \|\mathrm{e}^{sH_m}\|_2}{p!}. \tag{3.4.21}$$

- For $p = 0$ the analogous result is directly obtained: Combine (3.4.20) and (3.4.21) with $\|\mathrm{e}^{sH_k}\|_2 \leq \mathrm{e}^{t\mu_2(H_k)} \leq \mathrm{e}^{t\mu_2(A)}$ up to round-off and $\mu_2(A) \leq 0$, giving

$$|\delta_{p,k}(t)| \leq \beta \frac{t^p}{p!}, \quad \text{and} \quad \int_0^t |\delta_{p,k}(s)| \, \mathrm{d}s \leq \beta \frac{t^{p+1}}{(p+1)!}.$$

Together with (3.4.19) and (3.4.18) we conclude $\|l_{p,k}(t)\|_2 \leq t \cdot \mathrm{tol}$. □

3.5 Numerical experiments

The notation for the error $l_{p,m}(t)$, the estimate of the error norm $\zeta_{p,m}(t)$ and the tolerance tol have been introduced in (3.2.12) and (3.2.13). The notation $\zeta_{p,m}$ will be used for different choices of error estimates discussed in the previous section. Theorem 3.4.3 and Corollary 3.4.5 result in upper bounds on the error norm, $\|l_{p,m}(t)\|_2 \leq \zeta_{p,m}(t)$. The quadrature-based error estimates given in Remark 3.4.11 and 3.4.12 result in estimates for the error norm, $\|l_{p,m}(t)\|_2 \approx \zeta_{p,m}(t)$, and with additional conditions also give upper bounds.

For a fixed tolerance tol we use the notation $t(m)$ for the smallest time t with $\zeta_{p,m}(t) = t \cdot \mathrm{tol}$, see (3.2.13). This choice of $t(m)$ helps us to verify the tested error estimates for a time t which is of the most practical interest. With the help of a reference solution the true error norm per unit step can be tested by $\|l_{p,m}(t(m))\|_2/t(m)$.

We also consider the following previously known error estimates in our numerical experiments. The generalized residual estimate [HLS98] was recapitulated in Remark 3.4.11 and will be discussed in the numerical experiments. Furthermore, we test the performance of the error bound given in [DMR09, Proposition 6]. This upper bound on the error norm applies to the Krylov approximation of $\varphi_p(-t\,A)v$ for $p \in \mathbb{N}_0$, a matrix $A \in \mathbb{R}^{n\times n}$ with a numerical range in the right complex half-plane (up to a potential shift), and $v \in \mathbb{R}^n$. In this case the matrix A can have real and complex eigenvalues, where the latter come in

complex conjugate pairs. Concerning the skew-Hermitian case, a similar error bound for the Krylov approximation to $\varphi_p(-\mathrm{i}\,t\,B)v$ for a Hermitian matrix $B \in \mathbb{R}^{n\times n}$ and $p \in \mathbb{N}_0$ is given separately in [DMR09, Proposition 8]. To evaluate these error bounds the eigenvalues of H_m and the terms $h_{m+1,m}$ and γ_m are used.

A series expansion for the error concerning φ-functions is given in [Sid98, Theorem 2] and the leading terms of this expansion can be used for error estimation, cf. [Sid98, NW12]. In general [Sid98] suggests to evaluate more than one term of this series to ensure reliability of the obtained error estimate, which requires further matrix-vector multiplications in the given large dimensional space. This can often be inefficient in terms of computational cost, cf. Remark 2.8.2 in Chapter 2, and we avoid this series expansion in the general case. However, when the Ritz values are real-valued the error bound in Corollary 3.4.4 (corresponding to the bound in Theorem 3.4.3) coincides with the leading term of the error series in [Sid98, Theorem 2]. Thus, the first term of the error series in [Sid98, Theorem 2] yields a reliable error bound in this case. For the convection-diffusion equation with parameter $\nu = 100$ in Subsection 3.5.1 below (the Ritz values have negligible imaginary parts in this case) the error bound of Theorem 3.4.3 performs well (comparable to the effective order estimate and better than the other error estimates considered, e.g., the generalized residual estimate), and this potentially carries over to the error estimates in [Sid98, NW12].

3.5.1 Convection-diffusion equation

Consider the following two-dimensional convection-diffusion equation with $t \geq 0$ and $x \in [0,1]^2$,

$$\partial_t u = Lu, \quad \text{with} \quad L = \Delta + \nu(\partial_{x_1} + \partial_{x_2}), \quad u = u(t,x),\ \nu \in \mathbb{R}. \tag{3.5.1}$$

Let $A \in \mathbb{R}^{n\times n}$ be obtained by the two-dimensional finite difference discretization of the operator L in (3.5.1) with zero Dirichlet boundary conditions and $N = 500$ inner mesh points in each spatial direction, hence, $n = N^2$. This test problem is similar to other convection-diffusion equations appearing in the study of Krylov subspace methods, see also [JAK20, EE06, FGS14, BK19] and others.

For the convection parameter we choose $\nu = 100, 500$ which results in a non-normal matrix A. Considering the spectrum of A the case $\nu = 100$ is closer to the Hermitian case and $\nu = 500$ is closer to the skew-Hermitian case. In both cases the numerical range of A is contained in the left complex plane, $\mu_2(A) \leq 0$.

We discuss error estimates for the Krylov approximation of the matrix exponential ($p = 0$) and a φ-function (for which we choose $p = 2$). For the case $p = 0$ the action of the matrix exponential $\mathrm{e}^{tA}v$ is approximated in the Krylov subspace $\mathcal{K}_m(A, v)$, see (3.2.8b). Analogously, for the case $p = 2$ we approximate $\varphi_p(tA)v$ as given in (3.2.8a). As a starting vector we choose the normalized vector $v = (1/N, \ldots, 1/N)^* \in \mathbb{R}^n$. In Figure 3.3 and 3.4 we compare the error bounds given in Theorem 3.4.3, Corollary 3.4.5 and [DMR09, Proposition 6], and the generalized residual estimate (Remark 3.4.11) and the effective order estimate (Remark 3.4.12), for the convection-diffusion equation. The error bound of Corollary 3.4.5 is applied with $\xi_{\max} = 0$ (the effect of $\xi_{\max}$ is negligible for the current examples). Concerning the error bound given in [DMR09, Proposition 8], we choose the parameter ε by minimizing [DMR09, eq. (39)], and $a = 0$.

For the case $\nu = 100$ the eigenvalues of H_m have a negligible imaginary part and the upper bound given in Theorem 3.4.3 constitutes a tight upper bound on the exact evaluation of the scaled defect integral, which yields a tight error bound. This error bound and the effective order estimate (Remark 3.4.12), which is based on a quadrature estimate on the defect integral, yield approximately the same results for the case $\nu = 100$. The performance of the generalized residual estimate (Remark 3.4.11) is similar to the performance of the error bound in [DMR09, Proposition 6], especially for larger choices of m. The error bound in Corollary 3.4.5 is only accurate for small m in the current example. The high accuracy of the error bound in Theorem 3.4.3 and the effective order estimates results in time steps $t(m)$ which are larger than the time steps suggested by generalized residual estimate and the error bound in [DMR09, Proposition 6], and significantly larger compared to the time steps given by Corollary 3.4.5. Comparing the cases $p = 0$ and $p = 2$, the time steps suggested by the error bounds of Corollary 3.4.5 and [DMR09, Proposition 6] are slightly smaller in relation to the time step prescribed by the effective order estimate for $p = 2$. Considering the true error for the time steps computed by the error bound in Theorem 3.4.3, the effective order estimate and the generalized residual estimate, the performance of these estimates only differs slightly between the cases $p = 0$ and $p = 2$.

For the case $\nu = 500$ the matrix H_m has eigenvalues with larger imaginary parts (especially for larger m). In this case the error bound in Theorem 3.4.3, is less tight, and the effective order estimate (Remark 3.4.12) performs best comparing to the other error estimates. Comparing the cases $p = 0$ and $p = 2$, we observe that the time steps suggested by the error bounds of Theorem 3.4.3, Corollary 3.4.5 and [DMR09, Proposition 6] are slightly smaller in relation to the time step of the effective order estimate for $p = 2$.

The criterion ac.est.1(t) given in Remark 3.4.8 is evaluated for $\nu = 100, 500$ and $p = 0, 2$ with $t(m)$ corresponding to Theorem 3.4.3 (see caption of Figure 3.3 and 3.4). For $\nu = 100$ we obtain ac.est.1($t(m)$) < 0.1 for any m tested and $p = 0, 2$. For $\nu = 500$ the smallest m with ac.est.1($t(m)$) > 0.1 is $m = 40$ and $m = 36$ for $p = 0$ and $p = 2$, respectively. The error bound in Theorem 3.4.3 conforms to an upper bound of the scaled defect integral, and in the case of ac.est.1($t(m)$) > 0.1 a more accurate estimate on the defect integral is likely to perform better. For $\nu = 500$ and $m = 40$ ($p = 0$) and $m = 36$ ($p = 2$) we observe that this is the case for the effective order estimate. Similar to the criterion ac.est.1(t), we test ac.est.2(t) given in Remark 3.4.10 for $t(m)$ according to Corollary 3.4.5. For $\nu = 100$ the smallest m with ac.est.2($t(m)$) > 0.1 is $m = 7$ for $p = 0, 2$ individually. Otherwise, for $\nu = 500$ the smallest m with ac.est.2($t(m)$) > 0.1 is $m = 8$ and $m = 7$ for $p = 0$ and $p = 2$, respectively.

3.5.2 Free Schrödinger equation, a skew-Hermitian problem

For the free Schrödinger equation we let A be a finite difference discretization of the Laplace operator, precisely, we choose A corresponding to L in (3.5.1) with $\nu = 0$ and $N = 500$. With A corresponding to a discretized Laplace operator, the vector $\mathrm{e}^{\mathrm{i}tA}v$ yields a solution to a discretized free Schrödinger equation with starting vector v. The free Schrödinger equation represents a skew-Hermitian problem, and following Remark 3.2.4 we approximate $\mathrm{e}^{\mathrm{i}tA}v$ in the Krylov subspace $\mathcal{K}_m(A, v)$ by $\beta V_m \mathrm{e}^{\mathrm{i}tH_m} e_1$. Analogously to the previous subsection, we choose the normalized starting vector $v = (1/N, \ldots, 1/N)^* \in \mathbb{R}^n$, and we also consider

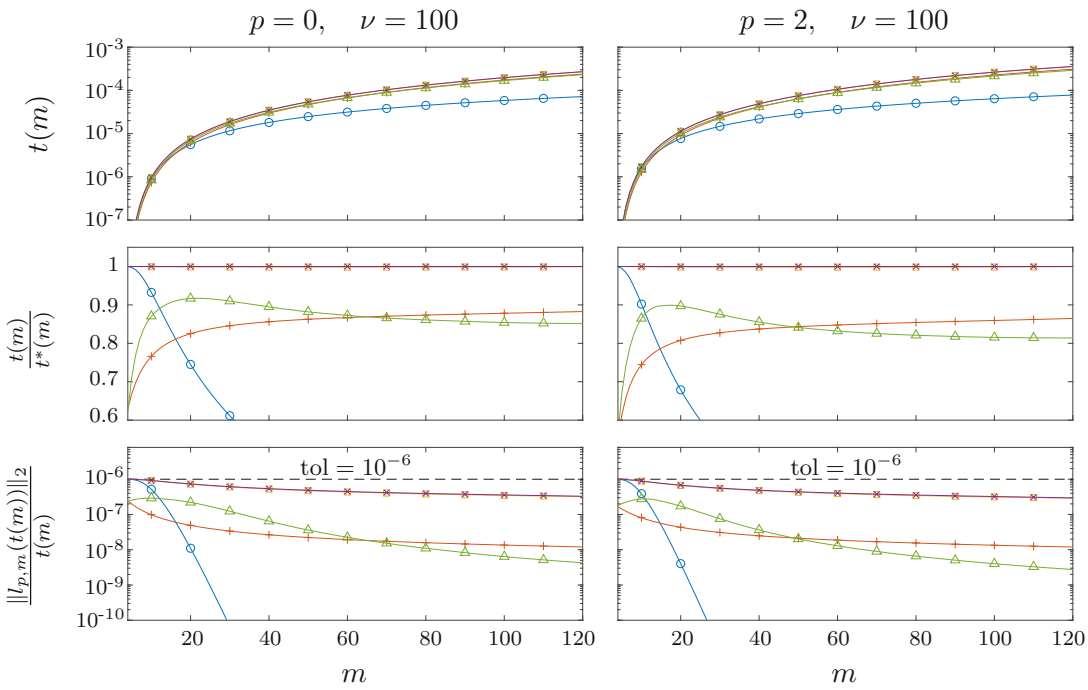


Figure 3.3: Convection-diffusion problem (3.5.1) for the parameter $\nu = 100$. We consider $p = 0$ (left) and $p = 2$ (right). Three rows of figures are presented here: The first row shows the time $t(m)$ which is the smallest t such that $\zeta_{p,m}(t) = t \cdot \mathrm{tol}$ for $\mathrm{tol} = 10^{-6}$ and $\zeta_{p,m}$ corresponding to the error bound given in Theorem 3.4.3 ($\times$), Corollary 3.4.5 ($\circ$), the generalized residual estimate given in Remark 3.4.11 ($+$), the effective order estimate given in Remark 3.4.12 ($\square$), and the error bound given in [DMR09, Proposition 6] ($\triangle$). For the second row we choose $t^*(m)$ as the largest time step $t(m)$ given by the currently discussed error estimate, and we show $t(m)/t^*(m)$ for $t(m)$ as chosen above. The third row shows the true error per unit step, $\|l_{p,m}(t(m))\|_2/t(m)$, for the time $t(m)$ as chosen above.

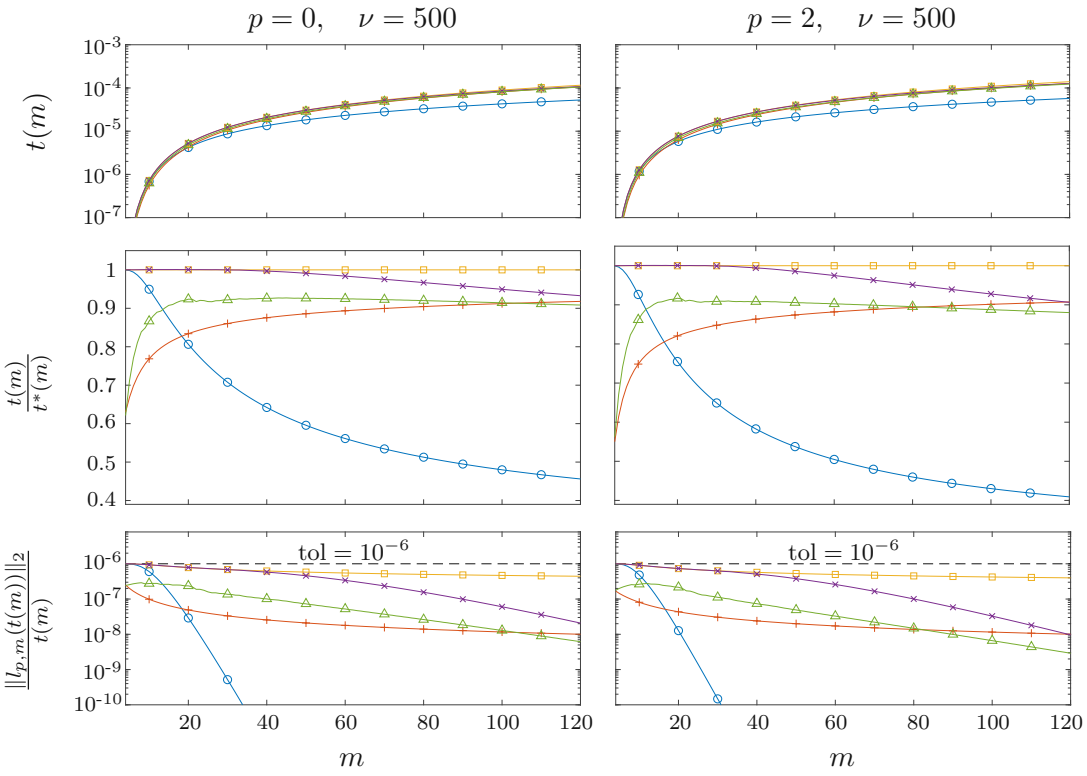


Figure 3.4: Convection-diffusion problem (3.5.1) for the parameter $\nu = 500$, similar to Figure 3.3. We consider $p = 0$ (left) and $p = 2$ (right). The first row of figures shows the time $t(m)$ which is the smallest t such that $\zeta_{p,m}(t) = t \cdot \text{tol}$ for $\text{tol} = 10^{-6}$ and $\zeta_{p,m}$ corresponding to the error bound given in Theorem 3.4.3 ($\times$), Corollary 3.4.5 ($\circ$), the generalized residual estimate given in Remark 3.4.11 ($+$), the effective order estimate given in Remark 3.4.12 ($\square$), and the error bound given in [DMR09, Proposition 6] ($\triangle$). For the second row we choose $t^*(m)$ as the largest time step $t(m)$ given by the currently discussed error estimate, and we show $t(m)/t^*(m)$ for $t(m)$ as chosen above. The third row shows the true error per unit step, $\|l_{p,m}(t(m))\|_2/t(m)$, for the time $t(m)$ as chosen above.

the Krylov approximation to $\varphi_p(itA)v$ for $p = 2$, i.e., $\beta V_m \varphi_p(\mathrm{i}\,t\,H_m)e_1$.

In Figure 3.5 the error bounds given in Corollary 3.4.5 (which coincides with the error bound given in Theorem 3.4.3 in the skew-Hermitian case) and [DMR09, Proposition 8] (the counterpart to [DMR09, Proposition 6] for the skew-Hermitian case), the effective order estimate (Remark 3.4.12), and the generalized residual estimate (Remark 3.4.11) are evaluated for the current example. For the parameter ε in [DMR09, Proposition 8] we choose $\varepsilon = m/t$ as suggested in the numerical experiments therein.

For the skew-Hermitian case, the effective order estimate (Remark 3.4.12) yields the largest time steps compared to the other error estimates. The error bound of Corollary 3.4.5 performs well for moderate m and better than the error bound in [DMR09, Proposition 8] for any of the tested m here. For larger m the generalized residual estimate performs better than the error bound of Corollary 3.4.5. Similar to examples of the previous subsection, the error bound of Corollary 3.4.5 performs better for the case $p = 0$ compared to $p = 2$. Similar results can be observed for the error bound of [DMR09, Proposition 8]. The performance of the effective order estimate and the generalized residual estimate only differs slightly between the cases $p = 0$ and $p = 2$.

We test ac.est.2(t) given in Remark 3.4.10 for $t(m)$ according to Corollary 3.4.5. The smallest m with ac.est.2$(t(m)) > 0.1$ is $m = 15$ and $m = 13$ for $p = 0$ and $p = 2$, respectively. Following Remark 3.4.10, the error bound given in Corollary 3.4.5 overestimates the error by a factor 1.1 (in an asymptotic sense) for these values of m, which fits to the results shown in Figure 3.5.

3.5.3 Free Schrödinger equation with a double well potential and a Gaussian wave packet as an initial state

In the following numerical experiment we choose a special starting vector which results in the matrix H_m having clustered eigenvalues, and we observe effects which were previously discussed in Subsection 3.4.1. Typically, this is related to regularity properties of the underlying initial state.

We consider the one-dimensional free Schrödinger equation with a double well potential,

$$\partial_t \psi = -\mathrm{i}H\psi, \quad \text{with} \quad H = -\Delta + V, \quad \psi = \psi(t,x) \in \mathbb{C},\ V = V(x) \in \mathbb{R}, \tag{3.5.2}$$

for $t \geq 0$, $x \in [-10, 10]$ and $V(x) = x^4 - 15x^2$. Let $B \in \mathbb{C}^{n\times n}$ be the discretized version of the Hamiltonian operator H in (3.5.2) with periodic boundary conditions using a finite difference scheme with a mesh of size $n = 10000$. With B Hermitian, the full problem $A = -\mathrm{i}B$ is skew-Hermitian (see Remark 3.2.4) with $\mu_2(A) = 0$. For the initial state of (3.5.2) we choose a Gaussian wave packet,

$$\psi(t = 0, x) = (0.2\pi)^{-1/4} \exp(-(x + 2.5)^2/(0.4)), \tag{3.5.3}$$

which is evaluated on the mesh and normalized to obtain a discrete starting vector $v \in \mathbb{R}^n$. This problem also appears in [IKS19, Sin19].

We discuss error estimates for the case $p = 0$ (Krylov approximation of $\mathrm{e}^{-\mathrm{i}tB}v$). The implementation of the skew-Hermitian problem is described in Remark 3.2.4. In Figure 3.6 the upper bound given in Corollary 3.4.5 (which coincides with the error bound given in

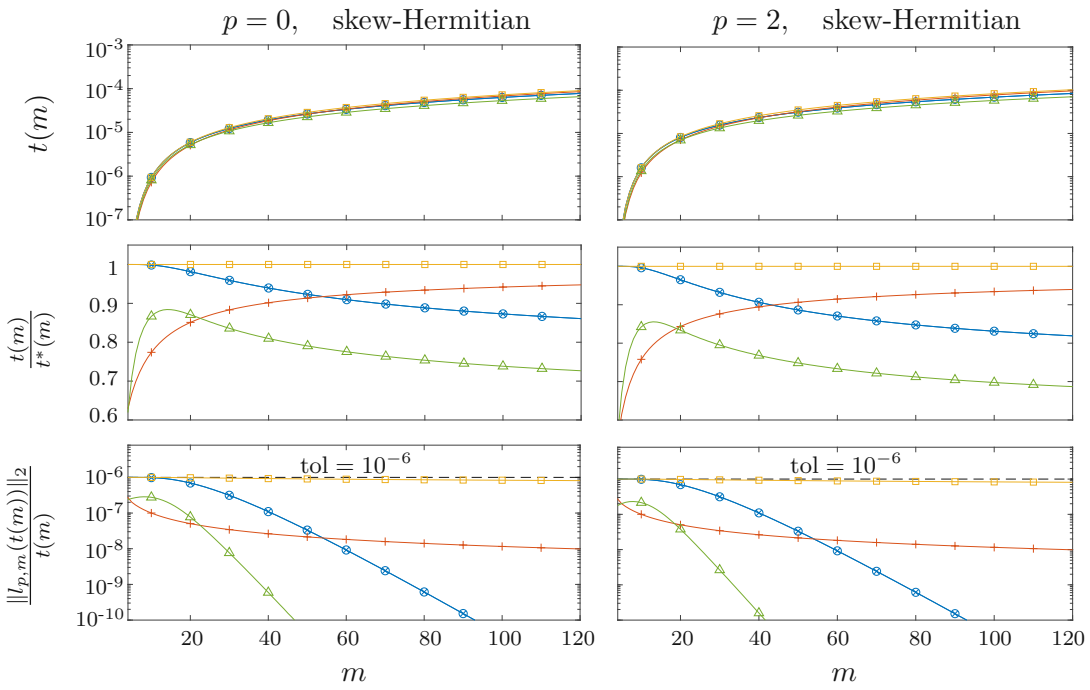


Figure 3.5: The skew-Hermitian problem $\varphi_p(\mathrm{i}\,A)v$ where A corresponds to the Laplace operator ((3.5.1) with $\nu = 0$) and $v = (1/N, \ldots, 1/N)^*$. Results are shown for $p = 0$ (left) and $p = 2$ (right). For $p = 0$ this problem is related to the free Schrödinger equation. The top row shows the time $t(m)$ which is the smallest t such that $\zeta_{p,m}(t) = t \cdot \mathrm{tol}$ for $\mathrm{tol} = 10^{-6}$ and $\zeta_{p,m}$ corresponding to the error bound given in Theorem 3.4.3 ($\times$), Corollary 3.4.5 ($\circ$), the generalized residual estimate given in Remark 3.4.11 ($+$), the effective order estimate given in Remark 3.4.12 ($\square$), and the error bound given in [DMR09, Proposition 8] ($\triangle$). The error bounds in Theorem 3.4.3 ($\times$) and Corollary 3.4.5 ($\circ$) coincide in the skew-Hermitian case. For the middle row we choose $t^*(m)$ as the largest time step $t(m)$ given by the currently discussed error estimate, and we show $t(m)/t^*(m)$ for $t(m)$ as chosen above. The bottom row shows the true error per unit step, $\|l_{p,m}(t(m))\|_2/t(m)$, for the time $t(m)$ as chosen above.

Theorem 3.4.3 for the skew-Hermitian case) and the error estimates given in Remark 3.4.11 and 3.4.12 are compared. Additionally, we consider the error bound given in [DMR09, Proposition 8] with the parameter choice $\varepsilon = m/t$.

The error bounds given in Corollary 3.4.5 and [DMR09, Proposition 8] are reliable but not tight for the current example. Thus, the time steps $t(m)$ which are suggested by these error bounds are significantly smaller than the time steps suggested by the quadrature-based error estimates (Remark 3.4.11 and 3.4.12), and comparing with the numerical experiments of the previous subsection, this seems to be highly affected by the choice of the starting vector. For the error bound in Corollary 3.4.5 this can be explained by the loss of order of the defect. However, the error bound in Corollary 3.4.5 shows a better performance compared to the error bound in [DMR09, Proposition 8].

In terms of accuracy the effective order estimate (Remark 3.4.12) performs significantly better compared to the error bounds in Corollary 3.4.5 and [DMR09, Proposition 8], and better compared to the generalized residual estimate (Remark 3.4.11). In terms of reliability we have argued that the effective order estimate and the generalized residual estimate constitute upper bounds on the error norm when the defect norm behaves sufficiently smooth. The defect norm $|\delta_{m,0}(t)|$, which is presented in the lower right corner of Figure 3.6, does have an oscillatory behavior in a specific time regime which can be related to the starting vector, cf. Subsection 3.4.1. For the time steps which are relevant for the current example, this does not critically affect the quadrature estimates on the defect integral related to Remark 3.4.11 and 3.4.12. Under certain conditions, e.g., a different choice for the tolerance tol, this oscillatory behavior of the defect can lead to failure of the error estimates given in Remark 3.4.11 and 3.4.12. However, the quadrature of the defect integral can be further improved in such cases to ensure a reliable error estimate.

3.6 Conclusions and outlook

In this work various a posteriori bounds and estimates on the error norm, which have their origin in an integral representation of the error using the defect (residual), are studied. We have characterized the accuracy of these error bounds by the positioning of Ritz values (i.e., eigenvalues of H_m) on the complex plane. The case of real Ritz values is the most favorable one to obtain a tight error bound via an integral on the defect norm (Corollary 3.4.4). A new error bound (Theorem 3.4.3) has shown to be tight if Ritz values are close to the real axis and in this case favorably compares with existing error bounds. We further recapitulate an existing error bound (Corollary 3.4.5) which remains relevant, especially for the case of Ritz values with a significant imaginary part. In addition for the error bound in Theorem 3.4.3 and Corollary 3.4.5, we have provided a criterion to quantify the achieved accuracy on the fly. For an illustration of the claims concerning the new error bound we primary refer to the numerical example given in Subsection 3.5.1. The quadrature-based error estimates in Subsection 3.4.1 (e.g., the generalized residual estimate) do not yield proven upper bounds on the error norm and we addressed special cases (e.g., the numerical example in Subsection 3.5.3) for which the reliability of these estimates can be problematic. These cases are also analyzed in terms of Ritz values in Subsection 3.4.1 and this relation can be of further interest for a numerical implementation. Nevertheless, in most cases the

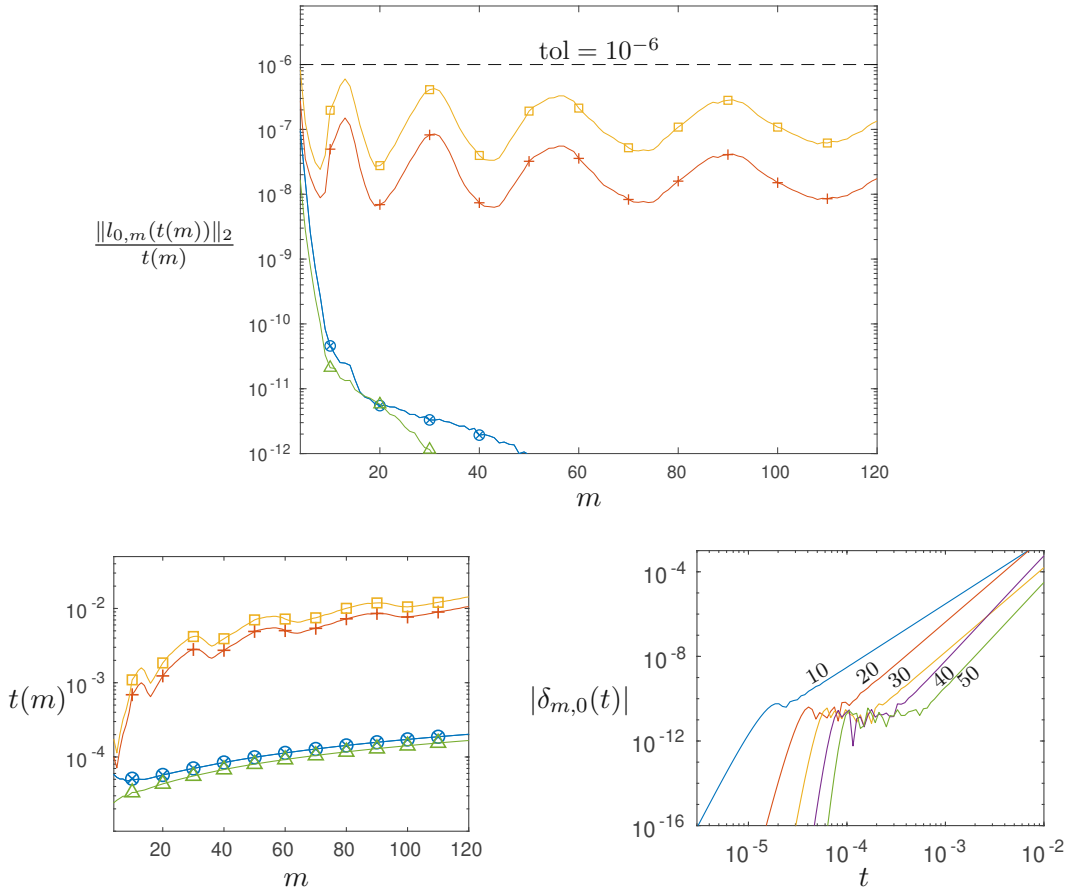


Figure 3.6: Results for the free Schrödinger problem with a double well potential and the starting vector given by (3.5.3). This figure shows the time $t(m)$ (bottom left), which is the smallest t so that $\zeta_{0,m}(t) = t \cdot \text{tol}$ for $\text{tol} = 10^{-6}$, the true error per unit step $\|l_{0,m}(t(m))\|_2/t(m)$ (top) and the defect norm $|\delta_{m,0}(t)|$ (bottom right) for $m \in \{10, 20, 30, 40, 50\}$. The results for $t(m)$ and $\|l_{0,m}(t(m))\|_2/t(m)$ are given for $\zeta_{0,m}$ being the upper norm bound given in Theorem 3.4.3 ($\times$), Corollary 3.4.5 ($\circ$), the generalized residual estimate given in Remark 3.4.11 ($+$), the effective order estimate given in Remark 3.4.12 ($\square$) and the error bound given in [DMR09, Proposition 8] ($\triangle$). The results for Theorem 3.4.3 ($\times$) and Corollary 3.4.5 ($\circ$) coincidence in the skew-Hermitian case.

quadrature-based estimates remain valid, whereat the effective order quadrature stands out in terms of performance.

We also remark that the theory provided in our work gives the possibility to adapt the choice of the error estimate on the fly to obtain an estimate which is as reliable, accurate and economic as possible. This is the topic of further work.

Appendix

3.A Properties of the Krylov subspace in exact and floating point arithmetic

Let $H_m = V_m^* A V_m$ and $V_m^* V_m = I_{m\times m}$ in exact arithmetic. For $z \in \mathrm{W}(H_m)$ (numerical range of H_m) there exists $x \in \mathbb{C}^m$ with

$$z = \frac{x^* H_m x}{x^* x} = \frac{x^* V_m^* A V_m x}{x^* V_m^* V_m x} = \frac{y^* A y}{y^* y}, \quad \text{for } y = V_m x, \tag{3.A.1}$$

whence $\mathrm{W}(H_m) \subseteq \mathrm{W}(A)$.

Similar results hold in floating point arithmetic with relative machine precision ε and certain additional assumptions. Assume there exists an orthonormal basis $\widehat{V}_m \in \mathbb{C}^{n\times m}$ and a perturbation $\widetilde{U}_m \in \mathbb{C}^{m\times m}$, which is sufficiently small in norm (i.e., there exists a moderate constant C_3 with $\|\widetilde{U}_m\|_2 \leq C_3\varepsilon$), with

$$H_m = \widehat{V}_m^* A \widehat{V}_m + \widetilde{U}_m. \tag{3.A.2}$$

With assumption (3.A.2) and basic properties of the numerical range we obtain

$$\mathrm{W}(H_m) \subseteq \mathrm{W}(\widehat{V}_m^* A \widehat{V}_m) + \mathrm{W}(\widetilde{U}_m). \tag{3.A.3}$$

Similar to (3.A.1) we obtain

$$\mathrm{W}(\widehat{V}_m^* A \widehat{V}_m) \subseteq \mathrm{W}(A). \tag{3.A.4}$$

Then we combine (3.A.3) and (3.A.4) and make use of $\|\widetilde{U}_m\|_2 \leq C_3\varepsilon$ to obtain

$$\mathrm{W}(H_m) \subseteq U_{C_3\varepsilon}(\mathrm{W}(A)),$$

with $U_{C_3\varepsilon}(\mathrm{W}(A))$ being the neighborhood of $\mathrm{W}(A)$ with a distance $C_3\varepsilon$.

In [Sim84, Theorem 5] the existence of the representation (3.A.2) is proven for the Lanczos method with a sufficiently small constant C_3 and the assumption that the Krylov basis is semiorthogonal.

For the general case of the Arnoldi method the representation (3.A.2) can be derived using (3.2.6), (3.2.7a) and an additional condition on the level of orthogonality of the Krylov basis, e.g., assuming that an orthonormal basis $\widehat{V}_m$ exists for which $\|\widehat{V}_m - V_m\|_2$ is small enough (see also [BLR00, Theorem 2.1] and references therein).

3.B Some properties of divided differences

Proof of ***Proposition 3.4.1.*** For $p \in \mathbb{N}_0$ and any $A \in \mathbb{C}^{m\times m}$, $w \in \mathbb{C}^m$, from the series representation (3.2.2) we obtain

$$\int_0^t s^p \varphi_p(sA)w \,\mathrm{d}s = \int_0^t \Big(\sum_{k=0}^{\infty} \frac{s^{k+p}A^k w}{(k+p)!}\Big)\,\mathrm{d}s = \sum_{k=0}^{\infty} \frac{t^{k+p+1}A^k w}{(k+p+1)!} = t^{p+1}\varphi_{p+1}(tA)w. \tag{3.B.1}$$

This identity carries over to divided differences in the following way. Let

$$\Theta_m = \begin{pmatrix} \lambda_1 & & & \\ 1 & \lambda_2 & & \\ & \ddots & \ddots & \\ & & 1 & \lambda_m \end{pmatrix} \in \mathbb{C}^{m\times m}.$$

As a consequence of the Opitz formula, see [Opi64] and remarks in [dB05, Proposition 25], we have

$$(\varphi_p)_t[\lambda_1,\dots,\lambda_m] = e_m^* \varphi_p(t\Theta_m)e_1. \tag{3.B.2}$$

Using (3.B.1) and (3.B.2) we obtain

$$\begin{aligned}\int_0^t s^p(\varphi_p)_s[\lambda_1,\dots,\lambda_m]\,\mathrm{d}s &= e_m^* \int_0^t s^p \varphi_p(s\Theta_m)e_1\,\mathrm{d}s = e_m^* t^{p+1}\varphi_{p+1}(t\Theta_m)e_1 \\ &= t^{p+1}(\varphi_{p+1})_t[\lambda_1,\dots,\lambda_m],\end{aligned}$$

which completes the proof. □

Remark 3.B.1. *We will make use of the following integral representation for divided differences, the so-called Hermite-Genocchi formula, [Hig08, eq. (B.25)]. With the differential operator* $(D^{(m-1)}f_t)(\lambda) = \frac{\mathrm{d}^{m-1}}{\mathrm{d}\lambda^{m-1}} f(t\lambda)$,

$$\begin{aligned} f_t[\lambda_1,\dots,\lambda_m] &= \int_{[\lambda_1,\dots,\lambda_m]} D^{(m-1)}f_t \\ &= \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{m-2}} D^{(m-1)} f\Big(\lambda_1 + \sum_{j=1}^{m-1} s_j(\lambda_{j+1}-\lambda_j)\Big)\mathrm{d}s_{m-1}\dots\mathrm{d}s_2\,\mathrm{d}s_1. \end{aligned} \tag{3.B.3}$$

Proof of ***Proposition 3.4.2.*** Applying (3.B.3) to the exponential function gives

$$\begin{aligned} |\exp_t[\lambda_1,\dots,\lambda_k]| &\le \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-2}} t^{k-1}\Big|\exp\Big(\lambda_1 + \sum_{j=1}^{k-1} s_j(\lambda_{j+1}-\lambda_j)\Big)\Big|\,\mathrm{d}s_{k-1}\dots\mathrm{d}s_2\,\mathrm{d}s_1 \\ &= \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-2}} t^{k-1}\exp\Big(\xi_1 + \sum_{j=1}^{k-1} s_j(\xi_{j+1}-\xi_j)\Big)\mathrm{d}s_{k-1}\dots\mathrm{d}s_2\,\mathrm{d}s_1 \\ &= \exp_t[\xi_1,\dots,\xi_k], \end{aligned}$$

which completes the proof. □

Proof of ***Proposition 3.4.6****.* We use (3.B.3) to obtain

$$\begin{aligned}
\exp_t[\lambda_1,\dots,\lambda_k] &= \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-2}} t^{k-1} \exp\Big(t\Big(\lambda_1 + \sum_{j=1}^{k-1} s_j(\lambda_{j+1}-\lambda_j)\Big)\Big) \mathrm{d}s_{k-1} \dots \mathrm{d}s_2\, \mathrm{d}s_1 \\
&= \int_0^1 \int_0^{s_1} \cdots \int_0^{s_{k-2}} t^{k-1} \\
&\quad \cdot \Big[\cos\Big(t\Big(\eta_1 + \sum_{j=1}^{k-1} s_j(\eta_{j+1}-\eta_j)\Big)\Big) + \mathrm{i}\sin\Big(t\Big(\eta_1 + \sum_{j=1}^{k-1} s_j(\eta_{j+1}-\eta_j)\Big)\Big)\Big] \\
&\quad \cdot \exp\Big(t\Big(\xi_1 + \sum_{j=1}^{k-1} s_j(\xi_{j+1}-\xi_j)\Big)\Big)\, \mathrm{d}s_{k-1} \dots \mathrm{d}s_2\, \mathrm{d}s_1 \\
&= (\cos(tx) + \mathrm{i}\sin(ty)) \exp_t[\xi_1,\dots,\xi_k] \\
&\qquad\qquad \text{for certain } x, y \in \mathrm{Conv}(\{\eta_1,\dots,\eta_k\}).
\end{aligned}$$

Here, in the last step we have used the Mean Value Theorem for the integral. In this way we end up with the estimate

$$|\exp_t[\lambda_1,\dots,\lambda_m]| = |\cos(tx) + \mathrm{i}\sin(ty)| \cdot \exp_t[\xi_1,\dots,\xi_m].$$

With $|tx|, |ty| \le \widetilde{\eta}_t < \pi/2$ we obtain

$$\cos(\widetilde{\eta}_t) \le \cos(tx) \le |\cos(tx) + \mathrm{i}\sin(ty)|,$$

which completes the proof. □

3.C A new asymptotic expansion of divided differences

Our goal is to derive an asymptotic expansion for $|\exp_t[\lambda_1,\dots,\lambda_m]|$, see Theorem 3.C.2 at the end of this section.

Let $\lambda_1,\dots,\lambda_m \in \mathbb{C}$. We use the shortcut κ_k for the divided differences of power functions,

$$\kappa_k = (\cdot)^{m-1+k}[\lambda_1,\dots,\lambda_m] \quad \text{for} \quad k \in \mathbb{N}_0, \tag{3.C.1}$$

where $(\cdot)^j : z \mapsto z^j$ for $j \in \mathbb{N}_0$. Note that

$$(\cdot)^j[\lambda_1,\dots,\lambda_m] = 0 \quad \text{for} \quad j = 0,\dots,m-2.$$

With the notation (3.C.1) and the series representation of the exponential function we obtain

$$\exp_t[\lambda_1,\dots,\lambda_m] = \sum_{j=0}^{\infty} \frac{t^j\,(\cdot)^j[\lambda_1,\dots,\lambda_m]}{j!} = t^{m-1} \sum_{k=0}^{\infty} \frac{t^k \kappa_k}{(m-1+k)!} \tag{3.C.2a}$$

$$= \frac{t^{m-1}}{(m-1)!} + \mathcal{O}(t^m) \quad \text{for} \quad t \to 0. \tag{3.C.2b}$$

We also introduce the notation

$$S_l = \sum_{j=1}^{m} \lambda_j^l, \quad l \in \mathbb{N}. \tag{3.C.3}$$

For κ_0, κ_1 and κ_2 we obtain the following formula.

Proposition 3.C.1. *For κ_k introduced in* (3.C.1) *we have*

$$\kappa_0 = 1, \quad \kappa_1 = S_1, \quad \kappa_2 = (S_1^2 + S_2)/2.$$

Proof. This follows from [dB05, eq. (27)]. □

To simplify the notation we write

$$f(t) = |\exp_t[\lambda_1, \ldots, \lambda_m]|.$$

The following asymptotic expansion of $f(t)$ for $t \to 0$ is motivated by the concept of effective order. We define the effective order by

$$\rho(t) = \frac{f'(t)\, t}{f(t)}, \tag{3.C.4a}$$

$$\text{satisfying} \quad \rho(t)/t = \big(\log(f(t))\big)'. \tag{3.C.4b}$$

The effective order of the function $f(t)$ can be understood as the slope of the double-logarithmic function

$$\ln(f(\mathrm{e}^\tau)) \quad \text{with} \quad \tau = \ln t, \quad \text{with derivative} \quad \frac{f'(\mathrm{e}^\tau)\, \mathrm{e}^\tau}{f(\mathrm{e}^\tau)}.$$

We now analyze the divided differences close to an asymptotic regime under the assumption $f(t) > 0$, which holds for sufficiently small $t > 0$. The effective order $\rho(t)$ is then well-defined by (3.C.4a). The following expansion (3.C.5) for $\rho(t)$ is be considered in an asymptotic sense for $t \to 0$; convergence of the series is not an issue here.

We make the ansatz

$$\rho(t) = \sum_{k=0}^{\infty} \rho_k t^k \tag{3.C.5}$$

Using (3.C.5) in (3.C.4b) we obtain

$$\frac{\rho(t)}{t} = \Big(\rho_0 \log(t) + \sum_{k=1}^{\infty} \rho_k t^k/k\Big)' = (\log(f(t)))'$$

$$c \exp\Big(\rho_0 \log(t) + \sum_{k=1}^{\infty} \rho_k t^k/k\Big) = f(t),$$

$$c\, t^{\rho_0} \exp\Big(\sum_{k=1}^{\infty} \rho_k t^k/k\Big) = f(t).$$

From (3.C.2b) we see that $c = 1/(m-1)!$ and $\rho_0 = m-1$, whence

$$\rho(t) = m - 1 + \sum_{k=1}^{\infty} \rho_k t^k, \tag{3.C.6}$$

and for sufficiently small t,

$$f(t) = |\exp_t[\lambda_1, \ldots, \lambda_m]| = \frac{t^{m-1}}{(m-1)!} \exp\Big(\sum_{k=1}^{\infty} \rho_k t^k/k\Big). \tag{3.C.7}$$

We aim for deriving a formula for the coefficients ρ_k. To avoid the square roots we choose $q(t) = f(t)^2$, such that $f'(t) = q'(t)/(2q(t)^{1/2})$. Due to (3.C.4a) the effective order $\rho(t)$ satisfies

$$q(t)\rho(t) = q'(t)t/2. \tag{3.C.8}$$

We proceed by rewriting $q(t)$ and $q'(t)$ to obtain a formulation for ρ_k ($k \geq 1$) via (3.C.8). From (3.C.2a),

$$q(t) = |\exp_t[\lambda_1, \ldots, \lambda_m]|^2 = t^{2(m-1)} \Big(\sum_{k=0}^{\infty} \frac{t^k \kappa_k}{(m-1+k)!}\Big)\Big(\sum_{\ell=0}^{\infty} \frac{t^\ell \overline{\kappa}_\ell}{(m-1+\ell)!}\Big).$$

The representation of $q(t)$ as well as $tq'(t)/2$ as a Cauchy product can be written in the form

$$q(t) = \frac{t^{2(m-1)}}{((m-1)!)^2} \sum_{k=0}^{\infty} \alpha_k t^k, \quad \text{and} \quad tq'(t)/2 = \frac{t^{2(m-1)}}{((m-1)!)^2} \sum_{k=0}^{\infty} \big((m-1)+k/2\big)\alpha_k t^k, \tag{3.C.9}$$

with coefficients α_k given by

$$\alpha_0 = 1, \quad \text{and} \quad \alpha_k = \sum_{j=0}^{k} \frac{((m-1)!)^2\, \kappa_j \overline{\kappa}_{k-j}}{(m-1+j)!\,(m-1+k-j)!} \quad \text{for } k \in \mathbb{N}.$$

With $\kappa_0 = 1$ (see Proposition 3.C.1) this can be written as

$$\alpha_k = \frac{2(m-1)!\,\mathrm{Re}(\kappa_k)}{(m-1+k)!} + \sum_{j=1}^{k-1} \frac{((m-1)!)^2\, \kappa_j \overline{\kappa}_{k-j}}{(m-1+j)!\,(m-1+k-j)!} \quad \text{for } k \in \mathbb{N}. \tag{3.C.10}$$

Furthermore, from (3.C.6) and (3.C.9) we obtain a representation of $q(t)\rho(t)$ in form of a Cauchy product,

$$q(t)\rho(t) = \frac{t^{2(m-1)}}{((m-1)!)^2} \sum_{k=0}^{\infty} \theta_k t^k, \quad \text{with } \theta_k = \sum_{j=0}^{k-1} \alpha_j \rho_{k-j} + (m-1)\alpha_k, \quad k \in \mathbb{N}_0. \tag{3.C.11}$$

We remark that (3.C.11) only holds for t small enough. With $\alpha_0 = 1$, in (3.C.11) we have

$$\theta_0 = m - 1, \quad \text{and} \quad \theta_k = \rho_k + \sum_{j=1}^{k-1} \alpha_j \rho_{k-j} + (m-1)\alpha_k, \quad k \in \mathbb{N}. \tag{3.C.12}$$

For the implicit equation (3.C.8) we combine (3.C.9) and (3.C.11) to obtain

$$\sum_{k=0}^{\infty} \theta_k t^k = \sum_{k=0}^{\infty} (m - 1 + k/2)\alpha_k t^k. \tag{3.C.13}$$

Comparing coefficients of t^k in (3.C.13) and using (3.C.12) we conclude

$$\theta_k = (m - 1 + k/2)\alpha_k, \quad \text{and} \quad \rho_k = \frac{k\alpha_k}{2} - \sum_{l=1}^{k} \alpha_l \rho_{k-l}, \quad k \geq 1. \tag{3.C.14}$$

From (3.C.14) we obtain a recursion for the coefficients ρ_k in the expansion (3.C.6) which can be resolved using (3.C.1) and (3.C.10).

We now evaluate the lower coefficients of $\rho(t)$. For α_1 and α_2, using Proposition 3.C.1 in (3.C.10) gives

$$\alpha_1 = \frac{2\ \mathrm{Re}(\kappa_1)}{m} = \frac{2\ \mathrm{Re}(S_1)}{m}, \quad \text{and} \quad \alpha_2 = \frac{|\kappa_1|^2}{m^2} + \frac{2\ \mathrm{Re}(\kappa_2)}{m(m+1)} = \frac{|S_1|^2}{m^2} + \frac{\mathrm{Re}(S_1^2 + S_2)}{m(m+1)}, \tag{3.C.15}$$

with S_1, S_2 according to definition (3.C.3) From the recursion in (3.C.14) we have

$$\rho_1 = \frac{\alpha_1}{2}, \quad \rho_2 = \frac{1}{2}(2\alpha_2 - \alpha_1^2), \tag{3.C.16}$$

and combining (3.C.15) with (3.C.16) we eventually obtain

$$\begin{aligned} \rho_1 &= \frac{\mathrm{Re}(S_1)}{m}, \\ \rho_2 &= \frac{|S_1|^2}{m^2} + \frac{\mathrm{Re}(S_1^2 + S_2)}{m(m+1)} - \frac{2\ \mathrm{Re}(S_1)^2}{m^2} = \frac{\mathrm{Im}(S_1)^2 - \mathrm{Re}(S_1)^2}{m^2} + \frac{\mathrm{Re}(S_1^2 + S_2)}{m(m+1)}. \end{aligned} \tag{3.C.17}$$

To study the influence of the real and imaginary parts of the nodes $\lambda_j = \xi_j + \mathrm{i}\eta_j$ we introduce the notation

$$S_{lk} = \sum_{j=1}^{m} \xi_j^l \eta_j^k, \quad l, k \in \mathbb{N}_0. \tag{3.C.18}$$

Basic computations, mostly binomial sums in (3.C.3), show

$$S_1 = S_{10} + \mathrm{i}S_{01}, \quad S_2 = S_{20} + 2\mathrm{i}S_{11} - S_{02}, \quad \text{and} \quad S_1^2 = S_{10}^2 + \mathrm{i}S_{10}S_{01} - S_{01}^2,$$

and

$$\mathrm{Im}(S_1) = S_{01}, \quad \mathrm{Re}(S_1) = S_{10}, \quad \mathrm{Re}(S_2) = S_{20} - S_{02}, \quad \text{and} \quad \mathrm{Re}(S_1^2) = S_{10}^2 - S_{01}^2. \tag{3.C.19}$$

Combining (3.C.17) with (3.C.19) gives

$$\rho_1 = \frac{S_{10}}{m}, \quad \text{and} \quad \rho_2 = \frac{S_{01}^2 - S_{10}^2}{m^2(m+1)} + \frac{S_{20} - S_{02}}{m(m+1)}. \tag{3.C.20}$$

After all these technicalities we arrive at the following asymptotic expansion.

Theorem 3.C.2. *Assume that for $\lambda_j = \xi_j + \mathrm{i}\eta_j$ at least one of the sequences $\{\xi_j\}_{j=1}^m$ and $\{\eta_j\}_{j=1}^m$ is not constant, and $\xi_j \leq 0$ for $j = 1, \dots, m$. Let $\operatorname{avg}(\xi) = \sum_{j=1}^m \xi_j/m$ be the average and $\operatorname{var}(\xi) = \sum_{j=1}^m (\xi_j - \operatorname{avg}(\xi))^2/m$ be the variance of $\{\xi_1, \dots, \xi_m\}$, and $\operatorname{var}(\eta)$ the variance of $\{\eta_1, \dots, \eta_m\}$. Then,*

(1)

$$|\exp_t[\lambda_1, \dots, \lambda_m]| = \frac{t^{m-1}}{(m-1)!} \exp\left(\rho_1 t + \rho_2 t^2/2 + \mathcal{O}(t^3)\right) \quad \text{for } t \to 0,$$

with

$$\rho_1 = \operatorname{avg}(\xi), \quad \rho_2 = \frac{\operatorname{var}(\xi) - \operatorname{var}(\eta)}{m+1},$$

and either $\rho_1 \neq 0$ or $\rho_2 \neq 0$.

(2) The derivative of the effective order $\rho(t)$ (see (3.C.4a)) satisfies $\rho'(t) = \rho_1 + \rho_2 t + \mathcal{O}(t^2)$ for $t \to 0$, and

$$\rho'(0+) < 0.$$

Proof. We use the expansion (3.C.7) for sufficiently small t. For the variance we obtain

$$\operatorname{var}(\xi) = \frac{1}{m}\sum_{j=1}^m (\xi_j - \operatorname{avg}(\xi))^2 = \frac{1}{m}\Big(\sum_{j=1}^m \xi_j^2 - \frac{1}{m}\big(\sum_{j=1}^m \xi_j\big)^2\Big).$$

The first coefficients ρ_1 and ρ_2 are given in (3.C.20). With the notation from (3.C.18) we observe $\operatorname{avg}(\xi) = S_{10}/m$ (for the average $\operatorname{avg}(\xi)$) and $\operatorname{var}(\xi) = (S_{20} - S_{10}^2/m)/m$, $\operatorname{var}(\eta) = (S_{02} - S_{01}^2/m)/m$ (for the variance $\operatorname{var}(\xi)$ and $\operatorname{var}(\eta)$, respectively), whence

$$\rho_1 = \operatorname{avg}(\xi), \quad \text{and} \quad \rho_2 = \frac{\operatorname{var}(\xi) - \operatorname{var}(\eta)}{m+1}.$$

With $\xi_1, \dots, \xi_m \leq 0$ for $j = 1, \dots, m$ we obtain $\rho_1 \leq 0$ and $\rho_1 = 0$ iff $\xi_1, \dots, \xi_m = 0$. For the case $\xi_1, \dots, \xi_m = 0$ we obtain $\operatorname{var}(\xi) = 0$ and

$$\rho_2 = -\frac{\operatorname{var}(\eta)}{m+1} \leq 0.$$

Here, $\rho_2 = 0$ only in the trivial case with $\xi_1, \dots, \xi_m = 0$ and a constant sequence $\eta_1, \dots, \eta_m$. This proves (a). For the proof of (b) we take the derivative of $\rho(t)$ in an asymptotic sense and make use of $\rho_1 \leq 0$ and $\rho_2 < 0$ iff $\rho_1 = 0$, see (a). □

3.D Auxiliary material

3.D.1 Auxiliary remarks on stopping criteria for the lucky breakdown

In Figure 3.7 a) we illustrate the criterion of Proposition 3.4.13 for the case $p = 0$ and a skew-Hermitian problem: Following Remark 3.2.4 we choose $A = \mathrm{i}\,B$ for a Hermitian matrix B to conform to the skew-Hermitian case. Let $B \in \mathbb{R}^{n\times n}$ be a diagonal matrix with

diagonal entries $(\underbrace{1,\dots,1}_{n-9 \text{ times}},2,\dots,10) \in \mathbb{R}^n$ and $n = 5000$. Then B has exactly 10 distinct eigenvalues and the rank of the respective Krylov subspace $\mathcal{K}_m(B,v)$ is at most 10 for any m, independently of the starting vector v. For the current example we choose the normalized starting vector $v = (1/\sqrt{n},\dots,1/\sqrt{n}) \in \mathbb{R}^n$. Thus, a breakdown of the Lanczos iteration occurs after 10 iteration steps when constructing $\mathcal{K}_m(B,v)$, and the approximation to $\mathrm{e}^{\mathrm{i}tB}v$ in the Krylov subspace $\mathcal{K}_m(B,v)$ for $m = 10$ is exact for any $t > 0$ up to round-off. In this case we have $\beta h_{k+1,k} \approx 10^{-32}$ for $k = 10$ and the stopping criterion of Proposition 3.4.13 correctly detects the lucky breakdown for any reasonable choice of tol.

Furthermore, we consider the case that a lucky breakdown nearly occurs. Results for the following setting are illustrated in Figure 3.7 b). We choose $B = \mathrm{tridiag}(-1,2,-1) \in \mathbb{R}^{n\times n}$ with $n = 1000$ and we choose the normalized starting vector $v = \psi_1 + \psi_{11} + \psi_{21}$, where ψ_j refers to the eigenvectors of B as in (3.4.15), Subsection 3.4.1. Thus, the vector v conforms to a linear combination of three eigenvectors of B. The Krylov approximation $\mathrm{e}^{\mathrm{i}tB}v$ in the Krylov subspace $\mathcal{K}_m(B,v)$ corresponds to the skew-Hermitian case, $A = \mathrm{i}\,B$ as given in Remark 3.2.4. When constructing the Krylov subspace $\mathcal{K}_m(B,v)$ a lucky breakdown could be expected for $m = 3$, but due to numerical inaccuracy (also depending on the accuracy of the provided eigenvectors ψ_j) no lucky breakdown occurs in computer arithmetic: In Figure 3.7 b) we observe that the error of the Krylov approximation is above round-off level for $m = 3$, and the accuracy of the Krylov approximation can be further improved by increasing the dimension of the Krylov subspace beyond $m = 3$. Appropriately, the criterion of Proposition 3.4.13 indicates that the error per unit step is smaller than $3\cdot 10^{-8}$ for $m = 3$ and no further iteration steps are required in case this satisfies the error tolerance.

