Rational Krylov subspace methods for φ-functions in exponential integrators

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Tanja Göckler

aus Lörrach

Tag der mündlichen Prüfung:	23. Juli 2014
Referent:	PD Dr. Volker Grimm
1. Korreferentin:	Prof. Dr. Marlis Hochbruck
2. Korreferent:	Prof. Dr. habil. Bernhard Beckermann

Acknowledgements

At this point, I wish to convey my genuine gratitude to a number of people whose support made this thesis possible.

First of all, I would like to express my special appreciation and thanks to my advisor PD Dr. Volker Grimm for the opportunity to work on this interesting topic and for his excellent supervision, constant support and motivation as well as for countless hours of helpful inspiring discussions during the research and writing process of this thesis. I am also very grateful to Prof. Dr. Marlis Hochbruck for being my second advisor and for giving me numerous valuable and constructive suggestions. Prof. Dr. habil. Bernhard Beckermann from University Lille in France deserves my honest thanks for serving as an external referee and for his very useful proposals for improvement.

I thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this work via GR 3787 / 1-1.

Moreover, I would like to thank my colleagues for the pleasant and fantastic working environment. A very special thank goes to Lena Martin for careful proofreading and many helpful suggestions for improvement of my dissertation. In addition, I am obliged to PD Dr. Markus Neher who convinced me to start a doctorate after my first state examination and proof-read parts of this thesis. I also acknowledge my friends for their help and encouragement during the period of my PhD studies.

Finally, I extend my sincere gratitude to my family for their invaluable and kind support.

Contents

Notation

We always denote matrices with bold capital letters $\boldsymbol{A}, \boldsymbol{B}, \ldots$ and vectors with bold small letters $\boldsymbol{v}, \boldsymbol{w}, \ldots$. Linear operators are designated by capital letters $A, B, \ldots$ and functions by small letters $f, g, v, \ldots$.

$\boldsymbol{I}$	identity matrix
$\boldsymbol{O}$	zero matrix
$\boldsymbol{0}$	zero vector
$\mathcal{P}_m$	space of polynomials with degree less than or equal to m
$\frac{\mathcal{P}_{m-1}}{q_{m-1}}$	$\left\{\frac{p_{m-1}(z)}{q_{m-1}(z)} : p_{m-1} \in \mathcal{P}_{m-1}\right\}$ for a fixed $q_{m-1} \in \mathcal{P}_{m-1}$
$\mathrm{diag}(\cdots)$	diagonal matrix
$\mathrm{tridiag}(\cdots)$	tridiagonal matrix
$p_{\boldsymbol{A}}^{\min}$	minimal polynomial of the matrix $\boldsymbol{A}$
$p_{\boldsymbol{A},\boldsymbol{v}}^{\min}$	minimal polynomial of $\boldsymbol{A}$ with respect to the vector $\boldsymbol{v}$
$\deg(p)$	degree of the polynomial p
$\mathrm{int}(\Gamma)$	interior of the closed contour Γ
$(\cdot,\cdot)$	inner product
$\lVert\cdot\rVert$	operator, matrix, or vector norm
$\lVert\cdot\rVert_2$	Euclidean vector norm
$\lVert\cdot\rVert_{\boldsymbol{M}}$	$\lVert\boldsymbol{M}^{1/2}\cdot\rVert_2$, where $\boldsymbol{M}$ is some positive definite Hermitian matrix
$\sigma(\boldsymbol{A})$	set of eigenvalues of $\boldsymbol{A}$
$W(\boldsymbol{A})$	field of values of $\boldsymbol{A}$ given by $\{(\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}) : \lVert\boldsymbol{x}\rVert = 1\}$
$\mathbb{C}_0^-$	closed left complex half-plane $\{z \in \mathbb{C} : \mathrm{Re}(z) \le 0\}$
$\mathbb{R}_0^+$	real numbers greater than or equal to zero
$\otimes$	Kronecker product
Δ, ∇	Laplace and Nabla operator
$\nabla_{\boldsymbol{n}}$	normal derivative
u'	derivative of u with respect to the time t
$\partial\Omega$	boundary of the domain Ω
$\mathcal{O}(\cdot)$	Landau notation for asymptotic behavior ($f(z) = \mathcal{O}\big(g(z)\big)$ as $z \to \xi$ means that there exist constants $C, \epsilon > 0$ with $\lvert f(z)\rvert \le C\lvert g(z)\rvert$ for $\lvert z-\xi\rvert < \epsilon$)
$C^n(\Omega)$	space of n-times continuously differentiable functions on Ω
$L^1(\Omega)$	space of Lebesgue integrable functions on Ω

$L^2(\Omega)$	space of quadratically Lebesgue integrable functions on Ω
$C_c^\infty(\Omega)$	space of infinitely differentiable functions with compact support on Ω
$H^k(\Omega)$	Sobolev space of k-times weakly differentiable L^2-functions on Ω
$H_0^1(\Omega)$	closure of $C_c^\infty(\Omega)$ in $H^1(\Omega)$
I	identity operator
$T(t)$	strongly continuous semigroup
$D(A)$	domain of the operator A
$\rho(A)$, $\sigma(A)$	resolvent set and spectrum of A
$\mathrm{Range}(\boldsymbol{A})$	image of $\boldsymbol{A}$
$\mathrm{Null}(\boldsymbol{A})$	null space of $\boldsymbol{A}$
$\boldsymbol{A}^H$	complex conjugate transpose of the matrix $\boldsymbol{A}$
$\boldsymbol{V}_m^+$	Moore-Penrose inverse of $\boldsymbol{V}_m$
$\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$	polynomial Krylov subspace
$\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$	rational Krylov subspace
$\mathcal{K}_{q+1,m}^\gamma(\boldsymbol{A}, \boldsymbol{v})$	extended Krylov subspace
$\mathcal{R}_m(\boldsymbol{A})$	rational matrix subspace
$\mathrm{span}\{\cdots\}$	set of all linear combinations of the vectors in brackets
$\dim(\cdot)$	dimension of the space
$\mathbb{D}$, $\overline{\mathbb{D}}$	open and closed unit disk
$\mathbb{T}$	interval $[-\pi, \pi)$ (real numbers modulo 2π)
$\omega(g, \delta)$	modulus of continuity defined as $\sup_{\lvert s-t\rvert \le \delta} \lvert g(s) - g(t)\rvert$
$\omega_r(\cdot\,,\cdot)$	rth modulus of smoothness
$\omega_\phi^r(\cdot\,,\cdot)$	weighted ϕ-modulus of smoothness
$\mathcal{T}_m$	set of real trigonometric polynomials of degree m
$\lfloor x \rfloor$	largest integer not greater than x
$\mathcal{F}f$	Fourier transform of f
$\mathcal{L}f$	Laplace transform of f
$BV(X)$	set of functions with bounded variation on the interval X
$\mathrm{Var}_X f$	variation of f on the interval X
$\mathrm{Var}_X^* f$	variation of a correction f^* of f
$\mathbb{1}_X$	indicator function of the interval X

Chapter 1

Introduction

1.1 Motivation

Many problems in science and engineering are modeled by partial differential equations. After a discretization in space, for example, by finite differences, finite elements or pseudospectral methods, such problems can be written as a semi-linear system of ordinary differential equations

$$\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{g}\big(t, \boldsymbol{u}(t)\big)\,, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0 \tag{1.1}$$

with functions $\boldsymbol{u} : \mathbb{R} \to \mathbb{R}^N$, $\boldsymbol{g} : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$, and a large sparse discretization matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$. The parameter N depends on the chosen space grid. Moreover, $\boldsymbol{u}_0 \in \mathbb{R}^N$ denotes the initial value and t is the time parameter. Typically, the nonlinear part $\boldsymbol{g}\big(t, \boldsymbol{u}(t)\big)$ is non-stiff and the linear part $\boldsymbol{A}\boldsymbol{u}(t)$ is stiff. As the matrix $\boldsymbol{A}$ in the linear part of (1.1) usually stems from the discretization of an unbounded linear differential operator, the norm of $\boldsymbol{A}$ grows for finer and finer space grids.

There is no precise definition of "stiffness". In the literature, for instance [33, 36, 51], one can find various descriptions. On the one hand, stiff ordinary differential equations might be characterized by the fact that the eigenvalues λ_i of the discretization matrix $\boldsymbol{A}$ satisfy

$$\max_i |\mathrm{Re}(\lambda_i)| \gg \min_i |\mathrm{Re}(\lambda_i)|\,. \tag{1.2}$$

On the other hand, we often say that a given problem is stiff, if the use of an explicit numerical integration scheme requires impractically small time steps to obtain the desired accuracy, so that for the efficient integration an implicit method is needed which allows for larger time steps, but is more costly. A third possible definition is that stiff problems may have fast (stiff) and slowly (non-stiff) varying components. This is, for example, the case in chemical kinetics, where very fast and slow reactions take place simultaneously.

In this context, the example of the wave equation shows the difficulty of measuring stiffness: For explicit schemes, the Courant-Friedrichs-Lewy (CFL) condition enforces a restriction of the time step size dependent on the spatial mesh. To overcome this drawback, we thus have to use an implicit time integration method. However, the characterization (1.2) does not apply in this case, since the discretization matrix $\boldsymbol{A}$, corresponding to the first order formulation of the wave equation, has purely imaginary eigenvalues.

In order to cover these different cases of stiffness, we study problems of the form (1.1) with a matrix $\boldsymbol{A}$ whose field of values $W(\boldsymbol{A})$ is located somewhere in the closed left complex half-plane. More precisely, $W(\boldsymbol{A})$ is generally widely distributed in the left half-plane and the norm of $\boldsymbol{A}$ can become arbitrarily large. Roughly speaking, we can denote $\boldsymbol{A}$ as a "stiff" matrix.

Up to now, many numerical time integration schemes have been designed to handle stiff differential equations. These include exponential integrators which were developed in the 1960s and are primarily attributable to Certaine [11], Pope [67], Lawson [50], and Nørsett [61]. The idea behind this very important class of integrators is to solve the linear part $\boldsymbol{A}\boldsymbol{u}(t)$ of the model problem (1.1) exactly by the matrix exponential and to integrate the remaining nonlinear part by an explicit scheme. Here, the so-called φ-functions come into play, which are closely related to the exponential function. These φ-functions are given as

$$\varphi_\ell(z) = \int_0^1 e^{(1-\theta)z} \frac{\theta^{\ell-1}}{(\ell-1)!}\, d\theta\,, \qquad \ell \geq 1\,.$$

In the simplest case, we approximate the nonlinearity $\boldsymbol{g}\big(t, \boldsymbol{u}(t)\big)$ by $\boldsymbol{g}(0, \boldsymbol{u}_0)$ with $\boldsymbol{u}_0 = \boldsymbol{u}(0)$ and obtain the exponential Euler method

$$\boldsymbol{u}(\tau) \approx e^{\tau \boldsymbol{A}}\boldsymbol{u}_0 + \tau\varphi_1(\tau\boldsymbol{A})\boldsymbol{g}(0, \boldsymbol{u}_0)\,, \qquad \varphi_1(z) = \frac{e^z - 1}{z}\,,$$

which involves the action of the matrix exponential $e^{\tau\boldsymbol{A}}$ on $\boldsymbol{u}_0$ and the entire φ_1-function evaluated at $\tau\boldsymbol{A}$ times $\boldsymbol{g}(0, \boldsymbol{u}_0)$. If $\boldsymbol{g}$ is constant, the scheme reproduces the exact solution.

Exponential integrators have the great advantage that even if $\|\boldsymbol{A}\|$ is large, this does not imply a restriction of the admissible time step size τ in the integration. Furthermore, the error bounds do not depend on $\|\boldsymbol{A}\|$ as well. Since, in general, $\boldsymbol{A}$ is a huge matrix, the required matrix functions cannot be computed directly. This is why exponential integrators have been regarded as impractical for a long time. But in the last decades, it became apparent that there is hope to overcome this drawback by approximating the occurring products of matrix functions with vectors in a suitable manner.

One possibility is to project $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ onto a subspace of dimension $m \ll N$, reducing the problem to the evaluation of a matrix function for a small $m \times m$-matrix, see for instance [16, 23, 85]. This is the basic idea of the well-known standard Krylov subspace method, where $f(\boldsymbol{A})\boldsymbol{v}$ is approximated by the action of a polynomial matrix function on the vector $\boldsymbol{v} \in \mathbb{R}^N$. Besides the standard Krylov subspace method, rational Krylov subspace techniques have been studied recently in, e.g., [29–31, 46, 52, 59, 60, 62, 69, 84]. As their name suggests, these methods are based on a rational approximation.

In order to retain the beneficial properties of exponential integrators, it is crucial and indispensable to approximate the occurring matrix functions in such a way that the approximation quality is independent of $\|\boldsymbol{A}\|$. Rational Krylov subspace methods represent a very promising approach in this direction: In contrast to the polynomial Krylov method, the convergence of the rational process is independent of $\|\boldsymbol{A}\|$. This reveals the rational Krylov subspace method as the optimal choice for our purposes. The following standard model problem illustrates these facts.

We consider the one-dimensional heat equation $u' = \Delta u$ on the interval $(0, 1)$ with initial function $u_0(x) = x(1-x)$ and homogeneous Dirichlet boundary conditions. The discretization with finite differences leads to the system of ordinary differential equations

$$\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)\,, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0\,, \qquad \boldsymbol{A} = (N+1)^2\,\mathrm{tridiag}(1, -2, 1) \in \mathbb{R}^{N\times N}\,,$$

where the parameter N specifies how many inner space points are chosen in the considered domain from 0 to 1. In Figure 1.1, we plot the error of the approximation to the exact solution $e^{\tau\boldsymbol{A}}\boldsymbol{u}_0$ at time $\tau = 0.05$ with the polynomial (red dashed line) and a rational (blue solid line) Krylov subspace method against the number of iteration steps for $N = 50$ on the left and $N = 200$ on the right-hand side. The convergence behavior of the standard Krylov approximation is strongly linked to the number N of inner grid points. The finer the discretization the later the method starts to converge. The rational Krylov subspace process clearly outperforms the polynomial Krylov approximation and achieves a high accuracy after only a few iterations independent of the value N.

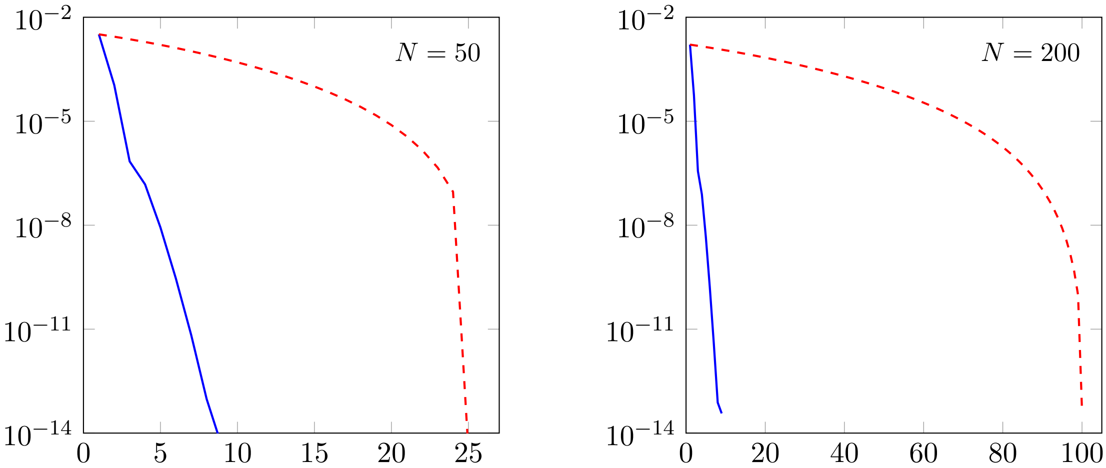


Figure 1.1: Comparison of the polynomial (red dashed line) and the rational (blue solid line) Krylov subspace method with $N = 50, 200$ discretization points.

However, the polynomial Krylov subspace method often works well in the first few iteration steps and then gives no further improvement. This is the case if the initial value u_0 of the differential equation fulfills specific smoothness properties. In our example involving the Laplace operator with homogeneous Dirichlet boundary conditions, smoothness refers to the order of differentiability of u_0 while preserving zero boundary conditions. In Figure 1.2, this effect is illustrated by replacing $u_0(x) = x(1 - x)$ with the "smoother" functions $u_0(x) = x^6(1 - x)^6$ (on the left) and $u_0(x) = x^8(1 - x)^8$ (on the right). The smoother the initial value the better the polynomial approximation performs in the first steps.

This observation has inspired the study of extended Krylov subspace methods in this thesis which combine the standard and the rational process in the following way: Initially, some polynomial Krylov steps are performed, until the convergence stagnates. Then the approximation is continued with the rational method. The first iteration steps are then cheap, since they usually involve only matrix-vector-products. As soon as the standard Krylov approximation does not further improve, the more efficient but more expensive rational Krylov process is used which requires solving a large linear system in each step.

For the approximation of $f(\boldsymbol{A})\boldsymbol{v}$ by an extended or a rational Krylov subspace process, several authors prove linear and superlinear convergence rates, e.g., [3, 4, 48]. These estimates depend on the geometry and the size of the field of values $W(\boldsymbol{A})$ of the matrix $\boldsymbol{A}$. They are very useful in the case that the geometry of $W(\boldsymbol{A})$ is known and that the field of values is a bounded set of moderate size. However, if $W(\boldsymbol{A})$ is huge, the error bounds suggest a very slow convergence. Therefore, these results are not suitable for our purposes. In this thesis, sublinear error bounds are derived which hold uniformly for all stiff matrices $\boldsymbol{A}$ with an arbitrarily large field of values in the left complex half-plane. Our bounds hold simultaneously for all reasonable space discretizations.

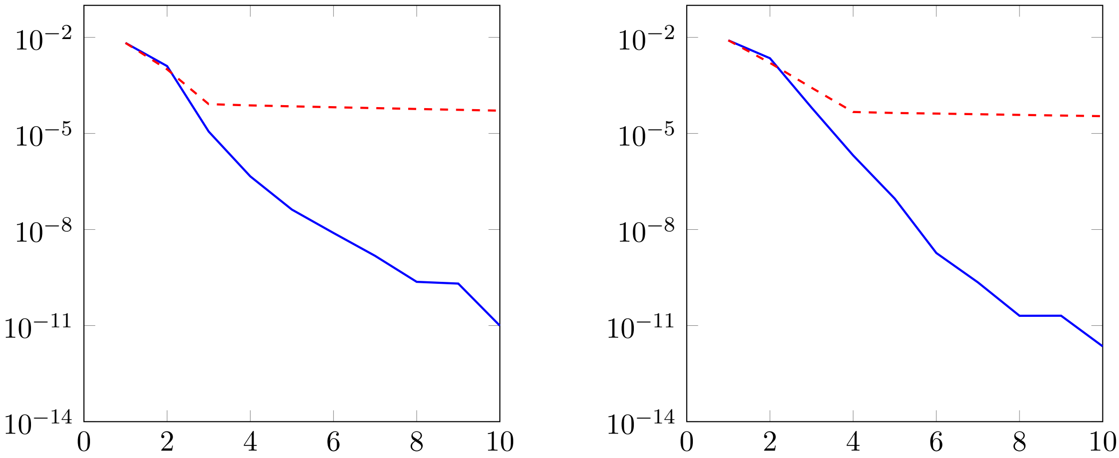


Figure 1.2: Convergence of the polynomial (red dashed line) and the rational (blue solid line) Krylov subspace approximation for the initial values $u_0(x) = x^6(1-x)^6$ (left) and $u_0(x) = x^8(1-x)^8$ (right).

1.2 Outline

The aim of this thesis is to analyze the convergence of rational and extended Krylov subspace methods for the approximation of the product of the matrix φ-functions and a vector, $\varphi(\boldsymbol{A})\boldsymbol{v}$, appearing in exponential integrators. With regard to semi-linear problems of the form (1.1) with stiff linear part, we are interested in error bounds that hold uniformly for all matrices $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ with field of values in the closed left complex half-plane. That means we are searching for convergence rates which are independent of the refinement of the spatial discretization.

In Chapter 2, we define matrix functions via the Jordan canonical form, polynomial interpolation and the Cauchy integral formula. Moreover, we give a brief overview of some fundamental properties of matrix functions.

The basic concepts and ideas of spatial discretization methods, strongly continuous semigroups and exponential integrators, especially exponential Runge-Kutta methods, are reviewed in Chapter 3.

Chapter 4 contains a description of standard and general rational Krylov subspaces as well as the approximation methods derived from these subspaces. In addition, the near-optimality property of Krylov subspace methods and the efficient computation of the approximation are discussed.

In the subsequent Chapter 5, a special class of rational Krylov subspace methods is introduced, namely the shift-and-invert Krylov subspace approximation, which uses one single repeated pole at $\gamma > 0$. We derive error bounds for a special class of functions and, in particular, for the φ-functions. Furthermore, suitable choices of γ are suggested.

A convergence analysis of the extended Krylov subspace approximation, whose convergence strongly depends on the abstract smoothness of the initial value, is presented in Chapter 6.

In Chapter 7, we turn to the approximation of matrix functions times a vector in a rational Krylov subspace with different simple poles, which are equidistantly distributed on a line in the right complex half-plane parallel to the imaginary axis.

All results obtained in Chapter 5, 6, and 7 are illustrated by several numerical experiments at the end of each chapter.

Finally, we give a short conclusion and a brief outlook in Chapter 8.

Chapter 2

Matrix functions

Based on the books [38] by Higham and [47] by Horn and Johnson, we give a brief overview on the theory of matrix functions in this chapter. The emphasis is placed on the different definitions of matrix functions, as needed for the numerical solution of ordinary or semi-discretized differential equations.

The most familiar matrix function is presumably the matrix exponential that can be used to express the solution of the homogeneous system

$$\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)\,, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0$$

by the formula $\boldsymbol{u}(t) = e^{t\boldsymbol{A}}\boldsymbol{u}_0$, where $\boldsymbol{A}$ is a constant matrix, whose entries do not depend on the time t. The probably most obvious definition of the matrix exponential $e^{t\boldsymbol{A}}$ is given by the well-known power series

$$e^{t\boldsymbol{A}} = \sum_{k=0}^{\infty} \frac{1}{k!}\, t^k \boldsymbol{A}^k\,,$$

which converges for all matrices $\boldsymbol{A}$, since we have

$$\|e^{t\boldsymbol{A}}\| \leq \sum_{k=0}^{\infty} \frac{1}{k!}\, t^k \|\boldsymbol{A}\|^k = e^{t\|\boldsymbol{A}\|} < \infty$$

for any sub-multiplicative matrix norm $\|\cdot\|$. This is just one of many possibilities for the computation of $e^{t\boldsymbol{A}}$. In the review [58] by Moler and Van Loan, the authors present twenty different ways to compute the exponential of a matrix.

For a general complex-valued function f and an arbitrary square matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$, there are several more or less equivalent ways to define a matrix function $f(\boldsymbol{A})$. In the following, we will confine ourselves to the three most important representations determined in terms of the Jordan canonical form, a Hermite interpolation polynomial as well as the Cauchy integral formula. We will see that the definition of $f(\boldsymbol{A})$ and its well-definedness are strongly related to the spectrum $\sigma(\boldsymbol{A})$ of $\boldsymbol{A}$, which is given by

$$\sigma(\boldsymbol{A}) := \{\lambda \in \mathbb{C} \,:\, \lambda \text{ eigenvalue of } \boldsymbol{A}\}\,.$$

Moreover, it will turn out that every matrix function $f(\boldsymbol{A})$ can be represented pointwise, that means for a fixed matrix $\boldsymbol{A}$, as a polynomial matrix function.

2.1 Jordan canonical form

Before we address the question of how a general function $f : \mathbb{C} \to \mathbb{C}$ can be extended to a mapping from $\mathbb{C}^{N\times N}$ to $\mathbb{C}^{N\times N}$, we consider the simple case of a polynomial $p \in \mathcal{P}_m$, where

$\mathcal{P}_m$ denotes the set of all polynomials with degree less than or equal to m. In this case, the polynomial function of a matrix $\boldsymbol{A}$ is defined by inserting $\boldsymbol{A}$ into the given polynomial.

Definition 2.1 *For $p(z) = a_m z^m + a_{m-1} z^{m-1} + \ldots + a_1 z + a_0 \in \mathcal{P}_m$ with $z \in \mathbb{C}$ and coefficients $a_0, \ldots, a_m \in \mathbb{C}$, the polynomial matrix function $p(\boldsymbol{A})$ is defined as*

$$p(\boldsymbol{A}) := a_m \boldsymbol{A}^m + a_{m-1} \boldsymbol{A}^{m-1} + \ldots + a_1 \boldsymbol{A} + a_0 \boldsymbol{I}\,.$$

It is well-known that every matrix $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ can be represented in Jordan canonical form

$$\boldsymbol{A} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}, \qquad \boldsymbol{J} = \operatorname{diag}(\boldsymbol{J}_{n_1}, \ldots, \boldsymbol{J}_{n_s})\,,$$

where $n_1 + \ldots + n_s = N$, $\boldsymbol{S} \in \mathbb{C}^{N \times N}$ is nonsingular, and $\boldsymbol{J} \in \mathbb{C}^{N \times N}$ is unique up to a permutation of the Jordan blocks $\boldsymbol{J}_{n_1}, \ldots, \boldsymbol{J}_{n_s}$. Each Jordan block is of the form

$$\boldsymbol{J}_{n_k} = \boldsymbol{J}_{n_k}(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix} \in \mathbb{C}^{n_k \times n_k}\,,$$

where the values λ_k are eigenvalues of the matrix $\boldsymbol{A}$, which are not necessarily distinct. By inserting $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}$ into a polynomial p, we find

$$p(\boldsymbol{A}) = \boldsymbol{S}p(\boldsymbol{J})\boldsymbol{S}^{-1}, \qquad p(\boldsymbol{J}) = \operatorname{diag}\big(p(\boldsymbol{J}_{n_1}), \ldots, p(\boldsymbol{J}_{n_s})\big)\,.$$

Using the Taylor expansion

$$p(z) = \sum_{j=0}^{m} \frac{p^{(j)}(\lambda_k)}{j!}\,(z - \lambda_k)^j$$

around the eigenvalue $\lambda_k \in \sigma(\boldsymbol{A})$ and writing $\boldsymbol{J}_{n_k}(\lambda_k) = \lambda_k \boldsymbol{I} + \boldsymbol{N}$, with the nilpotent matrix $\boldsymbol{N} = \boldsymbol{J}_{n_k}(0)$, we obtain

$$p(\boldsymbol{J}_{n_k}) = \sum_{j=0}^{m} \frac{p^{(j)}(\lambda_k)}{j!}\,(\lambda_k \boldsymbol{I} + \boldsymbol{N} - \lambda_k \boldsymbol{I})^j = \sum_{j=0}^{\min\{m, n_k - 1\}} \frac{p^{(j)}(\lambda_k)}{j!}\,\boldsymbol{N}^j\,,$$

since $\boldsymbol{N}^j = \boldsymbol{O}$ for $j \geq n_k$. Due to the special structure of $\boldsymbol{N}$, the matrix $\boldsymbol{N}^j$ has the value one on the jth upper diagonal and zeros elsewhere, and it follows that

$$p(\boldsymbol{J}_{n_k}) = \begin{bmatrix} p(\lambda_k) & p'(\lambda_k) & \cdots & \dfrac{p^{(n_k-1)}(\lambda_k)}{(n_k - 1)!} \\ & p(\lambda_k) & \ddots & \vdots \\ & & \ddots & p'(\lambda_k) \\ & & & p(\lambda_k) \end{bmatrix} \in \mathbb{C}^{n_k \times n_k}\,.$$

This shows that $p(\boldsymbol{A})$ is essentially determined by the derivatives of $p(z)$ at the eigenvalues of the matrix $\boldsymbol{A}$. It seems reasonable to generalize the definition of polynomial matrix functions $p(\boldsymbol{A})$ to arbitrary matrix functions $f(\boldsymbol{A})$. Therefore, we have to ensure that the required derivatives of f exist on $\sigma(\boldsymbol{A})$.

We recall that the minimal polynomial $p_{\boldsymbol{A}}^{\min}(z)$ of $\boldsymbol{A}$ is the unique monic polynomial of smallest degree such that $p_{\boldsymbol{A}}^{\min}(\boldsymbol{A}) = \boldsymbol{O}$. If we assume that $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ has r distinct eigenvalues $\lambda_1, \dots, \lambda_r$, this polynomial is given as

$$p_{\boldsymbol{A}}^{\min}(z) = \prod_{\lambda_k \in \sigma(\boldsymbol{A})} (z - \lambda_k)^{m_k} = \prod_{k=1}^{r} (z - \lambda_k)^{m_k} \,, \tag{2.1}$$

where the exponent m_k corresponds to the size of the largest Jordan block associated with the eigenvalue $\lambda_k \in \sigma(\boldsymbol{A})$. The minimal polynomial is a divisor of any other polynomial p with $p(\boldsymbol{A}) = \boldsymbol{O}$.

Definition 2.2 *Let $p_{\boldsymbol{A}}^{\min}$, as in (2.1), be the minimal polynomial of $\boldsymbol{A}$. A function f is said to be defined on the spectrum $\sigma(\boldsymbol{A})$ of $\boldsymbol{A}$, if $f^{(j)}(\lambda_k)$ exists for $j = 0, \dots, m_k - 1$ and $k = 1, \dots, r$.*

With these considerations in mind, we can state the next definition.

Definition 2.3 *Let f be defined on $\sigma(\boldsymbol{A})$ and let $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}$ be the Jordan canonical form of $\boldsymbol{A}$ with $\boldsymbol{J} = \operatorname{diag}(\boldsymbol{J}_{n_1}, \dots, \boldsymbol{J}_{n_s})$ and $\boldsymbol{J}_{n_k} = \boldsymbol{J}_{n_k}(\lambda_k) \in \mathbb{C}^{n_k \times n_k}$. Then we set*

$$f(\boldsymbol{A}) := \boldsymbol{S} f(\boldsymbol{J}) \boldsymbol{S}^{-1} = \boldsymbol{S} \operatorname{diag}\big(f(\boldsymbol{J}_{n_1}), \dots, f(\boldsymbol{J}_{n_s})\big)\, \boldsymbol{S}^{-1} \,,$$

where

$$f(\boldsymbol{J}_{n_k}) = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \cdots & \dfrac{f^{(n_k-1)}(\lambda_k)}{(n_k-1)!} \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix} \in \mathbb{C}^{n_k \times n_k} \,. \tag{2.2}$$

The matrix function $f(\boldsymbol{A})$ according to Definition 2.3 is well defined, that is, the definition does not depend on the particular Jordan canonical form (Horn and Johnson [47], Theorem 6.2.9). For a diagonalizable matrix $\boldsymbol{A}$, the blocks $f(\boldsymbol{J}_{n_k})$ are all of size one, and Definition 2.3 yields

$$f(\boldsymbol{A}) = \boldsymbol{S} \operatorname{diag}\big(f(\lambda_1), \dots, f(\lambda_N)\big)\, \boldsymbol{S}^{-1} \,, \qquad \boldsymbol{J} = \operatorname{diag}(\lambda_1, \dots, \lambda_N) \,.$$

In the case of multi-valued complex functions, such as the square root or the logarithm, it is common practice to use a single branch for the function f in each Jordan block, if an eigenvalue occurs in more than one block. These matrix functions are called primary. Taking distinct branches for the same eigenvalue in different Jordan blocks, a nonprimary matrix function is obtained. In this thesis, we will be only concerned with primary matrix functions.

2.2 Polynomial interpolation

A further representation of the matrix function $f(\boldsymbol{A})$ is based on polynomial interpolation. We assume again that $\lambda_1, \dots, \lambda_r$ are the distinct eigenvalues of the matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$. The first lemma of this section shows that a matrix polynomial $p(\boldsymbol{A})$ is completely determined by the values of p on the spectrum of $\boldsymbol{A}$.

Lemma 2.4 *Let $p^{\min}_{\boldsymbol{A}}(z) = (z-\lambda_1)^{m_1} \cdots (z-\lambda_r)^{m_r}$ be the minimal polynomial of $\boldsymbol{A}$ and let p, q be two polynomials. Then we have $p(\boldsymbol{A}) = q(\boldsymbol{A})$ if and only if*

$$p^{(j)}(\lambda_k) = q^{(j)}(\lambda_k) \quad \text{for} \quad j = 0, \dots, m_k - 1\,, \quad k = 1, \dots, r\,. \tag{2.3}$$

Proof. Higham [38], Theorem 1.3. ❑

Lemma 2.4 says that we may replace a given polynomial p in $p(\boldsymbol{A})$ by an arbitrary polynomial q satisfying (2.3) without changing the result. Similar to the considerations in the previous section, this property can be transferred to general functions. This leads to the following representation of a matrix function via polynomial interpolation.

Theorem 2.5 *Let f be defined on $\sigma(\boldsymbol{A})$ and let $p^{\min}_{\boldsymbol{A}}(z) = (z-\lambda_1)^{m_1} \cdots (z-\lambda_r)^{m_r}$ be the minimal polynomial of the matrix $\boldsymbol{A}$. We have $f(\boldsymbol{A}) = p(\boldsymbol{A})$ if and only if the $\nu := \sum_{k=1}^{r} m_k = \deg(p^{\min}_{\boldsymbol{A}})$ interpolation conditions*

$$p^{(j)}(\lambda_k) = f^{(j)}(\lambda_k) \quad \text{for} \quad j = 0, \dots, m_k - 1\,, \quad k = 1, \dots, r \tag{2.4}$$

are fulfilled.

Proof. We have to check that the definition of $f(\boldsymbol{A})$ via the Jordan canonical form in Definition 2.3 complies with the representation as matrix polynomial $p(\boldsymbol{A})$, where p has to fulfill the Hermite interpolation condition (2.4). The equivalence of both representations follows from the comparison of the individual blocks $f(\boldsymbol{J}_{n_k})$ defined in (2.2) corresponding to $f(\boldsymbol{A}) = \boldsymbol{S} f(\boldsymbol{J}) \boldsymbol{S}^{-1} = \boldsymbol{S}\,\mathrm{diag}\big(f(\boldsymbol{J}_{n_1}), \dots, f(\boldsymbol{J}_{n_s})\big)\,\boldsymbol{S}^{-1}$ and the blocks $p(\boldsymbol{J}_{n_k})$ corresponding to the polynomial $p(\boldsymbol{A}) = \boldsymbol{S}\,p(\boldsymbol{J})\,\boldsymbol{S}^{-1} = \boldsymbol{S}\,\mathrm{diag}\big(p(\boldsymbol{J}_{n_1}), \dots, f(\boldsymbol{J}_{n_s})\big)\,\boldsymbol{S}^{-1}$. ❑

By Theorem 2.5, every matrix function $f(\boldsymbol{A})$ can be written as a polynomial p in $\boldsymbol{A}$. The properties of p depend on the values of the function f and its derivatives on $\sigma(\boldsymbol{A})$. There exists a uniquely determined polynomial $p \in \mathcal{P}_{\nu-1}$ with $\nu = \deg(p^{\min}_{\boldsymbol{A}})$ that satisfies the Hermite interpolation condition (2.4). Theorem 2.5 implies further that $f(\boldsymbol{A})$ and $g(\boldsymbol{A})$ are equal if and only if

$$f^{(j)}(\lambda_k) = g^{(j)}(\lambda_k) \quad \text{for} \quad j = 0, \dots, m_k - 1\,, \quad k = 1, \dots, r\,.$$

An explicit formula for the Hermite interpolation polynomial that fulfills (2.4) is given by the Lagrange-Hermite formula

$$p(z) = \sum_{k=1}^{r} \left[\left(\sum_{j=0}^{m_k - 1} \frac{1}{j!} \Phi_k^{(j)}(\lambda_k)(z-\lambda_k)^j \right) \prod_{\substack{j=1 \\ j \neq k}}^{r} (z-\lambda_j)^{m_j} \right], \quad \Phi_k(z) = \frac{f(z)}{\displaystyle\prod_{\substack{j=1 \\ j \neq k}}^{r} (z-\lambda_j)^{m_j}}\,.$$

If $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ has N distinct eigenvalues, we have $m_k = 1$ for $k = 1, \dots, N$ in the minimal polynomial and the above formula reduces to the Lagrange interpolation polynomial

$$p(z) = \sum_{k=1}^{N} f(\lambda_k) \prod_{\substack{j=1 \\ j \neq k}}^{N} \frac{z - \lambda_j}{\lambda_k - \lambda_j}\,.$$

It is also possible to use divided differences for the computation of the interpolation polynomial. For this purpose, we define the tuple

$$(x_1,\dots,x_\nu) := (\underbrace{\lambda_1,\dots,\lambda_1}_{m_1},\underbrace{\lambda_2,\dots,\lambda_2}_{m_2},\dots,\underbrace{\lambda_r,\dots,\lambda_r}_{m_r})$$

that contains all distinct eigenvalues $\lambda_1,\dots,\lambda_r$ of $\boldsymbol{A}$ according to their multiplicity in the minimal polynomial in the specified order. Then

$$p(z) = \sum_{k=1}^{\nu} f[x_1,\dots,x_k] \prod_{j=1}^{k-1} (z - x_j)$$

is the desired Hermite interpolation polynomial. For a function f and points $x_k,\dots,x_{k+i}$, the divided differences are given as

$$\begin{cases} f[x_k] = f(x_k)\,, & \\ f[x_k,\dots,x_{k+i}] = \dfrac{f[x_{k+1},\dots,x_{k+i}] - f[x_k,\dots,x_{k+i-1}]}{x_{k+i}-x_k} & \text{for} \quad x_k \neq x_{k+i}\,, \\ f[x_k,\dots,x_{k+i}] = \dfrac{f^{(i)}(x_k)}{i!} & \text{for} \quad x_k = x_{k+i}\,. \end{cases}$$

Example 2.6 We consider the exponential function $f(z) = e^z$ for the matrix

$$\begin{aligned} \boldsymbol{A} = \begin{bmatrix} 4 & 1 & 0 & 1 \\ 1 & 1 & -1 & 3 \\ 0 & 1 & 4 & 1 \\ 1 & 3 & -1 & 1 \end{bmatrix} &= \begin{bmatrix} 0 & 2 & 0 & 1 \\ -\frac{1}{2} & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ \frac{1}{2} & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} -2 & 0 & 0 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0 & 2 & 0 & 1 \\ -\frac{1}{2} & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ \frac{1}{2} & 0 & 1 & 0 \end{bmatrix}^{-1} \\ &= \boldsymbol{SJS}^{-1} \end{aligned}$$

with minimal polynomial $p_{\boldsymbol{A}}^{\min}(z) = (z+2)(z-4)^3$. The corresponding Hermite interpolation polynomial that fulfills the required conditions $p(-2) = f(-2)$, $p^{(j)}(4) = f^{(j)}(4)$ for $j = 0, 1, 2$ is given by

$$\begin{aligned} p(z) &= e^4 + e^4(z-4) + \frac{1}{2}e^4(z-4)^2 + \frac{13e^4 - e^{-2}}{216}(z-4)^3 \\ &= \frac{13e^4 - e^{-2}}{216}z^3 - \frac{4e^4 - e^{-2}}{18}z^2 - \frac{e^4 + 2e^{-2}}{9}z + \frac{31e^4 + 8e^{-2}}{27}\,. \end{aligned}$$

By construction, we have $f(\boldsymbol{A}) = p(\boldsymbol{A})$ such that the matrix exponential $e^{\boldsymbol{A}}$ can be computed by inserting $\boldsymbol{A}$ into the polynomial p. Alternatively, one can use Definition 2.3 to obtain

$$e^{\boldsymbol{A}} = \boldsymbol{S} \begin{bmatrix} f(-2) & 0 & 0 & 0 \\ 0 & f(4) & f'(4) & \frac{1}{2}f''(4) \\ 0 & 0 & f(4) & f'(4) \\ 0 & 0 & 0 & f(4) \end{bmatrix} \boldsymbol{S}^{-1} = \begin{bmatrix} 2e^4 & e^4 & -e^4 & e^4 \\ e^4 & \frac{e^4+e^{-2}}{2} & -e^4 & \frac{e^4-e^{-2}}{2} \\ e^4 & e^4 & 0 & e^4 \\ e^4 & \frac{e^4-e^{-2}}{2} & -e^4 & \frac{e^4+e^{-2}}{2} \end{bmatrix}.$$

The exponential function is depicted in Figure 2.1 together with the associated Hermite interpolation polynomial p and its first two derivatives. ❍

Some useful properties of matrix functions are collected in the following theorem, which can be found as Theorem 1.13 in Higham [38]. The proof relies on the representation of $f(\boldsymbol{A})$ via polynomial interpolation.

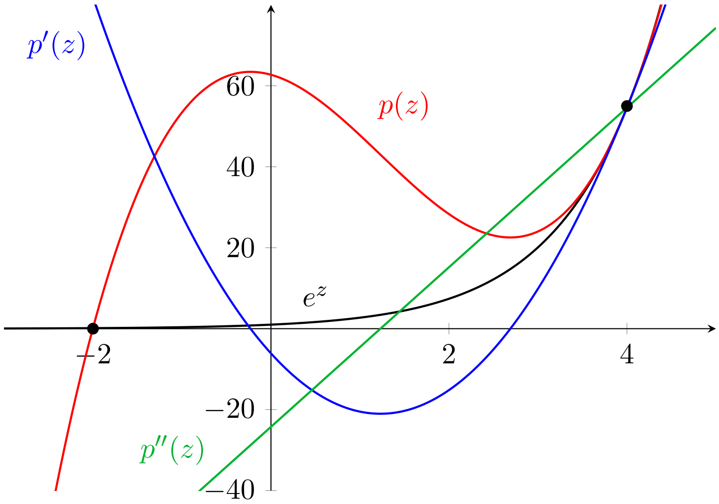


Figure 2.1: The exponential function and its Hermite interpolation polynomial p corresponding to the matrix $\boldsymbol{A}$.

Theorem 2.7 *Let the function f be defined on $\sigma(\boldsymbol{A})$, then*

1. *$f(\boldsymbol{A})$ and $\boldsymbol{A}$ commute with each other,*
2. *$f(\boldsymbol{A}^T) = f(\boldsymbol{A})^T$,*
3. *for invertible $\boldsymbol{X}$, we have $f(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{-1}) = \boldsymbol{X}f(\boldsymbol{A})\boldsymbol{X}^{-1}$,*
4. *$f(\lambda_k)$ are the eigenvalues of $f(\boldsymbol{A})$, where λ_k are the eigenvalues of $\boldsymbol{A}$,*
5. *$\boldsymbol{X}$ commutes with $f(\boldsymbol{A})$, if $\boldsymbol{X}$ commutes with $\boldsymbol{A}$.*

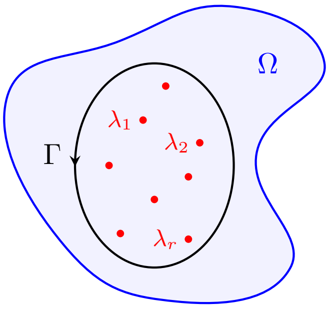
2.3 Cauchy integral formula

A third way of representing the matrix function $f(\boldsymbol{A})$ is via the Cauchy integral formula from complex analysis. The generalization of the Cauchy integral theorem from scalar functions to matrix functions gives rise to the following theorem.

Theorem 2.8 *Let $f : \Omega \to \mathbb{C}$ be analytic in the simply connected domain $\Omega \subset \mathbb{C}$ and let $\sigma(\boldsymbol{A}) \subset \Omega$. We have*

$$f(\boldsymbol{A}) = \frac{1}{2\pi i}\int_\Gamma f(\xi)(\xi\boldsymbol{I} - \boldsymbol{A})^{-1}\,d\xi\,,$$

where Γ is an arbitrary simple closed rectifiable curve that encloses $\sigma(\boldsymbol{A})$ in Ω and has winding number one.



For $\sigma(\boldsymbol{A}) \subset \mathrm{int}(\Gamma)$, the curve Γ is disjoint from the spectrum of $\boldsymbol{A}$ and the resolvent $(\xi\boldsymbol{I} - \boldsymbol{A})^{-1}$ in the integrand is well-defined. The resolvent is a matrix function as well. For better readability, we will often use the short notation $\frac{1}{\xi - \boldsymbol{A}}$ in the following instead of the equivalent expression $(\xi\boldsymbol{I} - \boldsymbol{A})^{-1}$.

Proof. [of Theorem 2.8] If we interpret the Cauchy integral as limit of Riemann sums in the normed space of matrices and consider the Jordan canonical form $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}$, we have

$$\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{A}}\,d\xi = \frac{1}{2\pi i}\int_\Gamma \boldsymbol{S}\,\frac{f(\xi)}{\xi - \boldsymbol{J}}\,\boldsymbol{S}^{-1}\,d\xi = \boldsymbol{S}\left(\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{J}}\,d\xi\right)\boldsymbol{S}^{-1}\,.$$

It follows further that

$$\begin{aligned}\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{J}}\,d\xi &= \frac{1}{2\pi i}\int_\Gamma f(\xi)\,\mathrm{diag}(\xi\boldsymbol{I} - \boldsymbol{J}_{n_1}, \ldots, \xi\boldsymbol{I} - \boldsymbol{J}_{n_s})^{-1}\,d\xi \\ &= \frac{1}{2\pi i}\int_\Gamma f(\xi)\,\mathrm{diag}\left(\frac{1}{\xi - \boldsymbol{J}_{n_1}}, \ldots, \frac{1}{\xi - \boldsymbol{J}_{n_s}}\right)d\xi \\ &= \mathrm{diag}\left(\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{J}_{n_1}}\,d\xi\,, \ldots, \frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{J}_{n_s}}\,d\xi\right).\end{aligned}$$

We have to show that the last expression is equal to $\mathrm{diag}\big(f(\boldsymbol{J}_{n_1}), \ldots, f(\boldsymbol{J}_{n_s})\big)$. With $\boldsymbol{J}_{n_k} = \lambda_k\boldsymbol{I} + \boldsymbol{N}$ and the Neumann series, we obtain

$$\begin{aligned}(\xi\boldsymbol{I} - \boldsymbol{J}_{n_k})^{-1} &= \big((\xi - \lambda_k)\boldsymbol{I} - \boldsymbol{N}\big)^{-1} = \frac{1}{\xi - \lambda_k}\left(\boldsymbol{I} - \frac{1}{\xi - \lambda_k}\boldsymbol{N}\right)^{-1} \\ &= \sum_{j=0}^{n_k - 1}\frac{1}{(\xi - \lambda_k)^{j+1}}\boldsymbol{N}^j\,.\end{aligned}$$

Cauchy's differentiation formula

$$\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{(\xi - \lambda_k)^{j+1}}\,d\xi = \frac{1}{j!}f^{(j)}(\lambda_k)$$

now yields

$$\begin{aligned}\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{J}_{n_k}}\,d\xi &= \sum_{j=0}^{n_k - 1}\left(\frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{(\xi - \lambda_k)^{j+1}}\,d\xi\right)\boldsymbol{N}^j \\ &= \sum_{j=0}^{n_k - 1}\frac{1}{j!}f^{(j)}(\lambda_k)\boldsymbol{N}^j = f(\boldsymbol{J}_{n_k})\end{aligned}$$

in accordance with Definition 2.3. ❑

The Cauchy integral representation in Theorem 2.8 is helpful for many theoretical results about matrix functions. Furthermore, this formula allows for the approximation of $f(\boldsymbol{A})$ by using a suitable quadrature rule, e.g., [75,83]. In contrast to the two previous representations of $f(\boldsymbol{A})$, the expression in terms of the Cauchy integral formula can be generalized to functions of operators (see [17], Section VII.3.6, Definition 9).

In general, the computation of $f(\boldsymbol{A})$ requires knowledge of the spectrum $\sigma(\boldsymbol{A})$ and the eigenvectors of the matrix $\boldsymbol{A}$. But in most cases, the eigenvalues and eigenvectors of $\boldsymbol{A}$ are not known precisely and can, if at all, only be computed approximately. Moreover, the computation of the Jordan canonical form is usually a numerically unstable process, since this form is very sensitive to perturbations. The Jordan canonical form is therefore commonly avoided in numerical analysis. Consequently, the representations of $f(\boldsymbol{A})$ based on the Jordan canonical form or the corresponding Hermite interpolation polynomial yield no adequate methods to determine $f(\boldsymbol{A})$ exactly. The Cauchy integral formula is typically not used for the exact computation of $f(\boldsymbol{A})$ as well.

Hence, one has to carefully design algorithms that compute accurate approximations to matrix functions for dense matrices of moderate size. This is still a subject of current research. Fortunately, for the matrix exponential and the matrix φ-functions, which are of interest in this thesis, algorithms for dense and moderate-sized matrices are known, cf. [2, 38, 77]. An overview about the computation of $f(\boldsymbol{A})$ for more general functions f and dense matrices $\boldsymbol{A}$ can be found in [38, 39].

In view of the numerical solution of differential equations, we are interested in the evaluation of the action of a matrix function $f(\boldsymbol{A})$ on a vector $\boldsymbol{v}$, without computing $f(\boldsymbol{A})$ explicitly. We will see that Krylov subspace methods are suitable to approximate $f(\boldsymbol{A})\boldsymbol{v}$ efficiently. Here, the methods for dense matrices cannot be applied. The occurring matrices $\boldsymbol{A}$ are sparse and large, but $f(\boldsymbol{A})$ is usually a large non-sparse matrix. The idea to circumvent the problems mentioned above is to project the large matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ onto some Krylov subspace of dimension $m \ll N$. This approach reduces $\boldsymbol{A}$ to a smaller matrix $\boldsymbol{S}_m$ of size $m \times m$ for which $f(\boldsymbol{S}_m)$ can be determined with standard algorithms for dense matrices of moderate size. We will come back to this issue in Chapter 4.

Chapter 3

Discretized evolution equations and exponential integrators

We are interested in the time integration of semi-linear problems of the form

$$u'(t) = Au(t) + g\big(t, u(t)\big)\,, \qquad u(0) = u_0\,, \tag{3.1}$$

which represents either a system of ordinary differential equations in $\mathbb{C}^N$, that stems from a suitable spatial discretization of a partial differential equation, or an abstract evolution equation on some Banach space with a linear, usually unbounded, differential operator A. Later on in this thesis, we will often restrict ourselves to Hilbert spaces, which are a special case of Banach spaces.

In the first case, we denote by $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ the discretization matrix and by $\boldsymbol{u} \in \mathbb{C}^N$ the approximation of the exact solution. For a finite-difference discretization, for example, this vector $\boldsymbol{u}$ contains approximate values of the solution u at certain grid points of the spatial domain. For a finite-element discretization, $\boldsymbol{u}$ is the coefficient vector of the nodal basis functions. In the discrete case, we thus write equation (3.1) as $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{g}\big(t, \boldsymbol{u}(t)\big)$, $\boldsymbol{u}(0) = \boldsymbol{u}_0$ with bold letters.

In the second case, one usually has to discretize the operator A by some kind of discretization process, such as finite-difference, pseudospectral, or finite-element methods. Therefore, we are concerned with matrices anyway. Nevertheless, we will study the following approximation methods in time for the abstract equation (3.1), in order to gain insight in the convergence behavior.

In what follows, we will consider stiff problems, where A is an unbounded operator on some Hilbert space H, or a huge matrix, whose norm can become arbitrarily large. Stiff discretization matrices $\boldsymbol{A}$ corresponding to the discretization of a partial differential equation may be characterized by a large field of values

$$W(\boldsymbol{A}) := \left\{ \frac{(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})} \,:\, \boldsymbol{x} \in \mathbb{C}^N,\, \boldsymbol{x} \neq \boldsymbol{0} \right\} = \left\{ (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) \,:\, \boldsymbol{x} \in \mathbb{C}^N,\, \|\boldsymbol{x}\| = 1 \right\}$$

located in the left complex half-plane, i.e., $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ with

$$\mathbb{C}_0^- := \{ z \in \mathbb{C} \,:\, \mathrm{Re}(z) \leq 0 \}\,,$$

see, for instance, [36, 51]. By $(\cdot\,,\cdot)$ we always denote a suitable inner product on $\mathbb{C}^N$ with associated norm $\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}\,,\boldsymbol{x})}$ for $\boldsymbol{x} \in \mathbb{C}^N$.

As a consequence of their bounded stability region, explicit integrators usually fail to integrate the linear part of (3.1) for stiff problems, unless impractically small time steps are used. Exponential integrators are an important class of numerical methods for the time integration of evolution equations which overcome this drawback. The name "exponential integrators" arises from the fact that these special integrators contain the matrix exponential or the operator exponential, i.e., the strongly continuous semigroup generated by A, and so-called φ-functions that are closely related to the exponential function.

Before we explain the ideas of exponential integrators and their construction based on the review [44] by Hochbruck and Ostermann, we outline the spatial discretization of partial differential equations, e.g., [9,55]. For fine space discretizations, the discretization matrices are huge and we might say that the matrix exponential corresponds to an approximation of the semigroup. We will need this correspondence later on, in order to understand the convergence of our methods. Therefore, we will summarize fundamental facts about strongly continuous semigroups and their generators following the books by Miklavčič [57], by Pazy [65], and by Engel and Nagel [19].

3.1 Spatial discretization

Using a spatial discretization, a partial differential equation is transformed into a system of ordinary differential equations of the form (3.1) with $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ and $\boldsymbol{u} \in \mathbb{C}^{N}$. This system of ordinary differential equations can then be solved with the help of standard time integration methods. As an example of a stiff system, we consider the two-dimensional heat equation with homogeneous Dirichlet boundary conditions

$$\begin{aligned} u' &= \Delta u && \text{for} \quad (x,y) \in \Omega\,,\ t \geq 0\,,\\ u(0,x,y) &= u_0(x,y) && \text{for} \quad (x,y) \in \Omega\,,\\ u(t,x,y) &= 0 && \text{for} \quad (x,y) \in \partial\Omega\,,\ t \geq 0 \end{aligned}$$

on the Hilbert space $L^2(\Omega)$ for a given domain $\Omega \subset \mathbb{R}^2$. The space $L^2(\Omega)$ contains all functions that are quadratically Lebesgue integrable on Ω, that is,

$$L^2(\Omega) := \{f \,:\, \Omega \to \mathbb{R} \,:\, \int_\Omega |f|^2 \, d(x,y) < \infty\}\,.$$

Moreover, Δ denotes the Laplacian $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ on $\mathbb{R}^2$.

In the following, we will focus on the finite-difference, the finite-element, and the spectral discretization, and present the basic ideas of these methods.

3.1.1 Finite differences

As domain Ω, we take a square. For the approximation, we discretize Ω by a uniform grid with mesh size $h = \frac{1}{n+1}$ and equidistant nodes $(x_i, y_j) = (x_0 + ih, y_0 + jh) \in \Omega$, $i,j = 0,\dots,n+1$, in each direction as depicted in Figure 3.1. If the solution u of the heat equation is sufficiently smooth, Taylor expansion yields

$$\frac{\partial^2}{\partial x^2} u(t,x_i,y_j) = \frac{1}{h^2}\big(u(t,x_i+h,y_j) - 2u(t,x_i,y_j) + u(t,x_i-h,y_j)\big) + \mathcal{O}(h^2)\,.$$

The same considerations apply to the second derivative of u with respect to the space variable y. Neglecting the remainder term of order h^2, which is small for a fine spatial

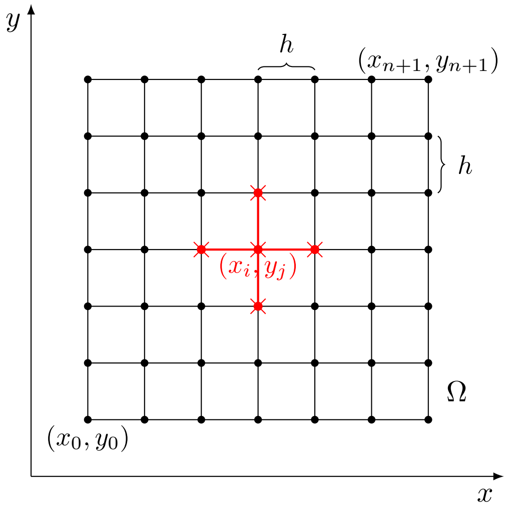


Figure 3.1: Two-dimensional grid for the finite-difference method on a square.

grid, we obtain

$$\Delta u(t, x_i, y_j) \approx \frac{1}{h^2}\big(u(t, x_i - h, y_j) + u(t, x_i, y_j - h) \\ - 4u(t, x_i, y_j) + u(t, x_i + h, y_j) + u(t, x_i, y_j + h)\big)\,.$$

The two-dimensional Laplacian is approximated by a so-called five-point-stencil, indicating that $\Delta u(t, x_i, y_j)$ is represented by a suitable linear combination of values of the function u at the point (t, x_i, y_j) itself and the four nearest neighbors in the spatial grid, see the cross-marked points in Figure 3.1. If we denote with $u_{i,j}$ and $u'_{i,j}$ the approximations to $u(t, x_i, y_j)$ and $u'(t, x_i, y_j)$, this procedure leads to the following system of ordinary differential equations

$$\begin{aligned} \frac{1}{h^2}(u_{i-1,j} + u_{i,j-1} - 4u_{i,j} + u_{i+1,j} + u_{i,j+1}) = u'_{i,j} \qquad &\text{for} \qquad (x_i, y_j) \in \Omega\,, \\ u_{i,j} = 0 \qquad &\text{for} \qquad (x_i, y_j) \in \partial\Omega\,. \end{aligned}$$

In matrix form, the system can be written as

$$\frac{1}{h^2}\begin{bmatrix} \boldsymbol{T} & \boldsymbol{I} & & \\ \boldsymbol{I} & \boldsymbol{T} & \ddots & \\ & \ddots & \ddots & \boldsymbol{I} \\ & & \boldsymbol{I} & \boldsymbol{T} \end{bmatrix} \cdot \begin{bmatrix} u_{1,1} \\ u_{2,1} \\ \vdots \\ u_{n,n} \end{bmatrix} = \begin{bmatrix} u'_{1,1} \\ u'_{2,1} \\ \vdots \\ u'_{n,n} \end{bmatrix}, \tag{3.2}$$

where $\boldsymbol{T}$ is the tridiagonal matrix

$$\boldsymbol{T} = \begin{bmatrix} -4 & 1 & & \\ 1 & -4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

and $\boldsymbol{I}$ is the identity matrix of dimension n. With regard to the domain Ω in Figure 3.1, the n^2 unknowns $u_{1,1}$, $u_{2,1}$, $\ldots$, $u_{n,n}$ are ordered row-wise from bottom left to top right.

This allows to reformulate the considered heat equation as the system

$$\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)\,, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0\,,$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $N = n^2$, is the block tridiagonal matrix $\boldsymbol{A} = \operatorname{tridiag}(\boldsymbol{I}, \boldsymbol{T}, \boldsymbol{I})$ and $\boldsymbol{u}(t) \in \mathbb{R}^N$ is a vector containing the values $u_{1,1}$, $u_{2,1}$, …, which are approximations to $u(t, x_1, y_1)$, $u(t, x_2, y_1)$, … on the inner grid points. A discrete approximation for the solution u is then determined by solving $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)$ with a time integration method for ordinary differential equations, where “discrete” means that the numerical solution is known only at certain points of the space domain Ω.

To reveal the stiff character of the semi-discrete system $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)$, we compute the eigenvalues of the discretization matrix $\boldsymbol{A}$. With the help of the Kronecker product $\otimes$, the matrix $\boldsymbol{A}$ is represented as

$$\boldsymbol{A} = \boldsymbol{I}\otimes\boldsymbol{L} + \boldsymbol{L}\otimes\boldsymbol{I}, \qquad \boldsymbol{L} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{n\times n}\,.$$

The eigenvalues of the tridiagonal matrix $\boldsymbol{L}$ are well-known, cf. [51], Section 2.10. They are given by $\lambda_k = -\frac{4}{h^2}\sin^2\left(\frac{k\pi}{2(n+1)}\right)$ for $k = 1, \dots, n$. For two matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ with eigenvalues $\mu_1, \dots, \mu_n$ and $\nu_1, \dots, \nu_n$, the eigenvalues of the Kronecker product $\boldsymbol{B}\otimes\boldsymbol{C}$ are $\mu_i\nu_j$ for $i, j = 1, \dots, n$. With the Jordan canonical form $\boldsymbol{L} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}$, $\boldsymbol{J} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, it follows that

$$\boldsymbol{I}\otimes\boldsymbol{L} + \boldsymbol{L}\otimes\boldsymbol{I} = (\boldsymbol{S}\otimes\boldsymbol{S})(\boldsymbol{I}\otimes\boldsymbol{J} + \boldsymbol{J}\otimes\boldsymbol{I})(\boldsymbol{S}\otimes\boldsymbol{S})^{-1}$$

by the properties of the Kronecker product. Since $\sigma(\boldsymbol{A}) = \sigma(\boldsymbol{I}\otimes\boldsymbol{J} + \boldsymbol{J}\otimes\boldsymbol{I})$, the eigenvalues of $\boldsymbol{A}$ are given by

$$-\frac{4}{h^2}\left(\sin^2\left(\frac{i\pi}{2(n+1)}\right) + \sin^2\left(\frac{j\pi}{2(n+1)}\right)\right)\,, \qquad i, j = 1, \dots, n\,.$$

For fine discretizations of the domain Ω, the mesh size h becomes small and $\sigma(\boldsymbol{A})$ contains negative eigenvalues of small as well as very large absolute value. This illustrates that we are concerned with a stiff problem.

If we want to measure the quality of the numerical solution $\boldsymbol{u}$, it is appropriate to scale the standard Euclidean norm $\|\cdot\|_2$ with the mesh size h of the space grid (e.g., [55], Section 6.1). This scaled Euclidean norm,

$$\|\boldsymbol{u}\|_h := h\|\boldsymbol{u}\|_2 = \left(h^2\sum_{i,j=1}^{n} u_{ij}^2\right)^{\frac{1}{2}} \approx \left(\int_\Omega u^2 d(x,y)\right)^{\frac{1}{2}}\,,$$

can be interpreted as a discrete L^2-norm. For an arbitrary matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$, the induced matrix norm $\|\cdot\|_h$ coincides with the spectral matrix norm $\|\cdot\|_2$, since

$$\|\boldsymbol{A}\|_h = \sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\|\boldsymbol{A}\boldsymbol{x}\|_h}{\|\boldsymbol{x}\|_h} = \sup_{\boldsymbol{x}\neq\boldsymbol{0}}\frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \|\boldsymbol{A}\|_2\,.$$

3.1.2 Finite elements

In contrast to the finite-difference method, where the exact solution is approximated at certain grid points, the finite-element method is based on taking an appropriate linear combination of some fixed nodal basis functions ϕ_i on subregions of the space domain Ω, that is,

$$u(t,x,y) \approx \sum_{i=1}^{N} u_i(t)\phi_i(x,y)\,.$$

We first state a variational formulation by rewriting our problem in a weak form. Subsequently, we discretize the problem with respect to the space variables x and y, which yields an approximate solution $\boldsymbol{u}(t) = \big(u_i(t)\big)_{i=1}^{N}$ in the finite-element space S_N of dimension N. In the simplest case, this finite-element space contains continuous, piecewise linear functions on a partition of the domain Ω containing triangular elements.

By $(\cdot\,,\cdot)_{L^2(\Omega)}$ and $\|\cdot\|_{L^2(\Omega)}$ we denote the inner product and its induced matrix norm

$$(v,w)_{L^2(\Omega)} = \int_\Omega vw\,d(x,y)\,, \qquad \|v\|_{L^2(\Omega)} = \sqrt{(v,v)_{L^2(\Omega)}}\,, \qquad v,w \in L^2(\Omega)\,.$$

We define the Sobolev space

$$H^k(\Omega) := \{v \in L^2(\Omega) \,:\, \frac{\partial^{\alpha+\beta}}{\partial^\alpha x\,\partial^\beta y}\,v(x,y) \in L^2(\Omega)\,,\ \alpha+\beta \le k\,,\ \alpha,\beta \in \mathbb{N}_0\}\,,$$

where the derivatives are understood in the weak sense. Moreover, we need the Sobolev space $H_0^1(\Omega)$, which is the closure of $C_0^\infty(\Omega)$ with respect to the Sobolev norm $\|\cdot\|_{H^1(\Omega)}$. Hereby, $C_0^\infty(\Omega)$ is the space of infinitely differentiable functions on Ω with compact support and the H^1-norm, induced by the inner product

$$(v,w)_{H^1(\Omega)} = \int_\Omega vw\,d(x,y) + \int_\Omega \nabla v\,\nabla w\,d(x,y)\,,$$

is given as

$$\|v\|_{H^1(\Omega)} = \Big(\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2\Big)^{\frac{1}{2}}\,,$$

where ∇v is a weak gradient vector that contains the weak partial derivatives of v. In fact, $H_0^1(\Omega)$ contains all functions whose weak derivative belongs to the space $L^2(\Omega)$ and that vanish on the boundary $\partial\Omega$ of the considered domain Ω.

To find the weak formulation for the heat equation, we multiply the differential equation by a test function $\phi \in H_0^1(\Omega)$, integrate over Ω, and apply Green's formula such that

$$\int_\Omega \phi u'\,d(x,y) = \int_\Omega \phi\,\Delta u\,d(x,y) = -\int_\Omega \nabla\phi\,\nabla u\,d(x,y) + \int_{\partial\Omega} \phi\,\nabla_{\boldsymbol{n}} u\,ds$$

for all $\phi \in H_0^1(\Omega)$, where $\nabla_{\boldsymbol{n}} u$ designates the normal derivative of u. The integral over $\partial\Omega$ vanishes for $\phi \in H_0^1(\Omega)$, since by assumption ϕ is equal to zero on $\partial\Omega$. Using the inner product on $L^2(\Omega)$, one can briefly note

$$(\phi, u')_{L^2(\Omega)} = -(\nabla\phi, \nabla u)_{L^2(\Omega)}\,.$$

Let $\phi_1, \dots, \phi_N$ be a basis of the finite-element space S_N. We replace $u(t,x,y)$ by a linear combination of these basis functions, that is, $u(t,x,y) \approx \sum_{i=1}^N u_i(t)\,\phi_i(x,y) \in S_N$, and search for coefficients $u_i(t)$ such that

$$\sum_{i=1}^N u_i'(t)(\phi_i, \phi_k)_{L^2(\Omega)} = -\sum_{i=1}^N u_i(t)(\nabla\phi_i, \nabla\phi_k)_{L^2(\Omega)} \quad \text{for} \quad k = 1, \dots, N\,. \tag{3.3}$$

If we approximate the initial function $u_0(x,y)$ by $\sum_{i=1}^N \gamma_i\,\phi_i(x,y) \in S_N$, we have $u_i(0) = \gamma_i$ for $i = 1, \dots, N$. With the vectors $\boldsymbol{u}(t) = \big(u_i(t)\big)_{i=1}^N$ and $\boldsymbol{u}_0 = (\gamma_i)_{i=1}^N$, equation (3.3) reads

$$\boldsymbol{M}\boldsymbol{u}'(t) = \boldsymbol{S}\boldsymbol{u}(t)\,, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0$$

in matrix notation, where $\boldsymbol{M}$ is the so-called mass matrix and $\boldsymbol{S}$ the stiffness matrix, whose entries are given by

$$(\boldsymbol{M})_{ij} = m_{ij} = (\phi_i, \phi_j)_{L^2(\Omega)}\,, \quad (\boldsymbol{S})_{ij} = s_{ij} = -(\nabla\phi_i, \nabla\phi_j)_{L^2(\Omega)}\,, \quad i,j = 1, \dots, N\,.$$

The solution of this system can then be approximated by standard methods for ordinary initial value problems. The matrix $\boldsymbol{M}$ is symmetric positive definite and thus invertible. Multiplying both sides with $\boldsymbol{M}^{-1}$ from the left yields the ordinary differential equation $\boldsymbol{u}'(t) = \boldsymbol{M}^{-1}\boldsymbol{S}\boldsymbol{u}(t)$ which has the form (3.1) with the stiff matrix $\boldsymbol{A} = \boldsymbol{M}^{-1}\boldsymbol{S}$ and $\boldsymbol{g} = \boldsymbol{0}$.

