TECHNISCHE UNIVERSITÄT WIEN

DISSERTATION

Model Order Reduction for Fractional Diffusion Problems

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

Prof. Dr. Joachim Schöberl
E101 – Institut für Analysis und Scientific Computing, TU Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Dipl. Ing. Tobias Danczul
Matrikelnummer: 01325217
[redacted]

Diese Dissertation haben begutachtet:

1. **Prof. Dr. Joachim Schöberl**
   Institut für Analysis und Scientific Computing, TU Wien
2. **Prof. Dr. Andrea Bonito**
   Department of Mathematics, Texas A&M University, College Station, TX 77845, USA.
3. **Prof. Dr. Gianluigi Rozza**
   SISSA mathLab, Mathematics Area, International School for Advanced Studies, Trieste, Italy

Wien, am 25. Oktober, 2021

Kurzfassung

In dieser Arbeit präsentieren wir eine universelle Methode zur Approximation von elliptischen und zeitabhängigen fraktionalen partiellen Differentialgleichungen. Ausgehend von einer finiten Elemente Diskretisierung wird der gewünschte Differentialoperator durch eine Matrix-Approximation $\mathbf{L}$ ersetzt. Dies erlaubt es uns, die diskrete Lösung als Matrix-Vektor Produkt der Form $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ zu interpretieren, wobei $\mathbf{b}$ ein Vektor und $f^{\boldsymbol{\tau}}$ eine parameterabhängige Funktion ist, wie z.B. die Potenzfunktion oder die Mittag-Leffler Funktion. Um den Rechenaufwand zu reduzieren, wird eine zusätzliche Approximationsebene in der Form einer rationalen Krylov Methode etabliert. Letztere projeziert die Matrix auf einen Unterraum von niedriger Dimension, der es erlaubt, das zugehörige Eigensystem direkt zu berechnen. Die Wahl des Unterraums hängt von einer Reihe unterschiedlicher Parameter ab, den sogenannten Polen. Ausgehend von dem dritten Zolotarëv Problem präsentieren wir ein breites Spektrum attraktiver Pol-Konfigurationen, die es erlauben, die Abbildung $\boldsymbol{\tau} \mapsto f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ für mehrfache Instanzen des Parameters gleichzeitig zu evaluieren. Wir beweisen exponentielle Konvergenz und stellen ein Fehlerzertifikat zur Qualitätssicherung zahlreicher Approximationen bereit, für die keine analytischen Ergebnisse vorhanden sind. Das Herzstück dieser Arbeit sind die sogenannten Zolotarëv Pole, die es ermöglichen, $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ gleichmäßig in $\boldsymbol{\tau}$ anzunähern, ohne dass die Approximation degeneriert, wenn beispielsweise die fraktionalen Parameter gegen eine ganze Zahl konvergieren.

Wir stellen die präsentierten Methoden in Verbindung mit ausgewählten Algorithmen aus der Literatur und zeigen, dass letztere als rationale Krylov Methoden interpretiert werden können. Diese theoretischen Einblicke erlauben es, unsere Resultate für neue Konvergenzbeweise heranzuziehen. Sie suggerieren die Implementierung neuer und die Verbesserung existenter Methoden und ermöglichen einen direkten Vergleich der Algorithmen. Unsere analytischen Erkenntnisse werden mit numerischen Experimenten untermauert. Wir führen einen systematischen Vergleich der erwähnten Methoden durch und präsentieren eine detaillierte Parameterstudie, die uns tiefe Einblicke in die Auswirkungen dieser Größen auf die Approximationseigenschaften von Lösungen fraktionaler Differentialgleichungen gewährleisten.

Abstract

In this thesis we present a unified framework to efficiently approximate solutions to fractional diffusion problems of elliptic and parabolic type. After finite element discretization, we take the point of view that the solution is obtained by a matrix-vector product of the form $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$, where $\mathbf{L}$ is the discretization matrix of the spatial operator, $\mathbf{b}$ a prescribed vector, and $f^{\boldsymbol{\tau}}$ a parametric function, such as a fractional power or the Mittag-Leffler function. To alleviate the computational expenses, a model order reduction strategy in the form of a rational Krylov method is applied which projects the matrix to a low-dimensional space where a direct evaluation of the eigensystem is feasible. The particular choice of the subspace depends on a collection of parameters, the so-called *poles*. On the basis of the third Zolotarëv problem, we propose a variety of attractive pole selection strategies which allow us to efficiently query the solution map $\boldsymbol{\tau} \mapsto f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ for multiple instances of the parameter. We either prove exponential convergence rates or provide the description of a computable error certificate to assess the quality of several poles where no analytical results are available. At the core of our exposition are the so-called Zolotarëv poles, which allow us to approximate $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ uniformly in the parameter $\boldsymbol{\tau}$ and do not degenerate as e.g., the fractional parameters approach an integer.

The proposed methods are set in correspondence with existing schemes from the fractional diffusion community. In particular, we prove that a large class of model order reduction strategies can be interpreted as rational Krylov method. These theoretical insights allow us to leverage our analysis to develop new convergence proofs for several of the studied schemes. They suggests how to design novel and improve available methods and allow for a direct comparison of the algorithms. The analytical findings are confirmed by numerical experiments, including a systematic comparison of the presented schemes and a parameter study which provides deep insights in the effect of the fractional parameters.

Danksagung

First and foremost I am deeply grateful to my supervisor Joachim Schöberl for his guidance and support from the very beginning of my academic career. Starting with my first Analysis course in 2013, he has been the driving force in the continuous expansion of my scientific interest across all engineering disciplines. His invaluable expertise as well as his trust in me are the reason I have come to this point which I would not have considered possible at the beginning of my studies. Secondly, I would like to offer my thanks to Clemens Hofreither for the fruitful collaboration throughout the last year. Our discussions encouraged me to consider the problems I have been facing at that time from a different perspective which significantly contributed to the progress of this thesis. Furthermore, I want to thank Prof. Andrea Bonito and Prof. Gianluigi Rozza for reviewing this thesis.

Special thanks go to my former fellow students who made my studies at TU Wien to one of the most enjoyable periods of my life. The mathematical and moral support I received from them is embracing. I would never want to miss the countless hours of joint exercise preparations or, in times where the level of motivation was not that high, the equally long nights at our favorite pub. No less grateful am I for my non-mathematician friends for bearing my bad math jokes, which have significantly increased since the beginning of my studies, and for helping me keep a healthy math-life balance.

I would also like to express my gratitude to my colleagues in the working group. They have always been the first ones to get in touch with for all arising problems and provided invaluable support for my life as a PhD student: Mathematically, personally, and culinary, in the form of several delicious birthday cakes. In particular, I want to thank my office mate and friend Michael Neunteufel for proof-reading this thesis and his service as "rubber duck" when listening to and answering many of the problems I have been facing throughout.

I am grateful for having such a wonderful family, including my parents Beate and Stefan, and my brothers Philipp and Lucas. Their everlasting support has always brought out the best in me even when things did not go as planed. Most importantly, I want to thank my girlfriend Lisa, who I have considered as part of this family for more than five years now. Ever since, you have been the anchor in my life. Your patience and faith in me has never ceased to amaze me and I happily look forward to our common future.

At last, I greatly acknowledge the support of the TU Wien, the Vienna School of Mathematics (VSM), and especially the financial and personal support of the Doctoral School "Dissipation and dispersion in nonlinear PDEs" through the Austrian Science Fund (FWF) with grant number F 65 and W1245.

Eidesstattliche Erklärung

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

Wien, am 25. Oktober, 2021

Tobias Danczul

Contents

List of Symbols

Notation	Description	Page List
$\mathcal{A}$	automatic poles	177
$\mathbf{A}$	stiffness matrix	81
arg	argument function	10
$\mathcal{B}_{\boldsymbol{\tau}}$	BURA poles	170
$B_\varepsilon(\mathbf{x})$	open Ball centered at $\mathbf{x}$ with radius r	10
$\mathcal{CB}$	complete Bernstein functions	109
$\lceil\cdot\rceil$	ceiling function	68
$\overline{\mathbb{C}}$	extended complex plane	10
$(\mathbb{A},\mathbb{B})$	condenser	129
$\mathrm{cap}(\mathbb{A},\mathbb{B})$	condenser capacity	130
$\frac{\partial \mathcal{U}}{\partial \mathbf{n}_s}$	conormal derivative	56
$(z_1,z_2;z_3,z_4)$	cross-ratio	139
$\mathcal{CS}$	Cauchy-Stieltjes functions	105
deg	polynomial degree	93
$\delta_{[\lambda_{\min},\lambda_{\max}]}$	spectral parameter	141
${}_R\partial_t^\alpha$	Riemann-Liouville fractional derivative	71
∂_t^α	(Caputo) fractional derivative	68
dist	distance function	10
$\langle f,v\rangle$	dual pairing	11
$\mathcal{E}$	EDS poles on $-\Sigma$	153
$\hat{\mathcal{E}}$	EDS poles on $\mathbb{R}_0^-$	154
$\lambda_{\max}$	largest eigenvalue of $\mathbf{L}$	92
$\lambda_{\min}$	smallest eigenvalue of $\mathbf{L}$	92
E_α	Mittag-Leffler function	22
$E_{\alpha,\beta}$	generalized Mittag-Leffler function	23
$\mathrm{e}_{\alpha,\beta}$	extended generalized Mittag-Leffler function	114
$\mathrm{ext}(\mathcal{C})$	exterior of the contour $\mathcal{C}$	27

Notation	Description	Page List
$\mathcal{F}$	fully automatic poles	179
$\lfloor \cdot \rfloor$	floor function	70
$f^{\boldsymbol{\tau}}$	parametric function	82
Γ	Gamma function	21
$\mathcal{G}$	weak greedy poles on $-\Sigma$	165
$\hat{\mathcal{G}}$	weak greedy poles on $\mathbb{R}_0^-$	165
$\mathbb{H}_{\Im<0}$	lower half plane	107
$\mathbb{H}_{\Im>0}$	upper half plane	107
$\mathcal{H}$	Hilbert space	11
$\mathbf{I}$	identity	10
$\mathbf{I}_{k+1}$	$(k+1) \times (k+1)$ identity matrix	94
$\Im$	imaginary part	10
$\mathrm{int}(\mathcal{C})$	interior of the contour $\mathcal{C}$	27
$(\cdot,\cdot)_{H^s_{\mathcal{L}}(\Omega)}$	associated interpolation scalar product	50
J^α	Riemann-Liouville integral	66
$\mathbf{L}$	matrix approximation of $\mathcal{L}$	82
$\|\cdot\|$	discrete L^2-norm	96
$(\cdot,\cdot)$	discrete L^2-scalar product	96
$\mathcal{L}$	diffusion operator	48
$\mathcal{L}^s$	fractional diffusion operator	50
$\mathscr{L}$	Laplace transform	19
$\mathbf{L}_{k+1}$	compression	95
$\mathcal{LS}$	Lapalce-Stieltjes functions	111
$\mathbf{M}$	mass matrix	81
$\|f\|_E$	maximum norm on E	97
$\mathrm{n}_\mathrm{c}^-, \mathrm{n}_\mathrm{c}^+$	cut-off parameters	157
Ξ	pole set $\Xi = \{\xi_0, \ldots, \xi_k\}$	92
$\mathcal{P}_1^0(\mathcal{T}_h)$	Lagrangian finite element space	81
$\mathcal{P}_N$	polynomials of maximum degree N	92
q_Ξ	monic polynomial with poles in Ξ	92

Notation	Description	Page List
q	sinc parameter	86
$r_{\Lambda,\Xi}$	rational function with roots in Λ and poles in Ξ	116
r_{Ξ}	rational function with roots in $-\Xi$ and poles in Ξ	119
$r_k^{\mathcal{B}_\tau}$	BURA	169
$\Re$	real part	10
$\rho_{[a,b]}$	convergence parameter	140
$\overline{\mathbb{R}}$	extended real line	10
$\mu_j^{(k)}$	rational Ritz value	96
$\mathcal{S}$	spectral poles on $-\Sigma$	158
$\hat{\mathcal{S}}$	spectral poles on $\mathbb{R}_0^-$	157
Σ	spectral interval of $\mathbf{L}$	92
supp	support	10
Θ_{C}	Cauchy-Stieltjes parameter set	114
Θ_{L}	Lapalce-Stieltjes parameter set	114
$T_{[\lambda_{\min},\lambda_{\max}]}$	Möbius transformation	141
$\mathcal{T}_{\mathrm{train}}$	train set	158
$\mathcal{T}_{\mathrm{train}}^{\mathrm{n_c^{\pm}}}$	train set with cut-off parameters	157
$\mathbf{U}$	matrix of eigenvectors of $\mathbf{L}$	83
$\mathcal{U}$	s-minimal extension	42
$\mathbf{u}_{k+1}$	rational Krylov approximation	94
$\mathbf{V}$	basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$	94
$\mathbf{V}^\dagger$	Moore-Penrose pseudo inverse	94
$\mathcal{V}_{k+1}^{Z}(\mathbf{L},\mathbf{b})$	reduced basis space	190
$\mathbf{w}(\zeta)$	parametric solution	190
$\mathbf{w}_{k+1}(\zeta)$	reduced basis approximation	190
Z	snapshot set $Z=\{\zeta_0,\ldots,\zeta_k\}$	190
$Z_k(\mathbb{A},\mathbb{B})$	Zolotarëv number	125
$\mathcal{Z}$	Zolotarëv poles on $-\Sigma$	146
$\hat{\mathcal{Z}}$	Zolotarëv poles on $\mathbb{R}_0^-$	146
$\mathcal{Z}_j^{(k)}$	Zolotarëv point	135

1 Introduction

Partial differential equations (PDEs) have been an outstanding tool in modern science to describe real-world phenomena across all engineering disciplines. Throughout the past twenty years, however, practitioners have been facing experimental setups which seemingly elude the reach of classical calculus. Standard PDEs lack the ability to adequately model nonlocal effects, in which two points at finite distance can interact. Fractional partial differential equations (FPDEs) have proven themselves as gateway to provide refined models which describe these physical processes more accurately. Their theoretical groundings as well as their confirmation in scientific experiments have sparked a remarkable amount of recent investigations, ranging from biomedicine [BOKG+14, YPK16, FKR+21, CGGG21] to image processing [GO09, GH15, AB17, AR19], and material science [Bat06, GRN+17, SGF20, FZ20]. A paradigm of a nonlocal operator that shall serve us as prototype throughout this thesis is the fractional Laplacian. On $\mathbb{R}^d$, $d \in \mathbb{N}$, there exist at least ten different approaches to define $(-\Delta)^s$, $s \in (0,1)$, which are all known to be equivalent [Kwa17]. These equivalences break down on bounded domains $\Omega \subset \mathbb{R}^d$ as there are several mathematically distinct ways to impose boundary conditions. Among these competing definitions of the fractional Laplacian, we are interested in the one obtained by spectral expansion.

The nonlocal nature of this operator has immediate consequences on some basic questions, such as the fractional Poisson problem:

$$\begin{aligned} (-\Delta)^s u &= f, \quad \text{in } \Omega, \\ u &= 0, \quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where Ω is a bounded Lipschitz domain and $f \in L^2(\Omega)$. Solutions to (1.1) for $\Omega = (0,1)^2$ and $f \equiv 1$ are depicted in Figure 1.1 and can be seen as showcase for the challenges that arise in the treatment of these problems.

- For small values of s, the fractional Laplacian $(-\Delta)^s$ is close to the identity operator, whence $u \approx 1$ in the interior of Ω. Towards the boundary the zero trace is imposed, which forces $u(\mathbf{x})$ to decrease rapidly as $\mathbf{x} \to \partial\Omega$. This is the reason why solutions to fractional diffusion problems exhibit limited regularity properties even if the underlying domain is smooth.

- The fractional exponent can be seen as parameter which is used to adapt the mathematical model to the observed data [BOKG+14, SV16]. As such, the precise value of s is typically unknown and needs to be determined experimentally in the course of a fitting procedure. Instead of solving (1.1) for one instance of s, one is typically interested in an approximation of the entire solution manifold $\{u(s) : s \in (0,1)\}$. We also refer to [AR19, ACR21] where the fractional parameter $s = s(\mathbf{x})$ is a function of a spatial variable which supports our point of view that $s \mapsto u(s)$ should be seen multi-query problem.

– Due of their nonlocal interactions, discretized FPDE systems have significantly less sparsity compared to discretized integer-order PDEs. Therefore, conventional localization techniques may fail to efficiently approximate such problems.

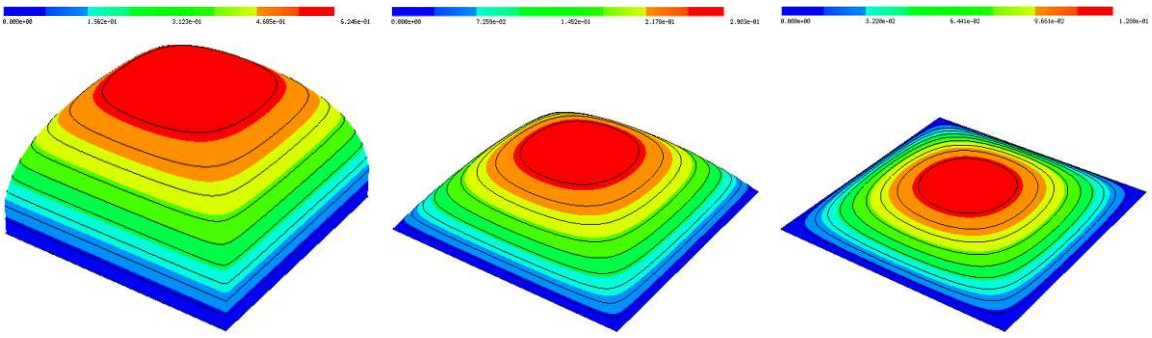

Figure 1.1: Solution u to (1.1) for $s = 0.2$ (left), $s = 0.5$ (middle), and $s = 0.8$ (right) on the unit square $\Omega = (0,1)^2$ with $f \equiv 1$.

The interest in fractional diffusion operators does not end in the stationary regime [SZB+18]. It has been observed in [BCdH08, FKR+21, FRW21] that the growth of tumors exhibits certain memory effects, making the problem global in time. Such phenomena can be modeled accurately using time-fractional differential operators. In combination with the fractional Laplacian, this leads us to the fractional heat equation

$$\begin{aligned} \partial_t^\alpha u + (-\Delta)^s u &= f, \quad \text{in } \Omega \times (0,T), \\ u &= 0, \quad \text{on } \partial\Omega \times (0,T), \\ u &= u_0, \quad \text{on } \partial\Omega \times \{0\}, \end{aligned} \tag{1.2}$$

where $\alpha \in (0,1]$ is the fractional time exponent, ∂_t^α the so-called Caputo fractional derivative of order α, $T \in \mathbb{R}^+$, $f \in L^\infty(0,T;L^2(\Omega))$ a forcing term, and $u_0 \in L^2(\Omega)$ some initial data. Adding to the difficulty of the nonlocal operator in space, the presence of ∂_t^α in (1.2) causes long-range interactions that require additional memory to store the history. The latter increases as T becomes large.

All these aspects need to be incorporated in the design of accurate, reliable, and yet computationally affordable numerical schemes. The amount of research published on this matter is vast and covers the treatment of elliptic problems [ILTA05, ILTA06, NOS15, BP15, Vab15, MN18, HLM+18, BLP19b, HMP21, DS21, DH21, HKL+21b, HKL+21a, Vab21a, Vab21b], space-fractional evolution equations [BLP17a, AM17, MR20b, Vab21c], time-fractional evolution equations [Lub88, JLZ15, KW21, FRW21], and fully space-time fractional diffusion problems [MN11, YTLI11, NOS16, BLP17b, Rie20, DHS21]. Needless to say, these references do not exhaust the rich literature on the subject. One way or another, either of the schemes listed above has to compensate for the nonlocality of the problem. Three classes of methods that we mention here explicitly are the following.

1. A conceptually straightforward approach is the accurate but expensive *discrete eigenfunction method* (DEM) [LPG+20, Hof20, BP15, ILTA05, ILTA06, YTLI11] which

relies on a matrix approximation $\mathbf{L} \in \mathbb{R}^{N\times N}$, $N \in \mathbb{N}$, of the spatial integer-order differential operator. The latter is used to interpret a discrete approximation of u as matrix-vector product $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$, where $f^{\boldsymbol{\tau}}$ is a matrix function that depends on a collection of parameters encoded in the vector $\boldsymbol{\tau} \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$, and $\mathbf{b} \in \mathbb{R}^N$ a vector that comes from the given data. Typical examples include

- $f^{\boldsymbol{\tau}}(\lambda) = f^s(\lambda) = \lambda^{-s}$ with $s \in \Theta = (0, 1)$,
- $f^{\boldsymbol{\tau}}(\lambda) = e^{-t\lambda^s}$ with $\boldsymbol{\tau} = (t, s) \in \Theta = \mathbb{R}^+ \times (0, 1)$,
- $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$, $\boldsymbol{\tau} = (\alpha, \beta, t, s) \in \Theta = \{(\alpha, \beta, t, s) \in (0, 1] \times \mathbb{R}^+ \times \mathbb{R}^+ \times [0, 1] : \beta \geq \alpha\}$, where $E_{\alpha,\beta}$ denotes the generalized Mittag-Leffler function.

The matrix $f^{\boldsymbol{\tau}}(\mathbf{L})$ is typically dense and its evaluation requires the knowledge of the entire eigensystem of $\mathbf{L}$. Having cubic complexity in N, this approach is only feasible if $\mathbf{L}$ is of moderate size.

2. Significantly more efficient are so-called quadrature schemes, which have been applied in [BP15, BP16, BLP17a, BLP17b, BGZ20, Rie20, DAC+21, AN21, DZ21], see also [DH21]. The idea is to rewrite $f^{\boldsymbol{\tau}}(\mathbf{L})$ via Cauchy's formula as a contour integral over a parametrized family of classical PDEs. The integral is discretized using a suitable quadrature whose evaluation boils down to the computation of multiple *local* problems which can be tackled using standard tools for elliptic PDEs. If $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$, the contour can be chosen as the negative real line, in which case the integral representation is known as Balakrishnan's formula [Bal60]. For time-dependent problems, one typically resorts to complex contours which in turn necessitates solutions to complex-valued problems, even if $\mathbf{L}$ and $\mathbf{b}$ are real.

3. The third and final approach we mention here is based on the harmonic extension technique developed in [CS07, ST10, CT10, CDDS11, BCdPS13]. The fractional differential equation is reinterpreted as local degenerate integer-order PDE on the semi-infinite cylinder $\mathcal{C}_\Omega := \Omega \times \mathbb{R}^+$. A natural approach consists of a $d + 1$ dimensional finite element method which takes advantage of the solution's rapid decay in the artificial direction, justifying truncation to a bounded domain of moderate size [NOS15, NOS16, BMN+18, BMS20, MR20b]; see also [AG18, ACN19]. Solutions to fractional diffusion problems can therefore be made available by resorting to well-known discretization methods for $d + 1$-dimensional degenerate integer-order problems.