In the setting of Figure 3.7 b) the small values for the error per unit step $\|l_{p,m}(t)\|_2/t$ for $m > 3$ can be further evaluated via a time-independent bound on the defect and the respective defect-based error bound: With Corollary 3.3.3 we have

$$\|l_{p,m}(t)\|_2 \le h_{m+1,m}t^{1-p}\max_{s\in[0,t]}|\delta_{p,m}(s)|. \tag{3.D.1}$$

Following Corollary 3.3.6 the defect satisfies

$$\delta_{p,m}(t) = \beta\gamma_m t^p(\varphi_p)_t[\lambda_1,\dots,\lambda_m]. \tag{3.D.2}$$

When the eigenvalues $\lambda_1,\dots,\lambda_m$ of H_m are distinct, then the Lagrange representation (cf. [dB05, Example 9]) for the divided differences in (3.D.2) yields

$$(\varphi_p)_t[\lambda_1,\dots,\lambda_m] = \sum_{j=1}^{m}\varphi_p(t\lambda_j)\Big/\prod_{l=1,\,l\neq j}^{m}(\lambda_j-\lambda_l).$$

With $|\varphi_p(t\lambda_j)| \le 1/p!$ for $\mathrm{Re}(\lambda_j) \le 0$ this implies

$$|\delta_{p,m}(t)| \le \beta\,\gamma_m\frac{t^p}{p!}\sum_{j=1}^{m}1\Big/\prod_{l=1,\,l\neq j}^{m}|\lambda_j-\lambda_l|. \tag{3.D.3}$$

Combining (3.D.1) and (3.D.3) yields

$$\|l_{p,m}(t)\|_2/t \le \beta\,h_{m+1,m}\gamma_m\frac{1}{p!}\sum_{j=1}^{m}1\Big/\prod_{l=1,\,l\neq j}^{m}|\lambda_j-\lambda_l|. \tag{3.D.4}$$

This error bound is illustrated in Figure 3.7 b).

3.D.2 The defect norm for φ-functions with $p > 0$

We have discussed effects of clustered nodes for the divided differences of the exponential function previously in Subsection 3.4.1. This can be relevant for the defect norm $|\delta_{p,m}(t)|$ for $p > 0$: With the identities in Corollary 3.3.6 we have

$$\delta_{p,m}(t) = \beta\gamma_m \exp_t[\lambda_1, \dots, \lambda_m, 0_p].$$

Thus, the defect norm for $p > 0$ corresponds to the scaled divided differences of the exponential function with p-many additional nodes equal to zero, and for sufficiently large t the defect norm behaves similar to $\exp_t[0_p] = t^{p-1}/(p-1)!$.

In Figure 3.8 we illustrate the defect norm for $p > 0$ for the free Schrödinger example with the starting vector described in case (a) previously in Subsection 3.4.1 and $m = 20$. The respective Ritz values $\lambda_1^{(a)}, \dots, \lambda_m^{(a)}$ are shown in the table in Figure 3.1 in Subsection 3.4.1. One of the Ritz values is close to zero, i.e., $\lambda_1^{(a)} \approx 0$, and we observe that $|\delta_{p,m}(t)|$ behaves similar to $|\exp_t[\mathrm{i}\,\lambda_1^{(a)}, 0_p]|$ for $t \in [10^1, 10^4]$ approximately. For $p = 1$ this relation is illustrated by showing $\gamma_m\big(\prod_{j=2}^m \lambda_j^{(a)}\big)^{-1} |\exp_t[\mathrm{i}\,\lambda_1^{(a)}, 0]|$ in Figure 3.8. The prefactor of this term is motivated by previous results in Subsection 3.4.1. We recall $|\exp_t[\mathrm{i}\,\lambda_1^{(a)}, 0_p]| = \mathcal{O}(t^p)$ and $|\delta_{p,m}(t)| = \mathcal{O}(t^{m+p-1})$, for $t \to 0$ respectively. For the error estimation via the defect integral (3.3.1b) the asymptotic regimes of the defect norm can have a significant impact on the performance of the estimate. However, the time regime which is relevant for the error estimation depends on the actual time step which further depends on problem sizes, and in many practical examples the defect norm for a moderate $p > 0$ behaves similar to the case $p = 0$ in the relevant time regime.

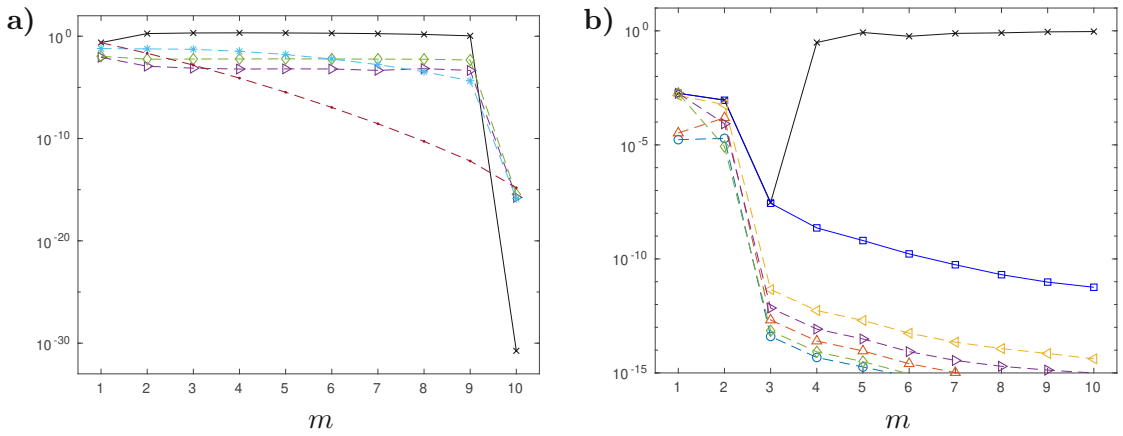


Figure 3.7: Numerical illustrations concerning the lucky breakdown and the Krylov approximation to the matrix exponential.

– Figure **a)** We choose a diagonal matrix $B \in \mathbb{R}^{n\times n}$ with diagonal entries $(1,\dots,1,2,\dots,10) \in \mathbb{R}^n$ and $n = 5000$, and a normalized starting vector $v = (1/\sqrt{n},\dots,1/\sqrt{n}) \in \mathbb{R}^n$. The line marked by $(\times)$ symbols shows $\beta h_{m+1,m}$ for the Krylov subspace $\mathcal{K}_m(B,v)$, and the dashed lines show the error per unit step $\|l_{p,m}(t)\|_2/t$ for the respective Krylov approximation to $\mathrm{e}^{\mathrm{i}tB}v$ with $t = 10^2, 10^1, 10^0, 10^{-1}$ marked by $(\triangleright, \diamond, *, \cdot)$, respectively. For $m = 10$ the value of $\beta h_{m+1,m}$ indicates a lucky breakdown, and respectively, the Krylov approximation in $\mathcal{K}_m(B,v)$ for $m = 10$ is exact up to round-off for any value of t shown here.

– Figure **b)** We choose a tridiagonal matrix $B = \mathrm{tridiag}(-1,2,-1) \in \mathbb{R}^{n\times n}$ with $n = 1000$ and a normalized starting vector $v = \psi_1 + \psi_{11} + \psi_{21} \in \mathbb{R}^n$, where ψ_j refers to the eigenvectors of B as in (3.4.15). The vector v corresponds to a linear combination of three eigenvectors of B which would cause a lucky breakdown after three iteration steps of the Lanczos method in exact arithmetic when constructing the Krylov subspace $\mathcal{K}_m(B,v)$. In computer arithmetic the Lanczos method can be continued beyond $m = 3$, and the accuracy of the Krylov approximation to $\mathrm{e}^{\mathrm{i}tB}v$ can be further improved by increasing the dimension of the Krylov subspace. The dashed lines show the error per unit step $\|l_{p,m}(t)\|_2/t$ for the respective Krylov approximation to $\mathrm{e}^{\mathrm{i}tB}v$ with $t = 10^5, 10^4, 10^3, 10^2, 10^1$ marked by $(\circ, \triangle, \triangleleft, \triangleright, \diamond)$, respectively. The line marked by $(\times)$ symbols shows $\beta h_{m+1,m}$ for the Krylov subspace $\mathcal{K}_m(B,v)$, and the line marked by $(\square)$ symbols shows the error bound given in (3.D.4). Both yield upper bounds for the error per unit step, whereat $\beta h_{m+1,m}$ corresponds to the stopping criterion for the lucky breakdown in Proposition 3.4.13.

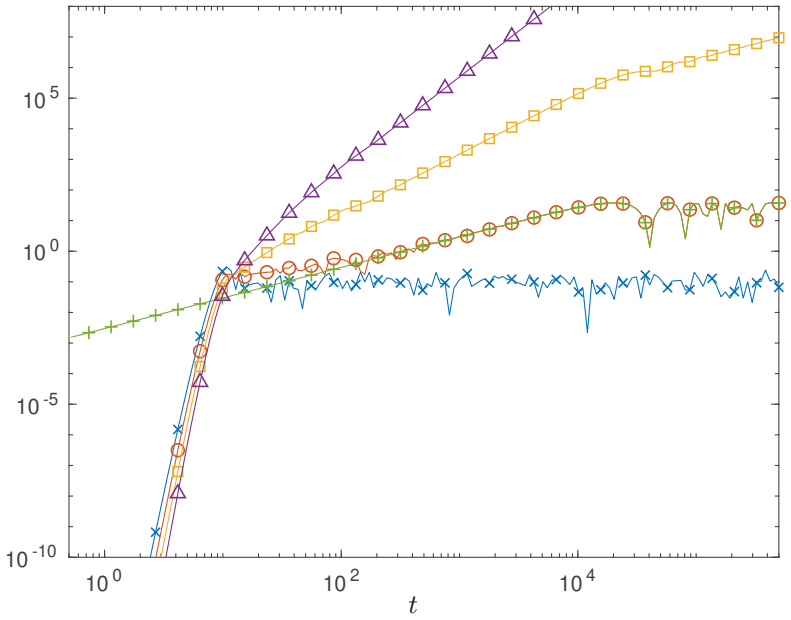


Figure 3.8: The defect norm $|\delta_{p,m}(t)|$ for $p = 0$ ($\times$), $p = 1$ ($\circ$), $p = 2$ ($\square$) and $p = 3$ ($\triangle$) for the free Schrödinger example with $m = 20$ and the randomized starting vector described in case (a), see also Figure 3.1 which includes the eigenvalues $\lambda_1^{(a)}, \ldots, \lambda_m^{(a)}$. With $\lambda_1^{(a)}$ being small the defect norm behaves similar to the divided differences $\exp_t[\lambda_1^{(a)}, 0_p]$ for sufficiently large t. This is illustrated for $p = 1$ by showing $\gamma_m \big(\prod_{j=2}^{m} \lambda_j^{(a)} \big)^{-1} \exp_t[\lambda_1^{(a)}, 0]$ marked by (+).

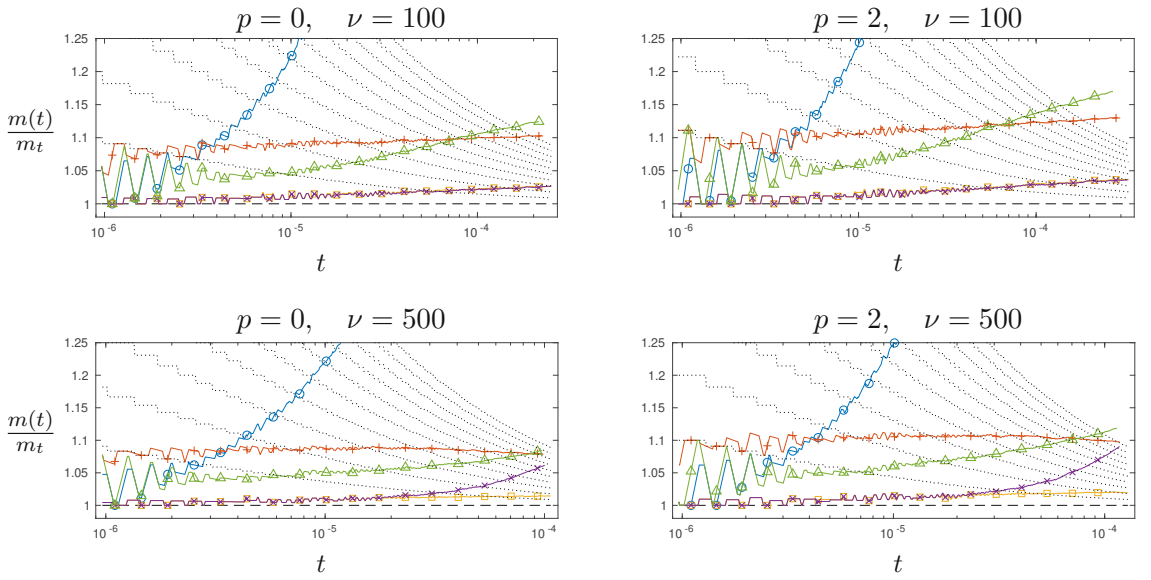


Figure 3.9: Convection-diffusion problem (3.5.1) for the parameter $\nu = 100$ (top) and $\nu = 500$ (bottom). For each choice of ν we consider $p = 0$ (left) and $p = 2$ (right). Each plot shows $m(t)/m_t$ where m_t is the smallest m such that $\|l_{p,m}(t)\|_2 \leq t \cdot \text{tol}$ for $\text{tol} = 10^{-6}$ and the exact error $l_{p,m}(t)$, and $m(t)$ is the smallest m such that $\zeta_{p,m}(t) \leq t \cdot \text{tol}$ for $\text{tol} = 10^{-6}$ and $\zeta_{p,m}$ corresponding to the following error estimates: The error bound given in Theorem 3.4.3 ($\times$), Corollary 3.4.5 ($\circ$), the generalized residual estimate given in Remark 3.4.11 ($+$), the effective order estimate given in Remark 3.4.12 ($\square$), and the error bound given in [DMR09, Proposition 6] ($\triangle$). Additionally, the dotted lines show $(m_t + k)/m_t$ for $k = 1, \ldots, 10$ as a reference. For a better visual impression the values of $m(t)/m_t$ are averaged over the time interval $[2^{-0.1}\, t, 2^{0.1}\, t]$. For the respective values of $m(t)$ see also Figure 3.3 and 3.4.

3.D.3 Numerical illustrations supplement to Section 3.5

Additional plots for the convection-diffusion problem (3.5.1)**.** In Figure 3.9 we provide additional numerical illustrations for the convection-diffusion problem (3.5.1) in Subsection 3.5. In this figure we show $m(t)/m_t$ over the time t, where m_t is the smallest Krylov dimension m such that $\|l_{p,m}(t)\|_2 \leq t \cdot \text{tol}$ and $m(t)$ is the smallest Krylov dimension m such that $\zeta_{p,m}(t) \leq t \cdot \text{tol}$ for $\text{tol} = 10^{-6}$ and $\zeta_{p,m}(t)$ corresponding to different choices of error estimates. For the values of $m(t)$ over t see also Figure 3.3 and 3.4 in Subsection 3.5.

Additional plots for the skew-Hermitian problem. In Figure 3.10 we provide an additional numerical illustration for the skew-Hermitian problem $\varphi_p(\mathrm{i}\, A)v$ as introduced in Subsection 3.5; namely, A corresponds to the Laplace operator ((3.5.1) with $\nu = 0$) and $v = (1/N, \ldots, 1/N)^*$. This figure shows $m(t)/m_t$ over the time t, where m_t is the smallest Krylov dimension m such that $\|l_{p,m}(t)\|_2 \leq t \cdot \text{tol}$ and $m(t)$ is the smallest Krylov dimension m such that $\zeta_{p,m}(t) \leq t \cdot \text{tol}$ for $\text{tol} = 10^{-6}$ and $\zeta_{p,m}(t)$ corresponding to different choices

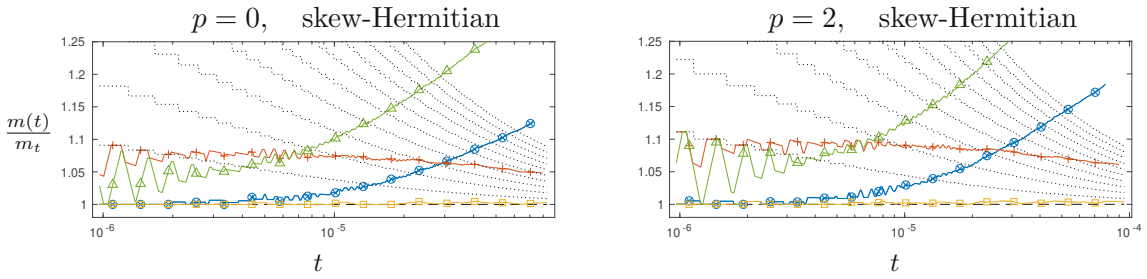


Figure 3.10: The skew-Hermitian problem $\varphi_p(\mathrm{i}\,A)v$ where A corresponds to the Laplace operator ((3.5.1) with $\nu = 0$) and $v = (1/N, \ldots, 1/N)^*$. Results are shown for $p = 0$ (left) and $p = 2$ (right). Each plot shows $m(t)/m_t$ where m_t is the smallest m such that $\|l_{p,m}(t)\|_2 \leq t \cdot \mathrm{tol}$ for $\mathrm{tol} = 10^{-6}$ and the exact error $l_{p,m}(t)$, and $m(t)$ is the smallest m such that $\zeta_{p,m}(t) \leq t \cdot \mathrm{tol}$ for $\mathrm{tol} = 10^{-6}$ and $\zeta_{p,m}$ corresponding to the following error estimates: The error bound given in Theorem 3.4.3 (×), Corollary 3.4.5 (○), the generalized residual estimate given in Remark 3.4.11 (+), the effective order estimate given in Remark 3.4.12 (□), and the error bound given in [DMR09, Proposition 8] (△). The error bounds in Theorem 3.4.3 (×) and Corollary 3.4.5 (○) coincide in the skew-Hermitian case. Additionally, the dotted lines show $(m_t + k)/m_t$ for $k = 1, \ldots, 10$ as a reference. For a better visual impression the values of $m(t)/m_t$ are averaged over the time interval $[2^{-0.1}\,t, 2^{0.1}\,t]$. For the respective values of $m(t)$ see also Figure 3.5.

of error estimates. For the respective values of $m(t)$ see also Figure 3.5 in Subsection 3.5.

4 A review of the Separation Theorem of Chebyshev-Markov-Stieltjes for polynomial and some rational Krylov subspaces

4.1 Introduction and historical context

In the present chapter we consider an Hermitian matrix $A \in \mathbb{C}^{n\times n}$ and a given vector $u \in \mathbb{C}^n$. The coefficients of u in the orthonormal eigenbasis of A are referred to as spectral coefficients, see (4.2.3) below. These coefficients rely on an underlying inner product on $\mathbb{C}^n$ which is specified in (4.2.2), and denoted as M-inner product in the sequel. Furthermore, Krylov subspaces in the sequel also rely on the M-inner product.

4.1.1 Historical context and previous works

For a polynomial or rational Krylov subspace of a matrix A with starting vector u, the spectral coefficients of u play a crucial role: The linear functional $f \mapsto (u, f(A)\,u)_{\mathsf{M}}$ can be understood as a Riemann-Stieltjes integral associated with a non-decreasing step function α_n. This step function is defined by the eigenvalues of A and the spectral coefficients of u, and many results concerning the theory of polynomial Krylov subspaces have their origin in the theory of orthogonal polynomials, namely, polynomials on the real axis which are orthogonal w.r.t. the Riemann-Stieltjes integral associated with α_n; see also [GM10] for a survey. We also refer to these polynomials as orthogonal polynomials associated with the distribution $\mathrm{d}\alpha_n$. In a similar manner, orthogonal rational functions describe rational Krylov subspaces.

For the polynomial case, the Lanczos method [Lan50] is used in practice to generate an M-orthonormal basis of the Krylov subspace and the associated Jacobi matrix, which corresponds to the representation of A in the respective Krylov subspace, see also [Saa11]. The respective M-orthonormal basis vectors satisfy a three-term recursion which conforms to the three-term recursion of the underlying orthogonal polynomials associated with $\mathrm{d}\alpha_n$; the Krylov basis and the orthogonal polynomials exist in an equivalent manner.

The Jacobi matrix associated with orthogonal polynomials for a given distribution plays a crucial role for Gaussian quadrature formulae for the respective Riemann-Stieltjes integral, which also goes by the name Gauss-Christoffel quadrature, cf. [Gau81]. For the Gauss-Christoffel quadrature formula with m quadrature nodes which integrates polynomials of degree $\leq 2m-1$ exactly, the quadrature nodes are given by the zeros of the $(m+1)$-th orthogonal polynomial and the quadrature weights are given by so called Christoffel numbers. Early works on quadrature formulae [Wil62, GW69] (historical remarks in [Gau81] also refer to earlier works of Goertzel) show that these quadrature nodes and weights can be

computed via the Jacobi matrix; the zeros of the $(m+1)$-th orthogonal polynomial coincide with eigenvalues of the respective Jacobi matrix, and the Christoffel numbers correspond to entries of its eigenvectors. In these works, the underlying distribution is not necessarily based on a matrix-vector pair as it is the case when considering a polynomial Krylov subspace; a reference to the Krylov setting is made later in [FF94, FH93] and also discussed in detail (including historical remarks) in [GM10, LS13]. In this context, the eigenvalues of the Jacobi matrix are also referred to as Ritz values, and m denotes the dimension of the Krylov subspace. Furthermore, the Christoffel numbers, which are given by entries of the eigenvectors of the Jacobi matrix, can be written as spectral coefficients of a vector $x \in \mathbb{R}^m$. In particularly, the vector x corresponds to the representation of the starting vector u in the Krylov subspace, i.e., the first unit vector scaled by the norm of u. Here, the spectral coefficients of $x \in \mathbb{R}^m$ denote its coefficients in the ℓ^2-orthonormal[1] eigenbasis of J_m.

The Separation Theorem of Chebyshev-Markov-Stieltjes (CMS Theorem) states that accumulated quadrature weights of a Gaussian quadrature formula (i.e., the accumulated Christoffel numbers) are bounded by Riemann-Stieltjes integrals over the interval between the left integral limit and the quadrature nodes. For details and historical remarks see [Sze85, Akh65, VA93]. In an equivalent manner, this statement can be formulated in a Krylov setting: The accumulated entries of eigenvectors of the Jacobi matrix (spectral coefficients of x) are bounded by Riemann-Stieltjes integrals associated with α_n over the interval between the left-most eigenvalue of A and the Ritz values. The step function α_n corresponds to accumulated spectral coefficients of u, and as a corollary, accumulated spectral coefficients of x yield bounds on sums of spectral coefficients of u and vice versa. Analogously, this statement can be formulated as an intertwining property of the distribution $\mathrm{d}\alpha_n$ and a distribution $\mathrm{d}\alpha_m$ associated with the step function α_m which is defined by Ritz values and spectral coefficients of x: Similar to $f \mapsto (u, f(A)\,u)_{\mathsf{M}}$ and α_n, the functional $f \mapsto (x, f(J_m)\,x)_2$ can be understood as a Riemann-Stieltjes integral associated with the step function α_m. The underlying Gaussian quadrature formula implies $(u, p(A)\,u)_{\mathsf{M}} = (x, p(J_m)\,x)_2$ for polynomials p of degree $\leq 2m-1$, and therefore, the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ have the same moments up to degree $2m-1$.

For distributions with the same moments, an intertwining property is stated in [KS53, Theorem 22.2], see also [Fis96, Theorem 2.2.5] and [LS13, Theorem 3.3.4]. Indeed, this intertwining property coincides with the result of the CMS Theorem. In the context of Krylov subspaces, the intertwining property of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ already appeared earlier in [FF94, Fis96, LS13]. For further remarks (including many historical remarks) on the moment problem we particularly refer to [LS13]. The identity $(u, p(A)\,u)_{\mathsf{M}} = (x, p(J_m)\,x)_2$ above corresponds to quadrature properties, and results from well-known identities for polynomials in the Krylov subspace in an equivalent manner; $p(A)u = V_m\,p(J_m)x$ for polynomials p of degree $\leq m-1$ and $p(A)u - V_m\,p(J_m)x \perp_{\mathsf{M}} V_m$ for polynomials p of degree m where $V_m \in \mathbb{C}^{m\times n}$ denotes the M-orthonormal Krylov basis written in matrix form.

The related theory in [Sze85, Akh65] applies in a slightly more general setting, namely, for Gaussian quadrature formulae which integrate polynomials of degree $\leq 2m-2$ exactly (where m is the number of quadrature nodes). This includes Gauss-Radau quadrature

[1] The notation 'ℓ^2-orthonormal' refers to a basis orthonormal w.r.t. the Euclidean inner product.

formulae where one of the m quadrature nodes is preassigned. The quadrature nodes and weights of Gauss-Radau quadrature formulae can be represented by the zeros of a so-called quasi-orthogonal polynomial and the Christoffel numbers associated with this polynomial, respectively. Similar to the Jacobi matrix, the recursion of the underlying set of polynomials constitutes a tridiagonal structure in matrix form, and the respective quadrature nodes and weights correspond to the eigenvalues and entries of eigenvectors, respectively, of this tridiagonal matrix. This relation between Gauss-Radau quadrature formulae and the eigendecomposition of this Jacobi-like tridiagonal matrix goes back to [Gol73] and is reviewed in detail in [GM10, Section 6.2].

In the present work we also consider rational Krylov subspaces, namely, subspaces spanned by $\{r(A)u\}$ where $r = p/q$ for polynomials p of degree $\leq m-1$ and a preassigned denominator polynomial q of degree $\leq m-1$ (here, p and q are complex polynomials and m again denotes the dimension of the Krylov subspace). For early works on rational Krylov subspaces we refer to [ER80, Ruh84], and for a review we refer to [Güt10]. The zeros of the denominator q are also referred to as poles in this context. Rational Krylov techniques using a single pole of multiplicity $m-1$ yield the most prominent cases, the resulting rational Krylov subspaces are also referred to as *Shift-and-Invert (SaI)* Krylov subspaces.

The rational Krylov subspace with preassigned denominator polynomial q and starting vector u is identical to the polynomial Krylov subspace with starting vector $q(A)^{-1}u$. The respective orthogonal polynomials (particularly, orthogonal polynomials associated with a scaled distribution $\mathrm{d}\widehat{\alpha}_n$) divided by the denominator polynomial q yield rational functions which are orthogonal w.r.t. the Riemann-Stieltjes integral associated with α_n (as given previously), cf. [DB07]. These orthogonal rational functions, evaluated at A as a matrix function and applied to u, provide an M-orthonormal basis of the rational Krylov subspace. Furthermore, results regarding Gaussian quadrature formulae carry over to the rational setting: The orthogonal rational functions which span the rational Krylov subspace of dimension m with a preassigned denominator q constitute a rational quadrature formula for the Riemann-Stieltjes integral associated with α_n, which integrates rational functions $r = p/|q|^2$ exactly for polynomials p of degree $\leq 2m-1$. For an overview on rational Gaussian quadrature see also [Gau93], and for the relation between rational Krylov subspaces and rational Gaussian quadrature we also refer to [LLRW08, Dec09, JR11].

The relation between a rational Krylov subspace with denominator q and starting vector u, and the polynomial Krylov subspace with starting vector $q(A)^{-1}u$ is more of a theoretical nature. In practice, various algorithms, covering different settings, are relevant to construct a rational Krylov subspace, and result in different sequences of M-orthonormal basis vectors of this subspace. To keep our results general, we do no restrict ourselves to a specific algorithm or an underlying recursion for the basis vectors in that concern. Assuming an M-orthonormal basis of a rational Krylov subspace is given, we refer to the representation of A in this basis as *Rayleigh quotient* $A_m \in \mathbb{C}^{m\times m}$. Furthermore, we reuse the notation $x \in \mathbb{C}^m$ for the representation of u in the given basis. As stated above, a rational Krylov subspace is closely related to orthogonal rational functions which constitute a rational Gaussian quadrature formula. In particular, the quadrature nodes and weights for this rational Gaussian quadrature formula correspond to the eigenvalues of the Rayleigh quotient A_m and the spectral coefficients of x, respectively. We remark that the eigenvalues of A_m (also referred to as rational Ritz values, which are real due to A_m being

Hermitian) and the spectral coefficients of x, which refer to the coefficients of x in the ℓ^2-orthonormal eigenbasis of A_m, are independent of the choice of the basis. Furthermore, the respective quadrature formula conforms to the identity $(u, r(A)\,u)_{\mathsf{M}} = (x, r(A_m)\,x)_2$ for rational functions $r = p/|q|^2$ as above.

Similar to the polynomial case, the functional $f \mapsto (x, f(A_m)\,x)_2$ can be understood as a Riemann-Stieltjes integral associated with α_m, which is now defined by eigenvalues of A_m and the spectral coefficients of x. The rational quadrature properties imply that $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ have $2m-1$ identical rational moments.

For rational Gaussian quadrature formulae, CMS type results depend on the choice of the denominator, and do not seem to be as popular as for the polynomial case. In [Li98] a separation theorem is given for a class of Laurent polynomials and an integral defined on the positive real axis. Here, Laurent polynomials correspond to rational functions with denominator $q(\lambda) = \lambda^{m/2}$ for even m. In a Krylov setting, this class of rational functions is related to some extended Krylov subspaces [DK98] for a matrix A with positive eigenvalues (i.e., the step function α_n is defined on the positive real axis). However, the results of [Li98] have not been applied in a Krylov setting yet.

More recently, [ZTK19] computes piecewise estimates on α_n based on a Shift-and-Invert Krylov subspace with a pole of multiplicity $m-1$ at zero (i.e., $q(\lambda) = \lambda^{m-1}$), for a matrix A with positive eigenvalues. In this work, a Shift-and-Invert representation is used instead of the Rayleigh quotient (see also [Güt10, Subsection 5.4.3]). The given estimates are based on an intertwining property of $\mathrm{d}\alpha_n$ and a distribution given by spectral properties of the Shift-and-Invert representation; the intertwining property goes back to the polynomial case, referring to [KS53, Theorem 22.2].

In the present work, we also consider Krylov techniques related to rational Gauss-Radau quadrature formulae. These quadrature formulae integrate rational function $r = p/|q|^2$ exactly, where p is a polynomial of degree $\leq 2m-2$, q is the given denominator, and one of the m quadrature nodes is preassigned, see also [LLRW08, JR13]. For rational Gauss-Radau quadrature formulae in a more general setting see also [Gau04, DBVD10, DB12]. Analogously to the Gauss-Radau quadrature formulae in the polynomial case, this slightly generalizes the previously discussed rational quadrature properties but can be treated similarly concerning the intertwining properties of the underlying distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$.

4.1.2 Applications

Computable estimates on spectral coefficients of u. A direct computation of eigenvalues and spectral coefficients requires access to the eigenbasis of the given matrix A which is not practical for problems of a large problem size n in general; typically, the full spectrum of A is not available. However, information on partly accumulated spectral coefficients, namely, the step function α_n on subsets of the spectrum of A, can be sufficient for some applications. CMS type results provide suitable estimates for this purpose, which can be evaluated using Krylov techniques. In particularly, this yields piecewise estimates on α_n covering the full spectrum of A. These estimates hold true independently of the convergence of individual (rational) Ritz values. However, more detailed information is provided for parts of the spectrum which are well resolved by (rational) Ritz values. We proceed to give

some applications based on estimates on α_n.

The eigenvalues of A together with the spectral coefficients of u have some relevance for the approximation of the action of a matrix function $f(A)u$, e.g., the matrix exponential function or the matrix inverse. Polynomial Krylov methods yield good approximations on matrix functions without any a priori information on the spectrum of A (such approximants are discussed in Chapter 2 and 3 for the matrix exponential function). However, further knowledge on the spectrum of A can help to improve the quality of the approximation (here we also refer to the introduction of [FH93]). In [FF94], piecewise estimates on α_n are applied to construct a polynomial preconditioner for the conjugate gradient method. This approach is based on the intertwining property of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$, where the latter is computed using a small number of Lanczos iterations in the progress (thus, α_m is based on a polynomial Krylov subspace here).

In [HPS09], the authors consider iterative bidiagonalization methods to solve ill-posed linear systems. In this work, effects of a noisy right-hand side on the projected problem are discussed. The ill-posed problems therein are associated with an underlying distribution (similar to $\mathrm{d}\alpha_n$ given previously in the present introduction), and due to problem assumptions and noise on the initial data this distribution is of a special structure which carries over to the projected problem. This process is closely related to the intertwining property of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ in the Lanczos case, and results in criteria to detect the noise level on the run, as introduced in [HPS09].

In [ZTK19], an inhomogeneous differential equation, arising in applications of dynamic analysis of structure, is diagonalized using eigenvectors of a large matrix. This requires computation of a moderate number of eigenvectors, namely, eigenvectors such that the external force vector is resolved with sufficient accuracy. The spectral decomposition of this vector is associated with a distribution $\mathrm{d}\alpha_n$, and estimates on this distribution allow to determine intervals which cover eigenvalues corresponding to the required eigenvectors. In [ZTK19], estimates on α_n are based on a Shift-and-Invert Lanczos method, and yield a pole selection strategy and stopping criteria for an eigenproblem solver based on rational Krylov methods.

In future works, estimates on α_n will be applied to design special rational approximations to the action of the exponential of skew-Hermitian matrices.

The structure of α_n roughly carries over to α_m. In Chapter 5, we consider a localized best approximation property of rational Krylov approximants to the action of a matrix exponential. In particularly, we consider the exponential of a skew-Hermitian matrix applied to a vector which is subject to some assumptions. Namely, strict increases of α_n are, up to a small perturbation, located in an interval. For some rational Krylov subspaces we illustrate that such properties carry over to the associated step function α_m. These ideas are based on theoretical results derived in the present chapter; and in Chapter 5 this approach motivates a localized best approximation result which can show a mesh-independent convergence (in a setting where the matrix exponential arises from a spatial discretization of a PDE (evolution equation)). In contrast to previously mentioned applications, computable estimates on α_n are not topical for the approach of Chapter 5.

Other applications. Apart from the Krylov setting, the CMS Theorem has applications in various fields, e.g., for a work on discretization of quantum systems see [Rei79].

Furthermore, bounds on distribution functions have some importance in probability theory and statistics; and various bounds are referenced to Chebyshev, Markov, Stieltjes and others. This includes variants of the CMS Theorem formulated in terms of moments, e.g., [Zel54] or more recently [Hür15]. Moment-matching methods also appear in the context of system theory [Ant05].

Krylov methods also have applications in the approximation to bilinear forms $(u, f(A)\,u)_{\mathsf{M}}$, where f is a given function, see also [LLRW08, GM10, JR11, JR13]. Due to the relation between $(u, f(A)\,u)_{\mathsf{M}}$ and a Riemann-Stieltjes integral associated with α_n, estimates on this bilinear form are directly related to quadrature formulae. However, these applications will not be further discussed in the present work.

4.1.3 Main contributions and overview of present work

We proceed to highlight the main contributions of the present chapter, including results or remarks which are considered to be new by the author.

- We introduce a new CMS type result for a class of rational Gaussian quadrature formulae, namely, quadrature formulae based on rational functions with a single real pole of higher multiplicity, see Theorem 4.4.11 in Subsection 4.4.3. To prove this result, we introduce rational majorants and minorants on Heaviside type functions in Proposition 4.4.12. In a Krylov setting, this theorem applies to the SaI Krylov subspace with a real shift. Our results include the case that the shift is located in the contour of the matrix spectrum; we consider a more general setting compared to [ZTK19]. An intertwining property of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ holds true up to a constant, see Proposition 4.4.13.

- For the setting of rational functions with a single complex pole of higher multiplicity, we introduce a new CMS type result which yields an upper bound on the Riemann-Stieltjes integral over the interval between neighboring quadrature nodes and at the boundary, see Proposition 4.4.19 in Subsection 4.4.4. This result applies to the SaI Krylov subspace with a single *complex* shift of higher multiplicity. To prove this upper bound, we make use of polynomial majorants on Heaviside type functions on the unit circle given in [Gol02]. Furthermore, we propose the use of an isometric Arnoldi method to compute the Rayleigh quotient of the SaI Krylov subspace with complex shift in a cost efficient way (comparable to the Lanczos method which applies when the shift is real), see Remark 4.2.7.

- Applying a CMS type result given in [Li98], we present an intertwining property for $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ in the setting of an extended Krylov subspace in Subsection 4.4.5.

Recalling results of [GM10] and others, we also apply the theory of quasi-orthogonal polynomials in a polynomial Krylov setting. This results in an Arnoldi-like decomposition where the residual is provided by a quasi-orthogonal polynomial; we refer to the respective representation as a *quasi-orthogonal residual (qor-)* Krylov representation for which one of the eigenvalues can be preassigned.

- The CMS Theorem is known to apply to Gauss-Radau quadrature formulae. In the present work, we specify these results in a Krylov setting; results in Section 4.4 for the polynomial case include the qor-Krylov setting, e.g., the intertwining property of $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ holds true when α_m is based on the qor-Krylov representation. This potentially leads to refined estimates on α_n in practice.

- Furthermore, we introduce a qor-Krylov approximation to the action of matrix functions in Subsection 4.3.1, comparable to the corrected Krylov scheme for the matrix exponential function given in [Saa92].

Various results for the polynomial case carry over to the rational case, and we introduce a rational qor-Krylov representation where one of the eigenvalues is preassigned, similar to [LLRW08, JR13].

- For the rational case, we introduce an efficient procedure to compute a rational qor-Krylov representation in Subsection 4.3.2.

- The CMS type result given in Subsection 4.4.3 and further estimates in Subsection 4.4.4 include the rational qor-Krylov case. Considering these CMS theorems, for some cases bounds on quadrature weights related to quadrature nodes at the right boundary (of the spectrum of A) are affected by α_n at the left boundary (of the spectrum of A) and vice versa, e.g., as in Corollary 4.4.16; α_n affects the bounds in a cycled sense at the boundaries. This is no longer the case when one of the nodes is preassigned at the boundary of the spectrum, see also Remark 4.4.17, and this potentially results in refined bounds.

- We introduce a rational qor-Krylov approximation to the action of matrix functions in Subsection 4.3.2.

Overview of present work In Section 4.2 we first recall some theory of orthogonal polynomials and the relation between orthogonal polynomials and the polynomial Krylov subspace. Here, polynomials are orthogonal w.r.t. an inner product on the vector space, which can be written as a Riemann-Stieltjes integral associated with a non-decreasing step function α_n. Furthermore, we recall some known results for rational Krylov subspaces based on the polynomial case. In Subsection 4.2.1 we provide some remarks on the SaI Krylov subspace. This includes a new approach to compute the SaI Krylov subspace with a complex shift based on the isometric Arnoldi method – a short-term recursion. In Section 4.3 we recall some theory on quasi-orthogonal polynomials which results in a polynomial and rational qor-Krylov representation in Subsection 4.3.1 and 4.3.2, respectively. Here, we also include some algorithmic details.

The main results of the present chapter concerning CMS theorems and intertwining properties of distributions are stated in Section 4.4. We first recall quadrature properties in Subsection 4.4.1 concerning polynomial and Gaussian quadrature formulae for the Riemann-Stieltjes integral associated with the step function α_n. Quadrature nodes and weights for these quadrature formulae are provided by the Jacobi matrix or the Rayleigh quotient of the respective Krylov subspace. The following results in Subsection 4.4.2–4.4.5

are stated for quadrature nodes and weights of respective quadrature formulae, and as such apply to eigenvalues and spectral coefficients for representations in the respective Krylov subspaces. In Subsection 4.4.2 we recall the classical CMS Theorem which applies to the polynomial Krylov setting. Besides other remarks in this subsection, we also specify the step function α_m and recall the intertwining property of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$. In Subsection 4.4.3 we introduce new results concerning rational Gaussian quadrature formulae for a class of rational functions with a single pole $s \in \mathbb{R}$ of higher multiplicity. This result applies to the SaI Krylov setting with a real shift, and the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ (whereat, $\mathrm{d}\alpha_m$ is now provided by the rational Krylov subspace) satisfy an intertwining property up to a constant shift. In Subsection 4.4.4 we proceed with a similar upper bound for the rational case with a single pole $s \in \mathbb{C}$ of higher multiplicity, which corresponds to a SaI Krylov setting with a complex shift. In Subsection 4.4.5 we apply CMS theorems given in [Li98] in the setting of an extended Krylov subspace, which yields results similar to the polynomial case. Previously discussed intertwining properties which correspond to CMS theorems are verified by numerical examples in Section 4.5.

4.2 Krylov subspace techniques and orthogonal polynomials

A basis of a Krylov subspace obtained by the Lanczos method is closely related to the theory of orthogonal polynomials. This relationship is explained in [GM10] and others and is reviewed here.

In the sequel, let $A \in \mathbb{C}^{n\times n}$ be a given Hermitian matrix, and let $u \in \mathbb{C}^n$ be a given initial vector. The polynomial Krylov subspace, with $m \leq n$, is denoted by

$$\mathcal{K}_m(A, u) = \operatorname{span}\{u, A\,u, \ldots, A^{m-1}u\} \ \subset \mathbb{C}^n. \tag{4.2.1}$$

Krylov subspace techniques rely on an inner product. Although the Euclidean inner product on the underlying vector space is practical in many cases, we consider a more general notation: For two vectors $x, y \in \mathbb{C}^n$ we define the M-inner product by[2]

$$(x, y)_{\mathsf{M}} = x^{\mathsf{H}} M y, \tag{4.2.2}$$

where $M \in \mathbb{C}^{n\times n}$ is an Hermitian[3] positive definite matrix which is given by the underlying problem setting. This notation includes the Euclidean inner product, namely, the case $M = I$ with[4] $(x, y)_{\mathsf{M}} = (x, y)_2$. In the current work, the motivation behind the M-inner product lies in problems which are based on discretized Hilbert spaces, e.g., for a FEM discretization of the Hilbert space L^2 (on a spatial domain) the inner product $(x, y)_{\mathsf{M}} = x^{\mathsf{H}} M y$ with M representing the mass matrix of the finite element space is a natural choice.

In the sequel we assume that A is Hermitian (self-adjoint) w.r.t. the M-inner product,

$$(A\,x, y)_{\mathsf{M}} = (x, A\,y)_{\mathsf{M}}, \quad x, y \in \mathbb{C}^n.$$

[2] The M-inner product given in (4.2.2) induces a vector norm, i.e., $\|x\|_{\mathsf{M}} = \sqrt{(x, x)_{\mathsf{M}}}$, which is equivalent to the Euclidean norm.

[3] The matrix M is Hermitian w.r.t. the Euclidean inner product, i.e., $M = M^{\mathsf{H}}$.

[4] By $(\cdot, \cdot)_2$ and $\|\cdot\|_2$ we denote the Euclidean inner product and norm, respectively.

Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ denote the eigenvalues and $q_1, \ldots, q_n \in \mathbb{C}^n$ the M-orthonormal eigenvectors of $A \in \mathbb{C}^{n\times n}$, i.e., $Aq_j = \lambda_j q_j$ with $(q_j, q_k)_\mathsf{M} = \delta_{jk}$, and let

$$w_j = (q_j, u)_\mathsf{M} \in \mathbb{C} \tag{4.2.3}$$

denote the corresponding spectral coefficients of the initial vector $u \in \mathbb{C}^n$, i.e.,

$$u = \sum_{j=1}^{n} w_j\, q_j.$$

In practice, the Lanczos method (cf. [Saa03]) delivers an M-orthonormal[5] basis $V_m = (v_1, \ldots, v_m) \in \mathbb{C}^{n\times m}$ of the Krylov subspace $\mathcal{K}_m(A, u)$, i.e.,

$$\operatorname{span}(V_m) = \mathcal{K}_m(A, u), \quad \text{and} \quad (V_m, V_m)_\mathsf{M} = I,$$

for which the starting vector u satisfies

$$(V_m, u)_\mathsf{M} = \beta_0\, e_1, \quad \beta_0 = \|u\|_\mathsf{M} \quad \text{and} \quad e_1 = (1, 0, \ldots, 0)^\mathsf{H} \in \mathbb{R}^m.$$

'Full rank' of $\mathcal{K}_m(A, u)$ means

$$\operatorname{rank}\left(u, A\,u, \ldots, A^{m-1}u\right) = m. \tag{4.2.4}$$

To proceed with the construction of $\mathcal{K}_{m+1}(A, u)$ at the m-th Lanczos iteration step we require (4.2.4) to hold also for $m+1$. Otherwise $\mathcal{K}_m(A, u)$ is an invariant subspace of A, and we refer to this case as a *lucky breakdown* after m steps. We remark that only if there exist at least m coefficients $w_j \neq 0$ with distinct eigenvalues λ_j, then (4.2.4) holds true for a respective[6] m. In the sequel we will assume that no lucky breakdown occurs: I.e., without loss of generality we assume that (4.2.4) holds true for $m \leq n$, hence we consider $w_j \neq 0$ with distinct eigenvalues λ_j for $j = 1, \ldots, n$. We further assume the ordering

$$\lambda_1 < \lambda_2 < \ldots < \lambda_n.$$

With (4.2.1) there exist polynomials $p_0, \ldots, p_{m-1}$ which satisfy

$$v_\ell = p_{\ell-1}(A)u, \quad \ell = 1, \ldots, m.$$

For these polynomials the orthonormal property of V_m, i.e., $(v_\ell, v_k)_\mathsf{M} = \delta_{\ell k}$, yields

$$(p_{\ell-1}(A)u, p_{k-1}(A)u)_\mathsf{M} = \delta_{\ell k}, \qquad \ell, k = 1, \ldots, m. \tag{4.2.5}$$

Various properties of Krylov subspaces have their origin in the theory of orthogonal polynomials for which we mainly refer to [Sze85, Akh65]. The theory therein can be formulated in terms of an integral-based inner product: Following [GM10], depending on u we consider the step function

$$\alpha_n(\lambda) = \begin{cases} 0, & \lambda < \lambda_1, \\ \sum_{j=1}^{\ell} |w_j|^2, & \lambda_\ell \leq \lambda < \lambda_{\ell+1}, \quad \ell = 1, \ldots, n-1, \\ \sum_{j=1}^{n} |w_j|^2, & \lambda_n \leq \lambda. \end{cases} \tag{4.2.6a}$$

[5] For two vectors $x, y \in \mathbb{C}^m$ an M-orthonormal basis V_m satisfies $(V_m\, x, V_m\, y)_\mathsf{M} = (x, y)_2$.

[6] See Proposition 4.A.1, Appendix 4.A.

We choose an interval (a, b) which includes $\lambda_1, \dots, \lambda_n$. For $f\colon \mathbb{R} \to \mathbb{C}$ we have

$$\sum_{j=1}^{n} |w_j|^2 f(\lambda_j) = \int_a^b f(\lambda)\, \mathrm{d}\alpha_n(\lambda), \tag{4.2.6b}$$

where the right-hand side is to be understood as a Riemann-Stieltjes integral. For the corresponding inner product we introduce the notation

$$(f, g)_{\alpha_n} = \int_a^b \overline{f}\,(\lambda) g(\lambda)\, \mathrm{d}\alpha_n(\lambda). \tag{4.2.6c}$$

In the eigenbasis of A the vector $p(A)u$, where p is a polynomial, has the representation

$$p(A)u = \sum_{j=1}^{n} p(\lambda_j)\, w_j\, q_j.$$

For two complex polynomials p and g the M-inner product of $p(A)u$ and $g(A)u$ reads

$$(p(A)u, g(A)u)_{\mathsf{M}} = \sum_{j=1}^{n} |w_j|^2\, \overline{p}(\lambda_j) g(\lambda_j). \tag{4.2.7}$$

With (4.2.6b), (4.2.6c) and (4.2.7) we have the equivalent formulations

$$(p, g)_{\alpha_n} = \int_a^b \overline{p}(\lambda) g(\lambda)\, \mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{n} |w_j|^2\, \overline{p}(\lambda_j) g(\lambda_j) = (p(A)u, g(A)u)_{\mathsf{M}}. \tag{4.2.8}$$

Thus, polynomials which satisfy (4.2.5) are indeed 'α_n-orthonormal', i.e.,

$$(p_\ell, p_k)_{\alpha_n} = \delta_{\ell k}, \quad \ell, k = 0, \dots, m-1. \tag{4.2.9}$$

We remark that the normalization factor β_0 as given previously satisfies the identities

$$\beta_0 = \big((u,u)_{\mathsf{M}}\big)^{1/2} = \big((1,1)_{\alpha_n}\big)^{1/2} = \Big(\int_a^b 1\, \mathrm{d}\alpha_n(\lambda)\Big)^{1/2}. \tag{4.2.10}$$

Three-term recursion, zeros of orthogonal polynomials, and the Jacobi matrix. Our assumption that a lucky breakdown does not occur for any $m < n$ corresponds to $w_j \neq 0$ and λ_j being distinct for $j = 1, \dots, n$ and entails that the step function α_n has n points of strict increase. Following [Sze85, Section 2.2] the respective inner product yields orthonormal polynomials $p_0, \dots, p_{n-1}$ of degree $0, \dots, n-1$, respectively. These polynomials enjoy a three-term recursion, see also [GM10, Section 2.2] or [Akh65, Sze85]:

Proposition 4.2.1. *Let $\beta_0 = (\int_a^b 1\, \mathrm{d}\alpha_n)^{1/2}$ as in (4.2.10). With $p_0 = 1/\beta_0$, $p_{-1} = 0$ and $m < n$ there exist $a_1, \dots, a_m \in \mathbb{R}$, $\beta_1, \dots, \beta_m > 0$ and α_n-orthonormal polynomials $p_0, \dots, p_m$ for which the three-term recursion*

$$\lambda\, p_{j-1}(\lambda) = \beta_{j-1} p_{j-2}(\lambda) + a_j p_{j-1}(\lambda) + \beta_j p_j(\lambda), \quad j = 1, \dots, m, \tag{4.2.11}$$

holds. Here, $a_j = (p_{j-1}, \lambda p_{j-1})_{\alpha_n}$, and $\beta_j > 0$ is fixed such that $(p_j, p_j)_{\alpha_n} = 1$.

In the sequel the notation $p_0, \ldots, p_m$ refers to the orthonormal polynomials from Proposition 4.2.1, where p_j is of degree j for $j = 0, \ldots, m$ due to the recursion (4.2.11).

Proposition 4.2.2 (See Section 3.3 in [Sze85]). *We recall the following well-known properties of the zeros of p_m;*

1. *The zeros $\theta_1, \ldots, \theta_m \in \mathbb{R}$ of p_m are distinct. Assume*

$$\theta_1 < \theta_2 < \ldots < \theta_m.$$

2. *The zeros of p_m and the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are interlacing. This means $\lambda_1 < \theta_1$, $\theta_m < \lambda_n$, and for $k = 1, \ldots, m-1$ there exists at least one $\lambda_{j(k)}$ with*

$$\theta_k < \lambda_{j(k)} < \theta_{k+1}.$$

The three-term recursion (4.2.11) can be represented in terms of the so-called symmetric *Jacobi matrix* J_m, whose eigenvalues coincide with the zeros of p_m: With $a_1, \ldots, a_m \in \mathbb{R}$ and $\beta_1, \ldots, \beta_m > 0$,

$$J_m = \begin{pmatrix} a_1 & \beta_1 & & & \\ \beta_1 & a_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{m-2} & a_{m-1} & \beta_{m-1} \\ & & & \beta_{m-1} & a_m \end{pmatrix} \in \mathbb{R}^{m \times m}. \tag{4.2.12}$$

Denoting $P(\lambda) = (p_0(\lambda), \ldots, p_{m-1}(\lambda))^{\mathsf{H}} \in \mathbb{C}^m$, the recursion (4.2.11) can be written in matrix form,

$$\lambda P(\lambda) = J_m\, P(\lambda) + \beta_m\, p_m(\lambda) e_m. \tag{4.2.13}$$

From (4.2.13) we observe that the zeros $\theta_1, \ldots, \theta_m$ of p_m are eigenvalues of J_m with non-normalized eigenvectors $P(\theta_j) = (p_0(\theta_j), \ldots, p_{m-1}(\theta_j))^{\mathsf{H}}$,

$$\theta_j P(\theta_j) = J_m\, P(\theta_j), \quad j = 1, \ldots, m.$$

We conclude that the matrix J_m has m distinct eigenvalues $\theta_1, \ldots, \theta_m \in \mathbb{R}$ which are indeed identical to the zeros of p_m and for which the properties from Proposition 4.2.2 hold true. We refer to $\theta_1, \ldots, \theta_m \in \mathbb{R}$ as *Ritz values.*

Polynomial Krylov subspace. We recall the usual denotation $V_m = (v_1, \ldots, v_m) \in \mathbb{C}^{n \times m}$ for the M-orthonormal basis of $\mathcal{K}_m(A, u)$ provided by the Lanczos method. We have $\operatorname{span}\{V_m\} = \mathcal{K}_m(A, u)$ and $(V_{m+1}, V_{m+1})_{\mathsf{M}} = I$ where V_{m+1} includes the subsequent basis vector v_{m+1}. The basis $\{v_1, \ldots, v_{m+1}\}$ satisfies a three-term recursion according to the Lanczos algorithm [Saa03, Section 6.6]. (The Lanczos algorithm in [Saa03, Section 6.6] relies on the Euclidean inner product but can be generalized in a direct manner.) Substituting A for λ in (4.2.11) and applying u yields a recursion for $p_0(A)u, \ldots, p_m(A)u$ which coincides with the Lanczos three-term recursion. Hence, $v_j = p_{j-1}(A)u$ for $j = 1, \ldots, m+1$ with the

orthonormal polynomials $p_0, \ldots, p_m$ from Proposition 4.2.1. Analogously to (4.2.13) the three-term recursion defining V_m can be written in matrix form,

$$A\,V_m = V_m\,J_m + \beta_m v_{m+1} e_m^{\mathsf{H}}. \tag{4.2.14}$$

We refer to $\beta_m\,v_{m+1}$ as a *residual*. With (4.2.14) and the M-orthogonality property of V_m the Jacobi matrix satisfies

$$J_m = (V_m, A\,V_m)_{\mathsf{M}}.$$

The tridiagonal structure of J_m implies $A^j\,u = \beta_0 V_m\,J_m^j\,e_1$ for $j = 0, \ldots, m-1$ and $\beta_0 V_m e_1 = u$ where $\beta_0 = \|u\|_{\mathsf{M}}$ [DK89, Saa92]. Thus,[7]

$$p(A)u = \beta_0 V_m\,p(J_m)e_1, \quad p \in \Pi_{m-1}. \tag{4.2.15a}$$

Furthermore, the corresponding deviation for a polynomial $p \in \Pi_m$ of exact degree m is in the span of the residual,

$$\beta_0 V_m\,p(J_m)e_1 - p(A)u \in \operatorname{span}\{v_{m+1}\} \perp_{\mathsf{M}} \mathcal{K}_m(A, u). \tag{4.2.15b}$$

Proposition 4.2.3. *With respect to the* M*-inner product the identity*

$$(u, p(A)u)_{\mathsf{M}} = \beta_0^2\,(e_1, p(J_m)e_1)_2, \qquad p \in \Pi_{2m-1} \tag{4.2.16}$$

holds true.

Proof. For $p \in \Pi_{2m-1}$ we can write $p = g_1\,g_2$ with $g_1 \in \Pi_{m-1}$ and $g_2 \in \Pi_m$, and

$$(u, p(A)u)_{\mathsf{M}} = (\overline{g}_1(A)u, g_2(A)u)_{\mathsf{M}} \quad \text{and} \quad (e_1, p(J_m)e_1)_2 = (\overline{g}_1(J_m)e_1, g_2(J_m)e_1)_2. \tag{4.2.17}$$

For $\overline{g}_1(A)u$ and $g_2(A)u$ we apply (4.2.15a) and (4.2.15b), respectively, to conclude

$$(\overline{g}_1(A)u, g_2(A)u)_{\mathsf{M}} = \beta_0^2 (V_m\,\overline{g}_1(J_m)e_1, V_m\,g_2(J_m)e_1)_{\mathsf{M}}. \tag{4.2.18}$$

With $(V_m, V_m)_{\mathsf{M}} = I$ we recall

$$(V_m\,\overline{g}_1(J_m)e_1, V_m\,g_2(J_m)e_1)_{\mathsf{M}} = (\overline{g}_1(J_m)e_1, g_2(J_m)e_1)_2. \tag{4.2.19}$$

Combining (4.2.17), (4.2.18) and (4.2.19) implies (4.2.16). □

Rational Krylov subspace. For rational Krylov subspaces we consider rational functions $r = p/q$ with a preassigned denominator q. The zeros of q are also referred to as the poles of r. Using the notation $s_1, s_2, \ldots \in \mathbb{C} \cup \pm\infty$ for the poles of r, for which we define

$$q_{m-1}(\lambda) = \prod_{j=1,\, s_j \neq \pm\infty}^{m-1} (\lambda - s_j). \tag{4.2.20}$$

[7] The denotation Π_j refers to the class of complex polynomials of degree $\leq j$.