We have not yet considered the question of how the nodal basis functions ϕ_i look like in our case. The first step consists of generating a suitable triangulation over the domain Ω. First, the given domain is approximated by a polygon Ω_T. The triangulation is composed of triangles K_i, $i = 1, \dots, E$, such that the triangles meet edge-to-edge and vertex-to-vertex and $\Omega = K_1 \cup K_2 \cup \ldots \cup K_E$. We denote by h_i the diameter of the circumcircle of the triangle K_i and by ρ_i the radius of the circle inscribed in K_i. We assume that the ratio of h_i to ρ_i is smaller than some constant, so that triangles with very small or large angles are avoided. In the following, the inner nodes of our mesh are denoted by $a_1, \dots, a_N$. In the simplest case, we construct piecewise linear basis functions with

$$\phi_i(a_j) = \begin{cases} 1\,, & i = j\,, \\ 0\,, & i \neq j \end{cases} \qquad \text{for} \quad i,j = 1, \dots, N\,,$$

so that ϕ_i is equal to zero on all triangles that do not contain the vertex a_i. Since $\phi_i = 0$ for all i with $a_i \in \partial\Omega_T$, the basis functions are in $H_0^1(\Omega_T)$, and moreover form a basis of the finite-element space S_N.

A simple example, where the rectangular domain Ω_T is discretized with congruent triangles, is shown in Figure 3.2. Additionally, we depict one of the basis functions ϕ_i, that takes the value one at the vertex a_i and is equal to zero on all other vertices.

Since m_{ij} and s_{ij} are only different from zero, if the vertices a_i and a_j belong to the same triangle, the mass matrix $\boldsymbol{M}$ and the stiffness matrix $\boldsymbol{S}$ are sparse. This is also demonstrated in the following example.

Example 3.1 We take the simple 4×5 - grid with mesh size h and inner vertices $a_1, \dots, a_6$, which is shown in Figure 3.3. If we want to compute, e.g., the entry $s_{11} = (\nabla\phi_1, \nabla\phi_1)_{L^2(\Omega)}$ of the stiffness matrix $\boldsymbol{S}$, it suffices to consider the six triangles $K_1, \dots, K_6$ adjacent to a_1 separately to obtain

$$s_{11} = -\sum_{j=1}^6 \int_{K_j} \nabla\phi_1\,\nabla\phi_1\,d(x,y) = -\frac{1}{2}h^2\left(\frac{2}{h^2} + \frac{1}{h^2} + \frac{1}{h^2} + \frac{2}{h^2} + \frac{1}{h^2} + \frac{1}{h^2}\right) = -4\,.$$

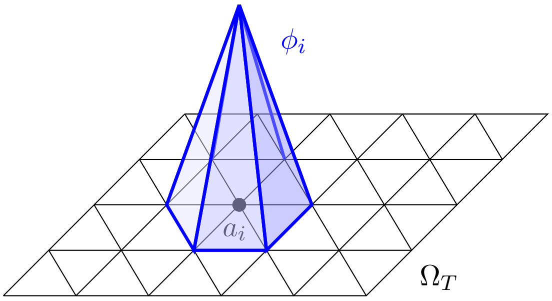


Figure 3.2: Basis function ϕ_i corresponding to the node a_i.

In order to determine s_{12}, we only have to investigate K_3 and K_4, since $\nabla\phi_1\nabla\phi_2$ is zero on the other triangles. This yields

$$s_{12} = -\frac{1}{2}h^2\left(-\frac{1}{h^2} - \frac{1}{h^2}\right) = 1\,.$$

Similar considerations apply to the remaining entries of the stiffness matrix $\boldsymbol{S}$. For the entries $(\boldsymbol{M})_{ij} = (\phi_i, \phi_j)_{L^2(\Omega)}$, $i,j = 1,\dots,6$, of the mass matrix, the computation is simplified by using a quadrature formula. If K is a triangle with area $|K|$ and (x_k, y_k), $k = 1,2,3$, are the midpoints of the sides, the quadrature rule

$$\int_K p(x,y)\,d(x,y) \approx \frac{|K|}{3}\big(p(x_1,y_1) + p(x_2,y_2) + p(x_3,y_3)\big)$$

is exact for any quadratic polynomial p. By making use of these facts, one can easily compute

$$\boldsymbol{M} = h^2 \cdot \frac{1}{12}\begin{bmatrix} 6 & 1 & 0 & 1 & 0 & 0 \\ 1 & 6 & 1 & 1 & 1 & 0 \\ 0 & 1 & 6 & 0 & 1 & 1 \\ 1 & 1 & 0 & 6 & 1 & 0 \\ 0 & 1 & 1 & 1 & 6 & 1 \\ 0 & 0 & 1 & 0 & 1 & 6 \end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix} -4 & 1 & 0 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 \\ 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & 0 & 1 & -4 \end{bmatrix}.$$

Note that $\boldsymbol{S}$ has the same structure as the matrix $\boldsymbol{A} = \boldsymbol{I} \otimes \boldsymbol{L} + \boldsymbol{L} \otimes \boldsymbol{I}$ from the finite-difference discretization. The eigenvalues of $\boldsymbol{M}^{-1}\boldsymbol{S}$ all lie in the left complex half-plane. Their proportionality to $\frac{1}{h^2}$ expresses again the stiff character of the heat equation. ❍

In practical computations, the inner products in the mass and the stiffness matrix, are computed in a clever way by an assembling process. We consider the single triangles element-wise, determine locally the corresponding integrals on each element, and assemble the derived information. For a fast and effective computation, the triangles are mapped by an affine transformation to a reference element $\hat{K}$, and the integration is performed by using an appropriate quadrature formula.

Analogously to the scaled Euclidean norm $\|\cdot\|_h = h\|\cdot\|_2$ for the finite-difference method, we have to use a suitable discrete L^2-norm. This can be motivated by the equality

$$\int_\Omega u^2\,d(x,y) \approx \sum_{i,j=1}^{N} u_i(t)\,u_j(t)\,(\phi_i,\phi_j)_{L^2(\Omega)} = \boldsymbol{u}(t)^T\boldsymbol{M}\boldsymbol{u}(t)\,.$$

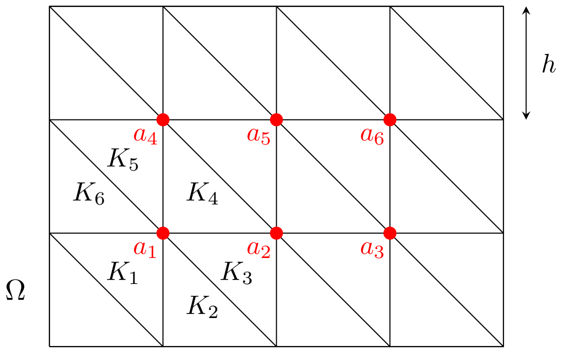


Figure 3.3: Regular grid used in Example 3.1.

Since the mass matrix $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ is symmetric and positive definite, there exists a unique matrix square root $\boldsymbol{M}^{1/2}$ which is also symmetric positive definite (cf. [26], Section 4.2.4). The matrix $\boldsymbol{M}$ is unitary diagonalizable by $\boldsymbol{M} = \boldsymbol{Q}\,\mathrm{diag}(\eta_1,\dots,\eta_N)\,\boldsymbol{Q}^T$ and the matrix square root $\boldsymbol{M}^{1/2}$ is thus given by $\boldsymbol{Q}\,\mathrm{diag}(\sqrt{\eta_1},\dots,\sqrt{\eta_N})\,\boldsymbol{Q}^T$. For $\boldsymbol{u} := \boldsymbol{u}(t) \in \mathbb{R}^N$, we therefore define

$$\|\boldsymbol{u}\|_{\boldsymbol{M}} = \sqrt{(\boldsymbol{u},\boldsymbol{u})_{\boldsymbol{M}}}\,, \qquad (\boldsymbol{u},\boldsymbol{u})_{\boldsymbol{M}} = \boldsymbol{u}^T\boldsymbol{M}\boldsymbol{u} = \|\boldsymbol{M}^{1/2}\boldsymbol{u}\|_2^2\,.$$

For an arbitrary matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$, we then obtain

$$\|\boldsymbol{A}\|_{\boldsymbol{M}} = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{M}}}{\|\boldsymbol{x}\|_{\boldsymbol{M}}} = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{M}^{1/2}\boldsymbol{A}\boldsymbol{M}^{-1/2}\boldsymbol{M}^{1/2}\boldsymbol{x}\|_2}{\|\boldsymbol{M}^{1/2}\boldsymbol{x}\|_2} = \|\boldsymbol{M}^{1/2}\boldsymbol{A}\boldsymbol{M}^{-1/2}\|_2\,.$$

3.1.3 Spectral methods

Later on in this thesis, we will also use a spectral discretization, but only for one-dimensional problems. For this reason, we mention only briefly the idea behind this discretization method. A detailed description is then given within the sections of the corresponding numerical experiments.

The spectral method is based on approximating the unknown solution u by a finite linear combination of the eigenfunctions $\psi_{j,k}(x,y)$ of the Laplacian with homogeneous Dirichlet boundary conditions. For instance, in the special case when $\Omega = (0,1)^2$, these eigenfunctions read

$$\psi_{j,k}(x,y) = C\cdot\sin(j\pi x)\sin(k\pi y)\,, \qquad j,k\in\mathbb{N}\,,$$

where the constant C is chosen such that $\int_\Omega \psi_{j,k}^2\,d(x,y) = 1$. The functions $\psi_{j,k}$ form an orthonormal basis of $L^2(\Omega)$. Because of

$$\Delta\psi_{j,k}(x,y) = -\pi^2(j^2+k^2)\psi_{j,k}(x,y)\,,$$

the corresponding eigenvalues are $-\pi^2(j^2+k^2)$ for $j,k\in\mathbb{N}$. A discretization is now obtained by substituting the ansatz

$$u(t,x,y) \approx \sum_{j,k=1}^{n} u_{j,k}(t)\psi_{j,k}(x,y)$$

in the given heat equation, which yields

$$\sum_{j,k=1}^{n} u'_{j,k}(t)\psi_{j,k}(x,y) = -\sum_{j,k=1}^{n} \pi^2(j^2+k^2)u_{j,k}(t)\psi_{j,k}(x,y)\,.$$

This represents a system of ordinary differential equations

$$\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)\,,$$

where $\boldsymbol{u}(t) \in \mathbb{R}^N$, $N = n^2$, contains the coefficients $u_{j,k}(t)$, that have to be determined. The discretization matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with entries $-\pi^2(j^2+k^2)$ for $j,k = 1,\dots,n$, and norm $\|\boldsymbol{A}\|_2 = 2\pi^2 N$.

In conclusion of this section, the following can be said: Not only for the heat equation, but also for general abstract or discretized differential equations, we should always keep in mind that we are concerned with unbounded operators or huge discretization matrices, whose norm can become arbitrarily large for very fine grids. In simple terms, one might say that the discretization matrix is approaching more and more the associated unbounded differential operator, if we refine the spatial grid. With regard to the approximation of the matrix exponential and related matrix φ-functions in exponential integrators, it will therefore be decisive to obtain error bounds that are not affected by the unboundedness of the operator A and that do not depend on the norm of the discretization matrix $\boldsymbol{A}$.

3.2 Strongly continuous semigroups

Let X be some Banach space. We denote by $\|\cdot\|$ the norm on X as well as the operator norm that is for a bounded operator $B : X \to X$ defined as

$$\|B\| = \sup_{\substack{x\in X\\ x\neq 0}} \frac{\|Bx\|}{\|x\|}\,.$$

In the following, we are concerned with the abstract semi-linear problem

$$u'(t) = Au(t) + g\big(t,u(t)\big)\,, \qquad u(0) = u_0\,, \tag{3.1}$$

where $A : D(A) \subseteq X \to X$ is a linear, in general unbounded, operator on X with domain of definition $D(A)$. If $g\big(t,u(t)\big) = 0$, (3.1) reduces to

$$u'(t) = Au(t)\,, \qquad u(0) = u_0\,. \tag{3.4}$$

For $u_0 \in D(A)$, we assume that there exists a unique solution $u(t)$ of (3.4). Then we can define an operator semigroup $T(t)$ such that

$$T(t)u_0 := u(t) \quad \text{for} \quad t \geq 0\,,$$

where $T(0) = I$ is the identity on X and the mapping $t \mapsto T(t)u_0$ is continuous from $\mathbb{R}_0^+$ to the Banach space X. If we choose $u(s)$ as initial value, the uniqueness of the solution implies that

$$T(t)u(s) = T(t)T(s)u_0 = u(t+s) = T(t+s)u_0\,,$$

indicating the semigroup property $T(t)T(s) = T(t+s)$. These fundamental facts are summarized in the following definition.

Definition 3.2 *A family $\big(T(t)\big)_{t\geq 0}$ of bounded linear operators on a Banach space X is called a strongly continuous semigroup (or shortly C_0-semigroup), if the following conditions are fulfilled:*

(a) We have $T(t+s) = T(t)T(s)$ for all $t, s \geq 0$ and $T(0) = I$.

(b) For every $x \in X$, the orbit map

$$\xi_x : \mathbb{R}_0^+ \to X\,, \quad t \mapsto T(t)x$$

is continuous.

It is well-known that right continuity of the orbit map at zero implies continuity of ξ_x on $[0, \infty)$. That is, we can replace part (b) in Definition 3.2 by the requirement $\lim_{t\searrow 0} \|T(t)x - x\| = 0$ for all $x \in X$. We immediately derive from Definition 3.2 that the operators commute, since

$$T(t)T(s) = T(t+s) = T(s+t) = T(s)T(t)\,.$$

Furthermore, one can easily prove by induction that

$$T(nt) = T(t + \ldots + t) = T(t)^n \quad \text{for} \quad n \in \mathbb{N}\,.$$

By the continuity of the orbit map ξ_x, it follows that $T(t)$ is locally bounded on compact intervals $[0, t_0]$, that is $\|T(t)x\| < \infty$ for all $t \in [0, t_0]$, $t_0 > 0$, and every $x \in X$. From the Uniform Boundedness Principle[1], we conclude that a C_0-semigroup is uniformly bounded on each compact interval of $\mathbb{R}_0^+$. This fact implies that every strongly continuous semigroup is exponentially bounded. More exactly, there exist constants $M \geq 1$ and $\omega \geq 0$ such that the inequality

$$\|T(t)\| \leq M e^{\omega t} \quad \text{for all} \quad t \geq 0 \tag{3.5}$$

holds true. To see this, we choose $t_0 > 0$ and $M \geq 1$ with $\|T(s)\| \leq M$ for all $s \in [0, t_0]$ and set $t = s + nt_0$ for $n \in \mathbb{N}_0$. Then

$$\|T(t)\| \leq \|T(s)\| \|T(t_0)\|^n \leq M^{n+1} \leq M e^{\omega n t_0} \leq M e^{\omega t} \quad \text{for all} \quad t \geq 0\,,$$

where $\omega = \ln(M)/t_0 \geq 0$. If the semigroup satisfies inequality (3.5), we also say that $T(t)$ is of type (M, ω). At this point, it should be noted that there are also semigroups which satisfy (3.5) with $\omega < 0$.

Since $T(t)u_0$ can be regarded as the unique solution of the abstract equation (3.4) with initial value u_0, we should analyze strongly continuous semigroups with respect to their differentiability. First, we point out that right differentiability of the orbit map ξ_x at $t = 0$ is equivalent to differentiability of ξ_x on $\mathbb{R}_0^+$. Its derivative is given by

$$\xi_x'(t) = T(t)\,\xi_x'(0) \quad \text{for all} \quad t \geq 0\,.$$

The right derivative $\xi_x'(0)$ at $t = 0$ yields an operator A that is called the infinitesimal generator of the C_0-semigroup.

[1] Uniform Boundedness Principle: If a set T of bounded linear operators is pointwise bounded, then it is uniformly bounded (e.g., Theorem 3.17 in [74]).

Definition 3.3 *The infinitesimal generator (or simply generator) $A : D(A) \subseteq X \to X$ of a strongly continuous semigroup $\big(T(t)\big)_{t\geq 0}$ is defined as*

$$Ax := \xi_x'(0) = \frac{d}{dt}T(t)x\Big|_{t=0} = \lim_{h\searrow 0}\frac{1}{h}\big(T(h)x - x\big)$$

for every x in the domain

$$D(A) := \{x \in X \,:\, \lim_{h\searrow 0}\frac{1}{h}\big(T(h)x - x\big) \text{ exists}\}\,. \tag{3.6}$$

The following example illustrates that every matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ generates, as a special case of a linear operator on $\mathbb{C}^N$, a C_0-semigroup. In analogy to the matrix exponential, it provides a motivation to think of the semigroup $T(t)$ as an operator exponential e^{tA}. For this reason, we will also write e^{tA} for $T(t)$.

Example 3.4 We consider a matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ on $X = \mathbb{C}^N$. It is well known that the solution of the initial value problem $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)$, $\boldsymbol{u}(0) = \boldsymbol{u}_0$, is given by

$$T(t)\boldsymbol{u}_0 := e^{t\boldsymbol{A}}\boldsymbol{u}_0\,, \qquad t \in \mathbb{R}\,.$$

One can easily check that $e^{t\boldsymbol{A}}$ satisfies the semigroup properties and that $\frac{d}{dt}\,e^{t\boldsymbol{A}} = \boldsymbol{A}e^{t\boldsymbol{A}}$ for all $t \in \mathbb{R}$. Consequently, $T(t)$ is not only a C_0-semigroup, but moreover a C_0-group (i.e., Definition 3.2 holds for $t \in \mathbb{R}$) with infinitesimal generator $\boldsymbol{A}$. If the field of values $W(\boldsymbol{A})$ is contained in the left complex half-plane, the semigroup is bounded by one, that is, $\|e^{t\boldsymbol{A}}\| \leq 1$ with $M = 1$ and $\omega = 0$ in (3.5), cf. Lemma 7.1 below. Such semigroups of type $(1, 0)$ are called contraction semigroups. ❍

It follows directly from the definition that the generator A of a strongly continuous semigroup $T(t)$ is a linear operator. Another important property is that the C_0-semigroup and its generator commute on $D(A)$. If $x \in D(A)$, then also $T(t)x \in D(A)$ and

$$\frac{d}{dt}T(t)x = T(t)Ax = AT(t)x \quad \text{for all} \quad t \geq 0\,.$$

The infinitesimal generator A is a closed and densely defined operator that determines the semigroup uniquely. Closedness means that $x_n \to x$ and $Ax_n \to y$ in X for any sequence $(x_n)_{n\in\mathbb{N}}$ in $D(A)$ implies $x \in D(A)$ and $Ax = y$. An operator A is called densely defined, if its domain $D(A)$ is dense in X.

It is also important to look at the spectral properties of the generator A. For this purpose, we recall the following definition.

Definition 3.5 *The resolvent set of a closed linear operator $A : D(A) \subseteq X \to X$ is defined by*

$$\rho(A) := \{z \in \mathbb{C} \,:\, zI - A \text{ is bijective}\}\,.$$

The spectrum of A is given as $\sigma(A) := \mathbb{C} \setminus \rho(A)$.

Assuming $z \in \rho(A)$, the inverse $R(z, A) := (zI - A)^{-1}$ exists. It is also called the resolvent of A. If A is closed, the resolvent $R(z, A)$ is closed as well. By definition, the domain of $R(z, A)$ is equal to X and we can conclude from the Closed Graph Theorem[2] that

$$R(z, A) \,:\, X \to D(A)\,, \qquad z \in \rho(A)$$

[2] Closed Graph Theorem: If X, Y are Banach spaces and $B : D(B) \subseteq X \to Y$ is a closed linear operator with $D(B) = X$, then B is bounded (e.g., Theorem 3.10 in [74]).

is a bounded operator on X. Furthermore, we have the identity

$$zR(z,A) - I = AR(z,A)\,, \qquad z \in \rho(A)$$

and, for all $x \in D(A)$, the resolvent commutes with A, i.e., $AR(z,A)x = R(z,A)Ax$. The mapping $z \mapsto R(z,A)$, $z \in \rho(A)$, is infinitely many times complex differentiable with

$$\frac{d^n}{dz^n} R(z,A) = (-1)^n\, n!\, R(z,A)^{n+1}\,, \qquad n \in \mathbb{N}\,.$$

For a strongly continuous semigroup of type (M,ω) with generator A, powers of the resolvent are bounded by

$$\|R(z,A)^n\| \le \frac{M}{(\mathrm{Re}(z) - \omega)^n} \quad \text{for} \quad \mathrm{Re}(z) > \omega\,, \quad n \in \mathbb{N}\,. \tag{3.7}$$

Conversely, the following theorem provides necessary and sufficient conditions, including (3.7), for A to generate a C_0-semigroup.

Theorem 3.6 (Hille-Yosida) *A linear operator $A : D(A) \subseteq X \to X$ is the infinitesimal generator of a C_0-semigroup with*

$$\|T(t)\| \le Me^{\omega t} \quad \textit{for} \quad M \ge 1, \quad \omega \in \mathbb{R} \quad \textit{and all} \quad t \ge 0$$

if and only if the following conditions are satisfied:

1. *A is a closed and densely defined operator.*
2. *For every $z \in \mathbb{C}$ with $\mathrm{Re}(z) > \omega$ it holds that $z \in \rho(A)$ and inequality (3.7) is satisfied for all $n \in \mathbb{N}$.*

Moreover, we mention another important generation theorem. For later purposes, this theorem is only required for the case of a Hilbert space H.

Theorem 3.7 (Lumer-Phillips) *Let A be a linear operator on some Hilbert space H. If $\mathrm{Re}(Ax,x) \le 0$ for all $x \in D(A)$ and $\mathrm{Range}(z_0 I - A) = H$ for some z_0 with $\mathrm{Re}(z_0) > 0$, then A is the infinitesimal generator of a C_0-semigroup of contractions on H.*

Resolvents can be used for the approximation of the semigroup $T(t)$ by a rational function. If A is an infinitesimal generator of a strongly continuous semigroup, a common resolvent based approximation is the implicit Euler scheme

$$e^{tA}x = T(t)x = \lim_{n\to\infty} \left(\frac{n}{t}\, R\left(\frac{n}{t}, A\right)\right)^n x = \lim_{n\to\infty} \left(I - \frac{t}{n}A\right)^{-n} x\,, \qquad x \in X\,. \tag{3.8}$$

Later on in this thesis, we will always be concerned with operators A generating a contraction semigroup of type $(1,0)$. In this case, we have $\|\frac{n}{t}R(\frac{n}{t},A)^n\| \le 1$ and the conditions in Brenner and Thomée [10] for the rational approximation of a semigroup are fulfilled. Under these assumptions, the implicit Euler method represents a special case of the results in [10]. For $x \in D(A^2)$, Brenner and Thomée have shown that the implicit Euler scheme is convergent of order one. More precisely, it holds

$$\left\| T(t)x - \left(I - \frac{t}{n}A\right)^{-n} x \right\| \le C\,\frac{t^2}{n}\,\|A^2 x\|\,.$$

A similar bound of this form can also be obtained for general semigroups of type (M, ω), if we consider a rescaled semigroup $A - \omega I$ and define a new norm being equivalent to $\|\cdot\|$ to avoid the constant M.

A second possible approximation to $T(t)$, that is strongly related to the implicit Euler method and represents a polynomial approach, is the explicit Euler scheme

$$e^{tA}x = T(t)x = \lim_{n\to\infty} \left(I + \frac{t}{n}A\right)^n x\,.$$

The existence of this limit is only ensured, if A is a bounded operator. Whereas (3.8) involves powers of the bounded resolvent, the explicit Euler formula contains powers of the possibly unbounded operator A, so that the explicit Euler scheme generally may fail to exist or converge, respectively. To guarantee that $(I + \frac{t}{n}A)^n x$ is well defined for all $n \in \mathbb{N}$, we must assume that $x \in \bigcap_{n=1}^{\infty} D(A^n)$, which imposes a strong restriction on x.

3.3 Exponential Runge-Kutta methods

Exponential integrators provide an interesting class of numerical methods for the time integration of stiff ordinary differential equations. The basic idea of these integrators is to separate the linear term of the differential equation $u'(t) = Au(t) + g\bigl(t, u(t)\bigr)$, that can be solved exactly, from the nonlinear part $g\bigl(t, u(t)\bigr)$. An important class of exponential integrators are exponential Runge-Kutta methods which rely on the variation of constants formula

$$u(t_n + \tau) = e^{\tau A}u(t_n) + \int_0^{\tau} e^{(\tau-\sigma)A} g\bigl(t_n + \sigma, u(t_n + \sigma)\bigr)\, d\sigma \tag{3.9}$$

for the solution of the semi-linear problem at time $t_{n+1} = t_n + \tau$. In the functional analytic framework, the notation $e^{\tau A}$ is here used for the C_0-semigroup $T(\tau) = e^{\tau A}$ generated by the linear operator A on some Banach space X, or respectively, for the matrix exponential in the finite dimensional case. The following results hold for operators A, that generate a strongly continuous semigroup, as well as for matrices, stemming from a spatial discretization of an abstract differential operator.

The construction of exponential Runge-Kutta schemes is similar to standard Runge-Kutta methods. We approximate the integral in (3.9) by a quadrature formula with nodes $0 \le c_i \le 1$ and weights $b_i(\tau A)$, $i = 1, \dots, s$, in which the nonlinearity is approximated and the semigroup is treated exactly. Since the integral involves the unknown solution u, internal stages are required. The internal stage values $U_{ni} \approx u(t_n + c_i\tau)$ are computed by another quadrature formula with the same nodes c_i and weights $a_{ij}(\tau A)$ applied to

$$u(t_n + c_i\tau) = e^{c_i\tau A}u(t_n) + \int_0^{c_i\tau} e^{(c_i\tau-\sigma)A} g\bigl(t_n + \sigma, u(t_n + \sigma)\bigr)\, d\sigma\,. \tag{3.10}$$

Assume we are given an approximation $u_n \approx u(t_n)$, this leads to the exponential Runge-Kutta scheme

$$\begin{aligned}
U_{ni} &= e^{c_i\tau A}u_n + \tau\sum_{j=1}^{s} a_{ij}(\tau A)G_{nj} \approx u(t_n + c_i\tau)\,,\\
G_{nj} &= g(t_n + c_j\tau, U_{nj}) \approx g\bigl(t_n + c_j\tau, u(t_n + c_j\tau)\bigr)\,,\\
u_{n+1} &= e^{\tau A}u_n + \tau\sum_{i=1}^{s} b_i(\tau A)G_{ni} \approx u(t_n + \tau)\,.
\end{aligned}$$

If we set A equal to zero, this one-step method reduces to the standard Runge-Kutta scheme with coefficients $a_{ij}(0)$ and $b_i(0)$. A desirable feature of numerical time integration methods is the preservation of equilibria u^* satisfying $Au^* + g(t, u^*) = 0$. By postulating $u^* = u_n = U_{ni}$ for all i and all n, it follows that the coefficients have to fulfill

$$\sum_{i=1}^{s} b_i(z) = \varphi_1(z)\,, \qquad \sum_{j=1}^{s} a_{ij}(z) = c_i\varphi_1(c_i z)\,. \tag{3.11}$$

Replacing $g\big(t_n + \sigma, u(t_n + \sigma)\big)$ in the variation of constants formulas (3.9) and (3.10) by an interpolation polynomial with nodes $c_1, \dots, c_s$, we can conclude that the weights $a_{ij}(z)$ and $b_i(z)$ of the exponential Runge-Kutta method may be expressed as a linear combination of the φ-functions

$$\varphi_\ell(z) = \int_0^1 e^{(1-\theta)z} \frac{\theta^{\ell-1}}{(\ell-1)!}\, d\theta\,, \qquad \ell \geq 1\,. \tag{3.12}$$

The first two of these φ-functions are given by

$$\varphi_1(z) = \frac{e^z - 1}{z}\,, \qquad \varphi_2(z) = \frac{e^z - 1 - z}{z^2}\,.$$

In the literature, a number of equivalent definitions for the φ-function can be found besides the integral representation (3.12), for example,

$$\varphi_\ell(z) = \frac{e^z - t_{\ell-1}(z)}{z^\ell}\,, \qquad t_{\ell-1}(z) = \sum_{k=0}^{\ell-1} \frac{z^k}{k!}\,, \tag{3.13}$$

where $t_{\ell-1}(z)$ is the $(\ell-1)$st order Taylor polynomial of the exponential function. These functions can be extended holomorphically to the point zero by $\varphi_\ell(0) = \frac{1}{\ell!}$ and are therefore analytic for all $z \in \mathbb{C}$. Moreover, the φ-functions fulfill the recurrence relation

$$\varphi_{\ell+1}(z) = \frac{\varphi_\ell(z) - \frac{1}{\ell!}}{z} \quad \text{for} \quad \ell \geq 0 \quad \text{with} \quad \varphi_0(z) := e^z\,.$$

By our assumption, A generates a strongly continuous semigroup on some Banach space X, so that we can use formula (3.12) and the bound (3.5) to conclude that the operator functions $\varphi_\ell(\tau A)$ are bounded on X by

$$\|\varphi_\ell(\tau A)\| \leq \int_0^1 \underbrace{\|e^{(1-\theta)\tau A}\|}_{\leq M e^{\omega\tau(1-\theta)}} \frac{\theta^{\ell-1}}{(\ell-1)!}\, d\theta \leq M\varphi_\ell(\omega\tau)\,.$$

The simplest exponential Runge-Kutta method is to take $s = 1$ and $c_1 = 0$, corresponding to an approximation of the nonlinearity in the integral by $g\big(t_n + \sigma, u(t_n + \sigma)\big) \approx g(t_n, u_n)$. This yields the exponential Euler method

$$u_{n+1} = e^{\tau A} u_n + \tau\varphi_1(\tau A) g(t_n, u_n)\,. \tag{3.14}$$

In practical applications, where the operator A is represented by a large matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ after a discretization in space, it is advantageous to replace the matrix exponential $e^{\tau \boldsymbol{A}}$ by the equivalent expression $\boldsymbol{I} + \tau\varphi_1(\tau\boldsymbol{A})\boldsymbol{A}$. This results in the representation

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + \tau\varphi_1(\tau\boldsymbol{A})\big(\boldsymbol{A}\boldsymbol{u}_n + \boldsymbol{g}(t_n, \boldsymbol{u}_n)\big)\,,$$

which can be evaluated more efficiently than (3.14), since only one product of a matrix function and a vector has to be computed instead of two. In the next chapter, we will discuss Krylov subspace methods that constitute a powerful tool for the approximation of such products of a matrix function with some vector.

The construction of exponential Runge-Kutta methods can best be explained by means of the linear problem

$$u'(t) = Au(t) + f(t)\,, \qquad u(0) = u_0\,. \tag{3.15}$$

In this case, the variation of constants formula (3.9) simplifies to

$$\begin{aligned} u(t_n+\tau) &= e^{\tau A}u(t_n) + \int_0^\tau e^{(\tau-\sigma)A} f(t_n+\sigma)\,d\sigma \\ &= e^{\tau A}u(t_n) + \tau \int_0^1 e^{(1-\theta)\tau A} f(t_n+\tau\theta)\,d\theta\,. \end{aligned} \tag{3.16}$$

Substituting $f(t_n+\tau\theta)$ by the Lagrange interpolation polynomial

$$f(t_n+\tau\theta) = \sum_{i=1}^{s} f(t_n+c_i\tau)\ell_i(\theta)\,, \qquad \ell_i(\theta) = \prod_{\substack{j=1\\ j\neq i}}^{s} \frac{\theta-c_j}{c_i-c_j}\,,$$

with $0 \le c_i \le 1$ and $c_i \neq c_j$ for $i \neq j$, we obtain the exponential quadrature rule

$$u_{n+1} = e^{\tau A}u_n + \tau \sum_{i=1}^{s} b_i(\tau A) f(t_n+c_i\tau)\,. \tag{3.17}$$

For $s = 2$, for example, we have the weights

$$\begin{aligned} b_1(z) &= \frac{1}{c_1-c_2}\,\varphi_2(z) - \frac{c_2}{c_1-c_2}\,\varphi_1(z)\,, \\ b_2(z) &= \frac{1}{c_2-c_1}\,\varphi_2(z) - \frac{c_1}{c_2-c_1}\,\varphi_1(z)\,, \end{aligned}$$

which satisfy the first requirement in (3.11). The exponential trapezoidal rule is obtained by taking the nodes $c_1 = 0$ and $c_2 = 1$.

For the convergence analysis of the exponential quadrature rule (3.17), we compute the Taylor expansion of $f(t_n+\sigma)$ around t_n in the variation of constants formula (3.16). Similarly, we expand the term $f(t_n+c_i\tau)$ in the numerical solution (3.17) in a Taylor series and compare both expansions. Solving the error recursion $e_n = u_n - u(t_n)$, it is proven in [43] by Hochbruck and Ostermann that the error is uniformly bounded by

$$\|u_n - u(t_n)\| \le C\tau^p \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \|f^{(p)}(\sigma)\|\,d\sigma\,, \qquad t_n \in [0,T]\,, \tag{3.18}$$

for $f^{(p)} \in L^1(0,T;X)$, if the method satisfies the order conditions

$$\varphi_j(\tau A) - \sum_{i=1}^{s} b_i(\tau A)\,\frac{c_i^{j-1}}{(j-1)!} = 0\,, \qquad j = 1,\dots,p\,.$$

The space $L^1(0,T;X)$ contains all functions $g : [0,T] \to X$ such that $\int_0^T \|g(s)\|\,ds < \infty$ for the norm $\|\cdot\|$ on X. The constant C in (3.18) depends only on T, but not on the chosen step size τ. Especially, for the exponential Euler method, we have the error bound

$$\|u_n - u(t_n)\| \le C\tau \sup_{0\le t\le t_n} \|f'(t)\|\,,$$

whenever the function $f : [0,T] \to X$ is differentiable with $\sup_{t\in[0,T]} \|f'(t)\| < \infty$.

For a more general semi-linear problem (3.1), involving the nonlinearity $g\big(t, u(t)\big)$, the construction and analysis becomes more complicated. In this case, we have expressions of the form $e^{\tau A}\big(g(t_n, u_n) - g\big(t_n, u(t_n)\big)\big)$ that have to be bounded in a suitable way. For a detailed description of exponential Runge-Kutta methods for this more general semi-linear problem, we refer the reader to [42].

Usually, the semi-linear problem in (3.1) arises from a fixed linearization of some evolution equation $u'(t) = F\big(t, u(t)\big)$, leading to $F\big(t, u(t)\big) = Au(t) + g\big(t, u(t)\big)$ with $A \approx \frac{\partial F}{\partial u}(t_0, u_0)$. By using a continuous linearization along the current numerical solution $u_n \approx u(t_n)$ instead, we obtain so-called exponential Rosenbrock methods, whose simplest representative is the exponential Rosenbrock-Euler method given by

$$u_{n+1} = u_n + \tau\varphi_1(\tau A_n)F(t_n, u_n)\,, \qquad A_n = \frac{\partial F}{\partial u}(t_n, u_n)\,.$$

Beside exponential one-step methods, it is also possible to construct exponential multistep methods that are related to explicit Adams methods. The idea behind these multistep methods is to exploit the information from previous time steps for the calculation of the next approximate value u_{n+1} (cf. Hochbruck and Ostermann [45]).

Chapter 4

Krylov subspace methods

In the previous chapter, we have seen that the application of an exponential integrator to the semi-discrete problem $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{g}\big(t, \boldsymbol{u}(t)\big)$ requires the evaluation of the product of the matrix φ-functions with some vector $\boldsymbol{v}$, that is, $\varphi_\ell(\tau\boldsymbol{A})\boldsymbol{v}$ for some $\ell \in \mathbb{N}$. In general, we are concerned with large and sparse discretization matrices $\boldsymbol{A} \in \mathbb{C}^{N\times N}$. Since the matrix function $f(\boldsymbol{A})$ is usually not sparse for arbitrary functions f defined on $\sigma(\boldsymbol{A})$, it is inconvenient to first compute $f(\boldsymbol{A})$ and then to multiply the result by the vector $\boldsymbol{v}$. Instead, we will approximate the action of $f(\boldsymbol{A})$ on the vector $\boldsymbol{v}$ by projecting the problem onto a suitable Krylov subspace of much smaller dimension than N.

Principally, there are two main options available: The use of a standard (polynomial) Krylov subspace yields an approximation of the form

$$f(\boldsymbol{A})\boldsymbol{v} \approx p_{m-1}(\boldsymbol{A})\boldsymbol{v}\,, \qquad p_{m-1} \in \mathcal{P}_{m-1}\,,$$

whereas rational Krylov subspace methods lead to an approximation

$$f(\boldsymbol{A})\boldsymbol{v} \approx r_{m-1}(\boldsymbol{A})\boldsymbol{v}\,, \qquad r_{m-1} = \frac{p_{m-1}}{q_{m-1}} \in \frac{\mathcal{P}_{m-1}}{q_{m-1}}\,.$$

Hereby, we denote by

$$\frac{\mathcal{P}_{m-1}}{q_{m-1}} = \left\{ \frac{p_{m-1}(z)}{q_{m-1}(z)} \,:\, p_{m-1} \in \mathcal{P}_{m-1} \right\}$$

the space of all rational functions with numerator polynomial of degree at most $m-1$ and a fixed chosen denominator polynomial $q_{m-1} \in \mathcal{P}_{m-1}$, whose roots have to be distinct from the eigenvalues of $\boldsymbol{A}$. The properties of the rational Krylov subspace approximation are determined by the particular choice of q_{m-1}.

Before we study standard and rational Krylov subspace methods on the basis of [30, 38, 41, 69, 70, 72, 73, 84], we first summarize the most important facts about orthogonal projections following [30] and [73]. After that, we discuss the near-optimality property and the efficient computation of Krylov subspace approximations, especially for the case of the matrix φ-functions (see [72, 77]).

4.1 Orthogonal projections

Since our approximation methods for $f(\boldsymbol{A})\boldsymbol{v}$ are based on the projection onto some Krylov subspace, we resume here the basic results on projectors. Later on, we will solely deal with orthogonal projections and, therefore, we only describe these.

A projector $\boldsymbol{P} \in \mathbb{C}^{N\times N}$ is a linear mapping from a vector space $\mathbb{C}^N$ to itself such that $\boldsymbol{P}^2 = \boldsymbol{P}$ is fulfilled. If $\boldsymbol{P}$ is a projector, the same holds true for the so-called complementary projector $\boldsymbol{I} - \boldsymbol{P}$. It is well-known that $\operatorname{Range}(\boldsymbol{P}) = \operatorname{Null}(\boldsymbol{I} - \boldsymbol{P})$, and conversely $\operatorname{Range}(\boldsymbol{I} - \boldsymbol{P}) = \operatorname{Null}(\boldsymbol{P})$. Furthermore, we have $\operatorname{Range}(\boldsymbol{P}) \cap \operatorname{Null}(\boldsymbol{P}) = \{\boldsymbol{0}\}$. This shows that a projector separates the vector space $\mathbb{C}^N$ into two complementary subspaces, that is, $\mathbb{C}^N = \operatorname{Range}(\boldsymbol{P}) \oplus \operatorname{Null}(\boldsymbol{P})$.

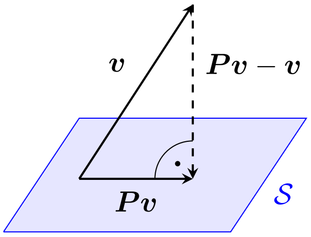


Figure 4.1: Orth. projection of $\boldsymbol{v}$ onto the subspace $\mathcal{S}$.

Let $(\cdot\,,\cdot)$ denote the inner product on $\mathbb{C}^N$. Then an orthogonal projector onto the subspace $\mathcal{S}$ is defined by the requirement that

$$\boldsymbol{Pv} \in \mathcal{S} \quad \text{and} \quad \boldsymbol{Pv} - \boldsymbol{v} \perp \mathcal{S}\,,$$

where orthogonality is meant with respect to the inner product on $\mathbb{C}^N$. Alternatively, a projector can be recognized as an orthogonal projector by the condition

$$(\boldsymbol{v}, \boldsymbol{Pw}) = (\boldsymbol{Pv}, \boldsymbol{w}) \quad \text{for all} \quad \boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^N\,.$$

That means $\boldsymbol{P}$ has to be self-adjoint with respect to the chosen inner product. For the Euclidean inner product $(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{w}^H\boldsymbol{v}$, we simply have $\boldsymbol{P} = \boldsymbol{P}^H$.

Given a matrix $\boldsymbol{V}_m \in \mathbb{C}^{N\times m}$, whose columns build an orthonormal basis of the m-dimensional subspace $\mathcal{S}$, the orthogonal projector $\boldsymbol{P}_m$ reads $\boldsymbol{V}_m\boldsymbol{V}_m^+$, where $\boldsymbol{V}_m^+ \in \mathbb{C}^{m\times N}$ is the Moore-Penrose pseudoinverse of $\boldsymbol{V}_m$.

Definition 4.1 *For $\boldsymbol{V}_m \in \mathbb{C}^{N\times m}$, the Moore-Penrose inverse of $\boldsymbol{V}_m$ is the unique solution of the four equations*

$$\begin{aligned} \boldsymbol{V}_m\boldsymbol{V}_m^+\boldsymbol{V}_m &= \boldsymbol{V}_m\,, & \qquad (\boldsymbol{V}_m^+\boldsymbol{V}_m)^* &= \boldsymbol{V}_m^+\boldsymbol{V}_m\,, \\ \boldsymbol{V}_m^+\boldsymbol{V}_m\boldsymbol{V}_m^+ &= \boldsymbol{V}_m^+\,, & \qquad (\boldsymbol{V}_m\boldsymbol{V}_m^+)^* &= \boldsymbol{V}_m\boldsymbol{V}_m^+\,, \end{aligned}$$

where $\boldsymbol{V}_m^+\boldsymbol{V}_m \in \mathbb{C}^{m\times m}$ and $\boldsymbol{V}_m\boldsymbol{V}_m^+ \in \mathbb{C}^{N\times N}$.

The notation $\boldsymbol{B}^*$ in Definition 4.1 indicates the adjoint of $\boldsymbol{B}$ with respect to the chosen inner product $(\cdot\,,\cdot)$ on $\mathbb{C}^N$ or $\mathbb{C}^m$, fulfilling the property

$$(\boldsymbol{Bv}, \boldsymbol{w}) = (\boldsymbol{v}, \boldsymbol{B}^*\boldsymbol{w}) \quad \text{for all} \quad \boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^N \quad \text{or} \quad \boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^m\,.$$

Hereby, the vector space $\mathbb{C}^m$ is always endowed with the standard Euclidean inner product, whereas the inner product on $\mathbb{C}^N$ is customized for the current application.

Any inner product $(\cdot\,,\cdot) : \mathbb{C}^N \times \mathbb{C}^N \to \mathbb{C}$ can be written as

$$(\boldsymbol{v}, \boldsymbol{w})_{\boldsymbol{M}} = \boldsymbol{w}^H\boldsymbol{Mv}\,, \qquad \boldsymbol{v}, \boldsymbol{w} \in \mathbb{C}^N$$

with a positive definite Hermitian matrix $\boldsymbol{M} \in \mathbb{C}^{N\times N}$ (e.g., [27]). The orthogonal projector onto $\mathcal{S}$ with respect to $(\cdot\,,\cdot)_{\boldsymbol{M}}$ is thus given as $\boldsymbol{P}_m = \boldsymbol{V}_m\boldsymbol{V}_m^+ = \boldsymbol{V}_m\boldsymbol{V}_m^H\boldsymbol{M}$. This can easily be verified by checking that $\boldsymbol{V}_m^+ = \boldsymbol{V}_m^H\boldsymbol{M}$ fulfills all conditions in Definition 4.1. Since $\boldsymbol{M}$ is a positive definite Hermitian matrix, we can conclude that

$$\|\boldsymbol{v}\|_{\boldsymbol{M}} = \sqrt{(\boldsymbol{v}, \boldsymbol{v})_{\boldsymbol{M}}} = \|\boldsymbol{M}^{1/2}\boldsymbol{v}\|_2\,.$$

Considering, for example, the finite-element method, the matrix $\boldsymbol{M}$ is given as the mass matrix. If $\mathbb{C}^N$ is equipped with the standard Euclidean inner product, we have $\boldsymbol{M} = \boldsymbol{I}$ and $\boldsymbol{V}_m^+ = \boldsymbol{V}_m^H$, so that $\boldsymbol{P}_m = \boldsymbol{V}_m\boldsymbol{V}_m^H$.

Of course, an orthogonal projector can also be defined in the continuous case: Let H be some Hilbert space with inner product $(\cdot,\cdot)$ and let $\mathcal{S} \subseteq H$ be an m-dimensional subspace onto which we want to project orthogonally. Since H is a Hilbert space, we can find an orthonormal basis $v_1, \ldots, v_m$ of $\mathcal{S}$, that we collect in the so-called quasi-matrix[1] $V_m = [v_1\, v_2\, \cdots\, v_m]$. For the projection operator, we use the analogue notation $P_m = V_m V_m^+$ with

$$V_m \;:\; \left\{ \begin{array}{ccc} \mathbb{C}^m & \to & H \\ \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} & \mapsto & \displaystyle\sum_{j=1}^m a_j v_j \end{array} \right. \qquad \text{and} \qquad V_m^+ \;:\; \left\{ \begin{array}{ccc} H & \to & \mathbb{C}^m \\ v & \mapsto & \begin{bmatrix} (v,v_1) \\ \vdots \\ (v,v_m) \end{bmatrix} \end{array} \right. .$$

Then the orthogonal projection of an arbitrary vector $v \in H$ onto $\mathcal{S}$ is given as

$$P_m v = V_m V_m^+ v = V_m \begin{bmatrix} (v,v_1) \\ \vdots \\ (v,v_m) \end{bmatrix} = \sum_{j=1}^m (v,v_j)v_j\,.$$

4.2 Standard Krylov subspace

Krylov methods have a long history in numerical analysis. The standard (also called polynomial) Krylov subspace goes back to 1931, when the Russian applied mathematician Aleksey N. Krylov studied the computation of eigenvalues by using the sequence $\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}^2\boldsymbol{v}, \ldots,$ in order to find the characteristic polynomial coefficients of the matrix $\boldsymbol{A}$, see [49]. For that reason, subspaces of the form as in the following definition have later been named as Krylov subspaces.

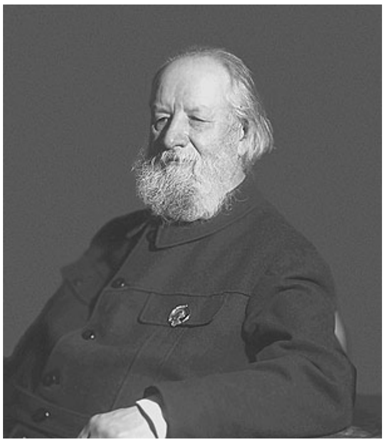


Figure 4.2: Krylov in the 1930s[2]

Definition 4.2 *For $m \geq 1$, the mth Krylov subspace of the matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ and the vector $\boldsymbol{v} \in \mathbb{C}^N$ is defined by*

$$\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}) := \operatorname{span}\{\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}, \ldots, \boldsymbol{A}^{m-1}\boldsymbol{v}\} = \{p(\boldsymbol{A})\boldsymbol{v} \,:\, p \in \mathcal{P}_{m-1}\}\,.$$

$\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$ contains all linear combinations of the images of the vector $\boldsymbol{v}$ under the kth power of the matrix $\boldsymbol{A}$ for $k = 0, \ldots, m-1$. These are exactly all matrix polynomials of

[1] The term “quasi-matrix” originates from Stewart [82].

[2] http://en.wikipedia.org/wiki/Aleksey_Krylov

degree smaller than or equal to $m-1$ times $\boldsymbol{v}$. The Krylov subspace is invariant under arbitrary shifts, i.e., $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_m(\boldsymbol{A}-\gamma\boldsymbol{I},\boldsymbol{v})$ for any $\gamma \in \mathbb{C}$.

Since we have seen in Chapter 2 that all matrix functions $f(\boldsymbol{A})$ can be represented by a matrix polynomial, it seems reasonable to use an approximation of the form

$$f(\boldsymbol{A})\boldsymbol{v} \approx p(\boldsymbol{A})\boldsymbol{v} \in \mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}), \qquad p \in \mathcal{P}_{m-1}.$$

Besides the minimal polynomial $p^{\min}_{\boldsymbol{A}}(z) = (z-\lambda_1)^{m_1}\cdots(z-\lambda_r)^{m_r}$, see (2.1), we define the minimal polynomial $p^{\min}_{\boldsymbol{A},\boldsymbol{v}}(z)$ of $\boldsymbol{A}$ with respect to the vector $\boldsymbol{v}$ to be the monic polynomial of lowest degree such that $p^{\min}_{\boldsymbol{A},\boldsymbol{v}}(\boldsymbol{A})\boldsymbol{v} = \boldsymbol{0}$. This polynomial has the form

$$p^{\min}_{\boldsymbol{A},\boldsymbol{v}}(z) = (z-\lambda_1)^{l_1}\cdots(z-\lambda_r)^{l_r} \quad \text{with} \quad l_k \le m_k, \quad k = 1,\dots,r,$$

and divides any polynomial p with $p(\boldsymbol{A})\boldsymbol{v} = \boldsymbol{0}$, especially the minimal polynomial of $\boldsymbol{A}$. The dimension of $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$ is for $m = 1,2,\dots$ equal to m, until m reaches $\deg(p^{\min}_{\boldsymbol{A},\boldsymbol{v}}) =: \mu$. For $m \ge \mu$, the Krylov subspace is invariant under $\boldsymbol{A}$ and we have $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_\mu(\boldsymbol{A},\boldsymbol{v})$, that is,

$$\mathcal{K}_1(\boldsymbol{A},\boldsymbol{v}) \subsetneq \mathcal{K}_2(\boldsymbol{A},\boldsymbol{v}) \subsetneq \ldots \subsetneq \mathcal{K}_\mu(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_{\mu+1}(\boldsymbol{A},\boldsymbol{v}) = \ldots.$$

We therefore call μ the invariance index of the Krylov subspace $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$. The next theorem shows that for every function f, which is defined on the spectrum of $\boldsymbol{A}$, we have $f(\boldsymbol{A})\boldsymbol{v} \in \mathcal{K}_\mu(\boldsymbol{A},\boldsymbol{v})$ (see, e.g., Higham [38], Theorem 13.2).

Theorem 4.3 *Let f be defined on $\sigma(\boldsymbol{A})$ and let $p^{\min}_{\boldsymbol{A},\boldsymbol{v}}(z) = (z-\lambda_1)^{l_1}\cdots(z-\lambda_r)^{l_r}$ be the minimal polynomial of $\boldsymbol{A}$ with respect to $\boldsymbol{v}$. If p is the unique Hermite interpolation polynomial with* $\deg(p) < \mu = l_1 + \ldots + l_r$ *satisfying*

$$p^{(j)}(\lambda_k) = f^{(j)}(\lambda_k) \quad \textit{for} \quad j = 0,\dots,l_k-1, \quad k = 1,\dots,r,$$

then $f(\boldsymbol{A})\boldsymbol{v} = p(\boldsymbol{A})\boldsymbol{v}$ holds true.

Usually, the invariance index μ can be quite large. In this case, we choose an approximation for $f(\boldsymbol{A})\boldsymbol{v}$ in the mth Krylov subspace of order $m < \mu$ by using a suitable compression of the matrix $\boldsymbol{A}$ onto $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$. Before we can define the Krylov subspace approximation of $f(\boldsymbol{A})\boldsymbol{v}$, we have to construct an orthonormal basis $\boldsymbol{V}_m = [\boldsymbol{v}_1\,\boldsymbol{v}_2\,\cdots\,\boldsymbol{v}_m] \in \mathbb{C}^{N\times m}$ of $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$. Here, we restrict ourselves to the Euclidean inner product and its induced matrix norm, the spectral norm, and write briefly $\|\cdot\|$ instead of $\|\cdot\|_2$. In the subsequent Section 4.5, we will then explain how the following facts and results can be easily adapted to a general inner product on $\mathbb{C}^N$. For the computation of the orthonormal basis $\boldsymbol{V}_m$, a stabilized Gram-Schmidt process is used, which is also known as Arnoldi procedure. It is described in Algorithm 4.4. Assuming that we are given an orthonormal basis $\boldsymbol{V}_m$, the new basis vector $\boldsymbol{v}_{m+1}$ is derived from

$$\widetilde{\boldsymbol{v}}_{m+1} = \boldsymbol{A}\boldsymbol{v}_m - \sum_{j=1}^{m} h_{j,m}\boldsymbol{v}_j, \qquad \boldsymbol{v}_{m+1} = \frac{1}{h_{m+1,m}}\widetilde{\boldsymbol{v}}_{m+1}, \qquad h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|.$$

The orthogonality condition $\boldsymbol{v}_j^H\boldsymbol{v}_{m+1} = 0$ for $j = 1,\dots,m$ implies that the coefficients must satisfy $h_{j,m} = \boldsymbol{v}_j^H\boldsymbol{A}\boldsymbol{v}_m$. This yields the Arnoldi decomposition

$$\boldsymbol{A}\boldsymbol{V}_m = \boldsymbol{V}_m\boldsymbol{H}_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^T = \boldsymbol{V}_{m+1}\widetilde{\boldsymbol{H}}_m, \tag{4.1}$$

where $\boldsymbol{e}_m$ denotes the mth unit vector in $\mathbb{C}^m$, $\boldsymbol{H}_m = (h_{i,j})_{i,j=1}^m \in \mathbb{C}^{m\times m}$ is an unreduced upper Hessenberg matrix and

$$\widetilde{\boldsymbol{H}}_m = \begin{bmatrix} & \boldsymbol{H}_m & \\ 0 & \cdots \; 0 & h_{m+1,m} \end{bmatrix} \in \mathbb{C}^{(m+1)\times m}\,.$$

A Hessenberg matrix is a special kind of matrix whose entries $h_{i,j}$ satisfy $h_{i,j} = 0$ for all $i > j+1$. We say that an upper Hessenberg matrix is unreduced, if the matrix has no zero subdiagonal entries. With the help of the orthogonality condition $\boldsymbol{V}_m^H \boldsymbol{V}_m = \boldsymbol{I}$, we obtain from (4.1) the compression

$$\boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m = \boldsymbol{H}_m \in \mathbb{C}^{m\times m}$$

of $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ onto $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ with respect to $\boldsymbol{V}_m$. If $\boldsymbol{A}$ is Hermitian or skew-Hermitian, the matrix $\boldsymbol{H}_m$ is Hermitian or skew-Hermitian as well, since

$$\begin{aligned} \boldsymbol{H}_m &= \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m = \boldsymbol{V}_m^H \boldsymbol{A}^H \boldsymbol{V}_m = \boldsymbol{H}_m^H && \text{for} \quad \text{Hermitian } \boldsymbol{A}\,, \\ \boldsymbol{H}_m &= \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m = -\boldsymbol{V}_m^H \boldsymbol{A}^H \boldsymbol{V}_m = -\boldsymbol{H}_m^H && \text{for} \quad \text{skew-Hermitian } \boldsymbol{A}\,. \end{aligned}$$

As a result, $\boldsymbol{H}_m$ has to be tridiagonal and the Arnoldi process reduces to a short three-term recurrence relation, which is then called the Hermitian or skew-Hermitian Lanczos algorithm, see [63] and the references therein.

Algorithm 4.4 Arnoldi process

```
given:  $\boldsymbol{A} \in \mathbb{C}^{N\times N}$,  $\boldsymbol{v} \in \mathbb{C}^N$
$\boldsymbol{v}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|$
for  $m = 1, 2, \dots$  do
    for  $j = 1, \dots, m$  do
        $h_{j,m} = \boldsymbol{v}_j^H \boldsymbol{A} \boldsymbol{v}_m$
    end for
    $\widetilde{\boldsymbol{v}}_{m+1} = \boldsymbol{A}\boldsymbol{v}_m - \sum_{j=1}^m h_{j,m} \boldsymbol{v}_j$
    $h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|$
    $\boldsymbol{v}_{m+1} = \widetilde{\boldsymbol{v}}_{m+1}/h_{m+1,m}$
end for
```

With these considerations, we can now define the Arnoldi approximation of order m to $f(\boldsymbol{A})\boldsymbol{v}$ as

$$f(\boldsymbol{A})\boldsymbol{v} \approx \boldsymbol{V}_m f(\boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m) \boldsymbol{V}_m^H \boldsymbol{v} = \|\boldsymbol{v}\| \boldsymbol{V}_m f(\boldsymbol{H}_m) \boldsymbol{e}_1\,. \tag{4.2}$$

A further motivation for the Arnoldi approximation, presented in [41] by Hochbruck and Lubich, is based on the Full Orthogonalization Method (FOM) in [71] applied to the linear system $(\xi \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}(\xi) = \boldsymbol{v}$, whose solution $\boldsymbol{x}$ depends on the shift ξ. The Full Orthogonalization Method constitutes an iterative technique for solving large linear systems. It determines an approximate solution $\boldsymbol{x}(\xi) \approx \boldsymbol{x}_m(\xi) = \boldsymbol{V}_m \boldsymbol{y}_m(\xi) \in \mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ such that the residual $\boldsymbol{r}_m(\xi) = \boldsymbol{v} - (\xi \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_m(\xi)$ is orthogonal to the Krylov subspace, which means

$$0 = \boldsymbol{V}_m^H \big(\boldsymbol{v} - (\xi \boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_m(\xi)\big) = \|\boldsymbol{v}\| \boldsymbol{e}_1 - (\xi \boldsymbol{I} - \boldsymbol{H}_m)\boldsymbol{y}_m(\xi)\,, \tag{4.3}$$

where we assume that $\xi \notin W(\boldsymbol{A})$. Since $\sigma(\boldsymbol{H}_m) \subseteq W(\boldsymbol{H}_m) \subseteq W(\boldsymbol{A})$ (see relation (4.6) below for $\boldsymbol{H}_m = \boldsymbol{S}_m$), the matrix $\xi \boldsymbol{I} - \boldsymbol{H}_m$ is invertible, and (4.3) is thus equivalent to $\boldsymbol{x}_m(\xi) = \|\boldsymbol{v}\| V_m(\xi \boldsymbol{I} - \boldsymbol{H}_m)^{-1}\boldsymbol{e}_1$. Hence, the Full Orthogonalization Method uses the approach

$$(\xi \boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v} = \boldsymbol{x}(\xi) \approx \boldsymbol{x}_m(\xi) = \|\boldsymbol{v}\| \boldsymbol{V}_m(\xi \boldsymbol{I} - \boldsymbol{H}_m)^{-1}\boldsymbol{e}_1 \,. \tag{4.4}$$

We assume that the function f is analytic in a neighborhood Ω of $W(\boldsymbol{A})$. The Arnoldi approximation to $f(\boldsymbol{A})\boldsymbol{v}$ is now obtained by replacing $(\xi \boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}$ in the Cauchy integral formula in Theorem 2.8 by the relation (4.4). This gives

$$f(\boldsymbol{A})\boldsymbol{v} = \frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{A}}\,\boldsymbol{v}\,d\xi \approx \|\boldsymbol{v}\|\boldsymbol{V}_m \frac{1}{2\pi i}\int_\Gamma \frac{f(\xi)}{\xi - \boldsymbol{H}_m}\,\boldsymbol{e}_1\,d\xi = \|\boldsymbol{v}\|\boldsymbol{V}_m f(\boldsymbol{H}_m)\boldsymbol{e}_1 \,,$$

where Γ is a closed contour in Ω which surrounds the field of values $W(\boldsymbol{A}) \supseteq W(\boldsymbol{H}_m)$.

Usually, (4.2) is a good approximation to $f(\boldsymbol{A})\boldsymbol{v}$ for $m \ll N$. The advantage of the Arnoldi approximation is that the problem has been reduced from dimension N to m, and we just have to evaluate a matrix function for the small matrix $\boldsymbol{H}_m$. This can be done by algorithms for dense matrices (see, e.g., Higham [38]).

For later purposes, especially in view of rational Krylov subspace methods, we are not only interested in a basis $\boldsymbol{V}_m$ of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ built by the Arnoldi process, but also in arbitrary orthonormal bases $\boldsymbol{V}_m$ of the Krylov subspace. In the following, we therefore generalize the Arnoldi approximation to any orthonormal basis of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$. For the sake of simplicity, we restrict ourselves to orthonormal bases with respect to the standard Euclidean inner product. However, all the following results remain valid for orthonormal bases with respect to an arbitrary inner product. The case of a not necessarily orthonormal basis and an arbitrary inner product, involving the Moore-Penrose inverse mentioned in the previous section, is described in Güttel's thesis [30]. In later applications, we will indicate, if a different inner product is used.

For an arbitrary orthonormal basis $\boldsymbol{V}_m$ of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$, the compression $\boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$ of the matrix $\boldsymbol{A}$ to the Krylov subspace does usually not coincide with the upper Hessenberg matrix $\boldsymbol{H}_m$ from the Arnoldi decomposition. This suggests the introduction of a new notation for the compression of $\boldsymbol{A}$ with respect to a general orthonormal basis given by $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$.

Definition 4.5 *Let $\boldsymbol{V}_m$ be an orthonormal basis of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ and let the function f be analytic on $W(\boldsymbol{A})$. The Krylov subspace approximation of order m to $f(\boldsymbol{A})\boldsymbol{v}$ is defined as*

$$f(\boldsymbol{A})\boldsymbol{v} \approx \boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v} \,, \tag{4.5}$$

where $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m \in \mathbb{C}^{m \times m}$ is the compression of $\boldsymbol{A}$ to $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$.

A justification of this definition follows from the fact that the Krylov subspace approximation is independent of the chosen orthonormal basis. To see this, we take two orthonormal bases $\boldsymbol{V}_m$ and $\boldsymbol{W}_m$ of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$. Then there exists a nonsingular matrix $\boldsymbol{U} \in \mathbb{C}^{m \times m}$ with $\boldsymbol{W}_m = \boldsymbol{V}_m \boldsymbol{U}$. Since

$$\boldsymbol{I} = \boldsymbol{W}_m^H \boldsymbol{W}_m = \boldsymbol{U}^H \boldsymbol{V}_m^H \boldsymbol{V}_m \boldsymbol{U} = \boldsymbol{U}^H \boldsymbol{U} \,,$$

the matrix $\boldsymbol{U}$ has to be unitary, i.e., $\boldsymbol{U}^{-1} = \boldsymbol{U}^H$, and with Theorem 2.7, part 3, we obtain the desired equality

$$\boldsymbol{W}_m f(\boldsymbol{W}_m^H \boldsymbol{A} \boldsymbol{W}_m)\boldsymbol{W}_m^H \boldsymbol{v} = \boldsymbol{V}_m \boldsymbol{U} f(\boldsymbol{U}^{-1}\boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m \boldsymbol{U})\boldsymbol{U}^{-1}\boldsymbol{V}_m^H \boldsymbol{v} = \boldsymbol{V}_m f(\boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m)\boldsymbol{V}_m^H \boldsymbol{v} \,.$$

The eigenvalues of the compression $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$ of $\boldsymbol{A}$ onto $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ are known as Ritz values of $\boldsymbol{A}$ (see [64]). They are also independent of the particular choice of the basis, since, for arbitrary orthonormal bases $\boldsymbol{V}_m$ and $\boldsymbol{W}_m$ of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$, the compressions $\boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$ and $\boldsymbol{W}_m^H \boldsymbol{A} \boldsymbol{W}_m$ are similar.

An important property of this compression is that the field of values of $\boldsymbol{S}_m$ is contained in the field of values of $\boldsymbol{A}$: For an arbitrary $\boldsymbol{x} \in \mathbb{C}^m$ with $\|\boldsymbol{x}\| = 1$, it holds

$$\boldsymbol{x}^H \boldsymbol{S}_m \boldsymbol{x} = \boldsymbol{x}^H \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m \boldsymbol{x} = \boldsymbol{y}^H \boldsymbol{A} \boldsymbol{y} \in W(\boldsymbol{A}) \,, \tag{4.6}$$

by setting $\boldsymbol{y} := \boldsymbol{V}_m \boldsymbol{x} \in \mathbb{C}^N$ with $\|\boldsymbol{y}\| = \|\boldsymbol{V}_m \boldsymbol{x}\| = \|\boldsymbol{x}\| = 1$, and so $W(\boldsymbol{S}_m) \subseteq W(\boldsymbol{A})$. This guarantees the existence of $f(\boldsymbol{S}_m)$ in Definition 4.5, if we assume that f is analytic on $W(\boldsymbol{A})$.

It is easy to check that $\boldsymbol{P}_m = \boldsymbol{V}_m \boldsymbol{V}_m^H$ is an orthogonal projector fulfilling

$$\mathrm{Range}(\boldsymbol{P}_m) = \mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v}) \,, \qquad \boldsymbol{P}_m^2 = \boldsymbol{P}_m \,, \qquad \boldsymbol{P}_m^H = \boldsymbol{P}_m \,.$$

The expression $\boldsymbol{A}_m = \boldsymbol{P}_m \boldsymbol{A} \boldsymbol{P}_m$ can thus be regarded as a restriction of $\boldsymbol{A}$ onto the Krylov subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ via orthogonal projection.

Lemma 4.6 *Let $\boldsymbol{V}_m$ be an orthonormal basis of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$. If f is analytic on $W(\boldsymbol{A})$ and, in addition, defined at zero, we find*

$$\boldsymbol{V}_m f(\boldsymbol{S}_m) \boldsymbol{V}_m^H \boldsymbol{v} = f(\boldsymbol{A}_m) \boldsymbol{v} \,,$$

where $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$, $\boldsymbol{A}_m = \boldsymbol{P}_m \boldsymbol{A} \boldsymbol{P}_m$, and $\boldsymbol{P}_m = \boldsymbol{V}_m \boldsymbol{V}_m^H$ is the orthogonal projector onto the Krylov subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$.

Proof. Because of $\boldsymbol{A}_m = \boldsymbol{V}_m \boldsymbol{S}_m \boldsymbol{V}_m^H$, we have $\sigma(\boldsymbol{A}_m) = \sigma(\boldsymbol{S}_m) \cup \{0\}$, where the eigenvalue zero occurs in the minimal polynomial $p_{\boldsymbol{A}_m}^{\min}$ with multiplicity one. Hence, the assumption on f ensures that $f(\boldsymbol{A}_m)$ is defined. Let now p be a polynomial that interpolates f at $\sigma(\boldsymbol{S}_m)$ and zero in the Hermite sense. Then the relation

$$\boldsymbol{V}_m f(\boldsymbol{S}_m) \boldsymbol{V}_m^H \boldsymbol{v} = \boldsymbol{V}_m p(\boldsymbol{S}_m) \boldsymbol{V}_m^H \boldsymbol{v} = p(\boldsymbol{A}_m) \boldsymbol{v} = f(\boldsymbol{A}_m) \boldsymbol{v}$$

yields the desired result. ❑

If the function f is analytic on $W(\boldsymbol{A})$, but not defined at zero, we still have the relation $\boldsymbol{V}_m f(\boldsymbol{S}_m) \boldsymbol{V}_m^H \boldsymbol{v} = p(\boldsymbol{A}_m) \boldsymbol{v}$, where p is the polynomial that interpolates f at $\sigma(\boldsymbol{S}_m)$ in the Hermite sense. Later, we will consider the φ-functions which are holomorphic on the whole complex plane and thus, in particular, defined at zero. For functions of this type, we can always write the Krylov subspace approximation of $\varphi(\boldsymbol{A})\boldsymbol{v}$ as $\varphi(\boldsymbol{A}_m)\boldsymbol{v}$ instead of $\boldsymbol{V}_m \varphi(\boldsymbol{S}_m) \boldsymbol{V}_m^H \boldsymbol{v}$. On the one hand, the notation $\boldsymbol{V}_m \varphi(\boldsymbol{S}_m) \boldsymbol{V}_m^H \boldsymbol{v}$ is more advantageous in view of computational issues and, on the other hand, the alternative representation $\varphi(\boldsymbol{A}_m)\boldsymbol{v}$ provides a shorter notation for the Krylov subspace approximation.

If f is a polynomial of degree less than or equal to $m-1$, the Krylov approximation (4.5) even yields the exact result (e.g., Saad [72], Lemma 3.1).