Each of the schemes mentioned above yields accurate approximations to solutions of fractional diffusion problems. Unfortunately, however, the computational effort of each single solve depends on the problem size N. Hence, their implementation is unfeasible if one is interested in computing solutions for multiple instances of the parameters. A remedy to this problem are *model order reduction strategies* [QR14], in short MOR strategies. The idea is to add an additional layer of approximation to reduce the computational costs by a significant margin while keeping the discretization error to a tolerable level. The efficiency of such methods is gained by the so-called offline-online decomposition of the computational routine. In the *offline stage*, one precomputes several potentially costly auxiliary quantities which are independent of the parameters. These preparations

allow for rapid simulations of varying $\boldsymbol{\tau}$ in the *online stage*. Reduced order models of this form have been successfully applied to many different branches of computational science, such as fluid dynamics [QR07, SR18, HSMR20, KNBR21], fluid-structure interaction [LQR12, NBR21], optimal control problems [NRMQ13, SBMR18, ZBF+20], and shape optimization [QR03, LR10, SJC20]. In the context of fractional diffusion, one of the pioneering works in this direction has been presented in [WGP17], where it was observed experimentally that solutions to fractional PDEs exhibit a low-rank structure. Further MOR strategies have been applied to the harmonic extension setting [ACN19] and quadrature schemes [DAC+21], whose empirical findings support the conjecture that the solution manifold to the fractional Poisson problem (1.1) is compressible. One of the first rigorous results in this direction were given in [BGZ20]. Coupling a MOR method with a quadrature scheme, it is shown that the surrogate converges uniformly for all $s \in [s_{\min}, s_{\max}]$ to the DEM approximation of (1.1) at exponential rates. Here, $0 < s_{\min} < s_{\max} < 1$ are user-provided fixed parameters that determine in which interval one wishes to query the solution map. We finally also mention [MN11, MN18, ABDN19] where rational Krylov methods have been applied to reduce the computational costs in the evaluation of the DEM approximation.

A broad spectrum of powerful algorithms exists which allow one to approximate solutions to fractional PDEs for a few particular instances of the parameters. What many MOR strategies are still lacking, however, is the ability to efficiently approximate the entire solution manifold for all admissible values of the parameters. Furthermore, many schemes are tailored to the particular problem setup and need to distinguish e.g., between the approximation of elliptic and evolutionary problems. The latter often include solutions to complex-valued problems even if the differential operator and the data are real. The desire for numerical schemes which mitigate these disadvantages is a philosophy we adopt in this thesis.

On the basis of [DS19, DS21, DH21, DHS21], we present a unified MOR method to approximate solutions to fractional diffusion problems of elliptic and parabolic type. Using the DEM as a starting point, a finite element method is applied to write the discrete solution as matrix-vector product of the form $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$. To diminish the computational costs, we establish a MOR strategy in terms of a rational Krylov method (RKM) which projects the matrix approximation to a low-dimensional space where a direct computation of the eigensystem is feasible. The particular choice of the subspace depends on a collection of parameters $\Xi := \{\xi_0, \dots, \xi_k\} \subset \mathbb{R}$, the so-called *poles*. The latter have a crucial impact on the performance of the RKM and need to be selected a priori. We propose several attractive pole selection strategies which allow us to approximate solutions to elliptic and parabolic fractional diffusion problems simultaneously. The core of this thesis is the analysis of these poles. One of our main results is the intriguing fact that the pole set Ξ can be chosen independently of the fractional parameters. Its proof is based on the observation that the parametric function $f^{\boldsymbol{\tau}}$, stemming from the particular problem, can be classified as Cauchy-Stieltjes, complete Bernstein, or Laplace-Stieltjes function. This unified point of view allows us to bound the approximation error by the *third Zolotarëv problem*. Inspired by these results, we prove uniform convergence rates when choosing the poles according to solutions of the third Zolotarëv problem. Unlike prior works, the approximation so obtained does not degenerate as the spatial fractional exponent approaches an integer. We also provide the description of an error certificate which allows us to assess the quality of a

large class of poles where no theoretical bounds for the error are available. While the scope of this thesis is limited to real matrices only, it can be easily seen that our results carry over to the complex Hermitian case.

In the final part of this thesis, we provide deep insights in several other MOR strategies advocated by the literature and show that these methods can be interpreted as variants of conventional rational Krylov methods. The theoretical insights so obtained allow us to leverage our analysis for RKMs to develop new convergence proofs for several of the studied schemes. They suggest how to design novel and improve available methods and allow for a direct comparison of the algorithms.

1.1 Structure of the Thesis

For the reader's convenience, we give a brief overview of the structure of this manuscript and provide a survey of the main components of each individual chapter.

- After some remarks on notation, we gather in Chapter 2 several well-known results to find a common ground for further discussions. We review the basics of Sobolev theory and recall the notion of Bochner integrals and Bochner spaces. We also provide a brief exposition on some special functions that shall be useful later on and remind the reader of the foundations of matrix functions.

- To encompass the full scope of this thesis, we review the theory of interpolation spaces in an abstract Hilbert space framework. As special cases, fractional Sobolev spaces are discussed together with several of their equivalent definitions.

- Given a generic differential operator $\mathcal{L}$, we introduce, in Chapter 4, its fractional power $\mathcal{L}^s$ of order $s \in (0,1)$ as operator of interpolation and provide three equivalent characterizations of the latter

  1. using the eigensystem of $\mathcal{L}$,
  2. as improper integral over parametric reaction-diffusion equations,
  3. as Dirichlet-to-Neumann map of a degenerate PDE on an artificially extruded cylinder $\mathcal{C}_\Omega = \Omega \times \mathbb{R}^+$.

  We study regularity properties of solutions to (1.1) and highlight the differences to the integer-order regime. Finally, several other possible definitions of $\mathcal{L}^s$ are discussed when $\mathcal{L} = -\Delta$.

- In Chapter 5, we give a brief survey of the foundations of fractional calculus. We introduce the Caputo fractional derivative of order $\alpha \in [0,1]$ as the time-fractional differential operator of our choice and consider its interaction with $\mathcal{L}^s$ in the context of fractional evolution equations.

- Chapter 6 is devoted to the discretization of fractional PDEs. For this purpose, we choose the finite element method as underlying discretization method for the spatial variable. Using the three charaterizations of $\mathcal{L}^s$, presented in Chapter 4, as a starting point, we introduce

1. the discrete eigenfunction method,
2. a quadrature scheme,
3. the harmonic extension method,

to approximate solutions to elliptic and parabolic problems of fractional diffusion type. We show that the computation of the DEM surrogate boils down to the evaluation of a matrix-vector product $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$, where $\mathbf{L}$ is a finite element matrix approximation of $\mathcal{L}$, $\mathbf{b}$ a vector stemming from the given data, and $f^{\boldsymbol{\tau}}$ a parametric matrix function. On the other hand, we rewrite the surrogates obtained by the quadrature scheme and the harmonic extension method as matrix-vecor product of the form $r^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$, where $r^{\boldsymbol{\tau}}$ is a rational function that can be seen as rational approximation of $f^{\boldsymbol{\tau}}$.

– The rational Krylov method enters the stage in Chapter 7 as the model order reduction scheme of our choice to approximate the discrete solution map $\boldsymbol{\tau} \mapsto f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ efficiently. For this purpose we establish, in dependency of the pole set $\Xi = \{\xi_0, \dots, \xi_k\} \subset \mathbb{R}$, the rational Krylov space $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ of dimension $k+1$ in which the rational Krylov approximation $\mathbf{u}_{k+1} \approx f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ is found via Rayleigh-Ritz extraction.

– We introduce, in Chapter 8, the notion of Cauchy-Stieltjes, complete Bernstein, and Laplace-Stieltjes functions and prove that $f^{\boldsymbol{\tau}}$ has membership in at least one of these classes. This unified point of view allows us to bound the rational Krylov error in terms of a particular rational approximation problem which is the key ingredient of our analysis.

– A connection between the rational approximation problem mentioned above and the third Zolotarëv problem is established in Chapter 9. After a concise survey of logarithmic potential theory, we derive explicit solutions to the third Zolotarëv problem which can be used to minimize the upper bound derived in Chapter 8.

– Chapter 10 constitutes the heart of this thesis and is devoted to the selection of poles $\Xi = \{\xi_0, \dots, \xi_k\}$ to build the rational Krylov space $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$. We present a broad spectrum of pole selection strategies that are suitable for the approximation of fractional PDEs. A systematical comparison of these poles provides deep insights in their strengths, weaknesses, and similarities. To quantify the performance of Ξ, we either provide rigorous analytical results or present a description of an error certificate which allows one to assess the quality of several pole sets even if no bounds for the error are available. A variety of numerical experiments are presented that underpin the main results of this chapter.

– In the final chapter of this thesis, we present a selection of MOR methods for fractional diffusion problems that are based on rational approximation. We demonstrate that several of these schemes can be interpreted as RKMs. This changed point of view allows us to develop new convergence proofs and suggests how to design novel and improve available algorithms.

1.2 Remarks on Notion

Throughout this thesis, the natural numbers $\mathbb{N}$ are understood as the set of all strictly positive integers and we set $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ by convention. We designate by $\mathbb{R}^+$ the set of all strictly positive real numbers and define $\mathbb{R}_0^+ := \mathbb{R}^+ \cup \{0\}$. The sets $\mathbb{R}^-$ and $\mathbb{R}_0^-$ are understood accordingly. By $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$ we denote the extended complex plane and set $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. For each $z \in \mathbb{C}$ we write $\Re z$ and $\Im z$ to denote the real and imaginary part of z, respectively. The argument $\arg(z)$ of $z \in \mathbb{C} \setminus \{0\}$ is defined as the angle $\phi \in (-\pi, \pi]$ of $z = re^{i\phi}$, $r \in \mathbb{R}^+$, in polar coordinates. The power function z^s is defined in the usual way $z^s := e^{s \ln(z)}$, where we use the branch-cut at $\mathbb{R}_0^-$ for the logarithm. Unless stated otherwise, the reader is encouraged to think of the letter λ as a scalar and real quantity, whereas z typically denotes a (possibly) complex one. For any function $f(z)$ that is defined on a subset $D \subset \mathbb{C}$ of the complex plane with $D \cap \mathbb{R} \neq \emptyset$, we again use the same letter f to refer to its restriction $f : D \cap \mathbb{R} \to \mathbb{C}$. If necessary, however, we shall write $f(\lambda)$ instead of $f(z)$ to emphasize which domain of definition is meant. Likewise, for any function $f(\lambda)$ defined on a subset of the real line that extends analytically to the complex plane we write $f(z)$ to denote its analytic continuation. The support of the function $f(z)$ is defined by

$$\operatorname{supp} f := \overline{\{z \in D : |f(z)| > 0\}},$$

where $|\cdot|$ labels the absolute value and $\overline{A}$ the closure of $A \subset \mathbb{C}$. We use the shorthand notation

$$1/A := \{1/a : a \in A\}, \qquad -A := \{-a : a \in A\}.$$

A property is said to hold almost everywhere (a.e.) in A if it holds in $A \setminus N$ where N is a subset of A with zero Lebesgue measure.

Consistently, we use bold lower-case letters for vectors and bold capital letters for matrices. Provided a symmetric and positive definite matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, we define

$$\|\mathbf{x}\|_{\mathbf{M}} := \sqrt{(\mathbf{x}, \mathbf{x})_{\mathbf{M}}}, \qquad (\mathbf{x}, \mathbf{y})_{\mathbf{M}} := \mathbf{x}^T \mathbf{M} \mathbf{y}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

The Euclidean norm and inner product on $\mathbb{R}^d$, $d \in \mathbb{N}$, is abbreviated by

$$\|\mathbf{x}\|_2 := \mathbf{x} \cdot \mathbf{y} := \|\mathbf{x}\|_{\mathbf{I}}, \qquad (\mathbf{x}, \mathbf{x})_2 := (\mathbf{x}, \mathbf{y})_{\mathbf{I}},$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ labels the unit matrix. The matrix $\mathbf{M}$ is said to be diagonalizable if there exists some $\mathbf{U} \in \mathbb{R}^{d \times d}$ invertible such that $\mathbf{U}\mathbf{M}\mathbf{U}^{-1} = \mathbf{D}$, where $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix with entries $\lambda_1, \ldots, \lambda_d$. We denote the open ball with center $\mathbf{x} \in \mathbb{R}^d$ and radius $r \in \mathbb{R}^+$ in $\mathbb{R}^d$ by

$$B_\varepsilon(\mathbf{x}) := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < r\}.$$

The distance between two subsets $A, B \subset \mathbb{C}$ of the complex plane is defined by

$$\operatorname{dist}(A, B) := \inf\{\|a - b\|_2 : a \in A, b \in B\}.$$

Furthermore, we write $a \preceq b$ to indicate $a \leq Cb$ for some constant $C \in \mathbb{R}^+$ that is independent of of a, b, the finite element mesh size h, the rational Krylov parameter k, and the quadrature parameter q.

For any Hilbert space $\mathcal{H}$ we denote with $(u, v)_{\mathcal{H}}$ the scalar product on $\mathcal{H}$ and $\|u\|_{\mathcal{H}} := \sqrt{(u, u)_{\mathcal{H}}}$ its norm. Whenever we refer to a Banach space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ as Hilbert space, we mean that its norm induces a scalar product $(\cdot, \cdot)_{\mathcal{H}}$ on $\mathcal{H}$, obtained by polarization identity, such that $(\mathcal{H}, (\cdot, \cdot)_{\mathcal{H}})$ is a Hilbert space. A linear operator $B : \mathcal{H} \to \hat{\mathcal{H}}$ between two Hilbert spaces is said to be bounded (or continuous) if

$$\|Bu\|_{\hat{\mathcal{H}}} \preceq \|u\|_{\mathcal{H}}$$

for all $u \in \mathcal{H}$. The sum of two Hilbert spaces, $\mathcal{H} + \hat{\mathcal{H}}$, is defined as the smallest vector space that contains both $\mathcal{H}$ and $\hat{\mathcal{H}}$. The dual space of $\mathcal{H}$ is always understood in the topological sense and is labeled as $\mathcal{H}'$. It is equipped with the operator norm

$$\|f\|_{\mathcal{H}'} := \sup_{v \in \mathcal{H}} \frac{\langle f, v \rangle}{\|v\|_{\mathcal{H}}},$$

where $\langle f, v \rangle := f(v)$ denotes the dual pairing. Its inner product is defined by

$$(g, f)_{\mathcal{H}'} := \langle g, \mathcal{R} f \rangle,$$

where $\mathcal{R} : \mathcal{H}' \to \mathcal{H}$ denotes the Riesz isomorphism defined by

$$(\mathcal{R} f, v)_{\mathcal{H}} = \langle f, v \rangle, \qquad v \in \mathcal{H}.$$

2 Preliminaries

In this chapter, we gather several important results that shall serve us as foundation for upcoming discussions. In the first section, we recall some common textbook knowledge on Sobolev spaces which provides the corner stone of classical interpolation theory. The latter makes heavy use of Hilbert-valued integrals whence we provide a concise survey of Bochner integrals and Bochner spaces in Section 2.2. In Section 2.3, we provide a succinct overview of the Laplace transform and some special functions that arise in the study of fractional evolution equations. Finally, for the numerical approximations presented in the second half of this thesis, we review the theory of matrix functions in Section 2.4 and state some results that are the building block of our analysis later on.

2.1 Classical Sobolev Theory

Throughout this thesis, we shall be concerned with function spaces defined on a subset Ω of $\mathbb{R}^d$, $d \in \mathbb{N}$. We limit ourselves to domains that allow for a local parametrization of the boundary using Lipschitz continuous functions.

Definition 2.1. *An open, bounded, and connected set $\Omega \subset \mathbb{R}^d$, is said to be a Lipschitz domain if for all $\mathbf{x} = (x_1, \dots, x_d) \in \partial\Omega$ there exists some $\varepsilon > 0$ and a bijective function $\Phi : B_1(0) \to B_\varepsilon(\mathbf{x})$ such that*

- *Φ and Φ^{-1} are Lipschitz continuous,*
- *$\Phi(\{\mathbf{x} \in B_1(0) : x_d = 0\}) = \partial\Omega \cap B_\varepsilon(\mathbf{x})$,*
- *$\Phi(\{\mathbf{x} \in B_1(0) : x_d < 0\}) = \Omega \cap B_\varepsilon(\mathbf{x})$,*
- *$\Phi(\{\mathbf{x} \in B_1(0) : x_d > 0\}) = (\mathbb{R}^d \setminus \overline{\Omega}) \cap B_\varepsilon(\mathbf{x})$.*

For the remainder of this thesis, the set Ω is always assumed to be a Lipschitz domain in the sense stated above. Sometimes we require stronger regularity assumptions on Ω, in which case Ω is said to be a C^∞-domain if its local parametrization Φ from Definition 2.1 is smooth.

The definition of *Sobolev spaces* on Ω relies on the theory of distributions [Rud74, Yos95, AF03, McL00, Eva10] and requires some further preparations. For this purpose, we denote by $C^k(\Omega)$, $k \in \mathbb{N}_0$, the function space consisting of all real-valued k-times continuously differentiable functions on Ω with range in $\mathbb{R}$ and set

$$C^\infty(\Omega) := \bigcap_{k \in \mathbb{N}_0} C^k(\Omega).$$

For brevity, we define $C(\Omega) := C^0(\Omega)$. The *gradient*, the *divergence*, and the *Laplacian* are defined in the usual way

$$\nabla u := \left(\frac{\partial u}{\partial \mathbf{x}_1}, \dots, \frac{\partial u}{\partial \mathbf{x}_d}\right), \qquad \operatorname{div}(w) := \sum_{j=1}^{d} \frac{\partial w_j}{\partial \mathbf{x}_j}, \qquad \Delta u := \operatorname{div}(\nabla u) = \sum_{j=1}^{d} \frac{\partial^2 u}{\partial^2 \mathbf{x}_j}.$$

The space of square-integrable functions on Ω is defined by

$$L^2(\Omega) := \{u : \Omega \to \mathbb{R} : \|u\|_{L^2(\Omega)} < \infty\}, \qquad \|u\|^2_{L^2(\Omega)} := \int_\Omega |u(\mathbf{x})|^2 \, d\mathbf{x}.$$

Endowed with its natural inner product

$$(u, v)_{L^2(\Omega)} := \int_\Omega u(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x},$$

$L^2(\Omega)$ is a Hilbert space. To introduce the notion of *weak differentiability*, we present the set of *test functions*

$$\mathcal{D}(\Omega) := C_0^\infty(\Omega) := \{v \in C^\infty(\Omega) : \operatorname{supp} v \subset \Omega \text{ compact}\}.$$

The space is equipped with the topology defined by the following notion of convergence:

$$v_j \to 0 \ :\Longleftrightarrow \ \exists K \subset \Omega \text{ compact: } \operatorname{supp} v_j \subset K \text{ and } \forall \gamma \in \mathbb{N}_0^d : D^\gamma v_j \to 0 \text{ uniformly in } K,$$

where we use multi-index notation

$$\gamma = (\gamma_1, \dots, \gamma_d), \qquad D^\gamma = \frac{\partial^{|\gamma|}}{\partial \mathbf{x}_1^{\gamma_1} \dots \partial \mathbf{x}_d^{\gamma_d}}, \qquad |\gamma| := \sum_{j=1}^{d} \gamma_j.$$

Its topological dual space

$$\mathcal{D}'(\Omega) := \{f : C_0^\infty(\Omega) \to \mathbb{R} : f \text{ is continuous and linear}\}$$

is called *the space of distributions*. We introduce the set of locally integrable functions

$$L^1_{\mathrm{loc}}(\Omega) := \{u : \Omega \to \mathbb{R} : \int_K u(\mathbf{x}) \, d\mathbf{x} < \infty \text{ for all } K \subset \Omega \text{ compact}\}.$$

Recall that any function $u \in L^1_{\mathrm{loc}}(\Omega)$ defines a distribution $\mathcal{F}_u \in \mathcal{D}'(\Omega)$ in the sense of

$$\langle \mathcal{F}_u, v \rangle := \int_\Omega u(\mathbf{x}) \, v(\mathbf{x}) \, d\mathbf{x}, \qquad v \in \mathcal{D}(\Omega),$$

where $\langle \mathcal{F}_u, v \rangle = \mathcal{F}_u(v)$ denotes the duality pairing. Distributions of this form are called *regular distributions*. For the remainder of this thesis, we identify any regular distribution with the function u by whom it is generated.

Clearly, any $u \in C_0^\infty(\Omega)$ is contained in $L^1_{\mathrm{loc}}(\Omega)$. Since u vanishes at the boundary of $\partial\Omega$, integration by parts reveals

$$\int_\Omega D^\gamma u(\mathbf{x}) v(\mathbf{x})\, d\mathbf{x} = (-1)^{|\gamma|} \int_\Omega u(\mathbf{x}) D^\gamma v(\mathbf{x})\, d\mathbf{x}, \qquad v \in C_0^\infty(\Omega).$$

This provides the motivation to introduce the *weak derivative* $D^\gamma u$ of the regular distribution $u \in L^1_{\mathrm{loc}}(\Omega)$ by

$$\langle D^\gamma u, v\rangle := \int_\Omega D^\gamma u(\mathbf{x}) v(\mathbf{x})\, d\mathbf{x} := (-1)^{|\gamma|} \int_\Omega u(\mathbf{x}) D^\gamma v(\mathbf{x})\, d\mathbf{x}, \quad v \in C_0^\infty(\Omega).$$

Clearly, if u is sufficiently smooth, then all classical and weak derivatives coincide. In light of these considerations, we introduce the Sobolev space of order one as

$$H^1(\Omega) := \{u \in L^2(\Omega) : \nabla u \in [L^2(\Omega)]^d\}, \qquad \|u\|^2_{H^1(\Omega)} := \|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)},$$

where the L^2-norm of vector-valued functions $v : \Omega \to \mathbb{R}^d$ is defined by

$$\|v\|_{L^2(\Omega)} := \int_\Omega \|v(\mathbf{x})\|_2^2\, d\mathbf{x}.$$

Occasionally, we shall write $H^0(\Omega) = L^2(\Omega)$. Sobolev spaces of arbitrary order $k \in \mathbb{N}$ are defined inductively

$$H^k(\Omega) := \{u \in L^2(\Omega) : \nabla u \in [L^2(\Omega)]^d\}, \qquad \|u\|^2_{H^k(\Omega)} := \|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{H^{k-1}(\Omega)},$$

The space $H^k(\Omega)$ equipped with

$$(u, v)_{H^k(\Omega)} := (u, v)_{L^2(\Omega)} + (\nabla u, \nabla v)_{H^{k-1}(\Omega)}$$

is a Hilbert space. An equivalent but less practical characterization reads [AF03]

$$H^k(\Omega) = \overline{C^\infty(\overline{\Omega})}^{\|\cdot\|_{H^k(\Omega)}}, \tag{2.1}$$

where $\overline{C^\infty(\overline{\Omega})}^{\|\cdot\|_{H^k(\Omega)}}$ denotes the closure of $C^\infty(\overline{\Omega})$ with respect to the norm $\|\cdot\|_{H^k(\Omega)}$. As a consequence, each $u \in H^k(\Omega)$ can be approximated by a sequence of smooth functions. While L^2-functions in general do not allow for pointwise evaluation, the following theorem shows that the trace evaluation of Sobolev functions is well-defined; see e.g., [AF03].

Theorem 2.2 (Trace theorem)**.** *There exists a linear operator* $\operatorname{tr} : H^1(\Omega) \to L^2(\partial\Omega)$ *with the properties*

$$\|\operatorname{tr} u\|_{L^2(\Omega)} \preceq \|u\|_{H^1(\Omega)}, \qquad \forall u \in H^1(\Omega) \cap C(\overline{\Omega}) : \operatorname{tr} u = u|_{\partial\Omega}.$$

This result allows one to include boundary conditions in the sense of

$$H^1_0(\Omega) := \{u \in H^1(\Omega) : \operatorname{tr} u = 0\}.$$

In accordance with (2.1), there holds

$$H^1_0(\Omega) = \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^1(\Omega)}}. \tag{2.2}$$

The availability of trace operations allow us to present the generalized integration by parts formula for Sobolev functions.