Here we admit $s_j = \pm\infty$ in order to include cases for which the denominator of r is of a smaller degree than its numerator. This can be used to constitute the so called extended Krylov subspace, see also [DK98] and will further be relevant in Subsection 4.3.2 below.[8]

We assume that the poles s_j are distinct from the eigenvalues $\lambda_1, \dots, \lambda_n$ of A, such that $q_{m-1}^{-1}(A)$ is well-defined. The rational Krylov subspace $\mathcal{Q}_m(A, u)$ with poles $s_1, \dots, s_{m-1}$ and q_{m-1} from (4.2.20), is defined by the span of $\{r(A)u \colon r = p/q_{m-1} \text{ for } p \in \Pi_{m-1}\}$, i.e.,[9]

$$\begin{aligned} \mathcal{Q}_m(A, u) &:= \operatorname{span}\{q_{m-1}^{-1}(A)u, A\, q_{m-1}^{-1}(A)u, \dots, A^{m-1} q_{m-1}^{-1}(A)u\} \\ &= \mathcal{K}_m(A, q_{m-1}^{-1}(A)u). \end{aligned} \tag{4.2.21a}$$

To simplify the notation we write

$$u_q = q_{m-1}^{-1}(A)u. \tag{4.2.21b}$$

With (4.2.21), the rational Krylov subspace $\mathcal{Q}_m(A, u)$ is identical to the polynomial Krylov subspace $\mathcal{K}_m(A, u_q)$. Let w_j be the spectral coefficient of u w.r.t. the eigenvalue λ_j, then $q_{m-1}^{-1}(\lambda_j)w_j$ is the corresponding spectral coefficient of u_q. Analogously to α_n in (4.2.6a), we introduce the step function

$$\widehat{\alpha}_n(\lambda) = \begin{cases} 0, & \lambda < \lambda_1 \\ \sum_{j=1}^{\ell} |q_{m-1}^{-1}(\lambda_j)w_j|^2, & \lambda_\ell \le \lambda < \lambda_{\ell+1}, \quad \ell = 1, \dots, n-1, \\ \sum_{j=1}^{n} |q_{m-1}^{-1}(\lambda_j)w_j|^2, & \lambda_n \le \lambda. \end{cases} \tag{4.2.22}$$

Analogously to (4.2.6c), the Riemann-Stieltjes integral associated with $\widehat{\alpha}_n$ defines an inner product,

$$(f, g)_{\widehat{\alpha}_n} = \int_a^b \overline{f}\,(\lambda) g(\lambda)\, \mathrm{d}\widehat{\alpha}_n(\lambda).$$

The $\widehat{\alpha}_n$-orthonormal polynomials given by Proposition 4.2.1 constitute a basis of $\mathcal{K}_m(A, u_q)$. For the existence of these orthonormal polynomials, analogously as before we assume that we have n coefficients $w_j \neq 0$ for distinct eigenvalues λ_j, together with $0 \neq q_{m-1}^{-1}(\lambda_j) \in \mathbb{C}$.

Let J_m and V_m be the Jacobi matrix and the M-orthonormal basis for $\mathcal{K}_m(A, u_q)$. For the eigenvalues $\theta_1, \dots, \theta_m$ of J_m the results of Proposition 4.2.2 remain valid.

The Jacobi matrix $J_m = (V_m, A\,V_m)_{\mathsf{M}}$ corresponds to a representation of A in the underlying rational Krylov subspace $\mathcal{Q}_m(A, u) = \mathcal{K}_m(A, u_q)$. However, V_m and J_m are more of a theoretical nature in this context. In practice, $u_q = q_{m-1}^{-1}(A)u$, is not directly available and the rational Krylov subspace is not constructed via its polynomial counterpart, but in an iterative manner. While the Lanczos method is by far the most prominent approach to construct a polynomial Krylov subspace, various iterative algorithms are relevant for the rational case. Choosing a proper algorithm to construct a rational Krylov subspace depends on the setting, e.g., the choice of poles. For computational details we also refer to [DB07, BG15, Güt10, BR09]. Unlike the polynomial case, where V_m refers to the

[8] For $0 \neq s_1, s_2, \dots \in \mathbb{C}$ we can exchange the factors of q_{m-1}, i.e., $(\lambda - s_j)$, with $(1 - \lambda/s_j)$ to obtain a definition of q_{m-1} which is equivalent to (4.2.20). This clarifies the convention $s_j = \pm\infty$ in (4.2.20).

[9] In the sequel, we also use $q_{m-1}^{-1}(\lambda)$ for $q_{m-1}(\lambda)^{-1} = 1/q_{m-1}(\lambda)$ to shorten the denotation.

orthonormal basis constructed by the Lanczos method, we choose the notation for the rational Krylov subspace independent of the underlying algorithm: We assume $U_m \in \mathbb{C}^{n\times m}$ is a given M-orthonormal basis of $\mathcal{Q}_m(A,u)$, i.e.,

$$U_m \in \mathbb{C}^{n\times m}, \quad \operatorname{span}\{U_m\} = \mathcal{Q}_m(A,u) \quad \text{and} \quad (U_m, U_m)_{\mathsf{M}} = I,$$

and we let A_m refer to the respective *Rayleigh quotient*

$$A_m = (U_m, A\,U_m)_{\mathsf{M}} \in \mathbb{C}^{m\times m}.$$

For instance, this notation covers rational Krylov bases and representations constructed as in Subsection 4.2.1. The matrix A_m is Hermitian w.r.t. the Euclidean inner product but in general not tridiagonal and does not coincide with J_m. Let us denote

$$K_m = (V_m, U_m)_{\mathsf{M}} \in \mathbb{C}^{m\times m}. \tag{4.2.23a}$$

U_m and V_m represent orthonormal bases of the same subspace, thus,

$$U_m = V_m (V_m, U_m)_{\mathsf{M}} = V_m K_m. \tag{4.2.23b}$$

By definition of the M-inner product we have $K_m^{\mathsf{H}} K_m = U_m^{\mathsf{H}} M V_m K_m$, and together with $V_m K_m = U_m$ (4.2.23b) this yields

$$K_m^{\mathsf{H}}\, K_m = (U_m, U_m)_{\mathsf{M}} = I. \tag{4.2.23c}$$

Furthermore, A_m and J_m are orthogonally similar matrices,

$$A_m = (U_m, A\,U_m)_{\mathsf{M}} = K_m^{\mathsf{H}}\,(V_m, A\,V_m)_{\mathsf{M}}\,K_m = K_m^{\mathsf{H}}\,J_m\,K_m, \tag{4.2.24}$$

therefore, the eigenvalues of A_m are equal to the Ritz values $\theta_1, \ldots, \theta_m$ corresponding to $\mathcal{K}_m(A, u_q)$.

We proceed with some identities in the rational Krylov subspace, a rational counterpart to (4.2.15). Assume that $q_{m-1}^{-1}(A_m)$ is well-defined, and let

$$x := (U_m, u)_{\mathsf{M}} \in \mathbb{C}^m.$$

Then,

$$r(A)u = U_m\, r(A_m)x \quad \text{for } r = p/q_{m-1} \text{ with } p \in \Pi_{m-1}. \tag{4.2.25a}$$

This result was given earlier in [Güt10, Lemma 4.6] and others. Furthermore, let $r = p/q_{m-1}$ for a polynomial $p \in \Pi_m$ of degree exactly m, then[10]

$$(U_m\, r(A_m)x - r(A)u) \perp_{\mathsf{M}} \operatorname{span}\{U_m\} = \mathcal{Q}_m(A,u). \tag{4.2.25b}$$

Following [Güt13, Remark 3.2] we conclude:

Proposition 4.2.4. *For $x = (U_m, u)_{\mathsf{M}} \in \mathbb{C}^m$ and rational functions $r = p/|q_{m-1}|^2$ with $p \in \Pi_{2m-1}$,*

$$(u, r(A)u)_{\mathsf{M}} = (x, r(A_m)x)_2. \tag{4.2.26}$$

[10] A proof of (4.2.25a) and (4.2.25b) is also provided in Proposition 4.A.3, Appendix 4.A.

Proof. For $r \in \Pi_{2m-1}/|q_{m-1}|^2$ we write $r = r_1\, r_2$ with $r_1 \in \Pi_{m-1}/\overline{q}_{m-1}$ and $r_2 \in \Pi_m/q_{m-1}$, and

$$(u, r(A)u)_{\mathsf{M}} = (\overline{r}_1(A)u, r_2(A)u)_{\mathsf{M}}, \quad \text{and} \quad (x, r(A_m)x)_2 = (\overline{r}_1(A_m)x, r_2(A_m)x)_2. \tag{4.2.27}$$

For $\overline{r}_1 \in \Pi_{m-1}/q_{m-1}$ and $r_2 \in \Pi_m/q_{m-1}$ we apply (4.2.25a) and (4.2.25b), respectively, to conclude

$$(\overline{r}_1(A)u, r_2(A)u)_{\mathsf{M}} = (U_m\, \overline{r}_1(A_m)x, U_m\, r_2(A_m)x)_{\mathsf{M}}. \tag{4.2.28}$$

Combining (4.2.27) with (4.2.28) and making use of $(U_m, U_m)_{\mathsf{M}} = I$ implies (4.2.26). □

4.2.1 Some remarks on the Shift-and-Invert (SaI) Krylov subspace

The poles s_j are not required to be distinct. A prominent example is the

*Shift-and-Invert (**SaI**) Krylov subspace,*

with $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$ for a single pole $s \in \mathbb{C}$ of multiplicity $m-1$.

Remark 4.2.5. *The rational Krylov subspace $\mathcal{Q}_m(A, u)$ with a single pole $s \in \mathbb{C}$ of multiplicity $m-1$ is identical to the polynomial Krylov subspace $\mathcal{K}_m(X, u)$ with $X = (A - s\,I)^{-1}$, i.e.,*

$$\mathcal{K}_m(X, u) = \operatorname{span}\{u, (A - sI)^{-1}u, \ldots, (A - sI)^{-(m-1)}u\}.$$

Note that $\mathcal{Q}_m(A, u) \subset \mathcal{K}_m(X, u)$ via the partial fraction decomposition for rational functions with denominator $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$, and $\mathcal{K}_m(X, u) \subset \mathcal{Q}_m(A, u)$ by normalizing. Thus, the rational Krylov subspace $\mathcal{Q}_m(A, u)$ can be constructed analogously as the polynomial Krylov subspace $\mathcal{K}_m(X, u)$. The matrix X is no longer Hermitian for $\operatorname{Im} s \neq 0$, and in this case the construction of the Krylov subspace $\mathcal{K}_m(X, u)$ requires the Arnoldi method, the counterpart of the Lanczos method for general matrices. Further computational details for the case $\operatorname{Im} s \neq 0$ are given in Remark 4.2.7 below. The Lanczos or Arnoldi method for $\mathcal{K}_m(X, u)$ generates an orthonormal basis U_m and an upper Hessenberg matrix $X_m = (U_m, XU_m)_{\mathsf{M}}$. With the subsequent basis vector u_{m+1} and $x_{m+1,m} = (X_{m+1})_{m+1,m}$, the Arnoldi decomposition of $\mathcal{K}_m(X, u)$ (similar to (4.2.14)) gives

$$(A - s\,I)^{-1}U_m = U_m\, X_m + x_{m+1,m}\, u_{m+1}\, e_m^{\mathsf{H}}. \tag{4.2.29}$$

With (4.2.29) and using the notation $y_m^{\mathsf{H}} = e_m^{\mathsf{H}}\, X_m^{-1}$, we obtain

$$A\,U_m = U_m(X_m^{-1} + s\,I) - x_{m+1,m}(A - s\,I)u_{m+1}y_m^{\mathsf{H}}. \tag{4.2.30}$$

For the Rayleigh quotient $A_m = (U_m, A\,U_m)_{\mathsf{M}}$, identity (4.2.30) implies

$$A_m = X_m^{-1} + s\,I - x_{m+1,m}\,(U_m, A\,u_{m+1})_{\mathsf{M}}\, y_m^{\mathsf{H}}. \tag{4.2.31}$$

This identity can be further simplified in view of numerical efficiency and stability (similar to [DK98, eq. (5.7)] for $s = 0$ or [Gri12, eq. (5.8)] for $s \in \mathbb{R}$): With A being Hermitian and the identity (4.2.30) we have

$$(U_m, A\,u_{m+1})_{\mathsf{M}} = (A\,U_m, u_{m+1})_{\mathsf{M}} = -x_{m+1,m}\big((A\,u_{m+1}, u_{m+1})_{\mathsf{M}} - \overline{s}\big)y_m. \tag{4.2.32}$$

Algorithm 4.1: An algorithm to compute an orthonormal basis U_m and the Rayleigh quotient A_m of $\mathcal{Q}_m(A,u)$ for a single pole $s \in \mathbb{C}$ of multiplicity $m-1$, the SaI case.

$X = (A - sI)^{-1}$;
if $s \in \mathbb{R}$ apply the Lanczos method for $\mathcal{K}_{m-1}(X,u)$;
else apply the Arnoldi method for $\mathcal{K}_{m-1}(X,u)$;
in both cases this returns β_0, U_m, $X_m = (U_m, XU_m)_{\mathsf{M}}$, β_m, u_{m+1};
$\kappa = (u_{m+1}, A\,u_{m+1})_{\mathsf{M}} \in \mathbb{R}$;
$y_m^{\mathsf{H}} = e_m^{\mathsf{H}}\, X_m^{-1}$;
$A_m = (X_m^{-1} + (X_m^{-1})^{\mathsf{H}})/2 + \mathrm{Re}(s)\, I + \beta_m^2(\kappa - \mathrm{Re}(s)) y_m\, y_m^{\mathsf{H}}$;
set $x = \beta_0 e_1$;
return x, U_m, A_m;

Combining (4.2.31) *and* (4.2.32) *together with* $\kappa = (u_{m+1}, A\,u_{m+1})_{\mathsf{M}} \in \mathbb{R}$ *yields*

$$A_m = X_m^{-1} + s\,I + x_{m+1,m}^2(\kappa - \overline{s}) y_m\, y_m^{\mathsf{H}}. \tag{4.2.33}$$

With A_m *and* $y_m\, y_m^{\mathsf{H}} \in \mathbb{C}^{m\times m}$ *being Hermitian we take the Hermitian part of* (4.2.33) *to obtain*

$$A_m = (X_m^{-1} + (X_m^{-1})^{\mathsf{H}})/2 + \mathrm{Re}(s)\, I + x_{m+1,m}^2(\kappa - \mathrm{Re}(s)) y_m\, y_m^{\mathsf{H}}.$$

This representation for A_m *is equivalent to* (4.2.31) *but it is better suited for numerical computation. A shift of the inverse of the Hessenberg matrix* X_m*, i.e.,* $X_m^{-1} + s\,I$*, is closely related to the Rayleigh quotient* A_m*, see also [Güt10, Subsection 5.4.3], but it does not conserve orthogonality. E.g., for* $s \notin \mathbb{R}$ *the matrix* $X_m^{-1} + s\,I$ *is not necessarily Hermitian.*
Note that $x = \beta_0\, e_1$ *for the SaI Krylov subspace.*

The procedure which is stated in Remark 4.2.5 is summarized in Algorithm 4.1.

In some works concerning the SaI Krylov subspace, the matrix $X_m^{-1} + sI$ appears in place of the Rayleigh quotient, e.g. [vdEH06, ZTK19]; for a comparison see also [Güt10, Subsection 5.4.3]. In the following remark we show that for $s \in \mathbb{R}$ the matrix $X_m^{-1} + sI$ satisfies an identity similar to (4.2.26).

Remark 4.2.6. *Let* $X = (A - sI)^{-1}$ *for a given shift* $s \in \mathbb{R}$*. Thus,* X *is Hermitian. Then, the matrix* $X_m = (V_m, X\,V_m)_{\mathsf{M}}$ *associated with the polynomial Krylov subspace* $\mathcal{K}_m(X,u)$ *satisfies* $(u, p(X)u)_{\mathsf{M}} = (x, p(X_m)x)_2$ *for* $p \in \Pi_{2m-1}$ *due to Proposition 4.2.3. Polynomials of* X *can be rewritten as rational functions of* A*, see also Remark 4.B.1 in Appendix 4.B. A polynomial in* X_m *can be rewritten in an analogous manner: We recall* $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$ *for the given shift* $s \in \mathbb{R}$*. For a given* $p \in \Pi_{2m-2}$ *we have* $r \in \Pi_{2m-2}/q_{m-1}^2$ *with* $p(X) = r(A)$ *and* $p(X_m) = r(X_m^{-1} + sI)$*. Thus, similar to* (4.2.26) *we have the identity*

$$(u, r(A)u)_{\mathsf{M}} = (x, r(X_m^{-1} + sI)x)_2, \qquad r \in \Pi_{2m-2}/|q_{m-1}|^2. \tag{4.2.34}$$

Here, we remark $|q_{m-1}| = q_{m-1}$ *for* $s \in \mathbb{R}$*. In* (4.2.34)*, the numerator is of degree* $2m-2$ *instead of* $2m-1$ *as in* (4.2.26)*.*

We proceed with some additional remarks on the SaI Krylov subspace with a complex shift $s \in \mathbb{C} \setminus \mathbb{R}$.

Remark 4.2.7. *As stated in Remark 4.2.5, the rational Krylov subspace with a single pole $s \in \mathbb{C}$ of multiplicity $m-1$ corresponds to the polynomial Krylov subspace $\mathcal{K}_m(X, u)$ with $X = (A - s\,I)^{-1} \in \mathbb{C}^{n\times n}$. Let us consider the case $s \in \mathbb{C} \setminus \mathbb{R}$.*

In contrast to the case $s \in \mathbb{R}$, the matrix X is not Hermitian for $s \in \mathbb{C} \setminus \mathbb{R}$, and thus, the Lanczos three-term recursion fails to construct the Krylov subspace $\mathcal{K}_m(X, u)$. The Arnoldi method can be applied in this case but results in additional computational cost compared to the Lanczos method. However, to preserve some favorable properties of the Lanczos method in the case of $s \in \mathbb{C} \setminus \mathbb{R}$, we can construct the Krylov subspace by applying an isometric Arnoldi method on a transformed matrix, using a Cayley transform: We recall that $A \in \mathbb{C}^{n\times n}$ is Hermitian w.r.t. the M-inner product. Then, the matrix

$$Z = (A - \overline{s}\,I)(A - s\,I)^{-1} \in \mathbb{C}^{n\times n}$$

is unitary w.r.t. the M-inner product, i.e., $(Z\,v, Z\,w)_{\mathsf{M}} = (v, w)_{\mathsf{M}}$ for $v, w \in \mathbb{C}^n$. We introduce the notation τ for the corresponding scalar Cayley transform

$$\tau(\lambda) = (\lambda - \overline{s})(\lambda - s)^{-1}, \qquad \tau\colon \mathbb{R} \to \mathbb{T} \setminus \{1\}, \tag{4.2.35}$$

where $\mathbb{T} \subset \mathbb{C}$ denotes the unit circle. The matrix Z has eigenvalues $\tau(\lambda_j)$ and eigenvectors q_j, where λ_j and q_j denote the eigenvalues and eigenvectors of A, respectively. The function τ as given in (4.2.35) is bijective, which implies that A and Z have the same number of distinct eigenvalues with nonzero spectral coefficients $w_j = (q_j, u)_{\mathsf{M}}$. From remarks stated previously in the current section, and Proposition 4.A.1 in Appendix 4.A, we conclude that the rank of $\mathcal{Q}_m(A, u)$ and the rank of $\mathcal{K}_m(Z, u)$ are identical. For the polynomial Krylov subspace $\mathcal{K}_m(Z, u)$ we observe $\mathcal{K}_m(Z, u) \subset \mathcal{Q}_m(A, u)$ by normalizing. Due to having the same rank, the rational Krylov subspace $\mathcal{Q}_m(A, u)$ and the polynomial Krylov subspace $\mathcal{K}_m(Z, u)$ are identical.

For $\mathcal{K}_m(Z, u)$ we consider the following setting: Let V_m denote an M-orthonormal basis of the Krylov subspace $\mathcal{K}_m(Z, u)$ with $(V_m, u)_{\mathsf{M}} = \beta_0 e_1$ and an upper Hessenberg matrix $Z_m = (V_m, Z\,V_m)_{\mathsf{M}} \in \mathbb{C}^{m\times m}$, and let v_{m+1} denote the subsequent basis vector with normalization factor $z_{m+1,m} = (Z_{m+1})_{m+1,m} > 0$, $\|v_{m+1}\|_{\mathsf{M}} = 1$ and $(V_m, v_{m+1})_{\mathsf{M}} = 0$, such that

$$Z\,V_m = V_m Z_m + z_{m+1,m}\, e_m^{\mathsf{H}} v_{m+1}. \tag{4.2.36}$$

Such a representation can be generated by a short term Arnoldi method, e.g., the isometric Arnoldi method [JR94, Algorithm 3.1, eq. (3.4) and (3.5)] introduced in [Gra93, JR94]. For further details we also refer to [BGF97, Sch08, BMV18]. We also recapitulate the isometric Arnoldi method in Algorithm 4.2. In contrast to the standard Arnoldi method, the isometric Arnoldi method is more efficient in terms of computational cost, comparable to the Lanczos algorithm for Hermitian matrices.

Let the decomposition (4.2.36) be given and set $U_m := V_m$, then U_m conforms to an orthonormal basis of the rational Krylov subspace $\mathcal{Q}_m(A, u)$ with denominator $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$, and $x = \beta_0 e_1$. Substituting Z and V_m in (4.2.36) yields

$$(A - \overline{s}\,I)(A - s\,I)^{-1} U_m = U_m Z_m + z_{m+1,m}\, e_m^{\mathsf{H}} v_{m+1}.$$

Algorithm 4.2: An isometric Arnoldi method to compute an orthonormal basis U_m and the $Z_m = (U_m, Z\,U_m)_{\mathsf{M}}$ of $\mathcal{K}_m(Z,u)$ for a unitary matrix Z, e.g., a Cayley transform $Z = (A - \overline{s}\,I)(A - s\,I)^{-1}$ where A is an Hermitian matrix and $s \in \mathbb{C} \setminus \mathbb{R}$. See Remark 4.2.7 and references therein.

$\beta_0 = \|u\|_{\mathsf{M}}, \quad v_1 = u/\beta_0, \quad \widehat{v} = v_1, \quad Z_m = I_{m\times m};$
for $k = 1 : m;$
  $w = Z v_k;$
  $\gamma = -(\widehat{v}, w)_{\mathsf{M}};$
  $v_{\text{next}} = w + \gamma\widehat{v};$
  $\sigma = \|v_{\text{next}}\|_{\mathsf{M}}; \quad$ // $= (1 - |\gamma|^2)^{1/2}$ in exact arithmetic;
  $v_{k+1} = v_{\text{next}}/\sigma;$
  if $k < m;$
    $(Z_m)_{:,[k:k+1]} \leftarrow (Z_m)_{:,[k:k+1]} \cdot \begin{pmatrix} -\gamma & \sigma \\ \sigma & \overline{\gamma} \end{pmatrix};$
    $\widehat{v} \leftarrow \sigma\widehat{v} + \overline{\gamma} v_{k+1};$
    $\widehat{v} \leftarrow \widehat{v}/\|\widehat{v}\|_{\mathsf{M}};$ // not required in exact arithmetic;
  else // $k = m;$
    $(Z_m)_{:,k} \leftarrow -\gamma (Z_m)_{:,k};$
    $z_{m+1,m} = \sigma;$
return $\beta_0, U_m = (v_1, \ldots, v_m), v_{m+1}, Z_m, z_{m+1,m};$

Similar to Remark 4.2.5, this provides a computable formulation for the Rayleigh quotient $A_m = (U_m, AU_m)_{\mathsf{M}}$. *With* $\kappa = (v_{m+1}, A\,v_{m+1})_{\mathsf{M}} \in \mathbb{R}$ *and* $y_m^{\mathsf{H}} := e_m^{\mathsf{H}}(I - Z_m)^{-1}$ *we have*

$$A_m = (\overline{s}I - sZ_m)(I - Z_m)^{-1} + z_{m+1,m}^2\big(\kappa - \overline{s}\big)y_m y_m^{\mathsf{H}}.$$

This procedure is summarized in Algorithm 4.3.

As an alternative approach to compute the SaI Krylov subspace with $s \in \mathbb{C} \setminus \mathbb{R}$, *we also remark that the matrix* $X = (A - s\,I)^{-1}$ *is in the class of so called* normal$(1,1)$ *matrices (cf. [BM00]), i.e., the* M*-adjoint of* X *corresponds to a rational function* $p(X)q(X)^{-1}$ *with* $p, q \in \Pi_1$, *namely,*

$$X^* = (A - \overline{s}\,I)^{-1} = (X^{-1} + (s - \overline{s})\,I)^{-1} = X\,(I + (s - \overline{s})\,X)^{-1},$$

due to $X^{-1} = A - s\,I$. *For* normal$(1,1)$ *matrices a short Arnoldi recurrence exists, see [BM00, BMV18], but we do not further discuss this approach in the current work.*

4.3 A review on quasi-orthogonal polynomials

The theory of quasi-orthogonal polynomials is for instance covered in [Sze85, Akh65, GM10]. We will refer to a special linear combination $\widehat{p}_m$ of p_{m-1} and p_m as a *quasi-orthogonal polynomial* of degree m, where $p_0, \ldots, p_m$ denote the orthonormal polynomials

Algorithm 4.3: An optimized algorithm to compute an orthonormal basis U_m and the Rayleigh quotient A_m of $\mathcal{Q}_m(A,u)$ for a single pole $s \in \mathbb{C}\setminus\mathbb{R}$ of multiplicity $m-1$, see Remark 4.2.7.

$Z = (A - \overline{s}\, I)(A - s\, I)^{-1}$;
apply the isometric Arnoldi method for $\mathcal{K}_m(Z,u)$, see Algorithm 4.2;
this returns β_0, U_m, $Z_m = (U_m, ZU_m)_{\mathsf{M}}$, $z_{m+1,m}$, u_{m+1};
$\kappa = (u_{m+1}, A\, u_{m+1})_{\mathsf{M}} \in \mathbb{R}$;
$y_m^{\mathsf{H}} = e_m^{\mathsf{H}}\,(I - Z_m)^{-1}$;
$A_m = (\overline{s}I - sZ_m)(I - Z_m)^{-1} + z_{m+1,m}^2\big(\kappa - \overline{s}\big)y_m y_m^{\mathsf{H}}$;
set $x = \beta_0 e_1$;
return x, U_m, A_m;

from the previous section. In the class of quasi-orthogonal polynomials we impose an additional condition, i.e., we require that

the quasi-orthogonal polynomial $\widehat{p}_m$ vanishes at a given $\xi \in \mathbb{R}$, i.e., $\widehat{p}_m(\xi) = 0$. (4.3.1)

Quasi-orthogonal polynomials also appear in the theory of Gauss-Radau quadrature formulae. Similar to the three-term recursion of the orthogonal polynomials, the underlying recursion of the polynomials $p_1, \ldots, p_{m-1}$ and $\widehat{p}_m$ constitutes a matrix T_m which coincides with the Jacobi matrix J_m up to one entry. It was already shown in [Wil62, GW69], that T_m provides quadrature nodes and weights of Gauss-Radau quadrature formulae associated with the underlying distribution (i.e., $\mathrm{d}\alpha_n$ in the present setting). In the context of Gauss-Radau quadrature formulae, the preassigned zero ξ corresponds to a preassigned quadrature node, see also [Gau04, GM10].

At the beginning of the present section we recall some theory on quasi-orthogonal polynomials. In Subsection 4.3.1 this theory will be applied to the polynomial Krylov subspace $\mathcal{K}_m(A,u)$. While keeping the orthonormal basis V_m of $\mathcal{K}_m(A,u)$ as before, we consider the modified matrix T_m (given by the underlying recursion; see (4.3.4) below) as a representation of A in $\mathcal{K}_m(A,u)$. This results in the matrix decomposition (4.3.6), where $\widehat{p}_m$ provides the residual. Thus, we also refer to T_m as a

*quasi-orthogonal residual (***qor-***) Krylov representation.*

The zero $\xi \in \mathbb{R}$ of $\widehat{p}_m$ which is preassigned constitutes an eigenvalue of the modified matrix T_m. The spectrum of T_m constitutes a step function α_m which is introduced properly in Section 4.4 below. Based on the CMS Theorem, the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ satisfy some intertwining property (in general, this result is known for the Gauss-Radau quadrature rule; in the Krylov setting we specify this result in Section 4.4 below). In the qor-Krylov setting, we can make use of the preassigned zero ξ to modify computable bounds on α_n, which potentially result in refined bounds. Furthermore, we consider the matrix T_m to approximate a matrix function $f(A)u$. This is referred to as qor-Krylov approximation, see (4.3.10) below. The qor-Krylov approximation can be understood as a corrected Krylov approximation, comparable to the corrected Krylov scheme for the matrix exponential in [Saa92]. In

the context of approximating matrix functions, making use of quasi-orthogonal polynomials is a new idea.

Later on in this section the theory of quasi-orthogonal polynomials will be applied to the case of a rational Krylov subspace. We also refer to [Gau04] for rational Gauss-Radau quadrature formulae, which are also applied in a Krylov setting in [LLRW08, JR13]. As in the polynomial case, we aim to refine estimates on α_n in the sequel, and we also introduce a rational qor-Krylov approximation. In Remark 4.3.8 below, we introduce a new procedure to efficiently compute the rational qor-Krylov representation, i.e., we rewrite a rational Krylov subspace with arbitrary complex poles as an extended Krylov subspace with a modified initial vector, and then construct the rational qor-Krylov representation based on results for the polynomial case.

We proceed to recall some theory on quasi-orthogonal polynomials. Let $p_0, \ldots, p_m$ be the sequence of orthonormal polynomials from Proposition 4.2.1. Let $\beta_{m-1} > 0$ be given as in Proposition 4.2.1, and let $\omega_m \in \mathbb{R}$ (to be fixed in the sequel, see (4.3.3b)). We define

$$\widehat{p}_m(\lambda) = (\lambda - \omega_m)p_{m-1}(\lambda) - \beta_{m-1}p_{m-2}(\lambda). \tag{4.3.2a}$$

The polynomial p_m satisfies the recursive identity (4.2.11) (for $j = m$). Thus, $\widehat{p}_m$ can be expressed as a linear combination of p_{m-1} and p_m,

$$\widehat{p}_m = \beta_m\, p_m + (a_m - \omega_m)p_{m-1}, \quad \text{hence,} \quad \widehat{p}_m \perp p_0, \ldots, p_{m-2}. \tag{4.3.2b}$$

With the orthogonality property (4.3.2b) we refer to $\widehat{p}_m$ as quasi-orthogonal polynomial of degree m.[11]

According to the requirement $\widehat{p}_m(\xi) = 0$ imposed above (see (4.3.1)) for a given $\xi \in \mathbb{R}$ with $p_{m-1}(\xi) \neq 0$, definition (4.3.2a) implies

$$0 = \widehat{p}_m(\xi) = (\xi - \omega_m)p_{m-1}(\xi) - \beta_{m-1}p_{m-2}(\xi). \tag{4.3.3a}$$

This fixes the value of ω_m,

$$\omega_m = \xi - \beta_{m-1}\frac{p_{m-2}(\xi)}{p_{m-1}(\xi)}. \tag{4.3.3b}$$

We now reuse the denotation $\theta_1, \ldots, \theta_m$ in a modified way: In the context of quasi-orthogonal polynomials, $\theta_1, \ldots, \theta_m \in \mathbb{R}$ denote the zeros of $\widehat{p}_m$. We assume the ordering $\theta_1 < \theta_2 < \ldots < \theta_m$.

Proposition 4.3.1 (See also Section 3.3 in [Sze85])**.** *Let $\widehat{p}_m$ be the quasi-orthogonal polynomial defined in* (4.3.2a)*, with ω_m from* (4.3.3b) *for a given $\xi \in \mathbb{R}$ with $p_{m-1}(\xi) \neq 0$.*

(i) The zeros $\theta_1, \ldots, \theta_m$ of $\widehat{p}_m$ are distinct.

(ii) Interlacing property of eigenvalues $\lambda_1, \ldots, \lambda_n$ and zeros of $\widehat{p}_m$: For $k = 1, \ldots, m-1$ there exists at least one $\lambda_{j(k)}$ with

$$\theta_k < \lambda_{j(k)} < \theta_{k+1}.$$

[11] In the case $a_m = \omega_m$ the polynomial $\widehat{p}_m$ in (4.3.2a) is identical to $\beta_m p_m$, thus, $\widehat{p}_m$ is an orthogonal polynomial.

(iii) At most one of the zeros $\theta_1, \ldots, \theta_m$ is located outside of $[\lambda_1, \lambda_n]$. E.g., in the case $\xi < \lambda_1$ we have $\theta_1 < \lambda_1 < \theta_2 < \ldots < \theta_m < \lambda_n$.

As a slight modification of the Jacobi matrix J_m from (4.2.12) we now define the symmetric tridiagonal matrix

$$T_m = \begin{pmatrix} a_1 & \beta_1 & & & \\ \beta_1 & a_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{m-2} & a_{m-1} & \beta_{m-1} \\ & & & \beta_{m-1} & \boldsymbol{\omega_m} \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad \text{with } \omega_m \text{ from (4.3.3b).} \tag{4.3.4}$$

With the recursion (4.2.11) and identity (4.3.2b) the sequence of orthonormal polynomials $P(\lambda) = (p_0(\lambda), \ldots, p_{m-1}(\lambda))^{\mathsf{H}} \in \mathbb{R}^m$ and $\widehat{p}_m$ satisfy

$$\lambda P(\lambda) = T_m\, P(\lambda) + \widehat{p}_m(\lambda) e_m. \tag{4.3.5}$$

Thus, the eigenvalues of T_m are exactly the zeros $\theta_1, \ldots, \theta_m$ of $\widehat{p}_m$.

4.3.1 Krylov methods and quasi-orthogonal polynomials

Let $p_0, \ldots, p_m$ be the orthonormal polynomials from Proposition 4.2.1, which provide the M-orthonormal Krylov basis vectors $v_j = p_{j-1}(A)u$ for $j = 1, \ldots, m+1$, and let $\widehat{v}_{m+1} = \widehat{p}_m(A)u$ with the quasi-orthogonal polynomial $\widehat{p}_m$ from (4.3.2). Analogously to (4.3.5) we have the matrix decomposition

$$A\, V_m = V_m\, T_m + \widehat{v}_{m+1}\, e_m^{\mathsf{H}}. \tag{4.3.6}$$

We refer to $\widehat{v}_{m+1} \in \mathbb{C}^n$ as the residual of (4.3.6), with

$$\widehat{v}_{m+1} \in \operatorname{span}\{v_m, v_{m+1}\} \perp_{\mathsf{M}} \mathcal{K}_{m-1}(A, u).$$

Proposition 4.3.2. *For $p \in \Pi_{m-1}$,*

$$\beta_0 V_m\, p(T_m) e_1 = p(A)u. \tag{4.3.7}$$

Proof. We prove $\beta_0 V_m T_m^j\, e_1 = A^j\, u$ for $j = 0, \ldots, m-1$ by induction. This holds true for $j = 0$. Assuming that it also holds true for some $j < m-1$, then

$$A^{j+1}u = A\, A^j u = \beta_0\, A\, V_m T_m^j\, e_1.$$

Together with identity (4.3.6) this gives

$$A^{j+1}u = \beta_0 V_m T_m^{j+1} e_1 + \beta_0 \widehat{v}_{m+1} e_m^{\mathsf{H}} T_m^j\, e_1.$$

Due to the tridiagonal structure of T_m we have $e_m^{\mathsf{H}}\, T_m^j\, e_1 = 0$ for $j = 0, \ldots, m-2$. Altogether, this implies $\beta_0 V_m\, T_m^j\, e_1 = A^j\, u$ for $j = 0, \ldots, m-1$, which completes the proof of (4.3.7). □

In addition to Proposition 4.3.2 we note that for $p \in \Pi_m$ exactly of degree m,

$$\beta_0 V_m\, p(T_m)e_1 - p(A)u \in \operatorname{span}\{v_m, v_{m+1}\} \perp_{\mathsf{M}} \mathcal{K}_{m-1}(A,u).$$

The following proposition is associated with identities of Gauss-Radau quadrature formulae, see also [GM10] or [Gau04, Subsection 3.1.4]. This relation is discussed in more detail in Section 4.4 below.

Proposition 4.3.3. *For $p \in \Pi_{2m-2}$,*

$$(u, p(A)u)_{\mathsf{M}} = \beta_0^2 (e_1, p(T_m)e_1)_2. \tag{4.3.8}$$

Proof. We write $p = g_1\, g_2$ with $g_1, g_2 \in \Pi_{m-1}$ and apply Proposition 4.3.2 to both terms,

$$(u, p(A)u)_{\mathsf{M}} = (\overline{g}_1(A)u, g_2(A)u)_{\mathsf{M}} = \beta_0^2 (V_m\, \overline{g}_1(T_m)e_1, V_m\, g_2(T_m)e_1)_{\mathsf{M}}.$$

With $(V_m, V_m)_{\mathsf{M}} = I$ this implies (4.3.8). □

We proceed by recapitulating results from [GM10, Subsection 6.2.1] and [Gol73, Section 7] which reveal an algorithm to construct T_m.

Remark 4.3.4 ([GM10, Gol73])**.** *Let J_{m-1} be the Jacobi matrix constructed by $m-1$ steps of the Lanczos method. After substituting ξ for λ in* (4.2.13)*, the Jacobi matrix J_{m-1} and $P(\xi) = (p_0(\xi), \ldots, p_{m-2}(\xi))^{\mathsf{H}} \in \mathbb{R}^{m-1}$ satisfy*

$$(J_{m-1} - \xi I)P(\xi) = -\beta_{m-1}\, p_{m-1}(\xi) e_{m-1}.$$

The solution $\delta = (\delta_1, \ldots, \delta_{m-1}) \in \mathbb{R}^{m-1}$ of the linear system

$$(J_{m-1} - \xi I)\delta = \beta_{m-1}^2 e_{m-1} \tag{4.3.9}$$

is given by

$$\delta_\ell = -\beta_{m-1} \frac{p_{\ell-1}(\xi)}{p_{m-1}(\xi)}, \quad \ell = 1, \ldots, m-1.$$

The eigenvalues of J_{m-1} are identical to the zeros of p_{m-1}, hence, with $p_{m-1}(\xi) \neq 0$ the matrix $(J_{m-1} - \xi I)$ is invertible. The solution $\delta \in \mathbb{R}^{m-1}$ of (4.3.9) *yields a computable formula for ω_m via* (4.3.3b)*,*

$$\omega_m = \xi + \delta_{m-1}.$$

Algorithm 4.4 represents a summary on Remark 4.3.4. In Figure 4.1 we show values of ω_m over ξ for a given example.

A quasi-orthogonal residual (qor-)Krylov approximation to matrix functions $f(A)u$. We refer to

$$\beta_0 V_m f(T_m)e_1 \approx f(A)u \tag{4.3.10}$$

as quasi-orthogonal residual (qor-)Krylov approximation, based on the construction of V_m and T_m according to Algorithm 4.4. We recall that only $m-1$ steps of the Lanczos iteration are required. This provides the orthonormal basis V_{m-1}, the subsequent basis vector v_m and

Algorithm 4.4: An algorithm to compute V_m and the qor-Krylov representation T_m for a given $\xi \in \mathbb{R}$ which is distinct to the eigenvalues of J_{m-1}.

apply the Lanczos method for $\mathcal{K}_{m-1}(A, u)$: this returns β_0, $V_{m-1}, J_{m-1}, \beta_{m-1}, v_m$;
set $\omega_m = \xi + \beta_{m-1}^2 e_{m-1}^{\mathsf{H}} (J_{m-1} - \xi I)^{-1} e_{m-1}$ and define T_m via (4.3.4);
set $V_m = (V_{m-1}, v_m)$;
return β_0, V_m, T_m;

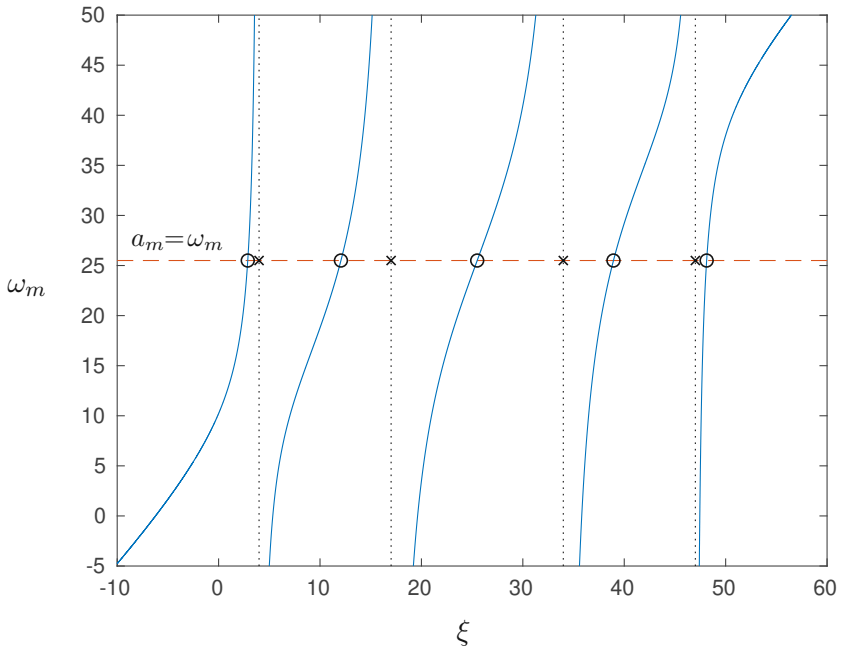


Figure 4.1: This figure shows the matrix entry ω_m of the qor-Krylov representation T_m, computed for different values of $\xi \in \mathbb{R}$ and $m = 5$. To compute the entries ω_m we follow Algorithm 4.4. As an example we choose A to be a $n \times n$ diagonal matrix with $n = 50$ and diagonal entries $(1, \ldots, n)$, and we choose $u = (1, \ldots, 1)^{\mathsf{H}} \in \mathbb{R}^n$. When the choice of ξ matches one of the eigenvalues of J_m (marked by ('$\circ$')), then the matrices T_m and J_m coincide (the matrix entry a_m of J_m is illustrated by the dashed horizontal line). On the other hand, when ξ coincides with an eigenvalue of J_{m-1} (marked by ('$\times$') and dotted vertical lines), then ω_m is undefined and Algorithm 4.4 fails. We remark that two neighboring eigenvalues of J_m enclose exactly one eigenvalue of J_{m-1}, cf. [Sze85, Theorem 3.3.2]. This property carries over to the eigenvalues of T_m via (4.3.2b) (indeed, the sign of $\widehat{p}_m$ corresponds to the sign of p_m at the zeros of p_{m-1} and at the boundary of $\mathbb{R}$). Thus, two neighboring eigenvalues of T_m enclose exactly one eigenvalue of J_{m-1} for any valid choice of ξ.

the Jacobi matrix J_{m-1}. The qor-Krylov approximation makes use of the orthonormal basis $V_m = (V_{m-1}, v_m)$, where the polynomial Krylov approximation, i.e, $\beta_0 V_{m-1} f(J_{m-1}) e_1 \approx f(A)u$, provides an approximation in the basis V_{m-1}.

The idea to 'correct' the Krylov approximation by including the subsequent basis vector (which is v_m at the $(m-1)$-th step) also appears in [Saa92], namely, the corrected Krylov scheme for the matrix exponential which is widely used in the Expokit package [Sid98] and others. Compared to the corrected Krylov scheme, the qor-Krylov approximation can be favorable if spectral properties of $f(A)$ are relevant, e.g., the mass conservation of $\mathrm{e}^{-\mathrm{i}tA}u$ carries over to the qor-Krylov approximation $\beta_0 V_m \mathrm{e}^{-\mathrm{i}tT_m} e_1$ due to T_m being Hermitian.

4.3.2 Rational Krylov methods and the theory of quasi-orthogonal polynomials

A rational Krylov subspace satisfies $\mathcal{Q}_m(A,u) = \mathcal{K}_m(A,u_q)$ for $u_q = q_{m-1}^{-1}(A)u$, where q_{m-1} denotes the denominator given by preassigned poles. Let J_m and V_m denote the Jacobi matrix and M-orthonormal basis of $\mathcal{K}_m(A,u_q)$. The procedure of Subsection 4.3.1 applies to the polynomial Krylov subspace $\mathcal{K}_m(A,u_q)$: For a given $\xi \in \mathbb{R}$ the matrix T_m is defined in (4.3.4) and satisfies the matrix decomposition (4.3.6) together with A and V_m.

In a practical setting, $u_q = q_{m-1}^{-1}(A)u$ is not directly available to construct the rational Krylov subspace via $\mathcal{K}_m(A,u_q)$. We proceed to generalize the qor-Krylov representation for the rational Krylov subspace: Let U_m be a given M-orthonormal basis of $\mathcal{Q}_m(A,u)$, i.e., $\operatorname{span}\{U_m\} = \mathcal{Q}_m(A,u)$ and $(U_m, U_m)_{\mathsf{M}} = I$. The respective Rayleigh quotient is $A_m = (U_m, A\,U_m)_{\mathsf{M}}$. With the orthonormal transformation $K_m = (V_m, U_m)_{\mathsf{M}} \in \mathbb{C}^{m\times m}$ we have $A_m = K_m^{\mathsf{H}} J_m K_m$ as given in (4.2.24). For a representation of T_m in the basis U_m we introduce the notation

$$B_m = K_m^{\mathsf{H}} T_m K_m. \tag{4.3.11}$$

The eigenvalues of B_m are equal to the eigenvalues $\theta_1, \ldots, \theta_m \in \mathbb{R}$ of T_m and satisfy Proposition 4.3.1. The Hermitian structure of T_m carries over to B_m.

Proposition 4.3.5. *With $x = (U_m, u)_{\mathsf{M}}$ we have*

$$r(A)u = U_m\, r(B_m)x \quad \textit{for} \;\; r \in \Pi_{m-1}/q_{m-1}. \tag{4.3.12}$$

Proof. Let $\zeta_0 = \|u_q\|_{\mathsf{M}}$, let V_m be the M-orthonormal basis of $\mathcal{K}_m(A,u_q)$, and let T_m be the respective qor-Krylov representation for a given $\xi \in \mathbb{R}$. Then Proposition 4.3.2 w.r.t. $\mathcal{K}_m(A,u_q)$ implies

$$p(A)u_q = \zeta_0\, V_m\, p(T_m)e_1, \quad p \in \Pi_{m-1}. \tag{4.3.13}$$

This implies $q_{m-1}(A)u_q = \zeta_0 V_m\, q_{m-1}(T_m)e_1$, and with the identities $q_{m-1}(A)u_q = u$ and $(V_m, V_m)_{\mathsf{M}} = I$ we arrive at

$$\zeta_0\, e_1 = q_{m-1}^{-1}(T_m)(V_m, u)_{\mathsf{M}}. \tag{4.3.14}$$

Let $r = p/q_{m-1}$ for $p \in \Pi_{m-1}$ then $r(A)u = p(A)u_q$, and with (4.3.13) we have

$$r(A)u = \zeta_0\, V_m\, p(T_m)e_1. \tag{4.3.15}$$

Inserting (4.3.14) into (4.3.15) gives

$$r(A)u = V_m\, p(T_m)q_{m-1}^{-1}(T_m)(V_m,u)_{\mathsf{M}} = V_m\, r(T_m)(V_m,u)_{\mathsf{M}}. \tag{4.3.16}$$

With $K_m\, K_m^{\mathsf{H}} = I$ (see (4.2.23c)) the matrix B_m in (4.3.11) satisfies $r(T_m) = K_m r(B_m) K_m^{\mathsf{H}}$, and together with $V_m K_m = U_m$ (4.2.23c) we have

$$V_m\, r(T_m)(V_m,u)_{\mathsf{M}} = U_m\, r(B_m)(U_m,u)_{\mathsf{M}}. \tag{4.3.17}$$

Combining (4.3.16) with (4.3.17) results in (4.3.12). □

The following proposition is associated with identities of rational Gauss-Radau quadrature formulae, see also [Gau04, § 3.1.4.4]. For more details on this relation see Section 4.4 below.

Proposition 4.3.6. *With* $x = (U_m,u)_{\mathsf{M}}$,

$$(u, r(A)u)_{\mathsf{M}} = (x, r(B_m)x)_2 \quad \textit{for} \quad r \in \Pi_{2m-2}/|q_{m-1}|^2. \tag{4.3.18}$$

Proof. For rational functions $r \in \Pi_{2m-2}/|q_{m-1}|^2$ we write $r = r_1\, r_2$, where $r_1 \in \Pi_{m-1}/\overline{q}_{m-1}$ and $r_2 \in \Pi_{m-1}/q_{m-1}$. With this notation we write

$$(u, r(A)u)_{\mathsf{M}} = (\overline{r}_1(A)u, r_2(A)u)_{\mathsf{M}}, \quad \text{and} \quad (x, r(B_m)x)_2 = (\overline{r}_1(B_m)x, r_2(B_m)x)_2. \tag{4.3.19}$$

For $r_1,\, r_2 \in \Pi_{m-1}/q_{m-1}$ we apply Proposition 4.3.5 to conclude

$$(\overline{r}_1(A)u, r_2(A)u)_{\mathsf{M}} = (U_m\, \overline{r}_1(B_m)x, U_m\, r_2(B_m)x)_{\mathsf{M}}. \tag{4.3.20}$$

Combining (4.3.19) with (4.3.20) and making use of $(U_m, U_m)_{\mathsf{M}} = I$ we conclude (4.3.18). □

The definition of B_m in (4.3.11) is of a theoretical nature. We propose a setup in which B_m can be computed efficiently. Let $\mathcal{Q}_{m-2}(A,u)$ be a rational Krylov subspace with arbitrary poles $s_1, \ldots, s_{m-3} \in \mathbb{C} \cup \{\pm\infty\}$. The poles $s_1, \ldots, s_{m-3}$ define the denominator q_{m-3} and we write $u_q = q_{m-3}^{-1}(A)u$. We also recall the identities

$$\mathcal{Q}_{m-2}(A,u) = \mathcal{K}_{m-2}(A,u_q) \quad \text{and} \quad \mathcal{K}_m(A,u_q) = \mathcal{K}_{m-2}(A,u_q) \oplus \operatorname{span}\{A\,u, A^2u\}.$$

We extend the rational Krylov subspace $\mathcal{Q}_{m-2}(A,u)$ by two additional polynomial Krylov steps, i.e.,

$$\mathcal{K}_m(A,u_q) = \mathcal{Q}_{m-2}(A,u) \,\oplus\, \operatorname{span}\{A\,u, A^2u\}, \quad \text{where} \quad u_q = q_{m-3}^{-1}(A)u. \tag{4.3.21}$$

The Krylov subspace $\mathcal{K}_m(A,u_q)$ can be referred to as an extended Krylov subspace, and some of the following results are related to [DK98, Section 5].

Proposition 4.3.7. *Let* m *be fixed and* $u_q = q_{m-3}^{-1}(A)u$ *for a given denominator* q_{m-3}. *Let* $U_{m-2} \in \mathbb{C}^{n\times m-2}$ *be a given* M*-orthonormal basis of* $\mathcal{Q}_{m-2}(A,u) = \mathcal{K}_{m-2}(A,u_q)$, *and* $A_{m-2} = (U_{m-2}, A\,U_{m-2})_{\mathsf{M}}$. *Let* $\mathcal{K}_m(A,u_q)$ *refer to the extended Krylov subspace given in* (4.3.21). *Let* $V_m = (v_1,\ldots,v_m) \in \mathbb{C}^{n\times m}$ *be the* M*-orthonormal basis of* $\mathcal{K}_m(A,u_q)$ *provided by the Lanczos method. Then the following statements hold true and provide a procedure to compute the qor-Krylov representation of* B_m *for a given* $\xi \in \mathbb{R}$ *and the basis* $\widetilde{U}_m$ *(given below) of the extended Krylov subspace* $\mathcal{K}_m(A,u_q)$.

(i) With $\widetilde{U}_m = (U_{m-2}, v_{m-1}, v_m) \in \mathbb{C}^{n\times m}$ *we have an* M*-orthonormal basis of the extended Krylov subspace* $\mathcal{K}_m(A, u_q)$*, i.e.,* $\operatorname{span}\{\widetilde{U}_m\} = \mathcal{K}_m(A, u_q)$ *and* $(\widetilde{U}_m, \widetilde{U}_m)_{\mathsf{M}} = I$*. Furthermore,* $\widetilde{U}_m$ *can be computed without reference to* u_q*.*

(ii) The Rayleigh quotient $\widetilde{A}_m = (\widetilde{U}_m, A\,\widetilde{U}_m)_{\mathsf{M}}$ *of the extended Krylov subspace is given by*

$$\widetilde{A}_m = \begin{pmatrix} \widetilde{A}_{m-1} & \beta_{m-1}e_{m-1} \\ \beta_{m-1}e_{m-1}^{\mathsf{H}} & a_m \end{pmatrix} \in \mathbb{C}^{m\times m}, \quad \text{with} \tag{4.3.22}$$

$$\widetilde{A}_{m-1} = \begin{pmatrix} A_{m-2} & \widetilde{a} \\ \widetilde{a}^{\mathsf{H}} & a_{m-1} \end{pmatrix} \in \mathbb{C}^{m-1\times m-1}, \quad \text{and} \quad \widetilde{a} = (U_{m-2}, A\,v_{m-1})_{\mathsf{M}} \in \mathbb{C}^{m-2}.$$

Furthermore, $a_m = (J_m)_{m,m}$*,* $a_{m-1} = (J_m)_{m-1,m-1}$ *and* $\beta_{m-1} = (J_m)_{m,m-1}$ *for the Jacobi matrix* J_m *of* $\mathcal{K}_m(A, u_q)$*. The matrix entries* a_m*,* a_{m-1}*,* β_{m-1}*, and* $\widetilde{a}$ *are computed in course of the orthogonalization procedure in (i).*

(iii) For the basis transformation $\widetilde{K}_m = (V_m, \widetilde{U}_m)_{\mathsf{M}}$ *we have*

$$\widetilde{K}_m = \begin{pmatrix} K_{m-2} & 0 \\ 0 & I_2 \end{pmatrix}, \quad \text{with} \quad I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{4.3.23}$$

and $K_{m-2} = (V_{m-2}, U_{m-2})_{\mathsf{M}}$*.*

(iv) Let T_m *be defined by* (4.3.4) *for* $\mathcal{K}_m(A, u_q)$*. Then,* $\widetilde{B}_m = \widetilde{K}_m^{\mathsf{H}}\, T_m\, \widetilde{K}_m$ *satisfies*

$$\widetilde{B}_m = \begin{pmatrix} \widetilde{A}_{m-1} & \beta_{m-1}\, e_{m-1} \\ \beta_{m-1}\, e_{m-1}^{\mathsf{H}} & \omega_m \end{pmatrix}, \tag{4.3.24a}$$

with

$$\omega_m = \xi + \beta_{m-1}^2 e_{m-1}^{\mathsf{H}} (\widetilde{A}_{m-1} - \xi I)^{-1} e_{m-1}. \tag{4.3.24b}$$

Proof.

(i) We have $\operatorname{span}\{U_{m-2}\} = \mathcal{K}_{m-2}(A, u_q) = \operatorname{span}\{V_{m-2}\}$, and by adding v_{m-1} and v_m to the basis we have $\operatorname{span}\{\widetilde{U}_m\} = \mathcal{K}_m(A, u_q)$. With $(U_{m-2}, U_{m-2})_{\mathsf{M}} = I$ and $v_{m-1}, v_m \perp_{\mathsf{M}} \mathcal{K}_{m-2}(A, u_q)$ this also implies $(\widetilde{U}_m, \widetilde{U}_m)_{\mathsf{M}} = I$.

The M-orthonormal basis $\widetilde{U}_m$ can be constructed without referring to u_q by the following procedure: We construct v_{m-1} by orthogonalizing $\widetilde{v} = A\,u$ w.r.t. U_{m-2} and normalizing. In a similar manner we construct v_m via Av_{m-1}. To demonstrate that this procedure yields the correct results, we argue as follows: We recall $u_q = q_{m-3}^{-1}(A)u$, hence, $u = q_{m-3}(A)u_q$. We introduce the notation $\widetilde{p}(\lambda) = \lambda\, q_{m-3}(\lambda)$, where $\lambda\, q_{m-3}(\lambda) = \lambda^{m-2} + \widetilde{p}_{m-3}(\lambda)$ for a polynomial $\widetilde{p}_{m-3} \in \Pi_{m-3}$. Let $\widetilde{v} = A\,u$, then $\widetilde{v} = \widetilde{p}(A)u_q$. We recall $v_j = p_{j-1}(A)u_q$ for the orthonormal polynomials $p_0, \ldots, p_m$ provided by Proposition 4.2.1. The polynomials $\widetilde{p}$ and p_{m-2} both have a positive real-valued leading coefficient. Hence, we obtain p_{m-2} by orthogonalizing $\widetilde{p}$ w.r.t. $p_0, \ldots, p_{m-3}$ and normalizing. Analogously, we obtain v_{m-1} by orthogonalizing $\widetilde{v} = A\,u$ w.r.t. U_{m-2} and normalizing as stated above.