Lemma 4.7 *Let $\boldsymbol{V}_m$ be an orthonormal basis of $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ and set $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$ and $\boldsymbol{A}_m = \boldsymbol{P}_m \boldsymbol{A} \boldsymbol{P}_m$ for $\boldsymbol{P}_m = \boldsymbol{V}_m \boldsymbol{V}_m^H$. Moreover, let $p \in \mathcal{P}_{m-1}$ be arbitrary. Then we have*

$$p(\boldsymbol{A})\boldsymbol{v} = \boldsymbol{V}_m p(\boldsymbol{S}_m) \boldsymbol{V}_m^H \boldsymbol{v} = p(\boldsymbol{A}_m)\boldsymbol{v}.$$

Proof. Since p is a polynomial of maximal degree $m-1$, it is sufficient to show the claim for the monomials $p_0(z)=1$, $p_1(z)=z$, …, $p_{m-1}(z)=z^{m-1}$, that means

$$\boldsymbol{A}^j\boldsymbol{v}=\boldsymbol{V}_m\boldsymbol{S}_m^j\boldsymbol{V}_m^H\boldsymbol{v}=\boldsymbol{A}_m^j\boldsymbol{v} \quad \text{for} \quad 0\le j\le m-1\,.$$

The statement is clearly true for $j=0$, since

$$\boldsymbol{A}^0\boldsymbol{v}=\boldsymbol{v}=\boldsymbol{P}_m\boldsymbol{v}=\boldsymbol{V}_m\boldsymbol{V}_m^H\boldsymbol{v}=\boldsymbol{V}_m\boldsymbol{S}_m^0\,\boldsymbol{V}_m^H\boldsymbol{v}=\boldsymbol{A}_m^0\boldsymbol{v}\,.$$

So, we assume that the assertion holds true for some k with $0\le k\le m-2$. Because of $\boldsymbol{A}^{k+1}\boldsymbol{v}\in\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$ for $0\le k\le m-2$, we obtain by induction

$$\begin{aligned}\boldsymbol{A}^{k+1}\boldsymbol{v}&=\boldsymbol{P}_m\boldsymbol{A}\boldsymbol{A}^k\boldsymbol{v}=(\boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m)\boldsymbol{A}^k\boldsymbol{v}\\&=(\boldsymbol{V}_m\boldsymbol{S}_m\boldsymbol{V}_m^H)\boldsymbol{V}_m\boldsymbol{S}_m^k\,\boldsymbol{V}_m^H\boldsymbol{v}=\boldsymbol{V}_m\boldsymbol{S}_m^{k+1}\boldsymbol{V}_m^H\boldsymbol{v}=\boldsymbol{A}_m^{k+1}\,\boldsymbol{v}\,,\end{aligned}$$

which proves the result. ❑

Not only for polynomials of degree less than the dimension of the Krylov subspace, but also if we perform the Krylov iteration until the invariance index μ is reached, the Krylov approximation (4.5) becomes exact. In this case, the Krylov subspace $\mathcal{K}_\mu(\boldsymbol{A},\boldsymbol{v})$ is invariant under multiplication with $\boldsymbol{A}$. This means that there exists a matrix $\boldsymbol{T}\in\mathbb{C}^{m\times m}$ with $\boldsymbol{A}\boldsymbol{V}_\mu=\boldsymbol{V}_\mu\boldsymbol{T}$ which is equal to $\boldsymbol{S}_\mu$, due to

$$\boldsymbol{S}_\mu=\boldsymbol{V}_\mu^H\boldsymbol{A}\boldsymbol{V}_\mu=\boldsymbol{V}_\mu^H\boldsymbol{V}_\mu\boldsymbol{T}=\boldsymbol{T}\,.$$

Using this relation, it follows for $\xi\notin W(\boldsymbol{A})\supseteq W(\boldsymbol{S}_\mu)$ that

$$\xi\boldsymbol{V}_\mu-\boldsymbol{A}\boldsymbol{V}_\mu=\xi\boldsymbol{V}_\mu-\boldsymbol{V}_\mu\boldsymbol{S}_\mu \quad\Longleftrightarrow\quad \frac{1}{\xi-\boldsymbol{A}}\boldsymbol{V}_\mu\boldsymbol{V}_\mu^H=\boldsymbol{V}_\mu\frac{1}{\xi-\boldsymbol{S}_\mu}\boldsymbol{V}_\mu^H$$

and hence

$$\frac{1}{\xi-\boldsymbol{A}}\boldsymbol{V}_\mu\boldsymbol{V}_\mu^H\boldsymbol{v}=\frac{1}{\xi-\boldsymbol{A}}\boldsymbol{v}=\boldsymbol{V}_\mu\frac{1}{\xi-\boldsymbol{S}_\mu}\boldsymbol{V}_\mu^H\boldsymbol{v}\,.$$

The Cauchy integral formula then implies

$$f(\boldsymbol{A})\boldsymbol{v}=\frac{1}{2\pi i}\int_\Gamma\frac{f(\xi)}{\xi-\boldsymbol{A}}\boldsymbol{v}\,d\xi=\frac{1}{2\pi i}\int_\Gamma\boldsymbol{V}_\mu\frac{f(\xi)}{\xi-\boldsymbol{S}_\mu}\boldsymbol{V}_\mu^H\boldsymbol{v}\,d\xi=\boldsymbol{V}_\mu f(\boldsymbol{S}_\mu)\boldsymbol{V}_\mu^H\boldsymbol{v}\,,$$

where Γ is a simple closed rectifiable curve enclosing $W(\boldsymbol{A})$. The Krylov subspace approximation in $\mathcal{K}_\mu(\boldsymbol{A},\boldsymbol{v})$ thus yields the exact result.

In general, we do not iterate until the invariance index μ is reached, we rather compute a polynomial approximation of $f(\boldsymbol{A})\boldsymbol{v}$ in $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$: By Theorem 2.5, $f(\boldsymbol{S}_m)$ can be represented as $p(\boldsymbol{S}_m)$, where p is a polynomial of degree less than or equal to $m-1$. This polynomial p interpolates f at the Ritz values of $\boldsymbol{A}$, that is, in the eigenvalues of $\boldsymbol{S}_m$ according to their multiplicity in the minimal polynomial of $\boldsymbol{S}_m$. With the help of Lemma 4.7, we conclude that

$$\boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{v}=\boldsymbol{V}_m p(\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{v}=p(\boldsymbol{A})\boldsymbol{v}\;\in\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})\,, \tag{4.7}$$

where $p\in\mathcal{P}_{m-1}$ is the polynomial that interpolates f at $\sigma(\boldsymbol{S}_m)$. This equality points out that the Krylov subspace approximation can be interpreted as a polynomial interpolation, where the nodes are the Ritz values of $\boldsymbol{A}$. The approximation quality depends on how well

$p(\boldsymbol{A})\boldsymbol{v} \in \mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ approximates $f(\boldsymbol{A})\boldsymbol{v}$, where p is the interpolation polynomial defined in (4.7).

For the Arnoldi approximation, in the special case that $f(z) = e^{\tau z}$ for $\tau > 0$, $\boldsymbol{A}$ is Hermitian negative semi-definite with eigenvalues in the interval $[-4\rho, 0]$, $\rho > 0$, and $\boldsymbol{v}$ is a vector with $\|\boldsymbol{v}\| = 1$, Theorem 2 in Hochbruck and Lubich [41] states the error bound

$$\|e^{\tau \boldsymbol{A}}\boldsymbol{v} - \boldsymbol{V}_m e^{\tau \boldsymbol{H}_m}\boldsymbol{e}_1\| \leq \begin{cases} 10\, e^{-\frac{m^2}{5\rho\tau}}\,, & \sqrt{4\rho\tau} \leq m \leq 2\rho\tau\,, \\ 10\, \dfrac{e^{-\rho\tau}}{\rho\tau} \left(\dfrac{e\rho\tau}{m}\right)^m\,, & m \geq 2\rho\tau\,, \end{cases} \tag{4.8}$$

which yields a superlinear convergence after $m \geq \sqrt{\|\tau \boldsymbol{A}\|}$ iteration steps. Furthermore, for skew-Hermitian matrices $\boldsymbol{A}$ with eigenvalues in an interval on the imaginary axis of length 4ρ, Theorem 4 in [41] yields

$$\|e^{\tau \boldsymbol{A}}\boldsymbol{v} - \boldsymbol{V}_m e^{\tau \boldsymbol{H}_m}\boldsymbol{e}_1\| \leq 12\, e^{-\frac{(\rho\tau)^2}{m}} \left(\frac{e\rho\tau}{m}\right)^m\,, \qquad m \geq 2\rho\tau\,,$$

which only leads to a substantial error reduction for $m \gg \rho\tau$.

These results show that the convergence of the Arnoldi method might set in very late, if the norm of $\boldsymbol{A}$ is large. We either must accept that very many Krylov iterations have to be performed or that a small time step size τ has to be chosen such that $\|\tau \boldsymbol{A}\|$ is of moderate size.

Later on, we will see that this restriction does not exist for rational Krylov subspace methods. For rational Krylov subspace methods, it is possible to obtain error bounds that are independent of $\|\boldsymbol{A}\|$. Given a large matrix $\boldsymbol{A}$, resulting from a discretization of a partial differential equation, rational Krylov methods have the favorable and very useful property that the convergence rate does not depend on the refinement of the space grid, whereas the convergence of the standard Krylov subspace method starts the later the finer the spatial domain is discretized.

4.3 Rational Krylov subspace

In contrast to the standard Krylov subspace, which contains powers of the matrix $\boldsymbol{A}$ times a vector $\boldsymbol{v}$, we now consider a different Krylov subspace based on rational matrix functions with $m-1$ poles distinct from $\sigma(\boldsymbol{A})$. The rational Krylov subspace method was developed by Ruhe, cf. [69], in the context of eigenvalue computations. It is a more general version of the spectral transformation Lanczos method [20], where the Lanczos algorithm is applied to the shifted and inverted matrix $(\gamma \boldsymbol{I} - \boldsymbol{A})^{-1}$ to approximate eigenvalues close to $\gamma \notin \sigma(\boldsymbol{A})$.

Definition 4.8 *Assume that the polynomial $q_{m-1} \in \mathcal{P}_{m-1}$ has no roots in $\sigma(\boldsymbol{A})$. Then the rational Krylov subspace of order $m \geq 1$ is defined by*

$$\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v}) := q_{m-1}(\boldsymbol{A})^{-1}\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v}) = \left\{ r(\boldsymbol{A})\boldsymbol{v} \,:\, r \in \frac{\mathcal{P}_{m-1}}{q_{m-1}} \right\}.$$

The assumption on the prescribed denominator polynomial q_{m-1} in Definition 4.8 ensures that the inverse $q_{m-1}(\boldsymbol{A})^{-1}$ exists. Using the fact that every matrix function $f(\boldsymbol{A})$ commutes with $\boldsymbol{A}$, cf. Theorem 2.7, we have $q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{A}^j\boldsymbol{v} = \boldsymbol{A}^j q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v}$ for all $j \geq 0$, so that

$$\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_m(\boldsymbol{A}, q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v})\,.$$

This natural link between the rational and the polynomial Krylov subspace, spanned by $\boldsymbol{A}$ and the modified initial vector $q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v}$, allows to transfer results from the standard to the rational Krylov subspace in a simple way.

Since $q_{m-1}(\boldsymbol{A})^{-1}$ has full rank, we can conclude

$$\dim\bigl(\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})\bigr) = \dim\bigl(q_{m-1}(\boldsymbol{A})^{-1}\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})\bigr) = \dim\bigl(\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})\bigr)\,.$$

Therefore, $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ has the same invariance index μ as $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{v})$. Analogously to the case of the polynomial Krylov subspace, we have $f(\boldsymbol{A})\boldsymbol{v} \in \mathcal{Q}_\mu(\boldsymbol{A},\boldsymbol{v})$ and nested spaces

$$\mathcal{Q}_1(\boldsymbol{A},\boldsymbol{v}) \subsetneq \mathcal{Q}_2(\boldsymbol{A},\boldsymbol{v}) \subsetneq \ldots \subsetneq \mathcal{Q}_\mu(\boldsymbol{A},\boldsymbol{v}) = \mathcal{Q}_{\mu+1}(\boldsymbol{A},\boldsymbol{v}) = \ldots\,,$$

if the denominator polynomials q_{m-1} and q_m of all two consecutive spaces differ only by a linear factor. Invariance in the rational case means that we can multiply every element $\boldsymbol{w}$ of $\mathcal{Q}_\mu(\boldsymbol{A},\boldsymbol{v})$ with an arbitrary matrix function $f(\boldsymbol{A})$ and the obtained product $f(\boldsymbol{A})\boldsymbol{w}$ always stays in $\mathcal{Q}_\mu(\boldsymbol{A},\boldsymbol{v})$, provided that the function f is defined on $\sigma(\boldsymbol{A})$.

The rational Krylov subspace $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ possesses an alternative representation, which is often useful. This is the subject of the following lemma.

Lemma 4.9 *If the denominator polynomial $q_{m-1} \in \mathcal{P}_{m-1}$ of the rational Krylov subspace $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ is of the form $q_{m-1}(z) = (z_1 - z)^{n_1}\cdots(z_s - z)^{n_s}$ with $z_k \notin \sigma(\boldsymbol{A})$, $k = 1,\ldots,s$, and $\sum_{k=1}^{s} n_k = m-1$, then $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ can be written as*

$$\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v}) = \left\{ r(\boldsymbol{A})\boldsymbol{v} \,:\, r \in \frac{\mathcal{P}_{m-1}}{q_{m-1}} \right\} = \left\{ a_0\boldsymbol{v} + \sum_{k=1}^{s}\sum_{j=1}^{n_s} \frac{a_{k,j}}{(z_k - \boldsymbol{A})^j}\,\boldsymbol{v} \,:\, a_0, a_{k,j} \in \mathbb{C} \right\}$$

$$= \operatorname{span}\left\{ \boldsymbol{v}\,, \frac{1}{(z_k - \boldsymbol{A})^j}\,\boldsymbol{v} \,:\, 1 \leq j \leq n_k\,, 1 \leq k \leq s \right\}.$$

Proof. Since the last equality is clear, it remains to show the second equality. To this end, we use partial fraction expansion. For every polynomial $p_{m-1} \in \mathcal{P}_{m-1}$, there are unique coefficients $a_0, a_{k,j} \in \mathbb{C}$ such that

$$\frac{p_{m-1}(z)}{q_{m-1}(z)} = a_0 + \sum_{k=1}^{s}\sum_{j=1}^{n_s} \frac{a_{k,j}}{(z_k - z)^j} \quad \text{for all} \quad z \neq z_k\,, \quad k = 1,\ldots,s\,.$$

Replacing $z \in \mathbb{C}$ by the matrix $\boldsymbol{A}$ and multiplying with the vector $\boldsymbol{v}$ from the right, we obtain the inclusion "$\subseteq$". The missing inclusion "$\supseteq$" is obtained by reducing the expression $h(z) := a_0 + \sum_{k=1}^{s}\sum_{j=1}^{n_s} \frac{a_{k,j}}{(z_k - z)^j}$ to the common denominator. This shows that $h \in \mathcal{P}_{m-1}/q_{m-1}$. ❑

In general, the prescribed denominator polynomial q_{m-1} of the rational Krylov subspace $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ is of the form

$$q_{m-1}(z) = \prod_{k=1}^{m-1}(z_k - z)$$

with a sequence $\{z_k\}$ of given poles, where the poles $z_k \in \mathbb{C} \backslash \sigma(\boldsymbol{A})$ do not have to be distinct.

Whereas there is only one standard Krylov subspace method because of the shift invariance, there are various rational Krylov subspace methods. This is due to the fact that we obtain a different method for any different choice of the poles $z_1, \ldots, z_{m-1}$. If all z_j are equal to a fixed shift $\gamma \in \mathbb{C}$, we obtain the shift-and-invert Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v}) = \mathcal{K}_m\big((\gamma\boldsymbol{I} - \boldsymbol{A})^{-1}, \boldsymbol{v}\big)$ (van den Eshof and Hochbruck [84]), also called restricted-denominator rational Krylov subspace (Moret and Novati [59,60,62]), or resolvent Krylov subspace in the case of operators (Grimm [29]). It is also possible to select the poles, for example, on a straight line, a parabola, or a hyperbola in the complex plane lying outside $W(\boldsymbol{A})$. In addition, we can combine the polynomial and rational Krylov subspace to an extended Krylov subspace

$$\mathcal{K}_{k,m}(\boldsymbol{A}, \boldsymbol{v}) := \operatorname{span}\{\boldsymbol{A}^{-k+1}\boldsymbol{v}, \ldots, \boldsymbol{A}^{-1}\boldsymbol{v}, \boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}, \ldots, \boldsymbol{A}^{m-1}\boldsymbol{v}\}\,, \qquad k \geq 1\,,\ m \geq 1\,,$$

introduced by Druskin and Knizhnerman in [15], provided that $\boldsymbol{A}$ is invertible. Of course, it is also conceivable to choose poles different from zero for the rational part of the extended subspace, see Chapter 6 below.

The rational Arnoldi decomposition presented in Algorithm 4.10 computes an orthonormal basis $\boldsymbol{V}_m = [\boldsymbol{v}_1\, \boldsymbol{v}_2\, \cdots\, \boldsymbol{v}_m]$ of $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$. This algorithm looks quite similar to the standard Arnoldi Algorithm 4.4, but the matrix $\boldsymbol{A}$ is replaced by the shifted inverse $(z_m\boldsymbol{I} - \boldsymbol{A})^{-1}$.

Algorithm 4.10 Rational Arnoldi process

```
given: $\boldsymbol{A} \in \mathbb{C}^{N \times N}$, $\boldsymbol{v} \in \mathbb{C}^N$, poles $z_1, z_2, \ldots \notin \sigma(\boldsymbol{A})$
$\boldsymbol{v}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|$
for $m = 1, 2, \ldots$ do
    for $j = 1, \ldots, m$ do
        $h_{j,m} = \boldsymbol{v}_j^H (z_m\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}_m$
    end for
    $\widetilde{\boldsymbol{v}}_{m+1} = (z_m\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}_m - \sum_{j=1}^{m} h_{j,m}\boldsymbol{v}_j$
    $h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|$
    $\boldsymbol{v}_{m+1} = \widetilde{\boldsymbol{v}}_{m+1}/h_{m+1,m}$
end for
```

The new basis vector $\boldsymbol{v}_{m+1}$ is obtained by orthogonalizing $(z_m\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}_m$ against the already known vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ and by scaling the received vector $\widetilde{\boldsymbol{v}}_{m+1}$ with the reciprocal of its norm. More precisely, we have

$$\widetilde{\boldsymbol{v}}_{m+1} = (z_m\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}_m - \sum_{j=1}^{m} h_{j,m}\boldsymbol{v}_j\,, \qquad h_{j,m} = \boldsymbol{v}_j^H (z_m\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}_m \tag{4.9}$$

and $\widetilde{\boldsymbol{v}}_{m+1} = h_{m+1,m}\boldsymbol{v}_{m+1}$ with $h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|$.

In contrast to Algorithm 4.10, the rational Arnoldi algorithm originally considered by Ruhe in [69,70] uses

$$\begin{aligned}
\boldsymbol{w} &= \boldsymbol{V}_m \boldsymbol{y}_m\,,\\
h_{j,m} &= \boldsymbol{v}_j^H \left(\boldsymbol{I} - \frac{1}{z_m}\boldsymbol{A}\right)^{-1} \boldsymbol{A}\boldsymbol{w}\,, \quad j = 1,\dots,m\\
\widetilde{\boldsymbol{v}}_{m+1} &= \left(\boldsymbol{I} - \frac{1}{z_m}\boldsymbol{A}\right)^{-1} \boldsymbol{A}\boldsymbol{w} - \sum_{j=1}^{m} h_{j,m}\boldsymbol{v}_j\,,
\end{aligned}$$

where the vector $\boldsymbol{y}_m$ should be chosen such that $(\boldsymbol{I} - \frac{1}{z_m}\boldsymbol{A})^{-1}\boldsymbol{A}\boldsymbol{w} \in \mathcal{Q}_{m+1}(\boldsymbol{A},\boldsymbol{v})\backslash\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$.

Theoretically, it could happen that the rational Arnoldi process 4.10 yields $\boldsymbol{v}_{m+1} = \boldsymbol{0}$ and thus breaks down, although the invariance index μ has not yet been reached. This is the case, if the previously computed vector is of the form $\boldsymbol{v}_m = p_{m-1}(\boldsymbol{A})q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v}$ and the new pole z_m is a root of the polynomial $p_{m-1} \in \mathcal{P}_{m-1}$. Such a breakdown might also appear in the rational Krylov decomposition by Ruhe, if $\boldsymbol{y}_m$ has not been suitably selected. However, in numerical experiments we never have observed such an "unlucky breakdown".

To obtain a formula similar to the standard Arnoldi decomposition (4.1), we multiply (4.9) by $z_m\boldsymbol{I} - \boldsymbol{A}$ and reorder the terms to obtain

$$\boldsymbol{A}\sum_{j=1}^{m+1} h_{j,m}\boldsymbol{v}_j = z_m \sum_{j=1}^{m+1} h_{j,m}\boldsymbol{v}_j - \boldsymbol{v}_m\,.$$

In matrix notation, this relation can also be written as

$$\boldsymbol{A}\boldsymbol{V}_m\boldsymbol{H}_m + h_{m+1,m}\boldsymbol{A}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^T = \boldsymbol{V}_m(\boldsymbol{H}_m\boldsymbol{D}_m - \boldsymbol{I}) + z_m h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^T\,,$$

where $\boldsymbol{H}_m = (h_{i,j})_{i,j=1}^m \in \mathbb{C}^{m\times m}$ and $\boldsymbol{D}_m = \operatorname{diag}(z_1,\dots,z_m)$.

For the shift-and-invert Krylov subspace $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_m\big((\gamma\boldsymbol{I} - \boldsymbol{A})^{-1},\boldsymbol{v}\big)$ with $z_j = \gamma$ for $j = 1,\dots,m-1$, we have the decomposition

$$(\boldsymbol{A} - \gamma\boldsymbol{I})\boldsymbol{V}_m\boldsymbol{H}_m = -\boldsymbol{V}_m + (\gamma\boldsymbol{I} - \boldsymbol{A})\,h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^T \tag{4.10}$$

or

$$(\gamma\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{V}_m = \boldsymbol{V}_m\boldsymbol{H}_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^T\,. \tag{4.11}$$

This is just the standard Arnoldi decomposition, where $\boldsymbol{A}$ is replaced by $(\gamma\boldsymbol{I} - \boldsymbol{A})^{-1}$. Consequently, the rational process with a fixed pole γ only breaks down with $h_{m+1,m} = 0$, if the invariance index μ is reached, as is the case for the standard Arnoldi algorithm (e.g., Saad [73], Proposition 6.6). This is called a "lucky breakdown", since then $f(\boldsymbol{A})\boldsymbol{v}$ belongs to the computed subspace $\mathcal{Q}_\mu(\boldsymbol{A},\boldsymbol{v})$. For the shift-and-invert Krylov subspace, a rearrangement of (4.10) yields the compression

$$\widehat{\boldsymbol{H}}_m := \boldsymbol{V}_m^H\boldsymbol{A}\boldsymbol{V}_m = \gamma\boldsymbol{I} - \boldsymbol{H}_m^{-1} - h_{m+1,m}\boldsymbol{V}_m^H\boldsymbol{A}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^T\boldsymbol{H}_m^{-1}\,, \tag{4.12}$$

whereas for the general case, we find

$$\widehat{\boldsymbol{H}}_m := \boldsymbol{V}_m^H\boldsymbol{A}\boldsymbol{V}_m = (\boldsymbol{H}_m\boldsymbol{D}_m - \boldsymbol{I})\boldsymbol{H}_m^{-1} - h_{m+1,m}\boldsymbol{V}_m^H\boldsymbol{A}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^T\boldsymbol{H}_m^{-1}\,.$$

By analogy with the standard Arnoldi approximation (4.2) above, the rational Arnoldi approximation reads

$$f(\boldsymbol{A})\boldsymbol{v} \approx \boldsymbol{V}_m f(\boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m)\boldsymbol{V}_m^H \boldsymbol{v} = \|\boldsymbol{v}\| \boldsymbol{V}_m f(\widehat{\boldsymbol{H}}_m)\boldsymbol{e}_1 \,. \tag{4.13}$$

A slightly different approximation for symmetric and negative semi-definite matrices and the function $f(z) = e^z$, based on the shift-and-invert Krylov subspace, is discussed by Hochbruck and van den Eshof in [84]. The authors consider the transformed function $g_\gamma(t) = e^{\gamma - t^{-1}}$ such that $g_\gamma((\gamma \boldsymbol{I} - \boldsymbol{A})^{-1}) = e^{\boldsymbol{A}}$ and therefore suggest the approximation

$$e^{\boldsymbol{A}}\boldsymbol{v} \approx \|\boldsymbol{v}\| \boldsymbol{V}_m g_\gamma(\boldsymbol{H}_m)\boldsymbol{e}_1 = \|\boldsymbol{v}\| \boldsymbol{V}_m e^{\gamma \boldsymbol{I} - \boldsymbol{H}_m^{-1}} \boldsymbol{e}_1 \,,$$

where $\boldsymbol{H}_m = \boldsymbol{V}_m^H(\gamma I - \boldsymbol{A})^{-1}\boldsymbol{V}_m$ is the matrix in the rational Arnoldi decomposition (4.11). This approximation is motivated by the fact that $e^{\boldsymbol{A}}\boldsymbol{v}$ is mainly determined by the eigenvalues of $\boldsymbol{A}$ with smallest modulus. If we apply the Lanczos process to the shifted and inverted matrix $(\gamma \boldsymbol{I} - \boldsymbol{A})^{-1}$, these important eigenvalues are detected faster. Because of relation (4.12), the matrix $\widehat{\boldsymbol{H}}_m$ can be seen as a rank-1 modification of $\gamma \boldsymbol{I} - \boldsymbol{H}_m^{-1}$. This means that the approximation in [84] does not coincide with the rational Arnoldi approximation defined in (4.13), but yields a similar approximation.

Reviewing our results, one can note that depending on how we compute an orthonormal basis $\boldsymbol{V}_m$ of $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v}) = \mathcal{K}_m(\boldsymbol{A}, q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v})$, the matrix $\boldsymbol{V}_m \in \mathbb{C}^{N \times m}$ has a different form. On the one hand, we might determine $\boldsymbol{V}_m$ by the standard Arnoldi Algorithm 4.4 using the matrix $\boldsymbol{A}$ and the vector $q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v}$ and, on the other hand, a computation via the rational Arnoldi Algorithm 4.10 would be possible. Both procedures lead to a different basis of the same subspace. This is why all the following statements will be formulated for a general orthonormal basis of $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$, which not necessarily coincides with the matrix $\boldsymbol{V}_m$ obtained by the rational Arnoldi process.

Definition 4.11 *Let $\boldsymbol{V}_m$ be an orthonormal basis of the rational Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$ and assume that the denominator $q_{m-1} \in \mathcal{P}_{m-1}$ has no roots in $W(\boldsymbol{A})$. For a function f analytic on $W(\boldsymbol{A})$, the rational Krylov subspace approximation reads*

$$f(\boldsymbol{A})\boldsymbol{v} \approx \boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v} \,, \qquad \boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m \,.$$

As for the standard Krylov subspace approximation, this definition is independent of the particular choice of the orthonormal basis $\boldsymbol{V}_m$ of $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$. Applying the rational Arnoldi process for the computation of $\boldsymbol{V}_m$, we have $\boldsymbol{S}_m = \widehat{\boldsymbol{H}}_m$. Analogously to the Ritz values above, we refer to the eigenvalues of the compression $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$ of $\boldsymbol{A}$ to $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$ as rational Ritz values of $\boldsymbol{A}$.

Defining $\boldsymbol{A}_m = \boldsymbol{P}_m \boldsymbol{A} \boldsymbol{P}_m$, where $\boldsymbol{P}_m = \boldsymbol{V}_m \boldsymbol{V}_m^H$ is the orthogonal projection onto the rational Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$, similar to Lemma 4.6, we find again that

$$\boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v} = f(\boldsymbol{A}_m)\boldsymbol{v} \,,$$

if the considered function f is, additionally, defined at zero.

The next theorem makes a statement about the exactness of the rational Krylov subspace approximation for all rational functions in $\mathcal{P}_{m-1}/q_{m-1}$. According to Lemma 4.7, this property is already known for the standard Krylov subspace and can directly be transferred to rational Krylov subspace methods (see Beckermann and Reichel [4], p. 21).

Lemma 4.12 *Let $\boldsymbol{V}_m$ be an orthonormal basis of $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ and $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A}\boldsymbol{V}_m$. For any rational function $r \in \mathcal{P}_{m-1}/q_{m-1}$, the rational Krylov subspace approximation is exact, that means*

$$r(\boldsymbol{A})\boldsymbol{v} = \boldsymbol{V}_m r(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v}\,,$$

provided that $r(\boldsymbol{S}_m)$ is defined.

Proof. Let $\boldsymbol{w} = q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v}$ and $\boldsymbol{V}_m$ be a basis of $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_m(\boldsymbol{A},\boldsymbol{w})$. For an arbitrary polynomial $p_{m-1} \in \mathcal{P}_{m-1}$, Lemma 4.7 gives

$$p_{m-1}(\boldsymbol{A})\boldsymbol{w} = \boldsymbol{V}_m p_{m-1}(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{w}\,.$$

Furthermore, we have

$$\boldsymbol{v} = q_{m-1}(\boldsymbol{A})\boldsymbol{w} = \boldsymbol{V}_m q_{m-1}(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{w} \quad \Longleftrightarrow \quad q_{m-1}(\boldsymbol{S}_m)^{-1}\boldsymbol{V}_m^H\boldsymbol{v} = \boldsymbol{V}_m^H \boldsymbol{w}$$

and therefore

$$\begin{aligned} r(\boldsymbol{A})\boldsymbol{v} &= p_{m-1}(\boldsymbol{A})q_{m-1}(\boldsymbol{A})^{-1}\boldsymbol{v} = p_{m-1}(\boldsymbol{A})\boldsymbol{w} = \boldsymbol{V}_m p_{m-1}(\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{w} \\ &= \boldsymbol{V}_m p_{m-1}(\boldsymbol{S}_m)q_{m-1}(\boldsymbol{S}_m)^{-1}\boldsymbol{V}_m^H\boldsymbol{v} = \boldsymbol{V}_m r(\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{v}\,, \end{aligned}$$

which concludes the proof. ❑

Lemma 4.12 is now exploited to show the following interpolation result (see, e.g., Güttel [30], Theorem 4.8), which is the analogue to relation (4.7) above in the polynomial case.

Lemma 4.13 *Assume that $f(\boldsymbol{S}_m)$ is defined. On the condition that the roots of the denominator polynomial q_{m-1} do not coincide with the eigenvalues of $\boldsymbol{S}_m$, we have*

$$\boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{v} = r(\boldsymbol{A})\boldsymbol{v}\,, \qquad \boldsymbol{S}_m = \boldsymbol{V}_m^H\boldsymbol{A}\boldsymbol{V}_m\,,$$

where the rational function $r \in \mathcal{P}_{m-1}/q_{m-1}$ interpolates f at the rational Ritz values of the matrix $\boldsymbol{A}$.

Proof. We define $g := q_{m-1}f$. By Theorem 2.5, we have $g(\boldsymbol{S}_m) = p_{m-1}(\boldsymbol{S}_m)$ for a polynomial $p_{m-1} \in \mathcal{P}_{m-1}$, where p_{m-1} interpolates g in the Hermite sense at the eigenvalues of the matrix $\boldsymbol{S}_m$, that is, at the rational Ritz values of $\boldsymbol{A}$. From this, we can conclude that $r := p_{m-1}/q_{m-1}$ interpolates the function f at $\sigma(\boldsymbol{S}_m)$ in the Hermite sense. For this reason, we obtain $f(\boldsymbol{S}_m) = p_{m-1}(\boldsymbol{S}_m)/q_{m-1}(\boldsymbol{S}_m)$ and thus

$$\boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{v} = \boldsymbol{V}_m\,\frac{p_{m-1}(\boldsymbol{S}_m)}{q_{m-1}(\boldsymbol{S}_m)}\,\boldsymbol{V}_m^H\boldsymbol{v} = \boldsymbol{V}_m r(\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{v} = r(\boldsymbol{A})\boldsymbol{v}$$

with the help of Lemma 4.12. ❑

In a similar way, other important properties of the polynomial Krylov subspace process can be transferred to the rational method. These properties include, for example, the exactness property $f(\boldsymbol{A})\boldsymbol{v} = \boldsymbol{V}_\mu f(\boldsymbol{S}_\mu)\boldsymbol{V}_\mu^H\boldsymbol{v}$, if the invariance index μ is reached.

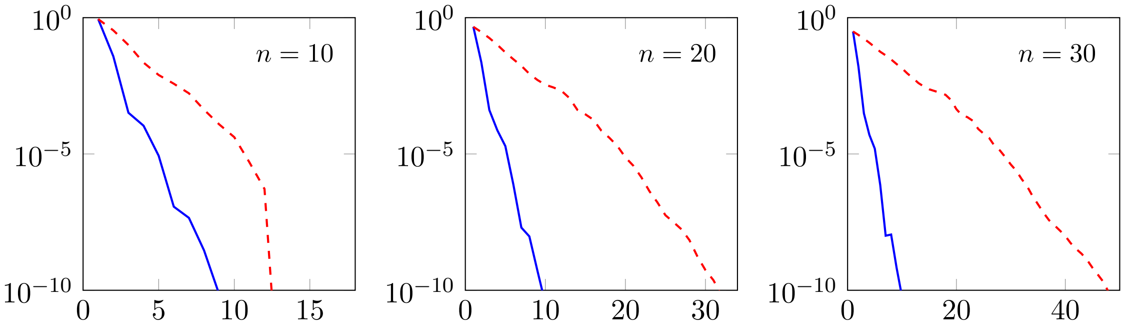


Figure 4.3: Plot of error versus iteration steps of the standard (red dashed line) and the rational (blue solid line) Krylov subspace method for the two-dimensional heat equation and $n = 10, 20, 30$.

Example 4.14 In order to demonstrate the advantages of the rational over the standard Krylov subspace method, we consider the discretization of the two-dimensional heat equation on $\Omega = (0,1)^2$ with homogeneous Dirichlet boundary conditions for the initial function

$$u(0,x,y) = u_0(x,y) = 100 \cdot \frac{x^2(1-x)^2y^2(1-y)^2}{\|x^2(1-x)^2y^2(1-y)^2\|_{L^2(\Omega)}}\,.$$

In Section 3.1.1, we have seen that the finite-difference discretization leads to the system of ordinary differential equations $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)$, $\boldsymbol{u}(0) = \boldsymbol{u}_0$, with solution $\boldsymbol{u}(t) = e^{t\boldsymbol{A}}\boldsymbol{u}_0$. We choose n inner grid points in each space direction and set $N = n^2$. Actually, we should write $\boldsymbol{A}_h$ instead of $\boldsymbol{A}$ to indicate that the discretization matrix depends on the mesh size $h = \frac{1}{n+1}$. This is the case in all our numerical experiments, but we will always omit the subscript h. In this example, the discretization matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ has the field of values

$$W(\boldsymbol{A}) = \left[-\frac{8}{h^2}\sin^2\left(\frac{n\pi}{2(n+1)}\right), -\frac{8}{h^2}\sin^2\left(\frac{\pi}{2(n+1)}\right)\right]$$

on the negative real line with the limit

$$W(\boldsymbol{A}) \to (-\infty, -2\pi^2] \quad \text{for} \quad h \to 0\,.$$

We approximate $e^{\tau\boldsymbol{A}}\boldsymbol{u}_0$ for $\tau = 0.05$ by the standard and a rational Krylov subspace method with denominator polynomial $q_{m-1}(z) = (1-z)^{m-1}$. For $n = 10, 20, 30$, the comparison of the obtained error curves is shown in Figure 4.3. It becomes evident that the rational Krylov subspace method achieves a high accuracy in a few iteration steps independent of the grid spacing $h = \frac{1}{n+1}$, whereas the convergence behavior of the polynomial method deteriorates the larger the value n is chosen. This effect can be explained by the error bound (4.8) that predicts a superlinear convergence only after $m \geq \sqrt{\|\tau\boldsymbol{A}\|}$ Krylov steps. Since $\|\tau\boldsymbol{A}\|$ is proportional to $\frac{\tau}{h^2}$, it is not surprising that the convergence is worse, if we decrease the mesh size h.

In order to prove a grid-independent convergence, we therefore need a uniform error bound that applies to all matrices with an arbitrary field of values on the negative real line. With regard to general evolution equations of parabolic or hyperbolic type, we have to look for error estimates that are uniform for all matrices with an arbitrarily large field of values in the left complex half-plane. This condition can never be fulfilled by a polynomial Krylov subspace method, but we will see later in this thesis that rational Krylov subspace methods are suitable to approximate uniformly matrix functions of such stiff matrices and hence guarantee a grid-independent convergence. ❍

4.4 Near-optimality of the Krylov subspace approximation

The exactness property of the standard and the rational Krylov subspace method in Lemma 4.7 and 4.12 can be used to bound the error $\|f(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v}\|$ of the Krylov subspace approximation by a scalar approximation problem on the field of values $W(\boldsymbol{A})$. For this purpose, we need a helpful theorem of Crouzeix [12].

Lemma 4.15 *For an arbitrary matrix $\boldsymbol{A}$ and any polynomial p, we have*

$$\|p(\boldsymbol{A})\| \le \mathcal{C} \sup_{z \in W(\boldsymbol{A})} |p(z)|\,, \qquad \mathcal{C} \le 11.08\,.$$

Crouzeix conjectures that this bound can be improved to $\mathcal{C} = 2$. By Runge's Theorem, cf. Corollary 12.1.2 in [28], the function f can be uniformly approximated by a polynomial on $W(\boldsymbol{A})$, if f is analytic in a neighborhood of $W(\boldsymbol{A})$. A density argument then makes it possible to generalize Lemma 4.15 to any function f that is analytic in a neighborhood of the field of values: Let Σ be a set with $W(\boldsymbol{A}) \subseteq \Sigma$, then

$$\|f(\boldsymbol{A})\| \le \mathcal{C} \sup_{z \in \Sigma} |f(z)| \tag{4.14}$$

holds true. This estimate is a very powerful tool, which enables us to formulate a well-known result concerning the near-optimality of the polynomial and the rational Krylov subspace approximation.

Theorem 4.16 *Let $\boldsymbol{S}_m = \boldsymbol{V}_m^H \boldsymbol{A} \boldsymbol{V}_m$ be the compression of $\boldsymbol{A}$ onto the standard Krylov subspace $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$ and let the function f be analytic in a neighborhood of $W(\boldsymbol{A})$. Then for any set $\Sigma \supseteq W(\boldsymbol{A})$, we have*

$$\|f(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v}\| \le 2\,\mathcal{C}\|\boldsymbol{v}\| \min_{p \in \mathcal{P}_{m-1}} \sup_{z \in \Sigma} |f(z) - p(z)|\,.$$

The same holds true, if $\widetilde{\boldsymbol{S}}_m = \widetilde{\boldsymbol{V}}_m^H \boldsymbol{A} \widetilde{\boldsymbol{V}}_m$ is the compression of $\boldsymbol{A}$ onto the rational Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$, that is,

$$\|f(\boldsymbol{A})\boldsymbol{v} - \widetilde{\boldsymbol{V}}_m f(\widetilde{\boldsymbol{S}}_m)\widetilde{\boldsymbol{V}}_m^H \boldsymbol{v}\| \le 2\,\mathcal{C}\|\boldsymbol{v}\| \min_{r \in \frac{\mathcal{P}_{m-1}}{q_{m-1}}} \sup_{z \in \Sigma} |f(z) - r(z)|\,.$$

Proof. Due to Lemma 4.7, we know in the first case that $p(\boldsymbol{A})\boldsymbol{v} = \boldsymbol{V}_m p(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v}$ for any polynomial $p \in \mathcal{P}_{m-1}$. Since $\boldsymbol{V}_m$ has orthonormal columns by assumption, we have $\|\boldsymbol{V}_m \boldsymbol{x}\| = \|\boldsymbol{x}\|$ for all $\boldsymbol{x} \in \mathbb{C}^N$. Moreover, $\boldsymbol{V}_m \boldsymbol{V}_m^H$ is the projection onto $\mathcal{K}_m(\boldsymbol{A}, \boldsymbol{v})$, such that $\boldsymbol{V}_m \boldsymbol{V}_m^H \boldsymbol{v} = \boldsymbol{v}$ and thus $\|\boldsymbol{V}_m^H \boldsymbol{v}\| = (\boldsymbol{v}^H \boldsymbol{V}_m \boldsymbol{V}_m^H \boldsymbol{v})^{1/2} = \|\boldsymbol{v}\|$. This yields

$$\begin{aligned}\|f(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v}\| &= \|f(\boldsymbol{A})\boldsymbol{v} - p(\boldsymbol{A})\boldsymbol{v} + \boldsymbol{V}_m p(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v} - \boldsymbol{V}_m f(\boldsymbol{S}_m)\boldsymbol{V}_m^H \boldsymbol{v}\| \\ &\le \|f(\boldsymbol{A}) - p(\boldsymbol{A})\|\|\boldsymbol{v}\| + \|p(\boldsymbol{S}_m) - f(\boldsymbol{S}_m)\|\|\boldsymbol{v}\| \\ &\le 2\,\mathcal{C}\|\boldsymbol{v}\| \sup_{z \in \Sigma} |f(z) - p(z)|\,,\end{aligned}$$

where the last inequality is obtained by (4.14) and the fact that $W(\boldsymbol{S}_m) \subseteq W(\boldsymbol{A})$. If we take the minimum over all polynomials $p \in \mathcal{P}_{m-1}$, the first statement of the theorem is proved. Analogously, using Lemma 4.12, the second estimate for the rational Krylov subspace approximation can be shown. ❑

For normal matrices, the bounds in Theorem 4.16 hold with $\mathcal{C} = 1$ and $\Sigma = \sigma(\boldsymbol{A}) \cup \sigma(\boldsymbol{S}_m)$: Since in this case $\boldsymbol{A}$ is unitary diagonalizable, that is, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^H$ with unitary $\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{D}$, it follows

$$\|f(\boldsymbol{A})\boldsymbol{v} - p(\boldsymbol{A})\boldsymbol{v}\| \leq \|\boldsymbol{U}\big(f(\boldsymbol{D}) - p(\boldsymbol{D})\big)\boldsymbol{U}^H\|\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\| \sup_{z \in \sigma(\boldsymbol{A})} |f(z) - p(z)|$$

for any polynomial $p \in \mathcal{P}_{m-1}$. Analogously, we derive

$$\|f(\boldsymbol{A})\boldsymbol{v} - r(\boldsymbol{A})\boldsymbol{v}\| \leq \|\boldsymbol{v}\| \sup_{z \in \sigma(\boldsymbol{A})} |f(z) - r(z)|$$

for any rational function $r \in \mathcal{P}_{m-1}/q_{m-1}$. The same applies for $\|f(\boldsymbol{A}_m)\boldsymbol{v} - p(\boldsymbol{A}_m)\boldsymbol{v}\|$ and $\|f(\boldsymbol{A}_m)\boldsymbol{v} - r(\boldsymbol{A}_m)\boldsymbol{v}\|$.

Theorem 4.16 has a far-reaching consequence: In order to find an error bound for the Krylov subspace approximation, we may study the scalar best approximation problem

$$\min_{p \in \mathcal{P}_{m-1}} \sup_{z \in \Sigma} |f(z) - p(z)| \qquad \text{or} \qquad \min_{r \in \frac{\mathcal{P}_{m-1}}{q_{m-1}}} \sup_{z \in \Sigma} |f(z) - r(z)|$$

on a set $\Sigma \supseteq W(\boldsymbol{A})$. Whereas for the polynomial problem there exists a unique best approximation, the rational best approximation exists, but it might not be unique (e.g., [87]).

The exponential function $f(z) = e^z$, for example, can be approximated by a truncated Taylor series, Chebyshev or Faber polynomials, see, for instance, [6,41,72,81]. Since polynomials are unbounded on unbounded domains, the results achieved with these approaches are only useful, if $W(\boldsymbol{A})$ is a bounded set of moderate size in the complex plane.

Possible rational approximations are Padé, Chebyshev-Padé or Carathéodory-Fejér approximations and methods which are based on the Faber transform for rational functions or the application of a quadrature rule to a suitable contour integral, e.g., [4,34,75,83]. A natural candidate for the latter is the Cauchy integral formula

$$f(z) = \frac{1}{2\pi i} \int_\Gamma \frac{f(\xi)}{\xi - z}\, d\xi\,,$$

where Γ is a path with parametrization $\Gamma(t)$ winding around $W(\boldsymbol{A})$. Applying the truncated trapezoidal rule with $m-1$ equally spaced nodes w_k of distance h, we obtain the rational approximation

$$f(z) \approx r(z) = \frac{h}{2\pi i} \sum_{k=1}^{m-1} \frac{f\big(\Gamma(w_k)\big)}{\Gamma(w_k) - z} \Gamma'(w_k)\,.$$

Another way to obtain upper bounds for the standard and the rational Krylov subspace approximation relies on a scalar problem of approximating f by an interpolating polynomial p or an interpolating rational function r. This approach uses the following formulas given in Walsh [87] (Chapter III, equation (4) and Chapter VIII, Theorem 2): Let the function f be analytic in a domain $\Sigma \subset \mathbb{C}$ and let $p \in \mathcal{P}_{m-1}$ interpolate f at the nodes $\alpha_1, \ldots, \alpha_m \in \Sigma$. If Γ is a contour in Σ encircling $\alpha_1, \ldots, \alpha_m$, we have the error representation

$$f(z) - p(z) = \frac{1}{2\pi i} \int_\Gamma \frac{s_{m,0}(z)}{s_{m,0}(\xi)} \frac{f(\xi)}{\xi - z}\, d\xi\,, \qquad z \in \operatorname{int}(\Gamma)\,,$$

where $s_{m,0}(z) = (z - \alpha_1) \cdots (z - \alpha_m)$. Similarly, the error for the approximation of f by a rational function r with interpolation nodes $\alpha_1, \ldots, \alpha_m$ and poles $\beta_1, \ldots, \beta_{m-1}$ distinct from the values α_k, $k = 1, \ldots, m$, is given by

$$f(z) - r(z) = \frac{1}{2\pi i} \int_\Gamma \frac{s_{m,m-1}(z)}{s_{m,m-1}(\xi)} \frac{f(\xi)}{\xi - z} \, d\xi \,, \qquad z \in \mathrm{int}(\Gamma) \,, \qquad z \notin \{\beta_1, \ldots, \beta_{m-1}\}$$

with

$$s_{m,m-1}(z) = \frac{(z - \alpha_1) \cdots (z - \alpha_m)}{(z - \beta_1) \cdots (z - \beta_{m-1})} \,.$$

This leads to the difficult task to select the nodes α_k, the poles β_k, and the contour Γ in such a way that

$$\frac{\max_{z \in \Sigma} |s_{m,j}(z)|}{\min_{\xi \in \Gamma} |s_{m,j}(\xi)|} \,, \qquad j \in \{0, m-1\} \tag{4.15}$$

is as small as possible, in order to give a reasonable error estimate. This leads to tools from logarithmic potential theory (see, for instance, [30]). However, it would go beyond the scope of this thesis to get into further details about this theory here.

In all these cases, it has to be ensured that the roots $z_1, \ldots, z_{m-1}$ of the prescribed denominator polynomial $q_{m-1}(z)$ of the rational Krylov subspace method must coincide with the poles of the chosen rational approximation $r(z)$. This means, in particular, that the poles have to be known in advance.

Moreover, there exist rational Krylov subspace methods that use an automated parameter selection. For functions of Cauchy-Stieltjes type[3], the authors in [32] suggest to choose the next pole z_m such that $|s_{m,m-1}(z_m)| = \min_{\xi \in \Gamma} |s_{m,m-1}(\xi)|$. This strategy aims to make the denominator in (4.15) as large as possible.

4.5 Generalization and computation of the Krylov subspace approximation

Previously, we only considered the Euclidean inner product. In this section, we generalize the Krylov subspace approximation to arbitrary inner products $(\cdot\,,\cdot)_{\boldsymbol{M}}$ on $\mathbb{C}^N$, where $\boldsymbol{M} \in \mathbb{C}^{N \times N}$ is a fixed positive definite Hermitian matrix. In this case, we have the very similar formula

$$f(\boldsymbol{A})\boldsymbol{v} \approx f(\boldsymbol{A}_m)\boldsymbol{v} = \boldsymbol{V}_m f(\boldsymbol{S}_m) \boldsymbol{V}_m^+ \boldsymbol{v} \tag{4.16}$$

for any function f that is analytic on $W(\boldsymbol{A})$ and defined at the point zero. Like before, $\boldsymbol{S}_m = \boldsymbol{V}_m^+ \boldsymbol{A} \boldsymbol{V}_m \in \mathbb{C}^{m \times m}$ is the compression and $\boldsymbol{A}_m = \boldsymbol{P}_m \boldsymbol{A} \boldsymbol{P}_m$ is the restriction of $\boldsymbol{A}$ onto the Krylov subspace. The matrix $\boldsymbol{P}_m = \boldsymbol{V}_m \boldsymbol{V}_m^+ = \boldsymbol{V}_m \boldsymbol{V}_m^H \boldsymbol{M}$ designates the orthogonal projection on the Krylov subspace with respect to the inner product $(\cdot\,,\cdot)_{\boldsymbol{M}}$ on $\mathbb{C}^N$ and $\boldsymbol{V}_m^+ = \boldsymbol{V}_m^H \boldsymbol{M}$ is the Moore-Penrose inverse of $\boldsymbol{V}_m$ defined in Definition 4.1 associated to $(\cdot\,,\cdot)_{\boldsymbol{M}}$. Using the approximation (4.16), the problem of computing $f(\boldsymbol{A})\boldsymbol{v}$ for a large N-by-N matrix $\boldsymbol{A}$ is reduced to the evaluation of $\boldsymbol{V}_m f(\boldsymbol{S}_m) \boldsymbol{V}_m^+ \boldsymbol{v}$. In this case, we only have to evaluate a matrix function for the small matrix $\boldsymbol{S}_m$ of size $m \times m$ with $m \ll N$. This can be done by standard algorithms for dense matrices.

[3] Functions of Cauchy-Stieltjes type can be written in the form $f(z) = \int_\Gamma (z-x)^{-1} d\mu(x)$ with some measure μ supported on a closed set $\Gamma \subset \mathbb{C}$.

Krylov subspaces can also be defined for operators A on some Hilbert space H. For this purpose, we replace the matrix $\boldsymbol{A}$ and the vector $\boldsymbol{v}$ by the operator A and a vector $v \in H$. The Krylov subspace approximation to the operator function $f(A)$ times v is defined analogously by $f(A_m)v = f(P_m A P_m)v$, where P_m is the orthogonal projector onto the considered Krylov subspace $\mathcal{K}_m(A, v)$ or $\mathcal{Q}_m(A, v)$.

While it is clear how $f(\boldsymbol{A}_m)\boldsymbol{v}$ is computed in the discrete case, we have to think about what is meant by the notation $f(A_m)v = f(P_m A P_m)v$ in the case of an operator. According to Section 4.1, the projection operator P_m reads $P_m = V_m V_m^+$, where V_m is the quasi-matrix $V_m = [v_1 \cdots v_m]$ with $v_j \in H$ for $j = 1, \ldots, m$, that contains an orthonormal basis of the Krylov subspace. The Moore-Penrose inverse V_m^+ applied to $v \in H$ is understood as the vector

$$V_m^+ v = \begin{bmatrix} (v, v_1) \\ (v, v_2) \\ \vdots \\ (v, v_m) \end{bmatrix} \in \mathbb{C}^m \,,$$

containing all inner products of v and the basis functions $v_1, \ldots, v_m$ of the space onto which we want to project. We also need the compression $V_m^+ A V_m$ of the operator A onto the Krylov subspace that is computed to

$$V_m^+ A V_m = \begin{bmatrix} (Av_1, v_1) & \cdots & (Av_m, v_1) \\ \vdots & & \vdots \\ (Av_1, v_m) & \cdots & (Av_m, v_m) \end{bmatrix} \in \mathbb{C}^{m \times m} \,,$$

which we also designate with $\boldsymbol{S}_m$ in analogy to the discrete case. Then the Krylov subspace approximation for operator functions is given by

$$f(A)v \approx f(A_m)v = V_m f(\boldsymbol{S}_m) V_m^+ v = V_m \, \boldsymbol{g} = \sum_{j=1}^{m} g_j v_j \in H \,,$$

where $\boldsymbol{g} = (g_j)_{j=1}^m = f(\boldsymbol{S}_m) V_m^+ v \in \mathbb{C}^m$ and $\boldsymbol{S}_m = V_m^+ A V_m \in \mathbb{C}^{m \times m}$. This shows that even in the case of operators, we only have to compute a matrix function $f(\boldsymbol{S}_m)$ of a small $m \times m$-matrix, which is then multiplied by the vector $V_m^+ v \in \mathbb{C}^m$.

For our purposes, mainly $f(z) = \varphi_\ell(z)$ is of great interest. In particular, we have seen that, generally, every time step in the exponential integrator requires the evaluation of the linear combination

$$\varphi_0(\tau \boldsymbol{A})\boldsymbol{w}_0 + \varphi_1(\tau \boldsymbol{A})\boldsymbol{w}_1 + \ldots + \varphi_s(\tau \boldsymbol{A})\boldsymbol{w}_s$$

of the matrix φ-functions acting on certain vectors $\boldsymbol{w}_j$. The approximation of $\varphi_\ell(\tau \boldsymbol{A})\boldsymbol{w}_j$ by a Krylov subspace method leads to expressions of the type $\boldsymbol{V}_m \varphi_\ell(\tau \boldsymbol{S}_m) \boldsymbol{V}_m^+ \boldsymbol{w}_j$.

For the computation of $\varphi_\ell(\boldsymbol{S}_m)$ times a vector $\boldsymbol{v}$, there exists an elegant way based on an idea of Saad [72]. He used an augmented matrix $\widetilde{\boldsymbol{S}}_m$ and calculated the matrix exponential of $\widetilde{\boldsymbol{S}}_m$ to obtain $\boldsymbol{b}^T \varphi_1(\boldsymbol{S}_m)$ for any vector $\boldsymbol{b} \in \mathbb{C}^m$. This result was generalized by Sidje [77] to the computation of $\varphi_\ell(\boldsymbol{S}_m)\boldsymbol{v}$ for an arbitrary vector $\boldsymbol{v} \in \mathbb{C}^m$ with the help of the augmented matrix

$$\widetilde{\boldsymbol{S}}_{m+\ell} = \begin{bmatrix} \boldsymbol{S}_m & \boldsymbol{v} & \\ & & \boldsymbol{I}_{\ell-1} \\ 0 & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{(m+\ell)\times(m+\ell)} \,,$$

where $\boldsymbol{I}_{\ell-1}$ is the identity matrix of dimension $\ell - 1$. Sidje has shown that the matrix exponential of this augmented matrix is

$$e^{\widetilde{\boldsymbol{S}}_{m+\ell}} = \begin{bmatrix} \varphi_0(\boldsymbol{S}_m) & \varphi_1(\boldsymbol{S}_m)\boldsymbol{v} & \varphi_2(\boldsymbol{S}_m)\boldsymbol{v} & \cdots & \varphi_\ell(\boldsymbol{S}_m)\boldsymbol{v} \\ & 1 & \dfrac{1}{1!} & \cdots & \dfrac{1}{(\ell-1)!} \\ & & 1 & \ddots & \vdots \\ & & & \ddots & \dfrac{1}{1!} \\ & & & & 1 \end{bmatrix}.$$

The desired vector $\varphi_\ell(\boldsymbol{S}_m)\boldsymbol{v} \in \mathbb{C}^m$ is then obtained by taking the first m entries of the last column of the matrix exponential of $\widetilde{\boldsymbol{S}}_{m+\ell}$. For a matrix $\boldsymbol{S}_m$ of moderate size, one can use the Matlab function `expm` to determine this matrix exponential of $\widetilde{\boldsymbol{S}}_{m+\ell}$. This Matlab function computes the matrix exponential by a Padé approximation with scaling and squaring (see Higham [37]).

Chapter 5

Shift-and-invert Krylov subspace approximation

In this chapter, we consider for $\gamma > 0$ the approximation of $f(\boldsymbol{A})\boldsymbol{v}$ in the so-called shift-and-invert Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$ with denominator $q_{m-1}(z) = (\gamma - z)^{m-1}$, that is

$$\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v}) = \mathcal{K}_m\big((\gamma \boldsymbol{I} - \boldsymbol{A})^{-1}, \boldsymbol{v}\big) = \operatorname{span}\left\{\boldsymbol{v}, \frac{1}{\gamma - \boldsymbol{A}}\,\boldsymbol{v}, \dots, \frac{1}{(\gamma - \boldsymbol{A})^{m-1}}\,\boldsymbol{v}\right\} \tag{5.1}$$

for an arbitrary vector $\boldsymbol{v} \in \mathbb{C}^N$ and a large matrix $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ with a field of values somewhere in the left complex half-plane, that means

$$W(\boldsymbol{A}) \subseteq \mathbb{C}_0^- = \{z \in \mathbb{C} \,:\, \operatorname{Re}(z) \leq 0\}\,.$$

For the chosen inner product $(\cdot\,,\cdot)$ on $\mathbb{C}^N$, the required condition $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ is equivalent to the relation $\operatorname{Re}(\boldsymbol{A}\boldsymbol{v}, \boldsymbol{v}) \leq 0$ for all $\boldsymbol{v} \in \mathbb{C}^N$. We designate by $\|\cdot\|$ the norm associated with the inner product on $\mathbb{C}^N$. As matrix norm, we always choose the induced norm that we denote with $\|\cdot\|$, too.

In the context of exponential integrators, we are particularly interested in the special case $f(z) = \varphi_\ell(z)$ and matrices $\boldsymbol{A}$ stemming from a spatial discretization of a partial differential equation. For fine discretizations, such matrices typically have a widely distributed field of values in the left complex half-plane and a huge norm. Exponential integrators have the favorable property that the temporal convergence results are independent of $\|\boldsymbol{A}\|$. Furthermore, they can be regarded as explicit schemes without a severe restriction of the time step size, even if the norm of the discretization matrix is very large. To preserve these benefits, it is important to approximate the matrix φ-functions independent of $\|\boldsymbol{A}\|$. We will derive error bounds for the shift-and-invert Krylov subspace method which satisfy this requirement.

In order to do this, we first approximate the matrix function $f(\boldsymbol{A})$ in the matrix subspace

$$\mathcal{R}_m(\boldsymbol{A}) = \operatorname{span}\left\{\boldsymbol{I}, \frac{1}{\gamma - \boldsymbol{A}}, \dots, \frac{1}{(\gamma - \boldsymbol{A})^{m-1}}\right\}\,. \tag{5.2}$$

This result can then be used to bound the error for the approximation of $f(\boldsymbol{A})\boldsymbol{v}$ in the rational Krylov subspace $\mathcal{R}_m(\boldsymbol{A})\boldsymbol{v} = \mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$. Since it will become apparent that $f(\boldsymbol{A})$ can be approximated uniformly in $\mathcal{R}_m(\boldsymbol{A})$ for every matrix $\boldsymbol{A}$ with $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$, this is also true for the shift-and-invert Krylov subspace approximation to $f(\boldsymbol{A})\boldsymbol{v}$. Whenever the notation $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$ is used in this chapter, we always mean the shift-and-invert Krylov subspace defined in (5.1) with denominator $q_{m-1}(z) = (\gamma - z)^{m-1}$.

We also discuss possible choices of the free parameter $\gamma > 0$, that speed up our convergence rate. Moreover, following the paper [29] by Grimm, we resume the results for abstract evolution equations and the approximation of operator functions $f(A)v$ in a so-called resolvent Krylov subspace spanned by powers of the resolvent $(\gamma I - A)^{-1}$ and the vector v.

5.1 Transformation to the unit disk

In order to estimate the error for the approximation of $f(\boldsymbol{A})$ in the rational matrix space $\mathcal{R}_m(\boldsymbol{A})$ defined in (5.2), we first reduce the problem to a scalar approximation problem on $\mathbb{C}_0^-$ by using Crouzeix's inequality (4.14). Afterwards, the problem of approximating $f(z)$ in the left complex half-plane is transformed to an approximation problem on the unit circle, where well-known results for trigonometric approximation can be applied.

In the following, we denote by $\mathbb{D}$ the open unit disk in the complex plane and by $\overline{\mathbb{D}}$ the closed unit disk with boundary $\partial\mathbb{D}$, that is,

$$\mathbb{D} := \{z \in \mathbb{C} \,:\, |z| < 1\}\,, \qquad \overline{\mathbb{D}} := \{z \in \mathbb{C} \,:\, |z| \le 1\}\,, \qquad \partial\mathbb{D} := \{z \in \mathbb{C} \,:\, |z| = 1\}\,.$$

Similar to Lemma 4.9, the next lemma provides an alternative representation of the matrix space $\mathcal{R}_m(\boldsymbol{A})$, that will be useful for later purposes.

Lemma 5.1 *We have* $\left(\frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}\right)^m \in \mathcal{R}_{m+1}(\boldsymbol{A})$ *for* $m = 0, 1, 2, \ldots$ *and* $\gamma > 0$. *Moreover, it holds that*

$$\mathcal{R}_m(\boldsymbol{A}) = \operatorname{span}\left\{\boldsymbol{I}, \frac{1}{\gamma-\boldsymbol{A}}, \ldots, \frac{1}{(\gamma-\boldsymbol{A})^{m-1}}\right\} = \operatorname{span}\left\{\boldsymbol{I}, \frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}, \ldots, \left(\frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}\right)^{m-1}\right\}.$$

Proof. We exploit the identity $\frac{\gamma}{\gamma-\boldsymbol{A}} - \frac{\boldsymbol{A}}{\gamma-\boldsymbol{A}} = \boldsymbol{I}$ to obtain

$$\frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}} = \frac{\gamma}{\gamma-\boldsymbol{A}} + \frac{\boldsymbol{A}}{\gamma-\boldsymbol{A}} = \frac{2\gamma}{\gamma-\boldsymbol{A}} - \boldsymbol{I} \in \mathcal{R}_2(\boldsymbol{A})\,. \tag{5.3}$$

Since the identity matrix $\boldsymbol{I}$ commutes with any matrix, we can conclude

$$\left(\frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}\right)^k = \left(\frac{2\gamma}{\gamma-\boldsymbol{A}} - \boldsymbol{I}\right)^k = \sum_{j=0}^{k}\binom{k}{j}(-1)^{k-j}(2\gamma)^j\left(\frac{1}{\gamma-\boldsymbol{A}}\right)^j \in \mathcal{R}_{k+1}(\boldsymbol{A})$$

by the binomial theorem. This proves the inclusion "$\supseteq$". To show the missing inclusion "$\subseteq$", we use (5.3) to find

$$\frac{1}{\gamma-\boldsymbol{A}} = \frac{1}{2\gamma}\left(\boldsymbol{I} + \frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}\right).$$

This gives

$$\begin{aligned}\left(\frac{1}{\gamma-\boldsymbol{A}}\right)^k &= \left(\frac{1}{2\gamma}\right)^k\left(\boldsymbol{I} + \frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}\right)^k \\ &= \left(\frac{1}{2\gamma}\right)^k\sum_{j=0}^{k}\binom{k}{j}\left(\frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}\right)^j \in \operatorname{span}\left\{\boldsymbol{I}, \frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}, \ldots, \left(\frac{\gamma+\boldsymbol{A}}{\gamma-\boldsymbol{A}}\right)^k\right\}\end{aligned}$$

and our lemma is proved. ❑

This lemma enables us to turn the best approximation problem

$$\inf_{r \in \frac{\mathcal{P}_{m-1}}{q_{m-1}}} \|f(\boldsymbol{A}) - r(\boldsymbol{A})\| = \inf_{a_k}\left\|f(\boldsymbol{A}) - \sum_{k=0}^{m-1} a_k \frac{1}{(\gamma-\boldsymbol{A})^k}\right\|,$$

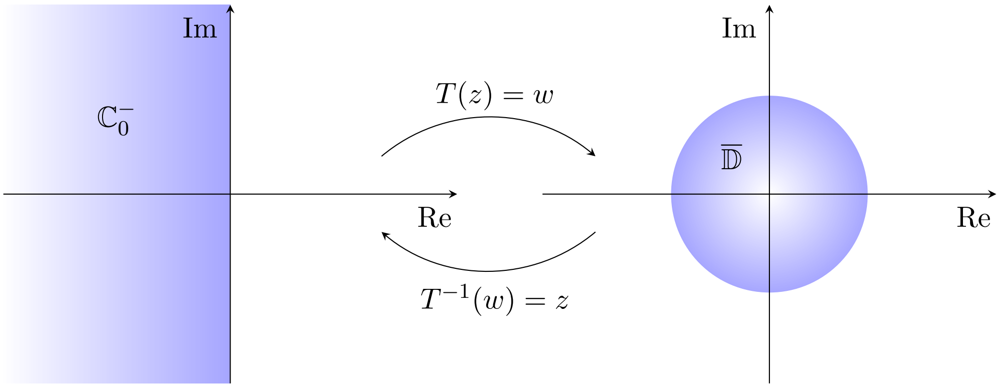


Figure 5.1: Transformation of the left complex half-plane $\mathbb{C}_0^-$ onto the unit disk $\overline{\mathbb{D}}\backslash\{-1\}$ and vice versa.

cf. Lemma 4.9, into the best approximation problem

$$\inf_{c_k} \left\| f(\boldsymbol{A}) - \sum_{k=0}^{m-1} c_k \left(\frac{\gamma + \boldsymbol{A}}{\gamma - \boldsymbol{A}} \right)^k \right\| = \inf_{c_k} \left\| f(\boldsymbol{A}) - \sum_{k=0}^{m-1} c_k \left(\frac{\boldsymbol{I} + \frac{1}{\gamma}\boldsymbol{A}}{\boldsymbol{I} - \frac{1}{\gamma}\boldsymbol{A}} \right)^k \right\| .$$

Due to Crouzeix's inequality (4.14), we know that for any function f that is analytic in a neighborhood of $W(\boldsymbol{A})$, the norm of the matrix function $f(\boldsymbol{A})$ can be bounded by $\|f(\boldsymbol{A})\| \leq \mathcal{C} \sup_{z\in\Sigma} |f(z)|$, $\mathcal{C} \leq 11.08$, for any set Σ with $W(\boldsymbol{A}) \subseteq \Sigma$. For matrices $\boldsymbol{A}$ whose field of values lies in a half-plane, von Neumann has shown (see Section 5.3 in [86]) that Crouzeix's inequality holds with $\mathcal{C} = 1$. By our assumption $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$, we thus have

$$\|f(\boldsymbol{A})\| \leq \sup_{z\in\mathbb{C}_0^-} |f(z)| \,, \tag{5.4}$$

whenever f is analytic in the left complex half-plane including the imaginary axis. This inequality is a powerful tool that makes it possible to reduce the approximation of $f(\boldsymbol{A})$ in the rational space $\mathcal{R}_m(\boldsymbol{A})$ to a polynomial approximation problem on the closed unit disk $\overline{\mathbb{D}}$ that is easier to handle. To see this, we set

$$T(z) := w = \frac{1 + \frac{z}{\gamma}}{1 - \frac{z}{\gamma}} \quad \Longleftrightarrow \quad z = T^{-1}(w) = \gamma \frac{w-1}{w+1} \,, \tag{5.5}$$

where T is the Möbius transformation that maps the left complex half-plane $\mathbb{C}_0^-$ onto the unit disk $\overline{\mathbb{D}}\backslash\{-1\}$.

We now define the set

$$\mathcal{M} := \{ f : \mathbb{C} \to \mathbb{C} \,:\, f \text{ analytic in } \mathbb{C}_0^- \text{ and } \exists\, c \in \mathbb{C} \text{ with } f(z) \xrightarrow{|z|\to\infty} c \,,\; z \in \mathbb{C}_0^- \} \,.$$

Then, for $f \in \mathcal{M}$, the transformed function $\widetilde{f}(w) := f\big(T^{-1}(w)\big)$ is analytic in $\overline{\mathbb{D}}\backslash\{-1\}$. Moreover, by setting $\widetilde{f}(-1) = c$, the function $\widetilde{f}$ can be extended continuously to the point -1. With the help of the Möbius transformation T, relation (5.5), and inequality (5.4), we now obtain the estimate

$$\begin{aligned}\left\| f(\boldsymbol{A}) - \sum_{k=0}^{m-1} c_k \left(\frac{\boldsymbol{I} + \frac{1}{\gamma}\boldsymbol{A}}{\boldsymbol{I} - \frac{1}{\gamma}\boldsymbol{A}} \right)^k \right\| &\leq \sup_{z \in \mathbb{C}_0^-} \left| f(z) - \sum_{k=0}^{m-1} c_k \left(\frac{1 + \frac{z}{\gamma}}{1 - \frac{z}{\gamma}} \right)^k \right| \\ &\leq \max_{w \in \overline{\mathbb{D}}} \left| f\left(\gamma \, \frac{w-1}{w+1} \right) - \sum_{k=0}^{m-1} c_k w^k \right| \\ &= \max_{w \in \overline{\mathbb{D}}} \left| \widetilde{f}(w) - \sum_{k=0}^{m-1} c_k w^k \right| .\end{aligned}$$

For the sake of brevity and a clear representation, we additionally introduce at this point the short notations $\widehat{f}(t) := f\big(T^{-1}(e^{it})\big)$ and $\mathbb{T} := [0, 2\pi)$.