Theorem 2.3. *Let $\mathbf{n} \in \mathbb{R}^d$ the outer normal vector to $\partial\Omega$, $u \in H^1(\Omega)$, and $v \in [H(\Omega)]^d$. Then there holds*

$$\int_\Omega \nabla u \cdot v \, d\mathbf{x} = -\int_\Omega u \operatorname{div} v \, d\mathbf{x} + \int_{\partial\Omega} (\operatorname{tr} v \cdot \mathbf{n}) \operatorname{tr} u \, ds.$$

Under suitable assumptions on the Sobolev order k, one can expect $u \in H^k(\Omega)$ to be continuous. To make matters precise, we introduce the norm

$$\|u\|^2_{C^k(\overline{\Omega})} := \sum_{|\gamma| \le k} \sup_{\mathbf{x} \in \overline{\Omega}} |D^\gamma u(\mathbf{x})|$$

to formalize the following result.

Theorem 2.4. *Let $k \in \mathbb{N}$ and $m \in \mathbb{N}_0$ with $k - \frac{d}{2} > m$. Then the embedding $H^k(\Omega) \subset C^m(\overline{\Omega})$ is continuous, i.e.,*

$$\|u\|_{H^k(\Omega)} \preceq \|u\|_{C^m(\overline{\Omega})}.$$

The following theorem gathers two important results that shall be frequently referred to in the further course of this thesis. It can be found in several textbooks, e.g., [AF03, Eva10].

Theorem 2.5. *Let $k \in \mathbb{N}_0$.*

1. *If $m \in \mathbb{N}$ with $m > k$, then the embedding $H^m(\Omega) \subset H^k(\Omega)$ is compact.*
2. *The embedding $H^1_0(\Omega) \subset H^1(\Omega)$ is continuous, i.e.,*

$$\|u\|_{H^1(\Omega)} \preceq \|\nabla u\|_{L^2(\Omega)}.$$

It turns out to be fruitful to introduce Sobolev spaces with negative exponents. This is usually done in the following manner.

Definition 2.6. *For all $k \in \mathbb{N}$ we define the space $H^{-k}(\Omega) := \big(H^k_0(\Omega)\big)'$ as the dual space of $H^k_0(\Omega)$.*

2.2 Bochner Theory

The purpose of this section is to generalize the theory of the Lebesgue integral and Sobolev spaces for functions of the form $u : (0,T) \to \mathcal{H}$, where $T \in \mathbb{R}^+ \cup \{\infty\}$ and $\mathcal{H}$ is a Hilbert space that we assume to be fixed throughout this section. This is a classical field of research and can be found in e.g., [Boc33, Hil53, Mik78, Yos95]. Reminiscent of standard Lebesgue theory, the integrability of $\mathcal{H}$-valued functions is built upon the theory of step functions, for which we introduce the indicator function of $I \subset \mathbb{R}$ by

$$\mathbb{1}_I(t) := \begin{cases} 1, & \text{if } t \in I, \\ 0, & \text{else.} \end{cases}$$

For simplicity, we assume $\mathcal{H}$ to be separable henceforth, meaning that $\mathcal{H}$ contains a dense subset of countable cardinality.

Definition 2.7. *A function $S : (0,T) \to \mathcal{H}$ is called step function (or simple) if there exists some $n \in \mathbb{N}$, a family of pairwise disjoint sets $(I_j)_{j=1}^n \subset (0,T)$ with finite Lebesgue measure, and $(h_j)_{j=1}^n \subset \mathcal{H}$, such that*

$$S(t) = \sum_{j=1}^{n} h_j \mathbb{1}_{I_j}(t).$$

We denote the set of all step functions with $\mathbb{S}((0,T);\mathcal{H})$.

The integral of functions contained in $\mathbb{S}((0,T);\mathcal{H})$ is defined in the expected fashion.

Definition 2.8. *Let $S \in \mathbb{S}((0,T);\mathcal{H})$ with*

$$S(t) = \sum_{j=1}^{n} h_j \mathbb{1}_{I_j}(t).$$

Then the Bochner integral of S is defined as

$$\int_0^T S(t)\,dt := \sum_{j=1}^{n} h_j \lambda(I_j), \tag{2.3}$$

where $\lambda(I_i)$ denotes the Lebesgue measure of I_j.

It can be readily verified that (2.3) defines a linear operator from $\mathbb{S}((0,T);\mathcal{H})$ to $\mathcal{H}$. To generalize this concept for arbitrary $u : (0,T) \to \mathcal{H}$, we agree on the following terminology.

Definition 2.9. *A function $u : (0,T) \to \mathcal{H}$ is called Bochner-integrable if there exists a sequence of step functions $(S_n)_{n\in\mathbb{N}} \subset \mathbb{S}((0,T);\mathcal{H})$ with the properties*

- *for almost all $t \in (0,T)$*

$$\lim_{n\to\infty} S_n(t) = u(t),$$

- *there holds*

$$\lim_{n\to\infty} \int_0^T \|S_n(t) - u(t)\|_{\mathcal{H}}\,dt = 0. \tag{2.4}$$

The space of Bochner-integrable functions is denoted by $L^1(0,T;\mathcal{H})$. For any $u \in L^1(0,T;\mathcal{H})$ we define the Bochner integral as

$$\int_0^T u(t)\,dt := \lim_{n\to\infty} \int_0^T S_n(t)\,dt. \tag{2.5}$$

One shows that

- the integral in (2.4) exists in the classical sense of Lebesgue,
- (2.5) is independent of the particular choice of the sequence of step functions,

whence the Bochner integral is well-defined. Reminiscent of classical Lebesgue theory, a convenient criterion for Bochner-integrability can be expressed in terms of *Bochner-measurable functions* [Mik78].

Definition 2.10. *A function $u : (0,T) \to \mathcal{H}$ is said to be Bochner-measurable (or strongly measurable) if there exists a sequence of step functions $(S_n)_{n\in\mathbb{N}} \subset \mathbb{S}((0,T);\mathcal{H})$ such that for almost all $t \in (0,T)$*

$$\lim_{n\to\infty} \|S_n(t) - u(t)\|_{\mathcal{H}} = 0.$$

Proposition 2.11. *There holds*

1. *$u : (0,T) \to \mathcal{H}$ is Bochner-measurable if and only if $t \mapsto (u(t),v)_{\mathcal{H}}$ is Lebesgue-measurable for all $v \in \mathcal{H}$,*
2. *$u \in L^1(0,T;\mathcal{H})$ if and only if $u : (0,T) \to \mathcal{H}$ is Bochner-measurable and $\|u(\cdot)\|_{\mathcal{H}} \in L^1((0,T))$.*

Further well-known properties of the Bochner integral are listed in the following lemma (cf. [Mik78]), where we introduce the convolution of $u, v \in L^1(0,T;\mathcal{H})$ as

$$(u * v)(t) := \int_{-\infty}^{\infty} \tilde{u}(t-\tau)\tilde{v}(\tau)\,d\tau,$$

with

$$\tilde{w}(t) := \begin{cases} w(t), & t \in (0,T), \\ 0, & t \notin (0,T), \end{cases} \qquad w \in L^1(0,T;\mathcal{H}).$$

Lemma 2.12 (Elementary properties)**.**

1. *For all $u \in L^1(0,T;\mathcal{H})$ there holds*

$$\left\| \int_0^T u(t)\,dt \right\|_{\mathcal{H}} \le \int_0^T \|u(t)\|_{\mathcal{H}}\,dt.$$

2. *If $\hat{\mathcal{H}}$ is another Hilbert space and $B : \mathcal{H} \to \hat{\mathcal{H}}$ a bounded linear operator, then*

$$B\int_0^T u(t)\,dt = \int_0^T Bu(t)\,dt.$$

3. *If $u \in L^1(0,T;\mathcal{H})$ and $v \in \mathcal{H}$, then $t \mapsto (u(t),v)_{\mathcal{H}}$ is Lebesgue-integrable and*

$$\left(\int_0^T u(t)\,dt, v \right)_{\mathcal{H}} = \int_0^T (u(t),v)_{\mathcal{H}}\,dt.$$

4. *If $u, v \in L^1(0,T;\mathcal{H})$, then $u * v \in L^1(0,T;\mathcal{H})$.*

A fundamental field of study in integration theory is the question under which conditions one can interchange the integral and the limit of a sequence of functions.

Theorem 2.13 (Lebesgue's dominated convergence theorem)**.** *Let $(u_n)_{n\in\mathbb{N}}$ be a sequence in $L^1(0,T;\mathcal{H})$ with $u_n(t) \to u(t)$ a.e. in $(0,T)$ and $g \in L^1((0,T))$ such that for all $n \in \mathbb{N}$ there holds $\|u_n(t)\|_{\mathcal{H}} \leq |g(t)|$ a.e. in $(0,T)$. Then $u \in L^1(0,T;\mathcal{H})$ and*

$$\lim_{n\to\infty} \int_0^T \|u(t) - u_n(t)\|_{\mathcal{H}}\, dt = 0, \qquad \lim_{n\to\infty} \int_0^T u_n(t)\, dt = \int_0^T u(t)\, dt.$$

The L^p-Bochner spaces are a straightforward generalization of their scalar counterpart.

Definition 2.14. *For each $p \in \mathbb{N} \cup \{\infty\}$ we define the Bochner space $L^p(0,T;\mathcal{H})$ as the vector space of equivalence classes of almost everywhere coinciding Bochner-measurable functions such that $\|u\|_{L^p(0,T;\mathcal{H})} < \infty$, where*

$$\|u\|_{L^p(0,T;\mathcal{H})} := \begin{cases} \left(\int_0^T \|u(t)\|_{\mathcal{H}}^p\, dt\right)^{\frac{1}{p}}, & \text{if } p \in \mathbb{N}, \\ \inf\{C \in \mathbb{R}^+ : \|u(t)\|_{\mathcal{H}} \leq C \text{ for almost all } t \in (0,T)\}, & \text{if } p = \infty. \end{cases}$$

There holds $L^p(0,T;\mathcal{H}) \subset L^q(0,T;\mathcal{H})$ if $q \leq p$. For all $p \in \mathbb{N}\cup\{\infty\}$ the space $L^p(0,T;\mathcal{H})$ is a Banach space. Only if $p = 2$, the norm $\|u\|_{L^2(0,T;\mathcal{H})}$ comes from an inner product

$$(u,v)_{L^2(0,T;\mathcal{H})} := \int_0^T (u(t), v(t))_{\mathcal{H}}\, dt.$$

Now that we are familiar with Lebesgue spaces, we limit ourselves to the Hilbert space case $p = 2$ and dedicate our attention to the definition of Bochner-Sobolev spaces. This requires the notion of differentiation for Hilbert-valued functions.

Definition 2.15. *A function $u : (0,T) \to \mathcal{H}$ is said to be differentiable in $t \in (0,T)$ if there exists some $v \in \mathcal{H}$ such that*

$$\lim_{\substack{\delta\to 0 \\ t+\delta\in(0,T)}} \frac{u(t+\delta) - u(t)}{\delta} = v \text{ in } \mathcal{H}.$$

We call $v =: u'$ the derivative of u. By $C^1((0,T);\mathcal{H})$ we denote the set of all differentiable functions $u : (0,T) \to \mathcal{H}$ whose derivatives are continuous.

Accordingly, the set $C^k((0,T);\mathcal{H})$, $k \in \mathbb{N}_0 \cup \{\infty\}$, is understood in the expected manner. This notion of differentiation can be essentially relaxed. We proceed as in the real-valued case.

Definition 2.16. *A function $u : (0,T) \to \mathcal{H}$ is said to be weakly differentiable if there exists some $v \in \mathcal{H}$ such that*

$$\forall w \in C_0^\infty((0,T)) : \quad \int_0^T v(t)w(t)\, dt = -\int_0^T u(t)\partial_t w(t)\, dt.$$

We call $v =: \partial_t u$ the weak derivative of u. Thereupon, the Bochner-Sobolev space of order $k \in \mathbb{N}$ is inductively defined by

$$H^1((0,T);\mathcal{H}) := \{u \in L^2(0,T;\mathcal{H}) : \partial_t u \in L^2(0,T;\mathcal{H})\}$$

and

$$H^k((0,T);\mathcal{H}) := \{u \in L^2(0,T;\mathcal{H}) : \partial_t u \in H^{k-1}((0,T);\mathcal{H})\}, \qquad k \geq 2.$$

Several important properties of Bochner-Sobolev spaces follow from the respective property of $\mathcal{H}$. The reader is encouraged to compare the following results for $H^1((0,T);\mathcal{H})$ with the corresponding ones of classical Sobolev spaces.

Lemma 2.17. *The space $H^1((0,T);\mathcal{H})$ equipped with the inner product*

$$(u,v)_{H^1((0,T);\mathcal{H})} := \int_0^T (u(t),v(t))_{\mathcal{H}} + (\partial_t u(t), \partial_t v(t))_{\mathcal{H}}\, dt$$

is a separable Hilbert space. Moreover,

1. *the embedding $H^1((0,T);\mathcal{H}) \subset C([0,T];\mathcal{H})$ is continuous,*
2. *there holds the integration by parts formula*

$$\int_0^T (\partial_t u(t), v(t))_{\mathcal{H}}\, dt = (u(T),v(T))_{\mathcal{H}} - (u(0),v(0))_{\mathcal{H}} - \int_0^T (\partial_t v(t), u(t))_{\mathcal{H}}\, dt$$

for all $u, v \in H^1((0,T);\mathcal{H})$.

2.3 Preliminaries from Fractional Calculus

In this section we recall the definition of the Laplace transform and remind the reader of some of its well-known properties. Moreover, we make ourselves familiar with the Gamma function and the (generalized) Mittag-Leffler function which are key ingredients in the study of fractional evolution equations.

2.3.1 The Laplace Transform

The Laplace transform is a crucial tool for the analysis of fractional evolution equations and allows one to transform the latter to algebraic equations which are typically easier to deal with. We take our definition from the classical work of Widder [Wid43], see also [Die10], where the interested reader may find a comprehensive treatment of this matter.

Definition 2.18. *Let $u : \mathbb{R}_0^+ \to \mathbb{C}$ be a function and $T, z_0, M \in \mathbb{R}^+$ with the property*

$$\forall t > T : \quad e^{-z_0 t}|u(t)| \leq M, \qquad \int_0^T |u(t)|\, dt < \infty. \tag{2.6}$$

Then the Laplace transform of u is defined by

$$\mathscr{L}[u](z) := \int_0^\infty e^{-zt} u(t)\, dt, \qquad \Re z > z_0.$$

For some elementary functions, the Laplace transform can be computed directly.

Example 2.19. *Consider the function $u(t) = e^{ct}$ for some $c \in \mathbb{R}$. Then u satisfies (2.6) with $z_0 = c$, $M = 1$, and arbitrary $T \in \mathbb{R}^+$. Its Laplace transform evaluates to*

$$\mathscr{L}[u](z) = \int_0^\infty e^{(c-z)t}\,dt = \frac{e^{(c-z)t}}{c-z}\bigg|_0^\infty = \frac{1}{z-c}$$

for all $z \in \mathbb{C}$ with $\Re z > c$.

We cite here some well-known properties of $\mathscr{L}$.

Lemma 2.20. *Let u_1 and u_2 denote two functions defined on $\mathbb{R}_0^+$ such that their Laplace transform exists.*

1. *The Laplace transform is a linear operator, i.e.,*

$$\mathscr{L}[cu_1 + u_2](z) = c\mathscr{L}[u_1](z) + \mathscr{L}[u_2](z)$$

   for all $c \in \mathbb{R}$.

2. *There holds*

$$\mathscr{L}[u_1 * u_2](z) = \mathscr{L}[u_1](z)\mathscr{L}[u_2](z). \tag{2.7}$$

3. *If $U_1(t) = \int_0^T u_1(t)\,dt$, then*

$$\mathscr{L}[U_1](z) = \frac{1}{z}\mathscr{L}[u_1](z).$$

4. *For all $k \in \mathbb{N}$ there holds*

$$\mathscr{L}[\partial_t^k u](z) = z^k \mathscr{L}[u](z) - \sum_{j=1}^{k} z^{k-j}\partial_t^{k-j}u(0). \tag{2.8}$$

The function u can be recovered from $\mathscr{L}[u]$ by the *inverse Laplace transform*, also known as *Bromwich integral*, *Fourier-Mellin integral*, or *Mellin's inverse formula*.

Theorem 2.21. *Let u satisfy (2.6) and $\gamma \in \mathbb{R}$ be larger than the real part of all singularities of $U(z) := \mathscr{L}[u](z)$. Then there holds*

$$u(t) = \mathscr{L}^{-1}[U](t) := \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} e^{zt}U(z)\,dz.$$

2.3.2 The Gamma Function

In integer-order calculus the factorial function plays prominent role because it allows one to succinctly comprehend the derivatives of several elementary functions. The *Gamma function* can be seen as counterpart to the factorial function in fractional-order calculus and is defined by

$$\Gamma(\lambda) := \int_0^\infty e^{-\zeta}\zeta^{\lambda-1}\,d\zeta, \qquad \lambda \in \mathbb{R}^+. \tag{2.9}$$

Elementary considerations of the theory of improper integrals reveal that the integral exists. The importance of Γ in fractional calculus is laid out in the following theorem.

Theorem 2.22. *For all $\lambda \in \mathbb{R}^+$ there holds $\Gamma(\lambda+1) = \lambda\Gamma(\lambda)$.*

Proof. This follows directly from the integration by parts formula, since

$$\begin{aligned}\Gamma(\lambda+1) &= \int_0^\infty e^{-\zeta}\zeta^{\lambda}\,d\zeta = \left(-e^{-\zeta}\zeta^{\lambda}\Big|_0^\infty + \lambda\int_0^\infty e^{-\zeta}\zeta^{\lambda-1}\,d\zeta\right)\\ &= \lambda\int_0^\infty e^{-\zeta}\zeta^{\lambda-1}\,d\zeta = \lambda\Gamma(\lambda).\end{aligned}$$

□

Since

$$\Gamma(1) = \int_0^\infty e^{-\zeta}\,d\zeta = 1,$$

it follows by induction that

$$\forall n \in \mathbb{N}: \quad \Gamma(n+1) = n!.$$

Therefore, the Gamma function can be seen as a natural generalization of the factorial function. Other useful values of the Gamma function are

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \qquad \Gamma(\lambda)\Gamma(1-\lambda) = \frac{\pi}{\sin(\pi\lambda)}, \qquad \lambda \in (0,1).$$

An extension of Theorem 2.22 is obtained by the equivalent identity

$$\Gamma(\lambda) = \frac{\Gamma(\lambda+1)}{\lambda}, \qquad \lambda \in \mathbb{R}^+. \tag{2.10}$$

Note that the right-hand side of (2.10) is meaningful not only for $\lambda \in \mathbb{R}^+$ but also if $\lambda \in (-1,0)$. Consequently, we may use (2.10) as definition for $\Gamma(\lambda)$ whenever $\lambda \in (-1,0)$. The latter is not included in the original definition (2.9) since the integral does not converge on $\mathbb{R}_0^-$. Provided this extended definition of Γ, we may return to (2.10) to inductively define $\Gamma(\lambda)$ for all values of $\lambda \in \mathbb{R} \setminus -\mathbb{N}_0$. The function so obtained is again called Gamma function, abbreviated with the same letter Γ, and is plotted in Figure 2.1. Its poles are given by the negative natural numbers including zero. For the later use, we compute the Laplace transform of the power function.

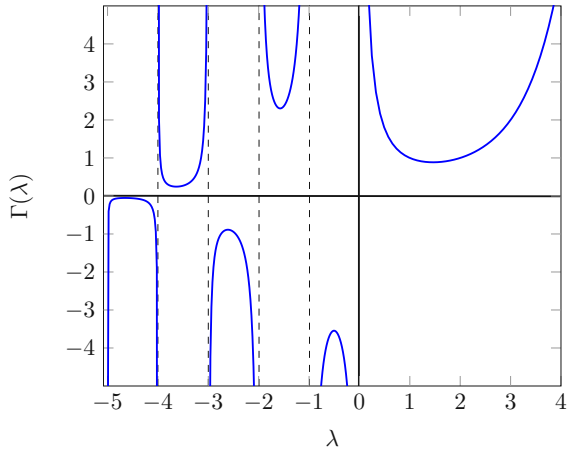

Figure 2.1: Gamma function $\Gamma(\lambda)$ on $(-5, 4] \setminus \{-4, -3, -2, -1, 0\}$.

Lemma 2.23. *Let $\alpha > -1$ and $u(t) = t^\alpha$. Then there holds for all $z \in \mathbb{C}$ with $\Re z > 0$*

$$\mathscr{L}[u](z) = \frac{\Gamma(1+\alpha)}{z^{1+\alpha}}.$$

Proof. This follows directly from the substitution $\zeta = zt$ since

$$\mathscr{L}[u](z) = \int_0^\infty e^{-zt} t^\alpha \, dt = \frac{1}{z^{\alpha+1}} \int_0^\infty e^{-\zeta} \zeta^\alpha \, d\zeta = \frac{\Gamma(1+\alpha)}{z^{1+\alpha}}. \qquad \square$$

One final result that we mention explicitly is the following integral identity, which is known as *Euler's integral of the first kind* or *Euler's Beta function*, see e.g., [Die10, Theorem D.6].

Lemma 2.24. *Let $\alpha, \beta \in \mathbb{R}^+$. Then there holds*

$$\int_0^1 (1-\zeta)^{\alpha-1} \zeta^{\beta-1} \, d\zeta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

2.3.3 The Mittag-Leffler Function

The Mittag-Leffler function has been considered in [Mai20] as "queen function" of fractional calculus. Its importance in fractional differential equations is comparable to the one of the exponential function in the theory of ordinary differential equations. Its origin goes back to the work of the Swedish mathematician Mittag-Leffler [ML03], who defined the Mittag-Leffler function as

$$E_\alpha(z) := \sum_{j=0}^\infty \frac{z^j}{\Gamma(\alpha j + 1)}, \qquad \alpha \in \mathbb{R}^+,$$

for all $z \in \mathbb{C}$. The *generalized* or *two-parameter Mittag-Leffler* function has been introduced 50 years later [Aga53] and reads

$$E_{\alpha,\beta}(z) := \sum_{j=0}^{\infty} \frac{z^j}{\Gamma(\alpha j + \beta)}, \qquad \alpha, \beta \in \mathbb{R}^+, \tag{2.11}$$

for all $z \in \mathbb{C}$. Note that

- $E_{\alpha,\beta}(0) = 1/\Gamma(\beta)$,
- $E_{\alpha,1}(z) = E_\alpha(z)$,
- $E_{1,1}(z) = E_1(z) = e^z$.

Therefore, E_α and $E_{\alpha,\beta}$ can be seen as generalizations of the exponential function. For some other choices of the parameters, the (generalized) Mittag-Leffler function can be expressed in terms of elementary functions, such as [Pod99, p. 17-18]

$$E_2(-z^2) = \cos(z), \qquad E_2(z^2) = \cosh(z), \qquad E_{1,2}(z^2) = \frac{e^z - 1}{z}.$$

Each of the functions listed above is an entire function. As the following theorem shows, the same applies to any admissible configuration of α and β.