(ii) In order to specify $\widetilde{A}_m = (\widetilde{U}_m, A\,\widetilde{U}_m)_{\mathsf{M}}$ we recall $\widetilde{U}_m = (U_{m-2}, v_{m-1}, v_m)$. The upper left submatrix of $\widetilde{A}_m$ is given by $A_{m-2} = (U_{m-2}, A\,U_{m-2})_{\mathsf{M}}$. In a similar manner we deduce $\widetilde{a}$, a_m, a_{m-1} and β_{m-1}. Here $a_m = (v_m, A\,v_m)_{\mathsf{M}}$ by the structure of $\widetilde{U}_m$ and with $J_m = (V_m, A\,V_m)_{\mathsf{M}}$ we also have $a_m = (J_m)_{m,m}$. Analogously, a_{m-1} and β_{m-1} are equal to entries of J_m. We introduce the notation $\widetilde{a} = (U_{m-2}, A\,v_{m-1})_{\mathsf{M}} \in \mathbb{C}^{m-2}$. The entries $(U_{m-2}, A\,v_m)_{\mathsf{M}}$ are zero due to $A\,v_m \in \operatorname{span}\{v_{m-1}, v_m, v_{m+1}\}$ being M-orthogonal to U_{m-2}.

(iii) The basis transformation $\widetilde{K}_m = (V_m, \widetilde{U}_m)_{\mathsf{M}}$ for $\widetilde{U}_m = (U_{m-2}, v_{m-1}, v_m)$ and $V_m = (v_1, \ldots, v_m)$, where $\widetilde{U}_m$ and V_m are M-orthonormal bases, indeed has the simple structure (4.3.23).

(iv) We proceed with the matrix entry ω_m of T_m in (4.3.4). Following Algorithm 4.4, ω_m evaluates to

$$\omega_m = \xi + \beta_{m-1}^2 e_{m-1}^{\mathsf{H}} (J_{m-1} - \xi I)^{-1} e_{m-1}, \tag{4.3.25}$$

where β_{m-1} refers to $(J_m)_{m,m-1}$ which is equal to $(\widetilde{A}_m)_{m,m-1}$, see (4.3.22). By the matrix structure of $\widetilde{K}_m$ (see (4.3.23)) we have $\widetilde{K}_{m-1} e_{m-1} = e_{m-1}$, thus,

$$e_{m-1}^{\mathsf{H}} (J_{m-1} - \xi I)^{-1} e_{m-1} = e_{m-1}^{\mathsf{H}} \widetilde{K}_{m-1}^{\mathsf{H}} (J_{m-1} - \xi I)^{-1} \widetilde{K}_{m-1} e_{m-1} \tag{4.3.26}$$

Furthermore, $\widetilde{K}_{m-1}^{\mathsf{H}} \widetilde{K}_{m-1} = I$ (4.2.23c) together with $\widetilde{A}_{m-1} = \widetilde{K}_{m-1}^{\mathsf{H}} J_{m-1} \widetilde{K}_{m-1}$ yield

$$e_{m-1}^{\mathsf{H}} \widetilde{K}_{m-1}^{\mathsf{H}} (J_{m-1} - \xi I)^{-1} \widetilde{K}_{m-1} e_{m-1} = e_{m-1}^{\mathsf{H}} (\widetilde{A}_{m-1} - \xi I)^{-1} e_{m-1}. \tag{4.3.27}$$

Combining (4.3.25) with (4.3.26) and (4.3.27) we conclude (4.3.24b).

Compare J_m (4.2.12) with T_m (4.3.4) to observe

$$T_m = J_m + (\omega_m - a_m) e_m e_m^{\mathsf{H}}.$$

With $\widetilde{K}_m^{\mathsf{H}}\, e_m = e_m$ and $\widetilde{A}_m = \widetilde{K}_m^{\mathsf{H}}\, J_m\, \widetilde{K}_m$ this implies

$$\widetilde{B}_m = \widetilde{K}_m^{\mathsf{H}}\, T_m\, \widetilde{K}_m = \widetilde{A}_m + (\omega_m - a_m) e_m e_m^{\mathsf{H}}. \tag{4.3.28}$$

With (4.3.22) and (4.3.28) we conclude (4.3.24a).

□

Remark 4.3.8. *The approach of Proposition 4.3.7 provides B_m for an extended Krylov subspace and can be slightly modified to fit for a fully rational Krylov subspace $\mathcal{Q}_m(A, u)$. Let $s_1, \ldots, s_{m-1} \in \mathbb{C} \cup \{\pm\infty\}$, where $s_{m-2}, s_{m-1} \in \mathbb{C}$, and let q_{m-1} be the respective denominator. We recall*

$$\mathcal{Q}_m(A, u) = \mathcal{K}_m(A, u_q), \quad \text{where} \quad u_q = q_{m-1}^{-1}(A) u. \tag{4.3.29}$$

We introduce the modified initial vector $\widehat{u}$ and denominator $\widehat{q}_{m-3}(A)$ as

$$\begin{aligned} \widehat{u} &= (A - s_{m-2} I)^{-1} (A - s_{m-1} I)^{-1} u, \quad \text{and} \\ \widehat{q}_{m-3}(A) &= (A - s_1 I)(A - s_2 I) \cdots (A - s_{m-3} I). \end{aligned} \tag{4.3.30}$$

Let $\mathcal{Q}_{m-2}(A,\widehat{u})$ be the rational Krylov subspace according to the initial vector $\widehat{u}$ and poles $s_1,\ldots,s_{m-3}$. Then

$$\mathcal{Q}_{m-2}(A,\widehat{u}) = \mathcal{K}_{m-2}(A,\widehat{q}_{m-3}^{-1}(A)\widehat{u}).$$

Due to (4.3.30)*, this initial vector satisfies $\widehat{q}_{m-3}^{-1}(A)\widehat{u} = u_q$ for u_q given in* (4.3.29)*. This implies*

$$\mathcal{Q}_{m-2}(A,\widehat{u}) = \mathcal{K}_{m-2}(A,u_q). \tag{4.3.31}$$

To apply Proposition 4.3.7 for the rational Krylov subspace $\mathcal{Q}_{m-1}(A,u)$ in (4.3.29)*, we represent $\mathcal{Q}_{m-1}(A,u)$ with poles $s_1,\ldots,s_{m-1}\in\mathbb{C}$ as an extended Krylov subspace of the form* (4.3.21)*. Substituting $\widehat{u}$ for the initial vector u in extended Krylov subspace in* (4.3.21)*, we have*

$$\mathcal{Q}_{m-2}(A,\widehat{u}) \oplus \operatorname{span}\{A\,\widehat{u}, A^2\widehat{u}\}. \tag{4.3.32}$$

We proceed to show that this accumulated vector space coincides with $\mathcal{Q}_m(A,u)$. Substituting

$$\mathcal{K}_m(A,u_q) = \mathcal{K}_{m-2}(A,u_q) \oplus \operatorname{span}\{A^{m-2}u_q, A^{m-1}u_q\}$$

for $\mathcal{K}_m(A,u_q)$ in (4.3.29)*, we have*

$$\mathcal{Q}_m(A,u) = \mathcal{K}_{m-2}(A,u_q) \oplus \operatorname{span}\{A^{m-2}u_q, A^{m-1}u_q\}. \tag{4.3.33}$$

Substituting $\widehat{u} = \widehat{q}_{m-3}(A)u_q$, we rewrite the right-hand term in (4.3.32) *to*

$$\operatorname{span}\{A\,\widehat{u}, A^2\widehat{u}\} = \operatorname{span}\{A\,\widehat{q}_{m-3}(A)u_q, A^2\widehat{q}_{m-3}(A)u_q\}.$$

The matrix polynomials $A\,\widehat{q}_{m-3}(A)$ and $A^2\widehat{q}_{m-3}(A)$ correspond to polynomials of degree $m-2$ and $m-1$, respectively, and this implies

$$\mathcal{K}_{m-2}(A,u_q) \oplus \operatorname{span}\{A\,\widehat{u}, A^2\widehat{u}\} = \mathcal{K}_{m-2}(A,u_q) \oplus \operatorname{span}\{A^{m-2}u_q, A^{m-1}u_q\}.$$

Combining (4.3.31) *and* (4.3.33) *with this identity, we conclude*

$$\mathcal{Q}_m(A,u) = \mathcal{Q}_{m-2}(A,\widehat{u}) \oplus \operatorname{span}\{A\,\widehat{u}, A^2\widehat{u}\}.$$

Thus, this rational Krylov subspace corresponds to an extended Krylov subspace with initial vector $\widehat{u} = (A-s_{m-2}I)^{-1}(A-s_{m-1}I)^{-1}u$, and the approach of of Proposition 4.3.7 provides an algorithm to compute a rational qor-Krylov representation B_m of $\mathcal{Q}_m(A,u)$ without accessing $q_{m-1}^{-1}(A)u$.

Following Remark 4.3.8, the approach of Proposition 4.3.7 provides a procedure to compute the matrix B_m for a rational Krylov subspace. For the SaI Krylov subspace with a single pole $s\in\mathbb{C}$ of multiplicity $m-1$ and a fixed $\xi\in\mathbb{R}$ this is specified in Algorithm 4.5.

A rational qor-Krylov approximation to matrix functions $f(A)u$.
We refer to

$$U_m f(B_m)x \approx f(A)u \tag{4.3.34}$$

as a rational quasi-orthogonal residual (qor-)Krylov approximation.

Algorithm 4.5: An algorithm to compute the matrix B_m for the SaI Krylov subspace with a single pole $s \in \mathbb{C}$ of multiplicity $m-1$ and a preassigned Ritz value $\xi \in \mathbb{R}$. This algorithm follows Proposition 4.3.7 for a modified starting vector $\widehat{u} = X^2 u$ with $X = (A - s\,I)^{-1}$.

$\widehat{u} = X^2 u$ with $X = (A - s\,I)^{-1}$;
run Algorithm 4.1 to compute U_{m-2} and A_{m-2} for the SaI Krylov subspace $\mathcal{K}_{m-2}(X, \widehat{u})$;
$\widetilde{v} = A\widehat{u}$;
orthogonalize $\widetilde{v}$ with U_{m-2} and set $v_{m-1} = \widetilde{v}/\|\widetilde{v}\|_{\mathsf{M}}$;
$\widehat{v} = A\,v_{m-1}$;
for $j = 1, \ldots, m-2$;
    $y_j = (u_j, \widehat{v})_{\mathsf{M}}$;
    $\widehat{v} \leftarrow \widehat{v} - y_j u_j$;
$a_{m-1} = (v_{m-1}, \widehat{v}\,)_{\mathsf{M}}$ and $\widehat{v} \leftarrow \widehat{v} - a_{m-1} v_{m-1}$;
$\beta_{m-1} = \|\widehat{v}\|_{\mathsf{M}}$ and $v_m = \widehat{v}/\beta_{m-1}$;
$A_{m-1} = [\,A_{m-2}, y\,;\; y^{\mathsf{H}}, a_{m-1}\,]$;
$\omega_m = \xi + \beta_{m-1}^2 e_{m-1}^{\mathsf{H}} (A_{m-1} - \xi I)^{-1} e_{m-1}$;
$B_m = [\,A_{m-1}, \beta_{m-1} e_{m-1}\,;\; \beta_{m-1} e_{m-1}^{\mathsf{H}}, \omega_m\,]$;
$U_m = (U_{m-2}, v_{m-1}, v_m)$;
$x = (U_m, u)_{\mathsf{M}}$;
return x, U_m, B_m;

4.4 The Separation Theorem of Chebyshev-Markov-Stieltjes (CMS Theorem) for polynomial and some rational Krylov subspaces

The CMS Theorem states that the accumulated quadrature weights of Gaussian quadrature formulae are bounded by Riemann-Stieltjes integrals over the intervals between the left integral limit and the quadrature nodes. In Subsection 4.4.1 we first reformulate previously stated identities of the Krylov representation (namely, Proposition 4.2.3, 4.2.4, 4.3.3 and 4.3.6) as Gaussian quadrature formulae for the Riemann-Stieltjes integral associated with the step function α_n; this allows us to present results in the following subsections (which apply in the Krylov setting) for a more general setting, i.e., for Gaussian quadrature formulae. We also recall some notation for Gaussian quadrature formulae of Riemann-Stieltjes integrals, and we link classical notations to the previously introduced setting.

In Subsection 4.4.2 we recapitulate the CMS Theorem for the polynomial Krylov setting, and in Subsection 4.4.3–4.4.5 we introduce CMS type results for various rational Krylov settings.

Throughout the present chapter, we consider integrals associated with a non-decreasing step function α_n with n points of strict increase. However, most of the results in the present section hold true for integrals associated with non-decreasing continuous functions α in a similar manner; the case of α being a continuous is not discussed in detail in the present

work.

4.4.1 Gaussian quadrature formulae and Krylov subspaces. Historical context

The integral associated with the step function α_n is to be understood as a Riemann-Stieltjes integral. Gaussian quadrature formulae for Riemann-Stieltjes integrals are also referred to as Gauss-Christoffel quadrature formulae in the literature, for previous remarks see also Subsection 4.1.1. For the Gauss-Christoffel quadrature formula which integrates polynomials of degree $\leq 2m-1$ exactly, the quadrature nodes are given by the zeros of the associated orthogonal polynomial of degree m, and the quadrature weights are given by so called Christoffel numbers. Similar results hold for Gauss-Radau formulae for which the quadrature nodes and weights coincide with zeros of quasi-orthogonal polynomials and respective Christoffel numbers. We briefly recapitulate the relation between Gaussian quadrature formulae and the Jacobi matrix, which is also mentioned in Subsection 4.1.1; for further details on Gaussian quadrature formulae we refer to [Gau81] and others. Further below in the present subsection, we recall similar results for rational Gaussian quadrature formulae.

The Christoffel numbers and the eigendecomposition of the Jacobi matrix. For the orthonormal polynomials $p_0, \ldots, p_{m-1}$ associated with the distribution $\mathrm{d}\alpha_n$, see Proposition 4.2.1, we define

$$\rho_{m-1}(\lambda) = 1 \Big/ \sum_{k=0}^{m-1} p_k(\lambda)^2 \in \mathbb{R}.$$

We recall that the Ritz values $\theta_1, \ldots, \theta_m \in \mathbb{R}$ correspond to the zeros of p_m. The numbers $\rho_{m-1}(\theta_1), \ldots, \rho_{m-1}(\theta_m)$ are also referred to as *Christoffel numbers* in the literature.

We proceed to recall the relation between Christoffel numbers and entries of eigenvectors of the Jacobi matrix which goes back to [Wil62, GW69]. We introduce the denotation $c_1, \ldots, c_m \in \mathbb{R}$ for the spectral coefficients of the vector $\beta_0 e_1$ in the eigenbasis of J_m, which further correspond to the first components of the scaled eigenvectors: Let $\widehat{q}_1, \ldots, \widehat{q}_m \in \mathbb{R}^m$ denote the ℓ^2-orthonormal eigenvectors of J_m, i.e., $J_m \widehat{q}_j = \theta_j \widehat{q}_j$ for the Ritz values θ_j and $(\widehat{q}_j, \widehat{q}_k)_2 = \delta_{jk}$, then

$$c_j = \beta_0 \, (\widehat{q}_j, e_1)_2 \in \mathbb{R}. \tag{4.4.1}$$

The Christoffel numbers correspond to the first components of the eigenvectors of the Jacobi matrix: We recall the following results for the eigenvectors of J_m. Following Section 4.2, the eigenvector for the eigenvalue θ_j is given by

$$(p_0(\theta_j), \ldots, p_{m-1}(\theta_j))^\intercal \in \mathbb{R}^m. \tag{4.4.2}$$

For the first component of the eigenvector we have $p_0 = 1/\beta_0$. Thus, the first component of the j-th normalized eigenvector scaled by β_0 and squared satisfies

$$c_j^2 = 1 \Big/ \sum_{k=0}^{m-1} p_k(\theta_j)^2 \in \mathbb{R}, \quad j = 1, \ldots, m, \tag{4.4.3}$$

and for the Christoffel numbers we have the identity

$$c_j^2 = \rho_{m-1}(\theta_j), \quad j = 1, \ldots, m. \tag{4.4.4}$$

The Christoffel numbers are nonzero,[12] i.e., $c_j \neq 0$. Although c_j is real-valued, we also write $|c_j|^2$ in place of c_j^2.

Similar results hold for the spectrum of the qor-Krylov representation T_m introduced in Subsection 4.3.1. We reuse some notation associated with the spectrum of J_m for T_m: Corresponding to T_m the denotations $\theta_1, \ldots, \theta_m$ and $c_1, \ldots, c_m$ refer to the eigenvalues of T_m and the spectral coefficients of $\beta_0 e_1$ in the ℓ^2-orthonormal eigenbasis of T_m, respectively. For the qor-Krylov representation T_m we assume that the preassigned eigenvalue ξ is given such that the underlying quasi-orthogonal polynomial is well-defined, and we assume that the eigenvalues of T_m are included within the integral limits of the respective Riemann-Stieltjes integral. (See Proposition 4.3.1 for some details on the location of the eigenvalues of T_m.) Following (4.3.5), the eigenvectors of T_m conform to (4.4.2) when $\theta_1, \ldots, \theta_m$ refer to the respective eigenvalues. Similar to the case of the Jacobi matrix, the representation (4.4.3) and the identity (4.4.4) also hold true for T_m.

A review on Gaussian quadrature formulae for the Riemann-Stieltjes integral. We proceed to reformulate Proposition 4.2.3 and 4.2.4 as Gaussian quadrature formulae for the Riemann-Stieltjes integral associated with the step function α_n. We recall that α_n is based on the eigenvalues of A and the spectral coefficients of u.

For a complex-valued function $f\colon \mathbb{R} \to \mathbb{C}$, where we consider polynomials or rational functions later on, the following formulations are equivalent (see also (4.2.8)),

$$\int_a^b f(\lambda) \mathrm{d}\alpha_n(\lambda) = (u, f(A)u)_{\mathsf{M}} = \sum_{j=1}^{n} f(\lambda_j)|w_j|^2. \tag{4.4.5a}$$

In a similar manner, the orthonormal eigendecomposition of J_m yields

$$\beta_0^2 (e_1, f(J_m)e_1)_2 = \sum_{j=1}^{m} f(\theta_j)|c_j|^2. \tag{4.4.5b}$$

Identity (4.4.5b) also holds true for T_m if θ_j and c_j refer to the spectrum of T_m.

The Ritz values θ_j and Christoffel numbers $\rho_{m-1}(\theta_j)$ provide quadrature nodes and weights, respectively, for the Gaussian quadrature formulae which are also referred to as Gauss-Christoffel quadrature formulae in the literature, see also [Gau81]. We recapitulate classical results on Gaussian quadrature formulae using the notation $|c_j|^2$ for the Christoffel numbers, see (4.4.4).

Remark 4.4.1 (Gaussian quadrature property, e.g., Subsection 6.2 [GM10])**.** *The Ritz values $\theta_1, \ldots, \theta_m$ and the spectral coefficients $c_1, \ldots, c_m$ w.r.t. J_m constitute a Gaussian quadrature formula for the Riemann-Stieltjes integral* (4.2.6b)*,*

$$\int_a^b p(\lambda) \mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} p(\theta_j)|c_j|^2, \quad p \in \Pi_{2m-1}. \tag{4.4.6}$$

[12] The result $c_j \neq 0$ is clarified in Appendix 4.A, Proposition 4.A.2.

Here, the Ritz values and the spectral coefficients represent the quadrature nodes and quadrature weights, respectively. On the basis of results of the present work, identity (4.4.6) *can be verified via the identities for the inner product in* (4.2.6c) *and* (4.4.5),

$$\int_a^b p(\lambda)\mathrm{d}\alpha_n(\lambda) = (u, p(A)u)_{\mathsf{M}} = \beta_0^2(e_1, p(J_m)e_1)_2 = \sum_{j=1}^{m} p(\theta_j)|c_j|^2, \quad p \in \Pi_{2m-1}.$$

Analogously, the qor-Krylov representation T_m provides the following quadrature formula. Let $\theta_1, \ldots, \theta_m$ and $c_1, \ldots, c_m$ be the eigenvalues and spectral coefficients of T_m, then the identities (4.3.8) *for $p \in \Pi_{2m-2}$ together with* (4.4.5) *imply*

$$\int_a^b p(\lambda)\mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} p(\theta_j)|c_j|^2, \quad p \in \Pi_{2m-2}. \tag{4.4.7}$$

When $\xi = a$ (thus, $\theta_1 = a$) or $\xi = b$ (thus, $\theta_m = b$) is preassigned this is also referred to as a Gauss-Radau quadrature formula.

In view of Remark (4.4.1) we summarize results for the Jacobi matrix J_m and the qor-Krylov representation T_m. For these results we write out the Riemann-Stieltjes integral (4.2.6b) in terms of its sum representation.

Corollary 4.4.2. *Let $\theta_1, \ldots, \theta_m$ and $c_1, \ldots, c_m$ denote the eigenvalues and spectral coefficients, respectively, of either J_m or T_m, where the spectral coefficients c_j refer to the vector $\beta_0\, e_1$. Then,*

$$\int_a^b p(\lambda)\mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} p(\theta_j)|c_j|^2, \quad p \in \Pi_{2m-2}. \tag{4.4.8}$$

Rational Gaussian quadrature formulae and rational Krylov subspaces. For rational Krylov subspaces $\mathcal{Q}_m(A, u)$ we recall the definition of the Rayleigh quotient $A_m = (U_m, AU_m)_{\mathsf{M}}$, where U_m is an orthonormal basis of $\mathcal{Q}_m(A, u)$. Furthermore, the vector $x = (U_m, u)_{\mathsf{M}}$ and the rational qor-Krylov representation B_m (introduced in Subsection 4.3.2 via (4.3.11)) implicitly depend on U_m. In the sequel we consider U_m to be fixed, and we assume that B_m is well-defined. For the latter we refer to the conditions concerning the definition of T_m in Section 4.3. We proceed to reuse the denotation $\theta_1, \ldots, \theta_m$ for the eigenvalues of A_m ('rational' Ritz values), and $c_1, \ldots, c_m$ for the spectral coefficients of x in the orthonormal eigenbasis of A_m: Let $\widehat{q}_j \in \mathbb{C}^m$ denote the ℓ^2-orthonormal eigenvectors of A_m, i.e., $A_m\widehat{q}_j = \theta_j\widehat{q}_j$ and $(\widehat{q}_j, \widehat{q}_k)_2 = \delta_{jk}$, then

$$c_j = (\widehat{q}_j, x)_2 \in \mathbb{C}, \quad j = 1, \ldots, m. \tag{4.4.9}$$

We remark that the coefficients c_j are independent of the explicit choice of the orthonormal basis U_m, this is clarified in Proposition 4.A.4, Appendix 4.A. For a function $f\colon \mathbb{R} \to \mathbb{C}$, the eigendecomposition of A_m yields

$$(x, f(A_m)x)_2 = \sum_{j=1}^{m} f(\theta_j)|c_j|^2. \tag{4.4.10}$$

In the context of the rational qor-Krylov representation B_m the denotation $\theta_1, \dots, \theta_m$ and $c_1, \dots, c_m$ is reused accordingly, and an identity similar to (4.4.10) holds true for B_m when θ_j and c_j refer to the spectrum of B_m.

Remark 4.4.3. *Similar to Remark 4.4.1, the identity in* (4.2.26) *corresponds to the following rational Gaussian quadrature formula. Let* $\theta_1, \dots, \theta_m$ *and* $c_1, \dots, c_m$ *refer to the spectrum of* A_m*, then*

$$\int_a^b r(\lambda)\,\mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} r(\theta_j)|c_j|^2, \quad r \in \Pi_{2m-1}/|q_{m-1}|^2. \tag{4.4.11}$$

To demonstrate (4.4.11) *we recall the identities for the inner product in* (4.2.26), (4.4.5a), *and* (4.4.10),

$$\int_a^b r(\lambda)\mathrm{d}\alpha_n(\lambda) = (u, r(A)u)_{\mathsf{M}} = (x, r(A_m)x)_2 = \sum_{j=1}^{m} |c_j|^2 r(\theta_j), \quad r \in \Pi_{2m-1}/|q_{m-1}|^2.$$

The rational qor-Krylov representation B_m *provides the following quadrature formula via Proposition 4.3.6,*

$$\int_a^b r(\lambda)\mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} r(\theta_j)|c_j|^2, \quad r \in \Pi_{2m-2}/|q_{m-1}|^2.$$

When the preassigned eigenvalue of B_m *is set to one of the integral limits, i.e.,* $\theta_1 = a$ *or* $\theta_m = b$*, then this formula is also referred to as rational Gauss-Radau quadrature formula.*

We summarize the statements of Remark 4.4.3 concerning A_m and B_m.

Corollary 4.4.4. *Let* $\theta_1, \dots, \theta_m$ *and* $c_1, \dots, c_m$ *denote the eigenvalues and spectral coefficients, respectively, of either* A_m *or* B_m*, where the spectral coefficients refer to the vector* x*. Then,*

$$\int_a^b r(\lambda)\,\mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} r(\theta_j)|c_j|^2, \quad r \in \Pi_{2m-2}/|q_{m-1}|^2. \tag{4.4.12}$$

4.4.2 The CMS Theorem for the polynomial case

The CMS Theorem dates back to works of Chebyshev, Markov and Stieltjes in the 19th century and also goes by the name Chebyshev-Markov-Stieltjes inequalities. For further historical and technical remarks we refer to [Sze85, Section 3.41] (including an extensive survey of this theorem), [Akh65, Theorem 2.54], [VA93, Section 4], [LS13, Section 3], [Chi78] and others.

The Riemann-Stieltjes integral associated with α_n (4.2.6a) over a subset of (a, b) can be understood as a measure of such a subset. Namely, with $\alpha_n(a) = 0$ we consider $\alpha_n(\theta)$ to be the associated measure of the interval $(a, \theta]$ for $\theta \in (a, b)$. To simplify the notation in the sequel, we let $\mu_n(R)$ denote the measure of a subset R of (a, b) associated with α_n. More precisely, we first define

$$J(R) = \{j : \lambda_j \in R\} \subset \{1, \dots, n\}, \quad \text{for a set } R \subset (a, b). \tag{4.4.13a}$$

The sum of the spectral coefficients w_j over the index set $J(R)$ corresponds to the measure of the set R associated with α_n, and we define

$$\mu_n(R) = \sum_{j \in J(R)} |w_j|^2. \tag{4.4.13b}$$

Thus, we have $\mu_n((a,\theta]) = \alpha_n(\theta)$ for $\theta \in (a,b)$. Furthermore, we proceed to use the notation μ_n and α_n for the measure of an interval $(a,\theta]$ in an equivalent manner. Similarly, we use the notation $\alpha_n(\theta-)$ for the measure of the open interval (a,θ), i.e.,

$$\alpha_n(\theta-) := \lim_{\varepsilon \to 0+} \alpha_n(\theta-\varepsilon) = \mu_n((a,\theta)).$$

We proceed to recall the CMS Theorem. This theorem is based on the Gaussian quadrature properties (4.4.8) as in Corollary 4.4.2, and thus, the following results hold true when θ_j and c_j refer to the spectrum of the Jacobi matrix J_m or the qor-Krylov representation T_m.

Theorem 4.4.5 (Separation Theorem of Chebyshev-Markov-Stieltjes, see also Section 3.41 in [Sze85])**.** *Let $\theta_1, \ldots, \theta_m \in (a,b)$ and $c_1, \ldots, c_m \in \mathbb{C}$ satisfy the Gaussian quadrature property* (4.4.8)*, then*

$$\alpha_n(\theta_k) < |c_1|^2 + \ldots + |c_k|^2 < \alpha_n(\theta_{k+1}-), \qquad k = 1, \ldots, m-1. \tag{4.4.14}$$

We point out that for $k = m$ the bounds in (4.4.14) can be replaced by the following identity. The Gaussian quadrature property (4.4.8) for $p = 1$ implies

$$\sum_{j=1}^{m} |c_j|^2 = \alpha_n(b). \tag{4.4.15}$$

(This also results directly from $\|u\|_{\mathsf{M}} = \beta_0 \|e_1\|_2$.)

To recall a classical proof of the CMS Theorem we introduce the following polynomials.

Proposition 4.4.6 (Eq. (3.411.1) in [Sze85], part of Theorem (2.5.4) in [Akh65] and others[13])**.** *Let $\theta_1 < \ldots < \theta_m \in \mathbb{R}$ and let k be fixed with $1 \le k < m$. Then there exist polynomials $p_{\{+,k\}}$ and $p_{\{-,k\}} \in \Pi_{2m-2}$ which satisfy*[14]

$$p_{\{\pm,k\}}(\theta_j) = \begin{cases} 1, & j = 1, \ldots, k, \\ 0, & j = k+1, \ldots, m, \end{cases} \tag{4.4.16}$$

together with

$$p_{\{+,k\}}(\lambda) \ge \begin{cases} 1, & \lambda \le \theta_k, \\ 0, & \lambda > \theta_k, \end{cases} \quad \text{and} \quad p_{\{-,k\}}(\lambda) \le \begin{cases} 1, & \lambda < \theta_{k+1}, \\ 0, & \lambda \ge \theta_{k+1}. \end{cases} \tag{4.4.17}$$

Additionally, the inequalities in (4.4.17) *are strict inequalities for $\lambda \notin \{\theta_1, \ldots, \theta_m\}$.*

The polynomials of Proposition 4.4.6 are illustrated in Figure 4.2 for a numerical example.

With Proposition 4.4.6 we proceed to prove Theorem 4.4.5.

[13] A classical proof of Proposition 4.4.6 is recapitulated in Appendix 4.B.

[14] In the sequel statements concerning $p_{\{\pm,k\}}$ apply to $p_{\{+,k\}}$ and $p_{\{-,k\}}$ individually.

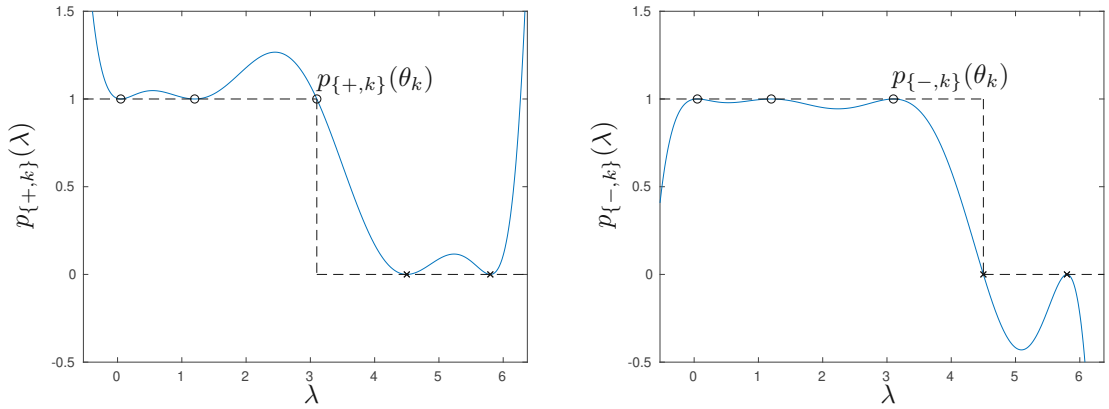


Figure 4.2: This figure illustrates the polynomials $p_{\{+,k\}}$ (left) and $p_{\{-,k\}}$ (right) given in Proposition 4.4.6 for given nodes $\theta_1, \ldots, \theta_m$ with $m = 5$. The identities (4.4.16) are illustrated for $\theta_1, \ldots, \theta_k$ ('$\circ$') and $\theta_{k+1}, \ldots, \theta_m$ ('$\times$') with $k = 3$, and the dashed line illustrates the bounds (4.4.17).

Proof of ***Theorem 4.4.5****.* Let $p_{\{\pm,k\}} \in \Pi_{2m-2}$ be given according to Proposition 4.4.6 for the eigenvalues $\theta_1 < \ldots < \theta_m$ and $k = 1, \ldots, m-1$. The polynomials $p_{\{\pm,k\}}$ satisfy (4.4.16), and this implies

$$\sum_{j=1}^{m} p_{\{\pm,k\}}(\theta_j)|c_j|^2 = \sum_{j=1}^{k} |c_j|^2. \tag{4.4.18}$$

On the other hand, identity (4.4.8) yields

$$\sum_{j=1}^{m} p_{\{\pm,k\}}(\theta_j)|c_j|^2 = \int_a^b p_{\{\pm,k\}}(\lambda)\, \mathrm{d}\alpha_n(\lambda).$$

Evaluating the Riemann-Stieltjes integral in this identity, we arrive at

$$\sum_{j=1}^{m} p_{\{\pm,k\}}(\theta_j)|c_j|^2 = \sum_{j=1}^{n} p_{\{\pm,k\}}(\lambda_j)|w_j|^2. \tag{4.4.19}$$

Let the index set $J((a, \theta_k]) \subset \{1, \ldots, n\}$ be given as in (4.4.13a). Then, the inequalities for $p_{\{+,k\}}$ in (4.4.17) imply

$$\sum_{j=1}^{n} p_{\{+,k\}}(\lambda_j)|w_j|^2 > \sum_{j \in J((a,\theta_k])} |w_j|^2 = \alpha_n(\theta_k). \tag{4.4.20}$$

This inequality is strict due to the interlacing property of the eigenvalues λ_j and θ_j, see Proposition 4.2.2 and 4.3.1. Combining (4.4.18), (4.4.19) and (4.4.20) yields the lower bound in (4.4.14).

Similarly to (4.4.20), the inequalities for $p_{\{-,k\}}$ in (4.4.17) imply

$$\sum_{j=1}^{n} p_{\{-,k\}}(\lambda_j)|w_j|^2 < \sum_{j\in J((a,\theta_{k+1}))} |w_j|^2 = \alpha_n(\theta_{k+1}-). \tag{4.4.21}$$

Combining (4.4.18), (4.4.19) and (4.4.21) yields the upper bound in (4.4.14). □

The inequalities (4.4.14) in Theorem 4.4.5 yield the following bounds on the measure of the intervals located between Ritz values. In the following, we use the notation μ_n for the measure as in (4.4.13b).

Corollary 4.4.7. *In the setting of Theorem 4.4.5, the following inequalities hold true.*

- *For indices j, k with $1 < j < k < m$,*

$$\mu_n([\theta_j, \theta_k]) < |c_j|^2 + |c_{j+1}|^2 + \ldots + |c_k|^2 < \mu_n((\theta_{j-1}, \theta_{k+1})) \tag{4.4.22a}$$

- *Furthermore, the accumulated spectral coefficients satisfy*

$$\mu_n([\theta_j, b)) < |c_j|^2 + \ldots + |c_m|^2 < \mu_n((\theta_{j-1}, b)), \quad j = 2, \ldots, m. \tag{4.4.22b}$$

Proof. Applying (4.4.14) twice (once we substitute $j-1$ for the index k therein) and subtracting, we observe

$$\alpha_n(\theta_k) - \alpha_n(\theta_j-) < |c_j|^2 + |c_{j+1}|^2 + \ldots + |c_k|^2 < \alpha_n(\theta_{k+1}-) - \alpha_n(\theta_{j-1}),$$

this shows (4.4.22a). Subtracting (4.4.14) for the index $j-1$ from (4.4.15), we arrive at

$$\alpha_n(b) - \alpha_n(\theta_j-) < |c_j|^2 + \ldots + |c_m|^2 < \alpha_n(b) - \alpha_n(\theta_{j-1}), \tag{4.4.23}$$

which entails (4.4.22b). □

We proceed to specify the intertwining property of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ which already appeared in the introduction of Subsection 4.1.1: Similarly to α_n in (4.2.6a), we introduce the step function

$$\alpha_m(\lambda) = \begin{cases} 0, & \lambda < \theta_1, \\ \sum_{j=1}^{\ell} |c_j|^2, & \theta_\ell \le \lambda < \theta_{\ell+1}, \quad \ell = 1, \ldots, m-1, \\ \sum_{j=1}^{m} |c_j|^2, & \theta_m \le \lambda. \end{cases} \tag{4.4.24}$$

For $f : \mathbb{R} \to \mathbb{C}$ the Riemann-Stieltjes integral associated with α_m reads

$$\int_a^b f(\lambda)\,\mathrm{d}\alpha_m(\lambda) = \sum_{j=1}^{m} |c_j|^2 f(\theta_j).$$

Thus, the quadrature property in Corollary 4.4.2 coincides with the identity

$$\int_a^b \lambda^j\,\mathrm{d}\alpha_n(\lambda) = \int_a^b \lambda^j\,\mathrm{d}\alpha_m(\lambda), \quad j = 0, \ldots, 2m-2. \tag{4.4.25}$$

The integral terms in (4.4.25) correspond to the moments of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$, and thus, the Gaussian quadrature property in Corollary 4.4.2 coincides with $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ having matching moments up to order $2m-2$. We define the auxiliary function

$$F(\lambda) = \alpha_n(\lambda) - \alpha_m(\lambda), \tag{4.4.26}$$

and remark the following properties of F. The step functions $\alpha_n(\lambda)$ and $\alpha_m(\lambda)$ are both increasing in λ, whereat the step function $\alpha_m(\lambda)$ has exactly m points of increase at $\lambda = \theta_1, \ldots, \theta_m$. Thus, the function $F(\lambda)$ is increasing for $\lambda \in (\theta_k, \theta_{k+1})$, $k = 1, \ldots, m-1$, and away from the boundaries $\lambda < \theta_1$ and $\lambda > \theta_m$. Furthermore, Theorem 4.4.5 yields

$$\alpha_n(\theta_k) - (|c_1|^2 + \ldots + |c_k|^2) < 0 < \alpha_n(\theta_{k+1}-) - (|c_1|^2 + \ldots + |c_k|^2), \qquad k = 1, \ldots, m-1.$$

The accumulated coefficients c_k correspond to the step function α_m (4.4.24), namely,

$$|c_1|^2 + \ldots + |c_k|^2 = \alpha_m(\theta_k) = \alpha_m(\theta_{k+1}-), \tag{4.4.27}$$

and we observe the inequalities

$$F(\theta_k) < 0 < F(\theta_{k+1}-) \quad \text{for} \quad k = 1, \ldots, m-1. \tag{4.4.28}$$

More precisely, the inequalities (4.4.28) are equivalent to the assertion of the CMS Theorem (Theorem 4.4.5).

To clarify the intertwining property of $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ in this context: The CMS Theorem relies on quadrature properties which correspond to (4.4.25), i.e., $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ having matching moments, and the result of the CMS Theorem corresponds to (4.4.28), which can be understood as an intertwining property of $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$.

Besides these remarks, the function F is further used in the following subsection to rewrite CMS type results for rational cases, and in Section 4.5 below where we verify results of the present section for numerical examples.

Remark 4.4.8. *In the present chapter, the measure α_n is introduced based on eigenvalues λ_j of A and the spectral coefficients w_j of the initial vector u in the eigenbasis of A as in* (4.2.6a)*. Thus, the bounds given by the CMS Theorem reveal bounds for the accumulated spectral coefficients w_j. To simplify the notation we proceed with the setting of the Jacobi matrix J_m, i.e., the eigenvalues θ_j and spectral coefficients c_j refer to the spectrum of the Jacobi matrix. In a similar manner such results also hold for the qor-Krylov representation T_m as specified below. The bounds on α_n provided by the CMS Theorem are computable, i.e., θ_j and c_j are available via an eigendecomposition of the Jacobi matrix which can be computed using the Lanczos method.*

We proceed in the setting of the Jacobi matrix. For its eigenvalues θ_j we define the index $\ell = \ell(k)$ for $k = 1, \ldots, m$, such that

$$\lambda_{\ell(k)} \leq \theta_k < \lambda_{\ell(k)+1}. \tag{4.4.29}$$

The positioning of the eigenvalues, which is specified in Proposition 4.2.2, implies $\ell(k) < \ell(k+1)$ for $k = 1, \ldots, m-1$ and $1 \leq \ell(k) < n$ for $k = 1, \ldots, m$.

With $\ell(k)$ defined in (4.4.29) *we have the representations*

$$\alpha_n(\theta_k) = \sum_{j=1}^{\ell(k)} |w_j|^2, \quad \text{and} \quad \alpha_n(\theta_{k+1}) = \sum_{j=1}^{\ell(k+1)} |w_j|^2, \quad k = 1, \ldots, m-1. \tag{4.4.30}$$

Note that $\alpha_n(\theta_{k+1}-) \leq \alpha_n(\theta_{k+1})$; to keep the notation simple, the case $\alpha_n(\theta_{k+1}-) < \alpha_n(\theta_{k+1})$ is not treated separately here. For the remainder of the present remark we assume

$$\lambda_{\ell(k)} \neq \theta_k, \quad k = 1, \ldots, m.$$

Thus, with (4.4.30) *Theorem 4.4.5 reads*

$$\sum_{j=1}^{\ell(k)} |w_j|^2 < \sum_{j=1}^{k} |c_j|^2 < \sum_{j=1}^{\ell(k+1)} |w_j|^2, \quad k = 1, \ldots, m-1. \tag{4.4.31}$$

Furthermore, for a set of eigenvalues of A located between two Ritz values θ_j and θ_k with $j < k$ we recall

$$\lambda_{\ell(j)} < \theta_j < \lambda_{\ell(j)+1} < \ldots < \lambda_{\ell(k)} < \theta_k, \quad k = 2, \ldots, m,$$

and with (4.4.30)*, the sum of spectral coefficients w_j associated with these eigenvalues corresponds to*

$$\sum_{\iota=\ell(j)+1}^{\ell(k)} |w_\iota|^2 = \alpha_n(\theta_k) - \alpha_n(\theta_j), \quad j < k. \tag{4.4.32}$$

Furthermore, combining this identity with (4.4.31) *or* (4.4.22)*, we obtain computable bounds on accumulated spectral coefficients of u. E.g., for $1 < j < k < m$ the inequality* (4.4.22a) *yields*

$$|c_{j+1}|^2 + \ldots + |c_{k-1}|^2 < \sum_{\iota=\ell(j)+1}^{\ell(k)} |w_\iota|^2 < |c_j|^2 + \ldots + |c_k|^2,$$

where the lower bound is trivial in the case $k = j + 1$.

We remark that the results of the present subsection can be generalized to the setting of the qor-Krylov representation T_m. For the qor-Krylov representation, the cases $\theta_1 < \lambda_1$ and $\lambda_n < \theta_m$ have to be considered explicitly in the notation, namely, the indices $\ell(1)$ and $\ell(m)$ have to be adapted accordingly for these cases.

Remark 4.4.9. *In the present work the measure α_n is based on the spectrum of A and has n points of strict increase. Thus, the identity of [Sze85, eq. (3.41.3)] which relies on a continuous measure does not hold true in the present case, i.e.,*

$$\text{in general, we do not find any point } y_k \in \mathbb{R} \text{ such that } \alpha_n(y_k) = \sum_{j=1}^{k} |c_j|^2.$$

Nevertheless, the inequalities in (4.4.31) *imply that there exist indices* ν_k *with* $\ell(k) < \nu_k \leq \ell(k+1)$ *and numbers* $\xi_k \in (0,1]$ *for* $k = 1, \dots, m-1$ *such that*

$$\sum_{j=1}^{\nu_k - 1} |w_j|^2 + \xi_k |w_{\nu_k}|^2 = \sum_{j=1}^{k} |c_j|^2,$$

This can give further theoretical insight on the estimates provided in Remark 4.4.8. Nevertheless, the indices ν_j *and scaling factors* ξ_j *are not computable in general.*

The indices ν_k *satisfy* $\lambda_{\nu_k} \in (\theta_k, \theta_{k+1}]$*, thus,*

$$\lambda_1 < \theta_1 < \lambda_{\nu_1} \leq \theta_2 < \lambda_{\nu_2} < \ldots \leq \theta_{m-1} < \lambda_{\nu_{m-1}} \leq \theta_m < \lambda_n.$$

For each spectral coefficient c_k*, this implies*

$$|c_1|^2 = \sum_{j=1}^{\nu_1 - 1} |w_j|^2 + \xi_1 |w_{\nu_1}|^2,$$

$$|c_k|^2 = (1 - \xi_{k-1}) |w_{\nu_{k-1}}|^2 + \sum_{j=\nu_{k-1}+1}^{\nu_k - 1} |w_j|^2 + \xi_k |w_{\nu_k}|^2, \quad k = 2, \dots, m-1, \quad \text{and}$$

$$|c_m|^2 = (1 - \xi_{m-1}) |w_{\nu_{m-1}}|^2 + \sum_{j=\nu_{m-1}+1}^{n} |w_j|^2.$$

4.4.3 The rational case with a single pole $s \in \mathbb{R}$ of higher multiplicity

In the present subsection we consider CMS type results for the setting of a rational Krylov subspace $\mathcal{Q}_m(A, u)$ with a single pole $s \in \mathbb{R}$, thus, we have the denominator $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$. This subspace corresponds to a SaI Krylov subspace; for previous remarks see also Subsection 4.2.1. Following Subsection 4.4.1, the eigenvalues $\theta_1, \dots, \theta_m \in (a, b)$ and spectral coefficients $c_1, \dots, c_m \in \mathbb{C}$ of the respective Rayleigh quotient A_m or qor-representation B_m satisfy the quadrature property (4.4.12) in Corollary 4.4.4. To provide results in a more general setting, the results in the remainder of the subsection are based on the quadrature property (4.4.12); we provide results for a class of rational Gaussian quadrature formulae which fit to the respective SaI Krylov setting.

Although the rational Krylov subspace corresponds to the polynomial Krylov subspace $\mathcal{K}_m(A, u_q)$ with starting vector $u_q = q_{m-1}^{-1}(A)u$, results of the previous subsection do not yield bounds associated with α_n, this is specified in the following remark.

Remark 4.4.10. *The rational Krylov subspace* $\mathcal{Q}_m(A, u)$ *with the respective denominator* q_{m-1} *is identical to* $\mathcal{K}_m(A, u_q)$ *with* $u_q = q_{m-1}^{-1}(A)u$*. This polynomial Krylov subspace is associated with the step function* $\widehat{\alpha}_n$ *given in* (4.2.22)*. Let* J_m *and* V_m *denote the Jacobi matrix and the* M*-orthonormal eigenbasis of* $\mathcal{K}_m(A, u_q)$ *constructed by the Lanczos method. In the setting of* $\mathcal{K}_m(A, u_q)$*, Theorem 4.4.5 yields bounds based on spectral coefficients of the vector* $x_q = (V_m, u_q)_{\mathsf{M}}$ *in the eigenbasis of* J_m *and the step function* $\widehat{\alpha}_n$*. This does not entail bounds based on spectral coefficients of* $x = (V_m, u)_{\mathsf{M}}$ *and the step function* α_n *in general.*

To simplify the notation in the sequel, we first define the indices k_1 and k_m such that

$$\theta_{k_m} < s < \theta_{k_1}, \quad k_1 = k_m + 1, \quad \text{in case of } s \in (\theta_1, \theta_m), \tag{4.4.33a}$$

and otherwise,

$$k_1 = 1 \quad \text{and} \quad k_m = m, \quad \text{in case of } s < \theta_1 \text{ or } s > \theta_m. \tag{4.4.33b}$$

Furthermore, we define the sets $I_k \subset \{1, \ldots, m\}$ and $R_k \subset \mathbb{R}$ for $k = 1, \ldots, m$ by

$$\begin{aligned} I_k &= \begin{cases} \{k_1, \ldots, k\}, & k_1 \le k \le m, \\ \{1, \ldots, k, k_1, \ldots, m\}, & 1 \le k < k_1, \end{cases} \quad \text{and} \\ R_k &= \begin{cases} (s, \theta_k], & \theta_k > s, \\ (a, \theta_k] \cup (s, b), & \theta_k < s. \end{cases} \end{aligned} \tag{4.4.34}$$

The set R_k is illustrated in Figure 4.3.

Let $\mu_n(R_k)$ and $\mu_n(R_k^{\circ})$ as in (4.4.13b) denote the measure of the sets R_k and [15]R_k°, respectively. Thus, we have

$$\mu_n(R_k) = \begin{cases} \mu_n((s, \theta_k]) &= \alpha_n(\theta_k) - \alpha_n(s), & \theta_k > s, \\ \mu_n((a, \theta_k] \cup (s, b)) &= \alpha_n(\theta_k) + \alpha_n(b) - \alpha_n(s), & \theta_k < s, \end{cases} \tag{4.4.35a}$$

and

$$\mu_n(R_k^{\circ}) = \begin{cases} \mu_n((s, \theta_k)) &= \alpha_n(\theta_k-) - \alpha_n(s), & \theta_k > s, \\ \mu_n((a, \theta_k) \cup (s, b)) &= \alpha_n(\theta_k-) + \alpha_n(b) - \alpha_n(s), & \theta_k < s. \end{cases} \tag{4.4.35b}$$

In the following theorem we provide a CMS type result for a class of rational Gaussian quadrature formulae which applies to the setting of SaI Krylov subspaces with a shift $s \in \mathbb{R}$.

Theorem 4.4.11 (A separation theorem for rational Gaussian quadrature formulae with a single single pole $s \in \mathbb{R}$ of higher multiplicity)**.** *Let $\theta_1, \ldots, \theta_m \in (a, b)$ and $c_1, \ldots, c_m \in \mathbb{C}$ satisfy the rational Gaussian quadrature properties* (4.4.12) *for $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$ with $s \in \mathbb{R}$. Let the index k_m be defined as in* (4.4.33)*. Let the sets $I_k \subset \{1, \ldots, m\}$ and $R_k \subset \mathbb{R}$ for $k = 1, \ldots, m$ be defined as in* (4.4.34)*, and let μ_n be defined as in* (4.4.13b) *(analogously,* (4.4.35)*). Additionally, define $R_{m+1} := R_1$. Then,*

$$\mu_n(R_k) < \sum_{j \in I_k} |c_j|^2 < \mu_n(R_{k+1}^{\circ}), \qquad k \in \{1, \ldots, m\} \setminus \{k_m\}. \tag{4.4.36}$$

The case $k = k_m$ is not discussed in Theorem 4.4.11. In this case we have $I_k = \{1, \ldots, m\}$ and the bounds (4.4.36) can be replaced by the identity

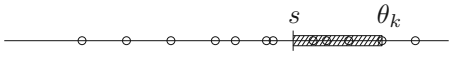
$$\sum_{j=1}^{m} |c_j|^2 = \alpha_n(b). \tag{4.4.37}$$

This identity corresponds to the identity (4.4.12) for $r = 1$ (or directly results from $\|u\|_{\mathsf{M}} = \|x\|_2$).

To prove Theorem 4.4.11, we first introduce rational functions which constitute bounds on a Heaviside type step function, similar to the polynomials given in Proposition 4.4.6.

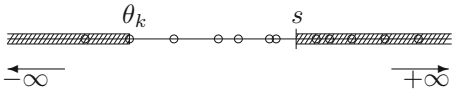
[15] In the sequel, we let R° denote the interior of a set R.

a) the case $\theta_k > s$:



$R_k = (s, \theta_k]$

b) the case $\theta_k < s$:



$R_k = (a, \theta_k] \cup (s, b)$

Figure 4.3: In Figure **a)** and **b)** we illustrate the set $R_k \subset \mathbb{R}$ given in (4.4.34) for a given sequence of nodes $\theta_1, \ldots, \theta_m$ ('$\circ$'), and a given pole s which satisfies $\theta_1 < s < \theta_m$. In Figure a) we choose the index k such that $\theta_k > s$, and in Figure b) we consider $\theta_k < s$. In each figure the set R_k is highlighted by a dashed area.

Proposition 4.4.12. *Let $\theta_1 < \ldots < \theta_m$ be a given sequence and let $s \in \mathbb{R}$ be a given pole which is distinct to $\theta_1, \ldots, \theta_m$. We make use of the denotations k_1 and k_m introduced in (4.4.33). Furthermore, let the sets $I_k \subset \{1, \ldots, m\}$ and $R_k \subset \mathbb{R}$ for $k = 1, \ldots, m$ be defined as in (4.4.34), and we define $R_{m+1} := R_1$.*

For $k \in \{1, \ldots, m\} \setminus \{k_m\}$ and $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$ there exist rational functions $r_{\{+,k\}}$ and $r_{\{-,k\}} \in \Pi_{2m-2}/|q_{m-1}|^2$ which satisfy[16]

$$r_{\{\pm,k\}}(\theta_j) = \begin{cases} 1, & j \in I_k, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4.38}$$

Furthermore, we have

$$r_{\{+,k\}}(\lambda) \geq \begin{cases} 1, & \lambda \in R_k, \\ 0, & \lambda \in \mathbb{R}_s \setminus R_k, \end{cases} \quad \text{and} \quad r_{\{-,k\}}(\lambda) \leq \begin{cases} 1, & \lambda \in R_{k+1}^{\circ}, \\ 0, & \lambda \in \mathbb{R}_s \setminus R_{k+1}^{\circ}, \end{cases} \tag{4.4.39}$$

where $\mathbb{R}_s = (a, b) \setminus \{s\}$. The inequalities in (4.4.39) are strict for $\lambda \notin \{\theta_1, \ldots, \theta_m\}$. Without loss of generality, we assume $(a, b) = \mathbb{R}$ in the present proposition.

Proof. See Appendix 4.B. □

Rational functions $r_{\{-,k\}}$ as introduced in Proposition 4.4.12 are illustrated in Figure 4.4 for a numerical example.

We proceed with the proof of Theorem 4.4.11.

*Proof of **Theorem 4.4.11**.* Let k_1 and k_m be given in (4.4.33), and let $k \in \{1, \ldots, m\} \setminus \{k_m\}$ be fixed. For the eigenvalues $\theta_1, \ldots, \theta_m$ we let $r_{\{\pm,k\}}$ denote the rational functions given in Proposition 4.4.12. We proceed to prove the lower bound in (4.4.36). The identities (4.4.38) imply

$$\sum_{j \in I_k} |c_j|^2 = \sum_{j=1}^{m} r_{\{\pm,k\}}(\theta_j) |c_j|^2. \tag{4.4.40a}$$

[16] Analogously to $p_{\{\pm,k\}}$, the denotation $r_{\{\pm,k\}}$ refers to $r_{\{+,k\}}$ and $r_{\{-,k\}}$ individually.

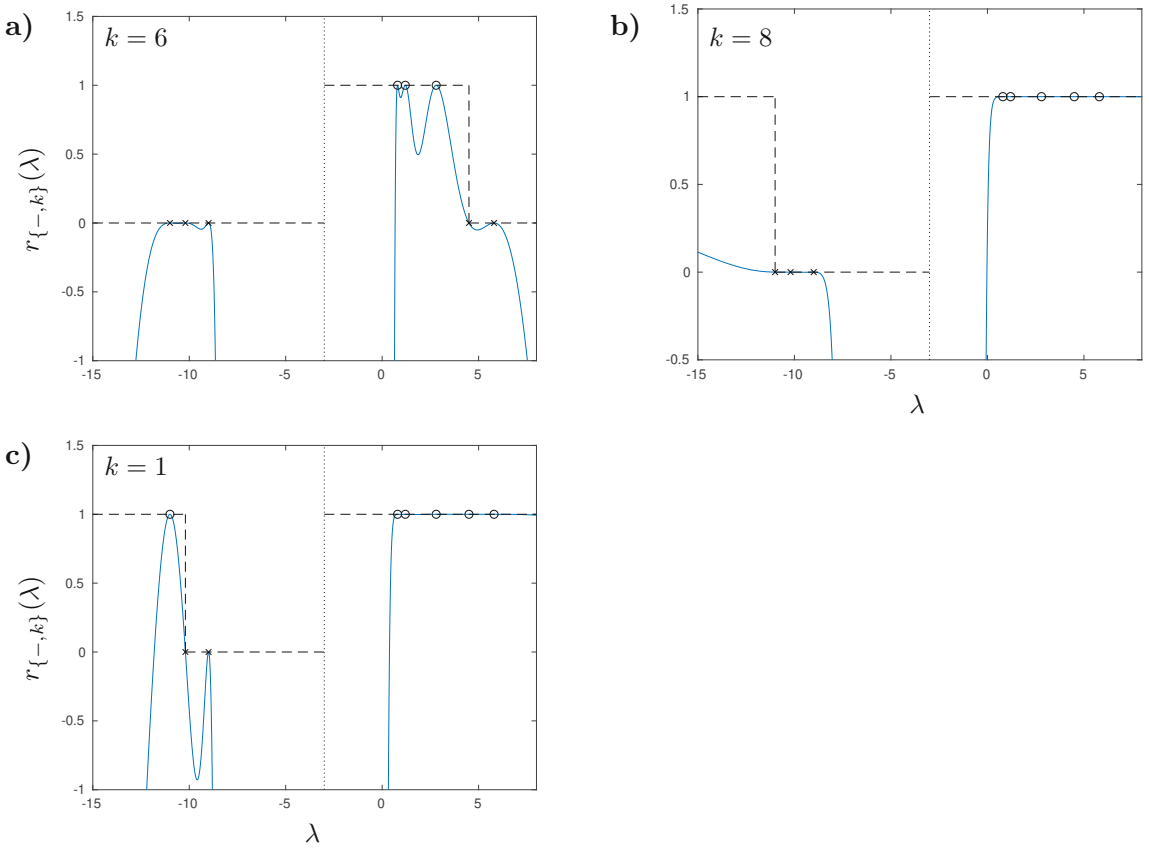


Figure 4.4: In Figure **a)**–**c)** we show $r_{\{-,k\}}$ for a given sequence of nodes $\theta_1, \ldots, \theta_m$ with $m = 8$, and a given pole s which is located between the nodes, i.e., $\theta_1 < s < \theta_m$. These figures show results for different choices of $k \in \{1, \ldots, m\} \setminus \{k_m\}$ where $k_m = 3$ (following (4.4.33b)). In each figure the symbols ('$\circ$') and ('$\times$') mark $r_{\{-,k\}}(\theta_j)$ for $j \in I_k$ and $j \notin I_k$, respectively. The dashed lines illustrate the upper bounds given in (4.4.39). Figure b) shows the special case $k = m$ for which the upper bound (4.4.39) relies on $R_{m+1} = R_1$. For additional illustrations considering $r_{\{\pm,k\}}$ we refer to Figure 4.13 and 4.14 in Appendix 4.B.