If we assume that $f \in \mathcal{M}$, the transformed function $\widetilde{f}$ is analytic on $\mathbb{D}$ and continuous on $\overline{\mathbb{D}}$ and we can conclude with the Maximum Modulus Theorem from complex analysis that

$$\max_{w \in \overline{\mathbb{D}}} \left| \widetilde{f}(w) - \sum_{k=0}^{m-1} c_k w^k \right| = \max_{w \in \partial\mathbb{D}} \left| \widetilde{f}(w) - \sum_{k=0}^{m-1} c_k w^k \right| = \max_{t \in \mathbb{T}} \left| \widehat{f}(t) - \sum_{k=0}^{m-1} c_k e^{itk} \right| .$$

So, we are now concerned with a trigonometric approximation problem for the 2π-periodic function $\widehat{f}(t)$ on $\mathbb{T}$. For this purpose, we split $\widehat{f}(t)$ into its real and imaginary part and apply a result of Achyèser [1] to the 2π-periodic real-valued functions $\mathrm{Re}[\widehat{f}(t)]$ and $\mathrm{Im}[\widehat{f}(t)]$. His result, formulated in the subsequent theorem, is based on Jackson's well-known theorem about the error of the best uniform approximation to a real-valued periodic function by a trigonometric polynomial. Achyèser's theorem involves the modulus of continuity $\omega(g, \delta)$, which we state in the following definition.

Definition 5.2 *The modulus of continuity $\omega(g, \delta)$, $\delta > 0$, of a real-valued continuous function g is defined as*

$$\omega(g, \delta) := \sup_{|s-t| \leq \delta} |g(s) - g(t)| .$$

In particular, if g is a 2π-periodic and continuous function, we have

$$\omega(g, \delta) = \sup_{0 < h \leq \delta} \max_{t \in \mathbb{T}} |g(t+h) - g(t)| .$$

Theorem 5.3 (Achyèser [1], Section 4) *Let g be a 2π-periodic real-valued function that is n times continuously differentiable. Then there exists a trigonometric polynomial*

$$J_{m,n}(g, t) = \frac{a_0}{2} + \sum_{k=1}^{m-1} D_k(m, n)\big(a_k \cos(kt) + b_k \sin(kt)\big) , \tag{5.6}$$

where a_k, b_k are the Fourier coefficients of the function g and $D_k(m, n)$ are real constants that depend on m and n, such that

$$\max_{t \in \mathbb{T}} |g(t) - J_{m,n}(g, t)| \leq \frac{C(n)}{m^n} \, \omega\left(g^{(n)}, \frac{1}{m} \right)$$

with a constant $C(n)$ that depends only on n.

The proof of this theorem is based on the Jackson integral

$$J_n(x) = \frac{1}{2}\lambda_n \int_{\mathbb{T}} f(x+t)\left(\frac{\sin\left(\frac{nt}{2}\right)}{n\sin\left(\frac{t}{2}\right)}\right)^4 dt = \frac{1}{2}\lambda_n \int_{\mathbb{T}} f(x+t)K_n(t)\,dt\,, \qquad n \in \mathbb{N}\,,$$

where $\frac{1}{2}\lambda_n K_n(t)$ is the so-called Jackson kernel and the constant λ_n is defined by the relation $\int_{\mathbb{T}} \frac{1}{2}\lambda_n K_n(t)\,dt = 1$. In [1], Achyèser uses a slightly different representation of the Jackson integral by de la Vallée Poussin [13] that is given as

$$\widetilde{J}_n(x) = \frac{3}{2\pi}\int_{\mathbb{R}} f\left(x+\frac{2t}{n}\right)\left(\frac{\sin(t)}{t}\right)^4 dt\,, \qquad n \in \mathbb{N}\,.$$

The Jackson integrals $J_n(x)$ and $\widetilde{J}_n(x)$ are trigonometric polynomials of degree $2n-2$ (cf. [54], p. 55) respectively $2n-1$ (cf. [1]). They are applied in approximation theory in order to bound the best trigonometric approximation of a periodic function. Using the integral $\widetilde{J}_n(x)$, Achyèser has proved the bound in Theorem 5.3 only for the case that m is even. If we make use of the representation $J_n(x)$ instead of $\widetilde{J}_n(x)$, it turns out that the result is also valid for odd values of m.

The expression $J_{m,n}(g,t)$ in Theorem 5.3 looks quite similar to the $(m-1)$st partial Fourier sum of the function g that is given by

$$S_{m-1}g(t) = \frac{a_0}{2} + \sum_{k=1}^{m-1} a_k\cos(kt) + b_k\sin(kt)$$

with Fourier coefficients

$$a_k = \frac{1}{\pi}\int_0^{2\pi} g(t)\cos(kt)\,dt\,, \qquad b_k = \frac{1}{\pi}\int_0^{2\pi} g(t)\sin(kt)\,dt\,,$$

except for the fact that the sum in (5.6) contains additional constants $D_k(m,n)$. These constants are specified by the Jackson kernel.

Achyèser's Theorem 5.3 is only applicable to real-valued and not to complex-valued functions. Therefore, we have to subdivide our problem into two real approximation problems: one for the real part and one for the imaginary part. But since we want to examine the expression $\inf_{c_k}\max_{t\in\mathbb{T}}|\widehat{f}(t) - \sum_{k=0}^{m-1} c_k e^{itk}|$, it is not possible to simply approximate the real and imaginary part of $\widehat{f}(t)$ separately from each other. The trigonometric approximation of the two parts is coupled by the common coefficients c_k. Only the special form of the Fourier coefficients associated to $\mathrm{Re}[\widehat{f}(t)]$ and $\mathrm{Im}[\widehat{f}(t)]$ will allow us to consider these two terms independently. More precisely, we exploit the fact that if the Fourier coefficients of $\mathrm{Re}[\widehat{f}(t)]$ are given as a_k^{Re} and b_k^{Re}, then $\mathrm{Im}[\widehat{f}(t)]$ has the Fourier coefficients $a_k^{\mathrm{Im}} = -b_k^{\mathrm{Re}}$ and $b_k^{\mathrm{Im}} = a_k^{\mathrm{Re}}$. This is the content of the next lemma.

Lemma 5.4 *Let the complex function $\widetilde{f}(w)$ be analytic in $\mathbb{D}$ and continuous on $\overline{\mathbb{D}}$. Denote by a_k^{Re}, b_k^{Re} the Fourier coefficients of the real part $\mathrm{Re}[\widetilde{f}(e^{it})] = \mathrm{Re}[\widehat{f}(t)]$ and by a_k^{Im}, b_k^{Im} the Fourier coefficients of the imaginary part $\mathrm{Im}[\widetilde{f}(e^{it})] = \mathrm{Im}[\widehat{f}(t)]$. Then the Fourier coefficients satisfy*

$$a_k^{\mathrm{Re}} = b_k^{\mathrm{Im}}\,, \qquad b_k^{\mathrm{Re}} = -a_k^{\mathrm{Im}}\,, \qquad k \geq 1\,.$$

Proof. For $k \geq 1$, the function $\widetilde{f}(w)w^{k-1} = f\big(T^{-1}(w)\big)w^{k-1}$ is holomorphic in $\mathbb{D}$ and continuous on $\overline{\mathbb{D}}$. By Cauchy's integral theorem, the integral of $\widetilde{f}(w)w^{k-1}$ over the closed curve $\partial\mathbb{D}$ is zero. This yields

$$\begin{aligned}
0 &= \int_{\partial\mathbb{D}} \widetilde{f}(w)w^{k-1}\,dw = \int_0^{2\pi} \widetilde{f}(e^{it})ie^{ikt}\,dt \\
&= \int_0^{2\pi} \big(\mathrm{Re}[\widehat{f}(t)] + i\,\mathrm{Im}[\widehat{f}(t)]\big)\big(i\cos(kt) - \sin(kt)\big)\,dt \\
&= -\int_0^{2\pi} \mathrm{Re}[\widehat{f}(t)]\sin(kt) + \mathrm{Im}[\widehat{f}(t)]\cos(kt)\,dt \\
&\quad + i\int_0^{2\pi} \mathrm{Re}[\widehat{f}(t)]\cos(kt) - \mathrm{Im}[\widehat{f}(t)]\sin(kt)\,dt \\
&= -\pi\,(b_k^{\mathrm{Re}} + a_k^{\mathrm{Im}}) + i\,\pi\,(a_k^{\mathrm{Re}} - b_k^{\mathrm{Im}})\,.
\end{aligned}$$

This implies $a_k^{\mathrm{Re}} = b_k^{\mathrm{Im}}$ and $b_k^{\mathrm{Re}} = -a_k^{\mathrm{Im}}$, which proves the assertion. ❑

With regard to Theorem 5.3, the idea is now to choose the coefficients c_k in the approximation problem $\max_{t\in\mathbb{T}} |\widehat{f}(t) - \sum_{k=0}^{m-1} c_k e^{itk}|$ as

$$\begin{aligned}
&c_0^* = a_0 - ib_0\,, \quad c_k^* = D_k(m,n)(a_k - ib_k)\,, \\
&a_0 = \frac{a_0^{\mathrm{Re}}}{2}\,, \quad b_0 = -\frac{a_0^{\mathrm{Im}}}{2}\,, \quad a_k = a_k^{\mathrm{Re}}\,, \quad b_k = b_k^{\mathrm{Re}} \quad \text{for} \quad 1 \leq k \leq m-1\,,
\end{aligned} \tag{5.7}$$

where $D_k(m,n)$ are the coefficients of the trigonometric polynomial $J_{m,n}$ in Theorem 5.3. This enables us to separate the approximation problem for $\widehat{f}(t)$, $t \in \mathbb{T}$, into two independent subproblems for the real and imaginary part in the following way:

$$\begin{aligned}
\inf_{c_k}\max_{t\in\mathbb{T}}\Big|\widehat{f}(t) - \sum_{k=0}^{m-1} c_k e^{itk}\Big| &\leq \max_{t\in\mathbb{T}}\Big|\widehat{f}(t) - \sum_{k=0}^{m-1} c_k^* e^{itk}\Big| \\
&\leq \max_{t\in\mathbb{T}}\Big|\widehat{f}(t) - (a_0 - ib_0) - \sum_{k=1}^{m-1} D_k(m,n)(a_k - ib_k)\big(\cos(kt) + i\sin(kt)\big)\Big| \\
&= \max_{t\in\mathbb{T}}\Big|\mathrm{Re}[\widehat{f}(t)] - a_0 - \sum_{k=1}^{m-1} D_k(m,n)\big(a_k\cos(kt) + b_k\sin(kt)\big) \\
&\quad + i\,\mathrm{Im}[\widehat{f}(t)] + ib_0 - i\sum_{k=1}^{m-1} D_k(m,n)\big(-b_k\cos(kt) + a_k\sin(kt)\big)\Big|
\end{aligned}$$

$$\leq \max_{t\in\mathbb{T}}\Big|\mathrm{Re}[\widehat{f}(t)] - \frac{a_0^{\mathrm{Re}}}{2} - \sum_{k=1}^{m-1} D_k(m,n)\big(a_k^{\mathrm{Re}}\cos(kt) + b_k^{\mathrm{Re}}\sin(kt)\big)\Big| \tag{5.8}$$

$$+ \max_{t\in\mathbb{T}}\Big|\mathrm{Im}[\widehat{f}(t)] - \frac{a_0^{\mathrm{Im}}}{2} - \sum_{k=1}^{m-1} D_k(m,n)\big(a_k^{\mathrm{Im}}\cos(kt) + b_k^{\mathrm{Im}}\sin(kt)\big)\Big|\,, \tag{5.9}$$

where we have used Lemma 5.4 and (5.7) for the last inequality. The application of

Theorem 5.3 to the terms (5.8) and (5.9) finally gives

$$\inf_{c_k} \max_{t\in\mathbb{T}} \left| \widehat{f}(t) - \sum_{k=0}^{m-1} c_k e^{itk} \right| \leq \frac{C(n)}{m^n} \left[\omega\left(\mathrm{Re}[\,\widehat{f}\,]^{(n)}, \frac{1}{m}\right) + \omega\left(\mathrm{Im}[\,\widehat{f}\,]^{(n)}, \frac{1}{m}\right)\right] , \tag{5.10}$$

if $\widehat{f}$ is n times continuously differentiable. The moduli for the real and imaginary part in (5.10) can be combined to one modulus of continuity for $\widehat{f}^{(n)}$ on $\mathbb{T}$.

Lemma 5.5 *The two moduli of continuity in* (5.10) *fulfill the inequality*

$$\omega\left(\mathrm{Re}[\,\widehat{f}\,]^{(n)}, \frac{1}{m}\right) + \omega\left(\mathrm{Im}[\,\widehat{f}\,]^{(n)}, \frac{1}{m}\right) \leq 2\,\omega\left(\widehat{f}^{(n)}, \frac{1}{m}\right) ,$$

where the last modulus $\omega(\widehat{f}^{(n)}, \frac{1}{m})$ has to be understood as generalization of Definition 5.2 to complex-valued functions $g : \mathbb{R} \to \mathbb{C}$.

Proof. Because of

$$\left|\mathrm{Re}[\widehat{f}(t+h)]^{(n)} - \mathrm{Re}[\widehat{f}(t)]^{(n)}\right| \leq \left|\widehat{f}^{(n)}(t+h) - \widehat{f}^{(n)}(t)\right| ,$$

$$\left|\mathrm{Im}[\widehat{f}(t+h)]^{(n)} - \mathrm{Im}[\widehat{f}(t)]^{(n)}\right| \leq \left|\widehat{f}^{(n)}(t+h) - \widehat{f}^{(n)}(t)\right|$$

for all $t \in \mathbb{T}$ and $0 < h \leq \frac{1}{m}$, we conclude that

$$\underbrace{\sup_{0<h\leq\frac{1}{m}} \max_{t\in\mathbb{T}} \left|\mathrm{Re}[\widehat{f}(t+h)]^{(n)} - \mathrm{Re}[\widehat{f}(t)]^{(n)}\right|}_{=\,\omega(\mathrm{Re}[\,\widehat{f}\,]^{(n)}, \frac{1}{m})} \leq \underbrace{\sup_{0<h\leq\frac{1}{m}} \max_{t\in\mathbb{T}} \left|\widehat{f}^{(n)}(t+h) - \widehat{f}^{(n)}(t)\right|}_{=\,\omega(\widehat{f}^{(n)}, \frac{1}{m})} ,$$

$$\underbrace{\sup_{0<h\leq\frac{1}{m}} \max_{t\in\mathbb{T}} \left|\mathrm{Im}[\widehat{f}(t+h)]^{(n)} - \mathrm{Im}[\widehat{f}(t)]^{(n)}\right|}_{=\,\omega(\mathrm{Im}[\,\widehat{f}\,]^{(n)}, \frac{1}{m})} \leq \underbrace{\sup_{0<h\leq\frac{1}{m}} \max_{t\in\mathbb{T}} \left|\widehat{f}^{(n)}(t+h) - \widehat{f}^{(n)}(t)\right|}_{=\,\omega(\widehat{f}^{(n)}, \frac{1}{m})} .$$

This yields the desired inequality. ❏

For practical computations, it is in general easier to estimate the moduli for the real and imaginary part separately instead of investigating the modulus of continuity for the whole function $\widehat{f}^{(n)}$ on $\mathbb{T}$.

How the inequality (5.10) for the best trigonometric approximation of $\widehat{f}(t)$ can be used to bound the Krylov subspace approximation error for functions $f \in \mathcal{M}$ will be discussed in the next section. Furthermore, we will estimate the two moduli of continuity of the real and imaginary part for the special case $\widehat{f}(t) = \widehat{\varphi}_\ell(t)$, $\ell \geq 1$.

5.2 Error bounds for $f \in \mathcal{M}$

Let $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ and $f \in \mathcal{M}$. We summarize the ideas of the previous section and conclude that the best approximation of $f(\boldsymbol{A})$ in the rational matrix subspace $\mathcal{R}_m(\boldsymbol{A})$ can be bounded by

$$\begin{aligned}\inf_{r\in\frac{\mathcal{P}_{m-1}}{q_{m-1}}}\|f(\boldsymbol{A})-r(\boldsymbol{A})\| &\le \inf_{c_k}\sup_{z\in\mathbb{C}_0^-}\left|f(z)-\sum_{k=0}^{m-1}c_k\left(\frac{1+\frac{z}{\gamma}}{1-\frac{z}{\gamma}}\right)^k\right| \\ &\le \inf_{c_k}\max_{t\in\mathbb{T}}\left|\widehat{f}(t)-\sum_{k=0}^{m-1}c_k e^{itk}\right| \le \max_{t\in\mathbb{T}}\left|\widehat{f}(t)-\sum_{k=0}^{m-1}c_k^* e^{itk}\right| \le \frac{C(n)}{m^n}\,\omega\left(\widehat{f}^{(n)},\frac{1}{m}\right),\end{aligned} \tag{5.11}$$

whenever $\widehat{f}(t)\in C^n(\mathbb{T})$. Hereby, the coefficients c_k^* are chosen as suggested in (5.7). Recall that $\widehat{f}(t)$ is defined as $f\big(T^{-1}(e^{it})\big)$, where T designates the Möbius transformation which maps the left complex half-plane onto the unit disk. The estimate (5.11) is, in particular, valid for the special rational function

$$r^*(z)=\sum_{k=0}^{m-1}c_k^*\left(\frac{1+\frac{z}{\gamma}}{1-\frac{z}{\gamma}}\right)^k\in\frac{\mathcal{P}_{m-1}}{q_{m-1}}\,. \tag{5.12}$$

With these considerations we now derive an upper bound for the error $\|f(\boldsymbol{A})\boldsymbol{v}-f(\boldsymbol{A}_m)\boldsymbol{v}\|$ of the shift-and-invert Krylov subspace approximation, where $\boldsymbol{A}_m$ is given by $\boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m$ and $\boldsymbol{P}_m$ represents the orthogonal projection onto $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})=\mathcal{K}_m\big((\gamma\boldsymbol{I}-\boldsymbol{A})^{-1},\boldsymbol{v}\big)$. But first of all, we must guarantee that $W(\boldsymbol{A}_m)\subseteq\mathbb{C}_0^-$, such that the projected matrix $\boldsymbol{A}_m$ actually fits in our framework.

Lemma 5.6 *If $W(\boldsymbol{A})\subseteq\mathbb{C}_0^-$, the restriction $\boldsymbol{A}_m=\boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m$ of the matrix $\boldsymbol{A}\in\mathbb{C}^{N\times N}$ to the subspace $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ satisfies $W(\boldsymbol{A}_m)\subseteq\mathbb{C}_0^-$ as well.*

Proof. The relation $W(\boldsymbol{A}_m)\subseteq\mathbb{C}_0^-$ is equivalent to $\mathrm{Re}(\boldsymbol{A}_m\boldsymbol{x},\boldsymbol{x})\le 0$ for all $\boldsymbol{x}\in\mathbb{C}^N$. Setting $\boldsymbol{y}=\boldsymbol{P}_m\boldsymbol{x}$ and using the fact that $\boldsymbol{P}_m$ is self-adjoint, cf. Definition 4.1, it follows

$$\mathrm{Re}(\boldsymbol{A}_m\boldsymbol{x},\boldsymbol{x})=\mathrm{Re}(\boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m\boldsymbol{x},\boldsymbol{x})=\mathrm{Re}(\boldsymbol{A}\boldsymbol{P}_m\boldsymbol{x},\boldsymbol{P}_m\boldsymbol{x})=\mathrm{Re}(\boldsymbol{A}\boldsymbol{y},\boldsymbol{y})\le 0\,,$$

since $W(\boldsymbol{A})\subseteq\mathbb{C}_0^-$ by assumption. ❑

Theorem 5.7 *Let $\boldsymbol{A}$ be a matrix with $W(\boldsymbol{A})\subseteq\mathbb{C}_0^-$ and suppose $f\in\mathcal{M}$. Moreover, for $\widehat{f}(t)=f\big(T^{-1}(e^{it})\big)$, we assume that the nth derivative $\widehat{f}^{(n)}$ exists and is continuous on $\mathbb{T}$. Then the error of the approximation $f(\boldsymbol{A}_m)\boldsymbol{v}$ to $f(\boldsymbol{A})\boldsymbol{v}$ in the shift-and-invert Krylov subspace $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ is bounded by*

$$\|f(\boldsymbol{A})\boldsymbol{v}-f(\boldsymbol{A}_m)\boldsymbol{v}\|\le 2\,\frac{C(n)}{m^n}\,\omega\left(\widehat{f}^{(n)},\frac{1}{m}\right)\|\boldsymbol{v}\|\,,$$

where $\boldsymbol{A}_m=\boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m$ and $\boldsymbol{P}_m$ is the orthogonal projection onto $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$.

Proof. Because of Lemma 5.6 and Lemma 4.12, we know that $W(\boldsymbol{A}_m)\subseteq\mathbb{C}_0^-$ and that

$$r(\boldsymbol{A})\boldsymbol{v}=r(\boldsymbol{A}_m)\boldsymbol{v}\quad\text{for every}\quad r\in\frac{\mathcal{P}_{m-1}}{q_{m-1}}\quad\text{with}\quad q_{m-1}(z)=(\gamma-z)^{m-1}\,.$$

We now consider the special rational function r^* defined in (5.12). Then $r^*(\boldsymbol{A})\boldsymbol{v}$ belongs to $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$ and we have $r^*(\boldsymbol{A})\boldsymbol{v}=r^*(\boldsymbol{A}_m)\boldsymbol{v}$. Since r^* is determined solely by f and not by the matrices $\boldsymbol{A}$ or $\boldsymbol{A}_m$, the error bound (5.11) can be applied to $\|f(\boldsymbol{A}) - r^*(\boldsymbol{A})\|$ as well as to $\|f(\boldsymbol{A}_m) - r^*(\boldsymbol{A}_m)\|$. Using the triangle inequality and (5.11), we are now able to conclude that

$$\begin{aligned}\|f(\boldsymbol{A})\boldsymbol{v} - f(\boldsymbol{A}_m)\boldsymbol{v}\| &\leq \|f(\boldsymbol{A}) - r^*(\boldsymbol{A})\| \, \|\boldsymbol{v}\| + \|f(\boldsymbol{A}_m) - r^*(\boldsymbol{A}_m)\| \, \|\boldsymbol{v}\| \\ &\leq 2 \, \frac{C(n)}{m^n} \, \omega \left(\widehat{f}^{(n)}, \frac{1}{m} \right) \|\boldsymbol{v}\|\end{aligned}$$

holds true. ❑

In a nutshell, this theorem says that the convergence rate of the shift-and-invert Krylov subspace approximation for $f(\boldsymbol{A})\boldsymbol{v}$, $f \in \mathcal{M}$, depends only on the smoothness properties of the transformed function $f\big(T^{-1}(w)\big)$ on the boundary of the unit circle. The obtained error bound is completely independent of $\|\boldsymbol{A}\|$. Consequently, we have a uniform convergence for all matrices $\boldsymbol{A}$ with a field of values somewhere in the left complex half-plane.

5.3 Error bounds for the φ-functions

The findings of Section 5.2 raise the question what is to be expected for $f = \varphi_\ell$. For $\ell \geq 1$, the modulus $|\varphi_\ell(z)|$ tends to zero for all $|z| \to \infty$ with $\mathrm{Re}(z) \leq 0$, so that $\varphi_\ell \in \mathcal{M}$ with $c = 0$. In contrast, we have $\varphi_0 \notin \mathcal{M}$, since $|e^z| \to 0$ for $\mathrm{Re}(z) \to -\infty$, whereas $|e^z| = 1$ on the imaginary axis. Thus, Theorem 5.7 can only be used for the φ_ℓ-functions with $\ell \geq 1$. According to the previous considerations, we have to analyze the real and imaginary part of $\widehat{\varphi}_\ell(t) = \varphi_\ell\big(T^{-1}(e^{it})\big)$ with respect to their differentiability. The next lemma will be helpful for this purpose.

Lemma 5.8 *Both functions $x \sin\left(\frac{1}{x}\right)$ and $x \cos\left(\frac{1}{x}\right)$, $x \in \mathbb{R}$, are $\frac{1}{2}$-Hölder continuous on any bounded domain.*

Proof. We show the assertion only for the first function $x \sin\left(\frac{1}{x}\right)$, since the proof for the second function $x \cos\left(\frac{1}{x}\right)$ can be conducted analogously. We start with

$$\left| y \sin\left(\frac{1}{y}\right) - x \sin\left(\frac{1}{x}\right) \right|^2 = \left| y^2 \sin^2\left(\frac{1}{y}\right) - 2xy \sin\left(\frac{1}{y}\right) \sin\left(\frac{1}{x}\right) + x^2 \sin^2\left(\frac{1}{x}\right) \right| .$$

Due to $2xy = x^2 + y^2 - (y-x)^2$ and $\left|\sin\left(\frac{1}{x}\right)\right| \leq 1$, we get with the triangle inequality

$$\left| y \sin\left(\frac{1}{y}\right) - x \sin\left(\frac{1}{x}\right) \right|^2 \leq 2 \left| y^2 \sin\left(\frac{1}{y}\right) - x^2 \sin\left(\frac{1}{x}\right) \right| + |y - x|^2 .$$

On a bounded domain, the function $x^2 \sin\left(\frac{1}{x}\right)$ is Lipschitz continuous, because its derivative is bounded. It thus follows

$$\left| y \sin\left(\frac{1}{y}\right) - x \sin\left(\frac{1}{x}\right) \right| \leq C \, |y - x|^{\frac{1}{2}}$$

by taking the square root. ❑

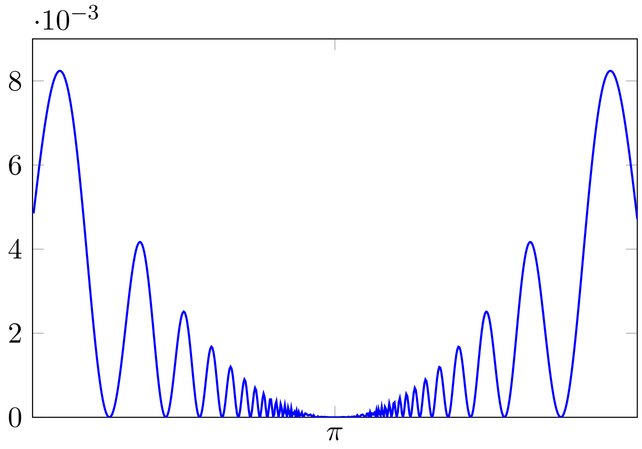


Figure 5.2: Plot of $\mathrm{Re}[\widehat{\varphi}_\ell(t)]$ for $t \in [3, 3.28]$, $\ell = 2$, and $\gamma = 1$.

Theorem 5.9 *Let $\boldsymbol{A}$ satisfy $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ and let $\boldsymbol{P}_m$ be the orthogonal projection onto the shift-and-invert Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$. For the restriction $\boldsymbol{A}_m = \boldsymbol{P}_m \boldsymbol{A} \boldsymbol{P}_m$ of $\boldsymbol{A}$ to $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$, we have the error bound*

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\| \le \frac{C(\ell,\gamma)}{m^{\frac{\ell}{2}}} \|\boldsymbol{v}\|, \qquad \ell \ge 1.$$

Proof. Due to Theorem 5.7, our task is to study the modulus of continuity $\omega(\widehat{\varphi}_\ell^{(n)}, \frac{1}{m})$. This is equivalent to examine separately the real and imaginary part of $\widehat{\varphi}_\ell(t)$ with respect to its smoothness. Here, we restrict ourselves to the analysis of $\mathrm{Re}[\widehat{\varphi}_\ell(t)]$. The imaginary part can be estimated analogously. Before we start our investigation, it is worthwhile to mention that

$$\frac{e^{it} - 1}{e^{it} + 1} = i \tan\left(\frac{t}{2}\right),$$

so that $\widehat{\varphi}_\ell(t) = \varphi_\ell\left(i\gamma \tan\left(\frac{t}{2}\right)\right)$. As composition of infinitely many times differentiable functions on $\mathbb{T}\backslash\{\pi\}$, $\mathbb{T} = [0, 2\pi)$, the function $\mathrm{Re}[\widehat{\varphi}_\ell(t)] = \mathrm{Re}\big[\varphi_\ell\big(i\gamma\tan\left(\frac{t}{2}\right)\big)\big]$ is also infinitely often differentiable for $t \neq \pi$. For this reason, we analyze $\mathrm{Re}[\widehat{\varphi}_\ell(t)]$ on the interval $\mathbb{T}_\pi := (\pi - \sigma, \pi + \sigma)$ for $0 < \sigma < \pi$, cf. Figure 5.2. Hereby, it is necessary to distinguish whether the index ℓ of the φ_ℓ-function is odd or even. In the following, we will analyze these two cases one after another.

(i) ℓ even: Using the representation (3.13) for $\varphi_\ell\big(i\gamma\tan\left(\frac{t}{2}\right)\big)$, we find

$$\mathrm{Re}[\widehat{\varphi}_\ell(t)] = (-1)^{\frac{\ell}{2}} \left[\frac{\cos\left(\gamma \tan\left(\frac{t}{2}\right)\right)}{\left(\gamma\tan\left(\frac{t}{2}\right)\right)^\ell} - \sum_{k=0}^{\frac{\ell}{2}-1} (-1)^k \frac{\gamma^{2k-\ell}}{(2k)!} \tan^{2k-\ell}\left(\frac{t}{2}\right) \right].$$

There is no need to worry about the second term, containing the sum over k, since $\tan^{2k-\ell}\left(\frac{t}{2}\right)$ is infinitely often differentiable on $\mathbb{T}_\pi$ for all $k \in \{0, \ldots, \frac{\ell}{2} - 1\}$. Differentiating the first term with respect to t,

$$\frac{d}{dt}\frac{\cos\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\left(\gamma\tan\left(\frac{t}{2}\right)\right)^{\ell}} = -\frac{1}{2}\left[\ell\gamma\frac{\cos\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\left(\gamma\tan\left(\frac{t}{2}\right)\right)^{\ell+1}} + \gamma\frac{\sin\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\left(\gamma\tan\left(\frac{t}{2}\right)\right)^{\ell}}\right.$$
$$\left. + \frac{\ell}{\gamma}\frac{\cos\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\left(\gamma\tan\left(\frac{t}{2}\right)\right)^{\ell-1}} + \frac{1}{\gamma}\frac{\sin\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\left(\gamma\tan\left(\frac{t}{2}\right)\right)^{\ell-2}}\right],$$

we see that we obtain a linear combination of similar expressions, where the numerator is partly replaced by $\sin\left(\gamma\tan\left(\frac{t}{2}\right)\right)$ and where the exponent of the denominator has changed by $+1$, 0, -1 or -2. The different terms are continuously differentiable as long as the exponent in the denominator is greater than 2. The functions $\sin\left(\gamma\tan\left(\frac{t}{2}\right)\right)/\tan^2\left(\frac{t}{2}\right)$ and $\cos\left(\gamma\tan\left(\frac{t}{2}\right)\right)/\tan^2\left(\frac{t}{2}\right)$ have no continuous derivative, but their derivatives are bounded on $\mathbb{T}_\pi$. This shows that $\mathrm{Re}[\widehat{\varphi}_\ell(t)] \in C^{\frac{\ell}{2}-1}(\mathbb{T}_\pi)$ and

$$\omega\left(\mathrm{Re}[\,\widehat{\varphi}_\ell\,]^{(\frac{\ell}{2}-1)}, \frac{1}{m}\right) \leq \frac{C(\ell,\gamma)}{m},$$

using the mean value theorem with $|\mathrm{Re}[\widehat{\varphi}_\ell(\xi)]^{(\frac{\ell}{2})}| \leq C(\ell,\gamma)$ for all $\xi \in \mathbb{T}_\pi$.

(ii) ℓ odd: Since now

$$\mathrm{Re}[\widehat{\varphi}_\ell(t)] = (-1)^{\frac{\ell-1}{2}}\left[\frac{\sin\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\left(\gamma\tan\left(\frac{t}{2}\right)\right)^{\ell}} + \sum_{k=0}^{\frac{\ell-1}{2}-1}(-1)^{k+1}\frac{\gamma^{2k-\ell+1}}{(2k+1)!}\tan^{2k-\ell+1}\left(\frac{t}{2}\right)\right],$$

the same considerations as in (i) apply with the exception that the first term $\sin\left(\gamma\tan\left(\frac{t}{2}\right)\right)/\tan^{\ell}\left(\frac{t}{2}\right)$ can only be differentiated continuously $\lfloor\frac{\ell}{2}\rfloor$ times for $t \in \mathbb{T}_\pi$, until we get expressions of the form

$$g_1(t) := \frac{\sin\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\gamma\tan\left(\frac{t}{2}\right)} \qquad \text{or} \qquad g_2(t) := \frac{\cos\left(\gamma\tan\left(\frac{t}{2}\right)\right)}{\gamma\tan\left(\frac{t}{2}\right)}.$$

The application of Lemma 5.8 with $x = 1/\left(\gamma\tan\left(\frac{t}{2}\right)\right)$ and $y = 1/\left(\gamma\tan\left(\frac{t+1/m}{2}\right)\right)$ gives for $j \in \{1,2\}$ the estimate

$$\left|g_j\left(t+\frac{1}{m}\right) - g_j(t)\right| \leq C\left|\frac{1}{\gamma\tan\left(\frac{t+1/m}{2}\right)} - \frac{1}{\gamma\tan\left(\frac{t}{2}\right)}\right|^{\frac{1}{2}} \leq \frac{C(\gamma)}{\sqrt{m}}, \qquad t \in \mathbb{T}_\pi,$$

where the second inequality follows by the mean value theorem. This yields

$$\omega\left(\mathrm{Re}[\,\widehat{\varphi}_\ell\,]^{\lfloor\frac{\ell}{2}\rfloor}, \frac{1}{m}\right) \leq \frac{C(\ell,\gamma)}{\sqrt{m}}.$$

Furthermore, note that $m^{-\lfloor\ell/2\rfloor}m^{-1/2} = m^{-\ell/2}$.

Together with Theorem 5.7 and a similar estimate for $\mathrm{Im}[\widehat{\varphi}_\ell(t)]$, the statement of the theorem is proved. ❑

For the best approximation of the matrix function $\varphi_\ell(\boldsymbol{A})$ in the rational matrix space $\mathcal{R}_m(\boldsymbol{A})$, the proof of Theorem 5.9 shows, in particular, that

$$\inf_{r\in\frac{\mathcal{P}_{m-1}}{q_{m-1}}} \|\varphi_\ell(\boldsymbol{A}) - r(\boldsymbol{A})\| \leq \|\varphi_\ell(\boldsymbol{A}) - r^*(\boldsymbol{A})\| \leq \frac{C(\ell,\gamma)}{m^{\frac{\ell}{2}}}, \tag{5.13}$$

where r^* is the rational function $r^*(z) = \sum_{k=0}^{m-1} c_k^* \bigl(\frac{1+z/\gamma}{1-z/\gamma}\bigr)^k$ in (5.12) with coefficients c_k^* defined by (5.7).

Remark 5.10 When applying numerical methods for the time integration of evolution equations, one is interested in the approximation of $\varphi_\ell(\tau\boldsymbol{A})\boldsymbol{v}$, where $\tau > 0$ denotes the step size in time. Since

$$\mathrm{Re}\bigl((\tau\boldsymbol{A})\boldsymbol{x},\boldsymbol{x}\bigr) = \tau\,\mathrm{Re}(\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}) \le 0 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{C}^N\,,$$

we have $W(\tau\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ in the case that $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$. As a result, replacing $\boldsymbol{A}$ by $\tau\boldsymbol{A}$ leads to exactly the same scalar approximation problem on the unit disk independent of the step size τ. This has the far-reaching consequence that all of the above theorems remain valid with the same constants for $\tau\boldsymbol{A}$ instead of $\boldsymbol{A}$, that is,

$$\|\varphi_\ell(\tau\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\tau\boldsymbol{A}_m)\boldsymbol{v}\| \le \frac{C(\ell,\gamma)}{m^{\frac{\ell}{2}}}\,\|\boldsymbol{v}\|\,.$$

5.4 Choice of the shift γ

So far, the shift $\gamma > 0$ has been regarded as a free parameter of the rational matrix subspace $\mathcal{R}_m(\boldsymbol{A})$ and the shift-and-invert Krylov subspace $\mathcal{Q}_m(\boldsymbol{A},\boldsymbol{v})$. This section aims to discuss a strategy for a best possible choice of γ. To this end, we will choose γ depending on the dimension m of the approximation space. In order to obtain suitable conditions for γ, which raise the asymptotic convergence rate, we need another modulus of smoothness, the rth modulus of smoothness $\omega_r(g,\delta)$. In contrast to the previous sections, where we used $\omega(g^{(n)},\delta)$ for $g \in C^n(\mathbb{T})$, this modulus is given by the following definition (see [14]).

Definition 5.11 *For $\delta > 0$, the rth modulus of smoothness $\omega_r(g,\delta)$ of a continuous function g is defined as*

$$\omega_r(g,\delta) := \sup_{0<h\le\delta} \max_{t\in\mathbb{T}} |\Delta_h^r(g,t)|\,, \qquad \Delta_h^r(g,t) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} g(t+kh)\,, \tag{5.14}$$

where $\Delta_h^r(g,t)$ is an rth order difference.

The first modulus of continuity $\omega_1(g,\delta)$ coincides with the modulus of continuity $\omega(g,\delta)$ in Definition 5.2. Compared to $\omega(g,\delta)$, which only measures the continuity properties of the function g, the modulus $\omega_r(g,\delta)$ is useful for measuring higher smoothness. Furthermore, we have the relations

$$\omega_r(g,\delta) \le \delta^r \max_{t\in\mathbb{T}} |g^{(r)}(t)| \qquad \text{and} \qquad \omega_{r+k}(g,\delta) \le \delta^r \omega_k(g^{(r)},\delta)\,, \tag{5.15}$$

see [14], p. 46.

In order to measure the quality of approximation for a 2π-periodic real-valued function g by its mth partial Fourier sum $S_m g(t)$, we make use of the Jackson-Stechkin inequality for the best approximation of g by a trigonometric polynomial $P \in \mathcal{T}_m$ (see, e.g., [14], Chapter 7, Theorem 2.3). The Jackson-Stechkin inequality reads

$$\inf_{P\in\mathcal{T}_m} \max_{t\in\mathbb{T}} |g(t) - P(t)| \le C(r)\,\omega_r\left(g, \frac{1}{m}\right)\,, \tag{5.16}$$

where $\mathcal{T}_m$ denotes the set of trigonometric polynomials of degree m. With the help of this inequality, we can formulate the following lemma.

Lemma 5.12 *The approximation of a real 2π-periodic and continuous function g by its mth partial Fourier sum $S_m g$ is bounded by*

$$|g(t) - S_m g(t)| \leq C(r)\, \ln(m)\, \omega_r\left(g, \frac{1}{m}\right) . \tag{5.17}$$

Proof. A relation between the approximation by the mth partial Fourier sum and the best trigonometric approximation can be found as Proposition 1.5.2 in Pinsky [66]. This proposition states

$$|g(t) - S_m g(t)| \leq \big(5 + \ln(m)\big) \inf_{P \in \mathcal{T}_m} \max_{t \in \mathbb{T}} |g(t) - P(t)| .$$

Due to the Jackson-Stechkin inequality (5.16), we obtain the desired result. ❏

The same ideas as in Section 5.1 for the estimate of $\inf_{c_k} \max_{t\in\mathbb{T}} |\widehat{f}(t) - \sum_{k=0}^{m-1} c_k e^{itk}|$ on page 60, but this time with $D_k(m,n) = 1$, lead for $f \in \mathcal{M}$ to the bound

$$\|f(\boldsymbol{A})\boldsymbol{v} - f(\boldsymbol{A}_m)\boldsymbol{v}\| \leq C(r)\ln(m)\left[\omega_r\left(\mathrm{Re}[\,\widehat{f}\,], \frac{1}{m}\right) + \omega_r\left(\mathrm{Im}[\,\widehat{f}\,], \frac{1}{m}\right)\right]\|\boldsymbol{v}\| . \tag{5.18}$$

Since we approximate the real and imaginary part of the function $\widehat{f}(t)$ by a trigonometric polynomial of degree less than or equal to $m-1$, it should read $m-1$ instead of m everywhere in inequality (5.18). It is nevertheless possible to write m, if we slightly change the generic constant $C(r)$.

In the case that $\widehat{f}$ is n-times continuously differentiable, we immediately get from (5.15) for $r = n$, $k = 1$ and Lemma 5.5 that

$$\|f(\boldsymbol{A})\boldsymbol{v} - f(\boldsymbol{A}_m)\boldsymbol{v}\| \leq \frac{C(n)}{m^n} \ln(m)\, \omega\left(\widehat{f}^{(n)}, \frac{1}{m}\right) \|\boldsymbol{v}\| .$$

Compared to Theorem 5.7, we have an additional factor $\ln(m)$. This is caused by setting the constants $D_k(m,n)$ in the trigonometric approximation equal to one and by considering the approximation of $\mathrm{Re}[\widehat{f}(t)]$ and $\mathrm{Im}[\widehat{f}(t)]$ by its $(m-1)$st partial Fourier sum, which is only the best trigonometric approximation in the L^2-norm, but not in the maximum norm. However, we do not have to worry about this extra factor, since $\ln(m)$ grows slower than any positive power m^β with $\beta > 0$.

We now aim to find an optimal shift γ for the case $f = \varphi_\ell$, $\ell \geq 1$, in (5.18). Of course, the estimate of $\omega_r(g,\delta)$ requires more effort than the estimate of $\omega(g^{(n)},\delta)$, since differences $\Delta_h^r(g,t)$ of order r are involved. These differences have to be bounded in a suitable manner and necessitate several case distinctions. Thus, one may wonder why we should now work with this apparently more complicated modulus of smoothness $\omega_r(g,\delta)$. The reason for this is the following: If we use the Jackson-Stechkin bound (5.16), we can conclude how the constants occurring in the estimates of the two moduli $\omega_r(\mathrm{Re}[\,\widehat{\varphi}_\ell\,], \frac{1}{m})$ and $\omega_r(\mathrm{Im}[\,\widehat{\varphi}_\ell\,], \frac{1}{m})$ depend on the parameters γ and r. From this, it is possible to deduce an appropriate parameter γ. In contrast, by analyzing the constants $C(\ell,\gamma)$ in the error estimates of the previous section, where the modulus $\omega(g^{(n)},\delta)$ for a function $g \in C^n(\mathbb{T})$ was used, we could not gain any insight into a suitable choice of γ.

For the shift γ, we choose the ansatz $\gamma(m) := m^\alpha$ with $\alpha \in \mathbb{R}$. This choice enables us to derive a condition for the parameter α, depending on ℓ and r, that raises our asymptotic

ℓ	i	j
ℓ even	$\frac{\ell}{2}$	$\frac{\ell}{2}, \dots, \ell$
	r	$1, \dots, \ell$
ℓ odd	$\frac{\ell}{2}$	$\frac{\ell}{2}$
	$\lfloor \frac{\ell}{2} \rfloor + 1$	$\lfloor \frac{\ell}{2} \rfloor, \dots, \ell$
	r	$1, \dots, \ell$

Table 5.1: Possible values of i and j in (5.19).

convergence rate of order $\mathcal{O}(m^{-\ell/2})$. The key advantage of the modulus $\omega_r(g, \delta)$ used here is that the parameter r is not bounded by the smoothness of $g \in C^n(\mathbb{T})$ and can thus be chosen greater than n.

The estimates of the two moduli $\omega_r(\mathrm{Re}[\widehat{\varphi}_\ell], \delta)$ and $\omega_r(\mathrm{Im}[\widehat{\varphi}_\ell], \delta)$ are not very difficult, since they partially rely on the ideas of the previous section. However, they are quite tedious and require a case distinction whether r, ℓ and $\lfloor \ell/2 \rfloor$ are even or odd. This is why we do not present here a detailed derivation of the results stated below.

The main ideas are as follows: As in the proof of Theorem 5.9, one has to treat separately the two cases $t \in \mathbb{T}_\pi = (\pi - \sigma, \pi + \sigma)$, $0 < \sigma < \pi$, and $t \in \mathbb{T} \backslash \{\pi\}$. In the first case, the same representation of $\mathrm{Re}[\widehat{\varphi}_\ell(t)]$ and $\mathrm{Im}[\widehat{\varphi}_\ell(t)]$ via formula (3.13) can be used. The occurring rth order differences are estimated by a suitable Taylor expansion. The calculation reveals that, thereby, the condition $r > \frac{\ell}{2} + 1$ has to be fulfilled. For $t \in \mathbb{T} \backslash \{\pi\}$ on the other hand, we represent $\widehat{\varphi}_\ell(t)$ by formula (3.12), use the bound

$$\max_{t \in [a,b]} |\Delta_h^r g(t)| \le h^r \max_{t \in [a,b]} |g^{(r)}(t)| \quad \text{for} \quad g \in C^r([a,b]) \,, \quad a, b \in \mathbb{R} \,,$$

and Faá die Bruno's formula, a generalization of the chain rule to higher derivatives. Altogether, the analysis shows that terms of the form

$$\frac{C(r)\delta^i}{\gamma^j} \qquad \text{and} \qquad C(r)\delta^r \gamma^k \,, \qquad k \in \{1, \dots, r\} \tag{5.19}$$

appear with a generic constant $C(r)$. Possible values of i and j are summarized in Table 5.1. The decisive prefactors, being most restrictive, are $\delta^{\ell/2}\gamma^{-\ell/2}$ and $\delta^r \gamma^r$. They lead with $\delta = \frac{1}{m}$ and $\gamma = m^\alpha$ to convergence rates of order $m^{-\ell/2 - \alpha\ell/2}$ and $m^{-r+\alpha r}$. By balancing these two expressions, we find the condition

$$-\frac{\ell}{2} - \alpha\frac{\ell}{2} = -r + \alpha r \qquad \text{or} \qquad \alpha = \frac{r - \frac{\ell}{2}}{r + \frac{\ell}{2}} \,.$$

This choice of α leads to the improved convergence rate

$$\mathcal{O}\left(m^{-\frac{\ell}{2}(1+\alpha)}\right) \,,$$

which tends to $\mathcal{O}(m^{-\ell})$ for $r \to \infty$. If we increase r, the rate of convergence becomes faster and faster. But the analysis shows that the occurring constants grow in the worst case like $\big(0.792r/\ln(r+1)\big)^r$. This results from an estimate using Faà di Bruno's formula in combinatorial form and the bound in [5] on Bell numbers. By counting partitions, these numbers indicate how many terms in Faà die Bruno's formula appear. As a result, the parameter r should not be chosen too large, in order to obtain a reasonable error bound.

For example, if we consider the φ_1-function, we have an improved rate of order $\mathcal{O}(m^{-6/7})$ for the choice $r = 3$ compared to $\mathcal{O}(m^{-\ell/2}) = \mathcal{O}(m^{-1/2})$. How the choice of γ affects the approximation behavior is illustrated by the following simple numerical experiment.

Example 5.13 We consider the approximation of $\varphi_4(\boldsymbol{A})\boldsymbol{v}$ in $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$ for a matrix $\boldsymbol{A} \in \mathbb{C}^{1\,000\times 1\,000}$ with $W(\boldsymbol{A}) \subseteq [-100\,i, 100\,i]$ and a random vector $\boldsymbol{v}$ of norm 1, which is generated by the Matlab function `randn`. The shift γ is chosen as one (black solid line) and, according to the discussion above, as $\gamma = m^\alpha$ with $\alpha = \frac{r-2}{r+2}$ for $r = 6, 12, 24$ and $m = 20$. In Figure 5.3, the approximation error is plotted against the number of iteration steps. As expected, we have a better convergence for larger r. The apprehension that overly large r have a negative effect cannot be observed here. Hence, the worst case estimate for $C(r)$ above seems to be too pessimistic. ❍

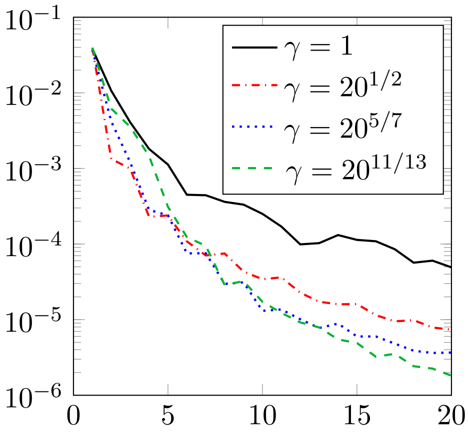


Figure 5.3: Error $\|\varphi_4(\boldsymbol{A})\boldsymbol{v} - \varphi_4(\boldsymbol{A}_m)\boldsymbol{v}\|$ versus m.

Remark 5.14 Probably, a result similar to Achyèser's Theorem 5.3 holds true for the rth modulus of smoothness which could not be found in standard literature. In this case, estimate (5.18) above would hold true without the factor $\ln(m)$.

5.5 The case of operators

The approximation of operator functions $f(A)v$, especially of $\varphi_\ell(A)v$, in the resolvent space $\mathcal{Q}_m(A, v)$ is analyzed in Grimm [29], where A is assumed to be a linear operator on a Hilbert space H satisfying Range$(\lambda I - A) = H$ for some λ with $\mathrm{Re}(\lambda) > 0$ and the dissipativity property $\mathrm{Re}(Ax, x) \le 0$ for every $x \in D(A)$. The notation $\mathcal{Q}_m(A, v)$ means that the matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ and the vector $\boldsymbol{v} \in \mathbb{C}^N$ in the rational Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$ are replaced by the operator A and $v \in H$. Using the Lumer-Phillips Theorem 3.7, we can conclude that A generates a strongly continuous contraction semigroup with $\|e^{\tau A}\| \le 1$ for all $\tau \ge 0$, where $\|\cdot\|$ designates the norm induced by the inner product on H.

We denote by $f_{(0)} : \mathbb{R} \to \mathbb{C}$ the restriction of a function $f : \mathbb{C} \to \mathbb{C}$ to $\mathrm{Re}(z) = 0$, that is, $f_{(0)}(\zeta) = f(i\zeta)$ for $\zeta \in \mathbb{R}$. Moreover, we assume that f is holomorphic and bounded on $\mathbb{C}_0^-$ and that $f_{(0)} \in C(\mathbb{R})$, $\mathcal{F}f_{(0)} \in L^1(\mathbb{R})$ and $\mathrm{supp}(\mathcal{F}f_{(0)}) \subseteq [0, \infty)$, where $\mathcal{F}f_{(0)}$ is the Fourier transform of $f_{(0)}$, given as

$$\mathcal{F}f_{(0)}(s) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixs} f_{(0)}(x)\, dx\,.$$

For a function f, fulfilling the above assumptions, we use the functional calculus (cf. [18])

$$f(A) := \int_0^\infty e^{sA} \mathcal{F} f_{(0)}(s)\, ds\,,$$

which defines a bounded linear operator $f(A)$ with $\|f(A)\| \le \|\mathcal{F} f_{(0)}\|_{L^1(\mathbb{R})}$. One can show that the φ_ℓ-functions, $\ell \ge 1$, satisfy the required assumptions and that $\varphi_\ell(A)$, defined via the functional calculus, coincides with the former definition

$$\varphi_\ell(A) = \int_0^1 e^{(1-s)A} \frac{s^{\ell-1}}{(\ell-1)!}\, ds\,,$$

see (3.12). It is shown in [29] that the best approximation of the operator function $f(A)$ in the resolvent subspace

$$\widetilde{\mathcal{R}}_m(A) = \operatorname{span}\{(\gamma I - A)^{-1}, \dots, (\gamma I - A)^{-m}\}\,, \qquad \gamma > 0$$

has the bound

$$\inf_{R \in \widetilde{\mathcal{R}}_m(A)} \|f(A) - R\| \le C\, \omega_\phi^r\left(\widetilde{f}, \frac{1}{\sqrt{m-1}}\right)\,, \qquad \widetilde{f}(s) = e^s \mathcal{F} f_{(0)}\left(\frac{s}{\gamma}\right)\,,$$

where $r < m-1$ and C is a constant independent of f. The quality of approximation is measured with the so-called weighted ϕ-modulus of smoothness $\omega_\phi^r(g,\delta)$ introduced by De Bonis, Mastroianni and Viggiano in [7] to characterize the K-functional[1] of a function g by its structural properties. The main part of this modulus is defined by

$$\Omega_\phi^r(g,\delta) := \sup_{0<h\le\delta} \|w \Delta_{h\phi}^r(g,\cdot)\|_{L^1[4r^2h^2, \frac{1}{h^2}]}$$

with the rth symmetric difference

$$\Delta_{h\phi}^r(g,t) = \sum_{j=0}^r (-1)^j \binom{r}{j} g\left(t + \frac{h\phi(t)}{2}(r-2j)\right)\,,$$

where $0 < \delta \le 1$, $\phi(t) = \sqrt{t}$ and $w(t) = e^{-t}$ for $t > 0$. For the complete modulus, we now compose $\omega_\phi^r(g,\delta)$ as

$$\omega_\phi^r(g,\delta) := \Omega_\phi^r(g,\delta) + \inf_{p \in \mathcal{P}_{r-1}} \|w(g-p)\|_{L^1(0,4r^2\delta^2)} + \inf_{q \in \mathcal{P}_{r-1}} \|w(g-q)\|_{L^1(\frac{1}{\delta^2},\infty)}\,.$$

A detailed analysis of $\omega_\phi^r\big(\widetilde{\varphi}_\ell, \frac{1}{\sqrt{m-1}}\big)$ for $\widetilde{\varphi}_\ell(s) = e^s \mathcal{F}\varphi_{\ell,(0)}\big(\frac{s}{\gamma}\big)$ and $r = \ell$ shows that

$$\inf_{R \in \widetilde{R}_m(A)} \|\varphi_\ell(A) - R\| \le \frac{C(\ell,\gamma)}{m^{\frac{\ell}{2}}}\,. \tag{5.20}$$

Like above, we define by P_m the orthogonal projection onto the resolvent Krylov subspace $\mathcal{Q}_m(A,v)$ and set $A_m = P_m A P_m$. Then the resolvent Krylov subspace approximation has the error bound

$$\|\varphi_\ell(A)v - \varphi_\ell(A_m)v\| \le \frac{C(\ell,\gamma)}{m^{\frac{\ell}{2}}}\|v\|\,.$$

[1] The general K-functional is given by $K_{r,\phi}(f,\delta^r)_{w,p} = \inf_{g^{(r-1)} \in AC_{\text{loc}}} \|w(f-g)\|_{L^p} + \delta^r \|w\phi^r g^{(r)}\|_{L^p}$, where $0 < \delta \le 1$, $1 \le p \le \infty$, w is a Laguerre weight of the form $w(t) = t^\alpha e^{-t}$ and AC_{loc} is the set of locally absolutely continuous functions.

The bounds here look quite similar to our bounds in Section 5.3 for the matrix case, except for the fact that the constants $C(\ell, \gamma)$ are different. But the modulus of continuity from Definition 5.2, used in the new derivation of this bound in Section 5.3, is more common and much easier to calculate than the weighted ϕ-modulus of smoothness in this section.

In [29], the shift γ has been regarded as a fixed constant. To improve the convergence rate in [29], we can proceed analogously to Section 5.4 to find an optimal γ. For this purpose, we examine by a tedious calculation how the terms occurring in the estimate of $\omega_\phi^r(\widetilde{\varphi}_\ell, \frac{1}{\sqrt{m-1}})$ depend on $\gamma > 0$. Looking for the most restrictive terms and setting again $\gamma = m^\alpha$, $\alpha \in \mathbb{R}$, we get a condition that the parameter α should fulfill, namely $\alpha = \frac{r-\ell}{r+\ell}$ for $\ell < r < m-1$. With this choice, we achieve an upgraded convergence rate of order

$$\mathcal{O}\left(m^{-\frac{\ell}{2}(1+\alpha)}\right), \qquad \alpha = \frac{r-\ell}{r+\ell}.$$

A sketch of the proof can be found in [25]. It should be noted that we do not end up with the same optimal value for γ and the same improved rate as in Section 5.4. But this is not surprising, since we used different moduli of smoothness. Again, we should keep in mind that the predicted convergence is indeed faster for larger values of r, but at the cost of a constant $C(\ell, \gamma)$ in (5.20) that grows like K^r for a fixed constant $K > 0$.

5.6 Numerical experiments

In our first experiment, we consider a simple test example that validates the theoretical bounds and that illustrates the faster convergence for φ_ℓ-functions of larger index ℓ . By the example of a one-dimensional wave equation, the influence of the choice of the shift γ on the convergence rate is demonstrated in Section 5.6.2. Finally, we compare the shift-and-invert Krylov subspace method with the implicit Euler and the Crank-Nicolson scheme for a finite-element discretization of a convection-diffusion equation.

5.6.1 Test example

We take a $5\,000 \times 5\,000$ - matrix with equidistant eigenvalues on the imaginary axis between $-1\,000\,i$ and $1\,000\,i$ that we collect in a diagonal matrix $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{5\,000})$. Performing an orthogonal similarity transform, we obtain a dense matrix $\boldsymbol{A} = \boldsymbol{Q}^H \boldsymbol{D} \boldsymbol{Q}$, where $\boldsymbol{Q}$ is chosen as an orthogonal test matrix from the Matlab gallery 'orthog' of type $k = 1$. The exact matrix function, that we need as a reference solution, can be easily computed by $\varphi_\ell(\boldsymbol{A}) = \boldsymbol{Q}^H \varphi_\ell(\boldsymbol{D}) \boldsymbol{Q}$. The initial vector $\boldsymbol{v} \in \mathbb{R}^{5\,000}$ is generated by `randn` and scaled such that $\|\boldsymbol{v}\|_2 = 1$. The matrix $\boldsymbol{A}$ is obviously normal and, therefore, we have $W(\boldsymbol{A}) = [-1\,000\,i, 1\,000\,i] \subseteq \mathbb{C}_0^-$ so that our error estimates apply.

In Figure 5.4 on the left-hand side, the error curves for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the shift-and-invert Krylov subspace $\mathcal{Q}_m(\boldsymbol{A}, \boldsymbol{v})$ with shift $\gamma = 1$ show a sublinear convergence behavior that is significantly faster for larger indices ℓ. From the top down, the plot shows the approximation error for $\ell = 2, 4, 6, 8$ (red, blue, green, black line) against the number of rational Krylov steps. To compare the convergence rate with our predicted convergence rate of order $m^{-\ell/2}$, we draw the same error curves (solid lines) in a double logarithmic scale together with some lines of order $\mathcal{O}(m^{-\ell/2})$, $\ell = 2, 4, 6, 8$, on the right-hand side of Figure 5.4.

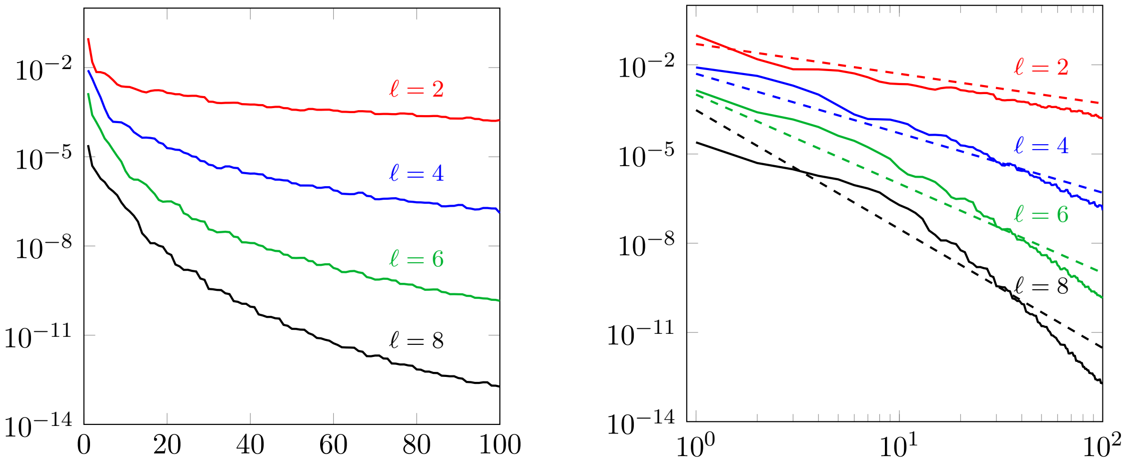


Figure 5.4: Left-hand side: Error $\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\|_2$ plotted against the number of shift-and-invert Krylov steps for $\gamma = 1$. Right-hand side: Same plot as on the left but in double logarithmic scale together with lines of order $\mathcal{O}(m^{-\ell/2})$, which correspond to the predicted convergence rate.

5.6.2 One-dimensional wave equation

In a second example, we approximate the solution of a one-dimensional wave equation with a source term $f(t,x)$ on the right hand side that is given as

$$\begin{aligned} u'' &= \frac{\partial^2}{\partial x^2} u + f(t,x) \qquad && \text{for} \quad x \in \Omega\,,\ t \geq 0\,, \\ u(0,x) &= u_0(x)\,,\ u'(0,x) = u_0'(x) \qquad && \text{for} \quad x \in \Omega\,, \end{aligned}$$

on the Hilbert space $H = L^2(\Omega)$ with $\Omega = (0,1)$. We assume homogeneous Dirichlet boundary conditions, $u(t,0) = u(t,1) = 0$, and use a spectral method for the spatial discretization. The functions $\psi_k(x) = \sqrt{2}\sin(k\pi x)$ are eigenfunctions of the second derivative operator with homogeneous Dirichlet boundary conditions to the eigenvalues $-(k\pi)^2$ and form an orthonormal basis of $L^2(\Omega)$ (cf. Section 6.2 below). Hence, $u(t,x)$ can be expanded in a generalized Fourier series

$$u(t,x) = \sum_{k=1}^{\infty} \widetilde{a}_k(t)\psi_k(x)\,, \qquad \psi_k(x) = \sqrt{2}\sin(k\pi x) \tag{5.21}$$

with the unknown Fourier coefficients

$$\widetilde{a}_k(t) = \int_0^1 u(t,x)\psi_k(x)\,dx\,.$$

We search for an approximate solution of the wave equation in the finite dimensional subspace $\operatorname{span}\{\psi_1(x),\ldots,\psi_N(x)\}$. For this purpose, we approximate the source term $f(t,x)$ by the truncated generalized Fourier series

$$f(t,x) \approx \sum_{k=1}^{N} f_k(t)\psi_k(x)\,, \qquad f_k(t) = \int_0^1 f(t,x)\psi_k(x)\,dx\,, \qquad k = 1,\ldots,N\,.$$

In order to approximate the unknown solution $u(t,x)$ in (5.21), we substitute the ansatz

$$u_N(t,x) = \sum_{k=1}^{N} a_k(t)\psi_k(x)$$

into the given partial differential equation which leads to

$$\sum_{k=1}^{N} a_k''(t)\psi_k(x) = -\sum_{k=1}^{N}(k\pi)^2 a_k(t)\psi_k(x) + \sum_{k=1}^{N} f_k(t)\psi_k(x)\,.$$

This is a system of ordinary differential equations

$$\boldsymbol{a}''(t) = -\boldsymbol{B}\boldsymbol{a}(t) + \boldsymbol{f}(t)$$

for the coefficients $a_k(t)$, where

$$\boldsymbol{a}(t) = \big(a_k(t)\big)_{k=1}^{N}\,, \qquad \boldsymbol{f}(t) = \big(f_k(t)\big)_{k=1}^{N}\,, \qquad \boldsymbol{B} = \operatorname{diag}\big(\pi^2, (2\pi)^2, \ldots, (N\pi)^2\big)\,.$$

The initial conditions $\boldsymbol{a}(0) = \big(a_k(0)\big)_{k=1}^{N}$ and $\boldsymbol{a}'(0) = \big(a_k'(0)\big)_{k=1}^{N}$ read

$$a_k(0) = \int_0^1 u_0(x)\psi_k(x)\,dx\,, \qquad a_k'(0) = \int_0^1 u_0'(x)\psi_k(x)\,dx\,, \qquad k = 1,\ldots,N\,.$$

Setting $\boldsymbol{v}(t) = \boldsymbol{a}(t)$ and $\widetilde{\boldsymbol{w}}(t) = \boldsymbol{a}'(t)$, the system $\boldsymbol{a}''(t) = -\boldsymbol{B}\boldsymbol{a}(t) + \boldsymbol{f}(t)$ can also be written as

$$\begin{bmatrix} \boldsymbol{v}(t) \\ \widetilde{\boldsymbol{w}}(t) \end{bmatrix}' = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{I} \\ -\boldsymbol{B} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(t) \\ \widetilde{\boldsymbol{w}}(t) \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{f}(t) \end{bmatrix}.$$

With the transformation $\boldsymbol{w}(t) = \boldsymbol{B}^{-1/2}\widetilde{\boldsymbol{w}}(t)$, we find the representation

$$\begin{aligned} \boldsymbol{y}'(t) = \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}' &= \begin{bmatrix} \boldsymbol{O} & \boldsymbol{B}^{1/2} \\ -\boldsymbol{B}^{1/2} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B}^{-1/2}\boldsymbol{f}(t) \end{bmatrix} \\ &= \boldsymbol{A}\boldsymbol{y}(t) + \boldsymbol{F}(t) \end{aligned} \tag{5.22}$$

with $\boldsymbol{B}^{1/2} = \operatorname{diag}\big(\pi, 2\pi, \ldots, N\pi\big)$ and initial value

$$\boldsymbol{y}(0) = \boldsymbol{y}_0 = \begin{bmatrix} \boldsymbol{v}_0 \\ \boldsymbol{w}_0 \end{bmatrix}, \qquad \boldsymbol{v}_0 = \big(a_k(0)\big)_{k=1}^{N}\,, \qquad \boldsymbol{w}_0 = \boldsymbol{B}^{-1/2}\big(a_k'(0)\big)_{k=1}^{N}\,.$$

The matrix $\boldsymbol{A} \in \mathbb{R}^{2N\times 2N}$ in (5.22) is skew-symmetric and hence has purely imaginary eigenvalues. Since $\boldsymbol{A}$ is a normal matrix, the field of values $W(\boldsymbol{A})$ is the convex hull of its eigenvalues and lies on the imaginary axis. More precisely, we have $W(\boldsymbol{A}) = [-iN\pi, iN\pi]$ and, thus, $\boldsymbol{A}$ fits in our framework above.

We choose $f(t,x) = f(x) = \sin(\pi x)$. Then the exponential Euler method yields the exact solution of the semi-discrete problem (5.22), that is,

$$\boldsymbol{y}(\tau) = e^{\tau\boldsymbol{A}}\boldsymbol{y}_0 + \tau\varphi_1(\tau\boldsymbol{A})\boldsymbol{F} = \tau\varphi_1(\tau\boldsymbol{A})(\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F}) + \boldsymbol{y}_0\,.$$

As initial functions, we use

$$u_0(x) = 0\,, \qquad u_0'(x) = 100\,x^2(1-x)^2\,,$$

such that $\boldsymbol{v}_0 = \boldsymbol{0}$. Because of

$$a_k'(0) = \int_0^1 u_0'(x)\psi_k(x)\,dx = \frac{200\sqrt{2}}{k^5\pi^5}\left(\cos(k\pi)(k^2\pi^2 - 12) - k^2\pi^2 + 12\right),$$

we further have

$$\boldsymbol{w}_0 = (w_{0,k})_{k=1}^N\,, \qquad w_{0,k} = \begin{cases} 0\,, & k \text{ even}\,, \\ \dfrac{400\sqrt{2}}{k\pi}\cdot\dfrac{12 - k^2\pi^2}{k^5\pi^5}\,, & k \text{ odd}\,. \end{cases}$$

The orthonormality of the chosen basis functions $\psi_k(x)$ on $L^2(\Omega)$ gives

$$\left(\int_0^1 \left(\sum_{k=1}^N a_k(t)\psi_k(x)\right)^2 dx\right)^{\frac{1}{2}} = \left(\sum_{k=1}^N a_k^2(t)\right)^{\frac{1}{2}} = \|\boldsymbol{a}\|_2\,,$$

so that we can measure the approximation error in the standard Euclidean norm. This identity is also known as Parseval's equality.