Lemma 2.25. *Let $\alpha, \beta \in \mathbb{R}^+$ and $z \in \mathbb{C}$. Then the power series (2.11) defining $E_{\alpha,\beta}(z)$ is absolutely convergent.*

Proof. See [Die10, Theorem 4.1]. □

Remark 2.26. *By direct substitution in (2.11), it is possible to define $E_{\alpha,\beta}$ also for $\alpha = 0$, which yields a power series with finite convergence radius. If $\beta \in \mathbb{R}^+$, there holds for all $z \in \mathbb{C}$ with $|z| < 1$*

$$E_{0,\beta}(z) := \sum_{j=0}^{\infty} \frac{z^j}{\Gamma(\beta)} = \frac{1}{\Gamma(\beta)} \sum_{j=0}^{\infty} z^j = \frac{1}{\Gamma(\beta)} \frac{1}{1-z}. \tag{2.12}$$

In the context of fractional differential equations, it turns out to be fruitful to see (2.12) as motivation to define

$$E_{0,\beta}(\lambda) := \frac{1}{\Gamma(\beta)} \frac{1}{1-\lambda}, \qquad \lambda \in \mathbb{R}_0^-.$$

Remark 2.27. *We mention that further generalizations of $E_{\alpha,\beta}$ exist in terms of the three-parameter Millag-Leffler function. However, the present scope of presentation is sufficient for the purpose of this thesis.*

To provide an intuitive illustration of the generalized Mittag-Leffler function, we plot $E_{\alpha,\beta}$ on the real line for different values of α and β in Figure 2.2. Unlike the classical exponential function, we note that $E_{\alpha,\beta}$ might be negative for some values of its arguments.

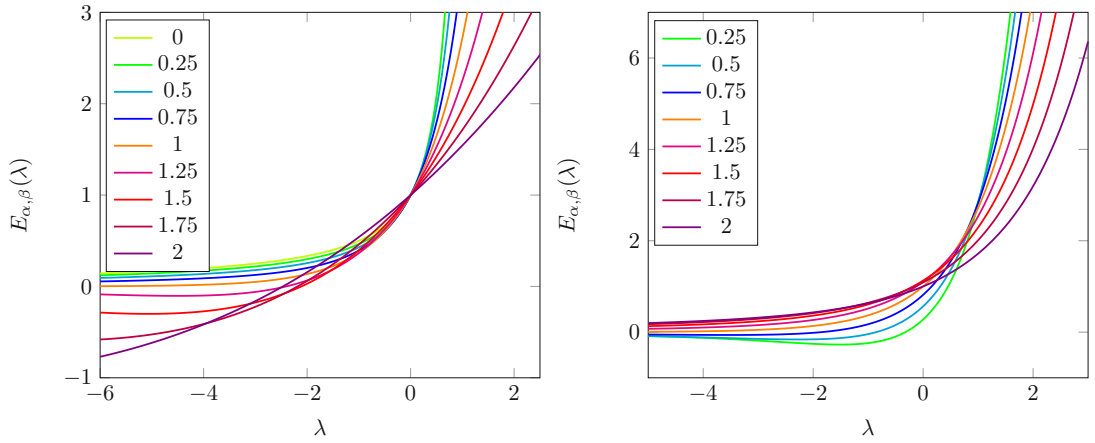

Figure 2.2: Generalized Mittag-Leffler function on the real line for $\alpha \in [0, 2]$, in the sense of Remark 2.26, and $\beta = 1$ (left) and $\alpha = 1$ and $\beta \in [0.25, 2]$ (right).

Remark 2.28. *We highlight that the numerical implementation of the generalized Mittag-Leffler function is a nontrivial task in itself. Stable algorithms that allow one to evaluate $E_{\alpha,\beta}$ for all admissible values of α and β have been proposed only recently and can be found in [WT07, PK09, Gar15].*

The asymptotic behaviour of $E_{\alpha,\beta}(z)$ in $\mathbb{C}$ is of importance in the analysis of fractional evolution equations. In case of the exponential function, it is well-known that for $z = re^{i\phi}$,

1. $e^z \to 0$ for $r \to \infty$ if $|\phi| > \frac{\pi}{2}$,
2. e^z remains bounded for $r \to \infty$ if $|\phi| = \frac{\pi}{2}$,
3. $|e^z| \to \infty$ for $r \to \infty$ if $|\phi| < \frac{\pi}{2}$.

The following theorem can be seen as generalization of the third point of the result above, where we write $\arg(z)$ to denote the argument ϕ of $z = re^{i\phi} \in \mathbb{C} \setminus \{0\}$. For simplicity, we restrict ourselves to the case $\alpha < 2$.

Theorem 2.29. *Let $\alpha \in (0, 2)$, $\beta \in \mathbb{R}^+$, $(z_n)_{n\in\mathbb{N}}$ a sequence in $S_\alpha := \{z \in \mathbb{C} : \arg(z) > \frac{\alpha\pi}{2}\}$ with $|z_n| \to \infty$. Then there holds $E_{\alpha,\beta}(z_n) \to 0$. Moreover, there exists some constant $c_{\alpha,\beta} \in \mathbb{R}^+$, only depending on α and β, such that*

$$\forall z \in S_\alpha : \quad |E_{\alpha,\beta}(z)| \leq \frac{c_{\alpha,\beta}}{1 + |z|}.$$

Proof. See [Pod99, Theorem 1.6]. □

The sectoral domain S_α from Theorem 2.29 is depicted in Figure 2.3. We see that for small values of α the area in which $E_{\alpha,\beta}(z)$ converges to zero, as $|z| \to \infty$, increases. A consequence of this observation is the following technical lemma, which turns out to be important for our analysis in Chapter 10.

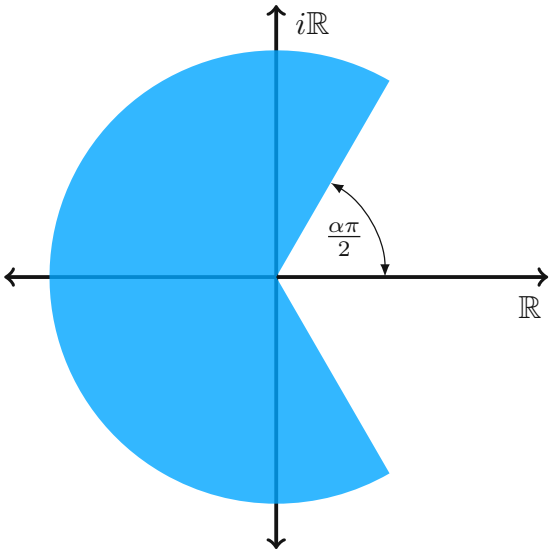

Figure 2.3: Sectoral domain S_α from Theorem 2.29 marked in blue.

Lemma 2.30. *Let $a \in \mathbb{R}^+$ be fixed, $\boldsymbol{\tau} = (\alpha, \beta, t, s) \in (0,1] \times \mathbb{R}^+ \times \mathbb{R}^+ \times (0,1]$ with $\beta \geq \alpha$, $s + \alpha < 2$, and $c_{\alpha,\beta}$ as in Theorem 2.29. Then there holds*

$$\int_{i\mathbb{R}} \left| \frac{E_{\alpha,\beta}(-t^\alpha z^s)}{a+z} \right| dz \leq 2c_{\boldsymbol{\tau}} := 2c_{\alpha,\beta} \left(a^{-1} + s^{-1} \ln(1 + t^{-\alpha}) \right).$$

Proof. Let $z \in i\mathbb{R} \setminus \{0\}$ with $\Im z > 0$ such that $z = re^{\frac{i\pi}{2}}$ for some $r \in \mathbb{R}^+$. Then there holds

$$\arg(t^\alpha z^s) = \arg\left(t^\alpha r^s e^{\frac{i\pi s}{2}} \right) = \frac{\pi s}{2}.$$

Therefore, $\arg(-t^\alpha z^s) = \pi(1 - \frac{s}{2})$. Theorem 2.29 reveals

$$|E_{\alpha,\beta}(-t^\alpha z^s)| \leq \frac{c_{\alpha,\beta}}{1 + t^\alpha |z|^s} \tag{2.13}$$

provided that $\pi(1 - \frac{s}{2}) > \frac{\alpha\pi}{2}$. The latter is satisfied since $\alpha + s < 2$. Analog computations show that (2.13) remains valid for $\Im z < 0$. We deduce

$$\begin{aligned}
\int_{i\mathbb{R}} \left| \frac{E_{\alpha,\beta}(-t^\alpha z^s)}{a+z} \right| dz &\leq c_{\alpha,\beta} \int_{i\mathbb{R}} \frac{1}{1 + t^\alpha |z|^s} \frac{1}{|a+z|} dz \\
&\leq 2c_{\alpha,\beta} \int_1^\infty \frac{dz}{z + t^\alpha z^{1+s}} + 2c_{\alpha,\beta} \int_0^1 \frac{dz}{a} \\
&= 2c_{\alpha,\beta} \int_1^\infty \frac{dz}{z^{1+s}(z^{-s} + t^\alpha)} + \frac{2c_{\alpha,\beta}}{a}.
\end{aligned}$$

Employing the substitution $\xi = z^{-s} + t^\alpha$ reveals

$$\int \frac{dz}{z^{1+s}(z^{-s} + t^\alpha)} = -\frac{\ln(z^{-s} + t^\alpha)}{s} \xrightarrow{z \to \infty} -\frac{\ln(t^\alpha)}{s}$$

and thus

$$\int_{i\mathbb{R}} \left| \frac{E_{\alpha,\beta}(-t^\alpha z^s)}{a+z} \right| dz \leq \frac{2c_{\alpha,\beta}}{s} \left(-\ln(t^\alpha) + \ln(1+t^\alpha)\right) + \frac{2c_{\alpha,\beta}}{a}$$
$$= 2c_{\alpha,\beta} \left(\frac{1}{s} \ln\left(\frac{1+t^\alpha}{t^\alpha} \right) + \frac{1}{a} \right),$$

which directly implies the conjecture. □

Two further properties that we cite here are the following well-known results that shall prove convenient in the further course of action.

Lemma 2.31. *Let $\alpha, \beta \in \mathbb{R}^+$, $t, \lambda \in \mathbb{R}_0^+$, $k \in \mathbb{N}_0$, $E^{(k)}_{\alpha,\beta}(t) = \partial_t^k E_{\alpha,\beta}(t)$, and $u(t) = t^{\alpha k+\beta-1} E^{(k)}_{\alpha,\beta}(-t^\alpha \lambda)$. Then there holds*

$$\mathscr{L}[u](z) = \frac{k! z^{\alpha-\beta}}{(z^\alpha + \lambda)^{k+1}}, \qquad \Re z > \lambda^{\frac{1}{\alpha}}.$$

Proof. See [Pod99, eq. (1.80)]. □

Lemma 2.32. *There holds for all $i \in \mathbb{N}_0$*

$$\int_0^t (t-\tau)^{\alpha-1} E_{\alpha,\alpha}(-(t-\tau)^\alpha \lambda) \tau^i \, d\tau = \Gamma(i+1) t^{\alpha+i} E_{\alpha,\alpha+i+1}(-t^\alpha \lambda).$$

Proof. See [MN11, MN18] and also [Pod99, p.25]. □

2.4 Matrix Functions

The theory of matrix functions is an integral part of this thesis and therefore deserves some discussion. Our interest lies in the analysis of expressions of the form $f(\mathbf{L})\mathbf{b}$, where $\mathbf{L} \in \mathbb{R}^{N\times N}$, $N \in \mathbb{N}$, is a matrix, $\mathbf{b} \in \mathbb{R}^N$ a vector, and f a complex-valued function. Several possibilities exist to introduce such functions of matrices [Hig08, HJ91]. For the scope of this thesis, we shall always assume that $\mathbf{L}$ is diagonalizable, in which case the following definition is a straightforward one.

Definition 2.33. *Let $\mathbf{L}$ be a diagonalizable matrix such that $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1}$ for some invertible $\mathbf{U} \in \mathbb{R}^{N\times N}$ and $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$. Assume that f is analytic in a neighborhood of the spectrum of $\mathbf{L}$. Then*

$$f(\mathbf{L}) := \mathbf{U} f(\mathbf{D}) \mathbf{U}^{-1}, \qquad f(\mathbf{D}) := \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_N)).$$

Clearly, $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of $\mathbf{L}$ and $\mathbf{U}$ can be choosen as matrix whose columns contain the eigenvectors of $\mathbf{L}$. A few immediate consequences that shall be useful in the sequel are listed in the following lemma.

Lemma 2.34. *Let f and g denote two functions that are analytic in a neighborhood of the spectrum of $\mathbf{L}$. Then there holds*

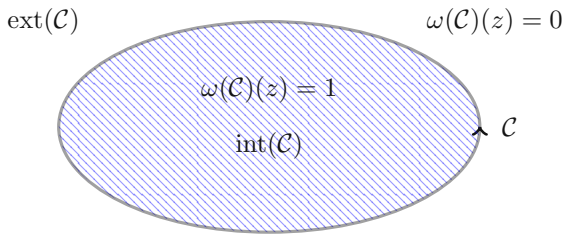

Figure 2.4: Contour $\mathcal{C}$ and its interior $\operatorname{int}(\mathcal{C})$, hatched in blue, where $\omega(\mathcal{C})(z) = 1$ holds. For all z in the unbounded component $\operatorname{ext}(\mathcal{C})$ we have $\omega(\mathcal{C})(z) = 0$.

1. $af(\mathbf{L}) + g(\mathbf{L}) = (af + g)(\mathbf{L})$ *for all* $a \in \mathbb{C}$,
2. $f(\mathbf{L})g(\mathbf{L}) = (fg)(\mathbf{L}) = g(\mathbf{L})f(\mathbf{L})$,
3. *the spectrum of* $f(\mathbf{L})$ *coincides with* $\{f(\lambda_1), \dots, f(\lambda_N)\}$.

One of the key ingredients in our analysis is the *Cauchy integral formula for matrix functions*. Before we state its result, we first present its well-known scalar variant, which requires some further terminology from complex integration [Hen93, Rud74]. A nonempty open subset of the complex plane is called *region*. If a region is connected, it is said to be a *component*. An *integration contour* $\mathcal{C}$ is a finite union of nonintersecting piecewise regular Jordan curves [Hen93], traversed in the positive sense, whose *winding number*

$$\omega(\mathcal{C})(z) := \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{d\zeta}{z - \zeta}, \qquad z \in \mathbb{C} \setminus \mathcal{C},$$

satisfies $\omega(\mathcal{C})(z) = 1$ if z is in the *interior* $\operatorname{int}(\mathcal{C})$ *of* $\mathcal{C}$ and $\omega(\mathcal{C})(z) = 0$ if z is in the *exterior* $\operatorname{ext}(\mathcal{C})$ *of* $\mathcal{C}$. Here, the interior and exterior of $\mathcal{C}$ are defined as the bounded and unbounded component of $\mathbb{C} \setminus \mathcal{C}$, respectively. With this at hand, the scalar Cauchy integral formula now reads as follows [Hen93, Rud74], cf. Figure 2.4.

Theorem 2.35. *Let* $\mathcal{C}$ *be an integration contour and* f *a function that is analytic in* $\operatorname{int}(\mathcal{C})$ *and extends continuously to* $\mathcal{C}$. *Then there holds for all* $z \in \operatorname{int}(\mathcal{C})$

$$f(z) = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{f(\zeta)}{z - \zeta}\, d\zeta. \tag{2.14}$$

If we now define the expression

$$\int_{\mathcal{C}} f(\zeta)(\mathbf{L} - \zeta \mathbf{I})^{-1}\, d\zeta$$

as limit of Riemann sums in the the matrix norm on the respective matrix space, one can prove the following straightforward generalization of (2.14), which is also known as *Cauchy-Dunford* or *Dunford-Taylor formula for matrices* [DS88, Güt10].

Theorem 2.36. *Let* $\mathcal{C}$ *be an integration contour such that the spectrum of* $\mathbf{L}$ *is contained in* $\operatorname{int}(\mathcal{C})$. *Assume that* f *is analytic in* $\operatorname{int}(\mathcal{C})$ *and extends continuously to* $\mathcal{C}$. *Then there holds*

$$f(\mathbf{L}) = \frac{1}{2\pi i} \int_{\mathcal{C}} f(\zeta)(\mathbf{L} - \zeta \mathbf{I})^{-1}\, d\zeta.$$

Proof. Let $a < b$, $\gamma : [a, b] \to \mathbb{R}$ a parametrization of $\mathcal{C}$, $(z_j)_{j=0}^n$ a partition of $[a, b]$ with $a = z_0 < \cdots < z_n = b$, and $(\eta_j)_{j=1}^n$ a corresponding sequence of nodes with $\eta_j \in [z_{j-1}, z_j]$ for all $j \in \{1, \dots, n\}$. Then there holds

$$\begin{aligned}\int_{\mathcal{C}} f(\zeta)(\mathbf{L} - \zeta\mathbf{I})^{-1}\, d\zeta &= \int_a^b f(\gamma(\zeta))\gamma'(\zeta)(\mathbf{L} - \gamma(\zeta)\mathbf{I})^{-1}\, d\zeta \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} (z_i - z_{i-1}) f(\gamma(\eta_i))\gamma'(\eta_i)(\mathbf{L} - \gamma(\eta_i)\mathbf{I})^{-1} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} (z_i - z_{i-1}) f(\gamma(\eta_i))\gamma'(\eta_i)\mathbf{U}(\mathbf{D} - \gamma(\eta_i)\mathbf{I})^{-1}\mathbf{U}^{-1}.\end{aligned}$$

After pulling $\mathbf{U}$ and $\mathbf{U}^{-1}$ out of the sum, we arrive at

$$\frac{1}{2\pi i} \int_{\mathcal{C}} f(\zeta)(\mathbf{L} - \zeta\mathbf{I})^{-1}\, d\zeta = \mathbf{U}\hat{\mathbf{D}}\mathbf{U}^{-1}, \tag{2.15}$$

where $\hat{\mathbf{D}} := \operatorname{diag}(\hat{\lambda}_1, \dots \hat{\lambda}_N)$ with

$$\hat{\lambda}_j = \frac{1}{2\pi i} \lim_{n\to\infty} \sum_{i=1}^{n} (z_i - z_{i-1}) f(\gamma(\eta_i))\gamma'(\eta_i)(\lambda_j - \gamma(\eta_i))^{-1}.$$

Recognizing the latter as Riemann sum for the corresponding scalar function, we deduce

$$\hat{\lambda}_j = \int_a^b f(\gamma(\zeta))\gamma'(\zeta)(\lambda_j - \gamma(\zeta))^{-1}\, d\zeta = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{f(\zeta)}{\lambda - \zeta}\, d\zeta = f(\lambda_j)$$

for all $j = 1, \dots, N$, where the last equality follows from Theorem 2.35. Recalling (2.15), we thus conclude

$$\frac{1}{2\pi i} \int_{\mathcal{C}} f(\zeta)(\mathbf{L} - \zeta\mathbf{I})^{-1}\, d\zeta = \mathbf{U} f(\mathbf{D})\mathbf{U}^{-1} = f(\mathbf{L}). \qquad \square$$

Theorem 2.36 remains valid if we deform the contour to the negative real line, in which case, under suitable regularity assumptions on f, there exists some real-valued function $\mu : \mathbb{R}^+ \to \mathbb{R}$ such that

$$f(\mathbf{L}) = \int_0^\infty \mu(\zeta)(\mathbf{L} + \zeta\mathbf{I})^{-1}\, d\zeta.$$

We state this result in a slightly more general version in the following theorem, where we replace $(\mathbf{L}+\zeta\mathbf{I})^{-1}$ with a generic matrix kernel $g(\mathbf{L}, \zeta)$. Its proof follows in complete analogy to the one of Theorem 2.36.

Theorem 2.37. *Assume that $f : \mathbb{R}^+ \to \mathbb{R}$ satisfies an integral representation of the form*

$$f(\lambda) = \int_0^\infty \mu(\zeta) g(\lambda, \zeta)\, d\zeta,$$

where $\mu : \mathbb{R}^+ \to \mathbb{R}$ and $g : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ are functions such that the integral is absolutely convergent. Then there holds

$$f(\mathbf{L}) = \int_0^\infty \mu(\zeta) g(\mathbf{L}, \zeta)\, d\zeta.$$

3 Abstract Interpolation Theory

One of our main interests lies in the study of fractional PDEs of the form

$$\begin{aligned} (-\Delta)^s u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{3.1}$$

where $s \in (0,1)$ and $f \in L^2(\Omega)$. Before we give a precise definition of the operator $(-\Delta)^s$, we need to specify its domain of definition. In the limit $s = 1$, we have $(-\Delta)^1 = -\Delta$, in which case we may multiply the PDE in (3.1) with a test function $v \in H_0^1(\Omega)$, integrate over Ω, and apply integration by parts to observe

$$\forall v \in H_0^1(\Omega): \quad \int_\Omega \nabla u \cdot \nabla v \, d\mathbf{x} = \int_\Omega f v \, d\mathbf{x}.$$

Standard results from elliptic PDEs can be applied to see that this weak formulation has a unique solution $u \in H_0^1(\Omega)$, whence $H_0^1(\Omega)$ is the right Hilbert space to study $(-\Delta)^s$ if $s = 1$. On the other hand, if we define $(-\Delta)^0 := \mathrm{I}$ to be the identity operator and neglect the boundary conditions, (3.1) reduces to $u = f \in L^2(\Omega)$. Therefore, $L^2(\Omega)$ provides the natural domain of definition of the fractional Laplacian if $s = 0$. For $s \in (0,1)$, one can expect that the domain of $(-\Delta)^s$ lies in between $L^2(\Omega)$ and $H_0^1(\Omega)$. The purpose of this chapter is to provide a mathematical framework for these purely heuristic considerations, which leads us to the study of interpolation spaces.

3.1 Interpolation Spaces

Interpolation spaces are a classical field of study [LM72, BL76, Tri78, Tar07] and can be defined between any two Banach spaces $(\mathcal{B}_0, \|\cdot\|_0)$ and $(\mathcal{B}_1, \|\cdot\|_1)$ that are linear subspaces of some larger vector space. One considers the spaces $\mathcal{B}_0 \cap \mathcal{B}_1$ and $\mathcal{B}_0 + \mathcal{B}_1$, which are, equipped with the norms

$$\begin{aligned} \|u\|_{\mathcal{B}_0 \cap \mathcal{B}_1} &:= \max\{\|u\|_0, \|u\|_1\}, \\ \|u\|_{\mathcal{B}_0 + \mathcal{B}_1} &:= \inf\{\|u_0\|_0 + \|u_1\|_1 : u = u_0 + u_1, u_0 \in \mathcal{B}_0, u_1 \in \mathcal{B}_1\}, \end{aligned}$$

Banach spaces themselves. A Banach space $\mathcal{B}$ that satisfies $\mathcal{B}_0 \cap \mathcal{B}_1 \subset \mathcal{B} \subset \mathcal{B}_0 + \mathcal{B}_1$ with continuous embeddings is said to be an interpolation space.

The subject simplifies if $\mathcal{B}_1 \subset \mathcal{B}_0$ are Hilbert spaces with dense and continuous embedding, in which case $\mathcal{B}_0 \cap \mathcal{B}_1 = \mathcal{B}_1$ and $\mathcal{B}_0 + \mathcal{B}_1 = \mathcal{B}_0$; see [CWHM15] for a detailed study of this particular setting. If the embedding is also compact, the Hilbert spaces are separable and one can resort to a countable basis of $\mathcal{B}_1$, which, due to density, also forms a basis of $\mathcal{B}_0$. These assumptions are sufficient for the purpose of our investigations which is why we include them in our definition of admissibility.

Definition 3.1. *A pair of Hilbert spaces $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ is called interpolation couple if $\mathcal{H}_1$ is dense in $\mathcal{H}_0$ and the embedding $\mathcal{H}_1 \subset \mathcal{H}_0$ is compact.*

The literature provides a large variety of different interpolation methods. A very natural one is based on spectral decomposition [LM72, Bra93, CWHM15].