Furthermore, the quadrature property (4.4.12) implies

$$\int_a^b r_{\{\pm,k\}}(\lambda)\,\mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} r_{\{\pm,k\}}(\theta_j)|c_j|^2.$$

Rewriting this Riemann-Stieltjes integral as in (4.4.5a), we arrive at

$$\sum_{j=1}^{n} r_{\{\pm,k\}}(\lambda_j)|w_j|^2 = \sum_{j=1}^{m} r_{\{\pm,k\}}(\theta_j)|c_j|^2. \tag{4.4.40b}$$

The inequalities in (4.4.39) for $r_{\{+,k\}}$ entail

$$\sum_{j=1}^{n} r_{\{+,k\}}(\lambda_j)|w_j|^2 > \sum_{j\in J(R_k)} |w_j|^2. \tag{4.4.41}$$

This inequality is strict due to the interlacing property of the eigenvalues λ_j and θ_j, see Proposition 4.2.2 and 4.3.1. Combine (4.4.40) and (4.4.41) to conclude with the lower bound in (4.4.36).

In a similar manner, $r_{\{-,k\}}$ reveals the upper bound in (4.4.36); the inequalities in (4.4.39) for $r_{\{-,k\}}$ yield

$$\sum_{j=1}^{n} r_{\{-,k\}}(\lambda_j)|w_j|^2 < \sum_{j\in J(R^{\circ}_{k+1})} |w_j|^2. \tag{4.4.42}$$

Indeed, the identities (4.4.40) together with (4.4.42) conclude the upper bound in (4.4.36). □

We reformulate the result of Theorem 4.4.11 in the following proposition.

Proposition 4.4.13. *In the setting of Theorem 4.4.11, the following inequality holds true,*

$$\alpha_n(\theta_k) \le |c_1|^2 + \ldots + |c_k|^2 + \gamma \le \alpha_n(\theta_{k+1}-), \quad k = 1, \ldots, m-1, \tag{4.4.43a}$$

with

$$\gamma = \alpha_n(s) - \alpha_m(s), \tag{4.4.43b}$$

where α_m is given in (4.4.24). The inequalities in (4.4.43a) are strict for $k \neq k_m$. Additionally, the case $k = m \neq k_m$ in (4.4.36) corresponds to

$$\alpha_n(\theta_m) \le |c_1|^2 + \ldots + |c_m|^2 + \gamma, \qquad \text{and} \quad \gamma \le \alpha_n(\theta_1-). \tag{4.4.44}$$

*Proof of **Proposition 4.4.13**.* We first prove (4.4.43). Here, we consider different cases for the index $k = 1, \ldots, m-1$.

- $k < k_1$ with k_1 as in (4.4.33); this case only occurs for $k_1 > 1$ which follows from $s \in (\theta_1, \theta_m)$. This case also implies $k_m = k_1 - 1$,

$$\alpha_m(s) = |c_1|^2 + \ldots + |c_{k_m}|^2, \quad \text{and} \quad \alpha_m(b) - \alpha_m(s) = |c_{k_1}|^2 + \ldots + |c_m|^2, \tag{4.4.45}$$

and with I_k as in (4.4.34),

$$\sum_{j\in I_k} |c_j|^2 = |c_1|^2 + \ldots + |c_k|^2 + \alpha_m(b) - \alpha_m(s). \tag{4.4.46}$$

– Additionally, let $k < k_1 - 1 = k_m$. This case implies $\theta_k, \theta_{k+1} < s$ and as given in (4.4.35),

$$\mu_n(R_k) = \alpha_n(\theta_k) + \alpha_n(b) - \alpha_n(s), \quad \text{and} \quad \mu_n(R^{\circ}_{k+1}) = \alpha_n(\theta_{k+1}-) + \alpha_n(b) - \alpha_n(s). \tag{4.4.47}$$

Substituting (4.4.46) and (4.4.47) in the inequalities (4.4.36), subtracting $\alpha_n(b)$ ($= \alpha_m(b)$) and adding $\alpha_n(s)$, we conclude (4.4.43) for $k < k_1 - 1$.

– Let $k = k_1 - 1$. Thus, $k = k_m \neq m$, and for this case the inequalities (4.4.36) do not apply; we show (4.4.43) in a direct manner:

For $k = k_m$ we have $|c_1|^2 + \ldots + |c_k|^2 = \alpha_m(s)$ as in (4.4.45). With this identity, the enclosed term in (4.4.43a) simplifies to

$$|c_1|^2 + \ldots + |c_k|^2 + \gamma = \alpha_n(s). \tag{4.4.48}$$

Due to α_n being an increasing function and $\theta_k < s < \theta_{k+1}$ for $k = k_1 - 1$ we have

$$\alpha_n(\theta_k) \leq \alpha_n(s) \leq \alpha_n(\theta_{k+1}-), \quad \text{for } k = k_1 - 1. \tag{4.4.49}$$

Combining (4.4.48) and (4.4.49), we conclude (4.4.43) for the case $k = k_1 - 1$.

- $k \geq k_1$. For this case we further distinguish between $s < \theta_m$ and $s > \theta_m$.

  – Let $s < \theta_m$ (this includes the case $s < \theta_1$). With I_k as in (4.4.34), we have

$$\sum_{j\in I_k} |c_j|^2 = |c_1|^2 + \ldots + |c_k|^2 - \alpha_m(s) \tag{4.4.50}$$

  Furthermore, this case implies $\theta_k, \theta_{k+1} > s$ (we recall $k < m$), and as in (4.4.35),

$$\mu_n(R_k) = \alpha_n(\theta_k) - \alpha_n(s), \tag{4.4.51a}$$

  and

$$\mu_n(R^{\circ}_{k+1}) = \alpha_n(\theta_{k+1}-) - \alpha_n(s). \tag{4.4.51b}$$

  Substituting (4.4.51) and (4.4.50) in the inequalities (4.4.36), we conclude (4.4.43) for $k \geq k_1$ and $s < \theta_m$.

  – Otherwise, for $k \geq k_1$ and $s > \theta_m$ our notation simplifies to $k_1 = 1$ and

$$\sum_{j\in I_k} |c_j|^2 = |c_1|^2 + \ldots + |c_k|^2. \tag{4.4.52}$$

  The case $s > \theta_m$ implies $\alpha_m(b) = \alpha_m(s)$, and due to $\alpha_m(b) = \alpha_n(b)$, we have

$$\alpha_n(b) = \alpha_m(s). \tag{4.4.53}$$

Furthermore, we have $\theta_k, \theta_{k+1} < s$, and $\mu_n(R_k)$ and $\mu_n(R^{\circ}_{k+1})$ correspond to (4.4.47) further above. Making use of (4.4.53) in (4.4.47) and substituting γ, we simplify

$$\mu_n(R_k) = \alpha_n(\theta_k) - \gamma, \quad \text{and} \quad \mu_n(R^{\circ}_{k+1}) = \alpha_n(\theta_{k+1}-) - \gamma. \tag{4.4.54}$$

Substituting (4.4.52) and (4.4.54) in the inequalities (4.4.36), we conclude (4.4.43) for $k \geq k_1$ and $s > \theta_m$.

We proceed with the proof of (4.4.44). The case $k = m \neq k_m$ only occurs for $s \in (\theta_1, \theta_m)$. Thus with $s < \theta_m$, $\mu_n(R_m)$ corresponds to (4.4.51a). Substituting $\mu_n(R_m)$ as in (4.4.51a) and the sum over I_m as in (4.4.50) in the lower bound in (4.4.36), we conclude the inequality on the left-hand side of (4.4.44).

To prove the inequality on the right-hand side of (4.4.44), we first recall $\theta_1 < s$, and as in (4.4.47)

$$\mu_n(R^{\circ}_1) = \alpha_n(\theta_1-) + \alpha_n(b) - \alpha_n(s). \tag{4.4.55}$$

Substituting (4.4.50) and (4.4.55) in the upper bound in (4.4.36) (for the case $k = m \neq k_m$ with $\mu_n(R^{\circ}_{m+1}) = \mu_n(R^{\circ}_1)$ due to convention), we arrive at

$$|c_1|^2 + \ldots + |c_m|^2 - \alpha_m(s) < \alpha_n(\theta_1-) + \alpha_n(b) - \alpha_n(s)$$

On the left-hand side we can further simplify $|c_1|^2 + \ldots + |c_m|^2 = \alpha_n(b)$ and subtract this term, which entails the inequality on the right-hand side of (4.4.44). □

Remark 4.4.14. *For the case $\alpha_m(s) = \alpha_n(s)$ the constant γ in Proposition 4.4.13 is zero, and the inequalities* (4.4.43a) *coincide with the inequalities given by Theorem 4.4.5, i.e., the CMS Theorem for polynomial Gaussian quadrature formulae. Furthermore, for this case the inequalities given in Corollary 4.4.7 hold true. Here, we highlight the case $s \notin (\lambda_1, \lambda_n)$ for the Gaussian quadrature formulae without preassigned nodes (this implies $\theta_j \in (\lambda_1, \lambda_n)$); a prominent case for which $\alpha_m(s) = \alpha_n(s)$ holds true a priori.*

Remark 4.4.15. *Following Remark 4.2.6, the SaI Krylov representation $X_m^{-1} + sI$ provides a Gaussian quadrature formula. For the case of a real shift $s \in \mathbb{R}$, the respective quadrature nodes are located on the real axis, and at least one eigenvalue λ_j is located between each neighboring pair of quadrature nodes. Thus, the result of Theorem 4.4.11 and its corollaries hold true in this setting. However, results concerning the SaI Krylov representation $X_m^{-1} + sI$ are not discussed in further detail in the present work.*

CMS type results for the SaI Krylov representation are also given in [ZTK19]. In the present work we include the case of a shift s being located inside the convex hull of the spectrum of A, which extends some results of [ZTK19].

We proceed to specify the results of Proposition 4.4.13 for the pole s being located in the convex hull of the rational Ritz values, i.e., $s \in (\theta_1, \theta_m)$. This case implies $k_m \neq m$, and substituting $\alpha_n(b) = |c_1|^2 + \ldots + |c_m|^2$ in (4.4.44), we observe

$$\alpha_n(\theta_m) - \alpha_n(b) \leq \gamma \leq \alpha_n(\theta_1-). \tag{4.4.56}$$

With these inequalities, we further specify the results of Proposition 4.4.13: The following corollary states some bounds on piecewise accumulated quadrature weights, similar to Corollary 4.4.7 in the previous subsection for the polynomial case.

Corollary 4.4.16. *Additionally to the setting of Theorem 4.4.11, we assume $s \in (\theta_1, \theta_m)$. Then Proposition 4.4.13 yields the following inequalities.*

- *The accumulated quadrature weights satisfy*

$$\mu_n([\theta_1, \theta_k]) \leq |c_1|^2 + \ldots + |c_k|^2 \leq \mu_n((a, \theta_{k+1}) \cup (\theta_m, b)), \quad k = 1, \ldots, m-1. \tag{4.4.57a}$$

- *For indices j, k with $1 < j < k < m$, the following piecewise accumulated quadrature weights satisfy*

$$\mu_n([\theta_j, \theta_k]) \leq |c_j|^2 + \ldots + |c_k|^2 \leq \mu_n((\theta_{j-1}, \theta_{k+1})). \tag{4.4.57b}$$

- *Furthermore, the accumulated quadrature weights satisfy*

$$\mu_n([\theta_j, \theta_m]) \leq |c_j|^2 + \ldots + |c_m|^2 \leq \mu_n((a, \theta_1) \cup (\theta_{j-1}, b)), \quad j = 2, \ldots, m. \tag{4.4.57c}$$

Proof. The inequalities (4.4.43a) in Proposition 4.4.13 yield

$$\alpha_n(\theta_k) - \gamma \leq |c_1|^2 + \ldots + |c_k|^2 \leq \alpha_n(\theta_{k+1}-) - \gamma.$$

Substituting (4.4.56) for γ, we arrive at

$$\alpha_n(\theta_k) - \alpha_n(\theta_1-) \leq |c_1|^2 + \ldots + |c_k|^2 \leq \alpha_n(\theta_{k+1}-) + \alpha_n(b) - \alpha_n(\theta_m).$$

This implies (4.4.57a).

To prove the inequalities in (4.4.57c), we first remark

$$|c_j|^2 + \ldots + |c_k|^2 = |c_1|^2 + \ldots + |c_k|^2 + \gamma - \big(|c_1|^2 + \ldots + |c_{j-1}|^2 + \gamma\big).$$

Applying (4.4.43a) twice (once we substitute $j-1$ for the index k therein) we observe

$$\alpha_n(\theta_k) - \alpha_n(\theta_j-) \leq |c_j|^2 + \ldots + |c_k|^2 \leq \alpha_n(\theta_{k+1}-) - \alpha_n(\theta_j),$$

which implies (4.4.57b).

To show (4.4.57c), apply (4.4.57a) for the index $j-1$ and subtract the result from $|c_1|^2 + \ldots + |c_m|^2 = \mu_n((a, b))$. □

Remark 4.4.17. *For the case $s \in (\theta_1, \theta_m)$ as in Corollary 4.4.16, bounds on quadrature weights related to the leftmost or rightmost quadrature nodes potentially depend on the measure of an interval including the opposite integral limit. This relation can be avoided by preassigning one of the quadrature nodes at the integral limit, using a rational Gauss-Radau formula associated with the spectrum of a rational qor-Krylov representation B_m in the Krylov setting.*

- *For a preassigned node $\xi < \lambda_1$, we have $\theta_1 = \xi$ and $\alpha_n(\theta_1) = 0$. Thus, the inequalities in* (4.4.57a) *correspond to*

$$\mu_n([a, \theta_k]) \leq |c_1|^2 + \ldots + |c_k|^2 \leq \mu_n((a, \theta_{k+1}) \cup (\theta_m, b)), \tag{4.4.58a}$$

and the inequalities in (4.4.57c) *correspond to*

$$\mu_n([\theta_j, \theta_m]) \leq |c_j|^2 + \ldots + |c_m|^2 \leq \mu_n((\theta_{j-1}, b)). \tag{4.4.58b}$$

- *For a preassigned node $\xi > \lambda_n$, we have $\theta_m = \xi$ and $\alpha_n(\theta_m) = \alpha_n(b)$. Thus, the inequalities in* (4.4.57a) *correspond to*

$$\mu_n([\theta_1, \theta_k]) \leq |c_1|^2 + \ldots + |c_k|^2 \leq \mu_n((a, \theta_{k+1})), \tag{4.4.58c}$$

and the inequalities in (4.4.57c) *correspond to*

$$\mu_n([\theta_j, b]) \leq |c_j|^2 + \ldots + |c_m|^2 \leq \mu_n((a, \theta_1) \cup (\theta_{j-1}, b)). \tag{4.4.58d}$$

We proceed to introduce a step function F_s which changes its sign at each rational Ritz value according to Proposition 4.4.13; with the step function F given in (4.4.26) we introduce

$$F_s(\lambda) = F(\lambda) - F(s). \tag{4.4.59}$$

Here, $F(s) = \gamma$ with γ as in Proposition 4.4.13. As previously stated in (4.4.27), we have

$$\alpha_m(\theta_k) = \alpha_m(\theta_{k+1}-) = |c_1|^2 + \ldots + |c_k|^2.$$

Then the inequalities (4.4.43a) correspond to

$$F_s(\theta_k) \leq 0 \leq F_s(\theta_{k+1}-), \quad k = 1, \ldots, m-1, \tag{4.4.60a}$$

whereat these inequalities are strict for $k \neq k_m$. Furthermore, the inequalities (4.4.44) correspond to

$$F_s(\theta_m) < 0, \quad \text{and} \quad 0 < F_s(\theta_1-), \tag{4.4.60b}$$

for $k_m \neq m$.

Remark 4.4.18. *In* (4.4.60)*, the special case $k = k_m$ holds true due to the identity* (4.4.37)*; the case $k \in \{1, \ldots, m\} \setminus \{k_m\}$ corresponds to the result of Theorem 4.4.11. Namely, the result of Theorem 4.4.11 conforms to the following inequality in an equivalent manner,*

$$F_s(\theta_k) \leq 0 \leq F_s(\theta_{k+1}-) \quad \textit{for} \quad k \in \{1, \ldots, m\} \setminus \{k_m\}, \ \textit{and with} \ \theta_{m+1} = \theta_1.$$

4.4.4 The rational case with a single pole $s \in \mathbb{C} \setminus \mathbb{R}$ of higher multiplicity

In the present subsection, we consider rational Gaussian quadrature formulae which satisfy the quadrature property (4.4.12) with $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$ for $s \in \mathbb{C} \setminus \mathbb{R}$. To specify, these quadrature formulae are exact for rational functions with denominator $|q_{m-1}(\lambda)|^2 = ((\lambda - \operatorname{Re} s)^2 + (\operatorname{Im} s)^2)^{m-1}$ where $\operatorname{Im} s \neq 0$, i.e., rational functions with complex-conjugate poles of higher multiplicity. Considering Krylov subspaces, these quadrature formulae are related to SaI Krylov subspaces with a complex shift $s \in \mathbb{C} \setminus \mathbb{R}$.

As a main result of the present subsection, the following Proposition yields upper bounds on the measure of the intervals between neighboring quadrature nodes, and the measure at the boundary of the spectrum.

Proposition 4.4.19. *Let $c_1, \ldots, c_m$ and $\theta_1 < \ldots < \theta_m$ satisfy the quadrature property* (4.4.12) *with $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$ for $s \in \mathbb{C} \setminus \mathbb{R}$. Then, with μ_n given in* (4.4.13)

$$\mu_n([\theta_k, \theta_{k+1}]) \leq |c_k|^2 + |c_{k+1}|^2, \quad k = 1, \ldots, m-1, \tag{4.4.61a}$$

and

$$\mu_n((a, \theta_1]) + \mu_n([\theta_m, b)) \leq |c_1|^2 + |c_m|^2. \tag{4.4.61b}$$

Before proving Proposition 4.4.19, we proceed with some auxiliary results. The results of the previous subsection do not apply for the case $s \in \mathbb{C} \setminus \mathbb{R}$. However, the present class of rational functions can be related to polynomials on the unit circle $\mathbb{T}$ and vice versa. To specify this relation, we recall the Cayley transform as in (4.2.35),

$$\tau(\lambda) = (\lambda - \overline{s})(\lambda - s)^{-1}, \quad \tau \colon \mathbb{R} \to \mathbb{T} \setminus \{1\}.$$

For a complex polynomial $p \in \Pi_{m-1}$ we consider $p(\tau(\lambda))$ as a function of λ; normalizing shows

$$p(\tau(\lambda)) = g(\lambda)/q_{m-1}(\lambda), \quad \text{for some } g \in \Pi_{m-1},$$

For $\lambda \in \mathbb{R}$ we conclude

$$|p(\tau(\lambda))|^2 = \overline{g}(\lambda)g(\lambda)/|q_{m-1}(\lambda)|^2, \quad \text{where } \ \overline{g}g \in \Pi_{2m-2}. \tag{4.4.62}$$

In the following corollary we introduce rational majorants on a Heaviside type step function, based on interpolating polynomials on the unit circle given in [Gol02, Lemma 4].

Corollary 4.4.20 (A corollary of Lemma 4 in [Gol02])**.** *Let $\theta_1, \ldots, \theta_m$ be a given sequence of nodes, and let $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$ for a given pole $s \in \mathbb{C} \setminus \mathbb{R}$.*

(i) Let $k \in \{1, \ldots, m-1\}$ be fixed. There exists a rational function $r_k \in \Pi_{2m-2}/|q_{m-1}|^2$ with

$$r_k(\theta_j) = \begin{cases} 1, & j \in \{k, k+1\}, \\ 0, & \text{otherwise}, \end{cases} \tag{4.4.63a}$$

and

$$r_k(\lambda) \geq \begin{cases} 1, & \lambda \in [\theta_k, \theta_{k+1}], \\ 0, & \lambda \in (-\infty, \theta_k) \cup (\theta_{k+1}, \infty). \end{cases} \tag{4.4.63b}$$

(ii) Additionally, there exists a function $r_m \in \Pi_{2m-2}/|q_{m-1}|^2$ with

$$r_m(\theta_j) = \begin{cases} 1, & j \in \{1, m\}, \\ 0, & \text{otherwise}, \end{cases} \tag{4.4.64a}$$

and

$$r_m(\lambda) \geq \begin{cases} 1, & \lambda \in (-\infty, \theta_1] \cup [\theta_m, \infty), \\ 0, & \lambda \in (\theta_1, \theta_m). \end{cases} \tag{4.4.64b}$$

Proof. The Cayley transform τ as in (4.2.35) reads

$$\tau(\lambda) = (\lambda - \overline{s})(\lambda - s)^{-1}, \quad \tau \colon \mathbb{R} \to \mathbb{T} \setminus \{1\}.$$

Simplifying this fraction yields

$$\tau(\lambda) = 1 + \frac{s - \overline{s}}{\lambda - s} = 1 + \frac{2\mathrm{i}\,\operatorname{Im} s}{\lambda - s}.$$

Here, $\tau \colon \mathbb{R} \to \mathbb{T} \setminus \{1\}$ is a continuous and bijective function, and with the previous representation, we observe

$$\tau(-\infty) = \begin{cases} 1 + \mathrm{i}0+, & \operatorname{Im} s < 0, \\ 1 + \mathrm{i}0-, & \operatorname{Im} s > 0. \end{cases}$$

Thus, τ maps $\mathbb{R}$ to $\mathbb{T} \setminus \{1\}$ in counter-clockwise and clockwise order for $\operatorname{Im} s < 0$ and $\operatorname{Im} s > 0$, respectively.

We proceed to define distinct points $\zeta_1, \dots, \zeta_m \in \mathbb{T}$ by

$$\zeta_j := \begin{cases} \tau(\theta_j), & \operatorname{Im} s < 0, \\ \tau(\theta_{m-j+1}), & \operatorname{Im} s > 0, \end{cases} \qquad j = 1, \dots, m. \tag{4.4.65}$$

The points $\zeta_1, \dots, \zeta_m \in \mathbb{T}$ are distinct, in counter-clockwise order, and the point 1 is located between ζ_1 and ζ_m on the unit circle. For the remainder of the proof we assume the case $\operatorname{Im} s < 0$ to simplify the notation. Thus, we consider $\zeta_j = \tau(\theta_j)$.

Additionally to (4.4.65), we define $\zeta_{m+1} := \zeta_1$. Let $p_k \in \Pi_{m-1}$ denote the complex polynomial given by [Gol02, Lemma 4] for the points ζ_j and a fixed index $k \in \{1, \dots, m\}$. Here, we also normalize p_k at ζ_k. Thus, p_k satisfies $|p_k(\zeta_k)| = |p_k(\zeta_{k+1})| = 1$. Following (4.4.62), the function $r_k(\lambda) := |p_k(\tau(\lambda))|^2$ conforms to a rational function $r_k \in \Pi_{2m-2}/|q_{m-1}|^2$.

We proceed to show (4.4.63) for the rational function r_k; let $k \in \{1, \dots, m-1\}$:

- With $r_k(\theta_j) = p_k(\zeta_j)$ the identities $|p_k(\zeta_k)| = |p_k(\zeta_{k+1})| = 1$ yield $r_k(\theta_k) = r_k(\theta_{k+1}) = 1$, and the identity $p_k(\zeta_j) = 0$ for $j \neq k, k+1$ yields $r_k(\theta_j) = 0$ for $j \neq k, k+1$, which shows (4.4.63a).

- Due to τ being a continuous and bijective function, the points ζ located on the unit circle between ζ_k and ζ_{k+1} (including ζ_k and ζ_{k+1}) are identical to the set $\{\zeta = \tau(\lambda) \mid \lambda \in [\theta_k, \theta_{k+1}]\}$. As a result of [Gol02, Lemma 4], the polynomial p_k satisfies $|p_k(\zeta)| \geq 1$ for ζ in this set of points, i.e., $|p(\tau(\lambda))| \geq 1$ for $\lambda \in [\theta_k, \theta_{k+1}]$. Thus, we have $r_k(\lambda) \geq 1$ for $\lambda \in [\theta_k, \theta_{k+1}]$; furthermore, r_k is positive for $\lambda \in \mathbb{R}$ due to $r_k(\lambda) = |p_k(\xi(\lambda))|^2$, which implies (4.4.63b).

We proceed to sketch the proof of (4.4.64) which corresponds to the case $k = m$. The polynomial p_m satisfies $|p_m(\zeta_m)| = |p_m(\zeta_1)| = 1$. Furthermore, the points ζ located between ζ_1 and ζ_m correspond to the set $\{\zeta = \tau(\lambda) \mid \lambda \in (-\infty, \theta_1) \cup (\theta_m, \infty)\} \cup \{1\} \subset \mathbb{T}$. Similar to previous arguments, this shows (4.4.64).

Considering the definition of ζ_j in (4.4.65), similar arguments hold for the case $\operatorname{Im} s > 0$. □

We proceed with the proof of Proposition 4.4.19.

Proof of ***Proposition 4.4.19****.* Let $k \in \{1, \dots, m-1\}$ be fixed, we prove (4.4.61a). For the nodes $\theta_1, \dots, \theta_m$ and k given, we let $r_k \in \Pi_{2m-2}/|q_{m-1}|^2$ denote the rational function given in Corollary 4.4.20 which satisfies (4.4.63). Due to (4.4.63a) we have

$$\sum_{j=1}^{m} |c_j|^2 r_k(\theta_j) = |c_k|^2 + |c_{k+1}|^2. \tag{4.4.66a}$$

The quadrature property (4.4.12) implies

$$\int_a^b r_k(\lambda) \, \mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} |c_j|^2 r_k(\theta_j), \tag{4.4.66b}$$

and the inequality (4.4.63b) yields

$$\int_a^b r_k(\lambda)\,\mathrm{d}\alpha_n(\lambda) \geq \mu_n([\theta_k, \theta_{k+1}]). \tag{4.4.66c}$$

Combining (4.4.66a)–(4.4.66c), we conclude (4.4.61a).

Analogously, making use of the rational function $r_m \in \Pi_{2m-2}/|q_{m-1}|^2$ (which satisfies the properties (4.4.64)) in combination with the quadrature property (4.4.12), we conclude (4.4.61b).

□

4.4.5 Results for an extended Krylov subspace

In the present subsection, we consider an extended Krylov subspace. Namely, the Krylov subspace of so called Laurent polynomials which also appears in [DK98] and corresponds to a rational Krylov subspace. Here, we also include a shift $s < \lambda_1$. This yields a rational Krylov subspace with denominator $q(\lambda) = (\lambda - s)^{\varrho-1}$ for $m = 2\varrho - 1$, i.e.,

$$\begin{aligned}\mathcal{Q}_{2\varrho-1}(A, u) &= \operatorname{span}\{(A-s)^{-\varrho+1}u, \ldots, (A-s)^{-1}u, u, Au, \ldots, A^{\varrho-1}u\} \\ &= \mathcal{K}_{2\varrho-1}(A, (A-s)^{-\varrho+1}u).\end{aligned} \tag{4.4.67}$$

Similar to previous sections, U_m denotes an M-orthonormal basis of the Krylov subspace and $A_m = (U_m, A\,U_m)_{\mathsf{M}}$ denotes the associated Rayleigh quotient. As previously, we let $x = (U_m, u)_{\mathsf{M}}$. An extended Lanczos recurrence to compute U_m and A_m in an efficient manner is given in [DK98, Section 5] and summarized in Algorithm 4.6.

Let $\theta_1, \ldots, \theta_m \in (a, b)$ and $c_1, \ldots, c_m \in \mathbb{C}$ denote the eigenvalues and spectral coefficients of the Rayleigh quotient A_m. Following Proposition 4.2.26 and Corollary 4.4.4, these eigenvalues and spectral coefficients satisfy the identity (4.4.12) for $r \in \Pi_{2m-1}/q^2$ with $q^2(\lambda) = (\lambda - s)^{2\varrho-2} = (\lambda - s)^{m-1}$, i.e.,

$$\int_a^b r(\lambda)\,\mathrm{d}\alpha_n(\lambda) = \sum_{j=1}^{m} r(\theta_j)|c_j|^2, \qquad r \in \Pi_{2m-1}/(\lambda - s)^{m-1}. \tag{4.4.68}$$

The CMS type results given in [Li98] apply to rational quadrature formulae which satisfy (4.4.68). i.e., a Gaussian quadrature formulae for so called Laurent polynomials.

We proceed to recapitulate results given in [Li98] for the setting of the extended Krylov subspace (4.4.67). To this end, we first recall the following rational functions introduced [Li98] which yield majorants and minorants on a Heaviside step function similar to the polynomials in Proposition 4.4.6.

Proposition 4.4.21 (Theorem 4 and 5 in [Li98])**.** *Let $\nu_1 < \ldots < \nu_m \in \mathbb{R}$ with $\nu_1 > 0$ and let k be fixed with $1 \leq k < m$. Then there exist rational functions $\widehat{r}_{\{+,k\}}$ and $\widehat{r}_{\{-,k\}} \in \Pi_{2m-2}/\lambda^{m-1}$ which satisfy*

$$\widehat{r}_{\{\pm,k\}}(\nu_j) = \begin{cases} 1, & j = 1, \ldots, k, \\ 0, & j = k+1, \ldots, m, \end{cases} \tag{4.4.69}$$

Algorithm 4.6: A summary of the extended Lanczos recurrence in [DK98, Section 5]; an algorithm to compute the M-orthogonal basis U_m and the Rayleigh quotient $A_m = (U_m, A\,U_m)_{\mathsf{M}}$ of the extended Krylov subspace given in (4.4.67). Here, $m = 2\varrho - 1$.

run Algorithm 4.1 to compute $\beta_0 = \|u\|_{\mathsf{M}}$, U_ϱ^{SaI} and $A_\varrho^{\mathrm{SaI}} = (U_\varrho^{\mathrm{SaI}}, A\,U_\varrho^{\mathrm{SaI}})_{\mathsf{M}}$ for the SaI Krylov subspace $\mathcal{K}_\varrho(X, u)$ with $X = (A - sI)^{-1}$;
$\widetilde{v} = Au$;
orthogonalize $\widetilde{v}$ with U_ϱ^{SaI} and set $U_\varrho = U_\varrho^{\mathrm{SaI}}$ and $u_{\varrho+1} = \widetilde{v}/\|\widetilde{v}\|_{\mathsf{M}}$;
$\widehat{v} = A\,u_{\varrho+1}$;
for $j = 1, \ldots, \varrho$;
  $y_j = (u_j, \widehat{v})_{\mathsf{M}}$;
  $\widehat{v} \leftarrow \widehat{v} - y_j u_j$;
$a_1 = (u_{\varrho+1}, \widehat{v}\,)_{\mathsf{M}}$ and $\widehat{v} \leftarrow \widehat{v} - a_1 u_{k+1}$;
$\beta_1 = \|\widehat{v}\|_{\mathsf{M}}$ and $u_{\varrho+2} = \widehat{v}/\beta_1$;
consider $u_{\varrho+1}, u_{\varrho+2}$, and a_1 and β_1 to be the result of two initial Lanczos steps, and continue the Lanczos procedure to compute $u_{\varrho+3}, \ldots, u_{2\varrho-1}$ and the Jacobi matrix $J_{\varrho-1}$ (using a total of $\varrho - 1$ Lanczos steps);
$A_m = [\, A_\varrho^{\mathrm{SaI}}, y\,e_1^{\mathsf{H}};\ e_1\,y^{\mathsf{H}}, J_{\varrho-1}\,]$, where $y\,e_1^{\mathsf{H}} \in \mathbb{C}^{\varrho\times\varrho-1}$;
$x = \beta_0\,e_1$;
return x, U_m, A_m;

together with

$$\widehat{r}_{\{+,k\}}(\lambda) \geq \begin{cases} 1, & \lambda \leq \nu_k, \\ 0, & \lambda > \nu_k, \end{cases} \quad \textit{and} \quad \widehat{r}_{\{-,k\}}(\lambda) \leq \begin{cases} 1, & \lambda < \nu_{k+1}, \\ 0, & \lambda \geq \nu_{k+1}. \end{cases} \tag{4.4.70}$$

Additionally, the inequalities in (4.4.17) *are strict inequalities for* $\lambda \notin \{\nu_1, \ldots, \nu_m\}$.

In the proof of Proposition 4.4.22 below, we apply these results for the shifted case with $s \leq a < \lambda_1$.

We proceed to recapitulate [Li98, eq. (4) in Theorem 1]. For the following proposition, we recall that $\lambda_1 < \theta_1$ holds true when θ_j refers to the eigenvalues of the Rayleigh quotient A_m, thus, for a pole $s < \lambda_1$ the condition $s < \theta_1$ is satisfied.

Proposition 4.4.22 (Eq. (4) in Theorem 1 in [Li98])**.** *Let* $\theta_1, \ldots, \theta_m \in (a, b)$ *and* $c_1, \ldots, c_m \in \mathbb{C}$ *satisfy* (4.4.68) *for* $r \in \Pi_{2m-2}/(\lambda - s)^{m-1}$ *and a pole* $s < \lambda_1, \theta_1$*. Then,*

$$\alpha_n(\theta_k) < |c_1|^2 + \ldots + |c_k|^2 < \alpha_n(\theta_{k+1}-), \qquad k = 1, \ldots, m-1. \tag{4.4.71}$$

Proof. The proof of this proposition is similar to the proof of Theorem 4.4.5, and is also provided in [Li98]. We proceed with a sketch of the proof.

We first introduce $\nu_j = \theta_j - s$ for $j = 1, \ldots, m$. The nodes ν_j are positive due to $s < \theta_1$ and for a fixed $k = 1, \ldots, m-1$ we let $\widehat{r}_{\pm,k} \in \Pi_{2m-2}/\lambda^{m-1}$ denote the rational functions given in Proposition 4.4.21. Based on these rational functions, we consider the rational

functions $r_{\pm,k}(\lambda) = \widehat{r}_{\pm,k}(\lambda - s)$ in the class $\Pi_{2m-2}/(\lambda - s)^{m-1}$; and based on properties of $\widehat{r}_{\pm,k}$ given in Proposition 4.4.21 the functions $r_{\pm,k}$ satisfy

$$r_{\{\pm,k\}}(\theta_j) = \begin{cases} 1, & j = 1, \ldots, k, \\ 0, & j = k+1, \ldots, m, \end{cases} \tag{4.4.72a}$$

together with

$$r_{\{+,k\}}(\lambda) \geq \begin{cases} 1, & \lambda \leq \theta_k, \\ 0, & \lambda > \theta_k, \end{cases} \quad \text{and} \quad r_{\{-,k\}}(\lambda) \leq \begin{cases} 1, & \lambda < \theta_{k+1}, \\ 0, & \lambda \geq \theta_{k+1}. \end{cases} \tag{4.4.72b}$$

The identity (4.4.72a) implies

$$\sum_{j=1}^{m} |c_j|^2 r_{\{\pm,k\}}(\theta_j) = |c_1|^2 + \ldots + |c_k|^2, \tag{4.4.73a}$$

Analogously, the inequalities in (4.4.72b) imply

$$\alpha_n(\theta_k) = \sum_{\{j:\lambda_j \leq \theta_k\}} |w_j|^2 < \sum_{j=1}^{n} |w_j|^2 r_{\{+,k\}}(\lambda_j), \tag{4.4.73b}$$

and

$$\alpha_n(\theta_{k+1}-) = \sum_{\{j:\lambda_j < \theta_{k+1}\}} |w_j|^2 > \sum_{j=1}^{n} |w_j|^2 r_{\{-,k\}}(\lambda_j). \tag{4.4.73c}$$

The right-hand sides of (4.4.73b) and (4.4.73c) can be understood as a Riemann-Stieltjes integral (4.4.5a), for which the quadrature property (4.4.68) yields

$$\sum_{j=1}^{n} |w_j|^2 r_{\{\pm,k\}}(\lambda_j) = \sum_{j=1}^{m} |c_j|^2 r_{\{\pm,k\}}(\theta_j), \tag{4.4.73d}$$

Combining the identities and inequalities in (4.4.73), we conclude (4.4.71); for further details we also refer to the proof of Theorem 4.4.5. □

As previously discussed in Remark 4.4.9; in the Krylov setting the measure α_n is not continuous and a property as in [Li98, eq. (3) in Theorem 1] does not hold in general.

The spectrum of the Rayleigh quotient A_m for the extended Krylov subspace given in (4.4.67) defines a measure α_m, as in (4.4.24). The result of Proposition 4.4.22 can be understood as an intertwining property of the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$, similar as in the polynomial case in Subsection 4.4.1. This property is illustrated for a numerical example in Section 4.5 below.

Considering a rational qor-Krylov setting, previous results for the qor-Krylov representation B_m also apply to the extended Krylov subspace (which does correspond to the rational Krylov subspace $\mathcal{Q}_{2\varrho-1}(A, u)$ with denominator $q(\lambda) = (\lambda - s)^{\varrho-1}$ as previously mentioned). Thus, Proposition 4.4.22 holds true for the qor-Krylov representation B_m (assuming $s \leq a < \theta_1$). However, these results are not specified here.

4.5 Numerical illustrations

In the present section we verify the results of Theorem 4.4.5 (Subsection 4.4.2) and 4.4.11 (Subsection 4.4.3), and Proposition 4.4.19 (Subsection 4.4.4) and 4.4.22 (Subsection 4.4.5) by numerical experiments.

For the present numerical examples, the notation θ_j and c_j refers to the quadrature nodes and weights, respectively, satisfying different polynomial and rational Gaussian quadrature formulae which originate from polynomial Krylov subspaces $\mathcal{K}_m(A, u)$ and rational Krylov subspaces $\mathcal{Q}_m(A, u)$ with different choices of poles. Here, the matrix $A \in \mathbb{R}^{n\times n}$ corresponds to the finite-difference discretization of the negative 1D Laplace operator with $n = 1200$, and u is a random starting vector which is normalized. The M-inner product corresponds to the Euclidian inner product.

For the polynomial case, the quadrature nodes and weights are based on the spectrum of the Jacobi matrix J_m which is computed using the Lanczos method. Considering the rational case, we show results for SaI Krylov subspaces with real and complex shifts. For the case of a real shift, we consider the Rayleigh quotient A_m as in Algorithm 4.1 and the rational qor-Krylov representation B_m as in Algorithm 4.5. For a complex shift we show an example using the Rayleigh quotient A_m. Furthermore, we show results for an extended Krylov subspace, for which the Rayleigh quotient A_m is computed using Algorithm 4.6.

The step functions α_n (4.2.6a) and α_m (4.4.24) are illustrated for numerical examples in Figure 4.5. The step function α_m is shown for the polynomial Krylov subspace and a SaI Krylov subspace with a shift $s \in \mathbb{R}$ located outside of the convex hull of the matrix spectrum, namely, $s < \lambda_1$. In both cases the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ satisfy an intertwining property. To provide a clear illustration of the results of the previous section we also show the function $F(\lambda)$ given in (4.4.26) for a numerical example concerning the polynomial case in Figure 4.6 a). In this figure, we observe that $F(\lambda)$ changes its sign at the Ritz values, and following (4.4.28), this verifies the result of Theorem 4.4.5. For the SaI Krylov subspace with a shift $s < \lambda_1$ we have $F(s) = 0$ which implies $F_s(\lambda) = F(\lambda)$ for the function $F_s(\lambda)$ as given in (4.4.59). Considering this example, the function $F = F_s$ is illustrated in Figure 4.6 b), and following Remark 4.4.18, the change of the sign of F_s at rational Ritz values verifies Theorem 4.4.11.

The case of a SaI Krylov subspace with a shift $s \in \mathbb{R}$ such that $\theta_1 < s < \theta_m$ is illustrated in Figure 4.7. As for the previous example, the change of the sign of F_s at rational Ritz values verifies Theorem 4.4.11 as stated in Remark 4.4.18. Here, Figure 4.7 a) illustrates F_s for the Rayleigh quotient A_m and Figure 4.7 b) illustrates F_s for a rational qor-Krylov representation with a preassigned eigenvalue $\xi \in \mathbb{R}$; this verifies the result of Theorem 4.4.11 for these cases.

In Figure 4.8 we consider a SaI Krylov subspace with a complex shift $s \in \mathbb{C} \setminus \mathbb{R}$. For this example, we illustrate $|c_k|^2 + |c_{k+1}|^2$ for $k = 1, \ldots, m-1$ and $|c_m|^2 + |c_1|^2$, which yield upper bounds on $\mu_n([\theta_k, \theta_{k+1}])$ for $k = 1, \ldots, m-1$ and $\mu_n((-\infty, \theta_1] \cup [\theta_m, \infty))$, respectively. This verifies the result of Proposition 4.4.19.

For the extended Krylov subspace as in Subsection 4.4.5, Proposition 4.4.22 yields an intertwining property for the distributions $\mathrm{d}\alpha_n$ and $\mathrm{d}\alpha_m$ as in the polynomial case; the changing sign of F as illustrated in Figure 4.9 verifies the result of Proposition 4.4.22.

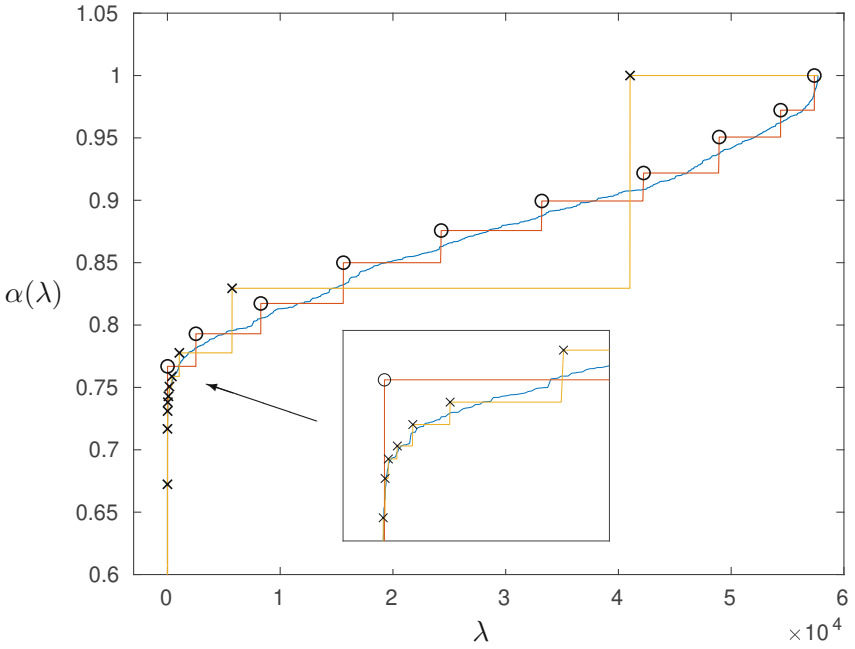


Figure 4.5: The matrix $A \in \mathbb{R}^{n\times n}$ is given by the finite-difference discretization of the negative 1D Laplace operator with $n = 1200$, and u is a random starting vector which is normalized. The continuous line without additional symbols illustrates the step function α_n associated with the eigenvalues and spectral coefficients of u in the eigenbasis of A. The symbols ('$\circ$') mark $\alpha_m(\theta_j)$ where θ_j are the Ritz values of the polynomial Krylov subspace $\mathcal{K}_m(A, u)$ with $m = 10$, and α_m is the respective step function given in (4.4.24). Similarly, the symbols ('$\times$') mark $\alpha_m(\theta_j)$ where θ_j refer to the rational Krylov subspace $\mathcal{Q}_m(A, u)$ with $m = 10$ and a single pole $s = -10^2$ of multiplicity $m - 1$.

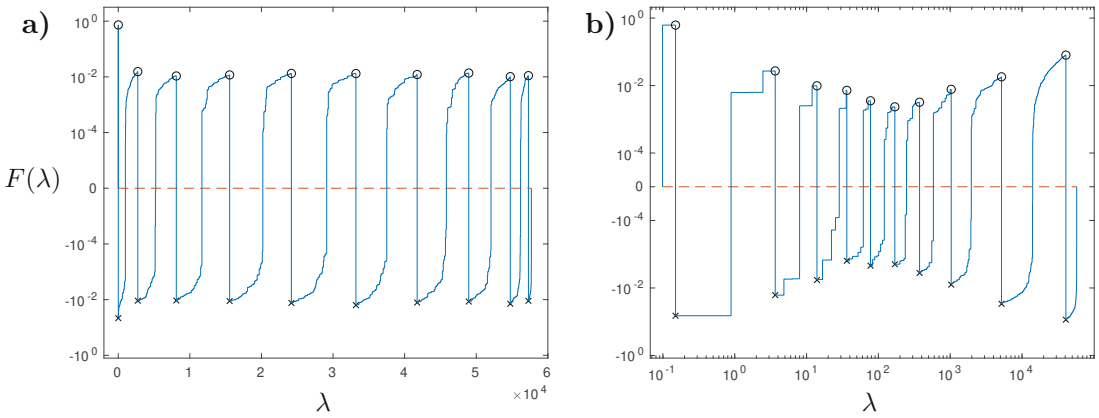


Figure 4.6: In Figure **a)** and **b)** the matrix $A \in \mathbb{R}^{n\times n}$ is given by the finite-difference discretization of the negative 1D Laplace operator with $n = 1200$. The starting vector u is chosen at random and is normalized. In these figures we show the function $F = \alpha_n - \alpha_m$ where α_m originates from different settings as stated below. The symbols ('$\circ$') and ('$\times$') mark $F(\theta_k-)$ and $F(\theta_k)$, respectively.
– Figure **a)** shows F with α_m given by spectral weights and Ritz values of the Jacobi matrix J_m for the polynomial Krylov subspace $\mathcal{K}_m(A, u)$ with $m = 10$. The y-axis is scaled logarithmically in positive and negative direction, namely, with range $(-10^0, -10^{-6}) \cup (10^{-6}, 10^0)$.
– Figure **b)** shows F where α_m refers to the spectrum of the Rayleigh quotient A_m for the rational Krylov subspace $\mathcal{Q}_m(A, u)$ with $m = 10$ and a single pole $s = -10^2$ of multiplicity $m-1$, thus, $s < \lambda_1$. Similar to Figure a) the y-axis is scaled logarithmically and covers $(-10^0, -10^{-5}) \cup (10^{-5}, 10^0)$. Additionally, the x-axis is scaled logarithmically in a classical sense.

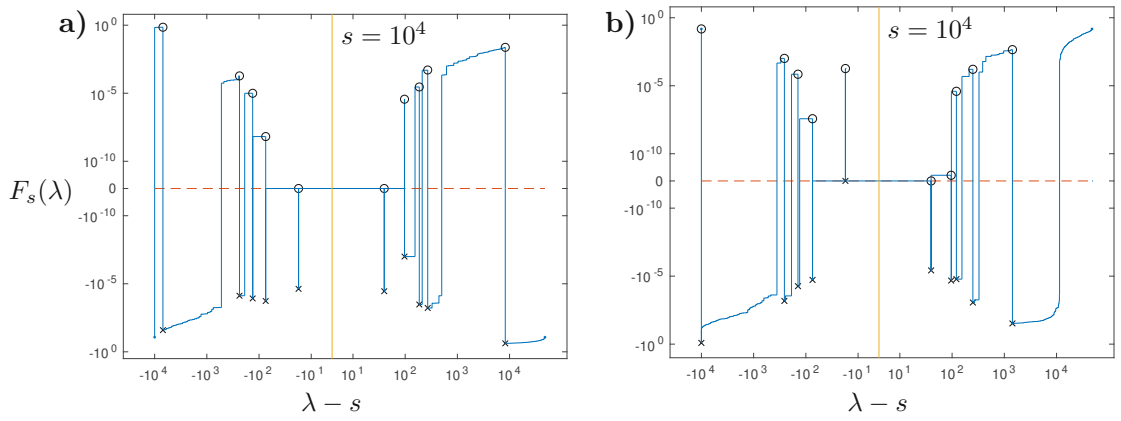


Figure 4.7: In Figure **a)** and **b)** the matrix $A \in \mathbb{R}^{n\times n}$ is given by the finite-difference discretization of the negative 1D Laplace operator with $n = 1200$. The starting vector u is chosen at random and is normalized. These figures show $F_s(\lambda) = F(\lambda) - F(s)$ for different settings, and the symbols ('$\circ$') and ('$\times$') mark $F_s(\theta_k-)$ and $F_s(\theta_k)$, respectively. Similar to Figure 4.6 the y-axis is scaled logarithmically and covers $(-10^0, -10^{-12}) \cup (10^{-12}, 10^0)$. With λ being the argument of the function $F_s(\lambda)$ as illustrated in the y-axis, the x-axis shows $\lambda - s$, i.e. the distance from the argument λ to the pole $s = 10^4$. Furthermore, the x-axis is scaled logarithmically with a range of approximately $(-10^4, -10^1) \cup (10^1, 10^4)$.
– Figure **a)** shows $F_s(\lambda)$ corresponding to the spectrum of A_m, where A_m is the Rayleigh quotient in the rational Krylov subspace $\mathcal{Q}_m(A, u)$ with $m = 10$ and a single pole $s = 10^4$ of multiplicity $m - 1$. Here, the pole s is enclosed by the eigenvalues of A_m.
– Figure **b)** shows $F_s(\lambda)$ where α_m corresponds to the spectrum of B_m, which is the rational qor-Krylov representation for which the eigenvalue $\theta_1 = -10$ is preassigned. For the underlying rational Krylov subspace $\mathcal{Q}_m(A, u)$ we have $m = 10$ and a single pole $s = 10^4$ of multiplicity $m - 1$. The pole s is enclosed by the eigenvalues of B_m.

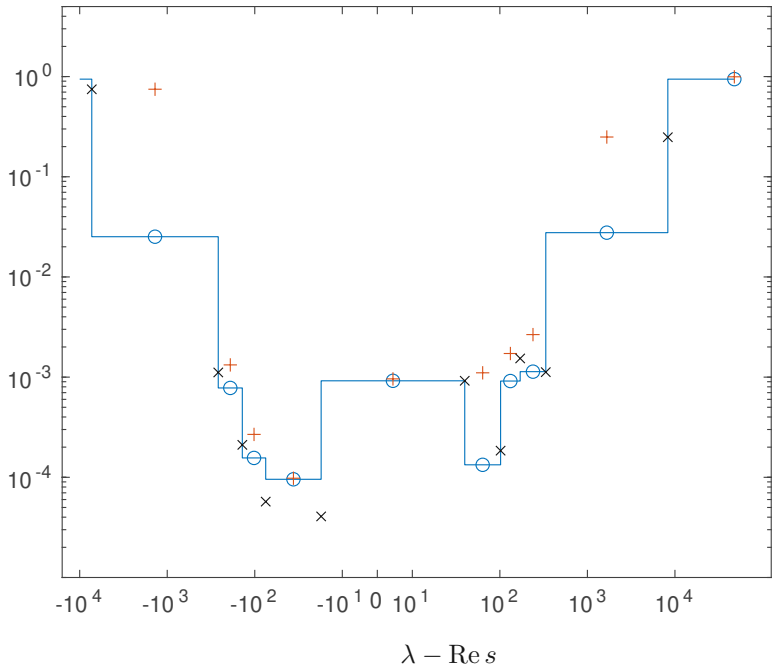


Figure 4.8: The matrix $A \in \mathbb{R}^{n\times n}$ is given by the finite-difference discretization of the negative 1D Laplace operator with $n = 1200$. The starting vector u is chosen at random and is normalized. For the present figure we consider the SaI Krylov subspace $\mathcal{Q}_m(A, u)$ with $m = 10$ and a complex shift $s = 10^4 - 10^2\mathrm{i}$. In the present caption, c_j and θ_k refer to the entries of eigenvectors and the eigenvalues of the respective Rayleigh quotient A_m. The symbols ('×') mark $|c_j|^2$ over θ_j. The symbols ('+') show $|c_k|^2 + |c_{k+1}|^2$ over the midpoint of the interval $[\theta_k, \theta_{k+1}]$ for $k = 1, \ldots, m-1$. Furthermore, the symbol ('+') located at the right boundary of the spectrum shows $|c_m|^2 + |c_1|^2$. The line marked by ('∘') shows the measure $\mu_n([\theta_k, \theta_{k+1}])$ over each interval $[\theta_k, \theta_{k+1}]$ for $k = 1, \ldots, m-1$, and the measure $\mu_n((-\infty, \theta_1] \cup [\theta_m, \infty))$ at the boundary. The y-axis is scaled logarithmically in a classical sense, and the x-axis shows $\lambda - \operatorname{Re} s$, i.e. the distance from the argument λ to the real part of the shift, i.e., $\operatorname{Re} s = 10^4$. Furthermore, the x-axis is scaled logarithmically with a range of approximately $(-10^4, -10^1) \cup (10^1, 10^4)$.

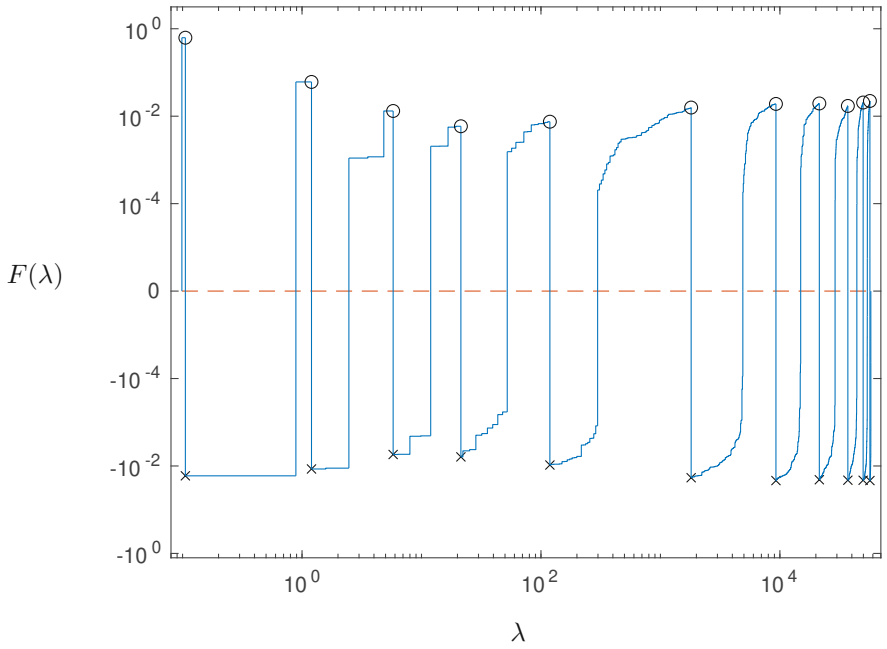


Figure 4.9: The matrix $A \in \mathbb{R}^{n\times n}$ is given by the finite-difference discretization of the negative 1D Laplace operator with $n = 1200$. The starting vector u is chosen at random and is normalized. In this figure we show the function $F = \alpha_n - \alpha_m$ where α_m corresponds to the spectrum of the Rayleigh quotient A_m given by the extended Krylov subspace (4.4.67) with $m = 11$ (thus, $\varrho = 6$) and the shift $s = -10 < \lambda_1$. The symbols ('$\circ$') and ('$\times$') mark $F(\theta_k-)$ and $F(\theta_k)$, respectively, where θ_k refers to the eigenvalues of A_m. The y-axis is scaled logarithmically in positive and negative direction, namely, with range $(-10^0, -10^{-6}) \cup (10^{-6}, 10^0)$. Additionally, the x-axis is scaled logarithmically in a classical sense.

Appendix

4.A Some properties of Krylov subspaces

Proposition 4.A.1. *Let $(q_1, \dots, q_n) \in \mathbb{C}^{n\times n}$ be an orthogonal eigenbasis of the matrix $A \in \mathbb{C}^{n\times n}$. Here, orthogonal is to be understood w.r.t. a given positive definite inner product. Let $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ be the corresponding eigenvalues, and $w_j = (q_j, u)_{\mathsf{M}} \in \mathbb{C}$ be the spectral coefficients of a given vector u. Then*

$$\operatorname{rank}\{u, Au, \dots, A^{m-1}u\} = m, \tag{4.A.1}$$

if and only if there exist at least m coefficients $w_j \neq 0$ with distinct λ_j.