In Figure 5.5, we see the obtained results using $N = 31, 63, 1\,023, 1\,048\,575$ ansatz functions (top left to bottom right) for the approximation of $\varphi_1(\tau\boldsymbol{A})(\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F})$ in the subspace $\mathcal{Q}_m(\tau\boldsymbol{A}, \boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F}) = \mathcal{K}_m\big((\gamma\boldsymbol{I} - \tau\boldsymbol{A})^{-1}, \boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F}\big)$ (blue solid and black dash-dotted line) and, for comparison, in the polynomial Krylov subspace $\mathcal{K}_m(\tau\boldsymbol{A}, \boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F})$ (red dashed line). The Krylov subspace approximation reads

$$\varphi_1(\tau\boldsymbol{A})(\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F}) \approx \varphi_1(\tau\boldsymbol{A}_m)(\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F}) = \boldsymbol{V}_m\varphi_1(\tau\boldsymbol{S}_m)\boldsymbol{V}_m^H(\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F})$$

with the compression $\boldsymbol{S}_m = \boldsymbol{V}_m^H\boldsymbol{A}\boldsymbol{V}_m$ and the restriction $\boldsymbol{A}_m = \boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m$ of the matrix $\boldsymbol{A}$, where $\boldsymbol{P}_m = \boldsymbol{V}_m\boldsymbol{V}_m^H$ is the orthogonal projection onto the rational or, respectively, standard Krylov subspace.

The approximation error is plotted against the number of iterations for $m = 1, \ldots, 20$, time step $\tau = 0.1$, and different shifts $\gamma = 1$ (blue solid line), $\gamma = 20^{3/5}$ (black dash-dotted line). The second shift is chosen as $\gamma = m^{(r-1/2)/(r+1/2)}$ according to the analysis in Section 5.4 with $m = 20$, $r = 2$ and, therefore, results in a noticeably faster convergence rate. As predicted by the error bounds above, the error curves corresponding to the rational method show a sublinear convergence behavior.

The polynomial method only leads to success for very coarse space discretizations with a small number ($N = 31, 63$) of basis functions $\psi_k(x)$, $k = 1, \ldots, N$. For larger values of N, we observe a stagnation of the error curve after the third iteration step, whereas the rational approximation performs well independent of the number of basis functions. This effect is, roughly speaking, caused by the smoothness properties of the initial value $\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F}$. In the next chapter, we will explain this observation in detail. Furthermore, we discuss how the standard and rational Krylov subspace method can be efficiently combined, in order to exploit the decrease of the less expensive standard Krylov subspace method in the first few iteration steps.

Moreover, we compare the performance of the standard and the shift-and-invert Krylov subspace method with respect to approximation error versus computing time in seconds. For the finest grid with $N = 1\,048\,575$ ansatz functions and time step $\tau = 0.1$ as above, the obtained results are shown in Figure 5.6. The computation has been conducted in the software environment Matlab, Release 2014a, under Ubuntu, Release 13.10, on a dual Xeon CPU workstation with a total of eight cores each running at 2.33 GHz.

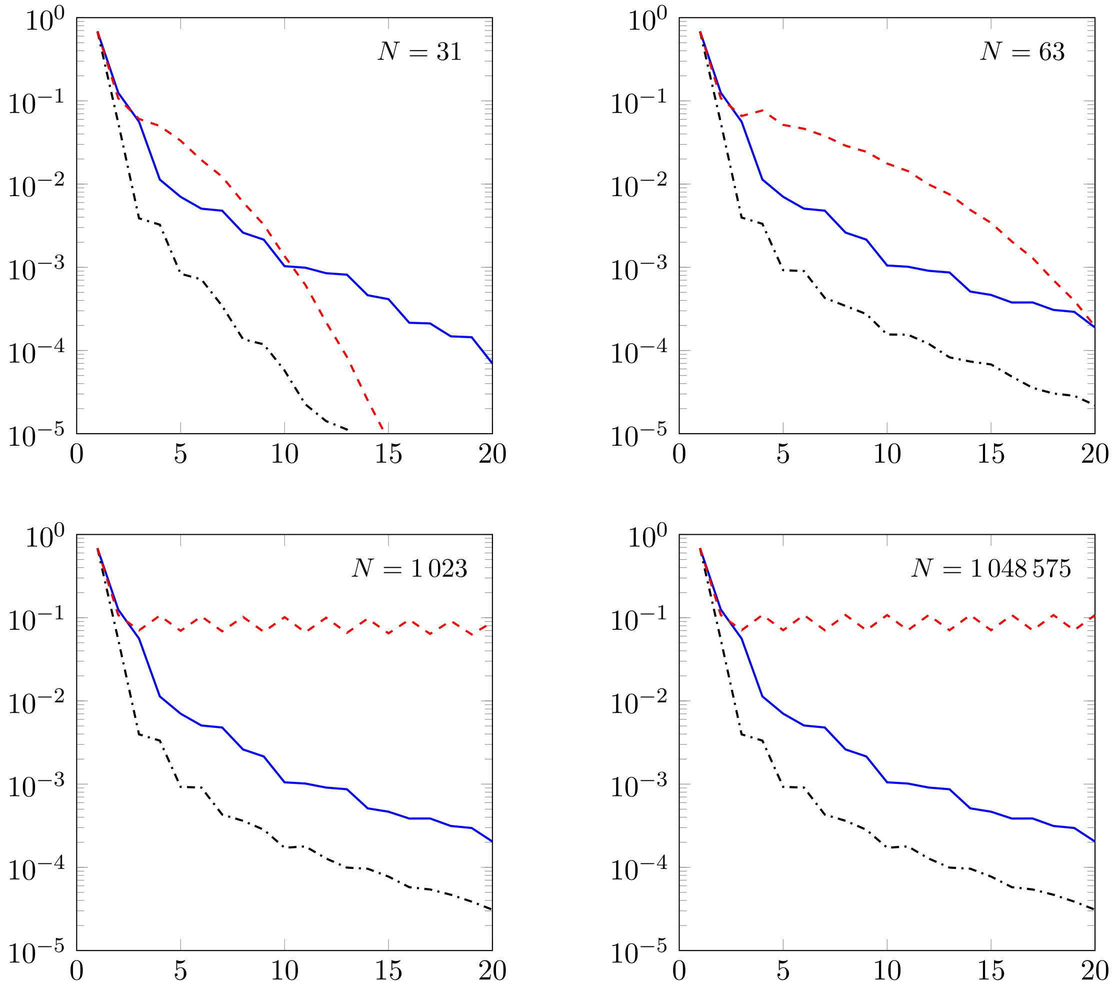


Figure 5.5: Plot of the error $\|\varphi_1(\tau \boldsymbol{A})(\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F}) - \varphi_1(\tau \boldsymbol{A}_m)(\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F})\|_2$ versus the dimension of the Krylov subspace for the standard Krylov subspace $\mathcal{K}_m(\tau\boldsymbol{A}, \boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F})$ (red dashed line) and the shift-and-invert Krylov subspace $\mathcal{Q}_m(\tau\boldsymbol{A}, \boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{F})$ with $\gamma = 1$ (blue solid line) and $\gamma = 20^{3/5}$ (black dash-dotted line) for $\tau = 0.1$, $N = 31, 63, 1\,023, 1\,048\,575$.

5.6.3 Convection-diffusion equation

As a third numerical example, we consider the convection-diffusion equation

$$\begin{aligned} u' &= d\Delta u - b^T \nabla u \qquad &&\text{for} \qquad (x, y) \in \Omega\,,\ t \geq 0\,, \\ u(0, x, y) &= u_0(x, y) \qquad &&\text{for} \qquad (x, y) \in \Omega \end{aligned}$$

with homogeneous Dirichlet boundary conditions on the unit square $\Omega = (0, 1)^2$ for the Hilbert space $H = L^2(\Omega)$. The coefficient $d > 0$ is called the diffusivity or diffusion coefficient, b^T represents the velocity, and u describes the concentration of a substance. The spatial discretization is done by finite elements using a regular triangulation with $n+2$ nodes in each space direction, such that we have a quadratic grid with $N = n^2$ inner nodes and mesh size $h = \frac{1}{n+1}$. We choose the standard N nodal linear basis functions ϕ_k, $k = 1, \ldots, N$, that take the value 1 at the kth vertex and 0 at all other nodes. According

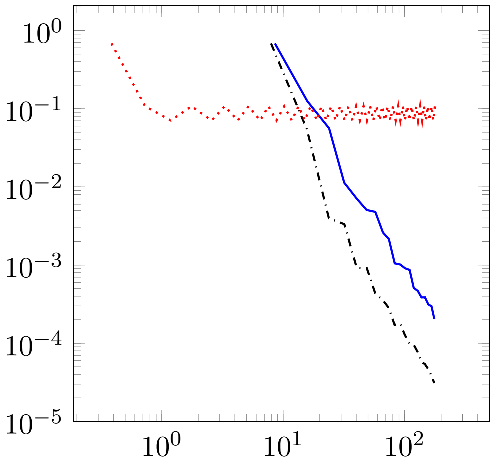


Figure 5.6: Plot of error versus computation time in seconds for the standard (red dotted line) and the shift-and-invert Krylov subspace method with $\gamma = 1$ (blue solid line) and $\gamma = 20^{3/5}$ (black dash-dotted line) for $N = 1\,048\,575$ and $\tau = 0.1$.

to Section 3.1.2, this leads to the system of ordinary differential equations

$$\boldsymbol{M}\boldsymbol{u}'(t) = \boldsymbol{S}\boldsymbol{u}(t)\,, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0\,, \tag{5.23}$$

where $\boldsymbol{M}$ represents the mass matrix, $\boldsymbol{S}$ is the stiffness matrix, and the coefficient vector $\boldsymbol{u}(t) = \big(u_k(t)\big)_{k=1}^N \in \mathbb{C}^N$ contains the coefficients of the approximation

$$u(t,x,y) \approx \sum_{k=1}^N u_k(t)\phi_k(x,y)\,.$$

By multiplication with $\boldsymbol{M}^{-1}$ from the left, equation (5.23) can be written as

$$\boldsymbol{u}'(t) = \boldsymbol{M}^{-1}\boldsymbol{S}\boldsymbol{u}(t) = \boldsymbol{A}\boldsymbol{u}(t)\,, \qquad \boldsymbol{u}(0) = \boldsymbol{u}_0$$

with exact solution $\boldsymbol{u}(\tau) = e^{\tau\boldsymbol{A}}\boldsymbol{u}_0$. Since the error analysis above only states estimates for the φ_ℓ-functions for $\ell > 0$ and not for the exponential function, we rewrite the solution as

$$\boldsymbol{u}(\tau) = \tau\varphi_1(\tau\boldsymbol{A})\boldsymbol{A}\boldsymbol{u}_0 + \boldsymbol{u}_0 \tag{5.24}$$

and approximate instead the action of the matrix φ_1-function of $\tau\boldsymbol{A}$ on $\boldsymbol{A}\boldsymbol{u}_0$ in the shift-and-invert Krylov subspace $\mathcal{Q}_m(\tau\boldsymbol{A}, \boldsymbol{A}\boldsymbol{u}_0)$.

The corresponding inner product $(\boldsymbol{v},\boldsymbol{w})_{\boldsymbol{M}} = \boldsymbol{w}^H\boldsymbol{M}\boldsymbol{v}$, $\boldsymbol{v},\boldsymbol{w} \in \mathbb{C}^N$, with associated norm $\|\boldsymbol{v}\|_{\boldsymbol{M}} = \sqrt{(\boldsymbol{v},\boldsymbol{v})_{\boldsymbol{M}}}$ is determined by the mass matrix $\boldsymbol{M}$, containing the L^2-inner products

$$(\phi_i,\phi_j)_{L^2(\Omega)} = \int_\Omega \phi_i\phi_j\, d(x,y)$$

of the finite-element basis ϕ_k, $k = 1,\dots,N$. In order to apply the results from above, we have to assure that the field of values with respect to the inner product $(\cdot,\cdot)_{\boldsymbol{M}}$ satisfies $W(\boldsymbol{A}) = W(\boldsymbol{M}^{-1}\boldsymbol{S}) \subseteq \mathbb{C}_0^-$, which is the statement of the following lemma.

Lemma 5.15 *The matrix $\boldsymbol{A} = \boldsymbol{M}^{-1}\boldsymbol{S}$ of the finite-element discretization fulfills*

$$\operatorname{Re}(\boldsymbol{A}\boldsymbol{v},\boldsymbol{v})_{\boldsymbol{M}} \le 0\,.$$

Proof. The stiffness matrix $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2$ stems from the spatial discretization of the differential operator $d\Delta - b^T\nabla$ in the convection diffusion equation, where $\boldsymbol{S}_1$ represents the part that belongs to the Laplacian $d\Delta$ and $\boldsymbol{S}_2$ is assigned to $-b^T\nabla$. Because of $d > 0$, the first part $\boldsymbol{S}_1$ is symmetric negative definite and has a field of values on the negative real axis. Moreover, since

$$\int_\Omega (b^T\nabla\phi_i)\phi_j \, d(x,y) = -\int_\Omega (b^T\nabla\phi_j)\phi_i \, d(x,y)\,,$$

the second matrix $\boldsymbol{S}_2$ is skew-symmetric and has a field of values on the imaginary axis. Denoting by $W_2(\cdot)$ the field of values with respect to the standard Euclidean inner product $(\cdot\,,\cdot)_2$, we have

$$W_2(\boldsymbol{S}) = \underbrace{W_2(\boldsymbol{S}_1)}_{\subseteq\mathbb{R}^-} + \underbrace{W_2(\boldsymbol{S}_2)}_{\subseteq i\mathbb{R}} \subseteq \mathbb{C}_0^-\,.$$

For every vector $\boldsymbol{v}$, we thus obtain

$$\operatorname{Re}(\boldsymbol{A}\boldsymbol{v},\boldsymbol{v})_{\boldsymbol{M}} = \operatorname{Re}(\boldsymbol{S}\boldsymbol{v},\boldsymbol{v})_2 \le 0\,,$$

which proves the result. ❑

We now approximate $\varphi_1(\tau\boldsymbol{A})\boldsymbol{A}\boldsymbol{u}_0$ in the rational Krylov subspace $\mathcal{Q}_m(\tau\boldsymbol{A},\boldsymbol{A}\boldsymbol{u}_0)$ for the time step size $\tau = 0.005$, diffusion coefficient $d = 0.05$ and velocity $b^T = [-50,-50]$ on a grid with $N = 10\,000$ inner nodes. The initial value u_0 has the shape of a small peak in the lower left corner of the unit square (see Figure 5.7 on the left), which moves to the upper right corner as time progresses and meanwhile diffuses a bit. More exactly, the initial value u_0 is given by

$$u_0(x,y) = \begin{cases} 10^5\cdot(x-0.05)^2\,(x-0.5)^2\,(y-0.05)^2\,(y-0.5)^2\,, & (x,y)\in[0.05,0.5]^2\,,\\ 0\,, & \text{elsewhere}\,. \end{cases}$$

The computation of an orthonormal basis $\boldsymbol{V}_m$ of $\mathcal{Q}_m(\tau\boldsymbol{A},\boldsymbol{A}\boldsymbol{u}_0)$ is done by a rational Arnoldi decomposition. In the most general sense, this process is, for a finite-element discretization, realized by Algorithm 5.16, which is here given in a simplified and easily readable form. For the numerical experiments, we made improvements concerning the stability and efficiency. For example, a reorthogonalization is used in each step and we compute $(\gamma\boldsymbol{M}-\tau\boldsymbol{S})^{-1}\boldsymbol{w}$ only once for each iteration of the loop over m. Not only in this numerical experiment, but also in all other experiments, we used improved versions of the presented algorithms for the numerical computations.

The basis $\boldsymbol{V}_m$ obtained from Algorithm 5.16 is orthonormal with respect to the $\boldsymbol{M}$-inner product such that $\boldsymbol{V}_m^H\boldsymbol{M}\boldsymbol{V}_m = \boldsymbol{I}$. In addition to the basis vector $\boldsymbol{v}_{m+1}$, we compute the auxiliary vector $\boldsymbol{w} = \boldsymbol{M}\boldsymbol{v}_{m+1}$ in each iteration step. This is motivated by the fact that, in the calculation of $h_{j,m}$, we have

$$(\gamma\boldsymbol{I}-\tau\boldsymbol{A})^{-1}\boldsymbol{v}_m = (\gamma\boldsymbol{I}-\tau\boldsymbol{M}^{-1}\boldsymbol{S})^{-1}\boldsymbol{v}_m = (\gamma\boldsymbol{M}-\tau\boldsymbol{S})^{-1}\boldsymbol{w}\,.$$

The projector onto the shift-and-invert Krylov subspace is given by $\boldsymbol{P}_m = \boldsymbol{V}_m\boldsymbol{V}_m^H\boldsymbol{M}$ and the restriction of $\boldsymbol{A} = \boldsymbol{M}^{-1}\boldsymbol{S}$ onto the rational subspace reads

$$\boldsymbol{A}_m = \boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m = \boldsymbol{V}_m\boldsymbol{S}_m\boldsymbol{V}_m^H\boldsymbol{M}\,, \qquad \boldsymbol{S}_m = \boldsymbol{V}_m^H\boldsymbol{M}\boldsymbol{A}\boldsymbol{V}_m = \boldsymbol{V}_m^H\boldsymbol{S}\boldsymbol{V}_m\,.$$

Algorithm 5.16 FE rational Arnoldi process

```
given:  mass matrix $\boldsymbol{M} \in \mathbb{C}^{N\times N}$,
        stiffness matrix $\boldsymbol{S} \in \mathbb{C}^{N\times N}$,
        initial value $\boldsymbol{v} \in \mathbb{C}^{N}$,  shift $\gamma > 0$
$\boldsymbol{w} = \boldsymbol{M}\boldsymbol{v}$
$\boldsymbol{v}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|_{\boldsymbol{M}}$,  $\boldsymbol{w} = \boldsymbol{w}/\|\boldsymbol{v}\|_{\boldsymbol{M}}$
for  $m = 1, 2, \dots$  do
    for  $j = 1, \dots, m$  do
        $h_{j,m} = \big((\gamma\boldsymbol{M} - \tau\boldsymbol{S})^{-1}\boldsymbol{w}, \boldsymbol{v}_j\big)_{\boldsymbol{M}}$
    end for
    $\widetilde{\boldsymbol{v}}_{m+1} = (\gamma\boldsymbol{M} - \tau\boldsymbol{S})^{-1}\boldsymbol{w} - \sum_{j=1}^{m} h_{j,m}\boldsymbol{v}_j$
    $\widetilde{\boldsymbol{w}} = \boldsymbol{M}\widetilde{\boldsymbol{v}}_{m+1}$
    $h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|_{\boldsymbol{M}}$
    $\boldsymbol{v}_{m+1} = \widetilde{\boldsymbol{v}}_{m+1}/h_{m+1,m}$
    $\boldsymbol{w} = \widetilde{\boldsymbol{w}}/h_{m+1,m}$
end for
```

With this, the rational Krylov subspace approximation is calculated as

$$\varphi_1(\tau\boldsymbol{A})\boldsymbol{A}\boldsymbol{u}_0 \approx \varphi_1(\tau\boldsymbol{A}_m)\boldsymbol{A}\boldsymbol{u}_0 = \boldsymbol{V}_m\varphi_1(\tau\boldsymbol{S}_m)\boldsymbol{V}_m^H\boldsymbol{M}\boldsymbol{A}\boldsymbol{u}_0 = \|\boldsymbol{A}\boldsymbol{u}_0\|_{\boldsymbol{M}}\boldsymbol{V}_m\varphi_1(\tau\boldsymbol{S}_m)\boldsymbol{e}_1 .$$

Even though the error estimates above provide no convergence result for the direct approximation of $e^{\tau\boldsymbol{A}}\boldsymbol{u}_0$ in the rational Krylov subspace $\mathcal{Q}_m(\tau\boldsymbol{A}, \boldsymbol{u}_0)$, relation (5.24) allows us to use the available rational Krylov subspace approximation for $\varphi_1(\tau\boldsymbol{A})\boldsymbol{A}\boldsymbol{u}_0$. On the right-hand side of Figure 5.7, we draw the approximation error (blue solid line)

$$E_m := \|e^{\tau\boldsymbol{A}}\boldsymbol{u}_0 - (\boldsymbol{u}_0 + \tau\varphi_1(\tau\boldsymbol{A}_m)\boldsymbol{A}\boldsymbol{u}_0)\|_{\boldsymbol{M}}$$

against the number of iteration steps for the shift $\gamma = 1$. In this situation, Theorem 5.9 yields the sublinear bound

$$E_m = \tau\|\underbrace{\varphi_1(\tau\boldsymbol{A})\boldsymbol{A}\boldsymbol{u}_0}_{=\frac{1}{\tau}(e^{\tau\boldsymbol{A}}\boldsymbol{u}_0 - \boldsymbol{u}_0)} - \varphi_1(\tau\boldsymbol{A}_m)\boldsymbol{A}\boldsymbol{u}_0\|_{\boldsymbol{M}} \leq \tau\,\frac{C(\gamma)}{\sqrt{m}}\|\boldsymbol{A}\boldsymbol{u}_0\|_{\boldsymbol{M}} . \tag{5.25}$$

Finally, we compare the performance of the shift-and-invert Krylov method with the implicit Euler and the Crank-Nicolson method. Applied to the system of ordinary differential equations $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t) = f\big(\boldsymbol{u}(t)\big)$, $\boldsymbol{u}_0 = \boldsymbol{u}(0)$, these schemes compute an approximation of $\boldsymbol{u}(\tau) = e^{\tau\boldsymbol{A}}\boldsymbol{u}_0$ via the recursions

$$\begin{aligned} &\text{implicit Euler:} && \boldsymbol{u}_{k+1}^{\text{imE}} = \boldsymbol{u}_k^{\text{imE}} + \tfrac{\tau}{m} f(\boldsymbol{u}_{k+1}^{\text{imE}}) , \\ &\text{Crank-Nicolson:} && \boldsymbol{u}_{k+1}^{\text{CN}} = \boldsymbol{u}_k^{\text{CN}} + \tfrac{\tau}{m} f\left(\tfrac{1}{2}(\boldsymbol{u}_k^{\text{CN}} + \boldsymbol{u}_{k+1}^{\text{CN}})\right) \end{aligned}$$

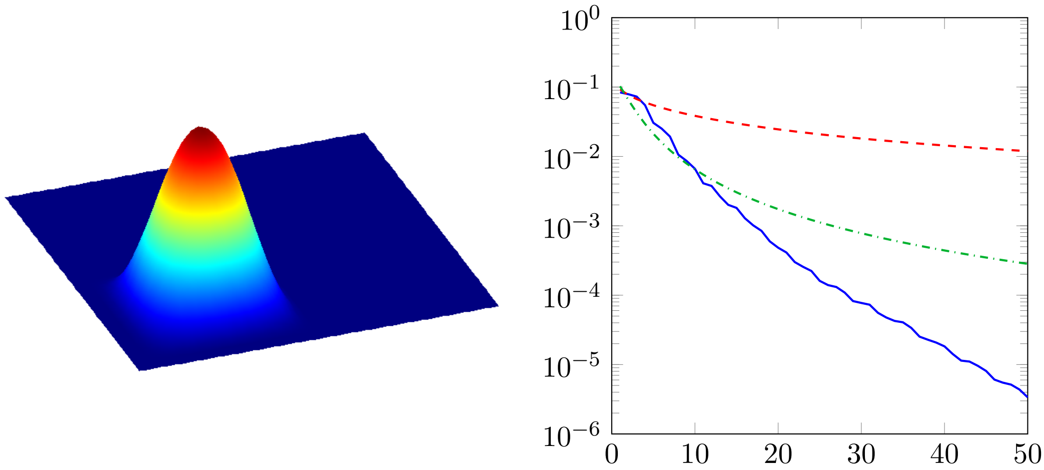


Figure 5.7: Left-hand side: Initial value u_0 on the unit square. Right-hand side: Plot of the approximation error E_m versus m (blue solid line) together with the error of the implicit Euler method (red dashed line) and the Crank-Nicolson scheme (green dash-dotted line).

for $k = 0, \dots, m-1$. In our case, we have

$$e^{\tau \boldsymbol{A}} \boldsymbol{u}_0 \approx \boldsymbol{u}_m^{\text{imE}} = \left(\boldsymbol{I} - \frac{\tau}{m} \boldsymbol{A} \right)^{-m} \boldsymbol{u}_0$$

and

$$e^{\tau \boldsymbol{A}} \boldsymbol{u}_0 \approx \boldsymbol{u}_m^{\text{CN}} = \left(\left(\boldsymbol{I} - \frac{\tau}{2m} \boldsymbol{A} \right)^{-1} \left(\boldsymbol{I} + \frac{\tau}{2m} \boldsymbol{A} \right) \right)^{m} \boldsymbol{u}_0 \,.$$

It is well-known that the implicit Euler scheme is convergent of order one, whereas the Crank-Nicolson method is convergent of order two, that means

$$\| \boldsymbol{u}(\tau) - \boldsymbol{u}_m^{\text{imE}} \| = \mathcal{O}\left(\frac{\tau}{m} \right) \qquad \text{and} \qquad \| \boldsymbol{u}(\tau) - \boldsymbol{u}_m^{\text{CN}} \| = \mathcal{O}\left(\frac{\tau^2}{m^2} \right) \,.$$

Via the relations

$$\boldsymbol{u}_m^{\text{imE}} \in \mathcal{K}_{m+1}\big((m\boldsymbol{I} - \tau \boldsymbol{A})^{-1}, \boldsymbol{u}_0 \big) \qquad \text{and} \qquad \boldsymbol{u}_m^{\text{CN}} \in \mathcal{K}_{m+1}\big((2m\boldsymbol{I} - \tau \boldsymbol{A})^{-1}, \boldsymbol{u}_0 \big) \,,$$

the implicit Euler and the Crank-Nicolson scheme are related to the shift-and-invert Krylov subspace method.

For the direct comparison of the three methods, we do not approximate the solution $e^{\tau \boldsymbol{A}} \boldsymbol{u}_0$ via the φ_1-function but directly by $e^{\tau \boldsymbol{A}_{m+1}} \boldsymbol{u}_0$, where $\boldsymbol{A}_{m+1}$ is the restriction of $\boldsymbol{A}$ to $\mathcal{K}_{m+1}\big((m\boldsymbol{I} - \tau \boldsymbol{A})^{-1}, \boldsymbol{u}_0 \big)$ or $\mathcal{K}_{m+1}\big((2m\boldsymbol{I} - \tau \boldsymbol{A})^{-1}, \boldsymbol{u}_0 \big)$, even though the error bounds derived above make no statement in this case.

On the left-hand side of Figure 5.8, we draw the error for the approximation of $e^{\tau \boldsymbol{A}} \boldsymbol{u}_0$ for $\tau = 0.005$ by the rational Krylov process with pole $\gamma = m$ (blue solid line) and by the implicit Euler method (red dashed line). On the right-hand side, the error curves for the shift-and-invert Krylov subspace approximation with $\gamma = 2m$ (blue solid line) and for the Crank-Nicolson scheme (green dash-dotted line) are depicted. In both cases, the rational Krylov subspace method converges significantly faster. The reason for this observation is the near-optimality property of Krylov subspace methods, which was discussed in Section 4.4. The implicit Euler and the Crank-Nicolson scheme can be interpreted as a fixed rational approximation in $\mathcal{K}_{m+1}\big((m\boldsymbol{I}-\tau\boldsymbol{A})^{-1},\boldsymbol{u}_0\big)$ or $\mathcal{K}_{m+1}\big((2m\boldsymbol{I}-\tau\boldsymbol{A})^{-1},\boldsymbol{u}_0\big)$. In contrast, by Theorem 4.16 the rational Krylov process automatically yields a near-best approximation in these subspaces.

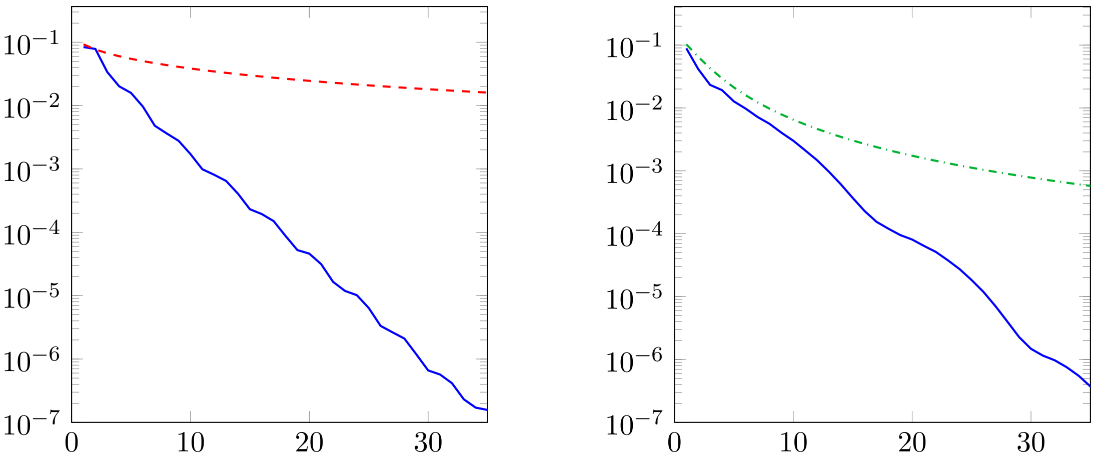


Figure 5.8: Left-hand side: Comparison of the implicit Euler scheme (red dashed line) with the shift-and-invert method for $\gamma = m$ (blue solid line). Right-hand side: Comparison of the Crank-Nicolson scheme (red dashed line) with the shift-and-invert method for $\gamma = 2m$ (blue solid line). In both pictures, the error is plotted against m.

Even if we choose the fixed shift $\gamma = 1$ and compare the performance of the shift-and-invert Krylov subspace approximation for $e^{\tau\boldsymbol{A}}\boldsymbol{u}_0$ via $\boldsymbol{u}_0+\tau\varphi_1(\tau\boldsymbol{A}_m)\boldsymbol{A}\boldsymbol{u}_0$ with the implicit Euler and the Crank-Nicolson method in Figure 5.7 on the right-hand side, the convergence rate of the rational Krylov process is considerably better.

In contrast to the fixed inverse $(\gamma\boldsymbol{I}-\tau\boldsymbol{A})^{-1}$ used in the rational Krylov decomposition, the inverse $(\boldsymbol{I}-\frac{\tau}{m}\boldsymbol{A})^{-1}$ for the implicit Euler or $(\boldsymbol{I}-\frac{\tau}{2m}\boldsymbol{A})^{-1}$ for the Crank-Nicolson method varies depending on the iteration index m. Once the inverse of $\gamma\boldsymbol{I}-\tau\boldsymbol{A}$ is known (e.g., in form of an LU decomposition), it can be applied in the rational Krylov algorithm independent of the number m of iteration steps. Conversely, $(\boldsymbol{I}-\frac{\tau}{m}\boldsymbol{A})^{-1}$ and $(\boldsymbol{I}-\frac{\tau}{2m}\boldsymbol{A})^{-1}$ have to be computed from scratch every time, if we raise the dimension m of the approximation subspace, in order to approximate the solution more accurately.

Chapter 6

Extended Krylov subspace approximation

If the initial value $\boldsymbol{v}$ satisfies specific smoothness conditions, it is often worthwhile to consider the extended Krylov subspace approximation of $f(\boldsymbol{A})\boldsymbol{v}$, which combines the standard and the rational Krylov subspace process. For a nonsingular and symmetric matrix $\boldsymbol{A}$, the extended Krylov subspace

$$\mathcal{K}_{q+1,m}(\boldsymbol{A},\boldsymbol{v}) = \operatorname{span}\{\boldsymbol{A}^q\boldsymbol{v},\dots,\boldsymbol{A}\boldsymbol{v},\boldsymbol{v},\boldsymbol{A}^{-1}\boldsymbol{v},\dots,\boldsymbol{A}^{-(m-1)}\boldsymbol{v}\}\,, \quad q \geq 0\,, \quad m \geq 1$$

was first proposed by Druskin and Knizhnerman [15], in order to approximate the matrix square root and related functions. Of course, the extended Krylov subspace method is generalizable to non-symmetric matrices. In [48], Knizhnerman and Simoncini study the approximation of Markov type functions by an extended Krylov subspace method for symmetric and non-symmetric matrices as well.

Because of the relation $\mathcal{K}_{q+1,m}(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_{q+m}(\boldsymbol{A},\boldsymbol{A}^{-(m-1)}\boldsymbol{v})$, the extended Krylov subspace can alternatively be seen as a standard Krylov subspace of order $q+m$ with the modified starting vector $\boldsymbol{A}^{-(m-1)}\boldsymbol{v}$. Since we would like to approximate matrix functions for matrices $\boldsymbol{A}$ with a field of values somewhere in the left complex half-plane and $\boldsymbol{A}$ thus may have zero eigenvalues, we will study a shifted version of this subspace, namely

$$\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v}) = \operatorname{span}\left\{\boldsymbol{A}^q\boldsymbol{v},\dots,\boldsymbol{A}\boldsymbol{v},\boldsymbol{v},\frac{1}{\gamma-\boldsymbol{A}}\,\boldsymbol{v},\dots,\frac{1}{(\gamma-\boldsymbol{A})^{m-1}}\,\boldsymbol{v}\right\}\,, \qquad \gamma > 0\,.$$

This subspace is of dimension $q+m$, if the invariance index has not been reached yet. The purely polynomial and purely rational part of this subspace are given by

$$\mathcal{K}^{\gamma}_{q+1,1}(\boldsymbol{A},\boldsymbol{v}) = \operatorname{span}\{\boldsymbol{A}^q\boldsymbol{v},\dots,\boldsymbol{A}\boldsymbol{v},\boldsymbol{v}\} = \mathcal{K}_{q+1}(\boldsymbol{A},\boldsymbol{v})$$

and

$$\mathcal{K}^{\gamma}_{1,m}(\boldsymbol{A},\boldsymbol{v}) = \operatorname{span}\left\{\boldsymbol{v},\frac{1}{\gamma-\boldsymbol{A}}\,\boldsymbol{v},\dots,\frac{1}{(\gamma-\boldsymbol{A})^{m-1}}\,\boldsymbol{v}\right\} = \mathcal{K}_m\big((\gamma\boldsymbol{I}-\boldsymbol{A})^{-1},\boldsymbol{v}\big)\,.$$

That is, $\mathcal{K}^{\gamma}_{1,m}(\boldsymbol{A},\boldsymbol{v})$ coincides with the shift-and-invert Krylov subspace in Chapter 5.

Assuming that $\boldsymbol{A}$ is a large discretization matrix with $W(\boldsymbol{A}) \subseteq \mathbb{C}^-_0$, as in the previous section, we are interested in error bounds for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ that guarantee a uniform convergence over all possible grids in space. This is an important property for the efficient application of the extended Krylov subspace method in exponential integrators.

In the first part of this chapter, it will turn out in which way the abstract smoothness of the continuous function v in a Hilbert space H, that is associated with the discrete vector $\boldsymbol{v}$, leads to a restriction for the index q of the polynomial part $\mathcal{K}^{\gamma}_{q+1,1}(\boldsymbol{A},\boldsymbol{v})$ of the extended Krylov subspace. Afterwards, we state error bounds for the extended Krylov subspace approximation and illustrate our results by several numerical experiments. The contents of this chapter can also be found in our paper [25].

6.1 Smoothness of the initial value

We first motivate to what extent the abstract smoothness of the initial value $v \in H$ is represented by the discretized vector $\boldsymbol{v}$ and that its smoothness plays a central role in the extended Krylov subspace approximation. In contrast to matrices, the differential operator A can only be applied to elements of a subspace $D(A) \subseteq H$, where $D(A)$ is the domain of A defined in (3.6). If we think of A, for example, as the Laplace operator with homogeneous Dirichlet boundary conditions, it is clear that A can only be applied to functions $v \in H$ that are twice differentiable and equal to zero at the boundary of the spatial domain Ω. Setting $D(A^0) := D(I) := H$, we define recursively spaces of smoother and smoother functions by

$$D(A^n) := \{v \in D(A^{n-1}) \,:\, A^{n-1}v \in D(A)\}\,, \qquad n = 1, 2, \dots\,.$$

Approximating the action of an operator function $f(A)$ on a vector $v \in H$ in the polynomial Krylov subspace $\mathcal{K}_{q+1}(A, v)$ or in the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(A, v)$, we would have to assume that $v \in D(A^q)$ to ensure that $Av, \dots, A^q v$ are defined in the case of operators. This smoothness requirement on v has a decisive effect on the discrete case that must not be neglected.

In order to illustrate this fact, we examine the one-dimensional Laplace operator $A = \frac{\partial^2}{\partial x^2}$ on the interval $\Omega = (0, 1)$ with homogeneous Dirichlet boundary conditions on the Hilbert space $H = L^2(\Omega)$. As initial value, we choose

$$v(x) = \frac{x^4(1-x)^4}{\|x^4(1-x)^4\|_{L^2(\Omega)}}\,, \qquad x \in \Omega\,. \tag{6.1}$$

Then v is infinitely often differentiable, but the forth derivative $v^{(4)}$ does no longer fulfill the zero boundary conditions, such that $v \in D(A^j)$ for $j \leq 2$ and $v \notin D(A^3)$. The discretized operator is given by the matrix $\boldsymbol{A} = \frac{1}{h^2}\,\mathrm{tridiag}(1, -2, 1) \in \mathbb{R}^{N \times N}$, where $h = \frac{1}{N+1}$ denotes the mesh size and N is the number of inner discretization points of Ω. This matrix arises from a finite-difference approximation of the second derivative. For the discretized initial value $\boldsymbol{v} \in \mathbb{R}^N$, we evaluate the continuous function in (6.1) at N inner grid points.

Let us consider the vectors $\boldsymbol{Av}$, $\boldsymbol{A}^2\boldsymbol{v}$ and $\boldsymbol{A}^3\boldsymbol{v}$ corresponding to the continuous counterparts $v^{(2)}$, $v^{(4)}$ and $v^{(6)}$. Figure 6.1 clearly shows that $\boldsymbol{Av}$ replicates the necessary zero boundary conditions very well. However, this is no longer the case for $\boldsymbol{A}^2\boldsymbol{v}$ and $\boldsymbol{A}^3\boldsymbol{v}$. This shows that a further multiplication of $\boldsymbol{A}^2\boldsymbol{v}$ with $\boldsymbol{A}$ is problematic. Moreover, we observe that the entries of $\boldsymbol{A}^k\boldsymbol{v}$ become larger and larger, if we increase the value $k \in \mathbb{N}$.

We compute the discrete L^2-norm of $\boldsymbol{A}^j\boldsymbol{v}$ for $j = 1, \dots, 4$ and plot them against the mesh size $h = \frac{1}{N+1}$ in Figure 6.3. The norms $\|\boldsymbol{A}^j\boldsymbol{v}\|$, corresponding to the well-defined abstract expressions Av and A^2v for $v \in D(A^2)$, stay at the same level for smaller values of h. As opposed to this, $\|\boldsymbol{A}^3\boldsymbol{v}\|$ and $\|\boldsymbol{A}^4\boldsymbol{v}\|$ grow rapidly for finer spatial meshes. The observed behavior corresponds to the fact that the associated continuous initial value v does neither belong to $D(A^3)$ nor to $D(A^4)$.

The situation is quite different, if we apply powers of $(\gamma\boldsymbol{I} - \boldsymbol{A})^{-1}$ to the initial value $\boldsymbol{v}$ instead. The boundedness and the smoothing property $(\gamma I - A)^{-1} \,:\, H \to D(A)$ of the resolvent is transferred to the discrete case. The vectors $(\gamma\boldsymbol{I} - \boldsymbol{A})^{-k}\boldsymbol{v}$ all replicate the homogeneous Dirichlet boundary conditions and remain uniformly bounded for all $k \in \mathbb{N}$, cf. Figure 6.2.

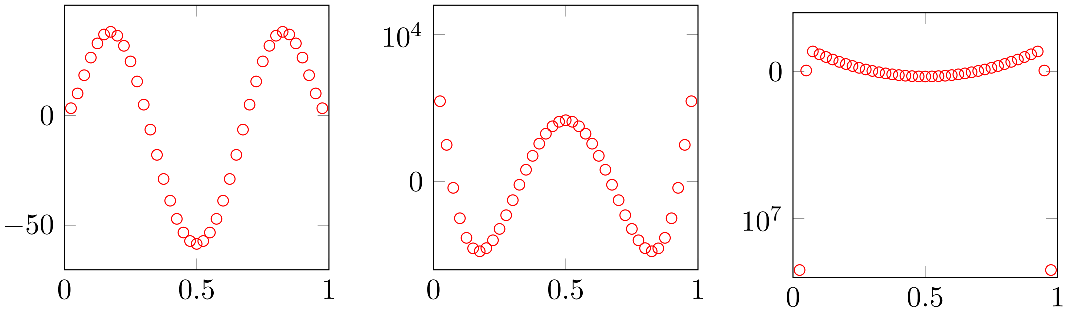


Figure 6.1: Plot of $\boldsymbol{A}\boldsymbol{v}$, $\boldsymbol{A}^2\boldsymbol{v}$ and $\boldsymbol{A}^3\boldsymbol{v}$ for $N = 39$.

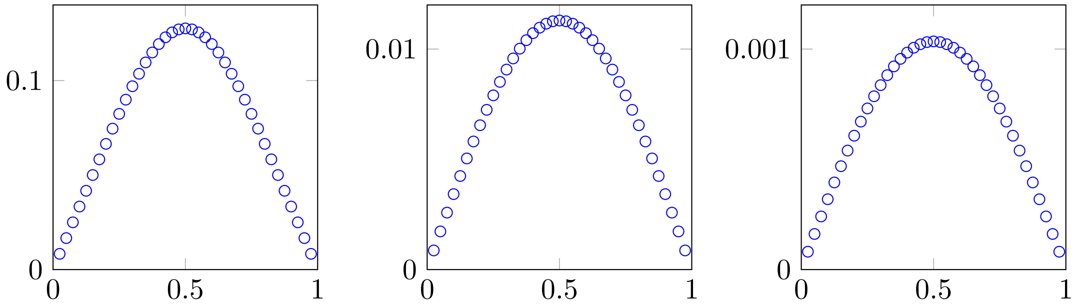


Figure 6.2: Plot of $(\gamma\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}$, $(\gamma\boldsymbol{I} - \boldsymbol{A})^{-2}\boldsymbol{v}$ and $(\gamma\boldsymbol{I} - \boldsymbol{A})^{-3}\boldsymbol{v}$ for $\gamma = 1$ and $N = 39$.

In summary, we can say that an approximation of $f(\boldsymbol{A})\boldsymbol{v}$ in the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$ is only reasonable in the case that the continuous counterpart satisfies the condition $v \in D(A^q)$. The size of the polynomial part of the extended subspace is therefore restricted by the maximal smoothness of the initial value v. In the following, we will establish error bounds for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in $\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$ that involve $\|\boldsymbol{A}^q\boldsymbol{v}\|$. For this reason, we will always assume that the vector $\boldsymbol{v}$ corresponds to a continuous value $v \in D(A^q)$, where A is the operator associated with $\boldsymbol{A}$. This requirement ensures that $\|\boldsymbol{A}^q\boldsymbol{v}\|$ does not grow for finer discretizations and yields a uniform grid-independent convergence for the extended Krylov approximation.

6.2 Motivation

The standard Krylov subspace approximation has the benefit that the computation is cheap, since we only have to evaluate matrix-vector multiplications. However, the convergence can be very slow. This is, in particular, the case if $\boldsymbol{A}$ stems from a fine space discretization and therefore has huge norm, or if the initial value is not sufficiently smooth. In contrast, it was illustrated in Chapter 5 that the shift-and-invert Krylov method has the great advantage that the convergence is independent of $\|\boldsymbol{A}\|$ and does not require any smoothness assumptions on the initial data. But nevertheless, its computation is more expensive, since we have to solve a large linear system in each iteration step. Thus, it may be worthwhile to start with some standard Krylov steps, as long as the approximation is improved, and then to continue with the rational Krylov subspace method. This leads to the idea of searching an approximation in the extended space $\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$.

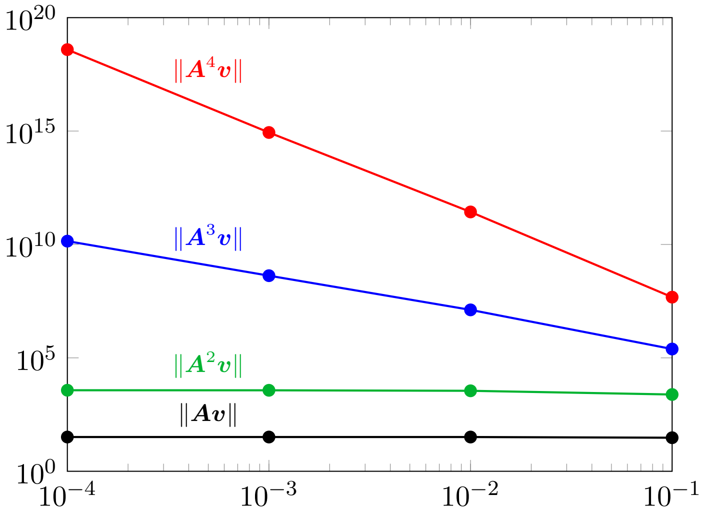


Figure 6.3: Plot of the discrete L^2-norm $\|\boldsymbol{A}^j\boldsymbol{v}\|$ for $j = 1, \dots, 4$ against the mesh size h.

To further motivate this idea, we return to the one-dimensional wave equation in Section 5.6.2. This time, we choose initial values with different smoothness properties. For the sake of simplicity, we set $f(t,x) = 0$ and consider the abstract equation

$$u'' = -Bu = \frac{\partial^2}{\partial x^2}\,u\,, \qquad u(0) = u_0\,, \qquad u'(0) = u_0'$$

on the Hilbert space $H = L^2(\Omega)$, where B is the negative Laplace operator $-\frac{\partial^2}{\partial x^2}$ with homogeneous Dirichlet boundary conditions on $\Omega = (0,1)$. The eigenvalues λ_k and eigenfunctions ψ_k of B are given by

$$\lambda_k = (k\pi)^2\,, \qquad \psi_k(x) = \sqrt{2}\sin(k\pi x)\,, \qquad k \in \mathbb{N}\,.$$

These eigenfunctions are orthonormal with respect to the L^2-inner product $(\cdot,\cdot)_{L^2(\Omega)}$. Since B is a positive and self-adjoint operator with a compact resolvent, the eigenfunctions ψ_k build an orthonormal basis of the Hilbert space H by the spectral theorem ([56], p. 95). So, every function in $L^2(\Omega)$ can be written as an infinite series of these basis functions, i.e., $u = \sum_{k=1}^\infty \widetilde{a}_k\psi_k$, with generalized Fourier coefficients $\widetilde{a}_k = (u,\psi_k)_{L^2(\Omega)}$. The action of the operator B on a function u is defined as

$$Bu = \sum_{k=1}^{\infty}(k\pi)^2\widetilde{a}_k\psi_k\,, \qquad \widetilde{a}_k = \int_0^1 u(t,x)\psi_k(x)\,dx$$

with domain

$$D(B) = \{u \in H \,:\, \sum_{k=1}^{\infty}(k\pi)^4\widetilde{a}_k^2 < \infty\}\,.$$

For $\alpha \in \mathbb{R}$, we define fractional powers by

$$B^\alpha u := \sum_{k=1}^{\infty}\lambda_k^\alpha\widetilde{a}_k\psi_k\,,$$

whose domain is

$$D(B^\alpha) = \{u \in H \,:\, \|B^\alpha u\|_{L^2(\Omega)} < \infty\}\,.$$

Setting $v = u$ and $w = B^{-1/2}u'$, we can rewrite the wave equation in an abstract form as

$$y'(t) = \begin{bmatrix} v(t) \\ w(t) \end{bmatrix}' = \begin{bmatrix} 0 & B^{1/2} \\ -B^{1/2} & 0 \end{bmatrix} \begin{bmatrix} v(t) \\ w(t) \end{bmatrix} = Ay(t)$$

with initial value

$$y(0) = y_0 = \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} = \begin{bmatrix} v(0) \\ w(0) \end{bmatrix} = \begin{bmatrix} u(0) \\ B^{-1/2}u'(0) \end{bmatrix}$$

and $D(A) = D(B^{1/2}) \times D(B^{1/2}) = H_0^1(0,1) \times H_0^1(0,1)$. As outlined in the previous subsections, we are interested in the domain $D(A^k) = D(B^{k/2}) \times D(B^{k/2})$, where $D(B^{k/2})$ contains all functions that are k times weakly differentiable and whose $(k-1)$st derivative is zero at the boundary points 0 and 1.

Let us consider the initial value

$$y_0^q = \begin{bmatrix} v_0^q \\ w_0^q \end{bmatrix}, \qquad v_0^q(x) = \frac{x^q(1-x)^q}{\|x^q(1-x)^q\|_{L^2(\Omega)}}, \qquad w_0^q(x) = 0\,,$$

for which we have $y_0^q \in D(A^q)$, but $y_0^q \notin D(A^{q+1})$, since $\frac{d^q}{dx^q}\, v_0^q(x)$ does no longer fulfill the required zero boundary conditions.

For the spectral discretization, we approximate $u = \sum_{k=1}^{\infty} \widetilde{a}_k \psi_k$ by a linear combination of the first N eigenfunctions $\psi_1, \dots, \psi_N$. More precisely, we use $u_N = \sum_{k=1}^{N} a_k \psi_k$ with coefficients a_k that have to be determined. This leads to the system of ordinary differential equations

$$\boldsymbol{y}'(t) = \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}' = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{B}^{1/2} \\ -\boldsymbol{B}^{1/2} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix} = \boldsymbol{A}\boldsymbol{y}(t)\,, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0^q = \begin{bmatrix} \boldsymbol{v}_0^q \\ \boldsymbol{w}_0^q \end{bmatrix} \tag{6.2}$$

with

$$\boldsymbol{B}^{1/2} = \operatorname{diag}(\pi, 2\pi, \dots, N\pi)\,, \qquad \boldsymbol{v}(t) = \big(a_k(t)\big)_{k=1}^N\,, \qquad \boldsymbol{w}(t) = \boldsymbol{B}^{-1/2}\big(a_k'(t)\big)_{k=1}^N\,.$$

The discretized initial values $\boldsymbol{v}_0^q$ and $\boldsymbol{w}_0^q$ contain the first N Fourier coefficients of the continuous counterparts v_0^q and w_0^q.

We now approximate the solution $e^{\tau\boldsymbol{A}}\boldsymbol{y}_0^q$ of (6.2) in the polynomial, the rational and the extended Krylov subspace for time step size $\tau = 0.1$, shift $\gamma = 1$, and $N = 1\,023$ basis functions. According to the smoothness of the initial value, the polynomial part of the extended Krylov subspace is restricted by the index q. For $q = 2, 3, 4, 5$, we see the obtained error curves plotted against the number of Krylov steps in Figure 6.4.

The polynomial steps of the extended Krylov decomposition are not performed in the indicated order of the subspace $\mathcal{K}_{q+1,m}^{\gamma}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$. Instead, we proceed from $\boldsymbol{y}_0^q$, $\boldsymbol{A}\boldsymbol{y}_0^q, \dots$ up to $\boldsymbol{A}^q\boldsymbol{y}_0^q$, such that the first $q+1$ steps coincide with the standard Krylov subspace procedure, cf. Section 6.4. Therefore, the blue solid and the red dashed line are identical at the beginning.

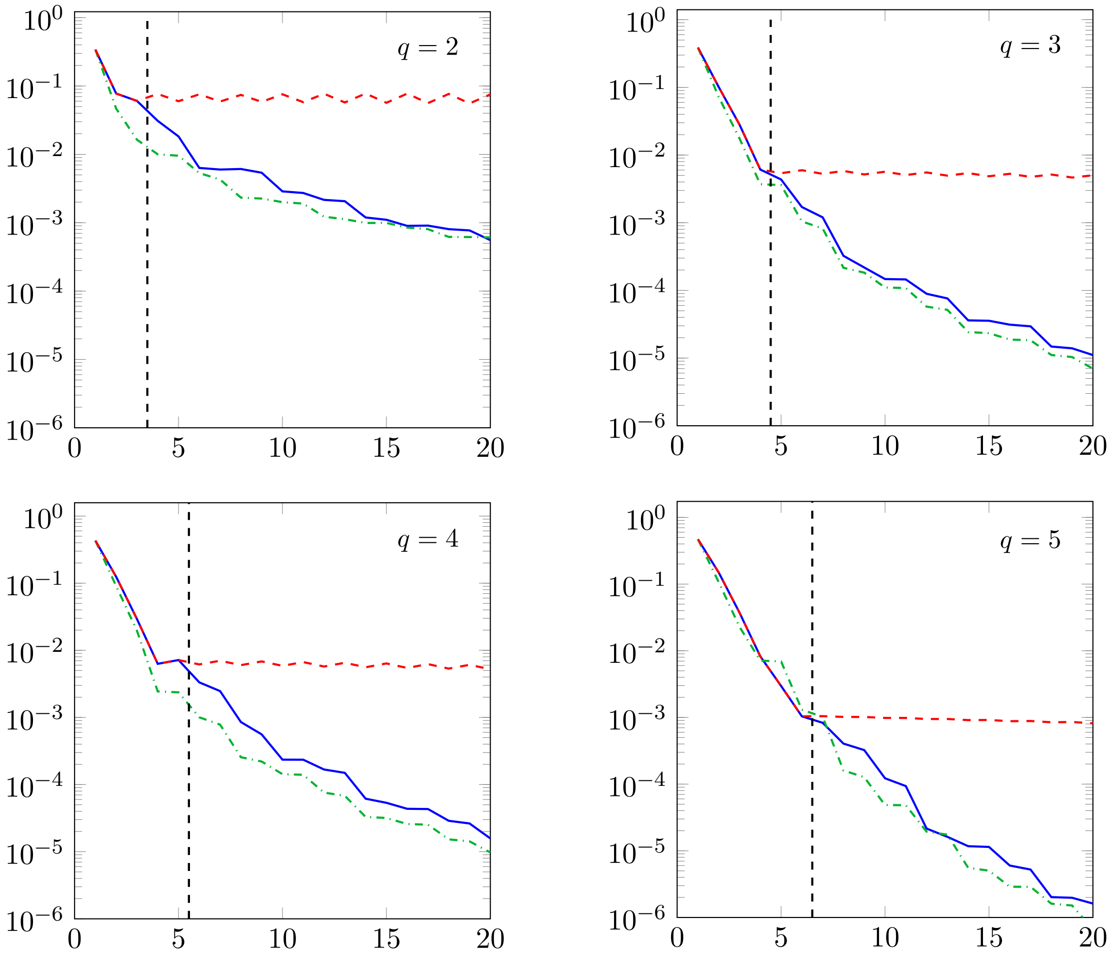


Figure 6.4: Plot of the error $\|e^{\tau \boldsymbol{A}} \boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{q+m}} \boldsymbol{y}_0^q\|_2$ versus the dimension of the Krylov subspace for the standard Krylov subspace $\mathcal{K}^{\gamma}_{q+m,1}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (red dashed line), the rational Krylov subspace $\mathcal{K}^{\gamma}_{1,q+m}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (green dash-dotted line), and the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (blue solid line) with initial vectors $\boldsymbol{y}_0^q$, $q = 2, 3, 4, 5$, whose continuous counterparts satisfy $y_0^q \in D(A^q)$ but $y_0^q \notin D(A^{q+1})$. The parameters are chosen as $\gamma = 1$, $\tau = 0.1$, and $N = 1\,023$.

A first observation is that the accuracy of the rational and the extended Krylov subspace approximation increases with the index q. The smoothness of the chosen initial function has thus an influence on the convergence rate and affects how well $e^{\tau \boldsymbol{A}} \boldsymbol{y}_0^q$ can be approximated in the considered Krylov subspaces. This dependence of the approximation quality on the smoothness of the initial value may be expected for a very fine discretization in space, in which case $\boldsymbol{A}$ represents more and more the differential operator A. It is astonishing that this behavior can be observed even for a coarse discretization using only $N = 1\,023$ basis functions.

The approximation in the polynomial Krylov subspace $\mathcal{K}^{\gamma}_{q+m,1}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ works fine until the subspace encloses vectors of the form $\boldsymbol{A}^l \boldsymbol{y}_0^q$ with $l > q$. This point is marked in Figure 6.4 by the vertical dashed line between the $(q + 1)$st standard Krylov step, that does not use the vector $\boldsymbol{A}^{q+1} \boldsymbol{y}_0^q$, and the $(q + 2)$nd step, that does use $\boldsymbol{A}^{q+1} \boldsymbol{y}_0^q$. This observation justifies the application of the extended Krylov subspace method, where we first perform $q+1$ polynomial Krylov steps and then continue with the rational Krylov subspace approximation.

The situation is a little bit different, if we use another initial value $\widetilde{y}_0^q = [\widetilde{v}_0^q, \widetilde{w}_0^q]^T$. We define the function $g(x) = x^{q+2}(x-\frac{1}{2})^q$ and set

$$\widehat{v}_0^q(x) = \begin{cases} g(x)\,, & x \in [0, \frac{1}{2}] \\ (-1)^{q+1} g(1-x)\,, & x \in (\frac{1}{2}, 1] \end{cases} \,, \qquad \widetilde{v}_0^q(x) = \frac{\widehat{v}_0^q(x)}{\|\widehat{v}_0^q(x)\|_{L^2(\Omega)}}\,.$$

This function can be differentiated q times and the qth weak derivative has a discontinuity at $x = \frac{1}{2}$, that is, $\widetilde{y}_0^q \in D(A^q)$ and $\widetilde{y}_0^q \notin D(A^{q+1})$. As before, the second component of the initial value is chosen as $\widetilde{w}_0^q = 0$.

In Figure 6.5, we see that the polynomial Krylov method might stagnate before the vector $\boldsymbol{A}^q \widetilde{\boldsymbol{y}}_0^q$ is used for the approximation of $e^{\tau \boldsymbol{A}} \widetilde{\boldsymbol{y}}_0^q$. This happens one iteration step earlier as expected, namely in the qth instead of the $(q+1)$st step. Nevertheless, this imposes no problem, since the abstract smoothness of the initial value is often not known in advance. In fact, we rely on heuristic methods to determine the point, where the polynomial method achieves no further improvement and where we should start to use the rational approximation. Two heuristic approaches will be presented in Section 6.5.1.

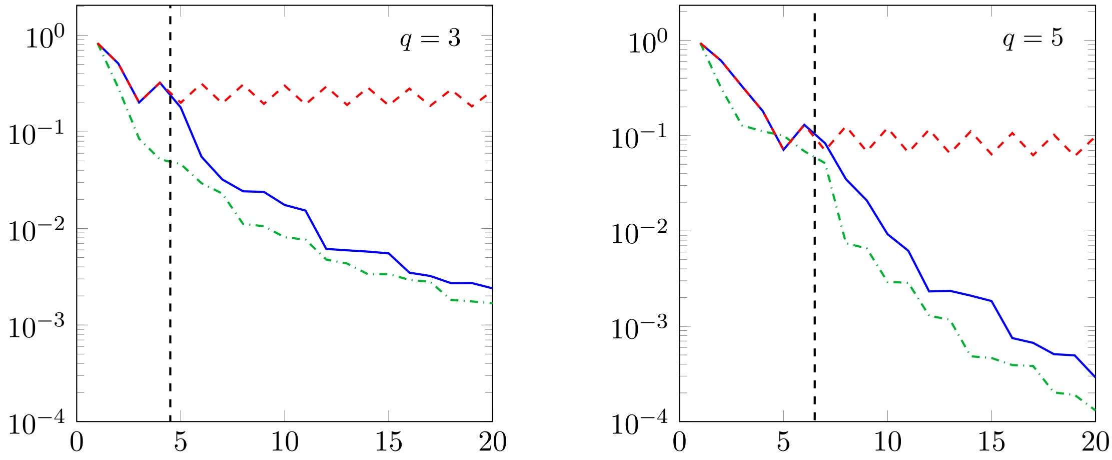


Figure 6.5: Plot of the error $\|e^{\tau \boldsymbol{A}} \widetilde{\boldsymbol{y}}_0^q - e^{\tau \boldsymbol{A}_{q+m}} \widetilde{\boldsymbol{y}}_0^q\|_2$ versus the dimension of the Krylov subspace for the standard Krylov subspace $\mathcal{K}^{\gamma}_{q+m,1}(\tau \boldsymbol{A}, \widetilde{\boldsymbol{y}}_0^q)$ (red dashed line), the rational Krylov subspace $\mathcal{K}^{\gamma}_{1,q+m}(\tau \boldsymbol{A}, \widetilde{\boldsymbol{y}}_0^q)$ (green dash-dotted line), and the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\tau \boldsymbol{A}, \widetilde{\boldsymbol{y}}_0^q)$ (blue solid line) with initial vectors $\widetilde{\boldsymbol{y}}_0^q$, $q = 3, 5$, whose continuous counterparts satisfy $\widetilde{y}_0^q \in D(A^q)$ but $\widetilde{y}_0^q \notin D(A^{q+1})$. The parameters are chosen as $\gamma = 1$, $\tau = 0.1$, and $N = 1\,023$.

Applying the extended Krylov subspace process, we can achieve a speed-up, since the standard Krylov method is, in general, cheap, due to the fact that only matrix-vector multiplications have to be performed. In contrast, the rational approximation usually performs better but is more expensive, because the rational Krylov decomposition requires the solution of a linear system $(\gamma \boldsymbol{I} - \tau \boldsymbol{A})^{-1} \boldsymbol{v}$ in each iteration step.

6.3 Error bounds

The derivation of error bounds for the extended Krylov subspace approximation strongly relies on the results of the shift-and-invert Krylov subspace method. As in Chapter 5, we assume that the matrix $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ satisfies

$$W(\boldsymbol{A}) \subseteq \mathbb{C}_0^- \tag{6.3}$$

with respect to some inner product on $\mathbb{C}^N$. Given a Hilbert space H with inner product $(\cdot\,,\cdot)$, such matrices typically originate from a differential operator A with $\operatorname{Re}(Av, v) \leq 0$ for all $v \in D(A)$ and $\operatorname{Range}(\lambda I - A) = H$ for some λ in the right complex half-plane, such that A generates a strongly continuous contraction semigroup on H by the Lumer-Phillips Theorem 3.7. Since the discretized operator $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ retains properties inherent to the continuous counterpart A, the assumption (6.3) is justified for a suitable spatial discretization. According to the discussion above, we state the following assumption on the initial vector $\boldsymbol{v}$.

Assumption 6.1 *Let $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ be a discretization matrix of some differential operator A including boundary conditions, which generates a strongly continuous contraction semigroup on H. Then we assume that the vector $\boldsymbol{v} \in \mathbb{C}^N$ stems from the discretization of a function $v \in D(A^q) \subseteq H$. Moreover, we suppose that the discretization is suitably chosen such that the smoothness requirements of the corresponding abstract problem are well reflected in the discrete case, that means $\|\boldsymbol{A}^q \boldsymbol{v}\| \leq K \|A^q v\|$, $v \in D(A^q)$, for a constant K which does not depend on the spatial grid.*

This assumption ensures that $\|\boldsymbol{A}^q \boldsymbol{v}\|$ is bounded irrespectively of the spatial mesh size and later guarantees that our error bounds hold uniformly, that is, independent of a refinement of the discretization in space.

Example 6.2 We consider a differential operator A on $L^2(\Omega)$, $\Omega = (0,1)$, that is given as the negative second derivative operator including homogeneous Dirichlet boundary conditions. Moreover, we take a vector $v \in D(A^q)$ which we represent in the eigenbasis $\psi_k(x) = \sqrt{2}\sin(k\pi x)$ of A via $v = \sum_{k=1}^{\infty} v_k(t)\psi_k(x)$. Using a spectral discretization with N ansatz function, we obtain the discretization matrix $\boldsymbol{A} = \operatorname{diag}(\pi^2, (2\pi)^2, \ldots, (N\pi)^2)$ and the vector $\boldsymbol{v} = \big(v_k(t)\big)_{k=1}^{N}$. In this case it is easy to see that

$$\|\boldsymbol{A}^q \boldsymbol{v}\|_2 = \left(\sum_{k=1}^{N} \left((k\pi)^{2q} v_k(t) \right)^2 \right)^{\frac{1}{2}} \leq \left(\sum_{k=1}^{\infty} \left((k\pi)^{2q} v_k(t) \right)^2 \right)^{\frac{1}{2}} = \|A^q v\|_{L^2(\Omega)} \,,$$

i.e., the inequality $\|\boldsymbol{A}^q \boldsymbol{v}\| \leq K \|A^q v\|$ in Assumption 6.1 holds with $K = 1$. ❍

We aim to derive error bounds for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$, $\ell \geq 0$, in the extended Krylov subspace $\mathcal{K}_{q+1,m}^{\gamma}(\boldsymbol{A}, \boldsymbol{v})$. A first step to this end is the following lemma that links the best approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the extended Krylov subspace $\mathcal{K}_{q+1,m}^{\gamma}(\boldsymbol{A}, \boldsymbol{v})$ with the best approximation of $\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v}$ in the rational space $\mathcal{K}_{1,m}^{\gamma}(\boldsymbol{A}, \boldsymbol{v})$. This relation makes it possible to fall back on the results that we established in Chapter 5 for the shift-and-invert Krylov subspace approximation. This time, we apply the bounds to the product of the matrix function $\varphi_{q+\ell}(\boldsymbol{A})$ and the vector $\boldsymbol{A}^q\boldsymbol{v}$.

Lemma 6.3 *Let $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ be a matrix with $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$. Then*

$$\inf_{\boldsymbol{z}\in\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})} \|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{z}\| \le \inf_{\boldsymbol{y}\in\mathcal{K}^{\gamma}_{1,m}(\boldsymbol{A},\boldsymbol{A}^q\boldsymbol{v})} \|\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \boldsymbol{y}\|$$

holds for any φ_ℓ-function with $\ell \ge 0$.

Proof. Our aim is to show that for any coefficients $b_0, \dots, b_{m-1} \in \mathbb{C}$ in (6.4), one can find coefficients $a_1, \dots, a_{m-1} \in \mathbb{C}$ and a polynomial $p_q \in \mathcal{P}_q$, so that

$$\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \underbrace{\sum_{k=0}^{m-1} b_k \frac{1}{(\gamma-\boldsymbol{A})^k}\boldsymbol{A}^q\boldsymbol{v}}_{\in\,\mathcal{K}^{\gamma}_{1,m}(\boldsymbol{A},\boldsymbol{A}^q\boldsymbol{v})} = \varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \underbrace{p_q(\boldsymbol{A})\boldsymbol{v} - \sum_{k=1}^{m-1} a_k \frac{1}{(\gamma-\boldsymbol{A})^k}\boldsymbol{v}}_{\in\,\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})}\,. \tag{6.4}$$

For this purpose, we consider the corresponding scalar problem. Since

$$p_q(z) - \sum_{k=1}^{m-1} a_k \frac{1}{(\gamma-z)^k} \in \frac{\mathcal{P}_{q+m-1}}{q_{m-1}}\,, \qquad q_{m-1}(z) = (\gamma-z)^{m-1}\,,$$

we have to prove that, for any polynomial $p_{m-1} \in \mathcal{P}_{m-1}$, we can find a polynomial $p_{q+m-1} \in \mathcal{P}_{q+m-1}$ with

$$\left(\varphi_{q+\ell}(z) - \frac{p_{m-1}(z)}{(\gamma-z)^{m-1}}\right) z^q = \varphi_\ell(z) - \frac{p_{q+m-1}(z)}{(\gamma-z)^{m-1}}\,.$$

Using formula (3.13) for the φ-functions, we obtain

$$\varphi_{q+\ell}(z)z^q = \frac{1}{z^{q+\ell}}\left(e^z - \sum_{k=0}^{q+\ell-1}\frac{z^k}{k!}\right) z^q = \varphi_\ell(z) - \sum_{k=0}^{q-1}\frac{z^k}{(k+\ell)!} = \varphi_\ell(z) - p_{q-1}(z)$$

for some polynomial $p_{q-1} \in \mathcal{P}_{q-1}$. This gives

$$\begin{aligned}\left(\varphi_{q+\ell}(z) - \frac{p_{m-1}(z)}{(\gamma-z)^{m-1}}\right) z^q &= \varphi_{q+\ell}(z)z^q - \frac{p_{m-1}(z)}{(\gamma-z)^{m-1}}\,z^q \\ &= \varphi_\ell(z) - p_{q-1}(z) - \frac{p_{m-1}(z)}{(\gamma-z)^{m-1}}\,z^q \\ &= \varphi_\ell(z) - \frac{p_{q+m-1}(z)}{(\gamma-z)^{m-1}}\end{aligned}$$

and thus the validity of the desired relation (6.4). Since (6.4) says that every element in $\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} + \mathcal{K}^{\gamma}_{1,m}(\boldsymbol{A},\boldsymbol{A}^q\boldsymbol{v})$ can be expressed as an element in $\varphi_\ell(\boldsymbol{A})\boldsymbol{v} + \mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$, we deduce the inclusion

$$\{\|\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \boldsymbol{y}\| \,:\, \boldsymbol{y}\in\mathcal{K}^{\gamma}_{1,m}(\boldsymbol{A},\boldsymbol{A}^q\boldsymbol{v})\} \subseteq \{\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{z}\| \,:\, \boldsymbol{z}\in\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})\}\,.$$

This yields the statement of the lemma. ❏

According to Lemma 6.3, it is now possible to reduce the problem of bounding the best approximation in the extended Krylov subspace to a problem of finding an estimate for the best approximation in the shift-and-invert Krylov subspace. This results in the following theorem.