3.1.1 Spectral Interpolation

For the sake of a more compact notation, let $(\cdot,\cdot)_i$ and $\|\cdot\|_i$ denote the inner products and norms on $\mathcal{H}_i$, $i = 0, 1$, respectively, and $\mathcal{H}_i'$ its dual space. Provided that $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ is an interpolation couple, there holds $\mathcal{H}_1 \subset \mathcal{H}_0$ and therefore $\mathcal{H}_0' \subset \mathcal{H}_1'$. Upon identifying $\mathcal{H}_0$ with its dual space, this yields the chain of inclusions

$$\mathcal{H}_1 \subset \mathcal{H}_0 \cong \mathcal{H}_0' \subset \mathcal{H}_1'. \tag{3.2}$$

It shall prove convenient to relabel $\mathcal{H}_{-1} := \mathcal{H}_1'$ and no longer distinguish between $\mathcal{H}_0$ and $\mathcal{H}_0'$ henceforth, i.e., we identify each function $f \in \mathcal{H}_0$ with the functional $v \mapsto (f, v)_0$, which we call f again, such that

$$\forall v \in \mathcal{H}_0 : \quad \langle f, v\rangle = (f, v)_0, \tag{3.3}$$

where $\langle\cdot,\cdot\rangle$ denotes the dual pairing. Under these premises, the triple $(\mathcal{H}_1, \mathcal{H}_0, \mathcal{H}_{-1})$ is said to be a *Gelfand triple*. The space $\mathcal{H}_{-1}$ equipped with

$$(g, f)_{-1} := (g, f)_{\mathcal{H}_{-1}} = \langle g, \mathcal{R}f\rangle, \qquad \|f\|_{-1} := \|f\|_{\mathcal{H}_{-1}} = \sqrt{(f, f)_{-1}} = \sup_{v\in\mathcal{H}_1} \frac{\langle f, v\rangle}{\|v\|_{\mathcal{H}_1}},$$

is a Hilbert space, where $\mathcal{R} : \mathcal{H}_{-1} \to \mathcal{H}_1$ denotes the Riesz isomorphism which identifies $\mathcal{H}_1$ with its dual space by

$$(\mathcal{R}f, v)_1 = \langle f, v\rangle, \qquad f \in \mathcal{H}_{-1}, v \in \mathcal{H}_1. \tag{3.4}$$

The following observation serves as starting point for further discussions and follows from standard arguments for self-adjoint and compact operators, cf. [CWHM15, Theorem 3.4].

Theorem 3.2. *The Riesz isomorphism $\mathcal{R}$ is self-adjoint and compact. There exists an orthogonal basis $(\tilde\varphi_j)_{j=1}^\infty$ of $\mathcal{H}_1$ where each $\tilde\varphi_j$ is an eigenfunction of $\mathcal{R}$ to the eigenvalue $\tilde\lambda_j$. There holds $\tilde\lambda_1 \geq \tilde\lambda_2 \geq \cdots > 0$ and $\tilde\lambda_j \to 0$ as $j \to \infty$.*

For convenience, we introduce the normalized family of eigenfunctions $\varphi_j := \tilde\varphi_j/\|\tilde\varphi_j\|_0 \in \mathcal{H}_1$ and set $\lambda_j := \tilde\lambda_j^{-1}$. By construction, we have $\varphi_j = \lambda_j\mathcal{R}\varphi_j$. Together with (3.4) and (3.3) we find

$$\forall v \in \mathcal{H}_1 : \quad (\varphi_j, v)_1 = \lambda_j(\mathcal{R}\varphi_j, v)_1 = \lambda_j\langle\varphi_j, v\rangle = \lambda_j(\varphi_j, v)_0$$

for all $j \in \mathbb{N}$. Since $\mathcal{H}_1$ is dense in $\mathcal{H}_0$, $(\varphi_j)_{j=1}^\infty$ is an orthonormal basis of $\mathcal{H}_0$ and the following result is valid.

Corollary 3.3. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple. Then there exists an orthonormal basis $(\varphi_j)_{j=1}^{\infty}$ of $\mathcal{H}_0$ and a sequence $0 < \lambda_1 < \lambda_2 < \dots$ with the property $\lambda_j \to \infty$ as $j \to \infty$ such that*

$$\forall v \in \mathcal{H}_1 : \quad (\varphi_j, v)_1 = \lambda_j (\varphi_j, v)_0.$$

Throughout what follows, let $u_j := (\varphi_j, u)_0$ denote the spectral components of $u \in \mathcal{H}_0$. Then

$$\|u\|_0^2 = (\sum_{j=1}^{\infty} u_j \varphi_j, \sum_{i=1}^{\infty} u_i \varphi_i)_0 = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} u_j u_i (\varphi_j, \varphi_i)_0 = \sum_{j=1}^{\infty} u_j^2 \tag{3.5}$$

for any $u \in \mathcal{H}_0$. Similarly, if $u \in \mathcal{H}_1$, there holds

$$\|u\|_1^2 = (\sum_{j=1}^{\infty} u_j \varphi_j, \sum_{i=1}^{\infty} u_i \varphi_i)_1 = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} u_j u_i (\varphi_j, \varphi_i)_1 = \sum_{j=1}^{\infty} \lambda_j u_j^2. \tag{3.6}$$

The following definition should now come as no surprise.

Definition 3.4. *Let $s \in [0,1]$ and $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ an interpolation couple. We define the interpolation norm $\|\cdot\|_{\overline{\mathcal{H}}^s}$ of $\overline{\mathcal{H}}$ by*

$$\|u\|_{\overline{\mathcal{H}}^s}^2 := \sum_{j=1}^{\infty} \lambda_j^s u_j^2, \qquad u_j = (\varphi_j, u)_0.$$

The interpolation space of $\overline{\mathcal{H}}$ is defined by

$$[\mathcal{H}_0, \mathcal{H}_1]_s := \{u \in \mathcal{H}_0 : \|u\|_{\overline{\mathcal{H}}^s} < \infty\}.$$

For all $s \in [0,1]$ the norm $\|\cdot\|_{\overline{\mathcal{H}}^s}$ comes from an inner product. It reads

$$(u, v)_{\overline{\mathcal{H}}^s} := \sum_{j=1}^{\infty} \lambda_j^s u_j v_j, \tag{3.7}$$

which makes $[\mathcal{H}_0, \mathcal{H}_1]_s$ a Hilbert space itself. For obvious reasons, we call (3.7) the *interpolation scalar product on* $[\mathcal{H}_0, \mathcal{H}_1]_s$. By construction, the interpolation space is strictly intermediate

$$\mathcal{H}_1 \subsetneq [\mathcal{H}_0, \mathcal{H}_1]_s \subsetneq \mathcal{H}_0$$

for all $s \in (0,1)$ and satisfies the interpolation condition $[\mathcal{H}_0, \mathcal{H}_1]_0 = \mathcal{H}_0$ and $[\mathcal{H}_0, \mathcal{H}_1]_1 = \mathcal{H}_1$ with equality of the norms. Several other properties of $[\mathcal{H}_0, \mathcal{H}_1]_s$ that match our understanding of interpolation are listed in the following lemma.

Lemma 3.5. *Let $0 < s_1 < s_2 < 1$ and $s \in [0,1]$. Then*

1. *$([\mathcal{H}_0, \mathcal{H}_1]_{s_1}, [\mathcal{H}_0, \mathcal{H}_1]_{s_2})$ is an interpolation couple,*

2. *there holds the reiteration property*

$$[[\mathcal{H}_0, \mathcal{H}_1]_{s_1}, [\mathcal{H}_0, \mathcal{H}_1]_{s_2}]_s = [\mathcal{H}_0, \mathcal{H}_1]_{(1-s)s_1+ss_2}$$

with equivalent norms,

3. *there holds for all* $u \in \mathcal{H}_1$

$$\|u\|_{\overline{\mathcal{H}}^s} \preceq \|u\|_0^s \|u\|_1^{1-s}.$$

Proof. See Proposition 6.1, Theorem 6.1, and Proposition 2.3 in [LM72], respectively. □

In view of (3.2), it is a natural question to ask whether the pairing $(\mathcal{H}_{-1}, \mathcal{H}_0)$ is admissible for space interpolation. We proceed in this direction with a positive result; cf. [Tar07, Chapter 41].

Proposition 3.6. *If* $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ *is an interpolation couple, then* $\overline{\mathcal{H}}' := (\mathcal{H}_{-1}, \mathcal{H}_0)$ *is an interpolation couple as well.*

Proposition 3.6 allows us to form the interpolation space of the dual pairing $\overline{\mathcal{H}}' = (\mathcal{H}_{-1}, \mathcal{H}_0)$, which raises the question how $[\mathcal{H}_{-1}, \mathcal{H}_0]_s$ relates to the interpolation space of the original interpolation couple $\overline{\mathcal{H}}$. To elaborate this in detail, let $(\varphi_j')_{j=1}^\infty \subset \mathcal{H}_0$ denote the system of $\mathcal{H}_{-1}$-orthonormal eigenfunctions with eigenvalues $(\lambda_j')_{j=1}^\infty$ associated to $\overline{\mathcal{H}}'$ in the sense of Corollary 3.3, such that

$$\forall v \in \mathcal{H}_0 : \quad (\varphi_j', v)_0 = \lambda_j'(\varphi_j', v)_{-1}.$$

Then the dual eigenpairs $(\varphi_j', \lambda_j')_{j=1}^\infty$ can be expressed in terms of the eigenpairs of the original interpolation couple $(\varphi_j, \lambda_j)_{j=1}^\infty$ in the following convenient manner; cf. [BP15, DS21].

Theorem 3.7. *For all* $j \in \mathbb{N}$ *there holds*

$$\varphi_j' = \sqrt{\lambda_j}\varphi_j, \qquad \lambda_j' = \lambda_j.$$

Proof. Following [BP15, Proposition 4.1], we invoke (3.3), (3.4), and Corollary 3.3 to observe for any $v \in \mathcal{H}_0$

$$(\varphi_j, v)_0 = \langle \varphi_j, v \rangle = (\varphi_j, \mathcal{R}v)_1 = \lambda_j(\varphi_j, \mathcal{R}v)_0 = \lambda_j(\varphi_j, v)_{-1}. \tag{3.8}$$

Choosing $v = \varphi_i$ for some $i \in \mathbb{N}$ we find

$$(\varphi_j, \varphi_i)_0 = (\sqrt{\lambda_j}\varphi_j, \sqrt{\lambda_j}\varphi_i)_{-1},$$

which shows that $(\sqrt{\lambda_j}\varphi_j)_{j=1}^\infty$ is a $\mathcal{H}_{-1}$-orthonormal system of eigenfunctions. Since

$$0 = (\varphi_j, v)_{-1} = \lambda_j^{-1}(\varphi_j, v)_0, \qquad v \in \mathcal{H}_0,$$

for all $j \in \mathbb{N}$ implies that $v = 0$, it is also a basis and the proof is complete. □

Theorem 3.7 is the key ingredient to show that the interpolation norm of the dual interpolation couple may be obtained from extrapolating the interpolation norm of the original one. Here and throughout the remainder of this chapter, we write $\|\cdot\|_{\overline{\mathcal{H}}'^s}$ to denote the interpolation norm of the dual interpolation couple $\overline{\mathcal{H}}' = (\mathcal{H}_{-1}, \mathcal{H}_0)$.

Theorem 3.8. *Let $s \in [0,1]$ and $f \in [\mathcal{H}_{-1}, \mathcal{H}_0]_{1-s}$. Then there holds*

$$\|f\|^2_{\overline{\mathcal{H}}'^{1-s}} = \sum_{j=1}^{\infty} \lambda_j^{-s} \langle f, \varphi_j \rangle^2. \tag{3.9}$$

Proof. It follows from (3.8) combined with (3.3) that

$$(f, \varphi_j)_{-1} = \frac{1}{\lambda_j} \langle f, \varphi_j \rangle. \tag{3.10}$$

Thanks to Theorem 3.7, we conclude

$$\|f\|^2_{\overline{\mathcal{H}}'^{1-s}} = \sum_{j=1}^{\infty} (\lambda'_j)^{1-s} (f, \varphi'_j)^2_{-1} = \sum_{j=1}^{\infty} \lambda_j^{2-s} (f, \varphi_j)^2_{-1} = \sum_{j=1}^{\infty} \lambda_j^{-s} \langle f, \varphi_j \rangle^2. \qquad \square$$

One readily verifies that the right-hand side of (3.9) coincides with the interpolation norm on $[\mathcal{H}_0, \mathcal{H}_1]'_s$, so that the following result is valid; see [LM72, Theorem 6.2] and [CWHM15].

Corollary 3.9. *For all $s \in [0,1]$ there holds*

$$[\mathcal{H}_0, \mathcal{H}_1]'_s = [\mathcal{H}_{-1}, \mathcal{H}_0]_{1-s}$$

and their norms coincide.

The discussion above allows us to interpret $[\mathcal{H}_0, \mathcal{H}_1]_s$ and $[\mathcal{H}_{-1}, \mathcal{H}_0]_s$ as one single scale of interpolation spaces

$$\mathcal{H}_s := \{ f \in \mathcal{H}_{-1} : \sum_{j=1}^{\infty} \lambda_j^s \langle f, \varphi_j \rangle < \infty \}$$

for all $s \in [-1,1]$; cf. [BP15]. We therefore define

$$[\mathcal{H}_0, \mathcal{H}_1]_{-s} := [\mathcal{H}_0, \mathcal{H}_1]'_s = [\mathcal{H}_{-1}, \mathcal{H}]_{1-s} \tag{3.11}$$

for all $s \in [0,1]$.

Remark 3.10. *As indicated at the beginning of this section, the requirements on the interpolation couple might be essentially relaxed. Many of the results presented above remain valid if the embedding $\mathcal{H}_1 \subset \mathcal{H}_0$ is only continuous. The treatment of these scenarios requires a generalized version of the spectral theorem in which case the spectrum of $\mathcal{R}$ is no longer discrete; see e.g., [LM72] for this particular setting.*

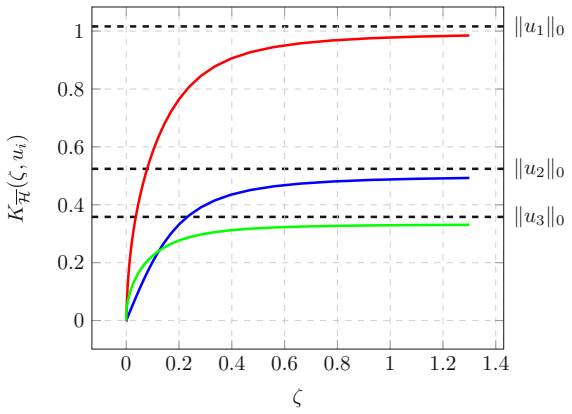

Figure 3.1: K-functional $K_{\overline{\mathcal{H}}}(\zeta, u_i)$ of $\overline{\mathcal{H}} = ((L^2(\Omega), \|\cdot\|_{L^2(\Omega)}), (H_0^1(\Omega), \|\nabla\cdot\|_{L^2(\Omega)}))$ for $u_1(\mathbf{x}) = 1$ (red), $u_2(\mathbf{x}) = \sin(\pi x)\sin(\pi y)$ (blue), and $u_3(\mathbf{x}) = xy$ (green) with $\mathbf{x} = (x, y) \in \Omega = (0, 1)^2$.

3.1.2 The K-Method

A conceptually different but equivalent approach for defining $[\mathcal{H}_0, \mathcal{H}_1]_s$ is obtained by the *K-method*, also referred to as *real method of interpolation* or *Peetre's method* [Pee63, BL76, Tri78, BS88, Bra93, McL00, Tar07, Lun09, CWHM15]. Unlike the spectral approach, the K-method does not rely on spectral theory and also works if $\mathcal{H}_0$ and $\mathcal{H}_1$ are only Banach spaces. The following definition is at the heart of its exposition.

Definition 3.11. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple. We define the K-functional of $\overline{\mathcal{H}}$ by*

$$K_{\overline{\mathcal{H}}} : \mathbb{R}^+ \times \mathcal{H}_0 \to \mathbb{R}$$
$$(\zeta, u) \mapsto K_{\overline{\mathcal{H}}}(\zeta, u) := \inf_{v \in \mathcal{H}_1} \sqrt{\|u - v\|_0^2 + \zeta^2 \|v\|_1^2}.$$

As illustrated in Figure 3.1, $K_{\overline{\mathcal{H}}}(\cdot, u)$ is a nonnegative, nondecreasing, concave, and continuous function on $\mathbb{R}^+$ for each $u \in \mathcal{H}_0$ [BS88]. Since $\mathcal{H}_1$ is dense in $\mathcal{H}_0$, we have

$$\lim_{\zeta \to 0^+} K_{\overline{\mathcal{H}}}(\zeta, u) = 0$$

and there holds the estimate

$$K_{\overline{\mathcal{H}}}(\zeta, u) \leq \min\{\|u\|_0, \zeta\|u\|_1\}. \tag{3.12}$$

Therefore,

$$\lim_{\zeta \to \infty} K_{\overline{\mathcal{H}}}(\zeta, u) = \|u\|_0.$$

The present form of the K-functional is less suited to directly access its value. A step towards a more explicit representation requires the well-known existence and uniqueness result by Lax-Milgram.

Theorem 3.12 (Lax-Milgram). *Let $\mathcal{H}$ be a Hilbert space, $f \in \mathcal{H}'$, and $a(\cdot,\cdot) : \mathcal{H} \times \mathcal{H}$ a bilinear form on $\mathcal{H}$. Assume that there exists some constants $c_1, c_2 \in \mathbb{R}^+$ such that for all $u, v \in \mathcal{H}$ there holds*

$$|a(u,v)| \leq c_1 \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}, \qquad a(u,u) \geq c_2 \|u\|_{\mathcal{H}}^2.$$

Then the problem: Find $u \in \mathcal{H}$ such that

$$\forall v \in \mathcal{H} : \quad a(u,v) = f(v)$$

has a unique solution $u \in \mathcal{H}$ and there holds the stability estimate

$$\|u\|_{\mathcal{H}} \leq \frac{1}{c_2} \|f\|_{\mathcal{H}'}.$$

Lax-Milgram's theorem allows us to derive a representation of the minimizer of $K_{\overline{\mathcal{H}}}(\zeta, u)$ in terms of the eigenpairs $(\varphi_j, \lambda_j)_{j=1}^{\infty}$ of $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$.

Lemma 3.13. *For all $\zeta \in \mathbb{R}^+$ and $u \in \mathcal{H}_0$ there exists a unique minimizer $\mathrm{v} = \mathrm{v}(\zeta, u) \in \mathcal{H}_1$ of the K-functional such that*

$$\sqrt{\|u - \mathrm{v}\|_0^2 + \zeta^2 \|\mathrm{v}\|_1^2} = \inf_{v \in \mathcal{H}_1} \sqrt{\|u - v\|_0^2 + \zeta^2 \|v\|_1^2}.$$

Moreover, v is the unique solution of the variational problem: Find $\mathrm{v} \in \mathcal{H}_1$ such that

$$\forall w \in \mathcal{H}_1 : \quad (\mathrm{v}, w)_0 + \zeta^2 (\mathrm{v}, w)_1 = (u, w)_0. \tag{3.13}$$

It can be expressed in terms of the excitations $u_j = (\varphi_j, u)_0$ of u by

$$\mathrm{v} = \sum_{j=1}^{\infty} \frac{u_j}{1 + \zeta^2 \lambda_j} \varphi_j. \tag{3.14}$$

Proof. Standard tools from calculus of variation reveal that the K-functional possesses a unique minimizer. To see that the latter satisfies (3.14), we proceed as in [Bra93, Theorem B.2] and make the ansatz

$$\mathrm{v} = \sum_{j=1}^{\infty} \mathrm{v}_j \varphi_j$$

with coefficients $(\mathrm{v}_j)_{j=1}^{\infty} \subset \mathbb{R}$. Invoking (3.5) and (3.6), we observe

$$\|u - \mathrm{v}\|_0^2 + \zeta^2 \|\mathrm{v}\|_1^2 = \sum_{j=1}^{\infty} \left((u_j - \mathrm{v}_j)^2 + \zeta^2 \lambda_j \mathrm{v}_j^2 \right). \tag{3.15}$$

Noting that each summand on the right-hand side of (3.15) is nonnegative, it follows

$$K_{\overline{\mathcal{H}}}^2(\zeta, u) = \sum_{j=1}^{\infty} \inf_{\mathrm{v}_j \in \mathbb{R}} \left((u_j - \mathrm{v}_j)^2 + \zeta^2 \lambda_j \mathrm{v}_j^2 \right).$$

For all $j \in \mathbb{N}$, the infimum is attained by $\mathrm{v}_j = u_j / (1 + \zeta^2 \lambda_j)$. We conclude that (3.14) is valid. Plugging (3.14) into (3.13) with $w = \varphi_j$ shows that v solves the variational formulation (3.13), which, according to Theorem 3.12, has a unique solution. This completes the proof. □

The value of $K_{\overline{\mathcal{H}}}(\zeta, u)$ can now be written in the following compact form.

Corollary 3.14. *Let $\zeta \in \mathbb{R}^+$, $u \in \mathcal{H}_0$, and v the unique solution to (3.13). Then there holds*

$$K^2_{\overline{\mathcal{H}}}(\zeta, u) = (u - \mathrm{v}, u)_0 = \sum_{j=1}^{\infty} \frac{\zeta^2 \lambda_j u_j^2}{1 + \zeta^2 \lambda_j}.$$

Proof. This is a direct consequence of (3.14). □

The identity (3.14) shows that $\mathrm{v} = \mathrm{v}(\zeta, u)$ is linear in u, that is,

$$\mathrm{v}(\zeta, cu + w) = c\mathrm{v}(\zeta, u) + \mathrm{v}(\zeta, w)$$

for all $c \in \mathbb{R}$ and $u, w \in \mathcal{H}_0$. Therefore

$$K^2_{\overline{\mathcal{H}}}(\zeta, cu) = \|cu - \mathrm{v}(\zeta, cu)\|_0^2 + \|\mathrm{v}(\zeta, cu)\|_1^2 = |c|\|u - \mathrm{v}(\zeta, u)\|_0^2 + |c|\|\mathrm{v}(\zeta, u)\|_1^2 = |c| K^2_{\overline{\mathcal{H}}}(\zeta, u).$$

Direct computations reveal that also

$$K_{\overline{\mathcal{H}}}(\zeta, u + w) \leq K_{\overline{\mathcal{H}}}(\zeta, u) + K_{\overline{\mathcal{H}}}(\zeta, w)$$

holds. Since $K_{\overline{\mathcal{H}}}(\zeta, u) = 0$ for all $\zeta \in \mathbb{R}^+$ if and only if $u = 0$, we conclude that $K_{\overline{\mathcal{H}}}(\zeta, \cdot)$ is a norm on $\mathcal{H}_0$. Thanks to (3.12) and the continuous embedding $\mathcal{H}_1 \subset \mathcal{H}_0$, we arrive at the following result.

Proposition 3.15. *For all $\zeta \in \mathbb{R}^+$ fixed, the functional $K_{\overline{\mathcal{H}}}(\zeta, \cdot) \in \mathcal{H}_0'$ is an equivalent norm on $\mathcal{H}_0$.*

Before we proceed with the main theorem of this section, we state a useful integral identity that can be found in [BS87, Chapter 10.4] and [Yos95, Chapter 9.11], see also [Bal60].