Proof. According to the eigendecomposition of A we have

$$A^\ell u = \sum_{j=1}^{n} \lambda_j^\ell w_j q_j \quad \text{for } \ell \in \mathbb{N}_0.$$

The matrix corresponding to the left-hand side of (4.A.1) takes the form of a Vandermonde matrix,

$$(u, Au, \dots, A^{m-1}u) = (q_1 w_1, q_2 w_2, \dots, q_n w_n) \begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{m-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{m-1} \end{pmatrix} \in \mathbb{C}^{n\times m}. \tag{4.A.2}$$

Let $n_1 \leq n$ be the number of nonzero coefficients w_j, thus, there exist indices $j(1), \dots, j(n_1)$ with $w_{j(1)}, \dots, w_{j(n_1)} \neq 0$. We define

$$\Theta_1 = \left(q_{j(1)} w_{j(1)}, q_{j(2)} w_{j(2)}, \dots, q_{j(n_1)} w_{j(n_1)}\right) \in \mathbb{C}^{n\times n_1}.$$

The orthogonality properties of $q_1, \dots, q_n$ imply $\operatorname{rank}(\Theta_1) = n_1$. For the corresponding rows of the Vandermonde matrix we introduce the notation

$$\Theta_2 = \begin{pmatrix} 1 & \lambda_{j(1)} & \lambda_{j(1)}^2 & \cdots & \lambda_{j(1)}^{m-1} \\ 1 & \lambda_{j(2)} & \lambda_{j(2)}^2 & \cdots & \lambda_{j(2)}^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_{j(n_1)} & \lambda_{j(n_1)}^2 & \cdots & \lambda_{j(n_1)}^{m-1} \end{pmatrix} \in \mathbb{C}^{n_1\times m}.$$

The identity in (4.A.2) can now be written as

$$\left(u, Au, \dots, A^{m-1}u\right) = \Theta_1 \Theta_2. \tag{4.A.3}$$

With $\Theta_1 \in \mathbb{C}^{n\times n_1}$ and $\operatorname{rank}(\Theta_1) = n_1$ we have $\operatorname{rank}(\Theta_1\Theta_2) = \operatorname{rank}(\Theta_2)$. Let $n_2 \leq n_1$ be the number of distinct eigenvalues within $\lambda_{j(1)}, \dots, \lambda_{j(n_1)}$, hence, we have indices $\ell(1), \dots, \ell(n_2)$

for which $\lambda_{j(\ell(1))}, \ldots, \lambda_{j(\ell(n_2))}$ are distinct. Then the Vandermonde matrix Θ_2 satisfies $\operatorname{rank}(\Theta_2) = \min\{m, n_2\}$, hence,

$$\operatorname{rank}(\Theta_1\Theta_2) = \operatorname{rank}(\Theta_2) = \min\{m, n_2\}. \tag{4.A.4}$$

Combining (4.A.3) with (4.A.4) we conclude

$$\operatorname{rank}\{u, Au, \ldots, A^{m-1}u\} = \min\{m, n_2\}.$$

We recall that n_2 is number of nonzero coefficients w_j with distinct λ_j, and (4.A.1) holds if and only if $n_2 \geq m$ which completes the proof. □

Proposition 4.A.2. *Let $w_1, \ldots, w_n \in \mathbb{C}$ with $w_j \neq 0$ and $\lambda_1 < \ldots < \lambda_n$ be given. Let $m < n$, and let $\theta_1 < \ldots < \theta_m$ and $|c_1|^2, \ldots, |c_m|^2$ denote quadrature nodes and quadrature weights, respectively, and assume*

$$\sum_{j=1}^{n} |w_j|^2 p(\lambda_j) = \sum_{j=1}^{m} |c_j|^2 p(\theta_j), \quad p \in \Pi_{2m-2}. \tag{4.A.5}$$

Then $c_j \neq 0$ for $j = 1, \ldots, m$.

Proof. We define $g_\ell(\lambda) = \prod_{j=1, j\neq\ell}^{m} (\lambda - \theta_j)^2 \in \Pi_{2m-2}$. The polynomial g_ℓ is zero only at the nodes $\theta_1, \ldots, \theta_{\ell-1}, \theta_{\ell+1}, \ldots, m$ and positive otherwise. Due to $n > m$ at least one λ_j is distinct to $\theta_1, \ldots, \theta_m$ and this yields

$$\sum_{j=1}^{n} |w_j|^2 g_\ell(\lambda_j) > 0.$$

Making use of the identity (4.A.5) and evaluating the right-hand side therein we conclude

$$\sum_{j=1}^{m} |c_j|^2 g_\ell(\theta_j) = g_\ell(\theta_\ell)|c_\ell|^2 > 0.$$

With $g_\ell(\theta_\ell) > 0$ this concludes $|c_\ell|^2 > 0$. □

Proposition 4.A.3 (*Identities for rational functions in the rational Krylov subspace*)**.** *Let $U_m \in \mathbb{C}^{n\times m}$ with $(U_m, U_m)_{\mathsf{M}} = I$ and $\operatorname{span}\{U_m\} = \mathcal{Q}_m(A, u)$ for the rational Krylov subspace $\mathcal{Q}_m(A, u)$ with denominator q_{m-1}. Let $A_m = (U_m, AU_m)_{\mathsf{M}}$ and $x = (U_m, u)_{\mathsf{M}}$.*

(i) The following identities hold true,

$$r(A)u = U_m\, r(A_m)\, x, \quad r \in \Pi_{m-1}/q_{m-1}.$$

(ii) Let $r = p/q_{m-1}$ with $p \in \Pi_m$ being a polynomial of degree exactly m, then

$$(U_m\, r(A_m)x - r(A)u) \perp_{\mathsf{M}} \operatorname{span}\{U_m\}. \tag{4.A.6}$$

Proof. We proceed similar to the proof of Proposition 4.3.5 in Subsection 4.3.2. We recall the identity $\mathcal{Q}_m(A,u) = \mathcal{K}_m(A,u_q)$ with $u_q = q_{m-1}^{-1}(A)u$. Let $\zeta_0 = \|u_q\|_{\mathsf{M}}$, let V_m be the M-orthonormal basis of $\mathcal{K}_m(A,u_q)$, and let J_m be the respective Jacobi matrix. Then the identity (4.2.15a) w.r.t. $\mathcal{K}_m(A,u_q)$ implies

$$p(A)u_q = \zeta_0\, V_m\, p(J_m)e_1, \quad p \in \Pi_{m-1}. \tag{4.A.7}$$

This implies $q_{m-1}(A)u_q = \zeta_0\, V_m\, q_{m-1}(J_m)e_1$, and with the identities $q_{m-1}(A)u_q = u$ and $(V_m, V_m)_{\mathsf{M}} = I$ we arrive at

$$\zeta_0\, e_1 = q_{m-1}^{-1}(J_m)(V_m,u)_{\mathsf{M}}. \tag{4.A.8}$$

Let $r = p/q_{m-1}$ with $p \in \Pi_{m-1}$ then $r(A)u = p(A)u_q$, and with (4.A.7) we have

$$r(A)u = \zeta_0\, V_m\, p(J_m)e_1. \tag{4.A.9}$$

Inserting (4.A.8) into (4.A.9) gives

$$r(A)u = V_m\, p(J_m) q_{m-1}^{-1}(J_m)(V_m,u)_{\mathsf{M}} = V_m\, r(J_m)(V_m,u)_{\mathsf{M}}. \tag{4.A.10}$$

With the identity $K_m\, K_m^{\mathsf{H}} = I$ (see (4.2.23c)) and (4.2.24) the matrix A_m satisfies $r(J_m) = K_m\, r(A_m) K_m^{\mathsf{H}}$, and together with $V_m K_m = U_m$ (4.2.23c) we have

$$V_m\, r(J_m)(V_m,u)_{\mathsf{M}} = U_m\, r(A_m)(U_m,u)_{\mathsf{M}}. \tag{4.A.11}$$

Combining (4.A.10) with (4.A.11) completes the proof of (i).

For a polynomial p of degree exactly m and w.r.t. $\mathcal{K}_m(A,u_q)$ the property (4.2.15b) writes

$$p(A)u_q - \zeta_0 V_m\, p(J_m)e_1 \perp_{\mathsf{M}} \operatorname{span}\{V_m\}, \quad p \in \Pi_m. \tag{4.A.12}$$

Let $r = p/q_{m-1}$, then the identities in (4.A.8) and (4.A.11) with $x = (U_m,u)_{\mathsf{M}}$ entail

$$\zeta_0 V_m\, p(J_m)e_1 = V_m\, r(J_m)(V_m,u)_{\mathsf{M}} = U_m\, r(A_m)x.$$

With $r(A)u = p(A)u_q$ this yields

$$p(A)u_q - \zeta_0 V_m\, p(J_m)e_1 = r(A)u - U_m\, r(A_m)x. \tag{4.A.13}$$

Making use of $\operatorname{span}\{V_m\} = \operatorname{span}\{U_m\}$ in (4.A.12) and substituting (4.A.13), we conclude (4.A.6). □

Proposition 4.A.4 (*The spectral coefficients c_j for the rational Krylov subspace and the choice of U_m). The spectral coefficients c_j of $x = (U_m,u)_{\mathsf{M}} \in \mathbb{C}^m$ in the orthonormal eigenbasis of $A_m = (U_m, A\,U_m)_{\mathsf{M}} \in \mathbb{C}^{m\times m}$ are independent of the explicit choice of the underlying orthonormal basis U_m of $\mathcal{Q}_m(A,u)$.*

Proof. We recall the representation of the spectral coefficients c_j given in (4.4.9),

$$c_j = (\widehat{q}_j, x)_2, \quad j = 1,\dots,m.$$

Here $\widehat{q}_j \in \mathbb{C}^m$ refer to the orthonormal eigenvectors of A_m. We further recall that the rational Krylov subspace $\mathcal{Q}_m(A,u)$ corresponds to the polynomial Krylov subspace $\mathcal{K}_m(A,u_q)$ with $u_q = q_{m-1}^{-1}(A)u$ for the denominator q_{m-1}. Let us recall the notation J_m and V_m for the Jacobi matrix and Krylov basis of $\mathcal{K}_m(A,u_q)$. Furthermore, we recall the orthonormal transformation $K_m = (V_m, U_m)_{\mathsf{M}} \in \mathbb{C}^{m\times m}$ given in (4.2.23a). With $U_m = V_m K_m$ (4.2.23b) the vector $x = (U_m, u)_{\mathsf{M}}$ corresponds to

$$x = K_m^{\mathsf{H}}(V_m, u)_{\mathsf{M}} =: K_m^{\mathsf{H}}\xi, \quad \text{and thus,} \quad c_j = (K_m\widehat{q}_j, \xi)_2. \tag{4.A.14}$$

With the identity $A_m = K_m^{\mathsf{H}} J_m K_m$ (4.2.24) and $\widehat{q}_j$ being eigenvectors of A_m, the vectors $K_m\widehat{q}_j$ for $j = 1,\dots,m$ correspond to the orthonormal eigenvectors of J_m. Thus, (4.A.14) implies that the coefficients c_j correspond to spectral coefficients of $\xi = (V_m,u)_{\mathsf{M}}$ in the orthonormal eigenbasis of J_m, and furthermore, the coefficients c_j are independent of the explicit choice of U_m. □

4.B Auxiliary functions for the CMS Theorem

Proof of ***Proposition 4.4.6****.* We recapitulate arguments of [Sze85, Akh65] and others. Let $\theta_1 < \ldots < \theta_m$ and $k \in \{1,\dots,m-1\}$ be given. We first prove the existence of a polynomial $p_{\{+,k\}}$ of degree $2m-2$ which satisfies (4.4.16) and (4.4.17). Let p be a polynomial of degree $2m-2$ subject to the conditions

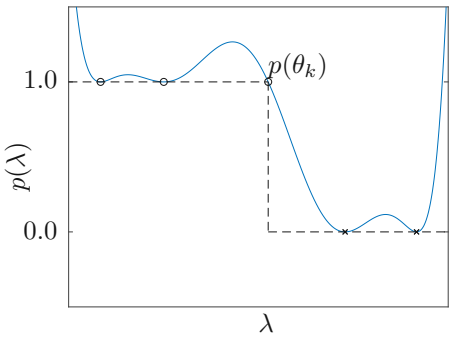


Figure 4.10: A numerical illustration of the polynomial p subject to the conditions (4.B.1); duplicated from Figure 4.2.

$$\begin{aligned}
&p(\theta_1) = 1, && p'(\theta_1) = 0,\\
&\quad\vdots && \quad\vdots\\
&p(\theta_{k-1}) = 1, && p'(\theta_{k-1}) = 0,\\
&p(\theta_k) = 1, &&\\
&p(\theta_{k+1}) = 0, && p'(\theta_{k+1}) = 0,\\
&\quad\vdots && \quad\vdots\\
&p(\theta_m) = 0, && p'(\theta_m) = 0.
\end{aligned} \tag{4.B.1}$$

Thus, we have m many conditions for the polynomial p and $m-1$ many conditions for its derivative, and such a polynomial uniquely exists. By the conditions (4.B.1) the polynomial p satisfies the identities (4.4.16).

To show that p satisfies the inequalities (4.4.17) considering $p_{\{+,k\}}$, we proceed to locate the zeros of p' which correspond to points of extreme values of p: The derivative p' is a polynomial of degree $2m-3$, and thus, has $2m-3$ zeros. By the conditions (4.B.1), we have $m-1$ many zeros of p' located at nodes. For each pair of neighboring nodes in $\{\theta_1,\dots,\theta_k\}$ and $\{\theta_{k+1},\dots,\theta_m\}$ the conditions (4.B.1) and Rolle's Theorem imply the existence of a zero of p' between the respective nodes. Thus, the derivative p' has $m-1$ many simple zeros located at nodes and $m-2$ many simple zeros located between nodes.

With $p(\theta_k) > p(\theta_{k+1})$ and with the respective changes of sign for p' we conclude that p satisfies the inequalities for $p_{\{+,k\}}$ in (4.4.17).

Furthermore, we have $p(\lambda) > 1$ for $\lambda \in (\theta_j, \theta_{j+1})$ with $j = 1, \dots, k-1$ and $\lambda < \theta_1$, and we have $p(\lambda) > 0$ for $\lambda \in (\theta_j, \theta_{j+1})$ with $j = k, \dots, m$ and $\lambda > \theta_m$. Thus, the inequalities for $p_{\{+,k\}}$ in (4.4.17) are strict for $\lambda \notin \{\theta_1, \dots, \theta_m\}$.

In a similar manner we conclude results for $p_{\{-,k\}}$.

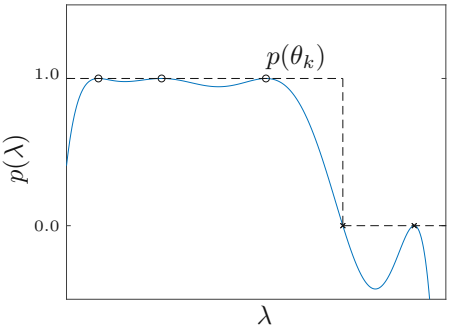


Figure 4.11: A numerical illustration of the polynomial p subject to the conditions (4.B.2); duplicated from Figure 4.2.

Let p be a polynomial of degree $2m-2$ subject to the conditions

$$\begin{aligned} &p(\theta_1) = 1, && p'(\theta_1) = 0, \\ &\quad\vdots && \quad\vdots \\ &p(\theta_k) = 1, && p'(\theta_k) = 0, \\ &p(\theta_{k+1}) = 0, && \\ &p(\theta_{k+2}) = 0, && p'(\theta_{k+2}) = 0, \\ &\quad\vdots && \quad\vdots \\ &p(\theta_m) = 0, && p'(\theta_m) = 0. \end{aligned} \tag{4.B.2}$$

Then similar arguments as previously show that such a polynomial p satisfies the identities (4.4.16) and inequalities (4.4.17) associated with $p_{\{-,k\}}$.

Thus, the polynomials subject to the conditions (4.B.1) and (4.B.2) satisfy the desired properties of $p_{\{+,k\}}$ and $p_{\{-,k\}}$, respectively, which completes the proof. □

Proof of ***Proposition 4.4.12****.* Let $a = -\infty$ and $b = \infty$ to simplify the notation.

For the given pole $s \in \mathbb{R}$ we define the transformation

$$x\colon \mathbb{R} \setminus \{s\} \to \mathbb{R} \setminus \{0\}, \qquad x(\lambda) := (\lambda - s)^{-1}. \tag{4.B.3}$$

For the case $\theta_1 < s < \theta_m$ the indices $k_1 > 1$ and $k_m = k_1 - 1$ are given in (4.4.33a) and the values $x(\theta_j)$ satisfy the ordering

$$x(\theta_{k_m}) < x(\theta_{k_m-1}) < \ldots < x(\theta_1) < 0 < x(\theta_m) < x(\theta_{m-1}) < \ldots < x(\theta_{k_1}). \tag{4.B.4a}$$

Otherwise, for $s < \theta_1$ (and $s > \theta_m$) we have $k_1 = 1$ and $k_m = m$ as in (4.4.33b), and

$$x(\theta_{k_m}) < \ldots < x(\theta_{k_1}) < 0, \quad s < \theta_1 \qquad \big(0 < x(\theta_{k_m}) < \ldots < x(\theta_{k_1}), \quad s > \theta_m\big). \tag{4.B.4b}$$

For the index k_m we recall and highlight

$$k_m = k_1 - 1, \quad \text{for } \theta_1 < s < \theta_m, \quad \text{and} \quad k_m = m, \quad \text{otherwise.} \tag{4.B.5}$$

For any of these cases we define the index mapping

$$\iota\colon \{1, \dots, m\} \to \{1, \dots, m\}, \quad \iota(j) := \begin{cases} k_1 - j, & 1 \le j < k_1, \\ m + k_1 - j, & k_1 \le j \le m. \end{cases}$$

The action of ι is illustrated in the following table,

j	1	2	$\cdots$	k_1-2	k_1-1	k_1	k_1+1	$\cdots$	$m-1$	m
$\iota(j)$	k_1-1	k_1-2	$\cdots$	2	1	m	$m-1$	$\cdots$	k_1+1	k_1

, (4.B.6)

where $k_m = k_1 - 1$ or $k_m = m$ as specified in (4.B.5). (Thus, in the case of $s < \theta_1$ or $s > \theta_m$ this gives

j	$1 = k_1$	2	$\cdots$	$m-1$	$m = k_m$
$\iota(j)$	m	$m-1$	$\cdots$	2	1

.) (4.B.7)

We remark that ι is involutory with $\iota(1) = k_m$ ($\iota(k_m) = 1$), and $\iota(m) = k_1$ ($\iota(k_1) = m$). Thus, this mapping is bijective and with $\iota(k_m) = 1$ we have

$$\iota(k) - 1 \in \{1, \ldots, m-1\}, \quad \text{for} \quad k \in \{1, \ldots, m\} \setminus \{k_m\}. \tag{4.B.8}$$

Let $\xi_1, \ldots, \xi_m$ denote the sequence of $x(\theta_j)$ arranged as in (4.B.4), i.e.,

$$\xi_j := x(\theta_{\iota(j)}), \quad \text{thus,} \quad \xi_1 = x(\theta_{k_m}) < \ldots < \xi_m = x(\theta_{k_1}).$$

We remark that ι being involutory implies

$$\xi_{\iota(j)} = x(\theta_j). \tag{4.B.9}$$

We recall the definition of the index set I_k given in (4.4.34), i.e.,

$$I_k = \begin{cases} \{1, \ldots, k, k_1, \ldots, m\}, & 1 \le k < k_1, \\ \{k_1, \ldots, k\}, & k_1 \le k \le m. \end{cases} \tag{4.B.10}$$

The set $\{x(\theta_j) \colon j \in I_k\}$ can be rewritten as follows: With $\xi_{\iota(j)} = x(\theta_j)$ (4.B.9) we have

$$\{x(\theta_j) \colon j \in I_k\} = \begin{cases} \{\xi_{\iota(1)}, \ldots, \xi_{\iota(k)}\} \cup \{\xi_{\iota(k_1)}, \ldots, \xi_{\iota(m)}\}, & 1 \le k < k_1, \quad \text{and} \\ \{\xi_{\iota(k_1)}, \ldots, \xi_{\iota(k)}\}, & k_1 \le k \le m. \end{cases} \tag{4.B.11}$$

We proceed to rewrite the indices of the sets on the right-hand side of this equation using (4.B.6). In particularly, the identities $\iota(1) = k_1 - 1$, $\iota(k_1) = m$ and $\iota(m) = k_1$ imply

$$\begin{aligned} (\iota(1), \ldots, \iota(k)) &= (k_1 - 1, k_1 - 2, \ldots, \iota(k)) \quad \text{for } k < k_1 \text{ and} \\ (\iota(k_1), \ldots, \iota(m)) &= (m, m-1, \ldots, k_1). \end{aligned}$$

Thus,

$$\{\xi_{\iota(1)}, \ldots, \xi_{\iota(k)}\} \cup \{\xi_{\iota(k_1)}, \ldots, \xi_{\iota(m)}\} = \{\xi_{\iota(k)}, \xi_{m-1}, \ldots, \xi_m\}, \quad 1 \le k < k_1. \tag{4.B.12a}$$

In a similar manner, identity (4.B.6) yields

$$(\iota(k_1), \ldots, \iota(k)) = (m, m-1, \ldots, \iota(k)), \quad \text{for} \quad k \ge k_1,$$

which implies

$$\{\xi_{\iota(k_1)}, \ldots, \xi_{\iota(k)}\} = \{\xi_{\iota(k)}, \ldots, \xi_m\}, \quad k_1 \le k \le m. \tag{4.B.12b}$$

The identity (4.B.11) together with (4.B.12), under consideration of the different cases for k, show

$$\{x(\theta_j)\colon j \in I_k\} = \{\xi_{\iota(k)}, \dots, \xi_m\}, \quad k = 1, \dots, m. \tag{4.B.13}$$

In a similar manner the set $R_k \subset \mathbb{R}$ in (4.4.34) for $k = 1, \dots, m$ satisfies

$$x(R_k) = [\xi_{\iota(k)}, +\infty) \setminus \{0\}, \quad \text{and} \quad x(\mathbb{R}_s \setminus R_k) = (-\infty, \xi_{\iota(k)}) \setminus \{0\}, \tag{4.B.14a}$$

where $\mathbb{R}_s = \mathbb{R}\setminus\{s\}$. The first identity in (4.B.14a) is illustrated in Figure 4.12. Analogously, the interior of R_k satisfies

$$x(R_k^{\circ}) = (\xi_{\iota(k)}, +\infty) \setminus \{0\}, \quad \text{and} \quad x(\mathbb{R}_s \setminus R_k^{\circ}) = (-\infty, \xi_{\iota(k)}] \setminus \{0\}. \tag{4.B.14b}$$

In the current setting we assume k satisfies $k \in \{1, \dots, m\} \setminus \{k_m\}$, thus, with (4.B.8) we have $\iota(k) - 1 \in \{1, \dots, m-1\}$. For the sequence $\xi_1 < \dots < \xi_m$ and the index $\iota(k) - 1$ we let $p_{\{+,\iota(k)-1\}}$ and $p_{\{-,\iota(k)-1\}}$ refer to the polynomials introduced in Proposition 4.4.6. Additionally, we define $g_{\{\pm,k\}}$ by

$$g_{\{\pm,k\}}(y) := 1 - p_{\{\mp,\iota(k)-1\}}(y), \qquad k \in \{1, \dots, m\} \setminus \{k_m\}. \tag{4.B.15}$$

The identities (4.4.16) for $p_{\{\pm,\iota(k)-1\}}$ write

$$p_{\{\pm,\iota(k)-1\}}(\xi_j) = \begin{cases} 1, & j = 1, \dots, \iota(k)-1, \\ 0, & j = \iota(k), \dots, m, \end{cases}$$

and this entails the following identities for $g_{\{\pm,k\}}$,

$$g_{\{\pm,k\}}(\xi_j) = \begin{cases} 0, & j = 1, \dots, \iota(k)-1, \\ 1, & j = \iota(k), \dots, m. \end{cases}$$

With (4.B.13) this conforms to the following identities for the nodes θ_j,

$$g_{\{\pm,k\}}(x(\theta_j)) = \begin{cases} 1, & j \in I_k, \\ 0, & \text{otherwise.} \end{cases} \tag{4.B.16}$$

In a similar manner the inequalities (4.4.17) for $p_{\{\pm,\iota(k)-1\}}$ read

$$p_{\{+,\iota(k)-1\}}(y) \geq \begin{cases} 1, & y \leq \xi_{\iota(k)-1}, \\ 0, & \xi_{\iota(k)-1} < y, \end{cases} \quad \text{and} \quad p_{\{-,\iota(k)-1\}}(y) \leq \begin{cases} 1, & y < \xi_{\iota(k)}, \\ 0, & \xi_{\iota(k)} \leq y, \end{cases} \tag{4.B.17}$$

and this entails

$$g_{\{+,k\}}(y) \geq \begin{cases} 0, & y < \xi_{\iota(k)}, \\ 1, & \xi_{\iota(k)} \leq y, \end{cases} \quad \text{and} \quad g_{\{-,k\}}(y) \leq \begin{cases} 0, & y \leq \xi_{\iota(k)-1}, \\ 1, & \xi_{\iota(k)-1} < y. \end{cases} \tag{4.B.18}$$

With (4.B.14a) the inequalities (4.B.18) for $g_{\{+,k\}}$ yield inequalities on the domain of x,

$$g_{\{+,k\}}(x(\lambda)) \geq \begin{cases} 0, & \lambda \in \mathbb{R}_s \setminus R_k, \\ 1, & \lambda \in R_k, \end{cases} \tag{4.B.19}$$

To rewrite the inequalities (4.B.18) for $g_{\{-,k\}}$ we proceed in a similar manner: We first consider the cases $s < \theta_1$ and $s > \theta_m$. For these cases the action of the mapping ι is illustrated in (4.B.7) and we observe

$$\iota(k) - 1 = \iota(k+1), \quad k = 1, \dots, m-1, \quad \text{and } s < \theta_1 \text{ or } s > \theta_m.$$

Thus, we have $\xi_{\iota(k)-1} = \xi_{\iota(k+1)}$ for these cases and the identities (4.B.14b) imply

$$x(R^{\mathrm{o}}_{k+1}) = (\xi_{\iota(k)-1}, +\infty) \setminus \{0\}, \quad \text{and} \quad x(\mathbb{R}_s \setminus R^{\mathrm{o}}_{k+1}) = (-\infty, \xi_{\iota(k)-1}] \setminus \{0\}.$$

Together with the inequalities for $g_{\{-,k\}}$ in (4.B.18), this shows the following inequalities in the domain of x,

$$g_{\{-,k\}}(x(\lambda)) \leq \begin{cases} 0, & \lambda \in \mathbb{R}_s \setminus R^{\mathrm{o}}_{k+1}, \\ 1, & \lambda \in R^{\mathrm{o}}_{k+1}, \end{cases} \qquad \begin{array}{l} \text{for } s < \lambda_1 \text{ or } s > \lambda_m, \text{ and} \\ k = 1, \dots, m-1. \end{array} \tag{4.B.20}$$

Similar results hold for the case $\theta_1 < s < \theta_m$ (thus, $k_m < m$): The illustration in (4.B.6) reveals

$$\iota(k) - 1 = \iota(k+1), \quad k \in \{1, \dots, m-1\} \setminus \{k_m\}, \quad \text{and} \quad \iota(m) - 1 = \iota(1).$$

Thus, with (4.B.14b) and the denotation $R_{m+1} = R_1$ we have

$$x(R^{\mathrm{o}}_{k+1}) = (\xi_{\iota(k)-1}, +\infty) \setminus \{0\}, \quad k \in \{1, \dots, m\} \setminus \{k_m\},$$

with similar results considering $x(\mathbb{R}_s \setminus R^{\mathrm{o}}_{k+1})$. With this identity, the inequalities for $g_{\{-,k\}}$ in (4.B.18) reveal inequalities similar to (4.B.20) for the case $\theta_1 < s < \theta_m$. Together with (4.B.20) for the case $s < \lambda_1$ or $s > \lambda_m$, we conclude with

$$g_{\{-,k\}}(x(\lambda)) \leq \begin{cases} 0, & \lambda \in \mathbb{R}_s \setminus R^{\mathrm{o}}_{k+1}, \\ 1, & \lambda \in R^{\mathrm{o}}_{k+1}, \end{cases} \qquad k \in \{1, \dots, m\} \setminus \{k_m\}, \tag{4.B.21}$$

We define the rational function $r_{\{\pm,k\}} \in \Pi_{2m-2}/q^2_{m-1}$ by

$$r_{\{\pm,k\}}(\lambda) := g_{\{\pm,k\}}(x(\lambda)).$$

Indeed, as demonstrated in Remark 4.B.1 further below, such a function is rational. In Figure 4.13 and 4.14 we plot the rational function $r_{\{\pm,m\}}$ and the respective auxiliary polynomial function $g_{\{\pm,m\}}$ for numerical examples. For further illustrations of $r_{\{-,m\}}$ we refer to Figure 4.4 in Subsection 4.4.3.

The rational functions $r_{\{\pm,k\}}$ satisfy the identities (4.4.38) and inequalities (4.4.39) which concludes the proof of Proposition 4.4.12: The identities (4.B.16) conclude the identities (4.4.38) for $r_{\{\pm,k\}}(\theta_j) = g_{\{\pm,k\}}(x(\theta_j))$. Analogously, (4.B.19) and (4.B.21) entail the inequalities (4.4.39).

Furthermore, we consider the inequalities (4.B.19) and (4.B.21) to be strict for $\lambda \neq \{\theta_1, \dots, \theta_m\}$. Indeed, for a given λ with $\lambda \neq \{\theta_1, \dots, \theta_m\}$ we have $y = x(\lambda) \neq \{\xi_1, \dots, \xi_m\}$ and the underlying inequalities for $p_{\pm,\iota(k)-1}$ in (4.B.17) are strict, which carries over to the inequalities (4.B.19) and (4.B.21). □

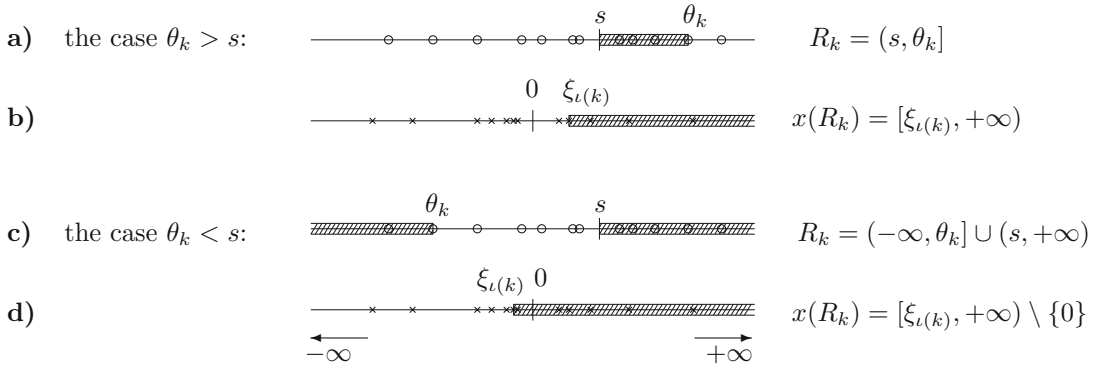


Figure 4.12: In these figures we illustrate the identity $x(R_k) = [\xi_{\iota(k)}, +\infty) \setminus \{0\}$ (4.B.14a) for given nodes $\theta_1, \ldots, \theta_m$. The pole s is given and satisfies $\theta_1 < s < \theta_m$. For the index k we consider two different cases, namely, we choose k such that $\theta_k > s$ in Figure a) and b), and we choose k such that $\theta_k < s$ in Figure c) and d).
– Figure **a)** and **c)** show the real axis with the nodes $\theta_1, \ldots, \theta_m$ ('$\circ$'). Furthermore, the set $R_k \subset \mathbb{R}$ given in (4.4.34) is highlighted by a dashed area.
– Figure **a)** and **c)** show the real axis with $\xi_1, \ldots, \xi_m$ ('$\times$'), i.e., the image of $\theta_1, \ldots, \theta_m$ under the transformation x (4.B.3) with $x(\theta_k) = \xi_{\iota(k)}$ (4.B.9). Furthermore, the dashed area highlights $x(R_k)$ which satisfies the identity (4.B.14a).

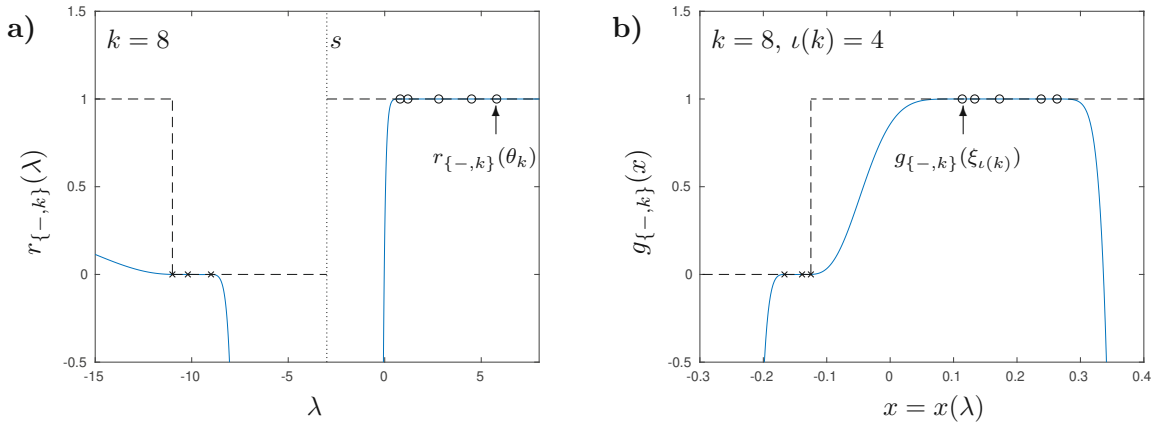


Figure 4.13: – In Figure **a)** we plot the rational function $r_{\{-,k\}}$ (introduced in Proposition 4.4.12) for a numerical example; we show $r_{\{-,k\}}(\lambda)$ over λ for a given pole $s = -3$, and given nodes $\theta_1, \dots, \theta_m$ with $m = 8$. We have $\theta_1 < s < \theta_m$, namely, $\theta_{k_m} < s < \theta_{k_1}$ with $k_m = 3$ and $k_1 = 4$. For $j \in I_k$ we mark $r_{\{-,k\}}(\theta_j)$ by ('$\circ$'), and for $j \notin I_k$ we mark $r_{\{-,k\}}(\theta_j)$ by ('$\times$'). The dashed lines illustrate the upper bounds of $r_{\{-,k\}}$ given in (4.4.39).
– In Figure **b)** we show the auxiliary polynomial function $g_{\{-,k\}}$ which appears in the proof of Proposition 4.4.12, namely, (4.B.15) therein. The nodes $\xi_1, \dots, \xi_m$ correspond to the image of the nodes θ_j under x, namely, $\xi_{\iota(j)} = x(\theta_j)$ for $j = 1, \dots, m$. For $k = 8$ we have $\iota(k) = 4$. The symbols ('$\times$') and ('$\circ$') mark $g_{\{-,m\}}(\xi_j)$ for $j = \iota(k)-1, \dots, m$ and $j = \iota(k), \dots, m$, respectively. The dashed lines illustrate the upper bounds of $g_{\{-,k\}}$ given in (4.B.18).

Remark 4.B.1. *Let $g \in \Pi_{2m-2}$ and let $x(\lambda) = (\lambda - s)^{-1}$, then*

$$r(\lambda) = g(x(\lambda)) \tag{4.B.22}$$

defines a rational function in λ, namely, $r \in \Pi_{2m-2}/q_{m-1}^2$ for $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$. To demonstrate this result we define

$$\widehat{g}(\lambda) = g\big((\lambda - s)^{-1}\big)(\lambda - s)^{2m-2}. \tag{4.B.23}$$

Expanding the right-hand side of (4.B.23) *shows $\widehat{g} \in \Pi_{2m-2}$. Substituting $x(\lambda)$ and $q_{m-1}(\lambda)$ in* (4.B.23)*, and dividing by $q_{m-1}(\lambda)^2$ reveals the representation*

$$\widehat{g}(\lambda)/q_{m-1}(\lambda)^2 = g(x(\lambda)),$$

and thus, with (4.B.22) *we have $r(\lambda) = \widehat{g}(\lambda)/q_{m-1}(\lambda)^2$. This shows $r \in \Pi_{2m-2}/q_{m-1}^2$.*

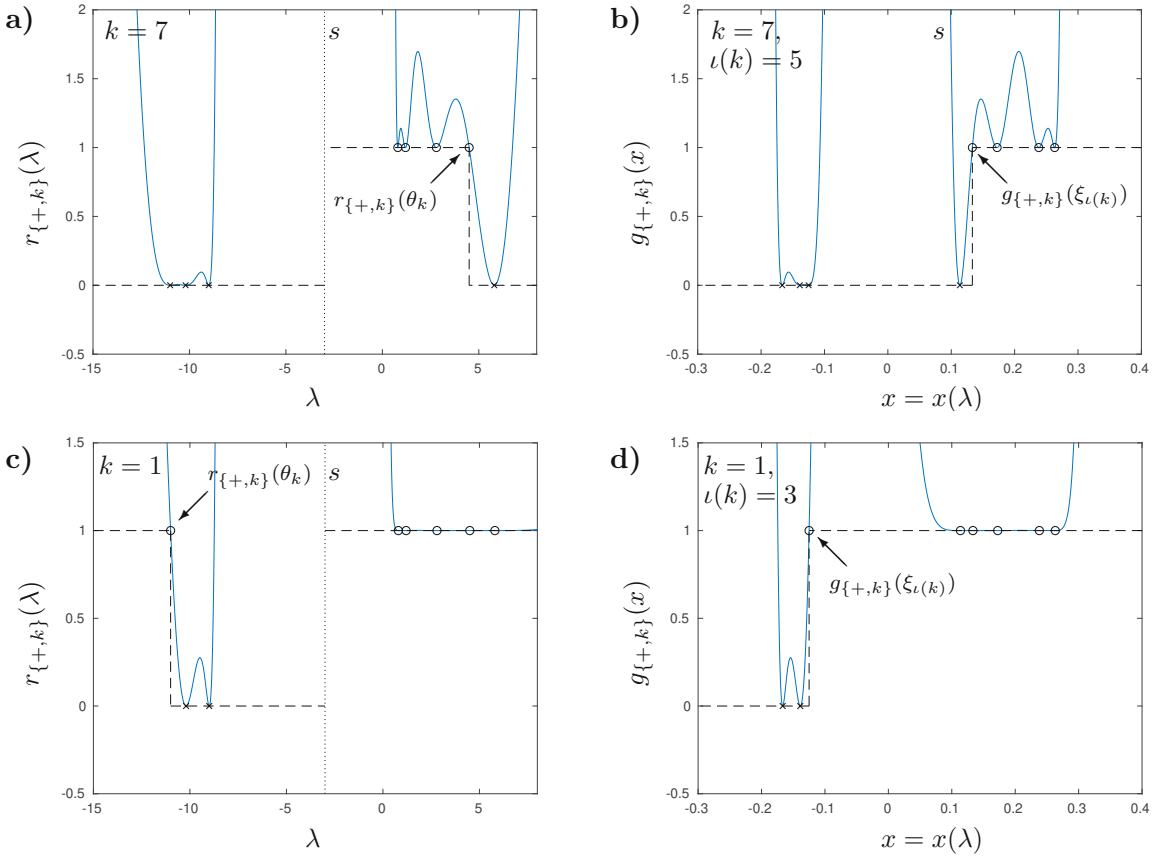


Figure 4.14: – In Figure **a)** and **c)** we plot the rational function $r_{\{+,k\}}$ (introduced in Proposition 4.4.12) for numerical examples. The nodes $\theta_1, \ldots, \theta_m$ with $m = 8$ and the pole s are given as in Figure 4.13 a). For the index k we choose $k = 7$ in Figure a) and $k = 1$ in Figure c). For $j \in I_k$ we mark $r_{\{+,k\}}(\theta_j)$ by ('$\circ$'), and for $j \notin I_k$ we mark $r_{\{+,k\}}(\theta_j)$ by ('$\times$'). The dashed lines illustrate the upper bounds of $r_{\{+,k\}}$ given in (4.4.39).
– In Figure **b)** and **c)** we show the auxiliary polynomial function $g_{\{+,k\}}$ which appears in the proof of Proposition 4.4.12, namely, (4.B.15) therein. The function $g_{\{+,k\}}$ with $k = 7$ plotted in Figure b) is associated with the function $r_{\{+,k\}}$ in Figure a), and analogously, such a relation is given for Figure d) and Figure c). For the index $\iota(k)$ which is relevant in the proof of Proposition 4.4.12 we remark $\iota(k) = 5$ for $k = 7$ and $\iota(k) = 3$ for $k = 1$. As in Figure 4.13 b), the nodes $\xi_1, \ldots, \xi_m$ correspond to the image of the nodes θ_j under x, namely, $\xi_{\iota(j)} = x(\theta_j)$ for $j = 1, \ldots, m$. The symbols ('$\times$') and ('$\circ$') mark $g_{\{+,m\}}(\xi_j)$ for $j = \iota(k) - 1, \ldots, m$ and $j = \iota(k), \ldots, m$, respectively. The dashed lines illustrate the upper bounds of $g_{\{+,k\}}$ given in (4.B.18).

5 A localized near-best approximation property for rational Krylov approximations to exponentials of skew-Hermitian matrices

5.1 Introduction

The matrix exponential yields an evolution operator to a linear, homogeneous system of differential equations and is relevant for many applications. In the present work we consider rational Krylov approximations to the action of exponentials of skew-Hermitian matrices, i.e., $\mathrm{e}^{-\mathrm{i}tA}u$ where $A \in \mathbb{C}^{n\times n}$ is Hermitian, $t > 0$ denotes a time step and $u \in \mathbb{C}^n$ denotes the initial vector. For the problem setting we refer to Section 5.2.

As the main result of the present chapter, we discuss a new near-best approximation property for the rational Krylov approximation (Proposition 5.3.2 in Section 5.3). More precisely, an error bound based on the approximation error of a scalar best approximation of the imaginary exponential on an interval covering a subset of the spectrum of A, which holds independently of the full spectrum of A. Comparing with the classical near-best approximation property which depends on a scalar best approximation on the convex hull of the spectrum of A, we refer to our result as a localized near-best approximation property. Our result is based on properties of the decomposition of the initial vector u in the eigenbasis of A (Assumption (A1) in Proposition 5.3.2). We further require that these spectral properties of the initial vector are preserved in the rational Krylov subspace (Assumption (A2) in Proposition 5.3.2). As a key aspect of the present work, we further discuss whether Assumption (A2) can be deduced from spectral properties of A and u, especially, when Assumption (A1) holds true. Such a statement is not proven here in a practical setting but seems to be valid for relevant examples (numerical experiments are presented in Section 5.4).

As an application for the localized near-best approximation property provided by Proposition 5.3.2, we discuss a problem setting where the rational Krylov approximation can show a grid-independent convergence rate, in contrast to a polynomial Krylov approximation: Let A and u correspond to a spatial grid discretization of an unbounded differential operator and an initial state, respectively. Then the error of the polynomial Krylov approximation to the matrix exponential typically correlates with the spectral norm of A, which increases with a refinement of the underlying grid. On the other hand, the rational Krylov approximation yields a grid-independent convergence rate in relevant cases. We give an overview on previous works on a grid-independent convergence rate for the rational Krylov approximation in Section 5.2. When the initial vector u is related to an initial state

of high regularity, then u fits to the setting of Assumption (A1). If we further assume that Assumption (A2) holds in this setting independently of a grid refinement (which is reasonable for the numerical experiments in Section 5.4), then Proposition 5.3.2 provides a grid-independent best-approximation property. This yields further insights on a grid-independent convergence rate for the rational Krylov approximation and a proper choice for the poles of the rational Krylov subspace, which will be topic of future work.

5.2 Problem setting and previous work

Let $\phi(t) \in \mathbb{C}^n$ be the solution of a large system of ordinary differential equations (ODEs), with an Hermitian matrix $A \in \mathbb{C}^{n\times n}$ and an initial vector $u \in \mathbb{C}^n$ at $t = 0$,

$$\phi'(t) = -\mathrm{i}\, A\phi(t), \quad \phi(0) = u, \quad -\mathrm{i}\, A \in \mathbb{C}^{n\times n} \text{ skew-Hermitian.} \tag{5.2.1}$$

A prominent example is a spatially discrete evolution equation of Schrödinger type, with a sparse matrix A, typically resulting from an ansatz based on localized basis functions on a given grid. To discuss the effects of a refinement of the underlying grid, we introduce a generalized inner product on $\mathbb{C}^n$ which is motivated by an inner product on the underlying function space, e.g., the L^2-inner product on the space of functions which are square-integrable on the underlying spatial domain. For two vectors $u, v \in \mathbb{C}^n$ we define the M-inner product by[1]

$$(u, v)_{\mathsf{M}} = u^{\mathsf{H}} M v, \tag{5.2.2}$$

where $M \in \mathbb{C}^{n\times n}$ is an Hermitian[2] positive definite matrix which is given by the underlying problem setting. In the case of A being based on a finite difference discretization of a differential operator, the M-inner product typically corresponds to a scaled Euclidean inner product, e.g.,[3] $(u, v)_{\mathsf{M}} = h\,(u, v)_2$ for a one-dimensional spatial domain with grid with h. With (5.2.2) this conforms to $M = h\,I$. Otherwise, in the case of an underlying finite element method (FEM) discretization, the matrix M typically corresponds to the mass matrix of the finite element space.

In the sequel, we assume that A is Hermitian (self-adjoint) w.r.t. the M-inner product. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ denote the eigenvalues of A, with $\lambda_1 \le \ldots \le \lambda_n$. As a typical case, A includes a discretization of the negative Laplacian $-\Delta$, with[4] $\|A\|_{\mathsf{M}} = \max_{j=1,\ldots,n} |\lambda_j| \to \infty$ for $n \to \infty$, i.e., A is unbounded for $n \to \infty$.

A basic assumption is $\lambda_1 \ge a \in \mathbb{R}$, with a independent of the problem size n.

Krylov approximation of the matrix exponential function. The solution of (5.2.1) is given by the matrix exponential,

$$\phi(t) = \mathrm{e}^{-\mathrm{i}tA} u = \sum_{k=0}^{\infty} \frac{(-\mathrm{i}tA)^k u}{k!}. \tag{5.2.3}$$

[1] The M-inner product given in (5.2.2) induces a vector norm, i.e., $\|u\|_{\mathsf{M}} = \sqrt{(u, u)_{\mathsf{M}}}$, which is equivalent to the Euclidean norm.

[2] The matrix M is Hermitian w.r.t. the Euclidean inner product, i.e., $M = M^{\mathsf{H}}$.

[3] By $(\cdot, \cdot)_2$ and $\|\cdot\|_2$ we denote the Euclidean inner product and norm in $\mathbb{C}^n$, respectively.

[4] The vector M-norm induces a matrix norm which is equivalent to the spectral norm.

For the time propagation of (5.2.1) it is not required to compute the matrix $\mathrm{e}^{-\mathrm{i}tA}$ directly which is rarely beneficial for efficiency reasons. Instead, in practice one approximates its action (5.2.3) applied to a given initial vector u. A common approach relies on approximation of $\phi(t)$ in a polynomial Krylov subspace

$$\mathcal{K}_m(A, u) = \operatorname{span}\{u, A\,u, \ldots, A^{m-1}u\}.$$

Such a type of approximation was early applied in the field of quantum mechanics [NW83, PL86], and a more extensive study is given in [Lub08]. For the exponential of a symmetric matrix, the Krylov approximation was early studied in [DK89, DK92, GS92, Saa92, HL97], and many other works are devoted to the case of more general matrices.

Let us introduce the notation for polynomial Krylov subspaces. By $V_m \in \mathbb{C}^{n\times m}$ we denote the M-orthonormal[5] basis of $\mathcal{K}_m(A, u)$ constructed by the Lanczos iteration starting from u, with

$$(V_m, u)_{\mathsf{M}} = \beta_0\, e_1, \quad \beta_0 = \|u\|_{\mathsf{M}}, \quad e_1 = (1, 0, \ldots, 0)^{\mathsf{T}} \in \mathbb{R}^m, \tag{5.2.4}$$

and let[6]

$$J_m = (V_m, A\,V_m)_{\mathsf{M}} \in \mathbb{R}^{m\times m} \tag{5.2.5}$$

denote the low-dimensional representation of A w.r.t. the Krylov subspace $\mathcal{K}_m(A, u)$. We refer to the polynomial Krylov approximation as

$$\beta_0 V_m\, \mathrm{e}^{-\mathrm{i}tJ_m} e_1 \approx \mathrm{e}^{-\mathrm{i}tA} u. \tag{5.2.6}$$

Here, $\mathrm{e}^{-\mathrm{i}tJ_m}$ typically represents a low-dimensional problem which can be treated in a direct manner in contrast to the original high-dimensional problem.

We use the denotation Π_{m-1} for the class of complex polynomials of degree $\le m-1$. The Krylov approximation (5.2.6) is near-optimal in the class of corresponding polynomial approximations. This result was stated early in [SL96] and others for the symmetric case and remains valid in a more general setting. An a priori estimate for the skew-Hermitian case is given in [Lub08, Theorem 2.8], and holds w.r.t. the M-inner product,

$$\|\beta_0 V_m\, \mathrm{e}^{-\mathrm{i}tJ_m} e_1 - \mathrm{e}^{-\mathrm{i}tA} u\|_{\mathsf{M}} \le 2 \min_{p\in\Pi_{m-1}} \max_{\lambda\in[\lambda_1,\lambda_n]} |p(\lambda) - \mathrm{e}^{-\mathrm{i}t\lambda}|\,\|u\|_{\mathsf{M}}.$$

This upper bound provides a rule of thumb for the Krylov dimension to obtain superlinear convergence. In general, a proper Krylov dimension m is related to the problem size (via $\|A\|_{\mathsf{M}}$) and the time step t, namely $m \approx t\|A\|_{\mathsf{M}}$ for an Hermitian matrix A and the approximation of $\lambda \mapsto \mathrm{e}^{-\mathrm{i}t\lambda}$, see also [GS92, HL97]. In our setting a grid refinement for the underlying discretization results in a larger norm $\|A\|_{\mathsf{M}}$ and respectively, a larger Krylov dimension m will be required to compute an accurate approximation in typical cases.

Over the last years, rational Krylov techniques have become more and more popular. These are based on a rational Krylov subspace

$$\mathcal{Q}_m(A, u) = \{r(A)u\colon\ r \in \Pi_{m-1}/q_{m-1}\},$$

[5] For two vectors $x, y \in \mathbb{C}^m$ an M-orthonormal basis V_m satisfies $(V_m\,x, V_m\,y)_{\mathsf{M}} = (x, y)_2$.

[6] The notation $(V_m, A\,V_m)_{\mathsf{M}}$ with $V_m = (v_1, \ldots, v_m) \in \mathbb{C}^{n\times m}$ refers to a $m \times m$-dimensional matrix for which the (j, k)-th entry corresponds to $(v_j, Av_k)_{\mathsf{M}}$.

with a preassigned set of poles $s_1, \dots, s_{m-1} \in \mathbb{C}$ defining the denominator $q_{m-1}(\lambda) = \prod_{j=1}^{m-1}(\lambda - s_j)$. For an early work on rational Krylov techniques we refer to [Ruh84]. Let $U_m \in \mathbb{C}^{n\times m}$ be a given M-orthonormal basis of $\mathcal{Q}_m(A, u)$ for which we define[7]

$$x = (U_m, u)_{\mathsf{M}} \in \mathbb{C}^m. \tag{5.2.7}$$

In contrast to the polynomial case, where V_m corresponds to the Krylov basis constructed by the Lanczos iteration, various algorithms exist to construct a rational Krylov subspace. Some specific algorithms for this purpose are given in more detail in Section 5.4. However, the notation U_m for the rational Krylov basis suits a rather general setting. For the corresponding low-dimensional representation of A w.r.t. $\mathcal{Q}_m(A, u)$ we introduce the denotation[8]

$$A_m = (U_m, A\,U_m)_{\mathsf{M}} \in \mathbb{C}^{m\times m}. \tag{5.2.8}$$

Similar to (5.2.6) the resulting rational Krylov approximation of the matrix exponential (5.2.3) is then given by

$$U_m\, \mathrm{e}^{-\mathrm{i}\,t A_m}\, x \approx \mathrm{e}^{-\mathrm{i}\,t A} u. \tag{5.2.9}$$

A review on rational Krylov methods for matrix functions is given in [Güt10].

For the algorithmic construction of a rational Krylov subspace, several applications of an inverse of full dimension n are required, relying on iterative techniques if applicable. This typically increases the computational cost compared to the polynomial case. Furthermore, the choice of poles will strongly influence the quality of the approximation and some a priori knowledge of the problem will be essential. On the other hand, making use of rational functions can significantly enhance the quality of the approximation.

For a dissipative system a rational approximations can benefit from the decaying nature of solutions. For this case, rational Krylov approximations with a single pole are studied in [vdEH06], and a priori convergence results independent of the matrix spectrum are given. The influence of the refinement of the underlying spatial discretization on the convergence behavior is typically not critical here. Similar results hold for sectorial operators [MN04].

For the skew-Hermitian case (5.2.1), on the other hand, solutions show an oscillatory behavior, and the approach from [vdEH06] and [MN04] is not directly applicable: Similar to the polynomial case we have a near-optimal convergence rate for the rational Krylov approximation in a class of rational functions r with denominator q_{m-1}, see also [Güt10, Theorem 4.10] or [DKZ09, BR09],

$$\|U_m\, \mathrm{e}^{-\mathrm{i}\,t A_m}\, x - \mathrm{e}^{-\mathrm{i}\,t A} u\|_{\mathsf{M}} \le 2 \min_{r\in\Pi_{m-1}/q_{m-1}} \max_{\lambda\in[\lambda_1,\lambda_n]} |r(\lambda) - \mathrm{e}^{-\mathrm{i}\,t\lambda}| \|u\|_{\mathsf{M}}. \tag{5.2.10}$$

For the dissipative case, an error bound of the type (5.2.10) depends on an approximation error of the negative exponential, i.e., $\max_{\lambda\in\mathbb{R}_+} |r(\lambda) - \mathrm{e}^{-t\lambda}|$, instead of the imaginary exponential, and an upper bound independent of the matrix spectrum is viable [vdEH06, Lemma 3.1 and following remarks]. Such a result does not appear to be plausible for the skew-Hermitian case based on (5.2.10). Nevertheless, desirable convergence properties

[7] With $(U_m, u)_{\mathsf{M}} = x \in \mathbb{C}^m$ we include the case where u is not the first basis vector of U_m. Note that $\|x\|_2 = \|u\|_{\mathsf{M}}$.

[8] In the sequel we refer to the matrix A_m as Rayleigh quotient.

have been observed in relevant applications of the skew-Hermitian case, e.g., for time integration of Maxwell equations in [Bot16] and [HPS+15]. For the skew-Hermitian case, a grid-independent convergence behavior seems to require further regularity properties of the initial vector u. An appropriate theory is introduced within a series of works[9] [GG10, Gri12, GG17] which we now briefly recapitulate. The work of [GG10] also considers the so-called φ-functions closely related to the exponential function; e.g.,

$$\mathrm{e}^{-\mathrm{i}tA}u = u + \varphi_1(-\mathrm{i}tA)(-\mathrm{i}tA)u, \quad \text{with} \quad \varphi_1(\lambda) = (\mathrm{e}^{\lambda} - 1)/\lambda. \tag{5.2.11}$$

The matrix φ-functions can be approximated by a truncated resolvent series ([GG10]) or in a resolvent Krylov subspace (i.e., a rational Krylov subspace with a single pole, [Gri12]), with a sublinear convergence rate independent of the underlying grid. For the resolvent series in [GG10] this result carries over to the respective matrix exponential function if additional regularity assumptions on the initial vector u are prescribed. In particular, Au is assumed to be bounded in norm independent of the grid. Intuitively, the importance of such bounds on Au can be understood from (5.2.11). With the smoothness operators introduced in [GG17] and regularity assumptions on u, results of [Gri12] carry over to the approximation of the matrix exponential in the resolvent Krylov subspace, essentially as in [GG17, Theorem 7.1].

5.3 A localized near-best approximation property for the rational Krylov approximation

As a main result of the present chapter we state a localized near-best approximation property for the rational Krylov approximation, see Proposition 5.3.2 below. This result provides a near-best approximation property which can hold independently of the full spectrum of A, and yields further insights on a potential grid-independent convergence rate of the rational Krylov approximation.

We proceed with formulating appropriate assumptions.