Theorem 6.4 *Suppose that $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$, and let $\boldsymbol{A}$ and $\boldsymbol{v}$ fulfill Assumption 6.1. Then the best approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$, $\ell \geq 0$, in the extended Krylov subspace $\mathcal{K}^\gamma_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$ is uniformly bounded by*

$$\inf_{\boldsymbol{z}\in\mathcal{K}^\gamma_{q+1,m}(\boldsymbol{A},\boldsymbol{v})} \|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{z}\| \leq \frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}} \|\boldsymbol{A}^q\boldsymbol{v}\| \leq K\, \frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}} \|A^q v\| \,,$$

where the constants K and $C(\ell,q,\gamma)$ are independent of the spatial grid.

Proof. For $\boldsymbol{y} = \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma-\boldsymbol{A})^k} \boldsymbol{A}^q\boldsymbol{v} \in \mathcal{K}^\gamma_{1,m}(\boldsymbol{A},\boldsymbol{A}^q\boldsymbol{v})$, we have

$$\|\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \boldsymbol{y}\| \leq \left\| \varphi_{q+\ell}(\boldsymbol{A}) - \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma-\boldsymbol{A})^k} \right\| \|\boldsymbol{A}^q\boldsymbol{v}\| \leq \frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}} \|\boldsymbol{A}^q\boldsymbol{v}\| \,,$$

by choosing the coefficients $b_0,\ldots,b_{m-1}$ according to relation (5.13) above. With Lemma 6.3, we obtain

$$\inf_{\boldsymbol{z}\in\mathcal{K}^\gamma_{q+1,m}(\boldsymbol{A},\boldsymbol{v})} \|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{z}\| \leq \inf_{\boldsymbol{y}\in\mathcal{K}^\gamma_{1,m}(\boldsymbol{A},\boldsymbol{A}^q\boldsymbol{v})} \|\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \boldsymbol{y}\| \leq \frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}} \|\boldsymbol{A}^q\boldsymbol{v}\| \,,$$

which proves the first inequality. The second inequality is just a consequence of Assumption 6.1 on $\boldsymbol{v}$ and $\boldsymbol{A}$. ❑

With "uniformly bounded" in Theorem 6.4, we mean that the error bound holds true for arbitrary matrices with a field of values in the left complex half-plane and that the bound is independent of the refinement of the space discretization. The latter is guaranteed by our assumption on the discretization matrix $\boldsymbol{A}$ and the initial vector $\boldsymbol{v}$, which is due to the close connection between the continuous expression $A^q v \in H$, that is only defined for $v \in D(A^q)$, and its discrete counterpart $\boldsymbol{A}^q\boldsymbol{v} \in \mathbb{C}^N$, as explained in Section 6.1. The smoothness of the initial data plays no important role for a coarse discretization, but is indispensable for fine discretizations. If we waive the requirement that $\boldsymbol{v}$ and $\boldsymbol{A}$ have to satisfy Assumption 6.1, the first inequality in Theorem 6.4 is still valid, since we always have $\|\boldsymbol{A}^q\boldsymbol{v}\| < \infty$ in the discrete case, compared with $\|A^q v\|$ that is only bounded for $v \in D(A^q)$. However, depending on the spatial mesh, $\|\boldsymbol{A}^q\boldsymbol{v}\|$ can become arbitrarily large.

In contrast to the best approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the extended Krylov subspace, which is usually unknown, the extended Krylov subspace approximation defined by

$$\varphi_\ell(\boldsymbol{A})\boldsymbol{v} \approx \varphi_\ell(\boldsymbol{A}_{q+m})\boldsymbol{v} \,, \qquad \boldsymbol{A}_{q+m} = \boldsymbol{P}_{q+m}\boldsymbol{A}\boldsymbol{P}_{q+m} \,,$$

can be computed efficiently. As before, $\boldsymbol{P}_{q+m}$ is the orthogonal projection onto $\mathcal{K}^\gamma_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$ with respect to the chosen inner product on $\mathbb{C}^N$. To bound $\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_{q+m})\boldsymbol{v}\|$, we need the fact that the extended Krylov subspace approximation is exact for all functions belonging to $\mathcal{P}_{q+m-1}/q_{m-1}$ with $q_{m-1}(z) = (\gamma - z)^{m-1}$, cf. Lemma 4.12.

Lemma 6.5 *Let $\boldsymbol{P}_{q+m}$ be the orthogonal projection onto the extended Krylov subspace $\mathcal{K}^\gamma_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$. Then for any $p_q \in \mathcal{P}_q$ and arbitrary coefficients $a_1,\ldots,a_{m-1}$, we have*

$$p_q(\boldsymbol{A})\boldsymbol{v} - \sum_{k=1}^{m-1} a_k \frac{1}{(\gamma-\boldsymbol{A})^k}\boldsymbol{v} = p_q(\boldsymbol{A}_{q+m})\boldsymbol{v} - \sum_{k=1}^{m-1} a_k \frac{1}{(\gamma-\boldsymbol{A}_{q+m})^k}\boldsymbol{v} \,,$$

where $\boldsymbol{A}_{q+m} = \boldsymbol{P}_{q+m}\boldsymbol{A}\boldsymbol{P}_{q+m}$.

Proof. Since $\boldsymbol{P}_{q+m}$ is the orthogonal projection onto the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$, we have the relations

$$\begin{aligned} \boldsymbol{P}_{q+m}\boldsymbol{A}^k\boldsymbol{v} &= \boldsymbol{A}^k\boldsymbol{v}\,, && 0 \le k \le q\,, \\ \boldsymbol{P}_{q+m}\,\frac{1}{(\gamma-\boldsymbol{A})^k}\,\boldsymbol{v} &= \frac{1}{(\gamma-\boldsymbol{A})^k}\,\boldsymbol{v}\,, && 1 \le k \le m-1\,. \end{aligned} \tag{6.5}$$

These properties of $\boldsymbol{P}_{q+m}$ immediately yield

$$\boldsymbol{A}^k\boldsymbol{v} = \boldsymbol{P}_{q+m}\boldsymbol{A}^k\boldsymbol{v} = \boldsymbol{P}_{q+m}\boldsymbol{A}^k\boldsymbol{P}_{q+m}\boldsymbol{v} = (\boldsymbol{P}_{q+m}\boldsymbol{A}\boldsymbol{P}_{q+m})^k\boldsymbol{v} = \boldsymbol{A}^k_{q+m}\boldsymbol{v}\,.$$

Thus, it remains to be shown that $(\gamma\boldsymbol{I}-\boldsymbol{A})^{-k}\boldsymbol{v} = (\gamma\boldsymbol{I}-\boldsymbol{A}_{q+m})^{-k}\boldsymbol{v}$ for $1 \le k \le m-1$. With the help of (6.5) and $\frac{\gamma}{\gamma-\boldsymbol{A}} - \boldsymbol{I} = \frac{\boldsymbol{A}}{\gamma-\boldsymbol{A}}$, the assertion is verified for $k=1$ by

$$\begin{aligned} \boldsymbol{v} &= \frac{\gamma}{\gamma-\boldsymbol{A}}\,\boldsymbol{v} - \boldsymbol{P}_{q+m}\left(\frac{\gamma}{\gamma-\boldsymbol{A}} - \boldsymbol{I}\right)\boldsymbol{v} = \frac{\gamma}{\gamma-\boldsymbol{A}}\,\boldsymbol{v} - \boldsymbol{P}_{q+m}\,\frac{\boldsymbol{A}}{\gamma-\boldsymbol{A}}\,\boldsymbol{v} \\ &= \frac{\gamma}{\gamma-\boldsymbol{A}}\,\boldsymbol{v} - \boldsymbol{P}_{q+m}\boldsymbol{A}\boldsymbol{P}_{q+m}\,\frac{1}{\gamma-\boldsymbol{A}}\,\boldsymbol{v} = (\gamma\boldsymbol{I}-\boldsymbol{A}_{q+m})\,\frac{1}{\gamma-\boldsymbol{A}}\,\boldsymbol{v} \end{aligned}$$

and multiplying both sides from the left with $(\gamma\boldsymbol{I}-\boldsymbol{A}_{q+m})^{-1}$. For $\gamma > 0$, this inverse exists, since analogously to Lemma 5.6, one can easily show that $\sigma(\boldsymbol{A}_{q+m}) \subseteq W(\boldsymbol{A}_{q+m}) \subseteq \mathbb{C}^-_0$. Via induction with hypothesis $(\gamma\boldsymbol{I}-\boldsymbol{A})^{-k}\boldsymbol{v} = (\gamma\boldsymbol{I}-\boldsymbol{A}_{q+m})^{-k}\boldsymbol{v}$ for some $k \le m-2$, we obtain with (6.5) that

$$\begin{aligned} \frac{1}{(\gamma-\boldsymbol{A}_{q+m})^k}\,\boldsymbol{v} &= \frac{1}{(\gamma-\boldsymbol{A})^k}\,\boldsymbol{v} \\ &= \boldsymbol{P}_{q+m}\,\frac{1}{(\gamma-\boldsymbol{A})^k}\,\boldsymbol{v} + \gamma\,\frac{1}{(\gamma-\boldsymbol{A})^{k+1}}\,\boldsymbol{v} - \gamma\boldsymbol{P}_{q+m}\,\frac{1}{(\gamma-\boldsymbol{A})^{k+1}}\,\boldsymbol{v} \\ &= \gamma\,\frac{1}{(\gamma-\boldsymbol{A})^{k+1}}\,\boldsymbol{v} - \boldsymbol{P}_{q+m}\,\frac{1}{(\gamma-\boldsymbol{A})^k}\left(\frac{\gamma}{\gamma-\boldsymbol{A}} - \boldsymbol{I}\right)\boldsymbol{v} \\ &= \gamma\,\frac{1}{(\gamma-\boldsymbol{A})^{k+1}}\,\boldsymbol{v} - \boldsymbol{P}_{q+m}\boldsymbol{A}\boldsymbol{P}_{q+m}\,\frac{1}{(\gamma-\boldsymbol{A})^{k+1}}\,\boldsymbol{v} \\ &= (\gamma\boldsymbol{I}-\boldsymbol{A}_{q+m})\,\frac{1}{(\gamma-\boldsymbol{A})^{k+1}}\,\boldsymbol{v}\,, \end{aligned}$$

which is equivalent to

$$\frac{1}{(\gamma-\boldsymbol{A}_{q+m})^{k+1}}\,\boldsymbol{v} = \frac{1}{(\gamma-\boldsymbol{A})^{k+1}}\,\boldsymbol{v}$$

and thus proves the desired claim. ❑

Theorem 6.6 *Let $\boldsymbol{A}$ be a matrix with $W(\boldsymbol{A}) \subseteq \mathbb{C}^-_0$ and assume that $\boldsymbol{v}$ and $\boldsymbol{A}$ satisfy Assumption 6.1. Moreover, let $\boldsymbol{P}_{q+m}$ be the orthogonal projection onto the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\boldsymbol{A},\boldsymbol{v})$ and $\boldsymbol{A}_{q+m} = \boldsymbol{P}_{q+m}\boldsymbol{A}\boldsymbol{P}_{q+m}$. Then*

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_{q+m})\boldsymbol{v}\| \le \frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}}\,\|\boldsymbol{A}^q\boldsymbol{v}\| \le K\,\frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}}\,\|A^q v\|\,,$$

holds uniformly for $\ell \ge 0$, where the constants K and $C(\ell,q,\gamma)$ are independent of the space discretization.

Proof. Because of $W(\boldsymbol{A}_{q+m}) \subseteq \mathbb{C}_0^-$, the matrix $\boldsymbol{A}_{q+m}$ fits in our framework and we are able to apply the previous results. We choose $\boldsymbol{z} \in \mathcal{K}_{q+1,m}^{\gamma}(\boldsymbol{A},\boldsymbol{v}) = \mathcal{K}_{q+1,m}^{\gamma}(\boldsymbol{A}_{q+m},\boldsymbol{v})$ with

$$\boldsymbol{z} = p_q(\boldsymbol{A})\boldsymbol{v} - \sum_{k=1}^{m-1} a_k \frac{1}{(\gamma - \boldsymbol{A})^k}\,\boldsymbol{v} = p_q(\boldsymbol{A}_{q+m})\boldsymbol{v} - \sum_{k=1}^{m-1} a_k \frac{1}{(\gamma - \boldsymbol{A}_{q+m})^k}\,\boldsymbol{v}\,,$$

cf. Lemma 6.5, such that, by relation (6.4), there exist coefficients $b_0,\dots,b_{m-1}$ with

$$\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{z} = \varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma - \boldsymbol{A})^k}\,\boldsymbol{A}^q\boldsymbol{v}$$

and

$$\varphi_\ell(\boldsymbol{A}_{q+m})\boldsymbol{v} - \boldsymbol{z} = \varphi_{q+\ell}(\boldsymbol{A}_{q+m})\boldsymbol{A}_{q+m}^q\boldsymbol{v} - \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma - \boldsymbol{A}_{q+m})^k}\,\boldsymbol{A}_{q+m}^q\boldsymbol{v}\,.$$

Using this, it follows

$$\begin{aligned}
\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_{q+m})\boldsymbol{v}\| &\le \|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{z}\| + \|\varphi_\ell(\boldsymbol{A}_{q+m})\boldsymbol{v} - \boldsymbol{z}\| \\
&= \left\|\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma - \boldsymbol{A})^k}\,\boldsymbol{A}^q\boldsymbol{v}\right\| \\
&\quad + \left\|\varphi_{q+\ell}(\boldsymbol{A}_{q+m})\boldsymbol{A}_{q+m}^q\boldsymbol{v} - \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma - \boldsymbol{A}_{q+m})^k}\,\boldsymbol{A}_{q+m}^q\boldsymbol{v}\right\| \\
&\le \left\|\varphi_{q+\ell}(\boldsymbol{A}) - \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma - \boldsymbol{A})^k}\right\| \|\boldsymbol{A}^q\boldsymbol{v}\| \\
&\quad + \left\|\varphi_{q+\ell}(\boldsymbol{A}_{q+m}) - \sum_{k=0}^{m-1} b_k \frac{1}{(\gamma - \boldsymbol{A}_{q+m})^k}\right\| \|\boldsymbol{A}^q\boldsymbol{v}\|\,,
\end{aligned}$$

since $\|\boldsymbol{A}_{q+m}^q\boldsymbol{v}\| = \|\boldsymbol{A}^q\boldsymbol{v}\|$ by Lemma 6.5. The coefficients $b_0,\dots,b_{m-1}$ are now selected according to the rational function r^* in relation (5.13) for the shift-and-invert Krylov subspace approximation. Since r^* is independent of the matrix argument, as long as the matrix has a field of values in $\mathbb{C}_0^-$, the estimate (5.13) can be applied to the first term containing $\boldsymbol{A}$ as well as to the second term with $\boldsymbol{A}_{q+m}$. Hence, we obtain

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_{q+m})\boldsymbol{v}\| \le \frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}}\,\|\boldsymbol{A}^q\boldsymbol{v}\| \le K\,\frac{C(\ell,q,\gamma)}{m^{\frac{q+\ell}{2}}}\,\|A^q v\|\,,$$

where we used Assumption 6.1 for the last inequality. ❏

Compared to the shift-and-invert Krylov subspace method in Chapter 5, we obtained the same sublinear convergence rate as if we would approximate $\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{v}$ instead of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$.

The statements above show that the extended Krylov subspace approximation is traced back to a rational approximation in the shift-and-invert Krylov subspace. Because of the inequality

$$\inf_{\boldsymbol{z}\in\mathcal{K}_{q+1,m}^{\gamma}(\boldsymbol{A},\boldsymbol{v})} \|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \boldsymbol{z}\| \le \inf_{\boldsymbol{y}\in\mathcal{K}_{1,m}^{\gamma}(\boldsymbol{A},\boldsymbol{A}^q\boldsymbol{v})} \|\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v} - \boldsymbol{y}\|$$

from Lemma 6.3, we can exploit the results for the shift-and-invert Krylov subspace approximation. In particular, we are able to refer to Section 5.4, where suitable choices of the shift $\gamma > 0$ were discussed. We have seen, that an improvement of the convergence rate is obtained by choosing the shift depending on the dimension m of the rational Krylov subspace, i.e., $\gamma = m^\alpha$ with an appropriate exponent α. The optimal γ for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the shift-and-invert subspace $\mathcal{K}_m\big((\gamma\boldsymbol{I} - \boldsymbol{A})^{-1}, \boldsymbol{v}\big) = \mathcal{K}_{1,m}^\gamma(\boldsymbol{A}, \boldsymbol{v})$ was given for $\alpha = (r - \frac{\ell}{2})/(r + \frac{\ell}{2})$ with $r > \frac{\ell}{2} + 1$. Since the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the extended space $\mathcal{K}_{q+1,m}^\gamma(\boldsymbol{A}, \boldsymbol{v})$ is related to the approximation of $\varphi_{q+\ell}(\boldsymbol{A})\boldsymbol{A}^q\boldsymbol{v}$ in the shift-and-invert space $\mathcal{K}_{1,m}^\gamma(\boldsymbol{A}, \boldsymbol{A}^q\boldsymbol{v})$, we now have

$$\gamma = m^\alpha\,, \qquad \alpha = \frac{r - \frac{q+\ell}{2}}{r + \frac{q+\ell}{2}}\,, \qquad r > \frac{q+\ell}{2} + 1\,,$$

which gives the improved convergence rate

$$\mathcal{O}\left(m^{-\frac{q+\ell}{2}(1+\alpha)}\right)\,.$$

6.4 Computation of the extended Krylov subspace approximation

The first step of the extended Krylov subspace method is to determine an orthonormal basis of $\mathcal{K}_{q+1,m}^\gamma(\boldsymbol{A}, \boldsymbol{v})$. For this purpose, we use Algorithm 6.7 that combines the standard and rational Arnoldi algorithm.

As already mentioned in the previous chapter, we are actually interested in the approximation of $\varphi_\ell(\tau\boldsymbol{A})\boldsymbol{v}$, where $\tau > 0$ denotes the time step size of the numerical time integration scheme. In this case, we have the error estimate

$$\|\varphi_\ell(\tau\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\tau\boldsymbol{A}_{q+m})\boldsymbol{v}\| \le \frac{C(\ell, q, \gamma)}{m^{\frac{q+\ell}{2}}}\,\tau^q\|\boldsymbol{A}^q\boldsymbol{v}\|\,, \qquad \ell \ge 0\,.$$

We thus search for an approximation $\varphi_\ell(\tau\boldsymbol{A}_{q+m})\boldsymbol{v}$ to $\varphi_\ell(\tau\boldsymbol{A})\boldsymbol{v}$ in the extended subspace $\mathcal{K}_{q+1,m}^\gamma(\tau\boldsymbol{A}, \boldsymbol{v})$. Since the polynomial part of this subspace fulfills

$$\mathcal{K}_{q+1,1}^\gamma(\tau\boldsymbol{A}, \boldsymbol{v}) = \mathcal{K}_{q+1,1}^\gamma(\boldsymbol{A}, \boldsymbol{v})\,,$$

it is common practice to exclude τ in the polynomial Arnoldi process. On the other hand, we have $(\gamma\boldsymbol{I} - \tau\boldsymbol{A})^{-1} = \frac{1}{\tau}(\frac{\gamma}{\tau}\boldsymbol{I} - \boldsymbol{A})^{-1}$ and therefore

$$\mathcal{K}_{1,m}^{\gamma/\tau}(\boldsymbol{A}, \boldsymbol{v}) = \mathcal{K}_{1,m}^\gamma(\tau\boldsymbol{A}, \boldsymbol{v}) \ne \mathcal{K}_{1,m}^\gamma(\boldsymbol{A}, \boldsymbol{v})\,.$$

This is why we cannot drop the time step size τ in the rational part of the extended Arnoldi process. Otherwise, we would have to take the scaled shift $\frac{\gamma}{\tau}$. We thus compute an orthonormal basis of

$$\begin{aligned}\mathcal{K}_{q+1,m}^\gamma(\tau\boldsymbol{A}, \boldsymbol{v}) &= \operatorname{span}\left\{\boldsymbol{A}^q\boldsymbol{v}, \dots, \boldsymbol{A}\boldsymbol{v}\,, \boldsymbol{v}, \frac{1}{\gamma - \tau\boldsymbol{A}}\,\boldsymbol{v}, \dots, \frac{1}{(\gamma - \tau\boldsymbol{A})^{m-1}}\,\boldsymbol{v}\right\}\\ &= \mathcal{K}_{q+1,1}^\gamma(\boldsymbol{A}, \boldsymbol{v}) + \mathcal{K}_{1,m}^\gamma(\tau\boldsymbol{A}, \boldsymbol{v})\,.\end{aligned}$$

Algorithm 6.7 Extended Arnoldi process

```
given: $\boldsymbol{A} \in \mathbb{C}^{N\times N},\ \boldsymbol{v} \in \mathbb{C}^N,\ \gamma > 0$
$\boldsymbol{v}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|$
for $m = 1, 2, \dots$ do
    if $m \le q$
        for $j = 1, \dots, m$ do
            $h_{j,m} = (\boldsymbol{A}\boldsymbol{v}_m, \boldsymbol{v}_j)$
        end for
        $\widetilde{\boldsymbol{v}}_{m+1} = \boldsymbol{A}\boldsymbol{v}_m - \sum_{j=1}^m h_{j,m}\boldsymbol{v}_j$
        $h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|$
        $\boldsymbol{v}_{m+1} = \widetilde{\boldsymbol{v}}_{m+1}/h_{m+1,m}$
    else
        for $j = 1, \dots, m$ do
            $h_{j,m} = \big((\gamma\boldsymbol{I} - \tau\boldsymbol{A})^{-1}\boldsymbol{v}_m, \boldsymbol{v}_j\big)$
        end for
        $\widetilde{\boldsymbol{v}}_{m+1} = (\gamma\boldsymbol{I} - \tau\boldsymbol{A})^{-1}\boldsymbol{v}_m - \sum_{j=1}^m h_{j,m}\boldsymbol{v}_j$
        $h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|$
        $\boldsymbol{v}_{m+1} = \widetilde{\boldsymbol{v}}_{m+1}/h_{m+1,m}$
    end if
end for
```

According to Algorithm 6.7, the orthonormal basis of the extended Krylov subspace is not calculated in the indicated order $\boldsymbol{A}^q\boldsymbol{v}$, $\boldsymbol{A}^{q-1}\boldsymbol{v}, \dots$ up to $(\gamma\boldsymbol{I} - \tau\boldsymbol{A})^{-m+1}\boldsymbol{v}$. Instead, we proceed as in the standard Arnoldi process: We first take the initial vector $\boldsymbol{v}$ and normalize this vector to obtain $\boldsymbol{v}_1$. Then we orthogonalize $\boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}^2\boldsymbol{v}, \dots, \boldsymbol{A}^q\boldsymbol{v}$ successively against all previously computed basis vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots$. If the iteration index $q+1$ is reached, we switch to the rational Arnoldi method and use the last computed vector $\boldsymbol{v}_{q+1}$ of the polynomial Arnoldi decomposition as starting vector. This process yields an orthonormal basis $\boldsymbol{V}_{q+m} = [\boldsymbol{v}_1\, \boldsymbol{v}_2 \cdots \boldsymbol{v}_{q+m}] \in \mathbb{C}^{N\times(q+m)}$ of $\mathcal{K}^{\gamma}_{q+1,m}(\tau\boldsymbol{A}, \boldsymbol{v})$ that is orthogonal with respect to the chosen inner product on $\mathbb{C}^N$.

Like before, for a function f analytic on $W(\boldsymbol{A})$ and the standard Euclidean inner product, the extended Krylov subspace approximation is computed as

$$f(\boldsymbol{A})\boldsymbol{v} \approx \|\boldsymbol{v}\|\boldsymbol{V}_{q+m}f(\boldsymbol{S}_{q+m})\boldsymbol{V}_{q+m}^H\,\boldsymbol{v} = \|\boldsymbol{v}\|\boldsymbol{V}_{q+m}f(\boldsymbol{S}_{q+m})\boldsymbol{e}_1$$

with a small matrix $\boldsymbol{S}_{q+m} = \boldsymbol{V}_{q+m}^H\,\boldsymbol{A}\boldsymbol{V}_{q+m} \in \mathbb{C}^{(q+m)\times(q+m)}$.

6.5 Numerical experiments

In the following numerical experiments, we approximate the solution of a Schrödinger equation on the unit square and of a wave equation on a non standard domain by the extended Krylov subspace method. Our numerical results are compared with the standard and the shift-and-invert Krylov subspace approximation.

6.5.1 Schrödinger equation

We consider the dimensionless time-dependent free Schrödinger equation on the unit square $\Omega = (0,1)^2$ given by

$$\begin{aligned} u' &= i\Delta u \qquad && \text{for} \quad (x,y) \in \Omega\,,\ t \geq 0\,,\\ u(0,x,y) &= u_0(x,y) && \text{for} \quad (x,y) \in \Omega \end{aligned}$$

with homogeneous Dirichlet boundary conditions for the Hilbert space $L^2(\Omega)$. If we split the solution u in its real and imaginary part, we obtain the system

$$\begin{aligned} (\operatorname{Re} u)' &= -\Delta(\operatorname{Im} u)\,,\\ (\operatorname{Im} u)' &= \Delta(\operatorname{Re} u)\,. \end{aligned}$$

Setting $v = \operatorname{Re} u$, $w = \operatorname{Im} u$ and defining the operator B as $-\Delta$ with homogeneous Dirichlet boundary conditions, we can write the Schrödinger equation as

$$y'(t) = Ay(t) \qquad \text{with} \qquad A = \begin{bmatrix} 0 & B \\ -B & 0 \end{bmatrix}, \qquad y(t) = \begin{bmatrix} v(t) \\ w(t) \end{bmatrix}$$

on $L^2(\Omega) \times L^2(\Omega)$. The domain of the operator A is given by $D(A) = D(B) \times D(B)$ with $D(B) = H_0^1(\Omega) \cap H^2(\Omega)$. As initial value, we choose

$$y(0) = y_0^q = \begin{bmatrix} v_0^q \\ w_0^q \end{bmatrix}, \quad v_0^q(x,y) = w_0^q(x,y) = \frac{x^{2q}(1-x)^{2q}y^{2q}(1-y)^{2q}}{\|x^{2q}(1-x)^{2q}y^{2q}(1-y)^{2q}\|_{L^2(\Omega)}}\,, \tag{6.6}$$

such that $y_0^q \in D(A^q)$ but $y_0^q \notin D(A^{q+1})$, since $A^q y_0^q$ does no longer fulfill the required zero boundary conditions. Therefore, we can refer to q as the maximal index of smoothness for the initial value y_0^q with respect to the differential operator A.

A finite-difference discretization on the standard grid (ih, jh) for $i,j = 0,\dots,n+1$, and mesh width $h = \frac{1}{n+1}$ gives with $N = n^2$ the $2N \times 2N$ - discretization matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{B} \\ -\boldsymbol{B} & \boldsymbol{O} \end{bmatrix}, \qquad \boldsymbol{B} = \frac{1}{h^2}(\boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T})\,, \qquad \boldsymbol{T} = \operatorname{tridiag}(-1,2,-1)\,, \tag{6.7}$$

where $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ represents the five-point stencil for the negative Laplacian in two dimensions. Moreover, we define the discretized initial values $\boldsymbol{v}_0^q$, $\boldsymbol{w}_0^q$ as the functions v_0^q, w_0^q evaluated at the inner grid points, i.e.,

$$\boldsymbol{y}_0^q = \begin{bmatrix} \boldsymbol{v}_0^q \\ \boldsymbol{w}_0^q \end{bmatrix} = \begin{bmatrix} v_0^q(ih,jh)_{i,j=1}^n \\ w_0^q(ih,jh)_{i,j=1}^n \end{bmatrix}.$$

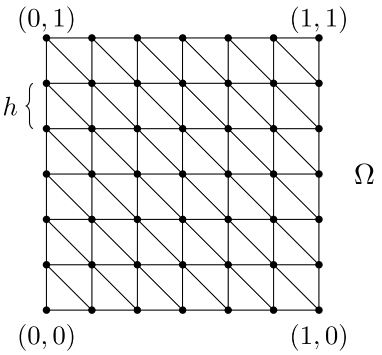


Figure 6.6: Regular triangulation of the unit square together with the used $(n + 2)^2$ grid nodes of the finite-difference and the finite-element discretization.

The appropriate inner product reads $(\boldsymbol{x}, \boldsymbol{z})_h = h^2 \boldsymbol{z}^H \boldsymbol{x}$ for $\boldsymbol{x}$, $\boldsymbol{z} \in \mathbb{C}^{2N}$ and the associated norm $\|\cdot\|_h$ is the standard Euclidean norm $\|\cdot\|_2$ scaled by h, that is, $\|\boldsymbol{x}\|_h = h\|\boldsymbol{x}\|_2$. Overall, this results in the semi-discrete evolution equation

$$\boldsymbol{y}'(t) = \boldsymbol{A}\boldsymbol{y}(t)\,, \qquad \boldsymbol{y}(t) = \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}, \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0^q$$

with exact solution $\boldsymbol{y}(\tau) = e^{\tau \boldsymbol{A}} \boldsymbol{y}_0^q$ at time τ.

For a regular triangulation of the unit square, the same matrix $\boldsymbol{A}$ arises by a linear finite-element discretization with mass lumping. As depicted in Figure 6.6, we have the same nodes as for the spatial grid of the finite-difference approximation before. Using the standard $N = n^2$ linear nodal basis functions ϕ_k and approximating the real part v and imaginary part w of the function u by

$$v(t, x, y) \approx \sum_{k=1}^{N} v_k(t)\phi_k(x, y)\,, \qquad w(t, x, y) \approx \sum_{k=1}^{N} w_k(t)\phi_k(x, y)\,,$$

we obtain the system of ordinary differential equations

$$\boldsymbol{M}\boldsymbol{y}'(t) = \boldsymbol{S}\boldsymbol{y}(t)\,, \qquad \boldsymbol{y}(t) = \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}, \qquad \boldsymbol{y}_0^q = \begin{bmatrix} \boldsymbol{v}_0^q \\ \boldsymbol{w}_0^q \end{bmatrix}$$

with the mass and stiffness matrices

$$\boldsymbol{M} = \begin{bmatrix} \widetilde{\boldsymbol{M}} & \boldsymbol{O} \\ \boldsymbol{O} & \widetilde{\boldsymbol{M}} \end{bmatrix} \in \mathbb{R}^{2N \times 2N}, \qquad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{O} & \widetilde{\boldsymbol{S}} \\ -\widetilde{\boldsymbol{S}} & \boldsymbol{O} \end{bmatrix} \in \mathbb{R}^{2N \times 2N}.$$

The idea of mass lumping is to replace the mass matrix with a diagonal approximation by applying a quadrature rule, rather than performing exact integrations, see [21], p. 83. This approach has the advantage that inverting the diagonal mass lumped matrix is very simple. More precisely, in order to compute the entries $(\widetilde{\boldsymbol{M}})_{ij}$ of the mass matrix, we take the quadrature formula

$$\int_K \phi_i \phi_j \, d(x, y) \approx \frac{|K|}{3} \sum_{k=1}^{3} \phi_i(\widetilde{x}_k, \widetilde{y}_k)\phi_j(\widetilde{x}_k, \widetilde{y}_k) =: (\widetilde{\boldsymbol{M}})_{ij}$$

for the integration over a triangle K with area $|K|$, where $(\widetilde{x}_k, \widetilde{y}_k)$, $k = 1, 2, 3$, denote the corners of K. Since $(\widetilde{\boldsymbol{M}})_{ij} = 0$ for $i \neq j$ and $(\widetilde{\boldsymbol{M}})_{ij} = h^2$ for $i = j$, the mass lumped matrix reads $\boldsymbol{M} = h^2\boldsymbol{I}$. The computation of the L^2-inner products $(\widetilde{\boldsymbol{S}})_{ij} = (\nabla\phi_i, \nabla\phi_j)_{L^2(\Omega)}$ reveals that $\widetilde{\boldsymbol{S}} = \boldsymbol{T} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{T}$ with the tridiagonal matrix $\boldsymbol{T}$ defined in (6.7). Hence, the stiffness matrix for our regular grid is given by $\boldsymbol{S} = h^2\boldsymbol{A}$, where $\boldsymbol{A}$ is the finite-difference discretization matrix in (6.7). Altogether, we end up with the same initial value problem

$$\boldsymbol{y}'(t) = \boldsymbol{M}^{-1}\boldsymbol{S}\boldsymbol{y}(t) = \frac{1}{h^2}\, h^2\boldsymbol{A} = \boldsymbol{A}\boldsymbol{y}(t)\,, \qquad \boldsymbol{y}(0) = \boldsymbol{y}_0^q$$

as before. Note that the inner product $(\boldsymbol{x}, \boldsymbol{z})_{\boldsymbol{M}} = \boldsymbol{z}^H\boldsymbol{M}\boldsymbol{x} = h^2\boldsymbol{z}^H\boldsymbol{x}$ for $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{C}^{2N}$, associated with the finite-element discretization, coincides with the inner product $(\cdot\,,\cdot)_h$ of the finite-difference method.

For the computation of the extended Krylov subspace approximation of $e^{\tau\boldsymbol{A}}\boldsymbol{y}_0^q$, we apply Algorithm 6.7 with the initial vector $\boldsymbol{y}_0^q$ and use the inner product $(\cdot\,,\cdot) = (\cdot\,,\cdot)_{\boldsymbol{M}}$. The projection $\boldsymbol{P}_{q+m}$ to $\mathcal{K}^{\gamma}_{q+1,m}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ is $\boldsymbol{P}_{q+m} = \boldsymbol{V}_{q+m}\boldsymbol{V}_{q+m}^H\,\boldsymbol{M}$, where $\boldsymbol{V}_{q+m} \in \mathbb{C}^{2N\times(q+m)}$ contains an $\boldsymbol{M}$-orthogonal basis of the Krylov subspace. Since $\boldsymbol{M}$ is in this special case given as $h^2\boldsymbol{I}$, it is just as well possible to take the standard Euclidean inner product $(\cdot\,,\cdot)_2$ and the projector $\widetilde{\boldsymbol{P}}_{q+m} = \widetilde{\boldsymbol{V}}_{q+m}\widetilde{\boldsymbol{V}}_{q+m}^H$, where the columns of $\widetilde{\boldsymbol{V}}_{q+m} = h\boldsymbol{V}_{q+m}$ are now orthonormal with respect to $(\cdot\,,\cdot)_2$. This leads to exactly the same approximation

$$e^{\tau\boldsymbol{A}}\boldsymbol{y}_0^q \approx \|\boldsymbol{y}_0^q\|_{\boldsymbol{M}}\boldsymbol{V}_{q+m}e^{\tau\boldsymbol{S}_{q+m}}\boldsymbol{e}_1 = \|\boldsymbol{y}_0^q\|_2\widetilde{\boldsymbol{V}}_{q+m}e^{\tau\widetilde{\boldsymbol{S}}_{q+m}}\boldsymbol{e}_1$$

with compressions $\boldsymbol{S}_{q+m} = \boldsymbol{V}_{q+m}^H\,\boldsymbol{M}\boldsymbol{A}\boldsymbol{V}_{q+m} = \boldsymbol{V}_{q+m}^H\,\boldsymbol{S}\boldsymbol{V}_{q+m}$ and $\widetilde{\boldsymbol{S}}_{q+m} = \widetilde{\boldsymbol{V}}_{q+m}^H\boldsymbol{A}\widetilde{\boldsymbol{V}}_{q+m}$.

In a first numerical experiment, we would like to illustrate how the smoothness of the initial value affects the convergence behavior of the Krylov subspace method. As mentioned above, we choose initial vectors $\boldsymbol{y}_0^q$ stemming from continuous values y_0^q defined in (6.6) with $y_0^q \in D(A^q)$ and $y_0^q \notin D(A^{q+1})$. In Figure 6.7, the approximation of $e^{\tau\boldsymbol{A}}\boldsymbol{y}_0^q$ in the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ (blue solid line) is compared with the approximation in the polynomial space $\mathcal{K}^{\gamma}_{q+m,1}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ (red dashed line) and in the rational space $\mathcal{K}^{\gamma}_{1,q+m}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ (green dash-dotted line) for step size $\tau = 0.005$, shift $\gamma = 1$, different numbers N of discretization points, and various smoothness indices q. From top left to bottom right, we have $N = 16\,129$ and $q = 6$, $N = 65\,025$ and $q = 5$, $N = 261\,121$ and $q = 4$, $N = 1\,046\,529$ and $q = 3$.

The standard Krylov subspace method behaves as expected. According to the abstract smoothness of the continuous initial value y_0^q, the polynomial Krylov process only works well in the first $q + 1$ steps and then stagnates. The vertical dashed line indicates the iteration step, at which the standard Krylov subspace involves vectors $\boldsymbol{A}^l\boldsymbol{y}_0^q$ with $l > q$. In contrast, the extended and the rational Krylov approximation both perform well.

It is also interesting to compare the performance of the three methods with respect to error versus computation time. This is shown in Figure 6.8 for $q = 5$, $N = 65\,025$ on the left and for $q = 4$, $N = 261\,121$ on the right. The convergence time comparison illustrates that the extended method clearly outperforms the standard and the rational Krylov subspace process. As above, the parameters are chosen as $\gamma = 1$ and $\tau = 0.005$. The computations have been conducted in the software environment Matlab, Release R2013b, under Ubuntu, Release 12.04, on a dual core processor with frequency 3 GHz on a desktop machine. In our example, one iteration of the standard Krylov subspace method only requires a multiplication of the large but sparse stiffness matrix $\boldsymbol{S}$ with some vector.

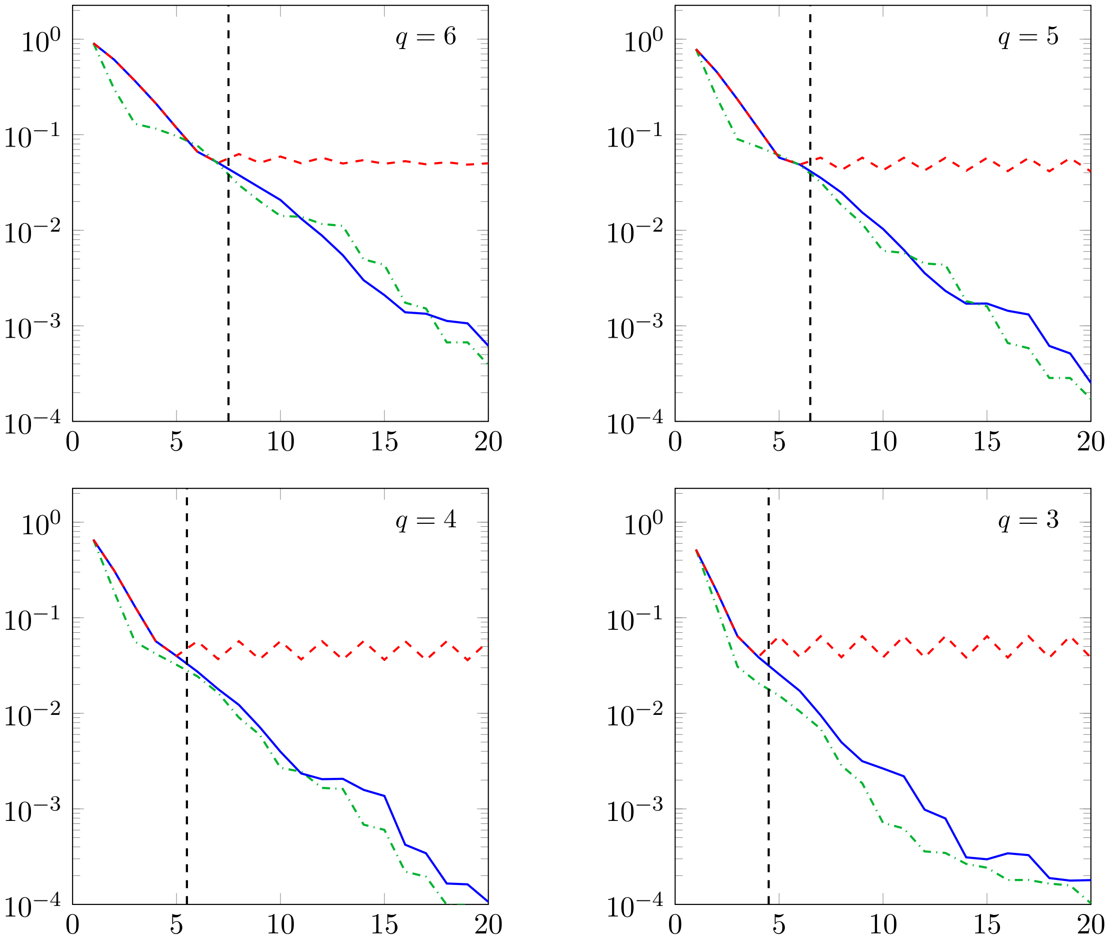


Figure 6.7: Plot of the error $\|e^{\tau \boldsymbol{A}}\boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{q+m}}\boldsymbol{y}_0^q\|_{\boldsymbol{M}}$ versus the dimension of the Krylov subspace for the standard Krylov subspace $\mathcal{K}_{q+m,1}^{\gamma}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ (red dashed line), the rational Krylov subspace $\mathcal{K}_{1,q+m}^{\gamma}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ (green dash-dotted line) and the extended Krylov subspace $\mathcal{K}_{q+1,m}^{\gamma}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ (blue solid line) for $N = 16\,129$, $q = 6$ (top left), $N = 65\,025$, $q = 5$ (top right), $N = 261\,121$, $q = 4$ (bottom left), $N = 1\,046\,529$, $q = 3$ (bottom right), and parameters $\gamma = 1$, $\tau = 0.005$ in each case.

Of course, this is much cheaper and faster than solving a linear system of the form

$$(\gamma \boldsymbol{I} - \tau \boldsymbol{A})^{-1}\boldsymbol{x} = (\gamma \boldsymbol{M} - \tau \boldsymbol{S})^{-1}\boldsymbol{M}\boldsymbol{x}$$

in each step of the rational process. This justifies the application of the standard Krylov subspace method in the first $q+1$ iteration steps, according to our index q of maximal smoothness, and the strategy to continue with the rational approximation afterwards. If, however, one would discretize the differential equation by finite elements without mass lumping, we would also have to solve a linear system of the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{M}^{-1}\boldsymbol{S}\boldsymbol{x}$ with a non-diagonal mass matrix $\boldsymbol{M}$ for the polynomial Krylov process. In this case, the standard Krylov subspace decomposition requires comparable numerical costs as the rational Krylov subspace method.

The special structure of the matrix $\boldsymbol{S}$ enables us to solve the linear systems with the help of the discrete sine transform. This is due to the fact that the tridiagonal matrix $\boldsymbol{T} = \operatorname{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ has eigenvalues λ_k and orthonormal eigenvectors $\boldsymbol{r}_k$ that are given as

$$\lambda_k = 4\sin^2\left(\frac{k\pi}{2(n+1)}\right),$$

$$\boldsymbol{r}_k = \frac{\sqrt{2}}{\sqrt{n+1}}\left[\ \sin\left(\frac{k\pi}{n+1}\right),\ \sin\left(\frac{2k\pi}{n+1}\right),\ \ldots,\ \sin\left(\frac{nk\pi}{n+1}\right)\ \right]^T$$

for $k = 1, \ldots, n$. These eigenvectors are related to the discrete sine transform which transforms a vector $(x_j)_{j=1}^n$ into a vector $(y_k)_{k=1}^n$ by

$$y_k = \sum_{j=1}^n x_j \sin\left(\frac{jk\pi}{n+1}\right), \qquad k = 1, \ldots, n.$$

If the eigenvectors $\boldsymbol{r}_k$ are collected in the orthogonal matrix $\boldsymbol{R}$, then $\boldsymbol{T}$ is diagonalizable via the transformation $\boldsymbol{T} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^T$ with $\boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. To compute $\boldsymbol{T}^{-1}\boldsymbol{b}$ for a vector $\boldsymbol{b} \in \mathbb{C}^n$, we therefore have to consider $\boldsymbol{T}^{-1}\boldsymbol{b} = \boldsymbol{R}\boldsymbol{D}^{-1}\boldsymbol{R}^T\boldsymbol{b}$. More exactly, we first apply a discrete sine transform to the vector $\boldsymbol{b}$, multiply with the diagonal matrix $\boldsymbol{D}^{-1}$ and finally transform the obtained result back via the inverse transformation. However, we are not interested in $\boldsymbol{T}^{-1}\boldsymbol{b}$, but in the solution of the linear system

$$(\gamma\boldsymbol{I} - \tau\boldsymbol{A})^{-1}\boldsymbol{x} = \begin{bmatrix} \gamma(\gamma^2\boldsymbol{I} + \tau^2\boldsymbol{B}^2)^{-1} & \tau\boldsymbol{B}(\gamma^2\boldsymbol{I} + \tau^2\boldsymbol{B}^2)^{-1} \\ -\tau\boldsymbol{B}(\gamma^2\boldsymbol{I} + \tau^2\boldsymbol{B}^2)^{-1} & \gamma(\gamma^2\boldsymbol{I} + \tau^2\boldsymbol{B}^2)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}.$$

That is, we have to compute $(\gamma^2\boldsymbol{I} + \tau^2\boldsymbol{B}^2)^{-1}\boldsymbol{x}_1 = \boldsymbol{z}_1$ and $(\gamma^2\boldsymbol{I} + \tau^2\boldsymbol{B}^2)^{-1}\boldsymbol{x}_2 = \boldsymbol{z}_2$. Since the eigenvalues of $\boldsymbol{B} = \frac{1}{h^2}(\boldsymbol{T}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{T})$ are $\lambda_{j,k} = \frac{1}{h^2}(\lambda_j+\lambda_k)$ with associated eigenvectors $\boldsymbol{r}_{j,k} = \boldsymbol{r}_j \otimes \boldsymbol{r}_k$ for $j, k = 1, \ldots, n$ (see Section 3.1.1), similar ideas can be used to compute $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ using the discrete sine transform.

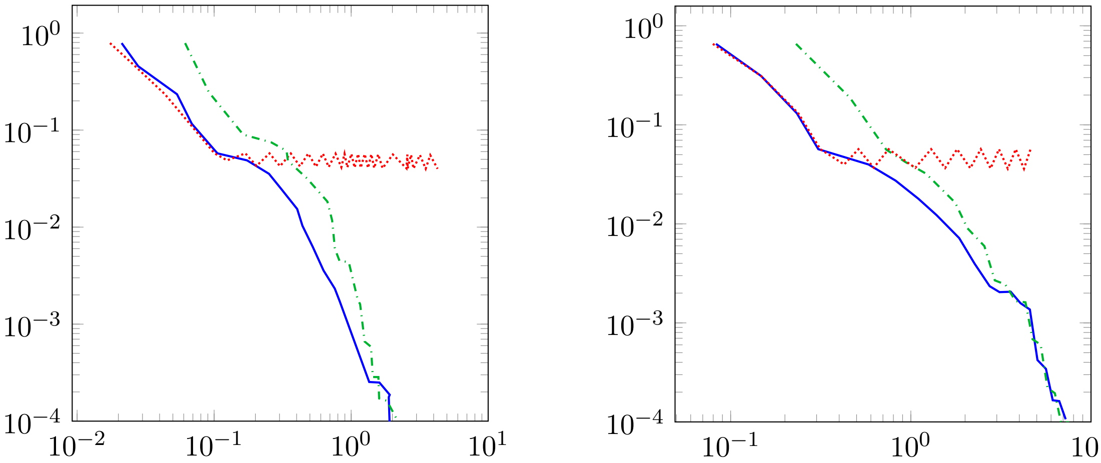


Figure 6.8: Plot of error versus computation time in seconds for the standard (red dotted line), the rational (green dash-dotted line) and the extended (blue solid line) Krylov subspace method for $q = 5$, $N = 65\,025$ on the left-hand side and $q = 4$, $N = 261\,121$ on the right-hand side as well as $\gamma = 1$, $\tau = 0.005$ in both cases.

An improvement of the convergence rate for the extended Krylov approximation based on the subspace $\mathcal{K}_{q+1,m}^{\gamma}(\tau\boldsymbol{A}, \boldsymbol{y}_0^q)$ is achieved by choosing the shift γ appropriately, that is,

$$\gamma = m^\alpha, \qquad \alpha = \frac{r - \frac{q+\ell}{2}}{r + \frac{q+\ell}{2}}, \qquad r > \frac{q+\ell}{2} + 1, \tag{6.8}$$

pursuant to Section 6.3. If we want to perform 20 extended Krylov steps for an initial value with smoothness index q, it holds $m = 20 - q$. Since the φ_0-function $\varphi_0(\tau \boldsymbol{A})\boldsymbol{y}_0^q = e^{\tau \boldsymbol{A}}\boldsymbol{y}_0^q$ is approximated, we have $\ell = 0$. In Figure 6.9, the error curves of the extended Krylov subspace approximation with $N = 1\,046\,529$, $\tau = 0.005$, and $q = 2, 3$ (left, right) are shown for $\gamma = 1$ (blue solid line) and $\gamma = 18^{3/5}$ (black dash-dotted line) on the left-hand side as well as $\gamma = 1$ (blue solid line) and $\gamma = 17^{5/11}$ (black dash-dotted line) on the right-hand side. These values of γ correspond in both cases to the choice $r = 4$ in (6.8). The method applied with $\gamma = 18^{3/5}$ and $\gamma = 17^{5/11}$, respectively, exhibits a faster convergence behavior than the method applied with the shift $\gamma = 1$. Of course, the adaption of the parameter γ only takes effect after the first $(q + 1)$st step, since the choice of γ has no influence on the polynomial approximation.

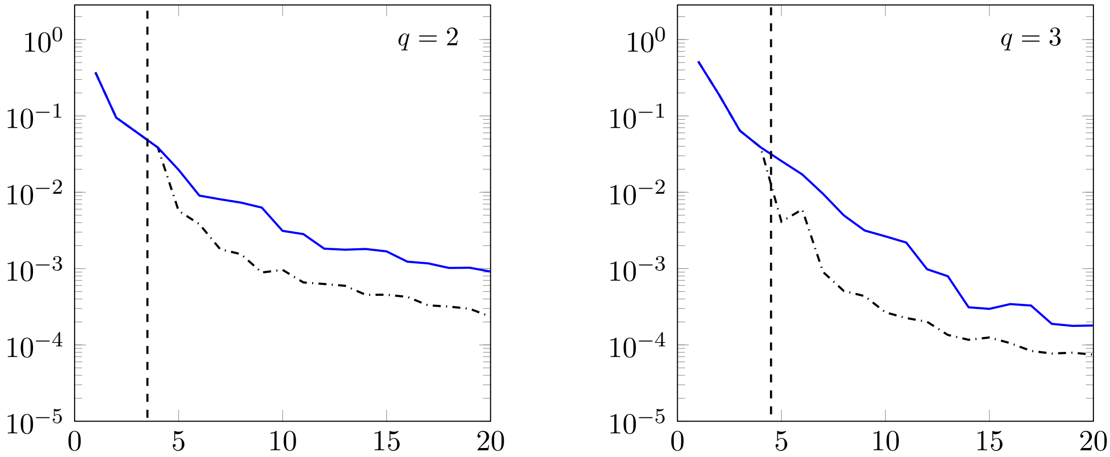


Figure 6.9: Comparison of the extended Krylov subspace approximation in $\mathcal{K}^{\gamma}_{q+1,m}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ with $N = 1\,046\,529$ for $q = 2$, $\gamma = 1$ (blue solid line), $\gamma = 18^{3/5}$ (black dash-dotted line) on the left and $q = 3$, $\gamma = 1$ (blue solid line), $\gamma = 17^{5/11}$ (black dash-dotted line) on the right. The error $\|e^{\tau \boldsymbol{A}}\boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{q+m}}\boldsymbol{y}_0^q\|_{\boldsymbol{M}}$ is plotted versus the dimension of the extended Krylov subspace for $\tau = 0.005$.

This method might not lead to success for time steps τ that are too large. This behavior is, roughly speaking, caused by the fact that the quality of the polynomial approximation is principally measured via Taylor expansion (e.g., [72]). If $\tau^k \|\boldsymbol{A}^k \boldsymbol{y}_0^q\|_{\boldsymbol{M}}$ is large for $k \le q$, it can happen that the approximation error does not decrease, although we have not yet reached the maximal index q of smoothness. This effect is depicted in Figure 6.10. Here, the approximation error in the extended (blue solid line), the polynomial (red dashed line) and in the shift-and-invert (green dash-dotted line) Krylov subspace are plotted versus the number of iteration steps. On the left-hand side, we take $\tau_1 = 0.005$, $N = 65\,025$, $\gamma = 1$, and the initial value $\boldsymbol{y}_0^4$ with continuous counterpart $y_0^4 \in D(A^4)$ but $y_0^4 \notin D(A^5)$. On the right-hand side, we use the same data but $\tau_2 = 0.05$. Comparing the quantities of $\tau_i^4 \|\boldsymbol{A}^4 \boldsymbol{y}_0^4\|_{\boldsymbol{M}}$ for $i = 1, 2$, we find

$$\tau_1^4 \|\boldsymbol{A}^4 \boldsymbol{y}_0^4\|_{\boldsymbol{M}} \approx 2.8\,, \qquad \tau_2^4 \|\boldsymbol{A}^4 \boldsymbol{y}_0^4\|_{\boldsymbol{M}} \approx 2.8 \cdot 10^4\,.$$

This explains why the polynomial method for the approximation of $e^{\tau_2 \boldsymbol{A}}\boldsymbol{y}_0^4$ even fails in the first iteration steps. For larger τ, the rational and extended Krylov subspace method still converge, but the improvement of the approximation quality is notably slower.

Finally, we are left with the question of how the smoothness of the initial value, that is

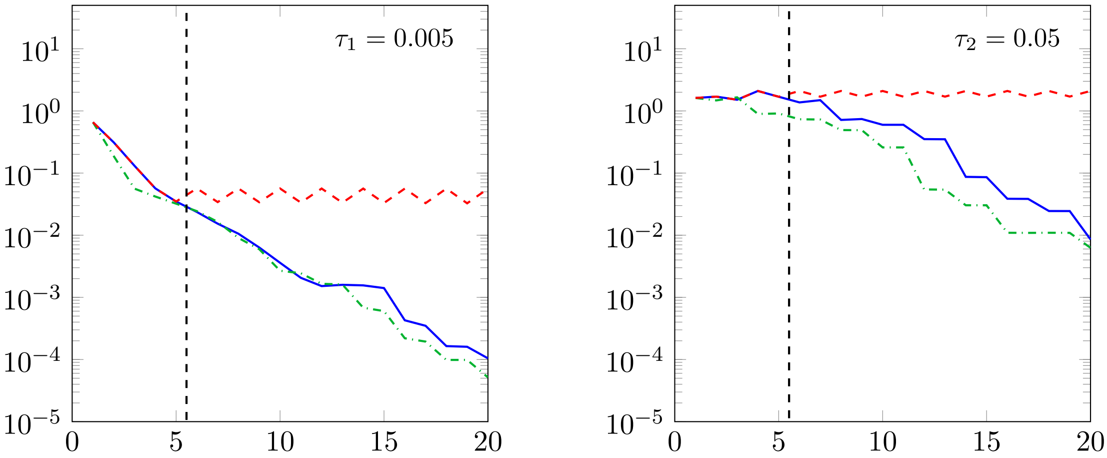


Figure 6.10: Plot of the error $\|e^{\tau_i \boldsymbol{A}} \boldsymbol{y}_0^4 - e^{\tau_i \boldsymbol{A}_{q+m}} \boldsymbol{y}_0^4\|_{\boldsymbol{M}}$ versus the dimension of the Krylov subspace for the standard Krylov subspace $\mathcal{K}_{q+m,1}^{\gamma}(\tau_i \boldsymbol{A}, \boldsymbol{y}_0^4)$ (red dashed line), the rational Krylov subspace $\mathcal{K}_{1,q+m}^{\gamma}(\tau_i \boldsymbol{A}, \boldsymbol{y}_0^4)$ (green dash-dotted line) and the extended Krylov subspace $\mathcal{K}_{q+1,m}^{\gamma}(\tau_i \boldsymbol{A}, \boldsymbol{y}_0^4)$ (blue solid line) with parameters $N = 1\,046\,529$, $\gamma = 1$, $q = 4$, and time step sizes $\tau_1 = 0.005$, $\tau_2 = 0.05$.

usually not known in advance, can be detected in general. The simplest heuristic approach to find the index q, at which it is reasonable to stop the polynomial approximation and to continue with the rational Krylov process, is to compute

$$E_m := \|e^{\tau \boldsymbol{A}_m} \boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{m-1}} \boldsymbol{y}_0^q\|_{\boldsymbol{M}} \,,$$

where $\boldsymbol{A}_m = \boldsymbol{P}_m \boldsymbol{A} \boldsymbol{P}_m$ is the restriction onto the polynomial subspace $\mathcal{K}_{m,1}^{\gamma}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$. If E_m is less than a given tolerance, we terminate the standard Krylov iteration and use the rational method.

In Figure 6.11, we plot the error of the standard Krylov method (red dashed line) together with E_m (cyan dash-dotted line) against m for $N = 65\,025$, $q = 5$ (on the left) and $N = 1\,046\,529$, $q = 3$ (on the right). The vertical dashed line indicates as before the point, where the polynomial approximation uses vectors $\boldsymbol{A}^l \boldsymbol{y}_0^q$ with $l > q$. In both pictures, E_m decreases in the first $q+1$ steps and then stagnates. The "zig-zag" behavior of the standard Krylov approximation error, which we also observed in some other numerical experiments (see, e.g., the one-dimensional wave equation in Section 5.6.2), has the consequence that the stagnation of E_m is located at around 10^{-1} and, therefore, might not allow for clear conclusions. It could lead to the wrong inference that the polynomial Krylov subspace approximation improves by $\mathcal{O}(10^{-1})$ with each iteration step. For this reason, a clearer indicator would be desirable. As an alternative heuristic detection, it is possible to take

$$\widetilde{E}_m := \min\{\|e^{\tau \boldsymbol{A}_m} \boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{m-1}} \boldsymbol{y}_0^q\|_{\boldsymbol{M}}, \|e^{\tau \boldsymbol{A}_m} \boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{m-2}} \boldsymbol{y}_0^q\|_{\boldsymbol{M}}\} \,,$$

which is depicted in Figure 6.11 by the orange solid line. The indication that the maximal index of smoothness has been reached is now more obvious and we can choose for example 10^{-2} or 10^{-3} as threshold value to stop the standard method and to start with the rational Krylov subspace approximation.

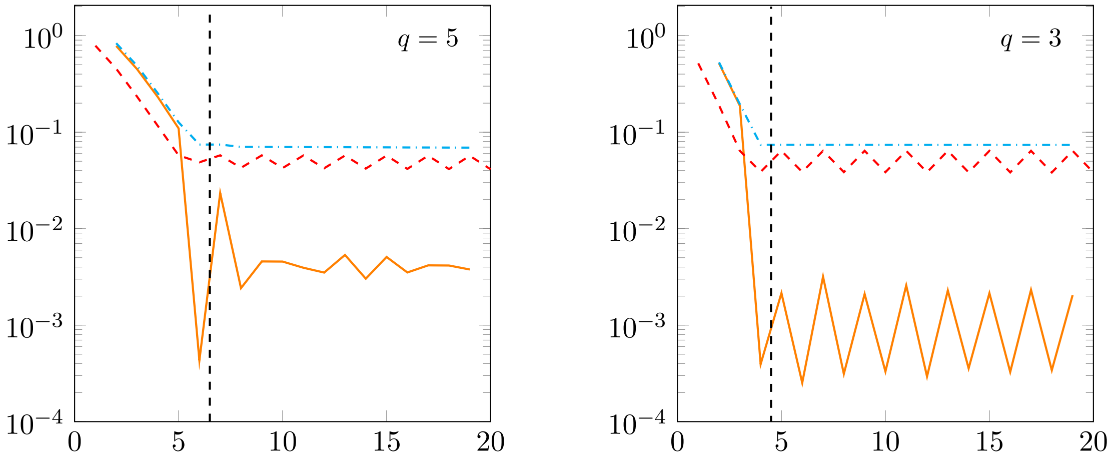


Figure 6.11: Plot of E_m (cyan dash-dotted line) and $\widetilde{E}_m$ (orange solid line) versus m together with the error $\|e^{\tau \boldsymbol{A}}\boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_m}\boldsymbol{y}_0^q\|_{\boldsymbol{M}}$ of the standard Krylov subspace approximation (red dashed line) for $N = 65\,025$, $q = 5$ (left), $N = 1\,046\,529$, $q = 3$ (right), and $\tau = 0.005$.

6.5.2 Wave equation on a non standard domain

For the non standard spatial domain Ω in Figure 6.13, consisting of two basins and a semicircular pipeline between them, we consider the wave equation with homogeneous Neumann boundary conditions

$$\begin{aligned}
&u'' = \Delta u - u && \text{for} \quad (x,y) \in \Omega\,,\ t \geq 0\,,\\
&u(0,x,y) = u_0(x,y)\,,\ u'(0,x,y) = u_0'(x,y) && \text{for} \quad (x,y) \in \Omega\,,\\
&\nabla_{\boldsymbol{n}} u = 0 && \text{for} \quad (x,y) \in \partial\Omega
\end{aligned}$$

on the Hilbert space $L^2(\Omega)$. In the following, we numerically approximate the solution u in the polynomial, the shift-and-invert, and the extended Krylov subspace. This time, we do not use the skew-symmetric first-order formulation as in Section 5.6.2, but instead the representation

$$y'(t) = \begin{bmatrix} u(t) \\ u'(t) \end{bmatrix}' = \begin{bmatrix} 0 & I \\ \Delta - I & 0 \end{bmatrix} \begin{bmatrix} u(t) \\ u'(t) \end{bmatrix} = Ay(t)\,, \qquad y(0) = y_0 = \begin{bmatrix} u_0 \\ u_0' \end{bmatrix},$$

where Δ is the Laplacian with homogeneous Neumann boundary conditions. This formulation requires the application of a suitable inner product. In the following, we assume that $z = [z_1, z_2]^T$ and $\widetilde{z} = [\widetilde{z}_1, \widetilde{z}_2]^T$ are two arbitrary vectors with $z_1, \widetilde{z}_1 \in H^1(\Omega)$, $\nabla_{\boldsymbol{n}} z_1 = \nabla_{\boldsymbol{n}} \widetilde{z}_1 = 0$ on $\partial\Omega$, and $z_2, \widetilde{z}_2 \in L^2(\Omega)$. For $B = \operatorname{diag}(-\Delta + I, I)$, we define the inner product

$$\begin{aligned}
(z, \widetilde{z})_B = (Bz, \widetilde{z})_{L^2(\Omega)\times L^2(\Omega)} &= \left(\begin{bmatrix} -\Delta + I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \begin{bmatrix} \widetilde{z}_1 \\ \widetilde{z}_2 \end{bmatrix} \right)_{L^2(\Omega)\times L^2(\Omega)}\\
&= \big((-\Delta + I) z_1, \widetilde{z}_1\big)_{L^2(\Omega)} + (z_2, \widetilde{z}_2)_{L^2(\Omega)}\\
&= \int_\Omega \nabla z_1 \nabla \widetilde{z}_1 \, d(x,y) + \int_\Omega z_1 \widetilde{z}_1 \, d(x,y) + \int_\Omega z_2 \widetilde{z}_2 \, d(x,y)\,.
\end{aligned}$$

The last equality follows from Green's formula and the assumed homogeneous Neumann boundary conditions for the functions z_1 and $\widetilde{z}_1$. Since $(\cdot\,,\cdot)_B$ is linear and

$$
\begin{aligned}
&(z,z)_B = \|z_1\|^2_{H^1(\Omega)} + \|z_2\|^2_{L^2(\Omega)} \geq 0\,,\\
&(z,z)_B = 0 \iff z_1 = z_2 = 0\,,\\
&(z,\widetilde{z})_B = (\widetilde{z},z)_B\,,
\end{aligned}
$$

the new inner product $(\cdot\,,\cdot)_B$ is well-defined. Since $-\Delta + I$ is positive and self-adjoint with respect to $(\cdot\,,\cdot)_{L^2(\Omega)}$, there exists a unique positive and self-adjoint square root $\sqrt{-\Delta+I}$ (e.g. Proposition 5.13 in Schmüdgen [76]), and we obtain

$$
\|Az\|^2_B = \|(\Delta - I)z_1\|^2_{L^2(\Omega)} + \|\sqrt{-\Delta+I}\, z_2\|^2_{L^2(\Omega)}\,.
$$

Hence, the domain of the operator A with respect to the new inner product is given as $D(A) = D(\Delta) \times D(\sqrt{-\Delta+I})$.

We want to solve the wave equation numerically using linear finite elements on the domain Ω shown in Figure 6.13 together with a coarse triangulation consisting of 229 nodes and 294 triangles. This triangulation is generated by the free Matlab Toolbox MESH2D[1]. We use smaller triangles at the corners of Ω and for the narrow pipe. Based on this coarse grid, we obtain finer and finer discretizations, if we subdivide every triangle into 4 smaller triangles. For this purpose, we take the vertices and edge midpoints of the larger triangles as new nodes of the sub-triangles.

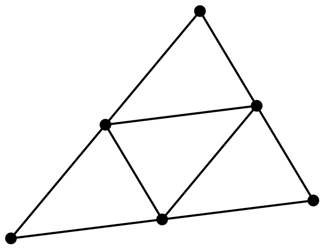


Figure 6.12: Refinement into 4 sub-triangles.

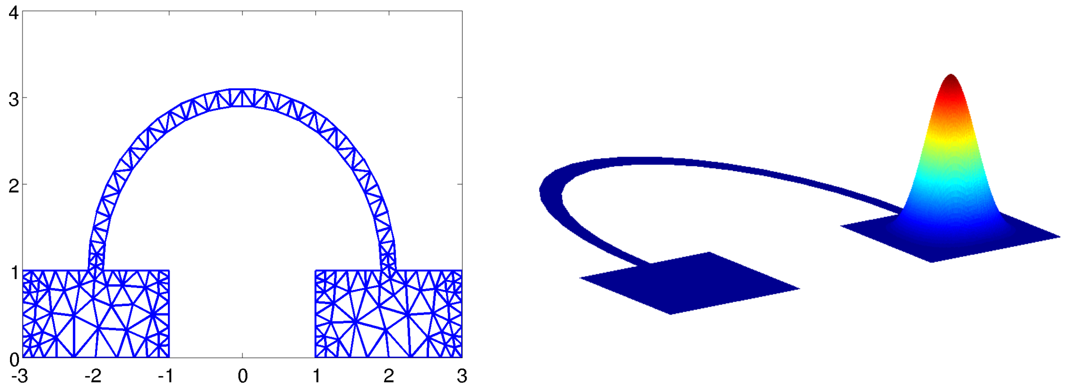


Figure 6.13: Left-hand side: Coarsest triangulation of the domain Ω with 229 nodes and 294 triangles. Right-hand side: Initial value $u_0^q = (u_0^q)'$ for smoothness index $q = 2$.