Lemma 3.16. *For all $s \in (0, 1)$ there holds*

$$\int_0^\infty \frac{\zeta^{-s}}{1 + \zeta} \, d\zeta = \frac{\pi}{\sin(\pi s)}.$$

We are now in position to prove the following conjecture which is the driving motivation of our interest in K-functionals (cf. [Bra93, Theorme B.2]).

Theorem 3.17. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple and $s \in (0, 1)$. Then there holds for all $u \in [\mathcal{H}_0, \mathcal{H}_1]_s$*

$$\|u\|^2_{\overline{\mathcal{H}}^s} = \frac{2\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-2s-1} K^2_{\overline{\mathcal{H}}}(\zeta, u) \, d\zeta. \tag{3.16}$$

Proof. We apply Corollary 3.14 to see that

$$\int_0^\infty \zeta^{-2s-1} K^2_{\overline{\mathcal{H}}}(\zeta, u) \, d\zeta = \int_0^\infty \sum_{j=1}^{\infty} \lambda_j u_j^2 \frac{\zeta^{1-2s}}{1 + \zeta^2 \lambda_j} \, d\zeta.$$

Each summand is nonnegative whence we may interchange the series with the integral to deduce

$$\int_0^\infty \zeta^{-2s-1} K^2_{\overline{\mathcal{H}}}(\zeta, u)\, d\zeta = \sum_{j=1}^\infty \lambda_j u_j^2 \left(\int_0^\infty \frac{\zeta^{1-2s}}{1+\zeta^2\lambda_j}\, d\zeta \right).$$

The transformation $\zeta \mapsto \zeta^2\lambda_j$ combined with Lemma 3.16 gives

$$\int_0^\infty \frac{\zeta^{1-2s}}{1+\zeta^2\lambda_j}\, d\zeta = \frac{\lambda_j^s}{2\lambda_j} \int_0^\infty \frac{\zeta^{-s}}{1+\zeta}\, d\zeta = \frac{\lambda_j^s}{2\lambda_j} \frac{\pi}{\sin(\pi s)}.$$

Hence, we finally arrive at

$$\int_0^\infty \zeta^{-2s-1} K^2_{\overline{\mathcal{H}}}(\zeta, u)\, d\zeta = \frac{\pi}{2\sin(\pi s)} \sum_{i=1}^\infty \lambda_j^s u_j^2 = \frac{\pi}{2\sin(\pi s)} \|u\|^2_{\overline{\mathcal{H}}^s}. \qquad \square$$

The square root of the integral on the right-hand side of (3.16) is often referred to as *K-norm*. We set

$$\|u\|^2_{K^s_{\overline{\mathcal{H}}}} := \frac{2\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-2s-1} K^2_{\overline{\mathcal{H}}}(\zeta, u)\, d\zeta.$$

Theorem 3.17 states that the interpolation norm and the K-norm coincide. A respective integral representation for the interpolation scalar product $(\cdot,\cdot)_{\overline{\mathcal{H}}^s}$ on $[\mathcal{H}_0, \mathcal{H}_1]_s$ is now easily derived.

Corollary 3.18. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple, $s \in (0,1)$, $u \in [\mathcal{H}_0, \mathcal{H}_1]_s$, and $\mathrm{v} = \mathrm{v}(\zeta)$ the unique solution of (3.13). Then there holds for all $w \in \mathcal{H}_0$*

$$(u, w)_{\overline{\mathcal{H}}^s} = \frac{2\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-2s-1} (u - \mathrm{v}(\zeta), w)_0\, d\zeta.$$

Proof. Due to the first identity in Corollary 3.14 there holds

$$\|u\|^2_{K_{\overline{\mathcal{H}}^s}} = \frac{2\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-2s-1} K^2_{\overline{\mathcal{H}}}(\zeta, u)\, d\zeta = \frac{2\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-2s-1} (u - \mathrm{v}(\zeta), u)_0\, d\zeta,$$

which shows that

$$(u, w)_{K_{\overline{\mathcal{H}}^s}} := \frac{2\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-2s-1} (u - \mathrm{v}(\zeta), w)_0\, d\zeta \tag{3.17}$$

satisfies $\|u\|^2_{K_{\overline{\mathcal{H}}^s}} = (u,u)_{K_{\overline{\mathcal{H}}^s}}$, i.e., (3.17) induces the K-norm. On the other hand, it follows from Theorem 3.17 that $(u,u)_{\overline{\mathcal{H}}^s} = \|u\|^2_{\overline{\mathcal{H}}^s} = \|u\|^2_{K^s_{\overline{\mathcal{H}}}} = (u,u)_{K_{\overline{\mathcal{H}}^s}}$. The claim now holds due to the polarization identity. $\square$

We highlight one final property that links the K-functional of the dual interpolation couple with the one of original pairing.

Lemma 3.19. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple and $\overline{\mathcal{H}}' = (\mathcal{H}_{-1}, \mathcal{H}_0)$. For each $\zeta \in \mathbb{R}^+$ and $f \in \mathcal{H}_{-1}$ let v' denote the minimizer of $K_{\overline{\mathcal{H}}'}(\zeta, f)$ in the sense of Lemma 3.13. Then there holds*

$$\mathrm{v}' = \sum_{j=1}^{\infty} \frac{\langle f, \varphi_j \rangle}{1 + \zeta^2 \lambda_j}. \tag{3.18}$$

In particular, v' coincides with the minimizer of $K_{\overline{\mathcal{H}}}(\zeta, f)$ if $f \in \mathcal{H}_0$.

Proof. Denoting with $(\varphi_j', \lambda_j')_{j=1}^{\infty}$ the eigenpairs of $\overline{\mathcal{H}}'$ in the sense of Corollary 3.3, we apply (3.14), Theorem 3.7, and (3.10) to deduce

$$\mathrm{v}' = \sum_{j=1}^{\infty} \frac{(f, \varphi_j')_{-1}}{1 + \zeta^2 \lambda_j'} \varphi_j' = \sum_{j=1}^{\infty} \lambda_j \frac{(f, \varphi_j)_{-1}}{1 + \zeta^2 \lambda_j} \varphi_j = \sum_{j=1}^{\infty} \frac{\langle f, \varphi_j \rangle}{1 + \zeta^2 \lambda_j^2} \varphi_j,$$

which proves (3.18). Clearly, if $f \in \mathcal{H}_0$, then $\langle f, \varphi_j \rangle = (f, \varphi_j)_0$ and the remainder of the proof follows from (3.14). □

Remark 3.20. *Another characterization of the interpolation space that is closely related to the K-method is the so-called J-method of interpolation [McL00, CWHM15]. However, since the latter leads to an integral representation of $\|u\|_{\overline{\mathcal{H}}^s}$ that follows from (3.16) by a simple substitution, we do not discuss this approach here.*

3.1.3 The Trace Method

The K-method allows one to characterize the interpolation norm as improper integral over the positive real line. The trace method, which is the main objective of this section, is inherently different and relies on the following weighted Bochner-Sobolev spaces.

Definition 3.21. *Let $s \in (0, 1)$ and $\mathcal{H}$ be a Hilbert space. We define the space $L_s^2(\mathbb{R}^+; \mathcal{H})$ of all Bochner-measurable functions $v : \mathbb{R}^+ \to \mathcal{H}$ such that*

$$\int_0^{\infty} \zeta^{1-2s} \|v(\zeta)\|_{\mathcal{H}}^2 \, d\zeta < \infty.$$

Further, we set

$$H_s^1(\mathbb{R}^+; \mathcal{H}) := \{v \in L_s^2(\mathbb{R}^+; \mathcal{H}) : v' \in L_2^2(\mathbb{R}^+; \mathcal{H})\},$$

where $v' = \partial_\zeta v$ is the weak derivative of v in the sense of Definition 2.16. Provided an interpolation couple $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$, we define the space

$$\mathbb{V}_s(\overline{\mathcal{H}}) := H_s^1(\mathbb{R}^+; \mathcal{H}_0) \cap L_s^2(\mathbb{R}^+; \mathcal{H}_1)$$

and endow it with the norm

$$\|v\|_{\mathbb{V}_s(\overline{\mathcal{H}})}^2 := \int_{\mathbb{R}^+} \zeta^{1-2s} \left(\|v(\zeta)\|_1^2 + \|v'(\zeta)\|_0^2 \right) d\zeta.$$

In line with [LM72, BL76, Tri78, Tar07], the purpose of this section is to characterize $[\mathcal{H}_0, \mathcal{H}_1]_s$ as space of trace functions. This is done in several steps. Following [CDDS11, Proposition 2.1], we first show that the trace operator

$$\begin{aligned} \operatorname{tr}_0 : \mathbb{V}_s(\overline{\mathcal{H}}) &\longrightarrow [\mathcal{H}_0, \mathcal{H}_1]_s, \\ v &\mapsto \operatorname{tr}_0 v := v(0), \end{aligned}$$

is continuous, i.e., we prove the existence of some constant $C \in \mathbb{R}^+$ such that

$$\| \operatorname{tr}_0 v\|_{\overline{\mathcal{H}}^s} \leq C \|v\|_{\mathbb{V}_s(\overline{\mathcal{H}})} \tag{3.19}$$

for all $v \in \mathbb{V}_s(\overline{\mathcal{H}})$. To this end, we note that any $v \in \mathbb{V}_s(\overline{\mathcal{H}})$ satisfies $v(\zeta) \in \mathcal{H}_0$ for all $\zeta \in \mathbb{R}^+$ and thus admits a representation of the form

$$v(\zeta) = \sum_{j=1}^{\infty} v_j(\zeta) \varphi_j, \qquad v_j(\zeta) = (\varphi_j, v(\zeta))_0.$$

Therefore, the norm on the right-hand side of (3.19) evaluates to

$$\begin{aligned} \|v\|^2_{\mathbb{V}_s(\overline{\mathcal{H}})} &= \int_{\mathbb{R}^+} \sum_{j=1}^{\infty} \zeta^{1-2s} \left(\lambda_j |v_j(\zeta)|^2 + |v_j'(\zeta)|^2 \right) d\zeta \\ &= \sum_{j=1}^{\infty} \int_{\mathbb{R}^+} \zeta^{1-2s} \left(\lambda_j |v_j(\zeta)|^2 + |v_j'(\zeta)|^2 \right) d\zeta, \end{aligned} \tag{3.20}$$

where the integral and series can be interchanged since each summand is nonnegative. We conclude that the spectral coefficients $v_j(\zeta)$ are contained in $H^1_s(\mathbb{R}^+) := H^1_s(\mathbb{R}^+; \mathbb{R})$ for all $j \in \mathbb{N}$. This space is amenable to trace evaluation, see e.g., [BM89], so that

$$\|v\|^2_{\mathbb{V}_s(\overline{\mathcal{H}})} \geq \sum_{j=1}^{\infty} |v_j(0)|^2 \inf_{\substack{\phi_j \in H^1_s(\mathbb{R}^+) \\ \phi_j(0)=1}} \int_{\mathbb{R}^+} \zeta^{1-2s} \left(\lambda_j |\phi_j(\zeta)|^2 + |\phi_j'(\zeta)|^2 \right) d\zeta. \tag{3.21}$$

These computations show that a proof of (3.19) is closely related to the minimization problem: Find $\phi_j \in H^1_s(\mathbb{R}^+)$ such that

$$\phi_j = \operatorname*{arg\,min}_{\substack{\phi_j \in H^1_s(\mathbb{R}^+) \\ \phi_j(0)=0}} \int_{\mathbb{R}^+} \zeta^{1-2s} \left(\lambda_j |\phi_j(\zeta)|^2 + |\phi_j'(\zeta)|^2 \right) d\zeta, \qquad j \in \mathbb{N}. \tag{3.22}$$

Standard tools from calculus of variation reveal that (3.22) has a unique minimizer that satisfies the *Euler-Lagrange equation*

$$\phi_j''(\zeta) - \frac{1-2s}{\zeta} \phi_j'(\zeta) - \lambda_j \phi_j(\zeta) = 0, \quad \text{in } \mathbb{R}^+, \tag{3.23a}$$

$$\phi_j(0) = 1. \tag{3.23b}$$

The ODE (3.23a) is a so-called *Bessel differential equation*. If $s = \frac{1}{2}$, two linear independent solutions are given by $e^{-\sqrt{\lambda_j}\zeta}$ and $e^{\sqrt{\lambda_j}\zeta}$. The integrability condition on $\phi_j \in H_s^1(\mathbb{R}^+)$ implies that $\phi_j(\zeta) = c_j e^{-\sqrt{\lambda_j}\zeta}$, $c_j \in \mathbb{R}$, is the solution we are looking for. If $s \neq \frac{1}{2}$, one has to resort to so-called *modified Bessel functions of first and second kind*. They are defined by

$$I_s(\zeta) := \sum_{j=0}^{\infty} \frac{\zeta^{2j+s}}{j!\Gamma(j+s+1)2^{2j+s}}, \qquad K_s(\zeta) := \frac{\pi}{2}\frac{I_{-s}(\zeta) - I_s(\zeta)}{\sin(\pi s)}, \qquad s \in (-1,1),$$

respectively. The key features of these functions are collected in the following lemma and can be found in [AS64, Section 9.6], see also [NOS15].

Lemma 3.22. *Let $s \in (-1,1)$. Then there holds*

1. *$\zeta^s K_s(\sqrt{\lambda_j}\zeta)$ and $\zeta^s I_s(\sqrt{\lambda_j}\zeta)$ are two linearly independent solutions of (3.23a),*
2. *$K_s(\zeta)$ decreases exponentially as $\zeta \to \infty$,*
3. *$I_s(\zeta)$ increases exponentially as $\zeta \to \infty$,*
4. *$K_s(\zeta)$ is real and positive on $\mathbb{R}^+$,*
5. *$K_s(\zeta)$ behaves like ζ^{-s} as $\zeta \to 0^+$. More precisely, there holds*

$$\lim_{\zeta \to 0^+} \zeta^s K_s(\zeta) = 2^{s-1}\Gamma(s), \tag{3.24}$$

6. *$K_s(\zeta) = K_{-s}(\zeta)$,*
7. *and finally,*

$$\frac{\partial}{\partial \zeta}\left(\zeta^s K_s(\zeta)\right) = -\zeta^s K_{1-s}(\zeta). \tag{3.25}$$

According to the first property in Lemma 3.22, the family of solutions to (3.23a) is given by

$$\phi_j(\zeta) = c_j \zeta^s K_s(\sqrt{\lambda_j}\zeta) + d_j \zeta^s I_s(\sqrt{\lambda_j}\zeta), \qquad c_j, d_j \in \mathbb{R}.$$

Since I_s increases exponentially and thus $I_s \notin H_s^1(\mathbb{R}^+)$, the only nontrivial contribution in this linear combination comes from K_s which is plotted in Figure 3.2. Hence, the solution we are seeking for is of the form

$$\phi_j(\zeta) = c_j \zeta^s K_s(\sqrt{\lambda_j}\zeta). \tag{3.26}$$

Thanks to (3.24), we can choose $c_j = 2^{1-s}\lambda_j^{\frac{s}{2}}/\Gamma(s)$ such that the initial condition $\phi_j(0) = 0$ is satisfied. For the later use, we collect these findings in the following proposition.

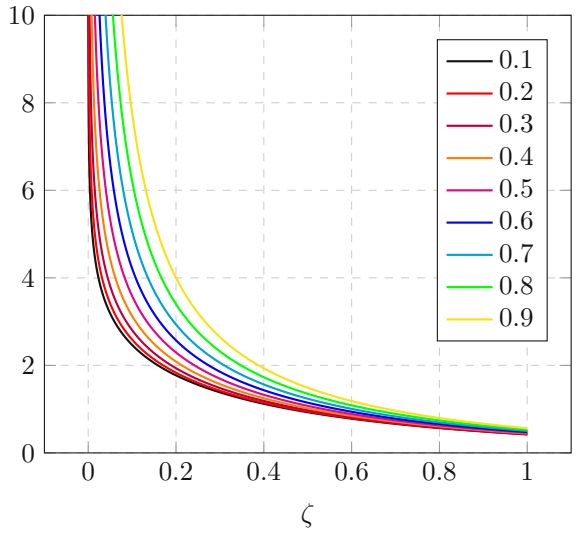

Figure 3.2: Modified Bessel functions of second kind $K_s(\zeta)$ on $[0,1]$ for different orders $s \in [0.1, 0.9]$.

Proposition 3.23. *Let $s \in (0,1)$ and $j \in \mathbb{N}$. Then the unique minimizer of (3.22) is given by*

$$\phi_j(\zeta) = \frac{2^{1-s}}{\Gamma(s)}(\sqrt{\lambda_j}\zeta)^s K_s(\sqrt{\lambda_j}\zeta). \tag{3.27}$$

The following technical lemma is instrumental for further discussions.

Lemma 3.24. *Let $s \in (0,1)$ and $j \in \mathbb{N}$. Then there holds*

$$\lim_{\zeta \to 0^+} \zeta^{1-2s}\phi_j'(\zeta) = -d_s\lambda_j^s, \qquad d_s := 2^{1-2s}\frac{\Gamma(1-s)}{\Gamma(s)}.$$

Proof. We apply (3.25) and the chain rule to observe

$$\phi_j'(\zeta) = \frac{2^{1-s}}{\Gamma(s)}\frac{\partial}{\partial\zeta}\left((\sqrt{\lambda_j}\zeta)^s K_s(\sqrt{\lambda_j}\zeta)\right) = -\frac{2^{1-s}}{\Gamma(s)}\sqrt{\lambda_j}(\sqrt{\lambda_j}\zeta)^s K_{1-s}(\sqrt{\lambda_j}\zeta).$$

Invoking (3.24) we obtain

$$\begin{aligned}\lim_{\zeta \to 0^+} \zeta^{1-2s}\phi_j'(\zeta) &= -\lim_{\zeta \to 0^+}\frac{2^{1-s}}{\Gamma(s)}\sqrt{\lambda_j^{1+s}}\zeta^{1-s}K_{1-s}(\sqrt{\lambda_j}\zeta) \\ &= -\lim_{\zeta \to 0^+}\frac{2^{1-s}}{\Gamma(s)}\lambda_j^s(\sqrt{\lambda_j}\zeta)^{1-s}K_{1-s}(\sqrt{\lambda_j}\zeta) = -2^{1-2s}\frac{\Gamma(1-s)}{\Gamma(s)}\lambda_j^s\end{aligned}$$

and the proof is complete. □

In view of (3.21), we are interested in the value of the integral with respect to the minimizer ϕ_j. This is the subject of the following lemma, see also [CDDS11, BCdPS13, NOS15].

Lemma 3.25. *Let $s \in (0,1)$, d_s as in Lemma 3.24, $j \in \mathbb{N}$, and ϕ_j defined by (3.26). Then there holds*

$$\int_{\mathbb{R}^+} \zeta^{1-2s} \left(\lambda_j |\phi_j(\zeta)|^2 + |\phi_j'(\zeta)|^2 \right) d\zeta = d_s \lambda_j^s.$$

Proof. We multiply the Bessel ODE (3.23a) with $\zeta^{1-2s}\phi_j$ and integrate over $\mathbb{R}^+$ to deduce

$$\int_0^\infty \zeta^{1-2s}\phi_j''(\zeta)\phi_j(\zeta) + (1-2s)\zeta^{2s}\phi_j'(\zeta)\phi_j(\zeta) - \zeta^{1-2s}\lambda_j|\phi_j(\zeta)|^2 \, d\zeta = 0.$$

Integration by parts of the first integrand and rearrangement the terms reveals

$$\int_0^\infty \zeta^{1-2s} \left(|\phi_j'(\zeta)|^2 + \lambda_j |\phi_j(\zeta)|^2 \right) d\zeta = \zeta^{1-2s}\phi_j'(\zeta)\phi_j(\zeta)\big|_{\zeta=0}^\infty.$$

From the properties 2. and 7. in Lemma 3.22 we see that both $\phi_j(\zeta)$ and $\phi_j'(\zeta)$ decrease exponentially as $\zeta \to \infty$. Therefore,

$$\lim_{\zeta \to \infty} \zeta^{1-2s}\phi_j'(\zeta)\phi_j(\zeta) = 0.$$

To compute the limit as $\zeta \to 0^+$, we apply Lemma 3.24 and $\phi_j(0) = 1$ to see that

$$\lim_{\zeta \to 0^+} \zeta^{1-2s}\phi_j'(\zeta)\phi_j(\zeta) = -d_s\lambda_j^s.$$

Hence,

$$\int_{\mathbb{R}^+} \zeta^{1-2s}(\lambda_j|\phi_j(\zeta)|^2 + |\phi_j'(\zeta)|^2) \, d\zeta = d_s\lambda_j^s. \qquad \square$$

As a consequence of (3.21), Proposition 3.23, and Lemma 3.25, we deduce the continuity of the trace operator.

Theorem 3.26. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple, $s \in (0,1)$, and d_s defined as in Lemma 3.24. Then there holds for all $v \in \mathbb{V}_s(\overline{\mathcal{H}})$*

$$\| \operatorname{tr}_0 v \|_{\overline{\mathcal{H}}^s} \leq \sqrt{d_s} \|v\|_{\mathbb{V}_s(\overline{\mathcal{H}})}.$$

As a by-product of the proof of Theorem 3.26, we obtain the following result, where we define the *s-minimal extension* $\mathcal{U} \in \mathbb{V}_s(\overline{\mathcal{H}})$ *of* $u \in [\mathcal{H}_0, \mathcal{H}_1]_s$ by

$$\mathcal{U}(\zeta) := \sum_{j=1}^\infty u_j \phi_j(\zeta)\varphi_j, \tag{3.28}$$

with ϕ_j as in (3.27).

Corollary 3.27. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple, $s \in (0,1)$, and $\mathcal{U}$ the s-minimal extension of $u \in [\mathcal{H}_0, \mathcal{H}_1]_s$. Then there holds*

$$\|\mathcal{U}\|_{\mathbb{V}_s(\overline{\mathcal{H}})} = \inf_{\substack{v \in \mathbb{V}_s(\overline{\mathcal{H}}) \\ \operatorname{tr}_0 v = u}} \|v\|_{\mathbb{V}_s(\overline{\mathcal{H}})}. \tag{3.29}$$

Proof. Due to Proposition 3.23, (3.21) holds with equality if $\mathcal{U}$ is the s-minimal extension of u. This immediately implies (3.29). □

The fact that $\mathrm{tr}_0 \mathcal{U} = u$ for any $u \in [\mathcal{H}_0, \mathcal{H}_1]_s$ shows that $\mathrm{tr}_0 : \mathbb{V}_s(\overline{\mathcal{H}}) \to [\mathcal{H}_0, \mathcal{H}_1]_s$ is surjective and thus

$$[\mathcal{H}_0, \mathcal{H}_1]_s = \mathrm{tr}_0(\mathbb{V}_s(\overline{\mathcal{H}}))$$

for all $s \in (0,1)$. One readily verifies that

$$\|u\|_{E_{\overline{\mathcal{H}}^s}} := \inf_{\substack{v \in \mathbb{V}_s(\overline{\mathcal{H}}) \\ \mathrm{tr}_0\, v = u}} \|v\|_{\mathbb{V}_s(\overline{\mathcal{H}})}$$

is a norm on $[\mathcal{H}_0, \mathcal{H}_1]_s$. Provided the right normalization constant, it coincides with the interpolation norm on $[\mathcal{H}_0, \mathcal{H}_1]_s$ for all $s \in (0,1)$.