Influence of the initial vector u on the quality of the Krylov approximation: a general consideration. Let $Q = (q_1, \dots, q_n) \in \mathbb{C}^{n\times n}$ represent an M-orthonormal eigenbasis of A with eigenvalues $\lambda_1 \le \dots \le \lambda_n \in \mathbb{R}$, i.e., $A\,q_j = \lambda_j q_j$ and $(q_j, q_k)_{\mathsf{M}} = \delta_{jk}$. For the initial vector u we define the spectral coefficients $w_1, \dots, w_n \in \mathbb{C}$ by $w_j = (q_j, u)_{\mathsf{M}}$. Then,

$$u = \sum_{j=1}^{n} w_j\, q_j, \quad \text{with} \quad \|u\|_{\mathsf{M}}^2 = \sum_{j=1}^{n} |w_j|^2. \tag{5.3.1}$$

Let $\widehat{p}_{m-1}$ denote the polynomial interpolant of $\lambda \mapsto \mathrm{e}^{-\mathrm{i}t\lambda}$ at the eigenvalues of J_m defined in (5.2.5), which are also known as *Ritz values*. For the error norm of the polynomial Krylov approximation we have ([Eri90])

$$\|\beta_0 V_m\, \mathrm{e}^{-\mathrm{i}tJ_m} e_1 - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} = \Big(\sum_{j=1}^{n} |\widehat{p}_{m-1}(\lambda_j) - \mathrm{e}^{-\mathrm{i}t\lambda_j}|^2\, |w_j|^2 \Big)^{1/2}. \tag{5.3.2a}$$

[9] See also [GH08] for trigonometric functions which are closely related to the exponential of a skew-Hermitian matrix.

A similar result holds for rational Krylov approximations: Let $\widehat{r}_{m-1}$ be the rational interpolant of $\lambda \mapsto \mathrm{e}^{-\mathrm{i}t\lambda}$ at the eigenvalues of A_m defined in (5.2.8), i.e., $\widehat{r}_{m-1} = \widetilde{p}_{m-1}/q_{m-1}$ where $\widetilde{p}_{m-1}$ is the Lagrange interpolant of $\lambda \mapsto q_{m-1}(\lambda)\mathrm{e}^{-\mathrm{i}t\lambda}$ at these eigenvalues. The identity $U_m\,\mathrm{e}^{-\mathrm{i}tA_m}\,x = \widehat{r}_{m-1}(A)u$ (see [Güt10, Theorem 4.8]) and (5.3.1) directly yield a counterpart of (5.3.2a) for the rational case, namely

$$\|U_m\,\mathrm{e}^{-\mathrm{i}tA_m}\,x - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} = \Big(\sum_{j=1}^{n} |\widehat{r}_{m-1}(\lambda_j) - \mathrm{e}^{-\mathrm{i}t\lambda_j}|^2\,|w_j|^2\Big)^{1/2}, \tag{5.3.2b}$$

with the Rayleigh quotient A_m defined in (5.2.8).

Some reasonable settings concerning the spectral coefficients of the initial vector u corresponding to a discretized initial state of higher regularity. We proceed with some remarks on the present setting, namely that (5.2.1) has its origin in a spatially discretized partial differential equation (PDE). In this context, the regularity of the underlying differential equation describes how many times the differential operator can be applied to the initial state in accord to the domain of definition of the underlying operator. For a proper spatial discretization this entails grid-independent bounds on $\|A^k u\|_{\mathsf{M}}$ in a typical setting, where $k \in \mathbb{N}_0$ depends on the level of regularity. Furthermore, we consider the convergence of eigenvalues λ_j and spectral coefficients w_j defined in (5.3.1) for a sufficiently fine grid, i.e., $n \to \infty$. We classify eigenvalues λ_j to be 'relevant' if the corresponding spectral coefficient w_j is of significant size. Typically, eigenvalues do not converge with a uniform speed, e.g., eigenvalues of the Laplace operator related to higher frequencies require a finer grid to converge. Nevertheless, once the relevant eigenvalues are roughly approximated, then grid-independent bounds on $\|A^k u\|_{\mathsf{M}} = (\sum_{j=1}^{n} \lambda_j^{2k}|w_j|^2)^{1/2}$ entail that the spectral coefficients of 'irrelevant' eigenvalues related to higher frequencies are appropriately small in size.

Motivated by regularity properties of an underlying initial state, we proceed to discuss different settings for the spectral coefficients $w_1, \ldots, w_n$ of the initial vector u, and the potential effects of a grid refinement on the error of the polynomial and rational Krylov approximation via the error representation in (5.3.2).

(W1) We first consider the initial vector $u \in \mathbb{C}^n$ to represent a discretization of a smooth (analytic) initial state. Technically, we assume there exists an interval $[a, b]$ which we consider to be grid-independent and for which the following assumptions hold true: We assume the interval $[a, b]$ covers the relevant eigenvalues, and therefore, $|w_j| \ll \|u\|_{\mathsf{M}}$ for any $\lambda_j \notin [a, b]$. Additionally, we assume that $|w_j|$ decays exponentially in j for $\lambda_j \notin [a, b]$.

The polynomial $\widehat{p}_{m-1}$ and the rational function $\widehat{r}_{m-1}$ need to provide an accurate approximation of $\lambda \mapsto \mathrm{e}^{-\mathrm{i}t\lambda}$ for sum terms related to relevant eigenvalues λ_j in (5.3.2), in order to guarantee an accurate Krylov approximation. In the polynomial case, $\widehat{p}_{m-1}(\lambda)$ increases polynomially in λ. Nevertheless, with the present assumptions of $|w_j|$ decaying exponentially the polynomial increase of $\widehat{p}_{m-1}(\lambda)$ is not critical and the polynomial approximation can show close to grid-independent convergence rates in this case. For the rational function $\widehat{r}_{m-1}$ in (5.3.2b) it is reasonable to assign

poles close to the relevant eigenvalues in $[a, b]$ such that relevant eigenvalues are first resolved in the Krylov subspace. Hence, we consider $\widehat{r}_{m-1}$ to be bounded on the 'irrelevant' part of the spectrum, $\lambda_j \notin [a, b]$. This lets us expect that both, polynomial as well as rational Krylov approximations, have the potential for grid-independent convergence rates in such cases.

(W2) A major topic of this work is to consider also perturbed states. We assume that the interval $[a, b]$ covers the relevant eigenvalues independently of the grid. Furthermore, let $\lambda_\ell, \dots, \lambda_n$ denote the 'irrelevant' eigenvalues, then we assume $(\sum_{j=\ell}^{n} |w_j|^2)^{1/2} < \varepsilon \|u\|_{\mathsf{M}}$ with $0 < \varepsilon \ll 1$ being *independent of the grid*.

Thus, the relevant eigenvalues are located in a specific part of the spectrum, and the contribution of the remaining eigenvalues to the initial vector constitutes a perturbation which is sufficiently small in norm. Such a perturbation may, for instance, originate from approximation of an (originally) analytical initial state, e.g., resulting from previous integration steps.

In this case the behavior of a polynomial approximation can significantly differ from the behavior of a rational approximation. Although the perturbation is assumed to be small, the polynomial increase of $\widehat{p}_{m-1}(\lambda)$ can scale up the error terms in (5.3.2a) related to larger eigenvalues λ_j, especially for $\lambda_j \notin [a, b]$. This effect is strengthened with the existence of larger eigenvalues λ_j, hence, the influence of the refinement of the underlying spatial discretization a grid can be critical for the approximation performance of $\widehat{p}_{m-1}(\lambda)$. On the other hand a rational function $\widehat{r}_{m-1}(\lambda)$, with a proper choice of poles, is bounded for larger choices of λ and does not critically scale up error terms related to $\lambda_j \notin [a, b]$, therefore, a grid-independent convergence can be expected for the rational approximation.

(W3) As a theoretical consideration we comment on the case of an initial vector u which is 'uniformly spread' over the spectrum of A. In this case both, the polynomial and rational Krylov approximation, need to give an accurate approximation on the full spectrum and for both the difficulty of this problem increases with the problem size, in particular on the choice of the underlying grid.

One of our main interests is case (W2), concerning rational approximation, for which we may expect a significantly improved performance compared to the polynomial case.

Stability properties of rational functions and a grid-independent convergence rate for bounded rational approximants to the matrix exponential. Here the term stability is used according to the context of numerical time integration (see also [HW02]), which fits well to the approximation of the exponential function. In a traditional context the stability study of a numerical integrator considers stability domains, subsets of the complex plane for which the time integration step satisfies specific bounds. In our setting we consider a bounded rational approximation of the exponential function on a subset of the imaginary axis. For a fixed $q_{m-1} \in \Pi_{m-1}$ we denote

$$R_{\varrho,b} = \{r = p/q_{m-1} \colon\ p \in \Pi_{m-1},\ \sup_{\lambda \in (b,+\infty)} |r(\lambda)| \leq \varrho\}. \tag{5.3.3}$$

A prominent example for functions in the class $R_{\varrho,b}$ arise from A-acceptable rational approximations to the exponential (which correspond to A-stable numerical integrators). This relation is further clarified in Remark 5.A.1, Appendix 5.A.

In the following proposition we give a grid-independent error bound for a rational approximation of the matrix exponential function with assumptions according to case (W2) and assuming that the rational approximant is properly bounded.

Proposition 5.3.1. *Let* $\lambda_1, \ldots, \lambda_{\ell-1} \in [a,b]$ *and* $(\sum_{j=\ell}^{n} |w_j|^2)^{1/2} < \varepsilon \, \|u\|_{\mathsf{M}}$. *Consider the class* $R_{\varrho,b}$ *defined in* (5.3.3) *with fixed denominator* q_{m-1}, *assuming that* $q_{m-1}(A)$ *is invertible. Then, for* $r \in R_{\varrho,b}$,

$$\|r(A)u - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} \leq \max_{\lambda\in[a,b]} |r(\lambda) - \mathrm{e}^{-\mathrm{i}t\lambda}| \, \|u\|_{\mathsf{M}} + \varepsilon(1+\varrho)\|u\|_{\mathsf{M}}. \tag{5.3.4}$$

Proof. See Appendix 5.A. □

With Assumption (W2), i.e., if $[a,b]$ and ε are independent of the underlying discretization, we refer to the error bound from Proposition 5.3.1 as *grid-independent.*

A new approach to motivate a grid-independent convergence rate for the rational Krylov approximation. Now we consider a rational Krylov approximation with denominator q_{m-1} defined via preassigned poles, and $R_{\varrho,b} \subseteq \Pi_{m-1}/q_{m-1}$. We recall $r(A)u = U_m \, r(A_m)x$ for x given in (5.2.7) and $r \in \Pi_{m-1}/q_{m-1}$, see [Güt10, Lemma 4.6] and others. For arbitrary $r \in R_{\varrho,b}$ this implies

$$\|U_m \, \mathrm{e}^{-\mathrm{i}tA_m} \, x - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} \leq \|r(A)u - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} + \|U_m r(A_m)x - U_m \mathrm{e}^{-\mathrm{i}tA_m} \, x \, \|_{\mathsf{M}}. \tag{5.3.5}$$

Due to U_m being an M-orthonormal basis, the second term on the right-hand side of this inequality satisfies

$$\|U_m r(A_m)x - U_m \mathrm{e}^{-\mathrm{i}tA_m} \, x \, \|_{\mathsf{M}} = \|r(A_m)x - \mathrm{e}^{-\mathrm{i}tA_m} \, x \, \|_2,$$

and (5.3.5) simplifies to

$$\|U_m \, \mathrm{e}^{-\mathrm{i}tA_m} \, x - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} \leq \|r(A)u - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} + \|r(A_m)x - \mathrm{e}^{-\mathrm{i}tA_m} \, x \, \|_2. \tag{5.3.6}$$

For the first term on the right-hand side of (5.3.6) we can apply Proposition 5.3.1. The second term is of a similar type, substituting A and u by A_m and x, respectively. Hence, we require conditions of Proposition 5.3.1 to be valid w.r.t. the spectral decomposition of x in the spectrum of A_m. We introduce the notation

$$\theta_1, \ldots, \theta_m \in \mathbb{R}, \quad \text{and} \quad Q_m \in \mathbb{C}^{m\times m}$$

for the eigenvalues and the ℓ^2-orthonormal[10] eigenbasis of A_m, respectively, i.e., $A_m = Q_m \, \Theta_m \, Q_m^{\mathsf{H}}$, with $\Theta_m = \mathrm{diag}(\theta_1, \ldots, \theta_m)$. We assume the ordering

$$\theta_1 < \ldots < \theta_m.$$

[10]The notation 'ℓ^2-orthonormal' refers to a basis orthonormal w.r.t. the Euclidean inner product.

Furthermore, we let $c_1, \dots, c_m \in \mathbb{C}$ denote the coefficients of x in the eigenbasis of A_m defined by $(c_1, \dots, c_m)^\intercal = Q_m^{\mathsf{H}} x$. We will also refer to these coefficients as the *spectral coefficients* w.r.t. the Krylov subspace.

We now give a new upper bound on the error norm for rational Krylov approximations of the matrix exponential.

Proposition 5.3.2. *Let U_m span the rational Krylov subspace $\mathcal{Q}_m(A, u)$ with poles given by q_{m-1}, and $x = (U_m, u)_{\mathsf{M}}$ and $A_m = (U_m, AU_m)_{\mathsf{M}}$. With the notation introduced before, we assume the following conditions to hold for $\varepsilon, \widehat{\varepsilon} > 0$ and some indices ℓ, k:*

(A1) $\lambda_1, \dots, \lambda_{\ell-1} \in [a, b]$ *with* $(\sum_{j=\ell}^{n} |w_j|^2)^{1/2} \le \varepsilon \|u\|_{\mathsf{M}}$, *and*

(A2) $\theta_1, \dots, \theta_{k-1} \in [a, b]$ *with* $(\sum_{j=k}^{m} |c_j|^2)^{1/2} \le \widehat{\varepsilon} \|u\|_{\mathsf{M}}$.

Let $R_{\varrho,b}$ be given by (5.3.3) *with denominator q_{m-1}. Then,*

$$\|U_m \mathrm{e}^{-\mathrm{i}tA_m} x - \mathrm{e}^{-\mathrm{i}tA} u\|_{\mathsf{M}} \le 2 \min_{r \in R_{\varrho,b}} \max_{\lambda \in [a,b]} |r(\lambda) - \mathrm{e}^{-\mathrm{i}t\lambda}| \|u\|_{\mathsf{M}} + (1+\varrho)(\varepsilon + \widehat{\varepsilon}) \|u\|_{\mathsf{M}}.$$

Proof. See Appendix 5.A. □

In Proposition 5.3.2, Assumption (A1), i.e., the choice of $[a, b]$ and ε, is based on properties of the underlying problem and can be assumed independently of the spatial discretization in a reasonable setting (W2). On the other hand, Assumption (A2) cannot be justified in an a priori sense without further considerations. However, the spectral coefficients c_j of x in the Krylov subspace are closely related to the spectral coefficients w_j of u.

As an example we proceed with a result based on Remark 4.4.14 and Corollary 4.4.7 (Section 4.4 in Chapter 4). With the assumption $a < \lambda_1$ in the present chapter, the notation a in the present chapter conforms to Chapter 4. However, the notation b has a different meaning in the present chapter and Chapter 4; we apply results of Chapter 4 with $b = \infty$.

Corollary 5.3.3. *[Remark 4.4.14 and Corollary 4.4.7] Consider the rational Krylov subspace with a single pole $s < \lambda_1$ of multiplicity $m-1$, i.e., $q_{m-1}(\lambda) = (\lambda - s)^{m-1}$. Let $\ell = \ell(k-1) \in \mathbb{N}$ be defined by*

$$\lambda_{\ell(k-1)} < \theta_{k-1} \le \lambda_{\ell(k-1)+1} \quad \text{for } 1 < k \le m. \tag{5.3.7}$$

Then the spectral coefficients in the Krylov subspace satisfy

$$\sum_{j=k}^{m} |c_j|^2 < \sum_{j=\ell(k-1)+1}^{n} |w_j|^2 \quad \text{for } k = 2, \dots, m. \tag{5.3.8}$$

Proof. Following Remark 4.4.14, the spectral coefficients c_j and the measure μ_n as given in (4.4.13) in Subsection 4.4.2 satisfy the inequalities given in Corollary 4.4.7. Particularly, the upper bound in (4.4.22b) in Corollary 4.4.7 yields

$$\sum_{j=k}^{m} |c_j|^2 < \mu_n((\theta_{k-1}, \infty)).$$

With the index ℓ as in (5.3.7), the measure μ_n as in (4.4.13) as satisfies

$$\mu_n([\theta_{k-1}, \infty)) = \sum_{j=\ell(k-1)+1}^{n} |w_j|^2,$$

which implies (5.3.8). □

Without going into detail, we remark that for the Ritz values $\theta_1, \ldots, \theta_m \in (\lambda_1, \lambda_n)$, the index $\ell(k-1)$ with $1 \leq \ell(k-1) < n$ is well-defined in (5.3.7), and $\theta_{k-1} \leq \lambda_{\ell(k-1)+1} < \theta_k$. In order to apply Corollary 5.3.3 we assume that there exist at least one Ritz value $\theta_k > b$, otherwise, with $a < \theta_1$ Assumption (A2) is trivially satisfied. We remark a special case for which Corollary 5.3.3 specifies $\widehat{\varepsilon}$ in (A2): Let (A1) be satisfied, and

$$\theta_{k-1} \leq b < \lambda_{\ell(k-1)+1} < \theta_k, \quad \text{then Corollary 5.3.3 implies} \quad \Big(\sum_{j=k}^{m} |c_j|^2\Big)^{1/2} < \varepsilon \|u\|_{\mathsf{M}}. \tag{5.3.9}$$

Thus, for the case (5.3.9) Assumption (A2) can be concluded from (A1) with $\widehat{\varepsilon} = \varepsilon$. In a general setting, $\lambda_{\ell(k-1)+1} \in [a, b]$ is plausible, and further properties of the Krylov subspace have to be considered to conclude (A2) in an a priori sense. Such arguments will not be followed here further. In Section 5.4 we present numerical examples which suggest that (A1) implies Assumption (A2) to hold with $\widehat{\varepsilon} \approx \varepsilon$ independently of the underlying grid width.

We proceed to give similar results for the rational Krylov subspace with a single pole $s \in \mathbb{R}$ of multiplicity $m-1$ which is located within the range of the eigenvalues of A, i.e., $\lambda_1 < s < \lambda_n$, where the choice $\lambda_1 < s < b$ is reasonable in the setting of Proposition 5.3.2. To be more precise, we assume $\theta_1 < s < \theta_m$ for the Ritz values, which certainly holds for a sufficiently large m when $\lambda_1 < s < \lambda_n$. We proceed with the following result based on Corollary 4.4.16 (Subsection 4.4.3 in Chapter 4).

Corollary 5.3.4. *[see Corollary 4.4.16] Consider the rational Krylov subspace with a single pole $s \in \mathbb{R}$ of multiplicity $m-1$ and $\lambda_1 < s < \lambda_n$. We further assume $\theta_1 < s < \theta_m$. Let the index k be given with $k > 1$, and let $\ell(k-1) \in \mathbb{N}$ and $\ell(1) \in \mathbb{N}$ be defined by* (5.3.7)*. Then*

$$\sum_{j=k}^{m} |c_j|^2 \leq \sum_{j=1}^{\ell(1)} |w_j|^2 + \sum_{j=\ell(k-1)+1}^{n} |w_j|^2. \tag{5.3.10}$$

Proof. The spectral coefficients c_j and the measure μ_n as given in (4.4.13) in Subsection 4.4.2 satisfy the inequalities given in Corollary 4.4.16. Particularly, the upper bound in (4.4.57c) in Corollary 4.4.16 yields

$$\sum_{j=k}^{m} |c_j|^2 \leq \mu_n((a, \theta_1) \cup (\theta_{k-1}, \infty)).$$

With the index ℓ as in (5.3.7), the measure μ_n as in (4.4.13) satisfies

$$\mu_n((a, \theta_1) \cup [\theta_{k-1}, \infty)) = \sum_{j=1}^{\ell(1)} |w_j|^2 + \sum_{j=\ell(k-1)+1}^{n} |w_j|^2,$$

which entails (5.3.8). □

The upper bound in (5.3.10) includes spectral coefficients w_j related to eigenvalues λ_j which are located on the left side of the spectrum. In the setting of Proposition 5.3.2 these spectral coefficients can be of larger size. Nevertheless, numerical illustrations in Section 5.4 suggest that Assumption (A2) holds with $\widehat{\varepsilon} \approx \varepsilon$ in the present setting.

Further choices for the poles of the rational Krylov subspace are not discussed in theory. However, for the numerical experiments in Section 5.4 we also consider Krylov subspaces according to a single complex pole or multiple poles, respectively.

We proceed with remarks on quasi-orthogonal residual (qor-)Krylov approximations.

The qor-Krylov approximation with a preassigned eigenvalue. The rational qor-Krylov approximation is introduced in Subsection 4.3.2 (Chapter 4) and is given by

$$U_m \mathrm{e}^{-\mathrm{i}tB_m} x \approx \mathrm{e}^{-\mathrm{i}tA} u, \tag{5.3.11}$$

where U_m denotes an orthonormal basis of the rational Krylov subspace $\mathcal{Q}_m(A, u)$, the vector x refers to the representation of u in the respective orthonormal basis, and $B_m \in \mathbb{C}^{m\times m}$ denotes the qor-Krylov representation which is introduced in Subsection 4.3.2 (Chapter 4). One of the eigenvalues of B_m is preassigned, and we denote it by $\xi \in \mathbb{R}$. In the present setting we reuse the denotations $\theta_1, \ldots, \theta_m$ and $c_1, \ldots, c_m$ for the eigenvalues of B_m and the respective spectral coefficients of x, respectively.

The error of the qor-Krylov approximation is bounded similar to Proposition 5.3.2: Following Proposition 4.3.5 in Subsection 4.3.2 (Chapter 4), the identity $U_m r(B_m)x = r(A)u$ holds true for $r \in \Pi_{m-1}/q_{m-1}$. Furthermore, we need to exclude the case $\theta_1 < a$, where θ_1 denotes the smallest eigenvalue of B_m (see also Proposition 4.3.1 in Section 4.3, Chapter 4 for more details considering the eigenvalues of B_m). Then the proof of Proposition 5.3.2 remains valid in the setting of the qor-Krylov approximation. Thus, with the assumptions and denotations of Proposition 5.3.2 we have

$$\|U_m\, \mathrm{e}^{-\mathrm{i}\,tB_m} x - \mathrm{e}^{-\mathrm{i}\,tA} u\|_{\mathsf{M}} \leq 2 \min_{r\in R_{\varrho,b}} \max_{\lambda\in[a,b]} |r(\lambda) - \mathrm{e}^{-\mathrm{i}\,t\lambda}| \|u\|_{\mathsf{M}} + (1+\varrho)(\varepsilon+\widehat{\varepsilon})\|u\|_{\mathsf{M}}. \tag{5.3.12}$$

For the qor-Krylov representation B_m, bounds on the spectral coefficients similar to Corollary 5.3.3 and 5.3.4 hold true. The main reason why we present the qor-Krylov approximation at this point is the possibility to preassign one of the eigenvalues $\theta_1, \ldots, \theta_m$, which we exploit to simplify bounds on the spectral coefficients, in order to deduce Assumption (A2) from Assumption (A1). However, in the setting of the following remark, Assumption (A2) with $\widehat{\varepsilon} = \varepsilon$ can be concluded from (A1) but the case $\theta_1 < a$ can occur, and the error bound (5.3.12) as such does not hold in this case.

Remark 5.3.5. *We consider the rational Krylov subspace with a single pole $s < \lambda_1$ of multiplicity $m-1$. Let one of the eigenvalues of B_m be preassigned at $\xi = b$, where b corresponds to the right interval boundary of $[a, b]$ according to* (5.3.12)*. Similar to the Ritz values, we assume the ordering $\theta_1 < \theta_2 < \ldots < \theta_m$ for the eigenvalues of B_m. Furthermore, we assume that at least one of the eigenvalues of B_m is larger than b, and*

let $k \leq m$ denote the smallest index such that $\theta_k > b$. Thus, we have $\theta_{k-1} = b$ for the preassigned eigenvalue $\xi = b$ of B_m. The index $\ell = \ell(k-1)$ as in (5.3.7) *is well-defined for the current setting, and similar to Corollary 5.3.3, we have*

$$\sum_{j=k}^{m} |c_j|^2 \leq \sum_{j=\ell(k-1)+1}^{n} |w_j|^2.$$

Thus, we can recreate the result of Corollary 5.3.3 for the spectrum of the qor-Krylov representation B_m in the current setting. Furthermore, with $\theta_{k-1} = b$ and $\ell(k-1)$ as given in (5.3.7) *we have $\lambda_{\ell(k-1)+1} > b$, and previous remarks, especially* (5.3.9)*, imply that Assumption* (A2) *with $\widehat{\varepsilon} = \varepsilon$ can be concluded from* (A1) *in this case.*

Unfortunately, the case $\theta_1 < a$ can occur, see Proposition 4.3.1 in Section 4.3. In the case of $\theta_1 < a$ the error bound (5.3.12) *does not hold as such and an a priori statement on the error of the qor-Krylov approximation cannot be given in general.*

As a second approach to gain theoretical benefit from the qor-Krylov approximation we consider the case of a single pole $s \in (\lambda_1, \lambda_n)$. Let B_m be the qor-Krylov representation with a preassigned eigenvalue $\xi = a$, where a corresponds to the left interval boundary of $[a, b]$ according to (5.3.12) and we further assume $a < \lambda_1$. In this case we have $\theta_1 = a$ and $\theta_2, \ldots, \theta_m \in (\lambda_1, \lambda_n)$, and the error bound in (5.3.12) holds true.

We proceed with bounds on the spectral coefficients in the setting of the qor-Krylov representation, based on results given in Subsection 4.4.3 in Chapter 4.

Corollary 5.3.6. *[see Remark 4.4.17] Let B_m be the qor-Krylov representation corresponding to the rational Krylov subspace with a single pole $s \in \mathbb{R}$ of multiplicity $m-1$ and a preassigned eigenvalue $\xi = a < \lambda_1$. Thus we have $\theta_1 = a$. We further assume $\theta_1 < s < \theta_m$. Furthermore, let $c_1, \ldots, c_m$ denote the spectral coefficients of x w.r.t. the ℓ^2-orthonormal eigenbasis of B_m. Let the index k be given with $k > 1$, and $\ell = \ell(k-1)$ be defined by* (5.3.7)*. Then,*

$$\sum_{j=k}^{m} |c_j|^2 \leq \sum_{j=\ell(k-1)+1}^{n} |w_j|^2. \tag{5.3.13}$$

Proof. Following Remark 4.4.17, we have

$$\sum_{j=k}^{m} |c_j|^2 \leq \mu_n((\theta_{k-1}, \infty)),$$

and with $\mu_n([\theta_{k-1}, \infty)) = \sum_{j=\ell(k-1)+1}^{n} |w_j|^2$, this implies (5.3.13). □

In contrast to Corollary 5.3.4, which corresponds to A_m, the upper bound in Corollary 5.3.6, which corresponds to B_m, does not include spectral coefficients $w_1, \ldots, w_{\ell(1)}$ which can be of larger size in the setting of Proposition 4.3.1. Nevertheless, Corollary 5.3.6 cannot be used to justify Assumption (A2) with $\widehat{\varepsilon} = \varepsilon$ from (A1) in an a priori manner (see also arguments following Corollary 5.3.3).

Remark 5.3.7. *We consider a special case for which Assumption* (A1) *implies* (A2) *with* $\widehat{\varepsilon} = \varepsilon$. *Let* B_m *be the qor-Krylov representation corresponding to the rational Krylov subspace with a single pole* $s = b$ *of multiplicity* $m-1$, *and a preassigned eigenvalue* $\xi = a < \lambda_1$. *Let there be an index* $k > 1$ *with* $\theta_{k-1} < s < \theta_k$ *as in* (A2). *We apply Theorem 4.4.11 (Subsection 4.4.3 in Chapter 4) for the index set* $I_m = \{k, \dots, m\}$ *as given in* (4.4.34). *The upper bound in* (4.4.36) *yields*

$$\sum_{j=k}^{m} |c_j|^2 < \mu_n(R_1^{\mathrm{o}}).$$

With $\theta_1 = a$, *the set* R_1^{o} *as given in* (4.4.34) *in Subsection 4.4.3 simplifies to* $R_1^{\mathrm{o}} = (s, \infty)$. *The measure* μ_n *is given in* (4.4.13) *in Subsection 4.4.2 and satisfies* $\mu_n((s,\infty)) = \sum_{j=\ell}^{n} |w_j|^2$ *for the index* ℓ *as in Assumption* (A1). *Thus, we have*

$$\sum_{j=k}^{m} |c_j|^2 < \sum_{j=\ell}^{n} |w_j|^2$$

and Assumption (A1) *implies* (A2) *with* $\widehat{\varepsilon} = \varepsilon$.

For the qor-Krylov approximation as in Remark 5.3.7, we can conclude (A2) in Proposition 5.3.2 based on Assumption (A1) in an a priori manner. However, the pole $s = b$ is not the best choice considering the performance of the Krylov approximation in some cases.

5.4 Numerical illustrations concerning the assumptions of Proposition 5.3.2

In this section we present the following numerical experiments: We consider matrices $A \in \mathbb{R}^{n\times n}$ corresponding to a finite-difference discretization of a Hamiltonian operator (related to a Schrödinger equation, see (5.4.2) below) for different problem sizes n. For each problem size we consider different initial vectors $u \in \mathbb{C}^n$ which fulfill Assumption (A1) for different choices of $\varepsilon > 0$: In Subsection 5.4.1 the initial vectors u correspond to a smooth initial state (in reference to a given differential equation) to which we add a randomized perturbation. The perturbation size directly affects the choice of ε corresponding to Assumption (A1). In Subsection 5.4.2 we consider u resulting from of a numerical time propagation step starting at a smooth initial state. The approximation error of the numerical time propagation step entails Assumption (A1) to hold with different choices of $\varepsilon > 0$. For each choice of A and u, we construct the respective rational Krylov subspace, whereat different choices of poles will be considered, and we test whether Assumption (A2) holds true with $\widehat{\varepsilon} \approx \varepsilon$.

The matrix A in the present test setting. For the spatial domain of the discretization we choose $[-10, 10] \subset \mathbb{R}$, and we define the grid width

$$h = 20/(n-1),$$

where the problem size n corresponds to the number of grid points. Thus, the respective grid points satisfy

$$\eta_j = 20\,(j-1)/(n-1) - 10, \qquad j = 1, \dots, n.$$

In the following we consider the problem sizes $n = 800, 1600, 2400$. For each n we consider $A \in \mathbb{R}^{n\times n}$ to be the respective finite-difference discretization of $-\Delta + V$ with $V = V(x) = 4x^4 - 15x^2 \in \mathbb{R}$, i.e.,

$$A = 1/h^2 \text{ tridiag}(-1, 2, -1) + \text{diag}(V(\eta_j)), \tag{5.4.1}$$

where $\text{diag}(V(\eta_j))$ denotes a diagonal matrix with diagonal entries $V(\eta_1), \ldots, V(\eta_n)$. In addition to (5.4.1), we set $A_{n,1} = A_{1,n} = -1/h^2$ conforming to periodic boundary conditions. In view of (5.2.1), we remark that $A \in \mathbb{R}^{n\times n}$, together with an initial vector $u \in \mathbb{C}^n$, corresponds to a discretized Schrödinger equation,

$$\phi'(t) = -\mathrm{i}A\phi(t), \quad \phi(0) = u, \quad \phi(t) \in \mathbb{C}^n, \quad t \geq 0, \tag{5.4.2}$$

whereat $V(x)$ given above represents a double well potential in this context. This problem also appears in Section 3.5.3 (Chapter 3) and references therein.

For the current problem setting we consider an inner product on $\mathbb{C}^n$ which fits to a discretized version of the L^2-inner product on the function space of square-integrable functions on the spatial domain $[-10, 10]$. Thus, we define

$$(u, v)_{\mathsf{M}} = h\,(u, v)_2, \quad u, v \in \mathbb{C}^n,$$

where h denotes the grid width.

For the eigenvalues of A, i.e., $\lambda_1 < \ldots < \lambda_n \in \mathbb{R}$, we have $\lambda_1 > a$ with $a = -9$ for any n, and $\lambda_n \approx 4.44 \cdot 10^4$, $6.35 \cdot 10^4$, $9.55 \cdot 10^4$ for $n = 800, 1600, 2400$, respectively, where λ_n refers to the rightmost eigenvalue of $A \in \mathbb{R}^{n\times n}$.

In the following, the vector $\phi_0 \in \mathbb{R}^n$ corresponds to a discretized Gaussian wave packet,

$$\phi_0 = (\phi_{0,1}, \ldots, \phi_{0,n})^{\mathsf{T}}, \quad \text{with} \quad \phi_{0,j} = (0.4\pi)^{-1/4} \exp(-(\eta_j + 2.5)^2/(0.8)). \tag{5.4.3a}$$

With $Q = (q_1, \ldots, q_n) \in \mathbb{R}^{n\times n}$ denoting the M-orthonormal eigenbasis of A as introduced before, we introduce the notation $w_1^0, \ldots, w_n^0$ for the spectral coefficients of ϕ_0 in this eigenbasis,

$$w_j^0 = (q_j, \phi_0)_{\mathsf{M}}, \quad j = 1, \ldots, n. \tag{5.4.3b}$$

5.4.1 A smooth initial vector with a randomized perturbation

For our first numerical experiments we construct the initial vector u by adding a randomized perturbation to the vector ϕ_0 given in (5.4.3a). This procedure is applied for each problem size, $n = 800, 1600, 2400$, and for different scaling factors $\delta = 10^{-3}, 10^{-4}, 10^{-5}$, which scale the perturbation size as stated below. This yields a total of 9 different initial vectors.

We construct these initial vectors $u \in \mathbb{R}^n$ such that Assumption (A1) holds true, whereat the choice of ε in (A1) depends on δ and is specified below. Considering (A1), the eigenvalues $\lambda_1 < \ldots < \lambda_n$ and spectral coefficients $w_1, \ldots, w_n$ refer to the spectrum of the matrix $A \in \mathbb{R}^{n\times n}$ given in the previous paragraph. In view of Assumption (A1) we perturb spectral coefficients related to eigenvalues which are located outside of an interval $[a, b]$. To this end, we choose

$$b = 2100, 2700, 3400 \text{ for } \delta = 10^{-3}, 10^{-4}, 10^{-5}, \text{ respectively,} \tag{5.4.4a}$$

and $a = -9$ as given in the previous paragraph. Let $Q = (q_1, \ldots, q_n) \in \mathbb{R}^{n\times n}$ denote the M-orthonormal eigenbasis of A as introduced before. With uniformly distributed random numbers $r_1, \ldots, r_n \in (0, 1)$ we define

$$\widehat{v} = \sum_{j=\ell}^{n} r_j\, q_j, \quad \text{where } \ell \in \mathbb{N} \text{ conforms to } \lambda_{\ell-1} < b < \lambda_\ell. \tag{5.4.4b}$$

For each choice of n and δ, we set

$$u = \phi_0 + \delta\, \widehat{v}/\|\widehat{v}\|_{\mathsf{M}}. \tag{5.4.5}$$

As introduced before, the notation $w_1, \ldots, w_n$ refers to the spectral coefficients of u in the M-orthonormal eigenbasis of A. The spectral coefficients $w_1, \ldots, w_n$, together with the spectral coefficients $w_1^0, \ldots, w_n^0$ of ϕ_0 given in (5.4.3b), are illustrated in Figure 5.1.

For each choice of n and δ, the spectral decomposition of u w.r.t. A satisfies Assumption (A1) with $\varepsilon = 1.02 \cdot \delta$, where the choice of b and ℓ in Assumption (A1) and (5.4.4) coincides and depends on δ.

Testing whether Assumption (A2) **holds with $\widehat{\varepsilon} \approx \varepsilon$ in this setting.** To this end we consider rational Krylov subspaces $\mathcal{Q}_m(A, u)$ with given poles. We construct the Rayleigh quotient $A_m \in \mathbb{C}^{m\times m}$ given in (5.2.8) or the qor-Krylov representation $B_m \in \mathbb{C}^{m\times m}$ (for a given preassigned eigenvalue $\xi \in \mathbb{R}$) and refer to $\theta_1 < \ldots < \theta_m \in \mathbb{R}$ as the eigenvalues of the respective matrix. For a single pole $s \in \mathbb{C}$ of multiplicity $m-1$ the Rayleigh quotient A_m is computed by Algorithm 4.1 (see Section 4.2 in Chapter 4), and the matrix B_m is computed by Algorithm 4.5 (see Subsection 4.3.2 in Chapter 4). For the action of the matrix inverse in Algorithm 4.1 and 4.5 we apply a direct solver (corresponding to the Matlab backslash operator) if not stated otherwise. For a numerical example concerning multiple poles we apply the `rat_krylov` procedure [BG15, Algorithm 3.1]. For any of these experiments the Krylov subspace is constructed w.r.t. the M-inner product.

Let $b \in \mathbb{R}$ correspond to (5.4.4a). We first assume that at least one of the eigenvalues $\theta_1 < \ldots < \theta_m$ is larger than b, and we define $k(m) \le m$ as the smallest index such that $\theta_{k(m)} > b$. Here the index $k(m)$ depends on m, the initial vector u and the choice of b. Let $c_1, \ldots, c_m$ denote the spectral coefficients of x as introduced before. We define

$$\zeta_m = \Big(\sum_{j=k(m)}^{m} |c_j|^2 \Big)^{1/2}, \quad \text{where } k(m) \text{ conforms to } \theta_{k(m)-1} \le b < \theta_{k(m)}. \tag{5.4.6}$$

Thus, Assumption (A2) holds true with $\widehat{\varepsilon} = \zeta_m$ and the respective choice of b and $k = k(m)$ for each $m \in \mathbb{N}$. In Figure 5.2 and 5.4 we show ζ_m over m for different choices of n and δ, and for rational Krylov subspaces with different sets of poles. For the case $\theta_1 < \ldots < \theta_m < b$, which is likely valid for smaller choices of m, we consider ζ_m to be zero and no symbols are added to the plot. The number of eigenvalues θ_j which are larger than b slowly increases with m and is equal to $m - k(m) + 1$. In the following figures the values ζ_m are marked by symbols (where ($\square$), ($\circ$) and ($\times$) correspond to results for $n = 800, 1600, 2400$, respectively). Additionally, values of ζ_m which refer to the same initial vector u and for which the value of $m - k(m) + 1$ matches are connected by a curve.

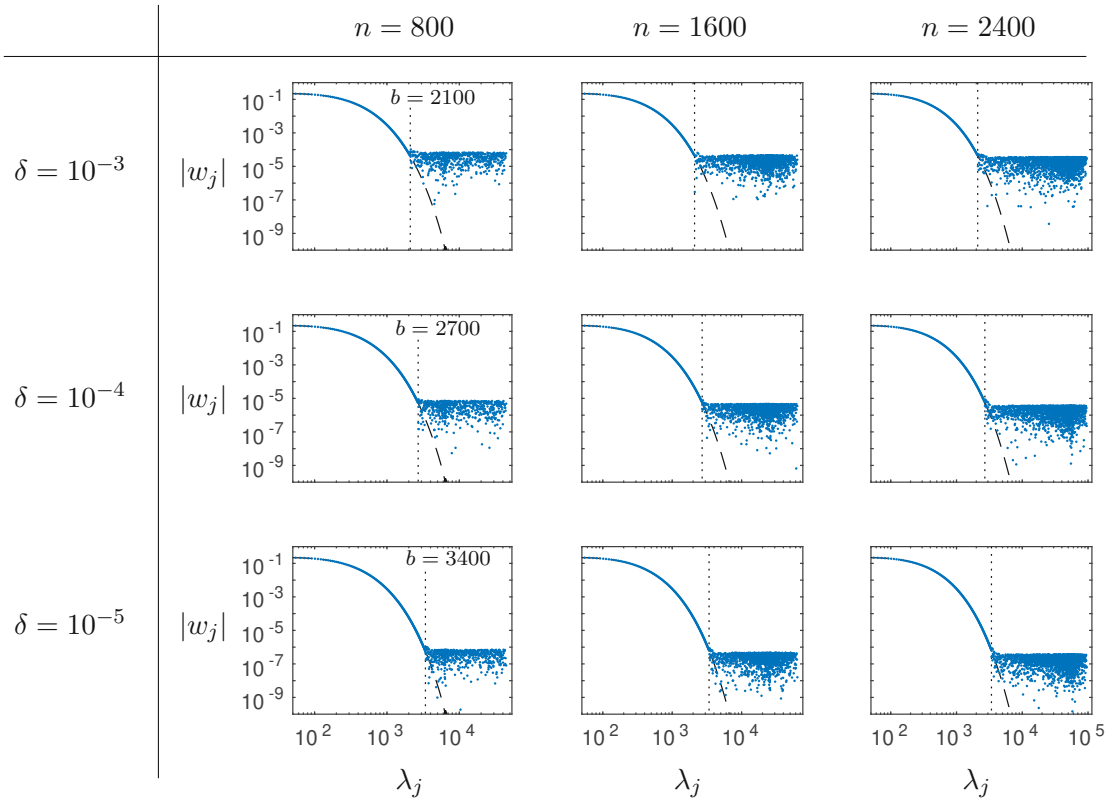


Figure 5.1: The plots show spectral coefficients of u given in (5.4.5) for different choices of n and δ. The dot marks in each plot illustrate the spectral coefficients $|w_j|$ of u versus the corresponding values of λ_j. The spectral coefficients $|w_j^0|$ of the vector χ for the respective choices of n, see (5.4.3), are located on the dashed line but not explicitly shown in the plot. The spectral coefficients of u correspond to the spectral coefficients of ϕ_0 for $\lambda_j < b$ (the dots cover the dashed line for $\lambda_j < b$) and perturbed by random values otherwise, see also (5.4.4). The dotted vertical line show b which depends on δ and is given in (5.4.4a).

In Corollary 5.3.3, 5.3.4 and 5.3.6 a motivation for the following statement is given: Assumption (A1) entails Assumption (A2) with $\widehat{\varepsilon} \approx \varepsilon$. For the present numerical examples Assumption (A1) holds with $\varepsilon \approx \delta$ and Assumption (A2) holds with $\widehat{\varepsilon} = \zeta_m$, thus, $\zeta_m \lessapprox \delta$ implies that this conclusion holds true.

In Figure 5.2 a) we consider the Rayleigh quotient A_m for $\mathcal{Q}_m(A, u)$ with a single pole $s = -10$ of multiplicity $m-1$. With $s < \lambda_1$ this conforms to the setting of Corollary 5.3.3. In Figure 5.2 b) and 5.2 c) we consider the Rayleigh quotient A_m for a single pole $s = 100$ (in b)) and $s = 400$ (in c)) with multiplicity $m-1$. These poles are located within the range of the Ritz values $\theta_1, \ldots, \theta_m$ for a large enough choice of m, thus, Corollary 5.3.4 applies. In Figure 5.2 d) we consider the qor-Krylov representation B_m for a single pole $s = 100$ with multiplicity $m-1$ and a preassigned eigenvalue $\xi = -10$. This conforms to the setting of Corollary 5.3.6.

For all of the plots in Figure 5.2 we observe $\zeta_m \approx \delta$ and $\zeta_m \to \delta$ for larger choices of m independent of the problem size n. We remark that in Figure 5.2 a)–d) the number of eigenvalues θ_j which are larger than b strictly increases with m for these examples, and the exact points of increase match the jumps of ζ_m.

Additionally, for the pole $s = 100$ we compare ζ_m corresponding to the Rayleigh quotient A_m in Figure 5.2 b) and the qor-Krylov representation B_m in Figure 5.2 d). These examples fit to the setting of Corollary 5.3.4 and 5.3.6, respectively, which yield bounds on ζ_m (set $k = k(m)$ in Corollary 5.3.4 and 5.3.6). The bound in Corollary 5.3.6 is tighter compared to the bound in Corollary 5.3.4. However, the results in Figure 5.2 b) and d) are similar with the exception that $\theta_m < b$ (hence, no eigenvalue θ_j is located outside of $[a, b]$) holds true for larger choices of m in Figure 5.2 d).

In Figure 5.3 we consider the qor-Krylov representation B_m for a single pole $s = b$ of multiplicity $m-1$ and a preassigned eigenvalue $\xi = -10$. Here, the pole s depends on the choice of the starting vector. In Remark 5.3.7 above, we show that Assumption (A1) implies (A2) with $\widehat{\varepsilon} = \varepsilon$ in this case. This corresponds to $\zeta_m \leq \delta$, which is verified for the numerical example in the plot. The property that Assumption (A1) implies (A2) allows to apply Proposition 5.3.2 in an a priori sense which is desirable in theory. However, the pole $s = b$ is potentially not the best choice for the accuracy of a Krylov approximation in practice. Approximately half of the eigenvalues θ_j are located outside of $[a, b]$ in this case, which can be due to the choice of the pole $s = b$. Considering the performance of a Krylov approximation to a matrix function, it can be favorably when eigenvalues θ_j are located close to the relevant eigenvalues of A.

In Figure 5.4 a) we consider the Rayleigh quotient A_m for $\mathcal{Q}_m(A, u)$ with a single pole $s = 100 + 100\,\mathrm{i} \in \mathbb{C}$ of multiplicity $m-1$. In Figure 5.4 b) we consider the Rayleigh quotient A_m for $\mathcal{Q}_m(A, u)$ corresponding to the denominator

$$q_{m-1}(\lambda) = \lambda^j(\lambda-100)^j(\lambda-200)^j(\lambda-300)^j(\lambda-400)^j, \text{ with } m = 5j+1 \text{ for } j \in \mathbb{N}. \quad (5.4.7)$$

Thus, the denominator in (5.4.7) corresponds to a rational Krylov subspace with multiple poles s_k of higher multiplicity, where $s_k \in \{0, 100, 200, 300, 400\}$ for $k = 1, \ldots, m-1$.

Theoretical bounds on spectral coefficients concerning the rational Krylov subspace in the setting of Figure 5.4 a), namely a single pole $s \in \mathbb{C} \setminus \mathbb{R}$, and Figure 5.4 b), namely

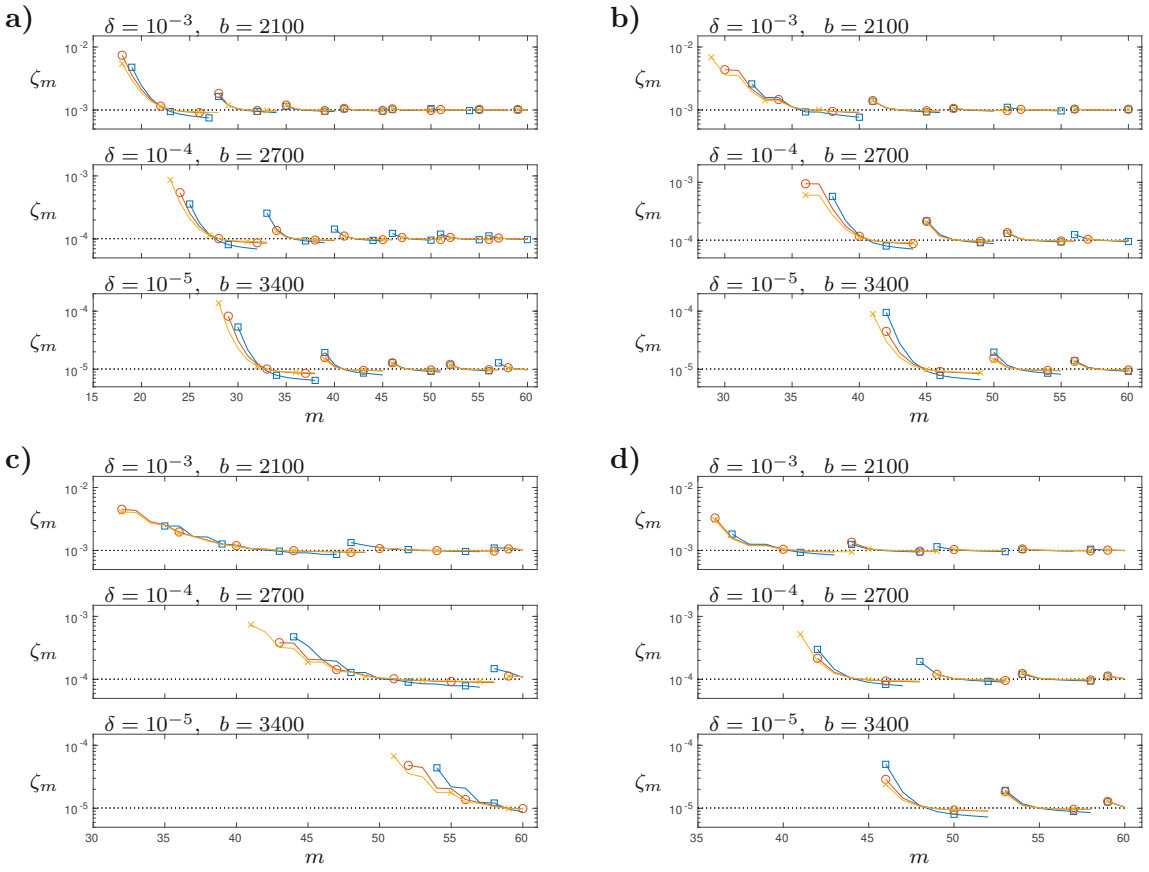


Figure 5.2: In **a)**–**d)** we show ζ_m given in (5.4.6). In each figure the three different graphs show result for different choices of δ, i.e., $\delta = 10^{-3}, 10^{-4}, 10^{-5}$ as denoted on top of each graph and illustrated by a dotted line. For each δ the choice of b is given in (5.4.4a). The different symbols refer to the different problem sizes, i.e., $n = 800$ ($\square$), $n = 1600$ ($\circ$) and $n = 2400$ ($\times$), and additionally values of ζ_m which refer to the same choice of n and where A_m (or B_m in Figure d)) has the same number of eigenvalues larger than b are connected by a curve. The initial vector u is given in (5.4.5).
– Figure **a)** The given results refer to the spectrum of the Rayleigh quotient A_m for the rational Krylov subspace with a single pole $s = -10$ of multiplicity $m-1$.
– Figure **b)** Similar to Figure a) with $s = 100$.
– Figure **c)** Similar to Figure a) with $s = 400$.
– Figure **d)** The given results refer to the spectrum of the qor-Krylov representation B_m for the rational Krylov subspace with a single pole $s = 100$ of multiplicity $m - 1$ and a preassigned eigenvalue at $\xi = -10$.

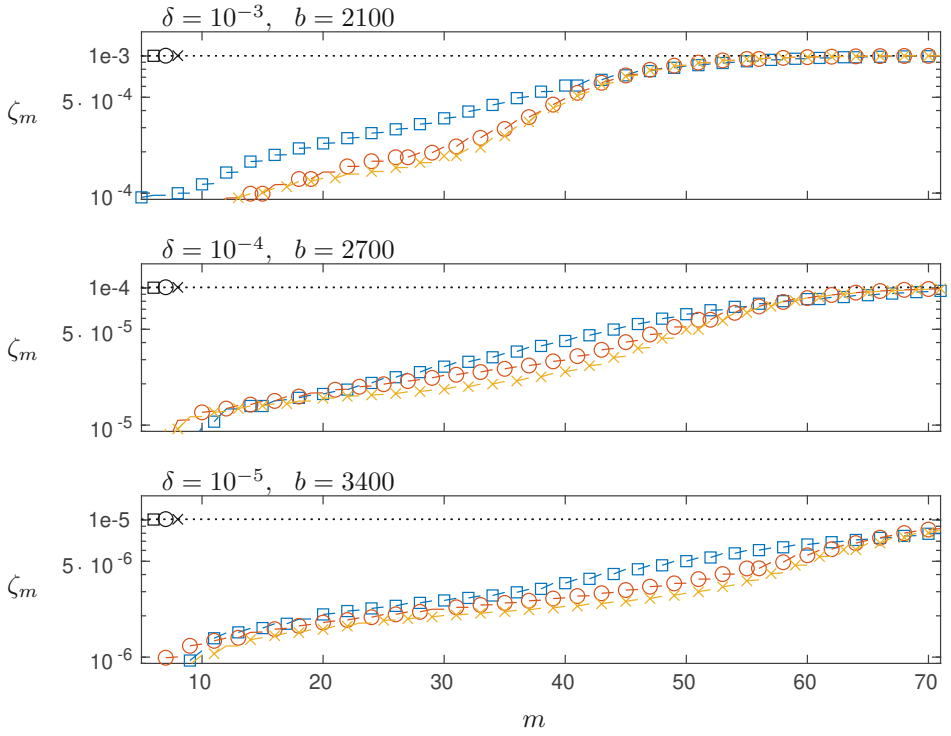


Figure 5.3: This plot shows ζ_m for the initial vector u given in (5.4.5) and the problem size $n = 800$ ($\square$), $n = 1600$ ($\circ$) and $n = 2400$ ($\times$). For details we refer to the caption of Figure 5.2. The given results refer to the spectrum of the qor-Krylov representation B_m for the rational Krylov subspace with a single pole $s = b$ of multiplicity $m - 1$, and a preassigned eigenvalue $\xi = -10$.

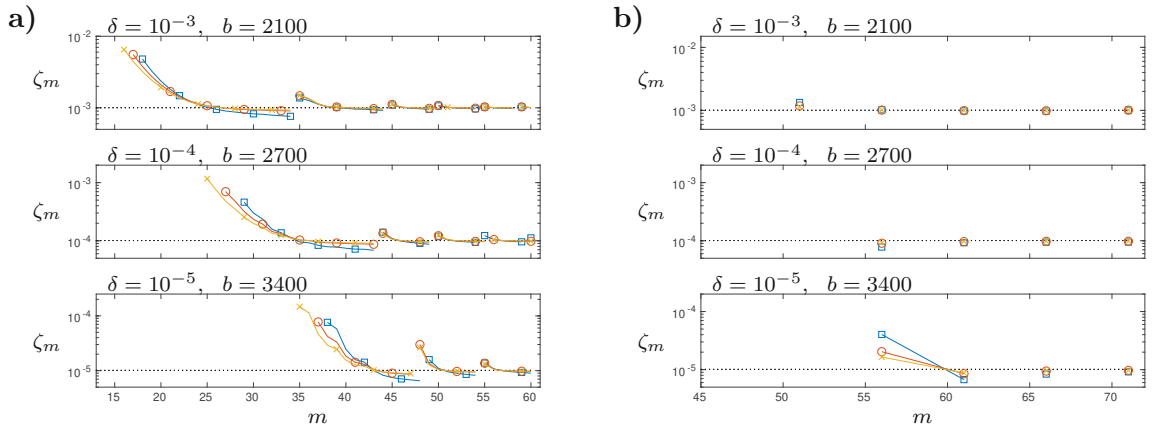


Figure 5.4: In **a)** and **b)** the plots show ζ_m for the initial vector u given in (5.4.5) and the problem size $n = 800$ (□), $n = 1600$ (○) and $n = 2400$ (×). For details we refer to the caption of Figure 5.2.
– Figure **a)** The given results refer to the spectrum of A_m for the rational Krylov subspace with a single pole $s = 100 + 100\,\mathrm{i}$ of multiplicity $m - 1$.
– Figure **b)** The given results refer to the spectrum of A_m for the rational Krylov subspace with multiple poles $s_k \in \{0, 100, 200, 300, 400\}$ for $k = 1, \ldots, m-1$ where each pole is of an equal multiplicity. Thus, we consider the Krylov dimensions $m = 6, 11, 16, \ldots$, where $m = 51$ and $m = 56$ are the smallest choices of m for which ξ_m is not equal zero and visible in the plots. The denominator q_{m-1} of this rational Krylov subspace is explicitly given in (5.4.7).

multiple poles, are not discussed in the present work. Nevertheless, similar to the results of Figure 5.2 we observe $\zeta_m \approx \delta$ and $\zeta_m \to \delta$ for larger choices of m in Figure 5.4.

5.4.2 The influence of a preceding inexact time propagation step

Similar to the previous numerical experiments we consider the rational Krylov approximation for an initial vector u and we discuss whether the assumptions of Proposition 5.3.2 are reasonable, especially whether Assumption (A2) can be deduced from Assumption (A1) with $\widehat{\varepsilon} \approx \varepsilon$. In the following examples we consider the initial vector u to originate from an approximated solution of the differential equation (5.4.2): Let $A \in \mathbb{R}^{n\times n}$ be defined in (5.4.1) with periodic boundary conditions and let $\phi_0 \in \mathbb{R}^n$ be defined in (5.4.3a). For a time step $\tau > 0$ which will be fixed later we define

$$\phi_\tau = \mathrm{e}^{-\mathrm{i}\tau A}\phi_0 \in \mathbb{C}^n. \tag{5.4.8}$$

The matrix exponential in (5.4.8) conforms to a time evolution operator for the differential equation (5.4.2): The solution $\phi(t) \in \mathbb{C}^n$ of (5.4.2) with initial vector $\phi(0) = \phi_0$ satisfies $\phi(\tau) = \phi_\tau$.