The functions u and u' are approximated by a linear combination of N linear nodal basis functions $\phi_k \in H^1(\Omega)$ as

$$
u(t,x,y) \approx \sum_{k=1}^{N} u_k(t)\phi_k(x,y)\,, \qquad u'(t,x,y) \approx \sum_{k=1}^{N} u'_k(t)\phi_k(x,y)\,.
$$

[1] http://www.mathworks.com/matlabcentral/fileexchange/25555-mesh2d-automatic-mesh-generation

In contrast to evolution equations with homogeneous Dirichlet boundary conditions, we additionally have to take basis functions $\phi_k(x,y) \in H^1(\Omega)$ for the nodes on the boundary $\partial\Omega$. For every test function $\phi_k \in H^1(\Omega)$, the integral over $\partial\Omega$ in Green's formula

$$\int_\Omega \phi_k \Delta u \, d(x,y) = -\int_\Omega \nabla\phi_k \nabla u \, d(x,y) + \int_{\partial\Omega} \phi_k \nabla_{\boldsymbol{n}} u \, ds$$

vanishes because of the prescribed homogeneous Neumann boundary conditions. This leads to the semi-discrete problem

$$\begin{bmatrix} \boldsymbol{M} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}' = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{M} \\ \boldsymbol{S}-\boldsymbol{M} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{v}(0) \\ \boldsymbol{w}(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_0 \\ \boldsymbol{w}_0 \end{bmatrix} \tag{6.9}$$

with initial values $\boldsymbol{v_0} = \big(u_k(0)\big)_{k=1}^N$ and $\boldsymbol{w}_0 = \big(u_k'(0)\big)_{k=1}^N$. The mass matrix $\boldsymbol{M}$ and the stiffness matrix $\boldsymbol{S}-\boldsymbol{M}$, which represents the discretization of $\Delta - I$, consist of the L^2-inner products

$$(\boldsymbol{M})_{ij} = (\phi_i, \phi_j)_{L^2(\Omega)}, \qquad (\boldsymbol{S})_{ij} = -(\nabla\phi_i, \nabla\phi_j)_{L^2(\Omega)}, \qquad i,j = 1,\dots,N.$$

Multiplying equation (6.9) from the left with the block diagonal matrix $\operatorname{diag}(\boldsymbol{M}^{-1}, \boldsymbol{M}^{-1})$, the semi-discrete formulation reads

$$\boldsymbol{y}'(t) = \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}' = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{I} \\ \boldsymbol{M}^{-1}(\boldsymbol{S}-\boldsymbol{M}) & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix} = \boldsymbol{A}\boldsymbol{y}(t), \quad \boldsymbol{y}(0) = \boldsymbol{y}_0 = \begin{bmatrix} \boldsymbol{v}_0 \\ \boldsymbol{w}_0 \end{bmatrix}.$$

Since $\boldsymbol{M}$ and $-\boldsymbol{S}$ are symmetric positive definite, the same holds true for the composition $-(\boldsymbol{S}-\boldsymbol{M})$ of both matrices. Therefore, the positive definite and symmetric matrix

$$\boldsymbol{B} = \begin{bmatrix} -(\boldsymbol{S}-\boldsymbol{M}) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{M} \end{bmatrix},$$

corresponding to the operator counterpart $B = \operatorname{diag}(-\Delta + I, I)$ from above, defines an inner product $(\cdot\,,\cdot)_{\boldsymbol{B}}$ on $\mathbb{C}^{2N}$ by

$$\left(\begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix}, \begin{bmatrix} \widetilde{\boldsymbol{z}}_1 \\ \widetilde{\boldsymbol{z}}_2 \end{bmatrix} \right)_{\boldsymbol{B}} = \begin{bmatrix} \widetilde{\boldsymbol{z}}_1^H & \widetilde{\boldsymbol{z}}_2^H \end{bmatrix} \begin{bmatrix} -(\boldsymbol{S}-\boldsymbol{M}) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix}$$
$$= -\widetilde{\boldsymbol{z}}_1^H(\boldsymbol{S}-\boldsymbol{M})\boldsymbol{z}_1 + \widetilde{\boldsymbol{z}}_2^H \boldsymbol{M}\boldsymbol{z}_2.$$

With respect to this inner product, we will show that $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ or, to be more precise, $W(\boldsymbol{A}) \subseteq i\,\mathbb{R}$. Thus, the matrix $\boldsymbol{A}$ fits in our framework above. This underlines the importance to consider arbitrary inner products.

Lemma 6.8 *For the inner product $(\cdot\,,\cdot)_{\boldsymbol{B}}$, we have $W(\boldsymbol{A}) \subseteq i\,\mathbb{R}$ or, respectively,*

$$\operatorname{Re}(\boldsymbol{A}\boldsymbol{z}, \boldsymbol{z})_{\boldsymbol{B}} = 0 \quad \textit{for all} \quad \boldsymbol{z} \in \mathbb{C}^{2N}.$$

Proof. For an arbitrary $\boldsymbol{z} = [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T]^T \in \mathbb{C}^{2N}$, a short calculation yields

$$\begin{aligned} (\boldsymbol{A}\boldsymbol{z}, \boldsymbol{z})_{\boldsymbol{B}} &= \left(\begin{bmatrix} \boldsymbol{O} & \boldsymbol{I} \\ \boldsymbol{M}^{-1}(\boldsymbol{S}-\boldsymbol{M}) & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix} \right)_{\boldsymbol{B}} \\ &= \left(\begin{bmatrix} \boldsymbol{z}_2 \\ \boldsymbol{M}^{-1}(\boldsymbol{S}-\boldsymbol{M})\boldsymbol{z}_1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix} \right)_{\boldsymbol{B}} \\ &= -\boldsymbol{z}_1^H(\boldsymbol{S}-\boldsymbol{M})\boldsymbol{z}_2 + \boldsymbol{z}_2^H(\boldsymbol{S}-\boldsymbol{M})\boldsymbol{z}_1 = 2i\operatorname{Im}(\boldsymbol{z}_2^H(\boldsymbol{S}-\boldsymbol{M})\boldsymbol{z}_1) \in i\,\mathbb{R}. \end{aligned}$$

This implies $\operatorname{Re}(\boldsymbol{A}\boldsymbol{z}, \boldsymbol{z})_{\boldsymbol{B}} = 0$ and the lemma is proved. ❑

We now approximate the exact solution $\boldsymbol{y}(\tau)$ of the semi-discrete problem at time τ in the polynomial, the shift-and-invert, and the extended Krylov subspace by

$$\boldsymbol{y}(\tau) = e^{\tau \boldsymbol{A}} \boldsymbol{y}_0 \approx e^{\tau \boldsymbol{A}_{q+m}} \boldsymbol{y}_0 = \boldsymbol{V}_{q+m} e^{\tau \boldsymbol{S}_{q+m}} \boldsymbol{V}_{q+m}^H \boldsymbol{B} \boldsymbol{y}_0 \,,$$

where $\boldsymbol{V}_{q+m} \in \mathbb{C}^{N\times(q+m)}$ is an orthonormal basis of the considered Krylov subspace of order $q+m$. With the orthogonal projection $\boldsymbol{P}_{q+m} = \boldsymbol{V}_{q+m}\boldsymbol{V}_{q+m}^H\boldsymbol{B}$, the restriction of $\boldsymbol{A}$ reads $\boldsymbol{A}_{q+m} = \boldsymbol{P}_{q+m}\boldsymbol{A}\boldsymbol{P}_{q+m}$ and the compression is given as $\boldsymbol{S}_{q+m} = \boldsymbol{V}_{q+m}^H\boldsymbol{B}\boldsymbol{A}\boldsymbol{V}_{q+m}$. The computation of $\boldsymbol{S}_{q+m}$ requires no calculation of $\boldsymbol{M}^{-1}(\boldsymbol{S}-\boldsymbol{M})$, since

$$\boldsymbol{B}\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O} & -(\boldsymbol{S}-\boldsymbol{M}) \\ \boldsymbol{S}-\boldsymbol{M} & \boldsymbol{O} \end{bmatrix} .$$

The initial values u_0^q and $(u_0^q)'$, whose index q refers to the abstract smoothness with respect to the differential operator A, describe a peak in the right basin, each given by the same function

$$g(x,y) = \begin{cases} (x-1)^{2q+1}(x-3)^{2q+1}y^{2q+1}(y-1)^{2q+1}\,, & (x,y) \in [1,3]\times[0,1]\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

We normalize the vector $[g,g]^T$ with respect to the norm $\|\cdot\|_B$ associated to the inner product $(\cdot\,,\cdot)_B$ with $B = \operatorname{diag}(-\Delta + I, I)$. This means that we scale the vector $[g,g]^T$ by the reciprocal of

$$\|[g,g]^T\|_B^2 = \int_\Omega \left(2g^2 + \nabla g \nabla g\right) d(x,y)\,.$$

By construction, $y_0^q = [g,g]^T/\|[g,g]^T\|_B$ belongs to $D(A^q)$ but not to $D(A^{q+1})$. The corresponding discrete initial value is designated by $\boldsymbol{y}_0^q = [(\boldsymbol{v}_0^q)^T, (\boldsymbol{w}_0^q)^T]^T = [(\boldsymbol{v}_0^q)^T, (\boldsymbol{v}_0^q)^T]^T$ for $\boldsymbol{v}_0^q, \boldsymbol{w}_0^q \in \mathbb{C}^N$.

On the basis of the considerations presented in the previous Section 6.1, we plot in Figure 6.14 the quantities $\|\boldsymbol{A}^2\boldsymbol{y}_0^2\|_{\boldsymbol{B}}$ and $\|\boldsymbol{A}^3\boldsymbol{y}_0^2\|_{\boldsymbol{B}}$ for different refinements of the triangulation. The value 1 on the horizontal axis corresponds to the coarsest grid shown in Figure 6.13 and the values 2, 3, 4 to the next three finer grids, that we obtain by subdividing each triangle into 4 smaller sub-triangles, where the new triangles are created by joining nodes introduced at the edge midpoints. As expected from the smoothness index $q = 2$, the norm $\|\boldsymbol{A}^2\boldsymbol{y}_0^2\|_{\boldsymbol{B}}$ stays at the same level, whereas $\|\boldsymbol{A}^3\boldsymbol{y}_0^2\|_{\boldsymbol{B}}$ increases, if we refine the spatial mesh.

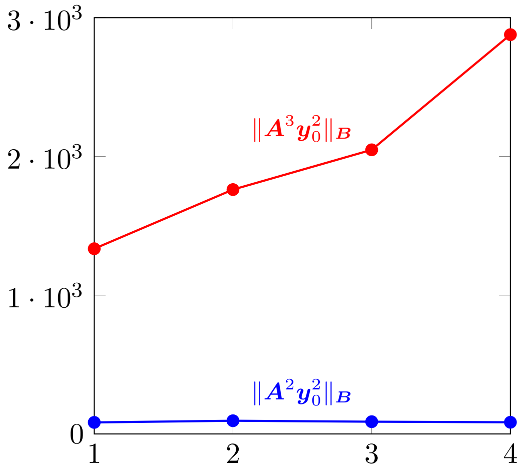


Figure 6.14: Values $\|\boldsymbol{A}^j\boldsymbol{y}_0^2\|$, $j = 2, 3$, for different refinements of the mesh.

In Figure 6.15, we draw the error curves for a coarse grid with 18 816 triangles and 10 057 nodes for $\tau = 0.1$, $q = 1, 2$ (left, right) and, in Figure 6.16, for a fine grid with 301 056 triangles, 153 121 nodes (on the left) and 1 204 224 triangles, 607 297 nodes (on the right) for the parameters $\tau = 0.05$ and $q = 1$. In the computations, the choice $\gamma = 15$ for the shift has turned out to be advantageous with regard to the multigrid solver used for the approximate solution of the linear systems in the rational Krylov decomposition.

The error curve for the approximation of $e^{\tau \boldsymbol{A}}\boldsymbol{y}_0^q$ in the polynomial subspace $\mathcal{K}^{\gamma}_{q+m,1}(\tau\boldsymbol{A},\boldsymbol{y}_0^q)$ is marked by a red dashed line, in the rational subspace $\mathcal{K}^{\gamma}_{1,q+m}(\tau\boldsymbol{A},\boldsymbol{y}_0^q)$ by a green dashed-dotted line and in the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\tau\boldsymbol{A},\boldsymbol{y}_0^q)$ by a blue solid line. As expected, the polynomial approximation virtually stagnates after $q+1$ Krylov steps, whereas the errors of the extended and the rational Krylov subspace process exhibit a sublinear convergence behavior.

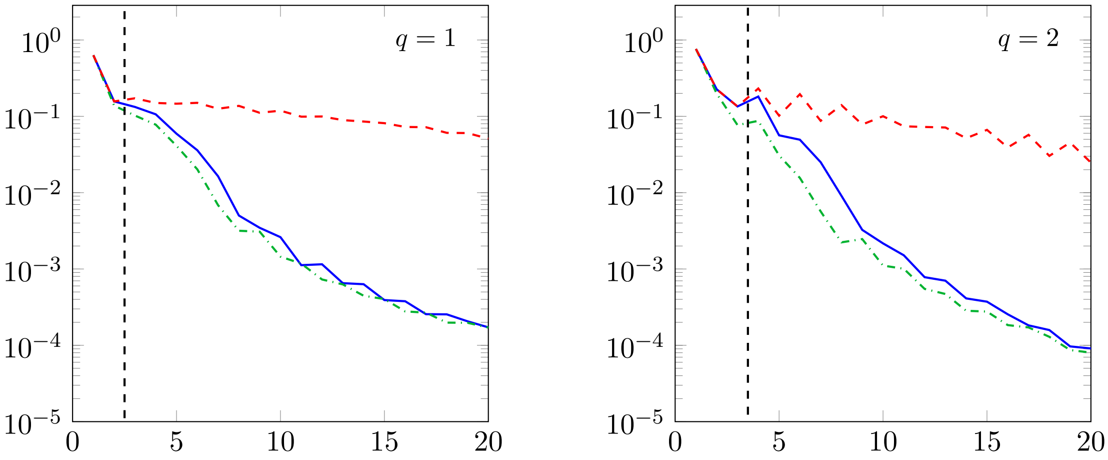


Figure 6.15: Plot of the error $\|e^{\tau \boldsymbol{A}}\boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{q+m}}\boldsymbol{y}_0^q\|_{\boldsymbol{B}}$ versus the dimension of the Krylov subspace for the standard Krylov subspace $\mathcal{K}^{\gamma}_{q+m,1}(\tau\boldsymbol{A},\boldsymbol{y}_0^q)$ (red dashed line), the rational Krylov subspace $\mathcal{K}^{\gamma}_{1,q+m}(\tau\boldsymbol{A},\boldsymbol{y}_0^q)$ (green dash-dotted line), and the extended Krylov subspace $\mathcal{K}^{\gamma}_{q+1,m}(\tau\boldsymbol{A},\boldsymbol{y}_0^q)$ (blue solid line) for a coarse grid with 18 816 triangles, 10 057 nodes, $\gamma = 15$, $\tau = 0.1$, and smoothness indices $q = 1, 2$.

In the rational Krylov subspace decomposition, we have to solve linear systems of the form $(\gamma\boldsymbol{I} - \tau\boldsymbol{A})\boldsymbol{x} = \boldsymbol{v}_m$, where $\boldsymbol{v}_m = [(\boldsymbol{v}_m^1)^T, (\boldsymbol{v}_m^2)^T]^T$ is the basis vector that is computed in the mth iteration step. Noting that

$$
\begin{aligned}
(\gamma\boldsymbol{I} - \tau\boldsymbol{A})^{-1}\boldsymbol{v}_m &= \begin{bmatrix} \gamma\boldsymbol{M} & -\tau\boldsymbol{M} \\ -\tau(\boldsymbol{S}-\boldsymbol{M}) & \gamma\boldsymbol{M} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{M} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{M} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_m^1 \\ \boldsymbol{v}_m^2 \end{bmatrix} \\
&= \begin{bmatrix} \big(\gamma^2\boldsymbol{M} - \tau^2(\boldsymbol{S}-\boldsymbol{M})\big)^{-1}(\gamma\boldsymbol{M}\boldsymbol{v}_m^1 + \tau\boldsymbol{M}\boldsymbol{v}_m^2) \\ \big(\gamma^2\boldsymbol{M} - \tau^2(\boldsymbol{S}-\boldsymbol{M})\big)^{-1}(\tau(\boldsymbol{S}-\boldsymbol{M})\boldsymbol{v}_m^1 + \gamma\boldsymbol{M}\boldsymbol{v}_m^2) \end{bmatrix},
\end{aligned}
\tag{6.10}
$$

we have to solve two linear systems with the matrix $\big(\gamma^2\boldsymbol{M} - \tau^2(\boldsymbol{S}-\boldsymbol{M})\big)^{-1}$ and different right-hand sides $\gamma\boldsymbol{M}\boldsymbol{v}_m^1 + \tau\boldsymbol{M}\boldsymbol{v}_m^2$ and $\tau(\boldsymbol{S}-\boldsymbol{M})\boldsymbol{v}_m^1 + \gamma\boldsymbol{M}\boldsymbol{v}_m^2$.

Since the matrices $\boldsymbol{M}$ and $-(\boldsymbol{S}-\boldsymbol{M})$ are both symmetric and positive definite, we can use a standard geometric multigrid method which is based on the following idea (see, for instance, Chapter V in [9]): Solving a large linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ stemming from a spatial discretization, we first apply a smoother (e.g., Gauss-Seidel or damped Jacobi iteration) to some initial guess $\boldsymbol{x}_0$ of the solution $\boldsymbol{x}$. This procedure removes high oscillations of the error vector $\boldsymbol{x} - \boldsymbol{x}_0$ on the actual fine grid, whereas low frequencies are only reduced very slowly. On that account, we change to a coarser grid, on which the low frequency

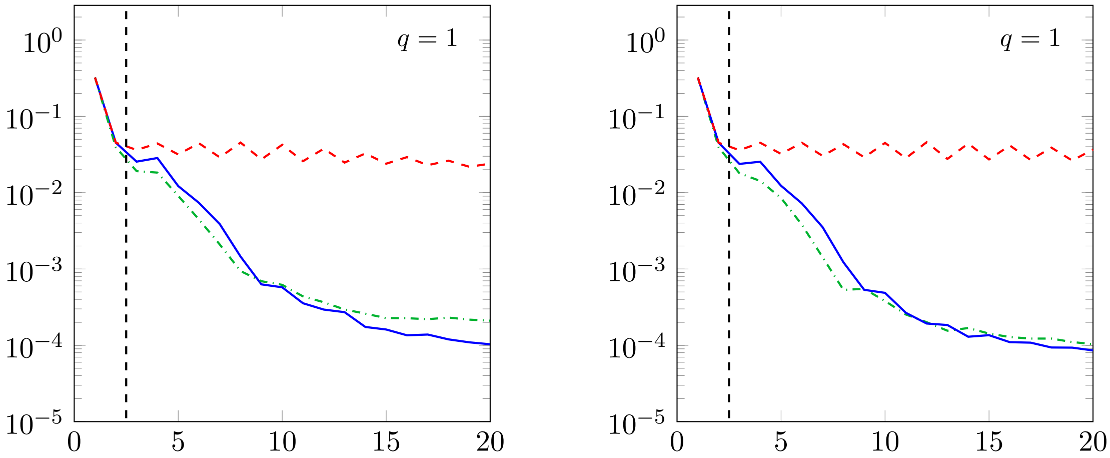


Figure 6.16: Plot of the error $\|e^{\tau \boldsymbol{A}} \boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{q+m}} \boldsymbol{y}_0^q\|_{\boldsymbol{B}}$ versus the dimension of the Krylov space for the standard Krylov subspace $\mathcal{K}_{q+m,1}^{\gamma}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (red dashed line), the rational Krylov subspace $\mathcal{K}_{1,q+m}^{\gamma}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (green dash-dotted line), and the extended Krylov subspace $\mathcal{K}_{q+1,m}^{\gamma}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (blue solid line) for fine grids with 153 121 nodes, 301 056 triangles (left) and with 607 297 nodes, 1 204 224 triangles (right) for $q = 1$, $\gamma = 15$, and $\tau = 0.1$.

parts appear as high frequencies and can be smoothed again. In this manner, we iterate from the finest grid to coarser and coarser grids, using suitable restrictions, solve the small linear system on the coarsest grid exactly, and finally prolongate the result back to the finest grid.

In contrast to the previous numerical example of the Schrödinger equation, where we used mass lumping to diagonalize the mass matrix $\boldsymbol{M}$, the standard Krylov subspace method requires here the solution of a linear system with a non-diagonal mass matrix $\boldsymbol{M}$ in each iteration step, since

$$\boldsymbol{A}\boldsymbol{v}_m = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{I} \\ \boldsymbol{M}^{-1}(\boldsymbol{S} - \boldsymbol{M}) & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_m^1 \\ \boldsymbol{v}_m^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_m^2 \\ \boldsymbol{M}^{-1}(\boldsymbol{S} - \boldsymbol{M})\boldsymbol{v}_m^1 \end{bmatrix}.$$

It became apparent in our numerical experiments that the multigrid method applied to the matrix $\boldsymbol{M}$ and the vector $(\boldsymbol{S} - \boldsymbol{M})\boldsymbol{v}_m^1$, required for the polynomial approximation, converges exceptionally fast. More precisely, the computed starting vector for the multigrid method via a nested iteration is already very accurate.

However, the situation is different for the linear systems of the form (6.10) in the rational Krylov decomposition. In this case, we have noticed that for the parameter choice $\gamma = 1$, the convergence of the geometric multigrid solver is quite slow and requires many iterations to obtain a residual error of order $\mathcal{O}(10^{-5})$. This drawback can be resolved by setting $\gamma = 15$. Then a few iteration steps suffice, most commonly, to reduce the error of the multigrid solution to $\mathcal{O}(10^{-8})$, that is small enough in our case, if we want to approximate $e^{\tau \boldsymbol{A}} \boldsymbol{y}_0^q$ with an accuracy of $\mathcal{O}(10^{-4})$.

Nevertheless, one step of the polynomial Krylov subspace method is performed more quickly than one iteration step of the rational approximation. So, it pays off to use the cheaper standard Krylov subspace process as long as it improves the approximation quality and then to proceed with the rational decomposition.

Computing for $q = 2$ and $\tau = 0.05$ recursively $\boldsymbol{y}_{k+1}^2 \approx e^{\tau \boldsymbol{A}} \boldsymbol{y}_k^2$, $k = 0, 1, \ldots$, with the extended Krylov method, we obtain an approximation to the exact solutions $u(t)$ and $u'(t)$ at time $\tau = 5$ and $\tau = 10$. In the first case, we iterate until $m = 100$ and, in the second case, we iterate until $m = 200$. The approximate solutions for $u(5)$ and $u(10)$ are depicted in Figure 6.18 and Figure 6.19.

Furthermore, we want to illustrate what happens, if we use a "wrong" inner product for which the field of values of $\boldsymbol{A}$ is not located in the left complex half-plane. For this purpose, we take the standard inner product for the finite-element discretization with respect to the block diagonal matrix $\widetilde{\boldsymbol{B}} = \operatorname{diag}(\boldsymbol{M}, \boldsymbol{M})$ instead of $\boldsymbol{B} = \operatorname{diag}(-(\boldsymbol{S} - \boldsymbol{M}), \boldsymbol{M})$ above. Then for arbitrary vectors $\boldsymbol{0} \neq \boldsymbol{z} = [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T]^T \in \mathbb{C}^{2N}$, we have

$$\frac{(\boldsymbol{A}\boldsymbol{z}, \boldsymbol{z})_{\widetilde{\boldsymbol{B}}}}{(\boldsymbol{z}, \boldsymbol{z})_{\widetilde{\boldsymbol{B}}}} = \frac{1}{(\boldsymbol{z}, \boldsymbol{z})_{\widetilde{\boldsymbol{B}}}} \cdot \left(2i \operatorname{Im}(\boldsymbol{z}_1^H \boldsymbol{M} \boldsymbol{z}_2) + \boldsymbol{z}_2^H \boldsymbol{S} \boldsymbol{z}_1\right) \subseteq W_{\widetilde{\boldsymbol{B}}}(\boldsymbol{A}) \,.$$

Choosing $\boldsymbol{z}$ in such a way that $\boldsymbol{z}_1 = -\boldsymbol{z}_2$, we can conclude that $W_{\widetilde{\boldsymbol{B}}}(\boldsymbol{A})$ contains quantities in the right complex half-plane, since $-\boldsymbol{S}$ is positive definite. In Figure 6.17, we plot the approximation error against the dimension of the Krylov subspace for the same parameters as in Figure 6.16, but using the "wrong" inner product with respect to $\widetilde{\boldsymbol{B}}$. In this case, no convergence can be observed anymore, regardless of whether we consider the extended, the shift-and-invert, or the polynomial Krylov subspace approximation.

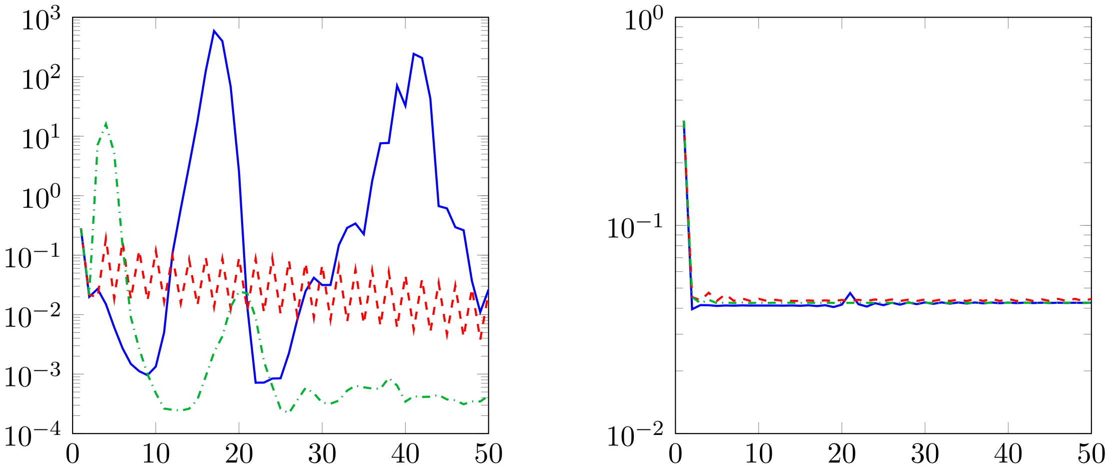


Figure 6.17: Plot of the error $\|e^{\tau \boldsymbol{A}} \boldsymbol{y}_0^q - e^{\tau \boldsymbol{A}_{q+m}} \boldsymbol{y}_0^q\|_{\widetilde{\boldsymbol{B}}}$ versus the dimension of the Krylov subspace for the standard Krylov subspace $\mathcal{K}_{q+m,1}^{\gamma}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (red dashed line), the rational Krylov subspace $\mathcal{K}_{1,q+m}^{\gamma}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (green dash-dotted line), and the extended Krylov subspace $\mathcal{K}_{q+1,m}^{\gamma}(\tau \boldsymbol{A}, \boldsymbol{y}_0^q)$ (blue solid line) for the grid with 153 121 nodes, 301 056 triangles (left) and with 607 297 nodes, 1 204 224 triangles (right) for $q = 1$, $\gamma = 15$, $\tau = 0.1$, and the "wrong" inner product with respect to $\widetilde{\boldsymbol{B}} = \operatorname{diag}(\boldsymbol{M}, \boldsymbol{M})$.

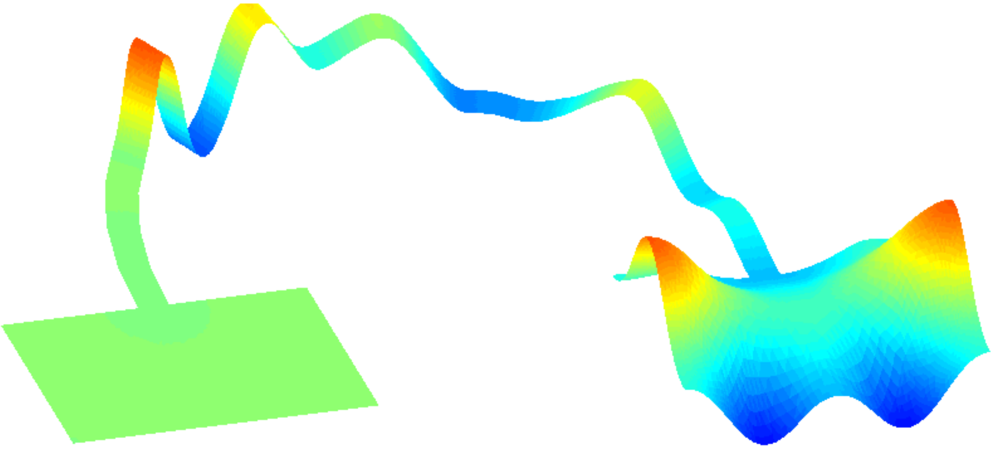


Figure 6.18: Approximation to the exact solution u at time $\tau = 5$.

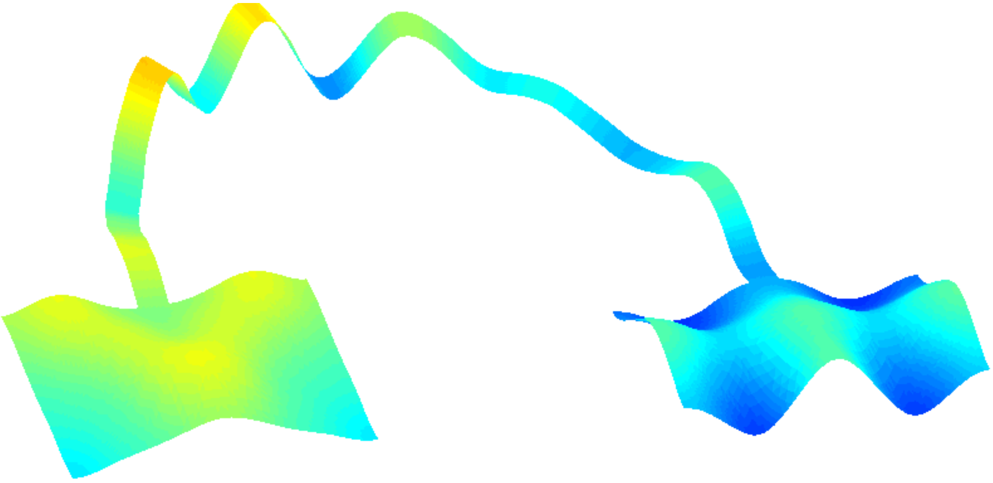


Figure 6.19: Approximation to the exact solution u at time $\tau = 10$.

Chapter 7

Rational Krylov subspace approximation with simple poles

So far, we have concentrated on the shift-and-invert Krylov subspace method, a special case of the rational Krylov approximation, that uses only one single repeated pole at $\gamma > 0$. This chapter will focus on a rational Krylov method for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$, $\ell \geq 1$, taking different simple poles in the right complex half-plane. As before, we consider matrices $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ of arbitrary dimension N that satisfy the condition $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ with respect to some inner product $(\cdot\,,\cdot)$ on the vector space $\mathbb{C}^N$.

In this chapter, we aim to speed-up the convergence of the Krylov approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$. Therefore, we choose $2m+1$ equidistant poles $z_k \in \mathbb{C}$ on the line $\mathrm{Re}(z) = \gamma > 0$ in the right complex half-plane. For this choice of poles, we will prove a faster convergence rate of order $\mathcal{O}(m^{-\ell})$. For comparison, we have seen that the error of the shift-and-invert Krylov subspace approximation behaves like $\mathcal{O}(m^{-\ell/2})$. Moreover, the resulting rational Krylov process is easily parallelizable, since for different poles z_k the $2m+1$ linear systems $(z_k\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}_k = \boldsymbol{v}$ can be solved independently in parallel, which leads to an additional speed-up.

The following results are based on our preprint [24]. Before we turn to the analysis of the rational Krylov subspace method with simple poles, we first need some preliminaries that we outline in the first section.

7.1 Preliminary notes

To guarantee the existence of $(z_k\boldsymbol{I} - \boldsymbol{A})^{-1}$ under our assumption $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$, the poles z_k of the rational Krylov subspace have to be located in the right complex half-plane. We choose $2m+1$ equidistant points $z_k = \gamma + ihk$, $k = -m, \dots, m$, with distance h on the line $\mathrm{Re}(z) = \gamma > 0$ (cf. Figure 7.1) and consider the rational Krylov subspace

$$\mathcal{Q}_{2m+2}(\boldsymbol{A}, \boldsymbol{v}) = \left\{ r(\boldsymbol{A})\boldsymbol{v} \;:\; r \in \frac{\mathcal{P}_{2m+1}}{q_{2m+1}} \right\}$$

with the prescribed denominator polynomial

$$q_{2m+1}(z) = \prod_{k=-m}^{m} (z_k - z)\,, \qquad z_k = \gamma + ihk\,, \qquad k = -m, \dots, m\,. \tag{7.1}$$

By Lemma 4.9, this subspace can also be written as

$$\mathcal{Q}_{2m+2}(\boldsymbol{A}, \boldsymbol{v}) = \mathrm{span}\left\{ \boldsymbol{v}, \frac{1}{z_{-m} - \boldsymbol{A}}\boldsymbol{v}, \frac{1}{z_{-m+1} - \boldsymbol{A}}\boldsymbol{v}, \dots, \frac{1}{z_m - \boldsymbol{A}}\boldsymbol{v} \right\}.$$

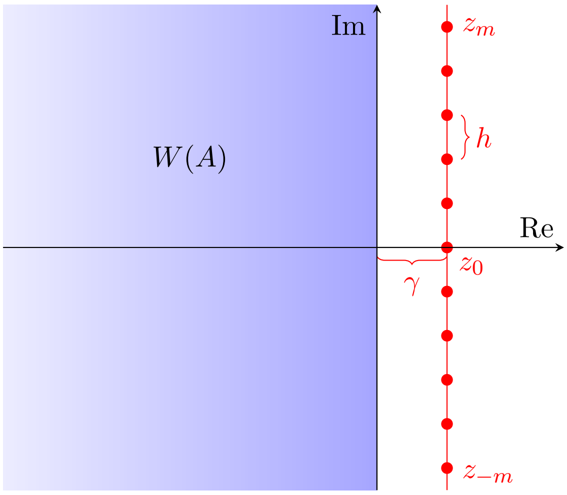


Figure 7.1: Locus of $W(\boldsymbol{A})$ and the roots z_k of the denominator polynomial q_{2m+1}.

Our first goal is to analyze the approximation of $\varphi_\ell(\boldsymbol{A})$ for $\ell \geq 1$ in the rational matrix subspace

$$\mathcal{R}_{2m+1}(\boldsymbol{A}) = \operatorname{span}\left\{ r(\boldsymbol{A}) \, : \, r \in \frac{\mathcal{P}_{2m}}{q_{2m+1}} \right\}.$$

We will establish uniform error estimates that hold true for matrices $\boldsymbol{A}$ with an arbitrarily large field of values in $\mathbb{C}_0^-$. Afterwards, these results will be used to bound the error $\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\|$ of the Krylov approximation. As before, $\boldsymbol{A}_m = \boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m$ is the restriction of $\boldsymbol{A}$ to $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$, $\boldsymbol{P}_m$ denotes the orthogonal projector onto the rational Krylov subspace, and $\boldsymbol{V}_m \in \mathbb{C}^{N\times(2m+2)}$ contains an orthonormal basis $[\boldsymbol{v}_1 \ \boldsymbol{v}_2 \ \cdots \ \boldsymbol{v}_{2m+2}]$ of $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$.

This basis $\boldsymbol{V}_m$ can be determined in parallel by assigning to every computing node the calculation of $\boldsymbol{x}_k = (z_k\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}$, that is required for the rational Krylov decomposition. This parallelization is possible by our choice of the poles leading to $2m+1$ decoupled linear systems for the $2m+1$ different poles z_k, $k = -m, \dots, m$. Subsequently, the computed solutions $\boldsymbol{x}_k$ are orthogonalized against each other to obtain an orthonormal basis of the Krylov subspace. As opposed to this, it is only possible to compute the basis of the shift-and-invert Krylov subspace with one fixed pole $\gamma > 0$ by a sequential process.

Whenever we write $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$ or $\mathcal{R}_{2m+1}(\boldsymbol{A})$, in what follows, we always mean the rational subspace with the fixed denominator polynomial q_{2m+1} defined in (7.1).

In the following, we state a few definitions and lemmas, in order to lay a groundwork for later examinations. The first lemma shows that the matrix exponential for arbitrary matrices with a field of values in the left complex half-plane is bounded by one.

Lemma 7.1 *For an arbitrary matrix $\boldsymbol{A} \in \mathbb{C}^{N\times N}$ satisfying $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ and all $\tau \geq 0$, it holds*

$$\|e^{\tau\boldsymbol{A}}\| \leq 1\,.$$

Proof. (cf. LeVeque [51], p. 300) We consider the initial value problem $\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}(t)$ with an arbitrary initial value $\boldsymbol{u}(0) = \boldsymbol{u}_0$. By our assumption on $W(\boldsymbol{A})$, we have

$$\frac{d}{dt}\,\|\boldsymbol{u}(t)\|^2 = \big(\boldsymbol{u}'(t), \boldsymbol{u}(t)\big) + \big(\boldsymbol{u}(t), \boldsymbol{u}'(t)\big) = \big(\boldsymbol{A}\boldsymbol{u}(t), \boldsymbol{u}(t)\big) + \big(\boldsymbol{u}(t), \boldsymbol{A}\boldsymbol{u}(t)\big)$$
$$= 2\,\mathrm{Re}\big(\boldsymbol{A}\boldsymbol{u}(t), \boldsymbol{u}(t)\big) \leq 0$$

and thus $\frac{d}{dt}\,\|\boldsymbol{u}(t)\| \leq 0$. It follows

$$\|e^{\tau\boldsymbol{A}}\boldsymbol{u}_0\| = \|\boldsymbol{u}(\tau)\| \leq \|\boldsymbol{u}(0)\| = \|\boldsymbol{u}_0\| \quad \text{for any} \quad \boldsymbol{u}_0 \in \mathbb{C}^N\,.$$

As a consequence, the induced matrix norm of $e^{\tau\boldsymbol{A}}$ is bounded by one. ❑

By equation (3.12), $\varphi_\ell(\boldsymbol{A})$ can be written as

$$\varphi_\ell(\boldsymbol{A}) = \int_0^1 e^{(1-\theta)\boldsymbol{A}}\,\frac{\theta^{\ell-1}}{(\ell-1)!}\,d\theta = \int_0^1 e^{s\boldsymbol{A}}\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}\,ds\,, \qquad \ell \geq 1\,.$$

For our particular purposes, the representation

$$\varphi_\ell(\boldsymbol{A}) = \int_0^\infty e^{s\boldsymbol{A}}\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}\cdot \mathbb{1}_{[0,1]}(s)\,ds\,, \qquad \ell \geq 1\,, \tag{7.2}$$

is useful, which looks, at first sight, more complicated than the previous formula. But this representation of the matrix φ-functions provides the advantage that the rational matrix functions $r(\boldsymbol{A}) \in \mathcal{R}_{2m+1}(\boldsymbol{A})$ possess a similar integral representation.

Lemma 7.2 *Every rational matrix function $r(\boldsymbol{A}) \in \mathcal{R}_{2m+1}(\boldsymbol{A})$ can be expressed as*

$$r(\boldsymbol{A}) = \frac{p_{2m}(\boldsymbol{A})}{q_{2m+1}(\boldsymbol{A})} = \sum_{k=-m}^{m} a_k\,\frac{1}{\gamma + ihk - \boldsymbol{A}} = \int_0^\infty e^{s\boldsymbol{A}} \sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\,ds\,,$$

where $q_{2m+1}(z) = \prod_{k=-m}^{m}(z_k - z)$ with $z_k = \gamma + ihk$, $\gamma > 0$.

Proof. The second equality follows directly from Lemma 4.9. In order to show the third equality, we first note that, by Lemma 7.1, we have

$$\|e^{s\boldsymbol{A}}e^{-(\gamma+ihk)s}\| \leq |e^{-\gamma s}|\|e^{s\boldsymbol{A}}\| \leq e^{-\gamma s} \quad \text{for all} \quad s \geq 0\,.$$

Consequently, the improper Riemann integral $\int_0^\infty e^{s\boldsymbol{A}}e^{-(\gamma+ihk)s}\,ds$ exists and we conclude that $\lim_{s\to\infty} e^{s\boldsymbol{A}}e^{-(\gamma+ihk)s} = \boldsymbol{O}$. This yields

$$\int_0^\infty e^{s\boldsymbol{A}}e^{-(\gamma+ihk)s}\,ds = -\frac{1}{\gamma + ihk - \boldsymbol{A}}\,e^{s\boldsymbol{A}}e^{-(\gamma+ihk)s}\Big|_0^\infty = \frac{1}{\gamma + ihk - \boldsymbol{A}} \tag{7.3}$$

and therefore

$$\int_0^\infty e^{s\boldsymbol{A}} \sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\,ds = \sum_{k=-m}^{m} a_k \int_0^\infty e^{s\boldsymbol{A}}e^{-(\gamma+ihk)s}\,ds = \sum_{k=-m}^{m} a_k\,\frac{1}{\gamma + ihk - \boldsymbol{A}}$$

as claimed above. ❑

Comparing equation (7.3) with the general formula of the Laplace transform, which is for a function $f : [0, \infty) \to \mathbb{C}$ defined as

$$\mathcal{L}f(t) = \int_0^\infty f(s) e^{-ts}\, ds\,,$$

the inverse $(z_k \boldsymbol{I} - \boldsymbol{A})^{-1}$ can be understood as the Laplace transform of $f(s) = e^{s\boldsymbol{A}}$ at the point $t = z_k = \gamma + ihk$ (e.g., [40], Section 11.1). In this context, the parameter $\gamma > 0$ causes an exponential damping that will be relevant for later purposes.

We have already seen above that if $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$, the same holds true for $\tau\boldsymbol{A}$, that is, $W(\tau\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ for $\tau \geq 0$. Without loss of generality, we thus state our theorems in the following for $\tau = 1$. All subsequent results remain valid for $\tau\boldsymbol{A}$, the matrix $\boldsymbol{A}$ has just to be replaced by the scaled matrix $\tau\boldsymbol{A}$ everywhere.

The estimate for the best approximation of $\varphi_\ell(\boldsymbol{A})$ in the subspace $\mathcal{R}_{2m+1}(\boldsymbol{A})$ will lead to a problem of best trigonometric approximation in the space L^1 on the unit circle $\mathbb{T}$. We refer to $\mathbb{T}$ as the real numbers with the identification of points modulo 2π, and the space $L^1(\mathbb{T})$ consists of all 2π-periodic functions satisfying $\int_{\mathbb{T}} |f(s)|\, ds < \infty$. To bound the error of the approximation problem on $L^1(\mathbb{T})$, we will need a similar modulus of smoothness as in Chapter 5 and the concept of bounded variation. Following [14], we state the following two definitions.

Definition 7.3 *A 2π-periodic function f is of bounded variation on the unit circle $\mathbb{T}$, or in short notation $f \in BV(\mathbb{T})$, if*

$$\mathrm{Var}_{\mathbb{T}} f := \sup \sum_{i=1}^{n-1} |f(x_{i+1}) - f(x_i)| < \infty\,,$$

where the supremum is taken over all partitions $x_1 < x_2 < \ldots < x_n$, $x_i \in \mathbb{T}$, $i = 1, \ldots, n$.

Functions of bounded variation can have a countable number of discontinuities α_i, whereby the left and right limits, $\lim_{s \nearrow \alpha_i} f(s)$ and $\lim_{s \searrow \alpha_i} f(s)$, have to exist at each point α_i. Furthermore, we will require the modified variation $\mathrm{Var}^*_{\mathbb{T}} f$, which is defined as the variation of a correction f^* of the function f. This corrected function f^* coincides with f except for the points α_i of discontinuity.

Definition 7.4 *For $f \in BV(\mathbb{T})$ with a countable set of discontinuities α_i, we define*

$$\mathrm{Var}^*_{\mathbb{T}} f := \mathrm{Var}_{\mathbb{T}} f^*\,,$$

where the corrected function f^* is some function that coincides with f on $\mathbb{T} \backslash \bigcup_i \{\alpha_i\}$ and takes values between $\lim_{s \nearrow \alpha_i} f(s)$ and $\lim_{s \searrow \alpha_i} f(s)$ at the points α_i of discontinuity.

With this newly defined variation, we can bound the rth modulus of smoothness $\omega_r(f, \delta)_{L^1(\mathbb{T})}$ with respect to the space $L^1(\mathbb{T})$,

$$\omega_r(f, \delta)_{L^1(\mathbb{T})} := \sup_{0 < h \leq \delta} \int_{\mathbb{T}} |\Delta_h^r(f, t)|\, dt\,, \qquad \Delta_h^r(f, t) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(t + kh)\,,$$

for a function $f \in L^1(\mathbb{T})$ (see [14], Theorem 9.3 in Chapter 2 on page 53 for $p = 1$). This modulus of smoothness has already been used in Section 5.4, cf. equation (5.14), for the maximum norm.

Lemma 7.5 *Let $f \in L^1(\mathbb{T})$ be a 2π-periodic function which can be corrected on a set of measure zero to a function g such that $g^{(r-2)}$ is absolutely continuous and the generalized (weak) derivative $g^{(r-1)}$ is of bounded variation on $\mathbb{T}$. Then the estimate*

$$\omega_r(f,\delta)_{L^1(\mathbb{T})} \le \delta^r \operatorname{Var}^*_{\mathbb{T}} g^{(r-1)}\,, \qquad \delta > 0$$

holds true.

The Stechkin inequality, obtained in the next lemma, enables us to bound the best approximation of a function $f \in L^1(\mathbb{T})$ by a trigonometric polynomial using the rth modulus of smoothness of f (see [14], Theorem 2.3 in Chapter 7 for the case $p = 1$).

Lemma 7.6 *For $r = 1, 2, \dots$ and a 2π-periodic functions $f \in L^1(\mathbb{T})$, there exists a constant $C(r)$ such that*

$$E_m(f) := \inf_{P\in\mathcal{T}_m} \int_{\mathbb{T}} |f(s) - P(s)|\, ds \;\le\; C(r)\,\omega_r\left(f, \frac{1}{m}\right)_{L^1(\mathbb{T})}\,,$$

where $\mathcal{T}_m$ denotes the set of all trigonometric polynomials of degree m with functions of the form

$$P(s) = \frac{\alpha_0}{2} + \sum_{k=1}^{m} \alpha_k \cos(ks) + \beta_k \sin(ks)\,, \qquad \alpha_k, \beta_k \in \mathbb{R}\,.$$

Since $\mathcal{T}_m$ is a finite dimensional space, there exists an element $P^* \in \mathcal{T}_m$ of best approximation for which the infimum in $E_m(f)$ is attained.

Herewith, all key requirements for the estimate of $\|\varphi_\ell(\boldsymbol{A}) - r(\boldsymbol{A})\|$ for $r(\boldsymbol{A}) \in \mathcal{R}_{2m+1}(\boldsymbol{A})$ in the next section are provided.

7.2 Error bounds

The first theorem in this section states an upper bound for the best approximation of $\varphi_\ell(\boldsymbol{A})$, $\ell \ge 1$, in the rational matrix space

$$\mathcal{R}_{2m+1}(\boldsymbol{A}) = \left\{ r(\boldsymbol{A}) \,:\, r \in \frac{\mathcal{P}_{2m}}{q_{2m+1}} \right\}\,, \qquad q_{2m+1}(z) = \prod_{k=-m}^{m} (z_k - z)\,,$$

where the poles z_k are given, as mentioned above, by $z_k = \gamma + ihk$ for $k = -m, \dots, m$ and shift $\gamma > 0$.

Theorem 7.7 *Let $\boldsymbol{A}$ be a matrix with $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$. Then for the best approximation of the matrix function $\varphi_\ell(\boldsymbol{A})$ for $\ell \ge 1$ in the subspace $\mathcal{R}_{2m+1}(\boldsymbol{A})$, it holds*

$$\inf_{r\in\frac{\mathcal{P}_{2m}}{q_{2m+1}}} \|\varphi_\ell(\boldsymbol{A}) - r(\boldsymbol{A})\| \le C_1(\ell,\gamma)\, \frac{e^{-\frac{\gamma\pi}{h}}}{1 - e^{-\frac{2\gamma\pi}{h}}} + C_2(\ell,\gamma)\, \frac{1}{(hm)^\ell}\,, \tag{7.4}$$

where the constants C_1 and C_2 depend only on γ and ℓ.

Proof. With the auxiliary results in Section 7.1, we are able to reduce the approximation problem on $\mathbb{C}^{N\times N}$ to a one-dimensional problem on the right semi-axis $[0,\infty)$. Splitting this one-dimensional problem into two appropriate subproblems on two disjoint subintervals of $[0,\infty)$, we finally obtain the two different terms on the right hand-side of the bound (7.4). In this bound, the second term arises from a trigonometric approximation on a finite spectrum and the first term is obtained due to the exponentially damping of the Laplace transform in (7.3), caused by the shift γ.

We first note that via partial fraction expansion, cf. Lemma 4.9, we have

$$\inf_{r\in\frac{\mathcal{P}_{2m}}{q_{2m+1}}} \|\varphi_\ell(\boldsymbol{A})-r(\boldsymbol{A})\| = \inf_{a_k}\left\|\varphi_\ell(\boldsymbol{A})-\sum_{k=-m}^{m} a_k\,\frac{1}{\gamma+ihk-\boldsymbol{A}}\right\|\,.$$

Using the integral representation of $\varphi_\ell(\boldsymbol{A})$ in (7.2) and of the rational function $r(\boldsymbol{A})$ in Lemma 7.2 as well as the inequality $\|e^{s\boldsymbol{A}}\|\le 1$ in Lemma 7.1, the following one-dimensional expression is obtained:

$$\begin{aligned}\left\|\varphi_\ell(\boldsymbol{A})-\sum_{k=-m}^{m} a_k\,\frac{1}{\gamma+ihk-\boldsymbol{A}}\right\| &\le \int_0^\infty \|e^{s\boldsymbol{A}}\|\left|\mathbb{1}_{[0,1]}(s)\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}-\sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\right| ds\\ &\le \int_0^\infty \left|\mathbb{1}_{[0,1]}(s)\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}-\sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\right| ds\,.\end{aligned}$$

For the rest of the proof, we always assume that the distance h between the equidistant poles z_k is chosen such that $h<\pi$, which implies that $\frac{\pi}{h}>1$. Now, the last integral is splitted into

$$\begin{aligned}&\int_0^\infty \left|\mathbb{1}_{[0,1]}(s)\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}-\sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\right| ds\\ &=\int_0^{\frac{\pi}{h}} \left|\mathbb{1}_{[0,1]}(s)\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}-\sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\right| ds+\int_{\frac{\pi}{h}}^\infty\left|\sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\right| ds \qquad (7.5)\\ &=\int_0^{\frac{\pi}{h}} \left|\mathbb{1}_{[0,1]}(s)\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}-\sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\right| ds+\sum_{j=1}^{\infty}\int_{(2j-1)\frac{\pi}{h}}^{(2j+1)\frac{\pi}{h}}\left|\sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s}\right| ds\,.\end{aligned}$$

In order to exploit the results for the best trigonometric approximation on $L^1(\mathbb{T})$ in the previous section, the first term, involving the indicator function $\mathbb{1}_{[0,1]}(s)$, must be extended to the interval $[-\frac{\pi}{h},\frac{\pi}{h})$ in a suitable way. A change of variables then leads to the required domain $\mathbb{T}=[-\pi,\pi)$ of integration. For this purpose, we have to find a new function that coincides with $\mathbb{1}_{[0,1]}(s)(1-s)^{\ell-1}/(\ell-1)!$ on $[0,\frac{\pi}{h})$ and that has sufficient smoothness properties. The last requirement is to guarantee a best possible trigonometric approximation. With regard to this, we define the function

$$g(s):=\begin{cases} C\displaystyle\int_{-1}^{s} e^{-\frac{1}{1-(2t+1)^2}}\,dt\,, & -1<s<0\,,\\ 1\,, & s\ge 0\,,\\ 0\,, & s\le -1\end{cases} \qquad (7.6)$$

with $C = (\int_{-1}^{0} e^{-\frac{1}{1-(2t+1)^2}}\,dt)^{-1}$. This function g is infinitely many times differentiable, that is, $g \in C^\infty(\mathbb{R})$. Furthermore, we set

$$f_\ell(s) := g(s) \cdot e^{\gamma s} \cdot \mathbb{1}_{[-1,1]}(s)\, \frac{(1-s)^{\ell-1}}{(\ell-1)!} \in C^{\ell-2}(\mathbb{R})\,.$$

This newly defined function belongs to $C^\infty(\mathbb{R}\backslash\{1\})$ and has a weak derivative of order $\ell-1$ (see Figure 7.2 for $\ell = 3$).

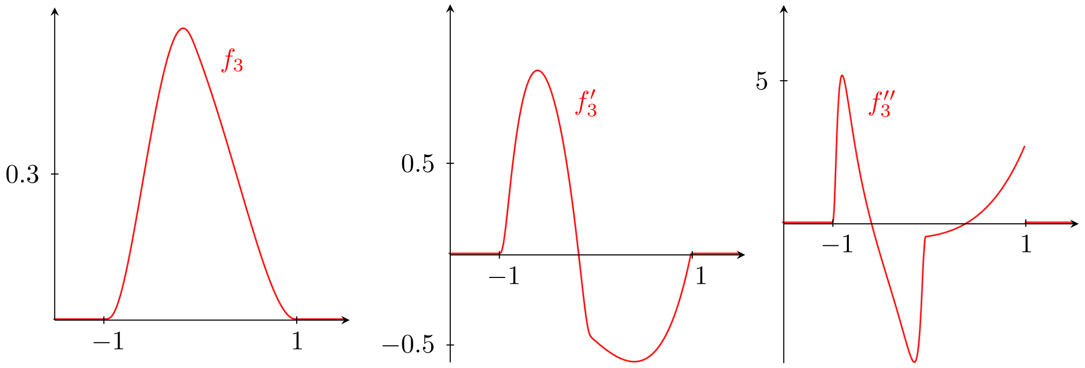


Figure 7.2: Plot of f_3, f_3', and the weak derivative f_3'' with jump discontinuity at 1.

Since $g(s) = 1$ for $s \geq 0$, it follows

$$\begin{aligned}
&\int_0^{\frac{\pi}{h}} \left| \mathbb{1}_{[0,1]}(s)\, \frac{(1-s)^{\ell-1}}{(\ell-1)!} - \sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s} \right| ds \\
&\quad = \int_0^{\frac{\pi}{h}} \left| e^{-\gamma s} e^{\gamma s} \cdot g(s) \cdot \mathbb{1}_{[0,1]}(s)\, \frac{(1-s)^{\ell-1}}{(\ell-1)!} - \sum_{k=-m}^{m} a_k e^{-\gamma s} e^{-ihks} \right| ds \\
&\quad = \int_0^{\frac{\pi}{h}} e^{-\gamma s} \left| f_\ell(s) - \sum_{k=-m}^{m} a_k e^{-ihks} \right| ds \leq \int_0^{\frac{\pi}{h}} \left| f_\ell(s) - \sum_{k=-m}^{m} a_k e^{-ihks} \right| ds \\
&\quad \leq \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \left| f_\ell(s) - \sum_{k=-m}^{m} a_k e^{-ihks} \right| ds\,.
\end{aligned}$$

The coefficients $a_{-m}, \dots, a_m$ are now chosen according to the best approximation for $f_\ell(s)$ in the space $\mathcal{T}_m$ of all real trigonometric polynomials of degree m. This leads to

$$\begin{aligned}
\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \left| f_\ell(s) - \sum_{k=-m}^{m} a_k e^{-ihks} \right| ds &= \min_{b_k} \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \left| f_\ell(s) - \sum_{k=-m}^{m} b_k e^{-ihks} \right| ds \\
&= \frac{1}{h} \cdot \min_{b_k} \int_{-\pi}^{\pi} \left| f_\ell\left(\frac{s}{h}\right) - \sum_{k=-m}^{m} b_k e^{-iks} \right| ds = \frac{1}{h} \cdot E_m\left(f_\ell\left(\frac{\cdot}{h}\right)\right)\,,
\end{aligned} \tag{7.7}$$

if the real function $f_\ell(\frac{\cdot}{h})\big|_{\mathbb{T}}$ is extended periodically to a function in $L^1(\mathbb{T})$. Since the best approximation of a real function has to be real as well, the complex coefficients a_k have to be taken in such a way that $a_{-k} = \overline{a_k}$. Then we have

$$a_k e^{-iks} + a_{-k} e^{iks} = 2\,\mathrm{Re}(a_k)\cos(ks) + 2\,\mathrm{Im}(a_k)\sin(ks)$$

and $\sum_{k=-m}^{m} a_k e^{-iks}$ is a real trigonometric polynomial. The application of Lemma 7.6 now yields

$$E_m\left(f_\ell\left(\frac{\cdot}{h}\right)\right) \le C(r)\,\omega_r\left(f_\ell\left(\frac{\cdot}{h}\right), \frac{1}{m}\right)_{L^1(\mathbb{T})}.$$

The $(\ell-2)$nd derivative of $f_\ell(\frac{\cdot}{h})$ is absolutely continuous and the $(\ell-1)$st weak derivative is of bounded variation on $\mathbb{T}$. Hence, we are able to apply Lemma 7.5 with r equal to ℓ, to obtain

$$\omega_r\left(f_\ell\left(\frac{\cdot}{h}\right), \frac{1}{m}\right)_{L^1(\mathbb{T})} \le \frac{1}{m^\ell}\operatorname{Var}^*_{\mathbb{T}}\, f_\ell^{(\ell-1)}\left(\frac{\cdot}{h}\right) = \frac{1}{m^\ell}\operatorname{Var}_{\mathbb{T}}\,\left(f_\ell^{(\ell-1)}\left(\frac{\cdot}{h}\right)\right)^*$$
$$=: \frac{1}{m^\ell}\operatorname{Var}_{\mathbb{T}}\, u_\ell(\cdot)\,,$$

where

$$u_\ell(\cdot) = \left(f_\ell^{(\ell-1)}\left(\frac{\cdot}{h}\right)\right)^* = \left(\frac{d^{\ell-1}}{ds^{\ell-1}}\left[f_\ell\left(\frac{\cdot}{h}\right)\right]\right)^* \in BV(\mathbb{T})\,.$$

According to Definition 7.4, the function $u_\ell(\cdot)$ is a suitable correction of the function $f_\ell^{(\ell-1)}(\frac{\cdot}{h})$ with jump discontinuity at h. In our case, we choose

$$u_\ell(s) := \begin{cases} f_\ell^{(\ell-1)}\left(\dfrac{s}{h}\right), & s \in \mathbb{T}\backslash\{h\}\,,\\ \dfrac{1}{2}\left(\lim\limits_{s\nearrow h} f_\ell^{(\ell-1)}\left(\dfrac{s}{h}\right) + \lim\limits_{s\searrow h} f_\ell^{(\ell-1)}\left(\dfrac{s}{h}\right)\right), & s = h\,. \end{cases}$$

Moreover, we define the transformed function

$$\widetilde{u}_\ell(s) := \begin{cases} f_\ell^{(\ell-1)}(s)\,, & s \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right)\backslash\{1\}\,,\\ \dfrac{1}{2}\left(\lim\limits_{s\nearrow 1} f_\ell^{(\ell-1)}(s) + \lim\limits_{s\searrow 1} f_\ell^{(\ell-1)}(s)\right), & s = 1\,, \end{cases}$$

which does no longer depend on h. Due to the definition of $f_\ell(s)$, the function $\widetilde{u}_\ell(s)$ is equal to zero for $s \le -1$ and $s > 1$. With the chain rule we obtain

$$\operatorname{Var}_{\mathbb{T}} u_\ell(\cdot) = \frac{1}{h^{\ell-1}}\operatorname{Var}_{[-\frac{\pi}{h},\frac{\pi}{h})}\widetilde{u}_\ell(\cdot) = \frac{1}{h^{\ell-1}}\operatorname{Var}_{[-1,1]}\widetilde{u}_\ell(\cdot)\,.$$

Altogether, we have

$$\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \left| f_\ell(s) - \sum_{k=-m}^{m} a_k e^{-ihks}\right| ds \le C_2(\ell,\gamma)\,\frac{1}{(hm)^\ell}\,, \qquad C_2(\ell,\gamma) = C(\ell)\operatorname{Var}_{[-1,1]}\widetilde{u}_\ell(\cdot)\,,$$

where the property $\widetilde{u}_\ell(\cdot) \in BV(\mathbb{T})$ assures that $C_2(\ell,\gamma) < \infty$. Consequently, the first term in (7.5) is covered. Our next task is to bound the second term. As mentioned above, we will take advantage of the fact that the parameter γ in the exponential function leads to a damping in the remaining domain of integration. Noting that

$$e^{-\gamma s} \le e^{-\gamma(2j-1)\frac{\pi}{h}} \quad \text{for} \quad s \in \left[(2j-1)\frac{\pi}{h}, (2j+1)\frac{\pi}{h}\right], \quad j = 1, 2, \dots\,,$$

we conclude by the periodicity of e^{ihks} and the geometric series that

$$\begin{aligned}\sum_{j=1}^{\infty}\int_{(2j-1)\frac{\pi}{h}}^{(2j+1)\frac{\pi}{h}}\left|\sum_{k=-m}^{m}a_k e^{-(\gamma+ihk)s}\right|ds &\le \sum_{j=1}^{\infty}e^{-\gamma(2j-1)\frac{\pi}{h}}\int_{(2j-1)\frac{\pi}{h}}^{(2j+1)\frac{\pi}{h}}\left|\sum_{k=-m}^{m}a_k e^{-ihks}\right|ds\\ &= e^{\gamma\frac{\pi}{h}}\left(\sum_{j=0}^{\infty}\left(e^{-2\gamma\frac{\pi}{h}}\right)^j-1\right)\cdot\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}\left|\sum_{k=-m}^{m}a_k e^{-ihks}\right|ds\\ &= \frac{e^{-\gamma\frac{\pi}{h}}}{1-e^{-2\gamma\frac{\pi}{h}}}\cdot\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}\left|\sum_{k=-m}^{m}a_k e^{-ihks}\right|ds\,.\end{aligned}$$

The integral term can be estimated by exploiting that the coefficients $a_{-m},\dots,a_m$ have been chosen in (7.7) according to the best approximation. This gives

$$\begin{aligned}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}\left|\sum_{k=-m}^{m}a_k e^{-ihks}\right|ds &= \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}\left|\sum_{k=-m}^{m}a_k e^{-ihks}-f_\ell(s)+f_\ell(s)\right|ds\\ &\le \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}|f_\ell(s)|\,ds+\min_{b_k}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}\left|f_\ell(s)-\sum_{k=-m}^{m}b_k e^{-ihks}\right|ds\\ &\le 2\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}|f_\ell(s)|\,ds\,,\end{aligned}$$

if we set $b_{-m}=\ldots=b_m=0$. Inserting the definition of f_ℓ, we get

$$\begin{aligned}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}|f_\ell(s)|\,ds &= \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}}\left|g(s)\cdot e^{\gamma s}\cdot\mathbb{1}_{[-1,1]}(s)\,\frac{(1-s)^{\ell-1}}{(\ell-1)!}\right|ds\\ &\le e^{\gamma}\int_{-1}^{1}\left|\frac{(1-s)^{\ell-1}}{(\ell-1)!}\right|ds \le e^{\gamma}\frac{2^\ell}{\ell!}\,.\end{aligned}$$

Thus, the second term in (7.5) fulfills the inequality

$$\sum_{j=1}^{\infty}\int_{(2j-1)\frac{\pi}{h}}^{(2j+1)\frac{\pi}{h}}\left|\sum_{k=-m}^{m}a_k e^{(-\gamma-ihk)s}\right|ds \le C_1(\ell,\gamma)\,\frac{e^{-\gamma\frac{\pi}{h}}}{1-e^{-2\gamma\frac{\pi}{h}}}\,,\qquad C_1(\ell,\gamma)=e^{\gamma}\frac{2^{\ell+1}}{\ell!}\,.$$

Finally, we obtain the desired error bound. ❑

The assumption $h<\pi$ in the proof of Theorem 7.7 is not mandatory. Just as well, we can choose another condition for the distance h between the simple poles $z_k=\gamma+ihk$. Then we need a suitable change of variables in the proof, in order to end up again with a trigonometric approximation problem on $L^1(\mathbb{T})$.

The attentive reader may have recognized that we just considered the approximation of $\varphi_\ell(\boldsymbol{A})$ in the space $\mathcal{R}_{2m+1}(\boldsymbol{A})$, containing all rational matrix functions of the form $p_{2m}(\boldsymbol{A})/q_{2m+1}(\boldsymbol{A})$ with $p_{2m}\in\mathcal{P}_{2m}$, whereas we are finally interested in an approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the Krylov subspace

$$\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})=\left\{r(\boldsymbol{A})\boldsymbol{v}\;:\;r\in\frac{\mathcal{P}_{2m+1}}{q_{2m+1}}\right\}\,,$$

whose numerator polynomial p_{2m+1} has a degree one larger than in the space $\mathcal{R}_{2m+1}(\boldsymbol{A})$. Compared to the subspace $\mathcal{R}_{2m+1}(\boldsymbol{A})\boldsymbol{v} = \mathcal{P}_{2m}(\boldsymbol{A})/q_{2m+1}(\boldsymbol{A})\boldsymbol{v}$, one could say that the rational Krylov subspace $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$ contains in addition the vector $\boldsymbol{v}$. This additional vector is required to ensure that the condition $r(\boldsymbol{A})\boldsymbol{v} = r(\boldsymbol{A}_m)\boldsymbol{v}$ holds for every rational function $r \in \mathcal{P}_{2m}/q_{2m+1}$, where $\boldsymbol{A}_m = \boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m$ and $\boldsymbol{P}_m$ is the orthogonal projection onto $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$. In particular, this requirement necessitates that $\boldsymbol{P}_m\boldsymbol{v} = \boldsymbol{v}$, which is only fulfilled if $\boldsymbol{v}$ is an element of the considered Krylov subspace. The following simple example illustrates this claim.

We consider $r(z) = \frac{1}{\gamma - z} \in \mathcal{P}_{2m}/q_{2m+1}$ and assume that $\boldsymbol{P}_m$ is the orthogonal projection onto $\mathcal{R}_{2m+1}(\boldsymbol{A})\boldsymbol{v}$. The requirement $r(\boldsymbol{A})\boldsymbol{v} = r(\boldsymbol{A}_m)\boldsymbol{v}$ is then equivalent to

$$\begin{aligned}
&\frac{1}{\gamma - \boldsymbol{A}}\,\boldsymbol{v} = \frac{1}{\gamma - \boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m}\,\boldsymbol{v} \\
&\qquad\qquad \Longleftrightarrow \quad (\gamma\boldsymbol{I} - \boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m)\,\frac{1}{\gamma - \boldsymbol{A}}\,\boldsymbol{v} = \frac{\gamma}{\gamma - \boldsymbol{A}}\,\boldsymbol{v} - \boldsymbol{P}_m\,\frac{\boldsymbol{A}}{\gamma - \boldsymbol{A}}\,\boldsymbol{v} = \boldsymbol{v} \\
&\qquad\qquad \Longleftrightarrow \quad \frac{\gamma}{\gamma - \boldsymbol{A}}\,\boldsymbol{v} + \boldsymbol{P}_m\left(\boldsymbol{I} - \frac{\gamma}{\gamma - \boldsymbol{A}}\right)\boldsymbol{v} = \boldsymbol{P}_m\boldsymbol{v} = \boldsymbol{v}\,,
\end{aligned}$$

showing that $\boldsymbol{P}_m\boldsymbol{v} = \boldsymbol{v}$ has to be fulfilled.

Due to the relation

$$\mathcal{R}_{2m+1}(\boldsymbol{A})\boldsymbol{v} + \operatorname{span}\{\boldsymbol{v}\} = \frac{\mathcal{P}_{2m}(\boldsymbol{A})}{q_{2m+1}(\boldsymbol{A})}\,\boldsymbol{v} + \operatorname{span}\{\boldsymbol{v}\} = \frac{\mathcal{P}_{2m+1}(\boldsymbol{A})}{q_{2m+1}(\boldsymbol{A})}\,\boldsymbol{v} = \mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})\,,$$

we study, from now on, the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ by $\varphi_\ell(\boldsymbol{A}_m)$ for $\ell \geq 1$, where $\boldsymbol{A}_m$ is the restriction of $\boldsymbol{A}$ to the rational Krylov subspace $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$. With the help of Theorem 7.7, we are able to bound the best approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the subspace $\mathcal{R}_{2m+1}(\boldsymbol{A})\boldsymbol{v}$. This raises the question whether Theorem 7.7 can nevertheless be applied for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$. The answer to this question is yes: Since $\mathcal{P}_{2m}/q_{2m+1} \subset \mathcal{P}_{2m+1}/q_{2m+1}$, adding the vector $\boldsymbol{v}$ to the subspace causes no problem.