Theorem 3.28. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple, $s \in (0,1)$, d_s defined as in Lemma 3.24, and $u \in [\mathcal{H}_0, \mathcal{H}_1]_s$. Then there holds*

$$\|u\|_{\overline{\mathcal{H}}^s} = \frac{1}{\sqrt{d_s}} \|u\|_{E_{\overline{\mathcal{H}}^s}}. \tag{3.30}$$

Proof. According to Corollary 3.27, (3.30) is equivalent to

$$\|u\|_{\overline{\mathcal{H}}^s} = \frac{1}{\sqrt{d_s}} \|\mathcal{U}\|_{\mathbb{V}_s(\overline{\mathcal{H}})},$$

where $\mathcal{U}$ is the s-minimal extension of u. Due to (3.20), the latter satisfies

$$\|\mathcal{U}\|^2_{\mathbb{V}_s(\overline{\mathcal{H}})} = \sum_{j=1}^{\infty} u_j^2 \int_{\mathbb{R}^+} \zeta^{1-2s} \left(\lambda_j |\phi_j(\zeta)|^2 + |\phi_j'(\zeta)|^2 \right) d\zeta.$$

By Lemma 3.24 it follows that

$$\|\mathcal{U}\|^2_{\mathbb{V}_s(\overline{\mathcal{H}})} = d_s \sum_{j=1}^{\infty} \lambda_j^s u_j^2 = d_s \|u\|^2_{\overline{\mathcal{H}}^s}$$

and the proof is complete. □

3.2 Fractional Sobolev spaces

In this section, we apply the abstract interpolation theory presented above to the case where $\mathcal{H}_0$ and $\mathcal{H}_1$ are Sobolev spaces. The latter turn out to provide the right framework to study fractional powers of differential operators and shall serve us as a starting point for our further discussions in Chapter 4. We state some of their intriguing properties and provide several equivalent characterizations which are frequently used in the literature.

3.2.1 The Space $H^s(\Omega)$

We start our discussion in the absence of boundary conditions on the bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$. Due to the first property in Theorem 2.5, the space $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$. Since $C_0^\infty(\Omega) \subset H^1(\Omega)$ and $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$, we conclude that $\overline{\mathcal{H}} = (L^2(\Omega), H^1(\Omega))$ is an interpolation couple so that the following definition is meaningful.

Definition 3.29. *For all $s \in (0,1)$ we define fractional Sobolev spaces and norms of $\overline{\mathcal{H}} = (L^2(\Omega), H^1(\Omega))$ by*

$$H^s(\Omega) := [L^2(\Omega), H^1(\Omega)]_s, \qquad \|u\|_{H^s(\Omega)} := \|u\|_{\overline{\mathcal{H}}^s}.$$

Fractional Sobolev spaces satisfy several properties that are reminiscent of the respective ones from classical Sobolev theory. The following lemma is a straightforward generalization of the integer-order case and allows one to apply density arguments.

Lemma 3.30. *For all $s \in (0,1)$ the space $C^\infty(\overline{\Omega})$ is dense in $H^s(\Omega)$.*

Proof. See [McL00, Theorem 3.25]. □

Theorem 2.4 shows that Sobolev functions $u \in H^k(\Omega)$ are at least continuous if $k > \frac{d}{2}$. A similar statement holds for $H^s(\Omega)$; see [NPV12, Theorem 8.2].

Theorem 3.31. *If $s > \frac{d}{2}$, then the embedding $H^s(\Omega) \subset C(\overline{\Omega})$ is continuous.*

Finally, we state the following compactness result which constitutes the fractional counterpart to the first claim in Theorem 2.5 and can be found in [McL00].

Theorem 3.32. *For all $s \in (0,1)$ the embedding $H^s(\Omega) \subset L^2(\Omega)$ is compact.*

Section 3.1 shows that the interpolation space of an interpolation couple $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ can be characterized in a spectral fashion, via improper integrals, and as space of trace functions. If $\mathcal{H}_0$ and $\mathcal{H}_1$ are Hilbert spaces of functions $u : \Omega \to \mathbb{R}$, several other characterizations exist that are frequently used in the literature [McL00, NPV12, BRS16, DL21]. One rather explicit way of defining $H^s(\Omega)$ is due to the almost simultaneous contributions of Aronszajn [Aro55], Gagliardo [Gag58], and Slobodeckij [Slo58]. As shown in [LM72, AF03], the space $H^s(\Omega)$ consists of all $u \in L^2(\Omega)$ such that

$$\int_\Omega \int_\Omega \frac{|u(\mathbf{x}) - u(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_2^{d+2s}} \, d\mathbf{x} \, d\mathbf{y} < \infty, \tag{3.31}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. The square root of (3.31) is often referred to as *Slobodeckij seminorm* and constitutes, after adding the term $\|u\|_{L^2(\Omega)}$, a norm on $H^s(\Omega)$ that is equivalent to $\|u\|_{H^s(\Omega)}$.

Under certain regularity assumptions on Ω, which are in particular satisfied for bounded Lipschitz domains, there holds the so-called *extension property*: For any $s \in (0,1)$ there exists an extension operator $\mathbb{E} : H^s(\Omega) \to H^s(\mathbb{R}^d)$ such that [NPV12, Theorem 5.4]

$$\mathbb{E}(u)|_\Omega = u, \qquad \|\mathbb{E}(u)\|_{H^s(\mathbb{R}^d)} \preceq \|u\|_{H^s(\Omega)}, \tag{3.32}$$

where

$$H^s(\mathbb{R}^d) := \{u \in L^2(\mathbb{R}^d) : \|u\|_{H^s(\mathbb{R}^d)} < \infty\}$$

and

$$\|u\|^2_{H^s(\mathbb{R}^d)} := \|u\|^2_{L^2(\mathbb{R}^d)} + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{|u(\mathbf{x}) - u(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}}\, d\mathbf{x}\, d\mathbf{y}.$$

Note that the inequality in (3.32) reveals that

$$u \mapsto \inf_{\substack{U \in H^1(\mathbb{R}^d) \\ U|_\Omega = u}} \|U\|_{H^s(\mathbb{R}^d)}$$

is an equivalent norm on $H^s(\Omega)$. Hence, we obtain yet another characterization (c.f. [LM72] Theorem 9.1 and 9.2) in the form of

$$H^s(\Omega) = \{U|_\Omega : U \in H^s(\mathbb{R}^d)\}.$$

Remark 3.33. *The space $H^s(\mathbb{R}^d)$ can be equivalently characterized by means of the Fourier transformation or as interpolation space between $L^2(\mathbb{R}^d)$ and $H^1(\mathbb{R}^d)$. Unlike in the case of bounded domains, however, the embedding $H^1(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$ is continuous but not compact (cf. Remark 3.10).*

3.2.2 The Spaces $H^s_0(\Omega)$ and $H^{\frac{1}{2}}_{00}(\Omega)$

We are interested in the inclusion of homogeneous Dirichlet boundary conditions in fractional Sobolev spaces. This is a delicate task that can be done in several mathematically distinct ways. For now, we proceed as before and define $H^s_0(\Omega)$ as interpolation space of the pairing $\overline{\mathcal{H}} = ((L^2(\Omega), \|\cdot\|_{L^2(\Omega)}), (H^1_0(\Omega), \|\nabla\cdot\|_{L^2(\Omega)}))$, which, due to Theorem 2.5, indeed is an interpolation couple.

Definition 3.34. *For all $s \in (0,1)$ we define fractional Sobolev spaces and norms of $\overline{\mathcal{H}} = (L^2(\Omega), H^1_0(\Omega))$ by*

$$H^s_0(\Omega) := [L^2(\Omega), H^1_0(\Omega)]_s, \qquad \|u\|_{H^s_0(\Omega)} := \|u\|_{\overline{\mathcal{H}}^s}.$$

Remark 3.35. *Sometimes it is useful to define $H^s_0(\Omega)$ also for values of $s > 1$. A natural approach to do this is obtained by means of the orthonormal system of eigenfunctions $(\varphi_j)_{j=1}^\infty \subset H^1_0(\Omega)$ satisfying*

$$(\nabla\varphi_j, \nabla v)_{L^2(\Omega)} = (\varphi_j, v)_{L^2(\Omega)}, \qquad v \in H^1_0(\Omega).$$

Recognizing that there is no reason to restrict s in Definition 3.4 to $[0,1]$, we define for all $s > 1$

$$H^s_0(\Omega) := \{u \in L^2(\Omega) : \|u\|_{H^s_0(\Omega)} < \infty\}, \qquad \|u\|^2_{H^s_0(\Omega)} := \sum_{j=1}^\infty \lambda_j^s u_j^2.$$

The space $(H^s(\Omega), \|\cdot\|_{H^s(\Omega)})$, $s > 1$, is understood accordingly.

The following observation is a natural one and can be found in e.g., [McL00].

Proposition 3.36. *For all $s \in (0,1)$ the embedding $H_0^s(\Omega) \subset H^s(\Omega)$ is continuous.*

In view of (2.2), it is a natural question to ask whether $H_0^s(\Omega)$ coincides with the closure of $C_0^\infty(\Omega)$ with respect to the fractional Sobolev norm $\|\cdot\|_{H^s(\Omega)}$. If $s \neq \frac{1}{2}$, this is indeed the case [LM72, Theorem 11.6].

Theorem 3.37. *Let $s \in (0,1) \setminus \{\frac{1}{2}\}$. Then there holds*

$$H_0^s(\Omega) = \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^s(\Omega)}}.$$

If $s = \frac{1}{2}$, there holds the strict inclusion

$$H_0^{\frac{1}{2}}(\Omega) \subsetneq \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^{\frac{1}{2}}(\Omega)}},$$

which shows that $H_0^{\frac{1}{2}}(\Omega)$ is a special case. The latter is called *Lions-Magenes space* and is often written as $H_{00}^{\frac{1}{2}}(\Omega)$. Its discrepancy to $\overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^{\frac{1}{2}}(\Omega)}}$ can be made more explicit by means of the identity [LM72, Theorem 11.7]

$$H_{00}^{\frac{1}{2}}(\Omega) = \left\{ u \in H^{\frac{1}{2}}(\Omega) : \int_\Omega \frac{|u(\mathbf{x})|^2}{\operatorname{dist}(\mathbf{x}, \partial\Omega)}\, d\mathbf{x} < \infty \right\}.$$

This shows $1 \in \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^{\frac{1}{2}}(\Omega)}}$ but $1 \notin H_{00}^{\frac{1}{2}}(\Omega)$. A characterization of $H_0^s(\Omega)$ which works for all $s \in (0,1)$ is obtained by

$$H_0^s(\Omega) = \{u \in H^s(\mathbb{R}^d) : \operatorname{supp} u \subset \overline{\Omega}\},$$

see [BSV15, Section 3.1.3] and references therein.

If $s = 1$, $[L^2(\Omega), H_0^1(\Omega)]_1 = H_0^1(\Omega)$ and there exists a well-defined trace operator according to Theorem 2.2. On the other hand, $[L^2(\Omega), H_0^1(\Omega)]_0 = L^2(\Omega)$ in which case a reasonable notion of traces does not exist. Only for s close to 1, one can thus hope for a meaningful trace operator. The following theorem addresses this matter.

Theorem 3.38. *For all $s \in (\frac{1}{2}, 1]$ there exists a linear operator $\operatorname{tr} : H^s(\Omega) \to L^2(\partial\Omega)$ with the properties*

$$\forall u \in H^s(\Omega) : \ \|\operatorname{tr} u\|_{L^2(\partial\Omega)} \preceq \|u\|_{H^s(\Omega)}, \qquad \forall u \in H^s(\Omega) \cap C(\overline{\Omega}) : \ \operatorname{tr} u = u|_{\partial\Omega}.$$

Proof. See [LM72, Theorem 9.4] and [McL00, Theorem 3.37 and 3.38]. □

Provided $s > \frac{1}{2}$, the space $H_0^s(\Omega)$ can be interpreted as kernel of the trace operator, i.e.,

$$H_0^s(\Omega) = \{u \in H^s(\Omega) : \operatorname{tr} u = 0\}, \qquad s \in (\tfrac{1}{2}, 1).$$

Except for the special case $s = \frac{1}{2}$, the interpolation space does not "see" the boundary conditions prescribed by $H_0^1(\Omega)$ in the absence of a well-defined trace; see [LM72, Theorem 11.1].

Theorem 3.39. *The space $C_0^\infty(\Omega)$ is dense in $H^s(\Omega)$ if and only if $0 < s \leq \frac{1}{2}$. In particular,*

$$\begin{cases} H_0^s(\Omega) = H^s(\Omega), & 0 < s < \frac{1}{2}, \\ H_0^s(\Omega) \subsetneq H^s(\Omega), & \frac{1}{2} \leq s < 1. \end{cases}$$

Remark 3.40. *In the special case $s = \frac{1}{2}$, there holds*

$$H_0^{\frac{1}{2}}(\Omega) = H_{00}^{\frac{1}{2}}(\Omega) \subsetneq \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^{\frac{1}{2}}(\Omega)}} = H^{\frac{1}{2}}(\Omega).$$

In accordance with the integer-order case, we conclude this section with the definition of fractional Sobolev spaces of negative order.

Definition 3.41. *For all $s \in (0,1)$ we define the negative fractional Sobolev space*

$$H^{-s}(\Omega) := \left(H_0^s(\Omega)\right)',$$

whose norm we denote with $\|\cdot\|_{H^{-s}(\Omega)}$.

4 Fractional Diffusion Operators

In this chapter we introduce the fractional powers $\mathcal{L}^s$, $s \in (0,1)$, of the diffusion operator

$$\begin{aligned} \mathcal{L} : H_0^1(\Omega) &\to H^{-1}(\Omega) \\ u &\mapsto -\operatorname{div}(\mathfrak{A}\nabla u) + \mathfrak{c}u, \end{aligned} \tag{4.1}$$

where

- $\mathfrak{A} \in L^\infty(\Omega; \mathbb{R}^{d\times d})$,
- $\mathfrak{A}$ is symmetric,
$$\mathfrak{A}(\mathbf{x}) = \mathfrak{A}(\mathbf{x})^T, \qquad \mathbf{x} \in \Omega,$$
- $\mathfrak{A}$ is uniformly positive definite,
$$\mathbf{y}^T\mathfrak{A}(\mathbf{x})\mathbf{y} \geq c_{\mathrm{p}}\|\mathbf{y}\|_2^2, \qquad (\mathbf{x},\mathbf{y}) \in \Omega \times \mathbb{R}^d, \tag{4.2}$$
for some $c_{\mathrm{p}} \in \mathbb{R}^+$,
- and $\mathfrak{c} \in L^\infty(\Omega)$ with $\mathfrak{c}(\mathbf{x}) \geq 0$ almost everywhere in Ω.

Our definition of $\mathcal{L}^s$ relies on the theory of interpolation operators. Leveraging our knowledge from Chapter 3, we present three equivalent characterizations of $\mathcal{L}^s$ which allow us to circumvent its nonlocal character at the cost of

1. an infinite family of local eigenvalue problems,
2. an improper integral over solutions to parametric reaction-diffusion problems,
3. a degenerate elliptic PDE in a $d+1$-dimensional domain.

We derive a weak formulation of the fractional diffusion problem: Find $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} \mathcal{L}^s u &= f, \quad \text{in } \Omega, \\ u &= 0, \quad \text{on } \partial\Omega. \end{aligned} \tag{4.3}$$

As special case, we study solutions to (4.3) when $\mathcal{L} = -\Delta$ and investigate several of its intriguing properties. In the final section of this chapter, we discuss two alternative definitions of the fractional powers of the Laplacian and comment on similarities and differences to the definition we pursue in this manuscript.

4.1 Definition and Characterizations

The literature advocates a variety of mathematically distinct definitions to introduce fractional powers of differential operators [DWZ17, BBNS18, LPG+20, DL21]. Only if $\Omega = \mathbb{R}^d$, they are known to be equivalent [Kwa17]. If Ω is bounded, there exist at least three nonequivalent ways to incorporate boundary conditions in the definition of $\mathcal{L}^s$. A natural one that shall serve us as *the* definition of the fractional diffusion operator is based on interpolation theory. To make matters precise, let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple and $[\mathcal{H}_0, \mathcal{H}_1]_s$ its interpolation space. Recalling (3.11), we consider the problem: Given $u \in [\mathcal{H}_0, \mathcal{H}_1]_s$ find $f \in [\mathcal{H}_0, \mathcal{H}_1]_{-s}$ such that

$$\forall v \in [\mathcal{H}_0, \mathcal{H}_1]_s : \quad \langle f, v\rangle = (u, v)_{\overline{\mathcal{H}}^s}. \tag{4.4}$$

Thanks to Riesz's representation theorem, there exists a unique solution $f \in [\mathcal{H}_0, \mathcal{H}_1]_{-s}$ that satisfies (4.4). Therefore, the following definition is well-defined.

Definition 4.1. *Let $\overline{\mathcal{H}} = (\mathcal{H}_0, \mathcal{H}_1)$ be an interpolation couple and $s \in [0, 1]$. We define the interpolation operator of $\overline{\mathcal{H}}$ of order s by*

$$\begin{aligned} \mathcal{L}^s_{\overline{\mathcal{H}}} : [\mathcal{H}_0, \mathcal{H}_1]_s &\to [\mathcal{H}_0, \mathcal{H}_1]_{-s} \\ u &\mapsto \mathcal{L}^s_{\overline{\mathcal{H}}} u := f, \end{aligned}$$

where f is the unique solution to (4.4).

Along with a suitable choice of $\mathcal{H}_0$ and $\mathcal{H}_1$, our goal is to define the fractional powers of (4.1) as operator of interpolation in the sense of Definition 4.1. To this end, we introduce the bilinear from $(\cdot,\cdot)_{H^1_{\mathcal{L}}(\Omega)} : H^1_0(\Omega) \times H^1_0(\Omega) \to \mathbb{R}$ associated to $\mathcal{L}$ by

$$(u, v)_{H^1_{\mathcal{L}}(\Omega)} := (\mathfrak{A}\nabla u, \nabla v)_{L^2(\Omega)} + (\mathfrak{c}u, v)_{L^2(\Omega)}. \tag{4.5}$$

The latter induces a norm on $H^1_0(\Omega)$ by

$$\|u\|_{H^1_{\mathcal{L}}(\Omega)} := \sqrt{(u, u)_{H^1_{\mathcal{L}}(\Omega)}}, \tag{4.6}$$

which makes $\left(H^1_0(\Omega), \|\cdot\|_{H^1_{\mathcal{L}}(\Omega)}\right)$ a Hilbert space. By construction, there holds

$$\langle \mathcal{L}u, v\rangle = (u, v)_{H^1_{\mathcal{L}}(\Omega)} \tag{4.7}$$

for all $v \in H^1_0(\Omega)$. It follows from (4.2), $\mathfrak{c} \geq 0$ a.e. in Ω, and the boundedness of $\mathfrak{A}$ and $\mathfrak{c}$ that

$$c_{\mathrm{p}}\|\nabla u\|_{L^2(\Omega)} \leq \|u\|_{H^1_{\mathcal{L}}(\Omega)} \leq \|\mathfrak{A}\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})}\|\nabla u\|_{L^2(\Omega)} + \|\mathfrak{c}\|_{L^\infty(\Omega)}\|u\|_{L^2(\Omega)}.$$

Thanks to the second claim in Theorem 2.5, we deduce

$$\|\nabla u\|_{L^2(\Omega)} \preceq \|u\|_{H^1_{\mathcal{L}}(\Omega)} \preceq \|\nabla u\|_{L^2(\Omega)},$$

whence (4.6) is an equivalent norm on $H^1_0(\Omega)$. We conclude that the pairing $\overline{\mathcal{H}} = ((L^2(\Omega), \|\cdot\|_{L^2(\Omega)}), (H^1_0(\Omega), \|\cdot\|_{H^1_{\mathcal{L}}(\Omega)}))$ is an interpolation couple, whose interpolation norm we denote by $\|\cdot\|_{H^s_{\mathcal{L}}(\Omega)}$. These considerations lead us to the central definition of this chapter.

Definition 4.2 (Fractional Diffusion Operator)**.** *Let $\mathcal{L}$ be defined by (4.1) and $s \in [0,1]$. We define the fractional power $\mathcal{L}^s$ of $\mathcal{L}$ as interpolation operator of the interpolation couple $\overline{\mathcal{H}} = \Big((L^2(\Omega), \|\cdot\|_{L^2(\Omega)}), (H_0^1(\Omega), \|\cdot\|_{H^1_{\mathcal{L}}(\Omega)})\Big)$ of order s. We call $\mathcal{L}^s$ the fractional diffusion operator.*

Remark 4.3. *Throughout this thesis, we restrict ourselves to homogeneous Dirichlet boundary conditions. For the treatment of nonhomogeneous boundary conditions of Dirichlet and Neumann type we refer to [APR17, HMP21, LPG+20].*

By construction, the fractional diffusion operator $\mathcal{L}^s$ satisfies

$$\langle \mathcal{L}^s u, v\rangle = (u,v)_{H^s_{\mathcal{L}}(\Omega)}, \qquad v \in H_0^1(\Omega), \tag{4.8}$$

where $(\cdot,\cdot)_{H^s_{\mathcal{L}}(\Omega)}$ denotes the interpolation scalar product of $\overline{\mathcal{H}}$. There holds

$$\mathcal{L}^0 = \mathrm{I}, \qquad \mathcal{L}^1 = \mathcal{L}.$$

Arguably the most prominent example of a fractional diffusion operator is the *fractional Laplacian* $(-\Delta)^s$, which is contained in Definition 4.2 as special case upon setting $\mathfrak{A}(\mathbf{x}) = \mathbf{I} \in \mathbb{R}^{d\times d}$ and $\mathfrak{c} \equiv 0$ in (4.1).

4.1.1 Spectral Representation

A more intuitive representation of fractional diffusion operators is obtained by spectral expansion. In view of Corollary 3.3, the action of $\mathcal{L}^s$ on $u \in H_0^s(\Omega)$ can be written by means of the L^2-orthonormal basis of eigenfunctions $(\varphi_j)_{j=1}^\infty \subset H_0^1(\Omega)$, satisfying

$$(\varphi_j, v)_{H^1_{\mathcal{L}}(\Omega)} = \lambda_j (\varphi_j, v)_{L^2(\Omega)}, \qquad v \in H_0^1(\Omega). \tag{4.9}$$

Since

$$f := \sum_{j=1}^\infty \lambda_j^s u_j \varphi_j, \qquad u_j = (\varphi_j, u)_{L^2(\Omega)},$$

satisfies

$$\langle f, v\rangle = \sum_{j=1}^\infty \sum_{i=1}^\infty \lambda_j^s u_j v_i \langle \varphi_j, \varphi_i\rangle = \sum_{j=1}^\infty \sum_{i=1}^\infty \lambda_j^s u_j v_i (\varphi_j, \varphi_i)_{L^2(\Omega)} = \sum_{j=1}^\infty \lambda_j^s u_j v_j = (u,v)_{H^1_{\mathcal{L}}(\Omega)}$$

for all $u, v \in H_0^1(\Omega)$, we arrive at the following spectral representation of the fractional diffusion operator.