The notation $w_j^0 = (q_j, \phi_0)_{\mathsf{M}}$ has been introduced in (5.4.3b) and refers to the spectral coefficients of ϕ_0 in the M-orthonormal eigenbasis of A. The matrix exponential in (5.4.8)

is unitary. Thus, it applies a phase shift to the spectral coefficients of ϕ_0 and we have $|(q_j, \phi_0)_{\mathsf{M}}| = |(q_j, \mathrm{e}^{-\mathrm{i}\tau A}\phi_0)_{\mathsf{M}}|$ for $\tau \geq 0$. For w_j^0 in (5.4.3b) and ϕ_τ in (5.4.8) this implies

$$|w_j^0| = |(q_j, \phi_\tau)_{\mathsf{M}}|. \tag{5.4.9}$$

We now consider u to be a numerical approximation of ϕ_τ. When this approximation is accurate and the spectral coefficients of ϕ_τ satisfy an upper bound as in Assumption (A1), then the spectral coefficients of u satisfy the same upper bound up to a small deviation: For $\delta \geq 0$,

$$\|u - \phi_\tau\|_{\mathsf{M}} \leq \delta \quad \text{implies} \quad \Big(\sum_{j=\ell}^{n} |w_j|^2\Big)^{1/2} \leq \Big(\sum_{j=\ell}^{n} |w_j^0|^2\Big)^{1/2} + \delta. \tag{5.4.10}$$

In Subsection 5.4.1 we have discussed Assumption (A1) for the vector ϕ_0 with an additional perturbation, and the results therein carry over to the vector ϕ_τ which spectral coefficients also correspond to $|w_j^0|$: The vector ϕ_τ satisfies Assumption (A1) for a proper choice of $[a, b]$ and ε, similarly as for ϕ_0. With (5.4.10) this carries over to u, when u denotes a sufficiently accurate approximation of ϕ_τ.

An initial vector u resulting from a preceding rational Krylov approximation to (5.4.8). For a given $\tau > 0$, the rational Krylov approximation (5.2.9) applied to the right-hand side of (5.4.8) yields an approximation to ϕ_τ. Here we consider the rational Krylov subspace $\mathcal{Q}_{\widetilde{m}}(A, \phi_0)$ with a single pole $s = -10$ of multiplicity $\widetilde{m} - 1$, where we choose $\widetilde{m} = 40$ for the Krylov dimension. The corresponding Krylov basis $U_{\widetilde{m}}$, the Rayleigh quotient $A_{\widetilde{m}}$ and x are provided by Algorithm 4.1 (see Chapter 4). This algorithm requires applications of the matrix inverse of $A - sI$, for which we either apply a direct solver or an iterative method. This results in two different choices of u:

- For the first approach we implement Algorithm 4.1 using a direct solver for the underlying matrix inverse. Let $U_{\widetilde{m}}$, $A_{\widetilde{m}}$ and x be the result of Algorithm 4.1, where the matrix inverses are computed by a

  $$\text{Cholesky decomposition, then we define} \quad u_{\mathrm{ds}} = U_{\widetilde{m}} \mathrm{e}^{-\mathrm{i}\tau A_{\widetilde{m}}} x \in \mathbb{C}^n, \tag{5.4.11}$$

  with τ-values specified in (5.4.13) below.

- For the second approach we apply an iterative method to compute the action of the matrix inverse of $A - sI$ in Algorithm 4.1, namely, using the Matlab `pcg` procedure realizing a preconditioned conjugate gradient method. For the current examples, using a direct solver is certainly the better choice and we choose the iterative variant solely for the purpose of testing. For `pcg` we choose the tolerance δ_{cg} specified below, following (5.4.13). As preconditioner we choose the diagonal entries of $A - sI$. Let $U_{\widetilde{m}}$, $A_{\widetilde{m}}$ and x be the result of Algorithm 4.1 using

  $$\texttt{pcg}, \text{ then we define} \quad u_{\mathsf{cg}} = U_{\widetilde{m}} \mathrm{e}^{-\mathrm{i}\tau A_{\widetilde{m}}} x \in \mathbb{C}^n, \tag{5.4.12}$$

  where τ is specified in (5.4.13). The approach of using an iterative method to compute the matrix inverse for the rational Krylov approximation is also discussed in [vdEH06, Section 5] and others.

Similar to the previous test setting, we consider the vectors $u_{\mathtt{cg}}$ and u_{ds} for different problem sizes $n = 800, 1600, 2400$, and $\delta = 10^{-3}, 10^{-4}, 10^{-5}$, respectively. Here δ refers to a tolerance on the approximation error as given in (5.4.10); we consider the following setting: Due to properties of the rational Krylov approximation, $u_{\mathtt{cg}}$ and u_{ds} yield an accurate approximation to ϕ_τ if the underlying time step τ is sufficiently small. Additionally, for $u_{\mathtt{cg}}$ the tolerance $\delta_{\mathtt{cg}}$ for `pcg` procedure has to be sufficiently small. For each choice of δ (the influence of n is negligible here) we choose τ and $\delta_{\mathtt{cg}}$ such that $\|u_{\mathtt{cg}} - \phi_\tau\|_{\mathsf{M}} \approx \delta$ holds true, i.e.,

$$\tau = 6.5 \cdot 10^{-3}, 3.8 \cdot 10^{-3}, 2.2 \cdot 10^{-3}, \text{ for } \delta = 10^{-3}, 10^{-4}, 10^{-5}, \text{ respectively,} \tag{5.4.13}$$

and $\delta_{\mathtt{cg}} = \delta/\tau \cdot 10^{-4}$. The deviation of $u_{\mathtt{cg}}$ and ϕ_τ has its origin partly in the inexact construction of the rational Krylov subspace (depending on $\delta_{\mathtt{cg}}$) and partly in the approximation of the matrix exponential (5.2.9) (scaling with the time step τ). For comparison reasons the same choice of τ is used for u_{ds}, and we have

$$\|u - \phi_\tau\|_{\mathsf{M}} \leq \delta, \quad \text{for} \quad u = u_{\mathtt{cg}}, u_{\mathrm{ds}}. \tag{5.4.14}$$

Let w_j denote the spectral coefficients of u in the M-orthonormal eigenbasis of A where $u = u_{\mathtt{cg}}, u_{\mathrm{ds}}$. The spectral coefficients w_j of $u = u_{\mathtt{cg}}$ and $u = u_{\mathrm{ds}}$ are illustrated in Figure 5.5 and 5.6, respectively. Similar to Subsection 5.4.1, we want u to comply with Assumption (A1) for a given interval $[a, b]$, and a suitable choice of ε. As previously we choose $a = -9$. In the present subsection we consider two different choices of b, which we refer to as b_1 and b_2: We choose b_1 s.t. $|w_j|$ shows a significant deviation from $|w_j^0|$ for $u = u_{\mathtt{cg}}, u_{\mathrm{ds}}$ likewise for $\lambda_j > b_1$ (considering Figure 5.5 and 5.6), namely,

$$b_1 = 1600, 2600, 3000, \text{ for } \delta = 10^{-3}, 10^{-4}, 10^{-5}, \text{ respectively.} \tag{5.4.15}$$

We choose b_2 such that the spectral coefficients of $u_{\mathtt{cg}}$ significantly distinguish from the spectral coefficients of u_{ds} (considering Figure 5.6), namely,

$$b_2 = 4000, 5000, 6000, \text{ for } \delta = 10^{-3}, 10^{-4}, 10^{-5}, \text{ respectively.} \tag{5.4.16}$$

Here, the perturbations on the spectral coefficients $|w_j|$ of u seem to be of different nature for $u = u_{\mathrm{ds}}$ and $u = u_{\mathtt{cg}}$: In the former case $|w_j|$ show an exponential decay in j, cf. Figure 5.5. (we did ascribe a decay on spectral coefficients to regularity properties of the underlying differential equation, which seem to carry over from the initial state when sub-steps of the numerical time propagation preserve regularity properties, i.e., when the matrix inverse is sufficiently accurate in the case of u_{ds}); In the later case multiple matrix applications of A are required when applying the conjugate gradient method, which result in a perturbation of the spectral coefficients $|w_j|$ for arbitrary eigenvalues λ_j, cf. Figure 5.6.

To specify ε in Assumption (A1) for the current setting, we introduce

$$\nu = \sum_{j=\ell}^{n} |w_j|, \quad \text{where } \ell \text{ conforms to } \lambda_{\ell-1} < b < \lambda_\ell. \tag{5.4.17}$$

Thus, for the spectral decomposition of u (with $u = u_{\mathrm{ds}}, u_{\mathtt{cg}}$ and different choices of n and δ) Assumption (A1) holds with $\varepsilon = \nu$ (for $b = b_1, b_2$ respectively). In all of the current test settings we observe $\nu \leq 1.2 \cdot \delta$.

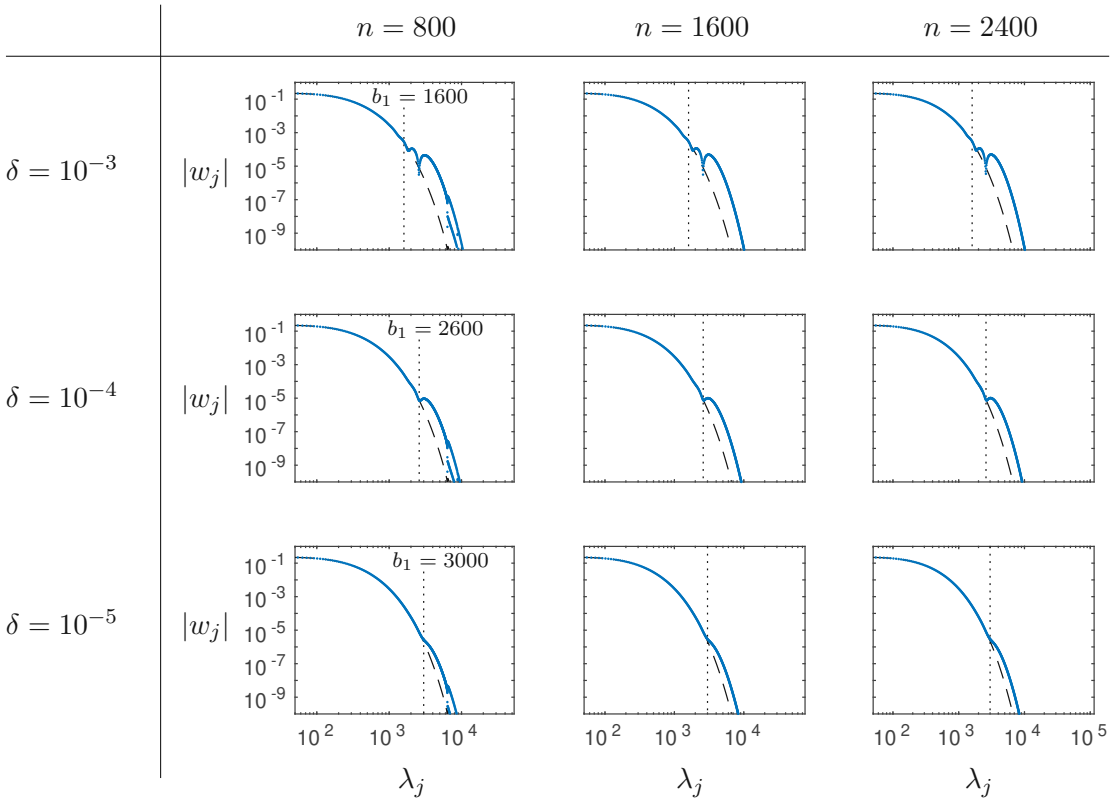


Figure 5.5: The different plots show the spectral coefficients of u_{ds} given in (5.4.11) for different choices of n and δ. Here δ refers to (5.4.14) and fixes the choice of the underlying time step τ as given in (5.4.13). The dot marks (which mainly conform to a curve) in each plot illustrate the spectral coefficients $|w_j|$ of $u = u_{\mathrm{ds}}$ versus the corresponding values of λ_j. The spectral coefficients $|w_j^0|$ of the vector ϕ_0 for the respective choices of n are located on the dashed line, and they are identical to the spectral coefficients of ϕ_τ, see also (5.4.8) and (5.4.9). The spectral coefficients of u_{ds} correspond to the spectral coefficients of ϕ_τ for $\lambda_j < b$, where $b = b_1$ is given in (5.4.15). For larger choices of λ_j the spectral coefficients $|w_j|$ significantly distinguish from the spectral coefficients of ϕ_τ.

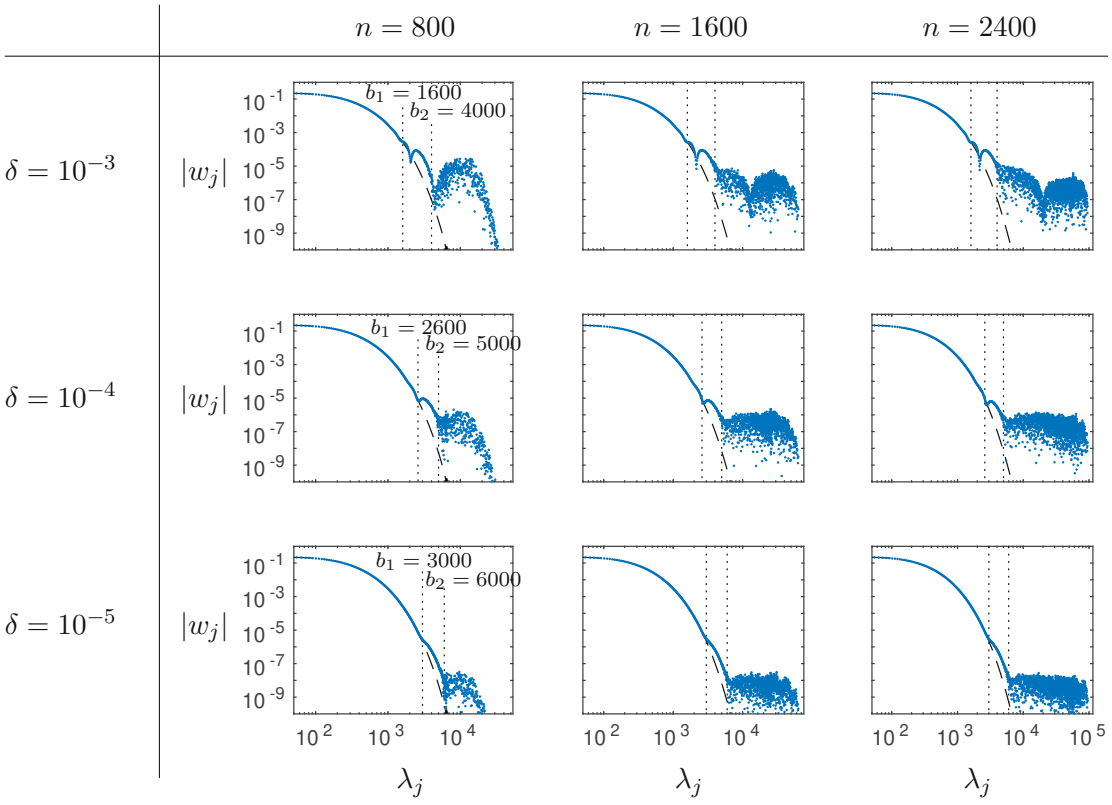


Figure 5.6: The different plots show the spectral coefficients of u_{cg} given in (5.4.12) for different choices of n and δ. Here δ refers to (5.4.14) and fixes the choice of the underlying time step τ (and δ_{cg}) as given in (5.4.13). The dot marks in each plot illustrate the spectral coefficients $|w_j|$ of $u = u_{\mathsf{cg}}$ versus the corresponding values of λ_j. As given in the caption of Figure 5.6 the spectral coefficients of ϕ_τ are located on the dashed line. The spectral coefficients of u correspond to the spectral coefficients of ϕ_τ for $\lambda_j < b_1$, where b_1 is given in (5.4.15). For larger choices of λ_j the spectral coefficients $|w_j|$ significantly distinguish from the spectral coefficients of ϕ_τ. For $b_1 < \lambda_j < b_2$ (b_2 is given in (5.4.16)) the spectral coefficients of u_{cg} and u_{ds} (shown in Figure 5.5) are similar, and for $\lambda_j > b_2$ the spectral coefficients of u_{cg} seem to be perturbed more randomly (similar to results of Figure 5.1 in Subsection 5.4.1).

Testing whether Assumption (A2) **holds with $\widehat{\varepsilon} \approx \varepsilon$ in this setting.** As stated previously Assumption (A1) holds with $\varepsilon = \nu$ for the present choices of u and b. In the following we consider the rational Krylov subspace $\mathcal{Q}_m(A, u)$ for a single pole $s = -10$ of multiplicity $m - 1$ and $u = u_{\mathsf{cg}}, u_{\mathrm{ds}}$, individually. For the spectral coefficients of u in the respective Krylov subspace Assumption (A2) holds true with $\widehat{\varepsilon} = \zeta_m$, where ζ_m is given in (5.4.6) for the respective choice of b. Thus, in the following numerical experiments $\zeta_m \approx \nu$ implies that Assumption (A2) follows from Assumption (A1) with $\widehat{\varepsilon} \approx \varepsilon$.

In Figure 5.7 a)–c) we illustrate ζ_m over m for $\mathcal{Q}_m(A, u)$ and the initial vector $u = u_{\mathsf{cg}}$ (with $b = b_1, b_2$ separately) and $u = u_{\mathrm{ds}}$ (with $b = b_1$), together with ν. In any of these figures we observe $\zeta_m \approx \nu$, which suggests that Assumption (A2) is reasonable to hold with $\widehat{\varepsilon} \approx \varepsilon$ when Assumption (A1) is holds true.

In Figure 5.7 b) we observe $\zeta_m \to \nu$ similar to results shown in Figure 5.2 and 5.4 in Subsection 5.4.1. In the setting of Figure 5.7 b) with $b = b_2$ and for the setting which is discussed in Subsection 5.4.1, a perturbation on $|w_j|$ is rather randomly distributed for $\lambda_j > b$. In contrast to Figure 5.7 b), the results of ζ_m in Figure 5.7 a) and c) with $b = b_1$ show steep jumps, which seem to be related to a decay in $|w_j|$ for $\lambda_j \approx b_1$. Furthermore, we recall that results of ξ_m are connected by a curve when the number of Ritz values which are larger than b matches for the respective values of m. Especially before the number of Ritz values which are larger than b increases (i.e., a jump in ζ_m in the figure) we observe $\zeta_m < \nu$ which can be beneficial for the error estimate in Proposition 5.3.2. Such effects can be relevant for future work but will not be further discussed here.

5.5 Summary

In the present chapter we have formulated a localized near-best approximation property for the rational Krylov approximation to the exponential of a skew-Hermitian matrix. When the action of the matrix exponential to the initial vector corresponds to the time evolution of a discretized differential equation and further assumptions are fulfilled which are related to regularity properties of the underlying problem, then the localized near-best approximation property can give insights on a grid-independent convergence rate of the rational Krylov approximation. We have discussed the necessary assumptions for this approach and numerically verified that these assumptions are realistic for a relevant numerical example, namely a Schrödinger-type equation.

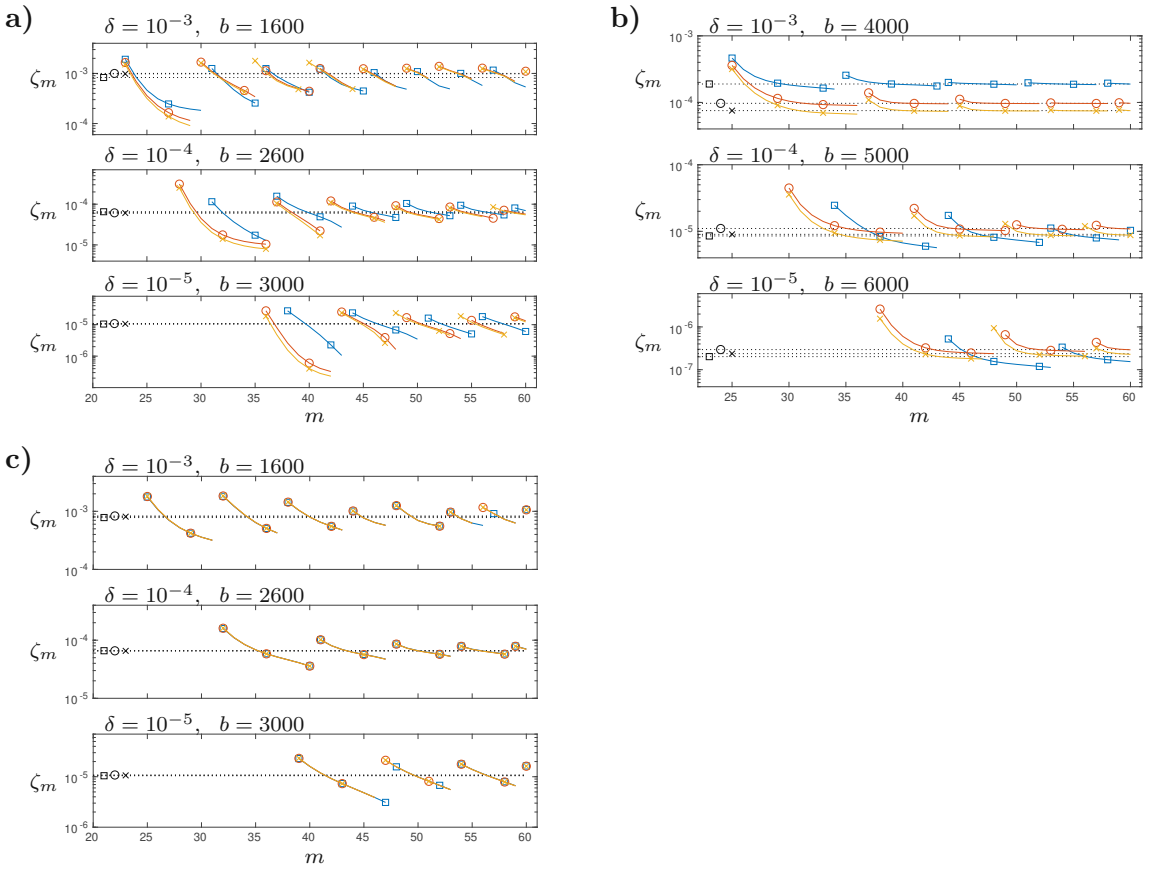


Figure 5.7: For **a)**–**c)** the plots show ζ_m (given in (5.4.6) with $b = b_1$ or $b = b_2$ depending on the sub-figure) for the initial vector u, where $u = u_{\mathsf{cg}}$ or $u = u_{\mathrm{ds}}$ depending on the sub-figure, and the problem size $n = 800$ (□), $n = 1600$ (○) and $n = 2400$ (×). Additionally, values of ζ_m which refer to the same choice of n and where A_m has the same number of eigenvalues larger than b are connected by a curve. The dotted lines show ν given in (5.4.17), where the symbol on the left-hand side of the dotted line refers to the underlying problem size n. The choice of δ refers to (5.4.14) and respective choices for the parameter τ and b, where the former is used to construct the initial vector $u = u_{\mathsf{cg}}, u_{\mathrm{ds}}$.
– Figure **a)**: The results refer to the spectrum of A_m for the rational Krylov subspace $\mathcal{Q}_m(A, u)$ with a single pole $s = -10$ of multiplicity $m-1$ and $u = u_{\mathsf{cg}}$ given in (5.4.11). For ζ_m and ν we choose $b = b_1$ given in (5.4.15).
– Figure **b)**: Similar to a) with $u = u_{\mathsf{cg}}$ given in (5.4.12), and $b = b_2$ given in (5.4.16).
– Figure **c)**: Similar to a) with $u = u_{\mathrm{ds}}$ given in (5.4.11), and $b = b_1$ given in (5.4.15).

Appendix

5.A Auxiliary material

Remark 5.A.1 (Bounded rational functions and stability properties of implicit methods)**.** *Implicit methods are well-studied in the field of numerical time integration. Stability properties of such methods give insights on possible restrictions on the choice of time steps and the underlying mesh size. The stability function of a method (see also [HW02, Definition 2.1]) corresponds to the respective numerical solution of* $x' = \lambda x$ *with* $x = x(t)$, $x(0) = 1$, *and* $\lambda \in \mathbb{C}$. *Thus, for an implicit method the stability function corresponds to a rational function which approximates the exponential function* $z \mapsto \mathrm{e}^z$ *for* $z = h\lambda \in \mathbb{C}$ *where* $h \in \mathbb{R}^+$ *denotes a time step. Let* $\widehat{r}_{m-1}$ *be a rational stability function which is not further specified here, but satisfies the assumption* $\widehat{r}_{m-1} = \widehat{p}_{m-1}/\widehat{q}_{m-1}$ *with* $\widehat{p}_{m-1}$ *and* $\widehat{q}_{m-1} \in \Pi_{m-1}$ *for a given* $m \in \mathbb{N}$,

$$\widehat{r}_{m-1}(z) \approx \mathrm{e}^z.$$

For the denominator $\widehat{q}_{m-1}$ *of* $\widehat{r}_{m-1}(z)$ *we write*

$$\widehat{q}_{m-1}(z) = \prod_{j=1}^{m-1} (z - \xi_j), \quad \text{where } \xi_1, \dots, \xi_{m-1} \in \mathbb{C} \text{ denote the poles of } \widehat{r}_{m-1}.$$

Then the stability domain related to $\widehat{r}_{m-1}$ *corresponds to*

$$S = \{z \in \mathbb{C} \,:\, |\widehat{r}_{m-1}(z)| \leq 1\}.$$

We are especially interested in bounded rational approximants to the exponential of a skew-Hermitian matrix $\mathrm{e}^{-\mathrm{i}tA}$. *To match the denotations, we write*

$$r(\lambda) = \widehat{r}_{m-1}(-\mathrm{i}\lambda), \qquad \text{thus,} \quad r(\lambda) \approx \mathrm{e}^{-\mathrm{i}t\lambda}, \qquad \lambda \in \mathbb{R}. \tag{5.A.1}$$

The respective denominator satisfies

$$\widehat{q}_{m-1}(-\mathrm{i}\lambda) = \prod_{j=1}^{m-1} (-\mathrm{i}\lambda - \xi_j) = (-\mathrm{i})^{m-1} \prod_{j=1}^{m-1} (\lambda - \mathrm{i}\xi_j).$$

Thus, we can represent the rational function r *in* (5.A.1) *as* $r = p/q_{m-1}$ *with* p *and* $q_{m-1} \in \Pi_{m-1}$, *whereat the denominator satisfies*

$$q_{m-1}(\lambda) = \prod_{j=1}^{m-1} (\lambda - s_j), \quad s_j = \mathrm{i}\xi_j. \tag{5.A.2}$$

Let $\xi_1, \dots, \xi_{m-1} \in \mathbb{C}$ *denote poles of a given stability function* $\widehat{r}_{m-1}$ *for which the respective stability domain covers the imaginary axis. Then the rational function* r *given in* (5.A.1) *is an element of the class of bounded rational functions* $R_{\varrho,b}$ *introduced in* (5.3.3) *with denominator* q_{m-1} *as given in* (5.A.2), $\varrho = 1$, *and any* $b \in \mathbb{R}$.

The stability domain of A-stable implicit methods (see also [HW02, Definition 3.3.]) covers the left half-plane, i.e., $S \supseteq \mathbb{C}_-$, *including the imaginary axis. Thus, stability functions of these methods yield examples for rational functions in the class* $R_{\varrho,b}$ *for the respective denominator.*

Proof of ***Proposition 5.3.1****.* By assumption we have

$$\lambda_1, \dots, \lambda_{\ell-1} \in [a,b], \quad \lambda_\ell, \dots, \lambda_n \in (b, +\infty), \quad \text{and} \quad \Big(\sum_{j=\ell}^{n} |w_j|^2 \Big)^{1/2} \leq \varepsilon \|u\|_{\mathsf{M}}. \tag{5.A.3}$$

For u given in (5.3.1) we define

$$u_1 = \sum_{j=1}^{\ell-1} w_j q_j, \quad \text{and} \quad u_2 = \sum_{j=\ell}^{n} w_j q_j.$$

With $u = u_1 + u_2$ the left-hand side of (5.3.4) satisfies

$$\|r(A)u - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} \leq \|r(A)u_1 - \mathrm{e}^{-\mathrm{i}tA}u_1\|_{\mathsf{M}} + \|r(A)u_2 - \mathrm{e}^{-\mathrm{i}tA}u_2\|_{\mathsf{M}}. \tag{5.A.4}$$

Similar to (5.3.1) we make use of the eigendecomposition of A to simplify the first term on the right-hand side of (5.A.4),

$$\|r(A)u_1 - \mathrm{e}^{-\mathrm{i}tA}u_1\|_{\mathsf{M}} = \Big(\sum_{j=1}^{\ell-1} |r(\lambda_j) - \mathrm{e}^{-\mathrm{i}t\lambda_j}|^2 |w_j|^2 \Big)^{1/2} \leq \max_{j=1,\dots,\ell-1} |r(\lambda_j) - \mathrm{e}^{-\mathrm{i}t\lambda_j}| \|u_1\|_{\mathsf{M}}.$$

Thus, with $\|u_1\|_{\mathsf{M}} \leq \|u\|_{\mathsf{M}}$ and $\lambda_j \in [a,b]$ for $j = 1, \dots, \ell-1$, we have

$$\|r(A)u_1 - \mathrm{e}^{-\mathrm{i}tA}u_1\|_{\mathsf{M}} \leq \max_{\lambda \in [a,b]} |r(\lambda) - \mathrm{e}^{-\mathrm{i}t\lambda}| \|u\|_{\mathsf{M}}. \tag{5.A.5}$$

Analogously, the second term on the right-hand side of (5.A.4) satisfies

$$\|r(A)u_2 - \mathrm{e}^{-\mathrm{i}tA}u_2\|_{\mathsf{M}} \leq \max_{j=\ell,\dots,n} |r(\lambda_j) - \mathrm{e}^{-\mathrm{i}t\lambda_j}| \|u_2\|_{\mathsf{M}}. \tag{5.A.6}$$

With $\lambda_\ell, \dots, \lambda_n \in (b, +\infty)$ as given in (5.A.3) and for $r \in R_{\varrho,b}$ given in (5.3.3) we conclude $|r(\lambda_j)| \leq \varrho$ for $j = \ell, \dots, n$, and together with $|\mathrm{e}^{-\mathrm{i}t\lambda_j}| = 1$ this implies

$$\max_{j=\ell,\dots,n} |r(\lambda_j) - \mathrm{e}^{-\mathrm{i}t\lambda_j}| \leq 1 + \varrho. \tag{5.A.7}$$

The upper bound in (5.A.3) implies $\|u_2\|_{\mathsf{M}} \leq \varepsilon \|u\|_{\mathsf{M}}$, and together with (5.A.6) and (5.A.7) this yields

$$\|r(A)u_2 - \mathrm{e}^{-\mathrm{i}tA}u_2\|_{\mathsf{M}} \leq (1+\varrho)\varepsilon \|u\|_{\mathsf{M}}. \tag{5.A.8}$$

Combining (5.A.4) with (5.A.5) and (5.A.8) completes the proof. □

Proof of ***Proposition 5.3.2****.* For an arbitrary $r \in R_{\varrho,b}$ we recall (5.3.6),

$$\|U_m \mathrm{e}^{-\mathrm{i}tA_m}x - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} \leq \|r(A)u - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} + \|r(A_m)x - \mathrm{e}^{-\mathrm{i}tA_m}x\|_2. \tag{5.A.9}$$

With Assumption (A1) in Proposition 5.3.2 the result of Proposition 5.3.1 applies to the first term on the right-hand side of (5.A.9),

$$\|r(A)u - \mathrm{e}^{-\mathrm{i}tA}u\|_{\mathsf{M}} \leq \max_{\lambda \in [a,b]} |r(\lambda) - \mathrm{e}^{-\mathrm{i}t\lambda}| \|u\|_{\mathsf{M}} + (1+\varrho)\varepsilon \|u\|_{\mathsf{M}}. \tag{5.A.10}$$

Assumption (A2) in Proposition 5.3.2 (with $\|x\|_2 = \|u\|_{\mathsf{M}}$) conforms to the assumptions of Proposition 5.3.1 for the second term on the right-hand side of (5.A.9) w.r.t. the Euclidean inner product. Thus, Proposition 5.3.1 yields

$$\|r(A_m)x - \mathrm{e}^{-\mathrm{i}tA_m}x\|_2 \leq \max_{\lambda\in[a,b]} |r(\lambda) - \mathrm{e}^{-\mathrm{i}t\lambda}|\|x\|_2 + (1+\varrho)\varepsilon\|x\|_2. \tag{5.A.11}$$

Combining (5.A.9), (5.A.10) and (5.A.11) (with $\|x\|_2 = \|u\|_{\mathsf{M}}$), and choosing the minimum over $r \in R_{\varrho,b}$, we complete the proof of Proposition 5.3.2. □

Bibliography

[AEEG08] M. Afanasjew, M. Eiermann, O. Ernst, and S. Güttel. Implementation of a restarted Krylov subspace method for the evaluation of matrix functions. *Linear Algebra Appl.*, 429(10):2293–2314, 2008.

[Akh65] N.I. Akhiezer. *The Classical Moment Problem and Some Related Questions in Analysis*. Oliver & Boyd, Edinburgh, first english edition, 1965.

[AKT14] W. Auzinger, O. Koch, and M. Thalhammer. Defect-based local error estimators for splitting methods, with application to Schrödinger equations, part II. higher-order methods for linear problems. *J. Comput. Appl. Math.*, 255:384–403, 2014.

[AMH11] A. Al-Mohy and N. Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM J. Sci. Comput.*, 33(2):488–511, 2011.

[Ant05] A.C. Antoulas. *Approximation of large-scale dynamical systems*, volume 6 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[BCOR09] S. Blanes, F. Casas, J. Oteo, and J. Ros. The Magnus expansion and some of its applications. *Phys. Rep.*, 470(5):151–238, 2009.

[BG15] M. Berljafa and S. Güttel. Generalized rational Krylov decompositions with an application to rational approximation. *SIAM J. Matrix Anal. Appl.*, 36(2):894–916, 2015.

[BGF97] A. Bunse-Gerstner and H. Faßbender. Error bounds in the isometric Arnoldi process. *J. Comput. Appl. Math.*, 86(1):53–72, 1997.

[BGH13] M. Botchev, V. Grimm, and M. Hochbruck. Residual, restarting and Richardson iteration for the matrix exponential. *SIAM J. Sci. Comput.*, 35(3):A1376–A1397, 2013.

[BK19] M. Botchev and L. Knizhnerman. ART: Adaptive residual-time restarting for Krylov subspace matrix exponential evaluations. *J. Comput. Appl. Math.*, 2019.

[BLR00] T. Braconnier, P. Langlois, and J. Rioual. The influence of orthogonality on the Arnoldi method. *Linear Algebra Appl.*, 309(1):307–323, 2000.

[BM00] T. Barth and T. Manteuffel. Multiple recursion conjugate gradient algorithms part I: Sufficient conditions. *SIAM J. Matrix Anal. Appl.*, 21(3):768–796, 2000.

[BM05] S. Blanes and P. Moan. Fourth- and sixth-order commutator-free Magnus integrators for linear and non-linear dynamical systems. *Appl. Numer. Math.*, 56:1519–1537, 2005.

[BMV18] B. Beckermann, C. Mertens, and R. Vandebril. On an economic arnoldi method for *bml*-matrices. *SIAM J. Matrix Anal. Appl.*, 39(2):737–768, 2018.

[Bot16] M. Botchev. Krylov subspace exponential time domain solution of Maxwell's equations in photonic crystal modeling. *J. Comput. Appl. Math.*, 293:20–34, 2016.

[BR09] B. Beckermann and L. Reichel. Error estimates and evaluation of matrix functions via the Faber transform. *SIAM J. Numer. Anal.*, 47(5):3849–3883, 2009.

[Chi78] T.S. Chihara. *An Introduction to Orthogonal Polynomials*. Dover, 1978.

[CKOR16] M. Caliari, P. Kandolf, A. Ostermann, and S. Rainer. The Leja method revisited: Backward error analysis for the matrix exponential. *SIAM J. Sci. Comput.*, 38(3):A1639–A1661, 2016.

[CM97] E. Celledoni and I. Moret. A Krylov projection method for systems of ODEs. *Appl. Numer. Math.*, 24(2):365–378, 1997.

[dB05] C. de Boor. Divided differences. *Surv. Approx. Theory*, 1:46–69, 2005.

[DB07] K. Deckers and A. Bultheel. Rational Krylov sequences and orthogonal rational functions. Technical report, Department of Computer Science, K.U.Leuvene, Katholieke Universiteit Leuven, Belgium, 01 2007.

[DB12] K. Deckers and A. Bultheel. The existence and construction of rational Gauss-type quadrature rules. *Appl. Math. Comput.*, 218(20):10299–10320, 2012.

[DBVD10] K. Deckers, A. Bultheel, and J. Van Deun. A generalized eigenvalue problem for quasi-orthogonal rational functions. Technical report, Department of Computer Science, K.U.Leuvene, Katholieke Universiteit Leuven, Belgium, August 2010. Report TW 571.

[Dec09] K. Deckers. *orthogonal rational functions: quadrature, recurrence and rational Krylov*. PhD thesis, Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium, February 2009.

[DGK98] V. Druskin, A. Greenbaum, and L. Knizhnerman. Using nonorthogonal Lanczos vectors in the computation of matrix functions. *SIAM J. Sci. Comput.*, 19(1):38–54, 1998.

[DK89] V. Druskin and L. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *USSR Comput. Math. Math. Phys.*, 29(6):112–121, 1989.

[DK92] V. Druskin and L. Knizhnerman. Error bounds in the simple Lanczos procedure for computing functions of symmetric matrices and eigenvalues. *Comput. Math. Math. Phys.*, 31(7):20–30, 1992.

[DK95] V. Druskin and L. Knizhnerman. Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic. *Numer. Linear Algebra Appl.*, 2(3):205–217, 1995.

[DK98] V. Druskin and L. Knizhnerman. Extended Krylov subspaces: Approximation of the matrix square root and related functions. *SIAM J. Matrix Anal. Appl.*, 19(3):755–771, 1998.

[DKZ09] V. Druskin, L. Knizhnerman, and M. Zaslavsky. Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts. *SIAM J. Sci. Comput.*, 31(5):3760–3780, 2009.

[DMR09] F. Diele, I. Moret, and S. Ragni. Error estimates for polynomial Krylov approximations to matrix functions. *SIAM J. Matrix Anal. Appl.*, 30(4):1546–1565, 2009.

[EE06] M. Eiermann and O. Ernst. A restarted Krylov subspace method for the evaluation of matrix functions. *SIAM J. Numer. Anal.*, 44:2481–2504, 2006.

[EEG11] M. Eiermann, O. Ernst, and S. Güttel. Deflated restarting for matrix functions. *SIAM J. Matrix Anal. Appl.*, 32(2):621–641, 2011.

[ER80] T. Ericsson and A. Ruhe. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Math. Comp.*, 35(152):1251–1268, 1980.

[Eri90] T. Ericsson. Computing functions of matrices using Krylov subspace methods. Technical report, Chalmers University of Technology, Department of Computer Science, Göteborg, Sweden, 1990.

[FF94] B. Fischer and R.W. Freund. On adaptive weighted polynomial preconditioning for Hermitian positive definite matrices. *SIAM J. Sci. Comput.*, 15(2):408–426, 1994.

[FGS14] A. Frommer, S. Güttel, and M. Schweitzer. Efficient and stable Arnoldi restarts for matrix functions based on quadrature. *SIAM J. Matrix Anal. Appl.*, 35(2):661–683, 2014.

[FH93] R.W. Freund and M. Hochbruck. Gauss quadratures associated with the Arnoldi process and the Lanczos algorithm. In *Linear Algebra for Large Scale and Real-Time Applications*, pages 377–380. Springer-Verlag, 1993.

[Fis96] B. Fischer. *Polynomial Based Iteration Methods for Symmetric Linear Systems*. Vieweg+Teubner Verlag, Wiesbaden, 1996.

[FTDR89] R.A. Friesner, L.S. Tuckerman, B.C. Dornblaser, and T.V. Russo. A method for exponential propagation of large systems of stiff nonlinear differential equations. *J. Sci. Comput.*, 4(4):327–354, 1989.

[Gau81] W. Gautschi. A survey of Gauss-Christoffel quadrature formulae. In P.L. Butzer and F. Feher, editors, *E.B. Christoffel: The Influence of His Work on Mathematics and Physics*, pages 72–147. Birkhäuser, Basel, Switzerland, 1981.

[Gau93] W. Gautschi. Gauss–type quadrature rules for rational functions. In *Numerical integration, IV (Oberwolfach, 1992)*, volume 112 of *Internat. Ser. Numer. Math.*, pages 111–130. Birkhäuser, Basel, Switzerland, 1993.

[Gau04] W. Gautschi. *Orthogonal polynomials: computation and approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004. Oxford Science Publications.

[GG10] V. Grimm and M. Gugat. Approximation of semigroups and related operator functions by resolvent series. *SIAM J. Numer. Anal.*, 48(5):1826–1845, 2010.

[GG13] T. Göckler and V. Grimm. Convergence analysis of an extended Krylov subspace method for the approximation of operator functions in exponential integrators. *SIAM J. Numer. Anal.*, 51(4):2189–2213, 2013.

[GG17] V. Grimm and T. Göckler. Automatic smoothness detection of the resolvent Krylov subspace method for the approximation of C_0-semigroups. *SIAM J. Numer. Anal.*, 55, 2017.

[GH08] V. Grimm and M. Hochbruck. Rational approximation to trigonometric operators. *BIT*, 48(2):215–229, 2008.

[GM10] G.H. Golub and G. Meurant. *Matrices, Moments and Quadrature with Applications*. Princeton series in applied mathematics. Princeton University Press, Princeton, NJ, USA, 2010.

[Gol73] G.H. Golub. Some modified matrix eigenvalue problems. *SIAM Rev.*, 15(2):318–334, 1973.

[Gol02] L. Golinskii. Quadrature formula and zeros of para-orthogonal polynomials on the unit circle. *Acta Math. Hungar.*, 96(3):169–186, 2002.

[Gra93] W.B. Gragg. Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle. *J. Comput. Appl. Math.*, 46(1):183–198, 1993.

[Gre89] A. Greenbaum. Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl.*, 113:7–63, 1989.

[Gri12] V. Grimm. Resolvent Krylov subspace approximation to operator functions. *BIT*, 52(3):639–659, 2012.

[GS92] E. Gallopoulos and Y. Saad. Efficient solution of parabolic equations by Krylov approximation methods. *SIAM J. Sci. Statist. Comput.*, 13(5):1236–1264, 1992.

[Gus91] K. Gustafsson. Control theoretic techniques for stepsize selection in explicit Runge-Kutta methods. *ACM Trans. Math. Software*, 17:533–554, 1991.

[Güt10] S. Güttel. *Rational Krylov Methods for Operator Functions*. PhD thesis, Technische Universität Bergakademie Freiberg, Germany, 2010. Thesis available as MIMS Eprint 2017.39.

[Güt13] S. Güttel. Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection. *GAMM-Mitt.*, 36(1):8–31, 2013.

[GVL89] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, second edition, 1989.

[GW69] G.H. Golub and J.H. Welsch. Calculation of Gauss quadrature rules. *Math. Comp.*, 23(106):221–221, 1969.

[Hel07] K. Held. Electronic structure calculations using dynamical mean field theory. *Adv. in Physics*, 56:829–926, 2007.

[HH05] M. Hochbruck and M. Hochstenbach. Subspace extraction for matrix functions. Technical report, Dept. of Math., Case Western Reserve University, 2005.

[Hig02] N. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

[Hig08] N. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[HJ85] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

[HL97] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 34(5):1911–1925, 1997.

[HLS98] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19(5):1552–1574, 1998.

[HO10] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.

[HPS09] I. Hnětynková, M. Plešinger, and Z. Strakoš. The regularizing effect of the Golub-Kahan iterative bidiagonalization and revealing the noise level in the data. *BIT*, 49(4):669–696, 2009.

[HPS+15] M. Hochbruck, T. Pažur, A. Schulz, E. Thawinan, and C. Wieners. Efficient time integration for discontinuous Galerkin approximations of linear wave equations. *Z. Angew. Math. Mech.*, 95(3):237–259, 2015.

[Hub63] J. Hubbard. Electron correlations in narrow energy bands. *Proc. Roy. Soc. London Ser. A*, 276(1365):238–257, 1963.

[Hür15] W. Hürlimann. An explicit version of the Chebyshev-Markov-Stieltjes inequalities and its applications. *J. Inequal. Appl.*, 2015(192), 2015.

[HW02] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Springer-Verlag, Berlin, 2002.

[IKS19] A. Iserles, K. Kropielnicka, and P. Singh. Compact schemes for laser-matter interaction in Schrödinger equation based on effective splittings of Magnus expansion. *Comput. Phys. Commun.*, 234:195–201, 2019.

[Jaf08] S. Jafari. Introduction to Hubbard model and exact diagonalization. *Iran. J. Phys. Res.*, 8, 2008.

[JAK20] T. Jawecki, W. Auzinger, and O. Koch. Computable upper error bounds for Krylov approximations to matrix exponentials and associated φ-functions. *BIT*, 60(1):157–197, 2020.

[Jaw22a] T. Jawecki. A review of the separation theorem of Chebyshev-Markov-Stieltjes for polynomial and some rational Krylov subspaces. preprint at https://arxiv.org/pdf/2205.01535.pdf, 2022.

[Jaw22b] T. Jawecki. A study of defect-based error estimates for the Krylov approximation of φ-functions. *Numer. Algorithms*, 90(1):323–361, 2022.

[JL15] Zh. Jia and H. Lv. A posteriori error estimates of Krylov subspace approximations to matrix functions. *Numer. Algorithms*, 69(1):1–28, 2015.

[JR94] C. Jagels and L. Reichel. A fast minimal residual algorithm for shifted unitary matrices. *Numer. Linear Algebra Appl.*, 1(6):555–570, 1994.

[JR11] C. Jagels and L. Reichel. Recursion relations for the extended Krylov subspace method. *Linear Algebra Appl.*, 434(7):1716–1732, 2011. Special Issue: NIU.

[JR13] C. Jagels and L. Reichel. The structure of matrices in rational Gauss quadrature. *Math. Comp.*, 82(284):2035–2060, 2013.

[KBC05] A.I. Kuleff, J. Breidbach, and L.S. Cederbaum. Multielectron wave-packet propagation: General theory and application. *J. Chem. Phys.*, 123(4):044111, 2005.

[KS53] S. Karlin and L.S. Shapley. Geometry of moment spaces. *Mem. Amer. Math. Soc.*, 12:93, 1953.

[KS10] L. Knizhnerman and V. Simoncini. A new investigation of the extended Krylov subspace method for matrix function evaluations. *Numer. Linear Algebra Appl.*, 17:615–638, 2010.

[Lan50] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255, 1950.

[Li98] X. Li. Separation theorem of Chebyshev–Markov–Stieltjes type for Laurent polynomials orthogonal on $(0, \infty)$. In A. Sri Ranga W.B. Jones, editor, *Orthogonal Functions: Moment Theory and Continued Fractions*, volume 199, pages 327–341. CRC Press, 1998.

[LLRW08] G. López Lagomasino, L. Reichel, and L. Wunderlich. Matrices, moments, and rational quadrature. *Linear Algebra Appl.*, 429(10):2540–2554, 2008.

[LS13] J. Liesen and Z. Strakoš. *Krylov subspace methods: Principles and Analysis*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2013.

[Lub08] C. Lubich. *From Quantum to Classical Molecular Dynamics; Reduced Models and Numerical Analysis*. Zurich lectures in advanced mathematics. Europ. Math. Soc., Zürich, 2008.

[MA06] N. Mohankumar and S.M. Auerbach. On time-step bounds in unitary quantum evolution using the Lanczos method. *Comput. Phys. Commun.*, 175:473–481, 2006.

[Mah93] G. Mahan. *Many-particle physics*. Physics of solids and liquids. Plenum Press, New York, second edition, 1993.

[MC10] N. Mohankumar and T. Carrington. A new approach for determining the time step when propagating with the Lanczos algorithm. *Comput. Phys. Commun.*, 181:1859–1861, 2010.

[MMS18] C. Musco, Ch. Musco, and A. Sidford. Stability of the Lanczos method for matrix function approximation. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 1605–1624, Philadelphia, PA, USA, 2018. Society for Industrial and Applied Mathematics.

[MN01] I. Moret and P. Novati. An interpolatory approximation of the matrix exponential based on Faber polynomials. *J. Comput. Appl. Math.*, 131(1):361–380, 2001.

[MN04] I. Moret and P. Novati. RD-rational approximations of the matrix exponential. *BIT*, 44:595–615, 2004.

[MNP84] A. McCurdy, K.C. Ng, and B.N. Parlett. Accurate computation of divided differences of the exponential function. *Math. Comp.*, 43:501–528, 1984.

[MVL03] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3–49, 2003.

[Nie07] J. Niehoff. *Projektionsverfahren zur Approximation von Matrixfunktionen mit Anwendungen auf die Implementierung exponentieller Integratoren*. PhD thesis, Heinrich-Heine-Universität Düsseldorf, 2007.

[NW83] A. Nauts and R.E. Wyatt. New approach to many-state quantum dynamics: The recursive-residue-generation method. *Phys. Rev. Lett.*, 51:2238–2241, 1983.

[NW12] J. Niesen and W.M. Wright. Algorithm 919: A Krylov subspace algorithm for evaluating the ϕ-functions appearing in exponential integrators. *ACM Trans. Math. Software*, 38(3):1–19, 2012.

[Opi64] G. Opitz. Steigungsmatrizen. *Z. Angew. Math. Mech.*, 44(S1):T52–T54, 1964.

[Pai76] C. Paige. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *IMA J. Appl. Math.*, 18(3):341–349, 1976.

[Pai80] C. Paige. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Algebra Appl.*, 34:235–258, 1980.

[Par98] B. Parlett. *The Symmetric Eigenvalue Problem*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1998.

[PKvdBS16] E. Pavarini, E. Koch, J. van den Brink, and G. Sawatzky. *Quantum Materials: Experiments and Theory*, volume 6 of *Modeling and Simulation*. Forschungszentrum Jülich, Jülich, Germany, September 2016.

[PL86] T.J. Park and J.C. Light. Unitary quantum time evolution by iterative Lanczos reduction. *J. Chem. Phys.*, 85:5870–5876, 1986.

[Rei79] W.P. Reinhardt. l^2 discretization of atomic and molecular electronic continua: Moment, quadrature and J-matrix techniques. *Comput. Phys. Commun.*, 17(1):1–21, 1979.

[Ruh84] A. Ruhe. Rational Krylov sequence methods for eigenvalue computation. *Linear Algebra Appl.*, 58:391–405, 1984.

[Saa92] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 29(1):209–228, 1992.

[Saa03] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2003.

[Saa11] Y. Saad. *Numerical methods for large eigenvalue problems*. Classics in applied mathematics ; 66. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2011.

[Sch08] K. Schäfer. *Krylov subspace methods for shifted unitary matrices and eigenvalue deflation applied to the Neuberger Operator and the matrix sign function*. PhD thesis, Bergische Universität Wuppertal, Germany, 2008.

[Sch15] M. Schweitzer. *Restarting and error estimation in polynomial and extended Krylov subspace methods for the approximation of matrix functions*. PhD thesis, Bergische Universität Wuppertal, Germany, 2015.

[Sid98] R. Sidje. Expokit: A software package for computing matrix exponentials. *ACM Trans. Math. Software*, 24(1):130–156, 1998.

[Sim84] H.D. Simon. Analysis of the symmetric Lanczos algorithm with reorthogonalization methods. *Linear Algebra Appl.*, 61:101–131, 1984.

[Sin19] P. Singh. Sixth-order schemes for laser-matter interaction in the Schrödinger equation. *J. Chem. Phys.*, 150(15):154111, 2019.

[SL96] D.E. Stewart and T.S. Leyk. Error estimates for Krylov subspace approximations of matrix exponentials. *J. Comput. Appl. Math.*, 72(2):359–369, 1996.

[Sze85] G. Szegö. *Orthogonal Polynomials*. American Mathematical Society, Providence, RI, USA, reprint., fourth edition, 1985.

[TE07] H. Tal-Ezer. On restart and error estimation for Krylov approximation of $W = F(A)V$. *SIAM J. Sci. Comput.*, 29(6):2426–2441, 2007.

[TEK84] H. Tal-Ezer and R. Kosloff. An accurate and efficient scheme for propagating the time dependent Schrödinger equation. *J. Chem. Phys.*, 81(9):3967–3971, 1984.

[VA93] W. Van Assche. The impact of Stieltjes' work on continued fractions and orthogonal polynomials. In G. van Dijk, editor, *Stieltjes Oeuvres Complètes - Collected Papers*, pages 5–37. Springer-Verlag, 1993.

[vdEH06] J. van den Eshof and M. Hochbruck. Preconditioning Lanczos approximations to the matrix exponential. *SIAM J. Sci. Comput.*, 27(4):1438–1457, 2006.

[VL77] C. Van Loan. The sensitivity of the matrix exponential. *SIAM J. Numer. Anal.*, 14(6):971–981, 1977.

[Wil62] H.S. Wilf. *Mathematics for the Physical Sciences*. Dover, 1962.

[WY17] H. Wang and Q. Ye. Error bounds for the Krylov subspace methods for computations of matrix exponentials. *SIAM J. Matrix Anal. Appl.*, 38(1):155–187, 2017.

[WZX16] G. Wu, L. Zhang, and T. Xu. A framework of the harmonic Arnoldi method for evaluating ϕ-functions with applications to exponential integrators. *Adv. Comput. Math.*, 42(3):505–541, 2016.

[Zel54] M. Zelen. Bounds on a distribution function that are functions of moments to order four. *J. Res. Nat. Bur. Standards*, 53(6):377–381, 1954.

[Zem03] J. Zemke. *Krylov Subspace Methods in Finite Precision : A Unified Approach.* PhD thesis, Technische Universität Hamburg, 2003.

[ZTK19] M. Zemaityte, F. Tisseur, and R. Kannan. Filtering frequencies in a shift-and-invert Lanczos algorithm for the dynamic analysis of structures. *SIAM J. Sci. Comput.*, 41(3):B601–B624, 2019.

Wissenschaftliche Publikationen

W. Auzinger, J. Dubois, K. Held, H. Hofstätter, T. Jawecki, A. Kauch, O. Koch, K. Kropielnicka, P. Singh, and C. Watzenböck. Efficient Magnus-type integrators for solar energy conversion in Hubbard models. *J. Comput. Math. Data Sci.*, 2:100018, 2022.

T. Jawecki. A study of defect-based error estimates for the Krylov approximation of φ-functions. *Numer. Algorithms*, 90(1):323–361, 2022.

W. Auzinger, T. Jawecki, O. Koch, P. Pukach, R. Stolyarchuk, and E.B. Weinmüller. Some aspects on [numerical] stability of evolution equations of stiff type; use of computer algebra. In *2021 IEEE XVII th International Conference on the Perspective Technologies and Methods in MEMS Design (MEMSTECH)*, pages 180–184, 2021.

C. Schattauer, L. Linhart, T. Fabian, T. Jawecki, W. Auzinger, and F. Libisch. Graphene quantum dot states near defects. *Phys. Rev. B*, 102:155430, 2020.

T. Jawecki, W. Auzinger, and O. Koch. Computable upper error bounds for Krylov approximations to matrix exponentials and associated φ-functions. *BIT*, 60(1):157–197, 2020.

Preprints

T. Jawecki and P. Singh. Unitarity of some barycentric rational approximants. preprint at https://arxiv.org/abs/2205.10606, 2022.

T. Jawecki. A review of the separation theorem of Chebyshev-Markov-Stieltjes for polynomial and some rational Krylov subspaces. preprint at https://arxiv.org/pdf/2205.01535.pdf, 2022.

Diploma Thesis

T. Jawecki. Bifurcation analysis via numerical continuation for nonlinear fourth-order partial differential equations. Master's thesis, TU Wien, Austria, 2017. available online at http://katalog.ub.tuwien.ac.at/AC13642458.

Lebenslauf

Persönliche Daten

Name	**Tobias Jawecki**
Geburtsdatum	21.06.1991
Geburtsort	Wien
Nationalität	Österreich
Email	tobias.jawecki@gmail.com

Ausbildung & wissenschaftliche Anstellungen

09/2021–08/2022	Universitätsassistent am Institut für Theoretische Physik, TU Wien, Österreich
03/2017–03/2020	Universitätsassistent am Institut für Analysis und Scientific Computing, TU Wien
seit 03/2017	Doktoratsstudium der Technischen Wissenschaften Technische Mathematik, TU Wien, Teilnahme am Doktoratskolleg TU-D
03/2015–03/2017	Diplomstudium Technische Mathematik, TU Wien
03/2011–03/2015	Bachelorstudium Technische Mathematik, TU Wien
06/2009	Matura

Wien, am 07.11.2022

Tobias Jawecki