Theorem 7.8 *Let the matrix $\boldsymbol{A}$ satisfy $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$ and let $\boldsymbol{A}_m = \boldsymbol{P}_m\boldsymbol{A}\boldsymbol{P}_m$ be the restriction of $\boldsymbol{A}$ to $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$ via orthogonal projection. Then the error of the rational Krylov subspace approximation can be bounded by*

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\| \leq 2\left[C_1(\ell,\gamma)\,\frac{e^{-\frac{\gamma\pi}{h}}}{1 - e^{-\frac{2\gamma\pi}{h}}} + C_2(\ell,\gamma)\,\frac{1}{(hm)^\ell}\right]\|\boldsymbol{v}\|\,,$$

where C_1 and C_2 are the same constants as in Theorem 7.7.

Proof. By Lemma 4.12, the Krylov subspace approximation in $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$ is exact for every rational function $r \in \mathcal{P}_{2m+1}/q_{2m+1}$, where $q_{2m+1}(z) = \prod_{k=-m}^{m}(\gamma + ihk - z)$. With $r(\boldsymbol{A})\boldsymbol{v} = r(\boldsymbol{A}_m)\boldsymbol{v}$, it follows

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\| \leq \|\varphi_\ell(\boldsymbol{A}) - r(\boldsymbol{A})\|\|\boldsymbol{v}\| + \|\varphi_\ell(\boldsymbol{A}_m) - r(\boldsymbol{A}_m)\|\|\boldsymbol{v}\|$$

for all $r \in \mathcal{P}_{2m+1}/q_{2m+1}$. We pick this rational function r especially from the space $\mathcal{P}_{2m}/q_{2m+1} \subseteq \mathcal{P}_{2m+1}/q_{2m+1}$ and use the already known relation $W(\boldsymbol{A}_m) \subseteq W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$.

In this case, we are able to estimate both terms on the right hand-side as in the proof of Theorem 7.7. This gives

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\| \leq 2\|\boldsymbol{v}\| \int_0^\infty \left| \mathbb{1}_{[0,1]}(s)\, \frac{(1-s)^{\ell-1}}{(\ell-1)!} - \sum_{k=-m}^{m} a_k e^{-(\gamma+ihk)s} \right| ds\,.$$

From here, the estimate of the one-dimensional approximation problem proceeds analogously to the proof of Theorem 7.7 and we end up with

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\| \leq 2 \left[C_1(\ell,\gamma)\, \frac{e^{-\frac{\gamma\pi}{h}}}{1-e^{-\frac{2\gamma\pi}{h}}} + C_2(\ell,\gamma)\, \frac{1}{(hm)^\ell} \right] \|\boldsymbol{v}\|\,,$$

which concludes the proof. ❑

If the first term of the error bound in Theorem 7.8 is small enough, the second term predicts a sublinear convergence behavior of order $\mathcal{O}\big((hm)^{-\ell}\big)$. The obtained error bound is completely independent of the matrix $\boldsymbol{A}$ and therefore guarantees a uniform approximation for arbitrary matrices whose field of values is located somewhere in the left complex half-plane. In the case that $\boldsymbol{A}$ represents a spatial discretization matrix of a differential operator, we proved a convergence rate that is entirely independent of the chosen mesh size and the refinement of the grid.

It has to be taken into account that only the second term of our error bound decreases with the dimension of the Krylov subspace. In order to obtain an efficient and useful error estimate, it is of great importance that the occurring free parameters are reasonably chosen. In contrast to the shift-and-invert and the extended Krylov subspace method, where we use a single repeated pole γ, we not only have to think about a suitable choice of the shift γ, but also about an appropriate selection for the distance h between the different poles of our rational Krylov subspace method. This will be the content of the next section.

7.3 Choice of the parameters γ and h

In this section, we want to investigate suitable choices for the free parameters γ and h. To this end, let us recall that the error for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the rational Krylov subspace $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$ behaves like

$$\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\| \leq 2 \left[C_1(\ell,\gamma)\, \frac{e^{-\frac{\gamma\pi}{h}}}{1-e^{-\frac{2\gamma\pi}{h}}} + C_2(\ell,\gamma)\, \frac{1}{(hm)^\ell} \right] \|\boldsymbol{v}\|\,.$$

For the subsequent discussion, we have to understand in what way the two occurring constants $C_1(\ell,\gamma)$ and $C_2(\ell,\gamma)$ depend on γ. Looking back to the proof of Theorem 7.7, we already know that

$$C_1(\ell,\gamma) = e^{\gamma}\frac{2^{\ell+1}}{\ell!}\,, \qquad C_2(\ell,\gamma) = C(\ell)\mathrm{Var}_{[-1,1]}\widetilde{u}_\ell(\cdot)\,,$$

where $\widetilde{u}_\ell(s)$ was defined as $f_\ell^{(\ell-1)}(s)$ for $s \in [-\frac{\pi}{h}, \frac{\pi}{h})\backslash\{1\}$ and as the mean of the left and right limit of $f_\ell^{(\ell-1)}(s)$ on the jump at the point 1. Hereby, the function f_ℓ was given as

$$f_\ell(s) = g(s)\cdot e^{\gamma s}\cdot \mathbb{1}_{[-1,1]}(s)\, \frac{(1-s)^{\ell-1}}{(\ell-1)!} \in C^{\ell-2}(\mathbb{R})$$

with the auxiliary function g defined in (7.6). To estimate the variation of $\widetilde{u}_\ell$, we exploit the fact that $\widetilde{u}_\ell$ is only supported on $[-1,1]$ and differentiable on $[-1,1)$. Moreover, it is well-known that for a function f, which is differentiable on $[a,b]$, we have

$$\operatorname{Var}_{[a,b]} f(\cdot) = \int_a^b |f'(s)|\, ds$$

(for instance, [68], Section 16.1). This leads to the inequality

$$\operatorname{Var}_{[-1,1]} \widetilde{u}_\ell(\cdot) \le \int_{-1}^{1} |f_\ell^{(\ell)}(s)|\, ds + \sup_{s\in[-1,1)} |f_\ell^{(\ell-1)}(s)|\,,$$

where the last term is required to cover the jump discontinuity of $\widetilde{u}_\ell$ at 1. Hence, we are now concerned with the estimate of $|f_\ell^{(\ell-1)}(s)|$ as well as $|f_\ell^{(\ell)}(s)|$ for $s \in [-1,1)$. For the function g in (7.6), we define the constant

$$C_g := \max_{k=0,\dots,\ell} \max_{s\in[-1,1]} |g^{(k)}(s)| < \infty\,,$$

which is independent of γ, since g itself does not depend on γ. The general Leibniz rule yields

$$\begin{aligned} |f_\ell^{(\ell-1)}(s)| &= \left| \sum_{k=0}^{\ell-1} \binom{\ell-1}{k} g^{(\ell-1-k)}(s) \left(e^{\gamma s} \frac{(1-s)^{\ell-1}}{(\ell-1)!} \right)^{(k)} \right| \\ &\le C_g \cdot \sum_{k=0}^{\ell-1} \binom{\ell-1}{k} \left| \left(e^{\gamma s} \frac{(1-s)^{\ell-1}}{(\ell-1)!} \right)^{(k)} \right|, \qquad s \in [-1,1)\,. \end{aligned}$$

Using the Leibniz rule once more, we obtain

$$\left| \left(e^{\gamma s} \frac{(1-s)^{\ell-1}}{(\ell-1)!} \right)^{(k)} \right| \le \sum_{j=0}^{k} \binom{k}{j} \left| (e^{\gamma s})^{(k-j)} \right| \left| \left(\frac{(1-s)^{\ell-1}}{(\ell-1)!} \right)^{(j)} \right| \le 2\, e^{\gamma} (1+\gamma)^k$$

by the binomial identity. The last inequality follows from the fact that $\left| \left(\frac{(1-s)^{\ell-1}}{(\ell-1)!} \right)^{(j)} \right| \le 2$ for all $j = 0,\dots,k$, $\ell \ge 1$, and $s \in [-1,1)$. Then a further application of the binomial identity provides the bound

$$|f_\ell^{(\ell-1)}(s)| \le 2\, C_g e^{\gamma} \cdot \sum_{k=0}^{\ell-1} \binom{\ell-1}{k} (1+\gamma)^k = 2\, C_g e^{\gamma} (2+\gamma)^{\ell-1}\,.$$

An analogous estimate for the second term $|f_\ell^{(\ell)}(s)|$ yields $|f_\ell^{(\ell)}(s)| \le 2\, C_g e^{\gamma}(2+\gamma)^{\ell}$. Altogether, this gives

$$\operatorname{Var}_{[-1,1]} \widetilde{u}_\ell(\cdot) \le \int_{-1}^{1} 2\, C_g e^{\gamma} (2+\gamma)^{\ell}\, ds + 2\, C_g e^{\gamma} (2+\gamma)^{\ell-1} \le 6\, C_g e^{\gamma} (2+\gamma)^{\ell}$$

and thus

$$C_2(\ell,\gamma) = C(\ell) \operatorname{Var}_{[-1,1]} \widetilde{u}_\ell(\cdot) \le C(\ell) e^{\gamma} (2+\gamma)^{\ell}$$

with a constant $C(\ell)$ that depends only on ℓ.

We are now prepared to discuss suitable choices for γ and h, by taking the whole expression

$$C_1(\ell,\gamma)\,\frac{e^{-\frac{\gamma\pi}{h}}}{1-e^{-\frac{2\gamma\pi}{h}}}+C_2(\ell,\gamma)\,\frac{1}{(hm)^\ell}\le\frac{2^{\ell+1}}{\ell!}\,\frac{e^{\gamma(1-\frac{\pi}{h})}}{1-e^{-\frac{2\gamma\pi}{h}}}+C(\ell)\,e^\gamma(2+\gamma)^\ell\frac{1}{(hm)^\ell} \tag{7.8}$$

into account. We first observe that the shift $\gamma > 0$ should neither be chosen too small nor too large, since otherwise either the first or the second term in (7.8) becomes quite large. Apart from this, a suitable choice for h is also not obvious. What is certain is that the parameter h is restricted by $0 < h < \pi$, in accordance with the assumption in the proof of Theorem 7.7.

The first term in (7.8) is independent of the iteration index and therefore does not decrease with m. A first possibility for the choice of the two parameters might be to define h and γ in such a way that

$$C_1(\ell,\gamma)\,\frac{e^{-\frac{\gamma\pi}{h}}}{1-e^{-\frac{2\gamma\pi}{h}}}\le tol\,,$$

where tol is a given tolerance, e.g., $tol = 10^{-4}$. We should avoid very small values for h, since the second term in (7.8) is of order $\mathcal{O}(h^{-\ell})$. Furthermore, we would have to expect instability problems in the orthogonalization process of the parallel rational Krylov subspace decomposition for small h, since $(z_k\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{v}\approx(z_{k+1}\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{v}$. This stability problem can be overcome by using the serial rational Arnoldi decomposition in Algorithm 4.10 or the rational Krylov subspace procedure by Ruhe. On the other hand, values of h greater than one unfortunately imply that the prefactor $e^\gamma(2+\gamma)^\ell$ of the second term is quite large. For the case $\ell = 1$, the correlation between h, γ, and $e^\gamma(2+\gamma)^\ell$ is demonstrated in the following table:

$C_1(\ell,\gamma)\frac{e^{-\frac{\gamma\pi}{h}}}{1-e^{-\frac{2\gamma\pi}{h}}}$	h	γ	$e^\gamma(2+\gamma)$
$\le 10^{-4}$	0.25	≥ 1	≥ 8.2
$\le 10^{-4}$	0.5	≥ 2.1	≥ 33.5
$\le 10^{-4}$	1	≥ 5	≥ 1038.9
$\le 10^{-4}$	2	≥ 18.6	$\ge 2.5\cdot 10^9$

Table 7.1: Correlation between h, γ, and $e^\gamma(2+\gamma)^\ell$ for $\ell = 1$.

A second possible strategy is to choose a fixed γ of moderate size, for example $\gamma = 1$, and to select h such that both terms in (7.8) equally decrease with the dimension m of the rational Krylov subspace. For this, we have to solve the equation $e^{-\gamma\pi/h} = (hm)^{-\ell}$ for h. If we rearrange this expression in a suitable way, we have

$$\begin{aligned} e^{-\frac{\gamma\pi}{h}}=\frac{1}{(hm)^\ell}\quad&\Longleftrightarrow\quad (hm)^\ell=\left(e^{\frac{\gamma\pi}{h\ell}}\right)^\ell\quad\Longleftrightarrow\quad hm=e^{\frac{\gamma\pi}{h\ell}}\\ &\Longleftrightarrow\quad \frac{\gamma\pi m}{\ell}=\frac{\gamma\pi}{h\ell}\,e^{\frac{\gamma\pi}{h\ell}}\,. \end{aligned}\tag{7.9}$$

At this point, the so-called Lambert W-function $W(z)$ comes into play. It is given as the inverse function of $f(x) = xe^x = z$ and therefore provides a solution of the equation $z = W(z)e^{W(z)}$. Identifying z with $\frac{\gamma\pi m}{\ell}$ and $W(z)$ with $\frac{\gamma\pi}{h\ell}$ in (7.9), the relation

$$W\Big(\frac{\gamma\pi m}{\ell}\Big) = \frac{\gamma\pi}{h\ell} \qquad \text{or} \qquad h = \frac{\gamma\pi}{\ell}\,\frac{1}{W\big(\frac{\gamma\pi m}{\ell}\big)} \tag{7.10}$$

is obtained. Using the well-known estimate (cf. [35])

$$\ln(z) - \ln\big(\ln(z)\big) \le W(z) \le \ln(z)\,, \qquad z \ge e\,,\quad z \in \mathbb{R}\,,$$

we find for $h = \frac{\gamma\pi}{\ell} W\big(\frac{\gamma\pi m}{\ell}\big)^{-1}$ the estimates

$$e^{-\frac{\gamma\pi}{h}} = e^{-\ell W\left(\frac{\gamma\pi m}{\ell}\right)} \le e^{-\ell\left[\ln\left(\frac{\gamma\pi m}{\ell}\right) - \ln\left(\ln\left(\frac{\gamma\pi m}{\ell}\right)\right)\right]} = \left(\frac{\ell}{\gamma\pi m}\right)^{\ell}\left(\ln\left(\frac{\gamma\pi m}{\ell}\right)\right)^{\ell} = \mathcal{O}\left(\frac{\ln(m)^{\ell}}{m^{\ell}}\right)$$

and

$$\frac{1}{(hm)^{\ell}} = \left(\frac{\ell}{\gamma\pi m}\right)^{\ell} W^{\ell}\left(\frac{\gamma\pi m}{\ell}\right) \le \left(\frac{\ell}{\gamma\pi m}\right)^{\ell}\left(\ln\left(\frac{\gamma\pi m}{\ell}\right)\right)^{\ell} = \mathcal{O}\left(\frac{\ln(m)^{\ell}}{m^{\ell}}\right)\,.$$

Consequently, the terms in (7.8) behave like

$$C_1(\ell,\gamma)\,\frac{e^{-\frac{\gamma\pi}{h}}}{1 - e^{-\frac{2\gamma\pi}{h}}} + C_2(\ell,\gamma)\,\frac{1}{(hm)^{\ell}} \le C(\ell,\gamma)\left(\frac{\ln(m)}{m}\right)^{\ell}\,. \tag{7.11}$$

7.4 Comparison with a fixed rational approximation

This section draws a comparison between the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$, $\ell \ge 2$, in the rational Krylov subspace $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$ and a fixed rational approximation using the same poles $z_k = \gamma + ihk$, that is,

$$\varphi_\ell(\boldsymbol{A})\boldsymbol{v} \approx \sum_{k=-m}^{m} \frac{c_k}{z_k - \boldsymbol{A}}\,\boldsymbol{v}$$

with given coefficients $c_{-m},\dots,c_m \in \mathbb{C}$. Such a fixed approximation is obtained with the help of the ideas in Stenger [80] as well as López-Fernandéz and Palencia [53] applied to the following useful representation of the φ_ℓ-functions by the Cauchy integral formula.

Lemma 7.9 *For $\ell \ge 2$ and $z \in \mathbb{C}_0^-$, we have*

$$\varphi_\ell(z) = \frac{1}{2\pi i}\int_\Gamma \frac{e^\xi}{\xi^\ell}\,\frac{1}{\xi - z}\,d\xi\,,$$

where Γ is a curve with parametrization $\Gamma(t) = \gamma + it$, $\gamma > 0$, and the parameter $t \in \mathbb{R}$ runs from $-\infty$ to $+\infty$.

Proof. We use the representation

$$\varphi_\ell(z) = \frac{e^z}{z^\ell} - \sum_{k=0}^{\ell-1}\frac{z^{k-\ell}}{k!}\,, \qquad \varphi_\ell(0) = \frac{1}{\ell!}\,,$$

for the φ_ℓ-functions, cf. relation (3.13). Since φ_ℓ is analytic in $\mathbb{C}$, the Cauchy integral formula yields

$$\varphi_\ell(z) = \frac{1}{2\pi i}\int_{\widetilde{\Gamma}} \frac{e^\xi}{\xi^\ell}\,\frac{1}{\xi - z}\,d\xi - \frac{1}{2\pi i}\int_{\widetilde{\Gamma}} \sum_{k=0}^{\ell-1}\frac{\xi^{k-\ell}}{k!}\,\frac{1}{\xi - z}\,d\xi\,, \tag{7.12}$$

where $\widetilde{\Gamma}$ is a simple closed rectifiable curve with winding number one around $z \in \mathbb{C}_0^-$. For the contour $\widetilde{\Gamma}$, we take particularly the boundary $\widetilde{\Gamma}_1 + \widetilde{\Gamma}_2$ of the left semicircle of the disk with center point $\gamma > 0$ and radius $R > 0$, which is parametrized by

$$\widetilde{\Gamma}_1(t) = \gamma + it\,, \quad t \in [-R, R]\,, \qquad \widetilde{\Gamma}_2(t) = \gamma + Re^{it}\,, \quad t \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right)\,,$$

such that $z \in \operatorname{int}(\widetilde{\Gamma}) = \operatorname{int}(\widetilde{\Gamma}_1 + \widetilde{\Gamma}_2)$. Using the residue theorem, we see that the second term on the right-hand side of (7.12) vanishes. Moreover, it is easy to verify that

$$\lim_{R\to\infty} \frac{1}{2\pi i}\int_{\widetilde{\Gamma}_2} \frac{e^\xi}{\xi^\ell}\,\frac{1}{\xi - z}\,d\xi = 0\,.$$

Since, for $\ell \geq 2$, we have

$$\lim_{R\to\infty}\left|\int_{\widetilde{\Gamma}_1} \frac{e^\xi}{\xi^\ell}\,\frac{1}{\xi - z}\,d\xi\right| \leq \lim_{R\to\infty}\int_{-R}^{R} \frac{e^\gamma}{|\gamma + it|^\ell}\,\frac{1}{|\gamma + it - z|}\,dt \leq \frac{e^\gamma}{\gamma}\int_{-\infty}^{\infty}\frac{1}{|\gamma + it|^\ell}\,dt < \infty\,,$$

the improper integral over $\widetilde{\Gamma}_1$ exists and the desired statement of the lemma is proved with $\Gamma(t) = \widetilde{\Gamma}_1(t)$ for $t \in \mathbb{R}$. ❑

The limit

$$\lim_{t\to\pm\infty} \frac{e^{\gamma+it}}{(\gamma + it)^\ell}\,\frac{1}{\gamma + it - z} = 0$$

of the integrand in Lemma 7.9 suggests an approximation of $\varphi_\ell(z)$ by the application of a truncated trapezoidal rule with step size h to the integral representation. More precisely, we consider the approximation

$$\varphi_\ell(z) \approx \frac{h}{2\pi}\sum_{k=-m}^{m} \frac{e^{\gamma+ihk}}{(\gamma + ihk)^\ell}\,\frac{1}{\gamma + ihk - z} =: S_m(\varphi_\ell, h, z)\,,$$

where $S_m(\varphi_\ell, h, z)$ is a rational function with the same poles $z_k = \gamma + ihk$ as in our rational Krylov subspace method above. The quadrature error of this approach will be estimated in the next lemma.

Lemma 7.10 *The error of the truncated trapezoidal rule applied to the integral representation of $\varphi_\ell(z)$, $\ell \geq 2$, in Lemma 7.9 is for all $z \in \mathbb{C}_0^-$ bounded by*

$$|\varphi_\ell(z) - S_m(\varphi_\ell, h, z)| \leq \widetilde{C}_1(\ell, \gamma)\,\frac{1}{e^{\frac{2\pi d}{h}} - 1} + \widetilde{C}_2(\ell, \gamma)\,\frac{1}{(hm)^{\ell-1}}\,, \qquad 0 < d < \gamma\,. \tag{7.13}$$

Proof. The proof follows the ideas of the proof of Theorem 1 given in [53]. First, we set

$$g(t) := \frac{e^{\gamma+it}}{(\gamma + it)^\ell}\,\frac{1}{\gamma + it - z}\,.$$

Then we estimate

$$\begin{aligned}\left|\int_{-\infty}^{\infty} g(t)\,dt - h\sum_{k=-m}^{m} g(kh)\right| &= \left|\int_{-\infty}^{\infty} g(t)\,dt - h\sum_{k=-\infty}^{\infty} g(kh) + h\sum_{|k|\geq m+1} g(kh)\right| \\ &\leq \left|\int_{-\infty}^{\infty} g(t)\,dt - h\sum_{k=-\infty}^{\infty} g(kh)\right| + h\sum_{|k|\geq m+1} |g(kh)|\,.\end{aligned}$$

The first term satisfies the assumptions of Theorem 4.1 in [80] and can be bounded by

$$\left|\int_{-\infty}^{\infty} g(t)\,dt - h\sum_{k=-\infty}^{\infty} g(kh)\right| \leq \frac{C(g,d)}{e^{\frac{2\pi d}{h}}-1}\,, \qquad 0<d<\gamma\,.$$

A calculation of the second term gives

$$\begin{aligned}h\sum_{|k|\geq m+1} |g(kh)| = h\sum_{k\geq m+1}\big(|g(kh)|+|g(-kh)|\big) &\leq 2h\,\frac{e^{\gamma}}{\gamma}\sum_{k=m+1}^{\infty}\frac{1}{|\gamma+ihk|^{\ell}} \\ &\leq 2h\,\frac{e^{\gamma}}{\gamma}\int_{mh}^{\infty}\frac{1}{s^{\ell}}\,ds = \frac{2e^{\gamma}}{\gamma(\ell-1)}\,\frac{1}{(hm)^{\ell-1}}\,.\end{aligned}$$

With $\widetilde{C}_1(\ell,\gamma) := C(g,d)$ and $\widetilde{C}_2(\ell,\gamma) = \frac{2e^{\gamma}}{\gamma(\ell-1)}$ the proof is finished. ❑

If we now replace $z\in\mathbb{C}_0^-$ in (7.13) by a matrix $\boldsymbol{A}$ with $W(\boldsymbol{A})\subseteq\mathbb{C}_0^-$ and use Crouzeix's inequality $\|f(\boldsymbol{A})\|\leq\mathcal{C}\sup_{z\in\mathbb{C}_0^-}|f(z)|$ with $\mathcal{C}=1$ (cf. Section 5.3 in von Neumann [86]), we find

$$\begin{aligned}\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - S_m(\varphi_\ell,h,\boldsymbol{A})\boldsymbol{v}\| &\leq \sup_{z\in\mathbb{C}_0^-}|\varphi_\ell(z)-S_m(\varphi_\ell,h,z)|\|\boldsymbol{v}\| \\ &\leq \left[\widetilde{C}_1(\ell,\gamma)\,\frac{1}{e^{\frac{2\pi d}{h}}-1} + \widetilde{C}_2(\ell,\gamma)\,\frac{1}{(hm)^{\ell-1}}\right]\|\boldsymbol{v}\|\end{aligned}$$

for $0<d<\gamma$ and $\ell\geq 2$. This bound looks quite similar to the result for the rational Krylov subspace approximation in Theorem 7.7. There, we proved an error bound of order $\mathcal{O}(m^{-\ell})$, whereas the convergence rate of the fixed rational approximation here is only $\mathcal{O}(m^{-(\ell-1)})$.

7.5 Numerical experiments

In what follows, we contrast the rational Krylov subspace approximation in $\mathcal{Q}_{2m+2}(\boldsymbol{A},\boldsymbol{v})$ with the fixed rational approximation from Section 7.4 and check the predicted convergence rates of Theorem 7.8 numerically. After that, the performance of a serial and a parallel implementation of the rational Krylov subspace process with simple poles is compared with respect to computing time. At the end, we come back to the wave equation considered in Section 6.5.2.

7.5.1 Comparison with the fixed rational approximation

The first numerical experiment is meant to demonstrate the superiority of the rational Krylov subspace approximation with simple poles $z_k = \gamma + ihk$ over the fixed rational approximation discussed in Section 7.4.

In Figure 7.3, the fixed rational approximation (red dashed line) of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ is compared to the rational Krylov subspace approximation in $\mathcal{Q}_{2m+2}(\boldsymbol{A}, \boldsymbol{v})$ (blue solid line) for $\ell = 2, 4$, a random vector $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_2 = 1$, and a dense normal matrix $\boldsymbol{A} \in \mathbb{C}^{1\,000\times 1\,000}$ with eigenvalues located on the boundary of the semicircle with radius 100 and midpoint 0 in the left complex half-plane. The parameters are chosen as $\gamma = 2$ and $h = 0.5$, such that the first term on the right in (7.8) is of order $\mathcal{O}(10^{-4})$ for $\ell = 2$ and of order $\mathcal{O}(10^{-9})$ for $\ell = 4$. We see that the rational Krylov subspace method is better than the approximation via the fixed rational matrix function $S_m(\varphi_\ell, h, \boldsymbol{A})$ times $\boldsymbol{v}$.

On the one hand, this observation can be explained by the fact that, according to our analysis above, the convergence rate of the rational Krylov subspace process behaves like $\mathcal{O}(m^{-\ell})$ and, in contrast, the rate of the fixed approximation decreases only like $\mathcal{O}(m^{-(\ell-1)})$. On the other hand, a faster convergence is to be expected by the near-optimality property of Krylov subspace methods, cf. Section 4.4.

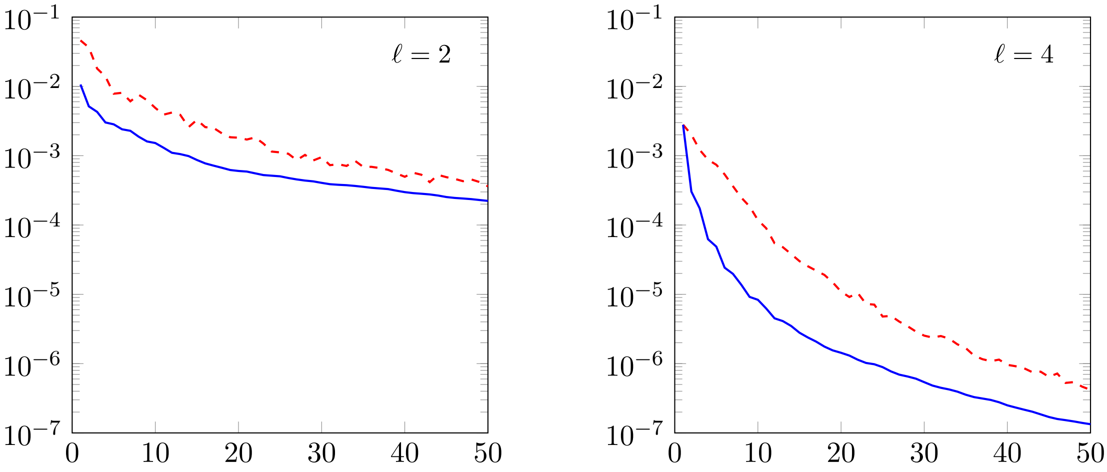


Figure 7.3: Comparison of the fixed rational approximation from Section 7.4 and the rational Krylov subspace approximation with simple poles. The errors $\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\|_2$ and $\|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - S_m(\varphi_\ell, h, \boldsymbol{A})\boldsymbol{v}\|_2$ are plotted versus m for $\gamma = 2$, $h = 0.5$, and $\ell = 2, 4$.

7.5.2 Convergence rate testing

For the same test matrix $\boldsymbol{A} \in \mathbb{C}^{1\,000\times 1\,000}$ as in the previous Section 7.5.1 and a random vector $\boldsymbol{v}$ of norm one, we check how well the predicted convergence rates match the actual approximation errors obtained in numerical experiments. By E_m we denote the error for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in the rational Krylov subspace $\mathcal{Q}_{2m+2}(\boldsymbol{A}, \boldsymbol{v})$, that is, $E_m = \|\varphi_\ell(\boldsymbol{A})\boldsymbol{v} - \varphi_\ell(\boldsymbol{A}_m)\boldsymbol{v}\|_2$, where $\boldsymbol{A}_m$ is the restriction of $\boldsymbol{A}$ to $\mathcal{Q}_{2m+2}(\boldsymbol{A}, \boldsymbol{v})$ via orthogonal projection.

For the parameter choice $\gamma = 2$ and $h = 0.5$, according to Table 7.1, we plot in Figure 7.4 on the left hand-side the values $\ln(E_m)/\ln(m)$ against the number m of iteration steps.

As expected from our error analysis, this quantity tends to $-\ell$, confirming that the convergence rate is of order $\mathcal{O}(m^{-\ell})$. Moreover, on the right-hand side of in Figure 7.4, we draw the approximation error for the choice $\gamma = 1$ and $h = \frac{\pi}{\ell} W(\frac{\pi m}{\ell})^{-1}$, as suggested in Section 7.3, together with curves of the predicted convergence order $\mathcal{O}\big(\ln(m)^\ell / m^\ell\big)$, cf. inequality (7.11).

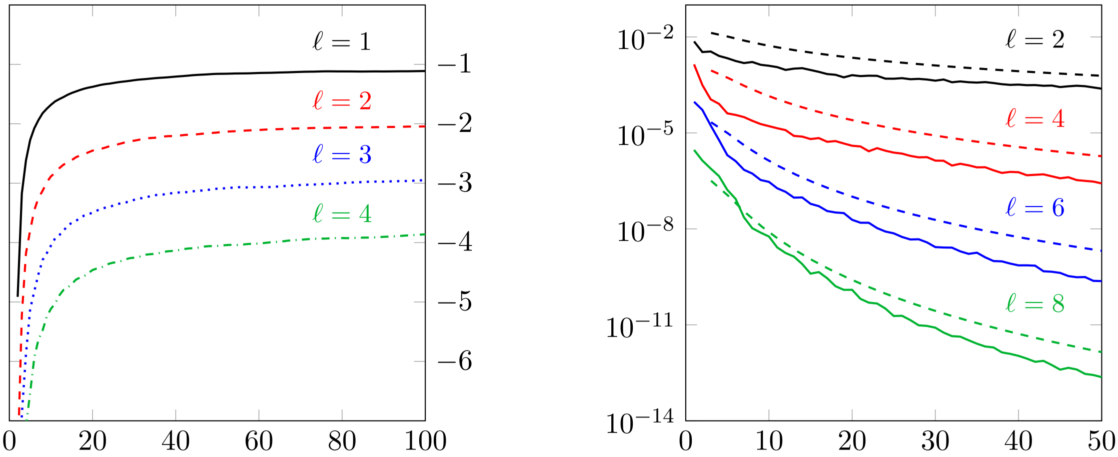


Figure 7.4: Left-hand side: Plot of $\ln(E_m)/\ln(m)$ versus m for $\gamma = 2$, $h = 0.5$. Right-hand side: Plot of the error E_m for the rational Krylov subspace method with simple poles of distance $h = \frac{\pi}{\ell} W(\frac{\pi m}{\ell})^{-1}$ and shift $\gamma = 1$ together with curves of order $\mathcal{O}\big(\ln(m)^\ell / m^\ell\big)$.

7.5.3 Parallel test example

Nowadays, there exist several variants of parallel computing with various fields of application. This technique is primary used either to treat extremely large problems, which cannot be handled on a single computer, or to save computation time by solving problems of medium or large size parallel in time. An implementation of a parallel Arnoldi method is presented, for example, by Booten, Meijer, te Riele and van der Vorst in [8]. In their algorithm, the computation of the matrix-vector products and of the inner products is performed in parallel.

Here, we focus our attention on the second case, namely, the parallelization in time: Given are different poles $z_k = \gamma + ihk$ for $k = -m, \ldots, m$ and $\gamma > 0$. This provides the advantage that the $2m + 1$ linear systems $(z_k \boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}$, involving the discretization matrix $\boldsymbol{A}$, are decoupled and, therefore, can be solved independently of each other in the rational Krylov subspace decomposition. This is why the rational Krylov subspace method with different simple poles is perfectly suited for a parallel implementation in time. In each iteration step, we solve p of the total $2m+1$ single linear systems simultaneously on p processors or kernels. In this way we achieve a tremendous speed-up which is illustrated by the following simple test example.

In the thesis by Skoogh [78], possible realizations of a parallel rational Krylov subspace algorithm are discussed in detail. His very general code, taken from Section 3.5 in [78] and adapted to our case, is shown in Algorithm 7.11. The presented parallel algorithm determines an orthonormal basis $\boldsymbol{V}_{2m+2} = [\boldsymbol{v}_1\, \boldsymbol{v}_2 \, \cdots \, \boldsymbol{v}_{2m+2}]$ of the rational Krylov subspace $\mathcal{Q}_{2m+2}(\boldsymbol{A}, \boldsymbol{v})$ with the help of a Gram-Schmidt process. In each loop, p linear systems

$\boldsymbol{w}_k = (z_k\boldsymbol{I} - \boldsymbol{A})^{-1}\widetilde{\boldsymbol{w}}_k$ are solved in parallel on p different processors (or kernels) for given starting vectors $\widetilde{\boldsymbol{w}}_k$. The resulting vectors $\boldsymbol{w}_k$ are then orthogonalized against each other and against all previously computed basis vectors $\boldsymbol{v}_i$. For the first iteration step with $j = 1$, the starting vectors $\widetilde{\boldsymbol{w}}_k$, $k = 1, \ldots, p$, are set to $\boldsymbol{v}_1$. In the next iterations of the while loop, each of the p processors can use its own orthogonalized $\widetilde{\boldsymbol{w}}_k$ of the previous step. Taking $\widetilde{\boldsymbol{w}}_k = \boldsymbol{v}_k$ instead, this would lead to a sequential algorithm, since the solution of $(z_k\boldsymbol{I} - \boldsymbol{A})\boldsymbol{w}_k = \boldsymbol{v}_k$ is only available, if $\boldsymbol{v}_k$ is known in advance. For our experiments, we simply choose $\widetilde{\boldsymbol{w}}_k = \boldsymbol{v} = \boldsymbol{v}_1$ for all k.

Algorithm 7.11 Parallel rational Krylov subspace process

```
given: $\boldsymbol{A} \in \mathbb{C}^{N\times N}$, $\boldsymbol{v} \in \mathbb{C}^N$
       set of poles $Z := \{z_k = \gamma + ihk\,,\ k = -m \ldots, m\}$
$\boldsymbol{v}_1 = \boldsymbol{v}/\|\boldsymbol{v}\|$
$j = 1$
while $j \le 2m + 1$
    choose vectors $\widetilde{\boldsymbol{w}}_k$, $k = 1, \ldots, p$
    choose poles $z_k \in Z$, $k = 1, \ldots, p$
    compute $\boldsymbol{w}_k = (z_k\boldsymbol{I} - \boldsymbol{A})^{-1}\widetilde{\boldsymbol{w}}_k$, $k = 1, \ldots, p$        (parallel step)
    for $k = 1, \ldots, p$ do
        for $i = 1, \ldots, j$ do
            $h_{i,j} = (\boldsymbol{w}_k, \boldsymbol{v}_i)$
        end for
        $\boldsymbol{w}_k = \boldsymbol{w}_k - \sum_{i=1}^{j} h_{i,j}\boldsymbol{v}_i$
        $h_{j+1,j} = \|\boldsymbol{w}_k\|$
        $\boldsymbol{v}_{j+1} = \boldsymbol{w}_k/h_{j+1,j}$
        $j = j + 1$
    end for
end while
```

Skoogh discusses several implementations of the parallel rational Krylov subspace algorithm. On the one hand, one can use an additional processor, also called master, that performs the orthogonalization of the computed vectors $\boldsymbol{w}_k$. On the other hand, it is also possible that each of the p so-called slave processors orthogonalizes its own computed vector. However, this requires the exchange of the orthogonalized vectors between the processors. More information and further details concerning, i.a., the advantages and disadvantages of the different implementations, especially the communication time, can be found in [78]. We will not go into any more detail here.

For our purposes, we use the first variant: A master handles the program control, manages the communication, and performs the orthogonalization, whereas every slave processor solves one of the occurring linear systems.

In order to demonstrate that a parallel computation of the rational Krylov subspace decomposition can essentially outperform the serial implementation, we consider a dense $1\,500 \times 1\,500$-matrix $\boldsymbol{A}$ with the field of values $W(\boldsymbol{A}) = [-1\,500, -1] \subseteq \mathbb{C}_0^-$ on the negative real axis and a random vector $\boldsymbol{v}$ of norm one. We apply a serial and a parallel version of the rational Krylov subspace method for the approximation of $\varphi_1(\tau\boldsymbol{A})\boldsymbol{v}$ and $\varphi_4(\tau\boldsymbol{A})\boldsymbol{v}$ with the parameter choice $\tau = 0.05$, $\gamma = 1$, and $h = 0.25$, in accordance with Table 7.1. The comparison is shown in Figure 7.2, where the approximation error is plotted against the computing time in seconds. The smallest error corresponds to the dimension 100 of the Krylov subspace, that is, $m = 49$ in $\mathcal{Q}_{2m+2}(\boldsymbol{A}, \boldsymbol{v})$.

The parallel version has been conducted on a local cluster of 13 heterogeneous workstation computers using MPI and the C programming language, whereas the serial process has been computed on one of these workstations. Even though the network is not a high-performance network suited for parallel computations, the parallel-in-time version of the rational Krylov subspace approximation with simple poles is enormously faster and thus more efficient for our considered test problem.

Moreover, we display in Figure 7.2 the approximation error obtained by the implicit Euler method applied to the system of ordinary differential equations

$$\boldsymbol{y}'(t) = \boldsymbol{A}\boldsymbol{y}(t) + \frac{t^{\ell-1}}{(\ell-1)!}\,\boldsymbol{v}\,, \qquad \boldsymbol{y}(0) = \boldsymbol{0}$$

with solution $\boldsymbol{y}(\tau) = \tau^\ell \varphi_\ell(\tau\boldsymbol{A})\boldsymbol{v}$. This serves as a reference for standard stiff integrators. For a reasonable comparison, the results of the implicit Euler scheme are scaled by the factor $\tau^{-\ell}$.

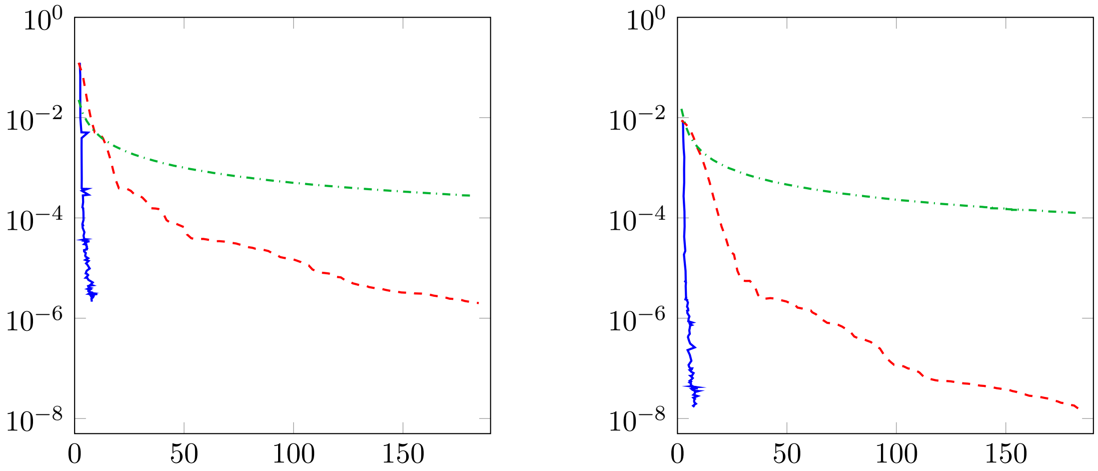


Table 7.2: Comparison of a serial (red dashed line) with a parallel (blue solid line) implementation of the rational Krylov subspace method and with the implicit Euler method (green dash-dotted line) for the approximation of $\varphi_1(\tau\boldsymbol{A})\boldsymbol{v}$ (top) and $\varphi_4(\tau\boldsymbol{A})\boldsymbol{v}$ (bottom), where $\boldsymbol{A}$ is a dense matrix with $W(\boldsymbol{A}) = [-1\,500, -1]$. The error is plotted versus computing time in seconds. The parameters are chosen as $\tau = 0.05$, $\gamma = 1$, and $h = 0.25$.

We performed the same numerical experiment up to the Krylov subspace dimension 450, requiring $\mathcal{O}(10^{12})$ floating point operations, which is equivalent to about 1 Tflop. For the approximation of $\varphi_1(\tau\boldsymbol{A})\boldsymbol{v}$ up to an accuracy of $1.044555 \cdot 10^{-9}$, the parallel version needed 33.53 seconds and the serial variant needed 837.6 seconds, which is approximately equal to 14 minutes.

Furthermore, we show in Figure 7.3 the error versus computing time for the approximation of $\varphi_4(\boldsymbol{A})\boldsymbol{v}$, where $\boldsymbol{A}$ is a dense matrix with a field of values $W(\boldsymbol{A}) = [-1\,500\,i, -i]$ on the imaginary axis. On the left-hand side, a local cluster of 13 heterogeneous workstations has been used for the computation. On the right-hand side, we computed the approximation on a single machine with 12 true kernels. Again, the parallel implementation exhibits a significant speed-up.

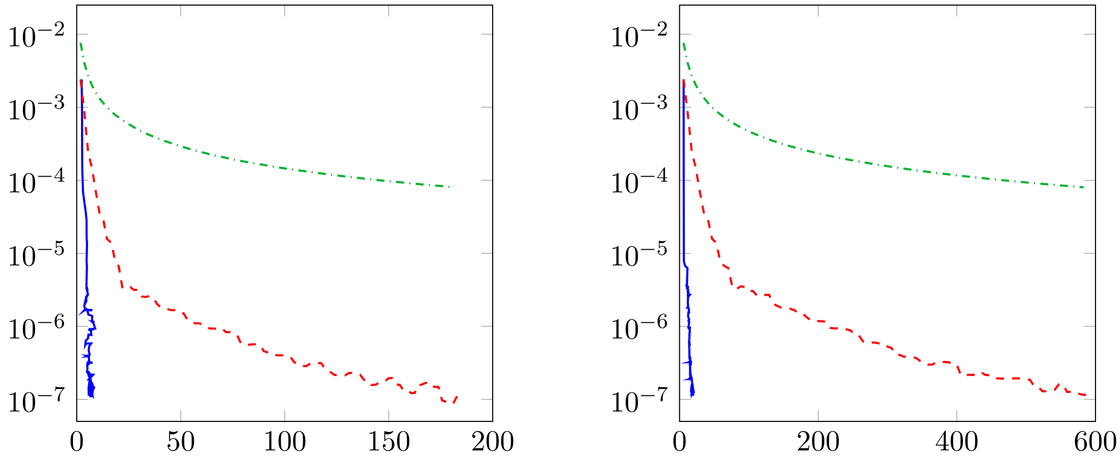


Table 7.3: Comparison of a serial (red dashed line) with a parallel (blue solid line) implementation of the rational Krylov subspace method and with the implicit Euler method (green dash-dotted line) for the approximation of $\varphi_4(\boldsymbol{A})\boldsymbol{v}$, where $\boldsymbol{A}$ is a dense matrix with $W(\boldsymbol{A}) = [-1\,500\,i, -i]$, $\gamma = 1$, and $h = 0.25$. The error is plotted versus computing time in seconds. On the left-hand side, a local cluster of 13 heterogeneous workstations has been used and, on the right-hand side, a single machine with 12 true kernels.

7.5.4 Wave equation on a non standard domain

In order to compare the shift-and-invert Krylov subspace method with the rational Krylov subspace approximation with simple poles, we return to the wave equation in Section 6.5.2,

$$\begin{aligned} &u'' = \Delta u - u && \text{for} \quad (x,y) \in \Omega\,,\ t \geq 0\,, \\ &u(0,x,y) = u_0(x,y)\,,\ u'(0,x,y) = u_0'(x,y) && \text{for} \quad (x,y) \in \Omega\,, \\ &\nabla_{\boldsymbol{n}} u = 0 && \text{for} \quad (x,y) \in \partial\Omega \end{aligned}$$

on $L^2(\Omega)$ with homogeneous Neumann boundary conditions for the non standard spatial domain shown in Figure 6.13 above. In Section 6.5.2, we have seen that for a finite-element discretization, the semi-discrete first order formulation is given as

$$\boldsymbol{y}'(t) = \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix}' = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{I} \\ \boldsymbol{M}^{-1}(\boldsymbol{S}-\boldsymbol{M}) & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}(t) \\ \boldsymbol{w}(t) \end{bmatrix} = \boldsymbol{A}\boldsymbol{y}(t)\,, \quad \boldsymbol{y}(0) = \boldsymbol{y}_0 = \begin{bmatrix} \boldsymbol{v}_0 \\ \boldsymbol{w}_0 \end{bmatrix}$$

with $(\boldsymbol{M})_{ij} = (\phi_i, \phi_j)_{L^2(\Omega)}$, $(\boldsymbol{S})_{ij} = -(\nabla\phi_i, \nabla\phi_j)_{L^2(\Omega)}$, where $\phi_k(x,y) \in H^1(\Omega)$ are the linear ansatz functions of the finite-element method. As initial value, we choose the vector $\boldsymbol{y}_0 = \boldsymbol{y}_0^1$ from Section 6.5.2.

This time, we do not approximate $e^{\tau \boldsymbol{A}}\boldsymbol{y}_0$, but $\varphi_1(\tau \boldsymbol{A})\boldsymbol{A}\boldsymbol{y}_0$ in the alternative representation $e^{\tau \boldsymbol{A}}\boldsymbol{y}_0 = \tau\varphi_1(\tau \boldsymbol{A})\boldsymbol{A}\boldsymbol{y}_0 + \boldsymbol{y}_0$. We compute approximations in the rational Krylov subspace $\mathcal{Q}_{2m+2}(\tau \boldsymbol{A}, \boldsymbol{A}\boldsymbol{y}_0)$ with simple poles of distance $h = 1$ and in the shift-and-invert Krylov subspace $\mathcal{K}_{2m+2}((\gamma \boldsymbol{I} - \tau \boldsymbol{A})^{-1}, \boldsymbol{A}\boldsymbol{y}_0)$. In both cases, we take $\gamma = 3$. For this choice of parameters, the fixed term in the error bound (7.4), which does not depend on m, is of size $\mathcal{O}(10^{-3})$.

A comparison of the two methods is shown in Figure 7.5 for two different choices of τ and numbers of triangles of the finite-element mesh. The blue solid error curve refers to the rational Krylov subspace with simple poles and the red dashed error curve to the Krylov space with one single pole. The error $\|\varphi_1(\tau \boldsymbol{A})\boldsymbol{A}\boldsymbol{y}_0 - \varphi_1(\tau \boldsymbol{A}_m)\boldsymbol{A}\boldsymbol{y}_0\|_{\boldsymbol{B}}$ is plotted against m, where $\boldsymbol{A}_m$ is the restriction of $\boldsymbol{A}$ to $\mathcal{Q}_{2m+2}(\tau \boldsymbol{A}, \boldsymbol{A}\boldsymbol{y}_0)$ or $\mathcal{K}_{2m+2}((\gamma \boldsymbol{I} - \tau \boldsymbol{A})^{-1}, \boldsymbol{A}\boldsymbol{y}_0)$, respectively. The error is measured with respect to $\|\cdot\|_{\boldsymbol{B}}$, where $\boldsymbol{B}$ was given as the block diagonal matrix $\boldsymbol{B} = \operatorname{diag}(-(\boldsymbol{S} - \boldsymbol{M}), \boldsymbol{M})$.

For the solution of the shifted linear systems $\big((\gamma + ihk)^2\boldsymbol{M} - \tau^2(\boldsymbol{S} - \boldsymbol{M})\big)\boldsymbol{x} = \boldsymbol{b}$, cf. (6.10) above, we can no longer use a multigrid method with Gauss-Seidel smoother. For complex shifts $\gamma + ihk$, the matrix here is not real symmetric and positive definite like the matrix $\gamma^2\boldsymbol{M} - \tau^2(\boldsymbol{S} - \boldsymbol{M})$ occurring in the shift-and-invert Krylov subspace approximation. Instead of the Gauss-Seidel iteration, we apply the Quasi Minimal Residual method (cf. [22]) for the smoothing process in the multigrid method as suggested in [79], Section 5.15.

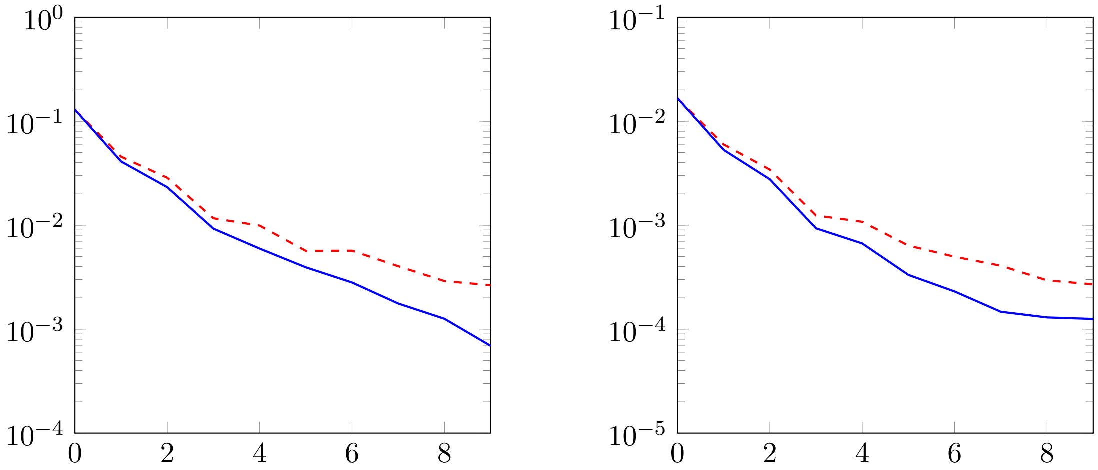


Figure 7.5: Plot of the error $\|\varphi_1(\tau \boldsymbol{A})\boldsymbol{A}\boldsymbol{y}_0 - \varphi_1(\tau \boldsymbol{A}_m)\boldsymbol{A}\boldsymbol{y}_0\|_{\boldsymbol{B}}$ versus m for the rational Krylov subspaces $\mathcal{Q}_{2m+2}(\tau \boldsymbol{A}, \boldsymbol{A}\boldsymbol{y}_0)$ (blue solid line) and $\mathcal{K}_{2m+2}((\gamma \boldsymbol{I} - \tau \boldsymbol{A})^{-1}, \boldsymbol{A}\boldsymbol{y}_0)$ for $\tau = 0.05$ and a mesh with 18 816 triangles, 10 057 nodes (left), $\tau = 0.01$ and a mesh with 301 056 triangles, 153 121 nodes (right), and parameters $\gamma = 3$, $h = 1$.

Chapter 8

Conclusion and outlook

The goal of this thesis was to study the convergence behavior of rational Krylov subspace methods for the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$. These φ-functions play a fundamental role in the application of exponential integrators for the time integration of evolution equations. Their efficient and reliable computation is presently of great interest and a subject of current research. For large stiff matrices $\boldsymbol{A}$, that arise from the discretization of some unbounded differential operator and have a field of values in the left complex half-plane, rational Krylov methods significantly outperform the well-established standard Krylov subspace iteration and other standard methods for stiff problems such as the implicit Euler or the Crank-Nicolson scheme.

We have analyzed the approximation of $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ in a rational Krylov subspace with one single repeated pole $\gamma > 0$ and with equidistant simple poles $\gamma + ihk$, $k = -m, \dots, m$, on the line $\mathrm{Re}(z) = \gamma$. For stiff matrices $\boldsymbol{A}$ with $W(\boldsymbol{A}) \subseteq \mathbb{C}_0^-$, sublinear convergence rates could be shown which do not depend on the norm of the discretization matrix $\boldsymbol{A}$. As a consequence, it is guaranteed that these rational Krylov subspace methods converge independent of the refinement of the spatial mesh. In contrast, the standard Krylov subspace approximation has error bounds that always involve $\|\boldsymbol{A}\|$, so that this method is not suited for our purposes.

However, if the initial value is smooth enough, it became apparent that it is often advantageous to first perform some cheaper standard Krylov steps and then to continue with the more efficient but usually more expensive rational Krylov subspace process. This has led us to study extended Krylov subspace methods. They represent a skillful combination of the polynomial and the rational Krylov subspace iteration which results in a faster grid-independent convergence, provided that the vector $\boldsymbol{v}$ satisfies certain smoothness properties.

Since the obtained error bounds for rational and extended Krylov subspace methods hold uniformly over all possible grids in space, the presented techniques constitute a promising approximation method for the φ-functions evaluated at large discretization matrices $\boldsymbol{A}$. In order to achieve an approximation to $\varphi_\ell(\boldsymbol{A})\boldsymbol{v}$ that is as optimal as possible, we also discussed suitable choices for the shift γ and the free parameter h, which determines the distance of the equidistant poles for the rational Krylov subspace process with different simple poles.

In the next step, it would be interesting to examine more precisely the influence of the smoothness of the initial data to the rational Krylov subspace approximation. The observations in our numerical experiments suggest that not only the convergence rate of the extended method but also the rate of the rational Krylov subspace approximation is affected by the smoothness properties of the initial vector.

In the case that the field of values of the matrix $\boldsymbol{A}$ has a special geometry, that is, $W(\boldsymbol{A})$ is, for example, a subset of the negative real line or lies in a sector in the left complex half-plane, better convergence rate are observed than predicted by our estimates. The next important step could therefore be to improve the error bounds depending on the opening angle of the sector, in which the field of values is located. Such matrices with a field of values in some sector arise from the discretization of parabolic problems involving sectorial operators.

Bibliography

[1] N. Achyèser. Über den Jacksonschen Approximationssatz. *Communications de la Société Mathématique de Kharkow*, 8(4):3–12, 1934.

[2] A. H. Al-Mohy and N. J. Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM Journal on Scientific Computing*, 33(2):488–511, 2011.

[3] B. Beckermann and S. Güttel. Superlinear convergence of the rational Arnoldi method for the approximation of matrix functions. *Numerische Mathematik*, 121(2):205–236, 2012.

[4] B. Beckermann and L. Reichel. Error estimates and evaluation of matrix functions via the Faber transform. *SIAM Journal on Numerical Analysis*, 47(5):3849–3883, 2009.

[5] D. Berend and T. Tassa. Improved bounds on bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30(2):185–205, 2010.

[6] L. Bergamaschi and M. Vianello. Efficient computation of the exponential operator for large, sparse, symmetric matrices. *Numerical Linear Algebra with Applications*, 7(1):27–45, 2000.

[7] M. C. De Bonis, G. Mastroianni, and M. Viggiano. K-functionals, moduli of smoothness and weighted best approximation on the semiaxis. In *Functions, series, operators (Budapest, 1999)*, pages 181–211. János Bolyai Math. Soc., Budapest, 2002.

[8] J. G. L. Booten, P. M. Meijer, H. J. J. te Riele, and H. A. van der Vorst. Parallel Arnoldi method for the construction of a Krylov subspace basis: an application in magnetohydrodynamics. In *HPCN*, Lecture Notes in Computer Science, pages 196–201. Springer, 1994.

[9] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.* Springer, Berlin, 4., überarbeitete und erweiterte Auflage, 2007.

[10] P. Brenner and V. Thomée. On rational approximations of semigroups. *SIAM Journal on Numerical Analysis*, 16(4):683–694, 1979.

[11] J. Certaine. The solution of ordinary differential equations with large time constants. In *Mathematical methods for digital computers*, pages 128–132. Wiley, New York, 1960.

[12] M. Crouzeix. Numerical range and functional calculus in Hilbert space. *Journal of Functional Analysis*, 244(2):668–690, 2007.

[13] C. de la Vallée Poussin. *Leçons sur l'approximation des fonctions d'une variable réelle: Professées à la Sorbonne.* Collection de monographies sur la théorie des fonctions. Gauthier-Villars, Paris, 1919.

[14] R. A. DeVore and G. G. Lorentz. *Constructive approximation.* Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen. Springer, Berlin, 1993.

[15] V. Druskin and L. Knizhnerman. Extended Krylov subspaces: approximation of the matrix square root and related functions. *SIAM Journal on Matrix Analysis and Applications*, 19(3):755–771, 1998.

[16] V. L. Druskin and L. A. Knizhnerman. Error bounds in the simple Lanczos procedure for computing functions of symmetric matrices and eigenvalues. *Computational Mathematics and Mathematical Physics*, 31:20–30, 1991.

[17] N. Dunford and J. T. Schwartz. *Linear operators, Part I.* Wiley classics library. Wiley Interscience Publ., Hoboken, New Jersey, 1988.

[18] H. Emamirad and A. Rougirel. A functional calculus approach for the rational approximation with nonuniform partitions. *Discrete and Continuous Dynamical Systems*, 22(4):955–972, 2008.

[19] K.-J. Engel and R. Nagel. *A Short Course on Operator Semigroups.* Universitext. Springer, 2006.

[20] T. Ericsson and A. Ruhe. The spectral transformation Lánczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems. *Mathematics of Computation*, 35(152):1251–1268, 1980.

[21] D. J. Estep, M. G. Larson, and R. D. Williams. *Estimating the error of numerical solutions of systems of reaction diffusion equations.* Memoirs of the American Mathematical Society. American Mathematical Society, Providence, R.I., 2000.

[22] R. W. Freund and N. M. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numerische Mathematik*, 60(3):315–339, 1991.

[23] E. Gallopoulos and Y. Saad. Efficient solution of parabolic equations by Krylov approximation methods. *SIAM Journal on Scientific and Statistical Computing*, 13(5):1236–1264, 1992.

[24] T. Göckler and V. Grimm. Uniform approximation of φ-functions in exponential integrators by a rational Krylov subspace method with simple poles. To appear in SIAM Journal on Matrix Analysis and Applications, 2014.

[25] T. Göckler and V. Grimm. Convergence analysis of an extended Krylov subspace method for the approximation of operator functions in exponential integrators. *SIAM Journal on Numerical Analysis*, 51(4):2189–2213, 2013.

[26] G. H. Golub and C. F. Van Loan. *Matrix computations.* Johns Hopkins studies in the mathematical sciences. Johns Hopkins University Pr., Baltimore, Md., 4th edition, 2013.

[27] A. Greenbaum. *Iterative methods for solving linear systems.* Frontiers in applied mathematics. SIAM, Philadelphia, 1997.

[28] R. E. Greene and S. G. Krantz. *Function Theory of One Complex Variable.* Graduate Studies in Mathematics. American Mathematical Society, 3rd edition, 2006.

[29] V. Grimm. Resolvent Krylov subspace approximation to operator functions. *BIT Numerical Mathematics*, 52(3):639–659, 2012.

[30] S. Güttel. Rational Krylov methods for operator functions, PhD thesis, TU Bergakademie Freiberg, Germany, 2010.

[31] S. Güttel. Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection. *GAMM Mitteilungen*, 36(1):8–31, 2013.

[32] S. Güttel and L. Knizhnerman. A black-box rational Arnoldi variant for Cauchy-Stieltjes matrix functions. *BIT Numerical Mathematics*, 53(3):595–616, 2013.

[33] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II : Stiff and Differential-Algebraic Problems.* Springer, Berlin, Heidelberg, 2nd revised edition, 2010.

[34] N. Hale, N. J. Higham, and L. N. Trefethen. Computing A^α, $\log(A)$, and related matrix functions by contour integrals. *SIAM Journal on Numerical Analysis*, 46(5):2505–2523, 2008.

[35] M. Hassani. Approximation of the Lambert W Function. *RGMIA Research Report Collection*, 8(4), 2005.

[36] D. J. Higham and L. N. Trefethen. Stiffness of ODEs. *BIT Numerical Mathematics*, 33(2):285–303, 1993.

[37] N. J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193, 2005.

[38] N. J. Higham. *Functions of matrices: theory and computation.* Society for Industrial and Applied Mathematics, Philadelphia, 2008.

[39] N. J. Higham and A. H. Al-Mohy. Computing matrix functions. *Acta Numerica*, 19:159–208, 2010.

[40] E. Hille and R. S. Phillips. *Functional analysis and semi-groups.* Colloquium publications. American Mathematical Society, Providence, R.I., revised edition, 1957.

[41] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 34(5):1911–1925, 1997.

[42] M. Hochbruck and A. Ostermann. Explicit exponential Runge-Kutta methods for semilinear parabolic problems. *SIAM Journal on Numerical Analysis*, 43(3):1069–1090, 2005.

[43] M. Hochbruck and A. Ostermann. Exponential Runge-Kutta methods for parabolic problems. *Applied Numerical Mathematics*, 53(2–4):323–339, 2005.

[44] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 5 2010.

[45] M. Hochbruck and A. Ostermann. Exponential multistep methods of Adams-type. *BIT Numerical Mathematics*, 51(4):889–908, 2011.

[46] M. Hochbruck, T. Pažur, A. Schulz, E. Thawinan, and C. Wieners. Efficient time integration for discontinuous Galerkin approximations of linear wave equations. To appear in Journal of Applied Mathematics and Mechanics, 2014.

[47] R. A. Horn and C. R. Johnson. *Topics in matrix analysis.* University Press, Cambridge [u.a.], 1st publ. edition, 1991.

[48] L. Knizhnerman and V. Simoncini. A new investigation of the extended Krylov subspace method for matrix function evaluations. *Numerical Linear Algebra with Applications*, 17(4):615–638, 2010.

[49] A. N. Krylov. On the numerical solution of the equation by which, in technical matters, frequencies of small oscillations of material systems are determined. *IZV. Akad. Nauk SSSR*, 7(4):491–539, 1931. in Russian.

[50] J. D. Lawson. Generalized Runge-Kutta processes for stable systems with large Lipschitz constants. *SIAM Journal on Numerical Analysis*, 4:372–380, 1967.

[51] R. J. LeVeque. *Finite difference methods for ordinary and partial differential equations : steady-state and time-dependent problems.* Society for Industrial and Applied Mathematics, Philadelphia, 2007.

[52] L. Lopez and V. Simoncini. Analysis of projection methods for rational function approximation to the matrix exponential. *SIAM Journal on Numerical Analysis*, 44(2):613–635, 2006.

[53] M. López-Fernández and C. Palencia. On the numerical inversion of the Laplace transform of certain holomorphic mappings. *Applied Numerical Mathematics*, 51(2-3), 2004.

[54] G. G. Lorentz. *Approximation of functions.* American Mathematical Society. AMS Chelsea Publishing, Providence, R.I., 1986.

[55] S. H. Lui. *Numerical analysis of partial differential equations.* Pure and applied mathematics: a Wiley series of texts, monographs, and tracts. Wiley, Hoboken, New Jersey, 2011.

[56] C. R. MacCluer. *Boundary Value problems and Fourier Expansions.* Dover Publications, Mineola, New York, revised edition, 2004.

[57] M. Miklavčič. *Applied Functional Analysis and Partial Differential Equations.* World Scientific Publishing Co. Pte. Ltd., Singapore, 1998.

[58] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[59] I. Moret. On RD-rational Krylov approximations to the core-functions of exponential integrators. *Numerical Linear Algebra with Applications*, 14(5):445–457, 2007.

[60] I. Moret and P. Novati. RD-rational approximations of the matrix exponential. *BIT Numerical Mathematics*, 44(3):595–615, 2004.

[61] S. P. Nørsett. An *A*-stable modification of the Adams-Bashforth methods. In *Conf. on Numerical Solution of Differential Equations (Dundee, 1969)*, pages 214–219. Springer, Berlin, 1969.

[62] P. Novati. Using the restricted-denominator rational Arnoldi method for exponential integrators. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1537–1558, 2011.

[63] B. N. Parlett. A new look at the Lanczos algorithm for solving symmetric systems of linear equations. *Linear Algebra and its Applications*, 29:323–346, 1980.

[64] B. N. Parlett. *The symmetric eigenvalue problem.* Prentice-Hall series in computational mathematics. Prentice-Hall, Englewood Cliffs, NJ, 1980.

[65] A. Pazy. *Semigroups of linear operators and applications to partial differential equations.* Applied mathematical sciences. Springer, New York, corrected 2nd printing, 1992.

[66] M. A. Pinsky. *Introduction to fourier analysis and wavelets.* Graduate studies in mathematics. American Mathematical Society, Providence, R.I, 2009.

[67] D. A. Pope. An exponential method of numerical integration of ordinary differential equations. *Communications of the ACM*, 6:491–493, 1963.

[68] H. M. Protter and C. B. Morrey. *A first course in real analysis.* Undergraduate texts in mathematics. Springer, New York, 2nd edition, 1991.

[69] A. Ruhe. Rational Krylov sequence methods for eigenvalue computation. *Linear Algebra and its Applications*, 58:391–405, 1984.

[70] A. Ruhe. The rational Krylov algorithm for nonsymmetric eigenvalue problems. III. Complex shifts for real matrices. *BIT Numerical Mathematics*, 34(1):165–176, 1994.

[71] Y. Saad. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of Computation*, 37(155):105–126, 1981.

[72] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 29(1):209–228, 1992.

[73] Y. Saad. *Iterative methods for sparse linear systems.* SIAM, Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition, 2003.

[74] M. Schechter. *Principles of Functional Analysis.* Graduate Studies in Mathematics. American Mathematical Society, 2nd edition, 2001.

[75] T. Schmelzer and L. N. Trefethen. Evaluating matrix functions for exponential integrators via Carathéodory-Fejér approximation and contour integrals. *Electronic Transactions on Numerical Analysis*, 29:1–18, 2007/08.

[76] K. Schmüdgen. *Unbounded self-adjoint operators on Hilbert space.* Graduate texts in mathematics. Springer, Dordrecht, 2012.

[77] R. B. Sidje. Expokit: A software package for computing matrix exponentials. *ACM Transactions on Mathematical Software*, 24(1):130–156, 1998.

[78] D. Skoogh. An implementation of a parallel rational Krylov algorithm, PhD thesis, Göteborg University and Chalmers University of Technology, Sweden, 1996.

[79] Y. Sphira. *Matrix-Based Multigrid: Theory and Applications.* Numerical Methods and Algorithms. Springer US, Boston, MA, 2nd edition, 2008.

[80] F. Stenger. Approximation via Whittaker's cardinal function. *Journal of Approximation Theory*, 17:222–240, 1976.

[81] D. E. Stewart and T. S. Leyk. Error estimates for Krylov subspace approximations of matrix exponentials. *Journal of Computational and Applied Mathematics*, 72(2):359–369, 1996.

[82] G. W. Stewart. *Afternotes goes to graduate school: lectures on advanced numerical analysis.* SIAM, Philadelphia, 1998.

[83] L. N. Trefethen, J. A. C. Weideman, and T. Schmelzer. Talbot quadratures and rational approximations. *BIT Numerical Mathematics*, 46(3):653–670, 2006.

[84] J. van den Eshof and M. Hochbruck. Preconditioning Lanczos approximations to the matrix exponential. *SIAM Journal on Scientific Computing*, 27(4):1438–1457, 2006.

[85] H. A. van der Vorst. An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A. *Journal of Computational and Applied Mathematics*, 18(2):249–263, 1987.

[86] J. von Neumann. Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. *Mathematische Nachrichten 4*, pages 258–281, 1951.

[87] J. L. Walsh. *Interpolation and approximation by rational functions in the complex domain.* Colloquium publications. American Mathematical Society, Providence, R.I., 5th edition, 1969.