Theorem 4.4. *For all $s \in [0,1]$ and $u \in H_0^s(\Omega)$ there holds*

$$\mathcal{L}^s u = \sum_{j=1}^\infty \lambda_j^s u_j \varphi_j. \tag{4.10}$$

Example 4.5. *We consider a one-dimensional example with $\mathcal{L} = -\Delta$ in $\Omega = (-L, L)$, $L > 0$, in which case the normalized solutions to the eigenproblem (4.9) are given in closed form. There holds*

$$\varphi_j(x) = \frac{1}{\sqrt{L}} \sin\left(\frac{j\pi x}{L}\right), \qquad \lambda_j = \frac{j^2\pi^2}{L^2},$$

for all $j \in \mathbb{N}$ such that

$$(-\Delta)^s u(x) = \sum_{j=1}^{\infty} \left(\frac{j^2\pi^2}{L^2}\right)^s \frac{u_j}{\sqrt{L}} \sin\left(\frac{j\pi x}{L}\right), \qquad u_j = \frac{1}{\sqrt{L}} \left(\sin\left(\frac{j\pi x}{L}\right), u\right)_{L_2((-L,L))}.$$

Let now $u(x) = \sin(\pi x)$ and $L = 1$. Then $u = \varphi_1$ and $(-\Delta)^s u$ evaluates to

$$\begin{aligned}(-\Delta)^s u(x) &= \sum_{j=1}^{\infty} (j\pi)^{2s} (\varphi_j, \varphi_1)_{L^2((-1,1))} \sin(j\pi x) \\ &= \pi^{2s} \|\varphi_1\|^2_{L^2((0,1))} \sin(\pi x) = \pi^{2s} \sin(\pi x).\end{aligned}$$

One of the characteristic properties of $\mathcal{L}^s$ is its nonlocal nature. We say that an operator $\mathfrak{L} : \mathcal{H} \to \hat{\mathcal{H}}$ between two Hilbert spaces $\mathcal{H}$ and $\hat{\mathcal{H}}$ of functions $u : \Omega \to \mathbb{R}$ is *local* if for all $\mathbf{x} \in \Omega$ and $u \in \mathcal{H}$ the value $(\mathfrak{L}u)(\mathbf{x})$ only depends on the values of $u|_{B_\varepsilon(\mathbf{x})}$ for all $\varepsilon > 0$. Operators of the form $\mathcal{N} : \mathcal{H} \to \hat{\mathcal{H}}$ which do not satisfy this condition are said to be *nonlocal*. While classical integer-order differential operators are local, $\mathcal{L}^s$, $s \in (0, 1)$, is a prototypical example of a nonlocal operator. This has the intriguing consequence that a function u might be identically zero in some neighborhood of $\mathbf{x} \in \Omega$ whereas $\mathcal{L}^s u(\mathbf{x}) \neq 0$. Moreover, the eigenvalues of $\mathcal{L}^s$ depend nonlinearly on the fractional exponent whence manipulations of the domain Ω directly affect the value of $\mathcal{L}^s u(\mathbf{x})$ for any $\mathbf{x} \in \Omega$. Although not explicitly visible in its expression, the reader should be reminded that there is always a domain Ω intrinsically linked to the operator $\mathcal{L}^s$.

Further properties of $\mathcal{L}^s$ follow directly from its spectral representation (4.10) and, among others, justify the abbreviation $\mathcal{L}^{-s} := (\mathcal{L}^s)^{-1}$.

Lemma 4.6.

1. *For all $s \in (0, 1)$ and $f \in H^{-s}(\Omega)$ there holds*

$$\mathcal{L}^{-s} f = \sum_{j=1}^{\infty} \lambda_j^{-s} \langle f, \varphi_j \rangle \varphi_j.$$

2. *For all $-1 < s_1 < s_2 < 1$ with $s_1 + s_2 \in (0, 1)$ there holds*

$$\mathcal{L}^{s_1} \mathcal{L}^{s_2} = \mathcal{L}^{s_2} \mathcal{L}^{s_1} = \mathcal{L}^{s_1+s_2}.$$

4.1.2 Integral Formulas

The spectral representation formula (4.10) is convenient to gain insights in some elementary properties of fractional diffusion operators. To design numerical approximations for these objects, however, alternative characterizations are often useful. In this section we show that $\mathcal{L}^s$ allows the interpretation as improper integral over parameterized reaction-diffusion problems. For this purpose, we state two important preliminary results. The first one can be found in [Bal60, BS87, Yos95]. A complete proof is provided in Section 8.1.1.

Theorem 4.7. *Let $s \in (0,1)$ and $\lambda \in \mathbb{R}^+$. Then there holds*

$$\lambda^s = \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{s-1} \frac{\lambda}{\lambda + \zeta} \, d\zeta. \tag{4.11}$$

An intuitive approach for defining fractional differential operators is to formally replace λ in (4.11) with $\mathcal{L}$. Recalling Remark 3.35, the following theorem states that this approach is indeed justified.

Proposition 4.8. *Let $s \in (0,1)$, $u \in H_0^2(\Omega)$, $\zeta \in \mathbb{R}^+$, and $\tilde{\mathrm{v}}(\zeta)$ the unique solution of the problem: Find $\tilde{\mathrm{v}}(\zeta) \in H_0^1(\Omega)$ such that*

$$\forall v \in H_0^1(\Omega): \quad (\tilde{\mathrm{v}}(\zeta), v)_{H^s_{\mathcal{L}}(\Omega)} + \zeta(\tilde{\mathrm{v}}(\zeta), v)_{L^2(\Omega)} = (u, v)_{H_0^1(\Omega)}.$$

Then there holds $\zeta^{s-1}\tilde{\mathrm{v}}(\zeta) \in L^1(\mathbb{R}^+; L^2(\Omega))$.

Proof. See [Bal60, BS87, Yos95] for a more general framework and the proof of [BP15, Theorem 2.1] for this particular setting. □

For the sake of a more explicit notation, we write $(\mathcal{L}+\zeta\,\mathrm{I})^{-1}\mathcal{L}u$ instead of $\tilde{\mathrm{v}}(\zeta)$ henceforth and understand the respective differential operators in the weak sense. In this terminology, Proposition 4.8 states that

$$\frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{s-1}(\mathcal{L} + \zeta\,\mathrm{I})^{-1}\mathcal{L}u \, d\zeta \tag{4.12}$$

exists as Bochner integral for any $u \in H_0^2(\Omega)$. In light of (4.11), the following theorem shows that (4.12) indeed coincides with our definition of $\mathcal{L}^s$ if $u \in H_0^2(\Omega)$.

Theorem 4.9. *Let $s \in (0,1)$ and $u \in H_0^2(\Omega)$. Then there holds*

$$\mathcal{L}^s u = \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{s-1}(\mathcal{L} + \zeta\,\mathrm{I})^{-1}\mathcal{L}u \, d\zeta \tag{4.13}$$

in the sense of Bochner.

Proof. Recalling Proposition 4.8, we apply the transformation $\zeta \mapsto \zeta^{-\frac{1}{2}}$ to deduce

$$\frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{s-1}(\mathcal{L} + \zeta\,\mathrm{I})^{-1}\mathcal{L}u \, d\zeta = \frac{2\sin(\pi s)}{\pi} \int_0^\infty \zeta^{1-2s}(\zeta^2\mathcal{L} + \mathrm{I})^{-1}\mathcal{L}u \, d\zeta.$$

Decomposing the integrand on the right-hand side in its spectral components, we find

$$\zeta^2(\zeta^2\mathcal{L}+\mathrm{I})^{-1}\mathcal{L}u = \sum_{j=1}^{\infty}\frac{\zeta^2\lambda_j u_j}{\zeta^2\lambda_j+1}\varphi_j = \sum_{j=1}^{\infty}u_j\varphi_j - \frac{u_j}{\zeta^2\lambda_j+1}\varphi_j = u-(\zeta^2\mathcal{L}+\mathrm{I})^{-1}u.$$

Therefore,

$$\frac{\sin(\pi s)}{\pi}\int_0^\infty \zeta^{s-1}(\mathcal{L}+\zeta\,\mathrm{I})^{-1}\mathcal{L}u\,d\zeta = \frac{2\sin(\pi s)}{\pi}\int_0^\infty \zeta^{-2s-1}(u-(\zeta^2\mathcal{L}+\mathrm{I})^{-1}u)\,d\zeta.$$

Denoting with $g(\zeta)$ the integrand on the right-hand side and $c_s := 2\sin(\pi s)/\pi$, we apply the third property in Lemma 2.12 combined with Corollary 3.18 to conclude for all $v\in H_0^s(\Omega)$

$$\begin{aligned}\left(c_s\int_0^\infty g(\zeta)\,d\zeta, v\right)_{L^2(\Omega)} &= c_s\int_0^\infty \zeta^{-2s-1}(u-(\zeta^2\mathcal{L}+\mathrm{I})^{-1}u, v)_{L^2(\Omega)}\,d\zeta\\ &= (u,v)_{H^s_{\mathcal{L}}(\Omega)} = (\mathcal{L}^s u, v)_{L^2(\Omega)}.\end{aligned}$$

Hence

$$\frac{\sin(\pi s)}{\pi}\int_0^\infty \zeta^{s-1}(\mathcal{L}+\zeta\,\mathrm{I})^{-1}\mathcal{L}u\,d\zeta = c_s\int_0^\infty g(\zeta)\,d\zeta = \mathcal{L}^s u. \qquad \square$$

Inspection of the proof above shows that the integral on the right-hand side of (4.13) can be reformulated as

$$\mathcal{L}^s u = \frac{2\sin(\pi s)}{\pi}\int_0^\infty \zeta^{-2s-1}(u-(\zeta^2\mathcal{L}+\mathrm{I})^{-1}u)\,d\zeta = \frac{2\sin(\pi s)}{\pi}\int_0^\infty \zeta^{1-2s}(\zeta^2\mathcal{L}+\mathrm{I})^{-1}\mathcal{L}u\,d\zeta,$$

which is used in [DS19, DS21, DH21] and [BP15] as starting point to define the fractional powers of $\mathcal{L}$, respectively.

Note that integral formulas for $\mathcal{L}^{-s}$ can be deduced from existing representations of $\mathcal{L}^s$ using the second claim in Lemma 4.6. Invoking Theorem 4.9, we find

$$\mathcal{L}^{1-s} = \frac{\sin(\pi(1-s))}{\pi}\int_0^\infty \zeta^{-s}(\mathcal{L}+\zeta\,\mathrm{I})^{-1}\mathcal{L}\,d\zeta.$$

Applying $\mathcal{L}^{-1}$ on both sides combined with $\sin(\pi(1-s)) = \sin(\pi s)$, we arrive at the following theorem; cf. [BP15, Theorem 2.1].

Theorem 4.10. *Let $s\in(0,1)$ and $f\in L^2(\Omega)$. Then there holds*

$$\mathcal{L}^{-s}f = \frac{\sin(\pi s)}{\pi}\int_0^\infty \zeta^{-s}(\mathcal{L}+\zeta\,\mathrm{I})^{-1}f\,d\zeta. \tag{4.14}$$

The identity (4.14) is known as *Balakrishnan's formula* [Bal60], *Dunford-Taylor representation* [Yag10, Section 2.7 (p. 92)], or *Kato's formula* [Kat60, Theorem 2 with simplification $\lambda=0$]. It is arguably among the most prominent representations of inverse fractional diffusion operators and has been applied in e.g., [BP16, BGZ20, DAC+21, DH21, DHS21].

Needless to say, the presented collection of representation formulas for $\mathcal{L}^s$ is far from complete. One final identity we want to mention is based on the so-called *heat semigroup*

[AF03, Section 7] and relies on the Gamma function Γ and its properties. We apply $\Gamma(1-s) = -s\Gamma(-s)$ (cf. Section 2.3.2), integration by parts, and the substitution $\zeta = \lambda t$, $\lambda \in \mathbb{R}^+$, to deduce for all $s \in (0,1)$

$$\begin{aligned} -s\Gamma(-s) = \Gamma(1-s) &= \int_0^\infty \zeta^{-s} e^{-\zeta}\, d\zeta \\ &= -\int_0^\infty \zeta^{-s} \frac{d}{d\zeta}\left(e^{-\zeta} - 1\right) d\zeta \\ &= -s\int_0^\infty \zeta^{-s-1}(e^{-\zeta} - 1)\, d\zeta = -s\lambda^{-s}\int_0^\infty t^{-s-1}(e^{-\lambda t} - 1)\, dt. \end{aligned}$$

Rearrangement of the terms yields

$$\lambda^s = \frac{1}{\Gamma(-s)} \int_0^\infty t^{-s-1}(e^{-\lambda t} - 1)\, dt, \qquad \lambda \in \mathbb{R}^+. \tag{4.15}$$

Purely formally for now, we replace λ with $\mathcal{L}$ to obtain a possible definition of $\mathcal{L}^s$ in terms of the Bochner integral

$$\frac{1}{\Gamma(-s)} \int_0^\infty t^{-s-1}(e^{-t\mathcal{L}}u(\mathbf{x},t) - u(\mathbf{x},t))\, dt, \tag{4.16}$$

where $U(\mathbf{x},t) := e^{-t\mathcal{L}}u(\mathbf{x})$ is the heat-semigroup, that is, the unique solution of the heat equation

$$\partial_t U + \mathcal{L}U = 0, \quad \text{in } \Omega \times \mathbb{R}^+, \tag{4.17a}$$
$$U = 0, \quad \text{on } \partial\Omega \times \mathbb{R}^+, \tag{4.17b}$$
$$U = u, \quad \text{on } \Omega \times \{0\}. \tag{4.17c}$$

The formula (4.16) has been applied in [CdTGG20] to introduce fractional powers of the Laplacian. After spectral decomposition, the identity (4.15) can be applied to (4.16) to show that the operator so obtained matches our understanding of $\mathcal{L}^s$.

Theorem 4.11. *Let $s \in (0,1)$ and $u \in H_0^s(\Omega)$. Then there holds*

$$\mathcal{L}^s u(\mathbf{x}) = \frac{1}{\Gamma(-s)} \int_0^\infty t^{-s-1}(U(\mathbf{x},t) - u(\mathbf{x},t))\, dt,$$

where U is the solution to (4.17).

Proof. See [ST10, Sti10] for a detailed proof. □

The counterpart to (4.15) for negative exponents reads

$$\lambda^{-s} = \frac{1}{\Gamma(s)} \int_0^\infty t^{s-1} e^{t\lambda}\, dt$$

and gives rise to the heat-semigroup representation for the inverse fractional diffusion operator

$$\mathcal{L}^{-s} f = \frac{1}{\Gamma(s)} \int_0^\infty t^{1-s} e^{-t\mathcal{L}} f\, dt,$$

where $e^{-t\mathcal{L}}f(\mathbf{x})$ satisfies (4.17) with f in place of u. This once more shows that the nonlocality of $\mathcal{L}^s$ can be compensated by means of an improper integral over local problems.

4.1.3 Harmonic Extension

The previous two sections show that $\mathcal{L}^s$ can be expressed as infinite series over the family of eigenpairs in (4.9) or as weighted integral over solutions to standard PDEs. An inherently different point of view is gained by means of the trace method. Initiated by the work of Caffarelli and Silvestre [CS07] and contributors [ST10, CT10, CDDS11, BCdPS13], it has been a driving force in the study of fractional diffusion problems [NOS15, BMN+18, AG18, ACN19, MR20b, BMS20]. We recall the definition of the extension space $\mathbb{V}_s(\overline{\mathcal{H}})$ of the couple $\overline{\mathcal{H}} = ((L^2(\Omega), \|\cdot\|_{L^2(\Omega)}), (H^1_0(\Omega), \|\cdot\|_{H^1_{\mathcal{L}}(\Omega)}))$ as

$$\mathbb{V}_s(\overline{\mathcal{H}}) = H^1_s(\mathbb{R}^+; L^2(\Omega)) \cap L^2_s(\mathbb{R}^+; H^1_0(\Omega))$$

If we consider the elements of $\mathbb{V}_s(\overline{\mathcal{H}})$ as functions in the extended variable $(\mathbf{x}, \zeta) \in \Omega \times \mathbb{R}^+$ with values in $\mathbb{R}$ rather than as a function in $\zeta \in \mathbb{R}^+$ with Hilbert-valued range, we can write its norm as

$$\begin{aligned}
\|v\|^2_{\mathbb{V}_s(\overline{\mathcal{H}})} &= \int_0^\infty \zeta^{1-2s} \left(\|v(\mathbf{x},\zeta)\|_{H^1_{\mathcal{L}}(\Omega)} + \|\partial_\zeta v(\mathbf{x},\zeta)\|_{L^2(\Omega)} \right) d\zeta \\
&= \int_0^\infty \zeta^{1-2s} \int_\Omega \|\nabla_{\mathbf{x}} v(\mathbf{x},\zeta)\|^2_{\mathfrak{A}(\mathbf{x})} + \mathfrak{c}(\mathbf{x}) |v(\mathbf{x},\zeta)|^2 + |\partial_\zeta v(\mathbf{x},\zeta)|^2 \, d\mathbf{x} \, d\zeta \\
&= \int_{\mathcal{C}_\Omega} \zeta^{1-2s} \left(\mathfrak{c}(\mathbf{x}) |v(\mathbf{x},\zeta)|^2 + \|\nabla_{\mathbf{x}} v(\mathbf{x},\zeta)\|^2_{\mathfrak{A}(\mathbf{x})} + |\partial_\zeta v(\mathbf{x},\zeta)|^2 \right) d(\mathbf{x},\zeta),
\end{aligned}$$

where $\|\mathbf{y}\|^2_{\mathfrak{A}(\mathbf{x})} := \mathbf{y}^T \mathfrak{A}(\mathbf{x}) \mathbf{y}$, $\nabla_{\mathbf{x}}$ denotes the gradient with respect to the $\mathbf{x}$-variable, and $\mathcal{C}_\Omega := \Omega \times \mathbb{R}^+$ a semi-infinite cylinder. Introducing the quantities

$$\hat{\mathfrak{c}}(\mathbf{x},\zeta) := \mathfrak{c}(\mathbf{x}), \qquad \hat{\mathfrak{A}}(\mathbf{x},\zeta) := \begin{pmatrix} \mathfrak{A}(\mathbf{x}) & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}, \tag{4.18}$$

we see that $\|v\|^2_{\mathbb{V}_s(\overline{\mathcal{H}})}$ can be written in more succinct form

$$\|v\|^2_{\mathbb{V}_s(\overline{\mathcal{H}})} = \int_{\mathcal{C}_\Omega} \zeta^{1-2s} \left(\hat{\mathfrak{c}} \|v\|^2_2 + \|\nabla v\|^2_{\hat{\mathfrak{A}}} \right) d(\mathbf{x},\zeta) =: \|v\|^2_{H^1_{\mathcal{L}^s}(\mathcal{C}_\Omega)}.$$

To emphasize this changed point of view, we write $\mathring{H}^1_s(\mathcal{C}_\Omega)$ instead of $\mathbb{V}_s(\overline{\mathcal{H}})$ henceforth. With this in mind, we recall that the s-minimal extension $\mathcal{U} \in \mathring{H}^1_s(\mathcal{C}_\Omega)$ of $u \in H^s_0(\Omega)$ satisfies

$$\|\mathcal{U}\|_{H^1_{\mathcal{L}^s}(\mathcal{C}_\Omega)} = \inf_{\substack{v \in \mathring{H}^1_s(\mathcal{C}_\Omega) \\ \operatorname{tr}_0 v = u}} \|v\|_{H^1_{\mathcal{L}^s}(\mathcal{C}_\Omega)},$$

where $\operatorname{tr}_0 v(\mathbf{x},\zeta) = v(\mathbf{x},0)$ is the trace operator evaluating v at the bottom of the cylinder $\mathcal{C}_\Omega$. Standard tools from calculus of variation reveal that $\mathcal{U}$ is a solution to the PDE: Find $\mathcal{U} \in \mathring{H}^1_s(\mathcal{C}_\Omega)$ such that

$$-\mathcal{L}\mathcal{U} + \frac{1-2s}{\zeta} \partial_\zeta \mathcal{U} + \partial^2_\zeta \mathcal{U} = 0, \quad \text{in } \mathcal{C}_\Omega, \tag{4.19a}$$

$$\mathcal{U} = 0, \quad \text{on } \partial\Omega \times \mathbb{R}^+, \tag{4.19b}$$

$$\mathcal{U} = u, \quad \text{on } \Omega \times \{0\}, \tag{4.19c}$$

where the derivatives are understood in the weak sense. Note that the PDE (4.19a) can be written in compact form

$$\operatorname{div}(\zeta^{1-2s}\hat{\mathfrak{A}}\nabla\mathcal{U}) + \zeta^{1-2s}\hat{\mathfrak{c}}\,\mathcal{U} = 0, \tag{4.20}$$

which is a standard elliptic PDE and thus uniquely solvable. Recognizing this fact, we introduce the *conormal derivative of* $\mathcal{U}$ as

$$\frac{\partial\mathcal{U}}{\partial\mathbf{n}_s}(\mathbf{x},\zeta) := -\lim_{\zeta\to 0^+}\zeta^{1-2s}\partial_\zeta\mathcal{U}(\mathbf{x},\zeta), \tag{4.21}$$

where $\mathbf{n}$ denotes the unit outer normal to $\mathcal{C}_\Omega$ at $\Omega\times\{0\}$. The quantity (4.21) can be seen as weighted normal derivative of $\mathcal{U}$ and allows us to write the fractional Laplacian in the following convenient manner; cf. [CS07, CDDS11].

Theorem 4.12. *Let $s\in(0,1)$, d_s defined as in Lemma 3.24, $u\in H^s_0(\Omega)$, and $\mathcal{U}$ the unique solution to (4.19). Then there holds*

$$\mathcal{L}^s u = -\frac{1}{d_s}\frac{\partial\mathcal{U}}{\partial\mathbf{n}_s}. \tag{4.22}$$

Proof. By definition of the s-minimal extension of u there holds

$$\mathcal{U}(\zeta) = \sum_{j=1}^{\infty} u_j\phi_j(\zeta)\varphi_j,$$

where ϕ_j is as in (3.27). Thanks to Lemma 3.24 we find

$$-\frac{1}{d_s}\lim_{\zeta\to 0^+}\zeta^{1-2s}\partial_\zeta\phi_j(\zeta) = \lambda_j^s,$$

so that (4.22) follows by superposition. □

The harmonic extension problem (4.19) is illustrated in Figure 4.1. It provides a powerful tool to localize $\mathcal{L}^s u$, so that the value of $\mathcal{L}^s u(\mathbf{x})$ only depends on the behaviour of $\mathcal{U}(\mathbf{x},\zeta)$ in an arbitrary small neighborhood of $(\mathbf{x},0)$. Note that for $s=\frac{1}{2}$ the PDE (4.19) is a standard elliptic problem (with weight equal to one) on the semi-infinite cylinder $\mathcal{C}_\Omega$. According to Theorem 4.12, the operator $\mathcal{L}^{\frac{1}{2}}$ maps the Dirichlet data u to the normal derivative of the solution $\mathcal{U}$ at the bottom of $\mathcal{C}_\Omega$. The same holds true if $s\neq\frac{1}{2}$ in consideration of the weight function ζ^{1-2s}, whence $\mathcal{L}^s$ can be seen as a *Dirichlet-to-Neumann operator*.

Under suitable manipulations of the boundary condition (4.19c), the action of the inverse fractional operator is obtained in a similar fashion. Consider the mixed boundary value problem: Find $\mathcal{U}\in\mathring{H}^1_s(\mathcal{C}_\Omega)$ such that

$$\operatorname{div}(\zeta^{1-2s}\hat{\mathfrak{A}}\nabla\mathcal{U}) + \zeta^{1-2s}\hat{\mathfrak{c}}\,\mathcal{U} = 0, \qquad \text{in } \mathcal{C}_\Omega, \tag{4.23a}$$

$$\mathcal{U} = 0, \qquad \text{on } \partial\Omega\times\mathbb{R}^+, \tag{4.23b}$$

$$\frac{\partial\mathcal{U}}{\partial\mathbf{n}_s} = d_s f, \quad \text{on } \Omega\times\{0\}, \tag{4.23c}$$

for some $f\in H^{-s}(\Omega)$, where d_s is the normalization from Lemma 3.24. Then $u=\mathcal{L}^{-s}f$ is recovered from (4.23) via trace evaluation [CDDS11, Lemma 2.2].

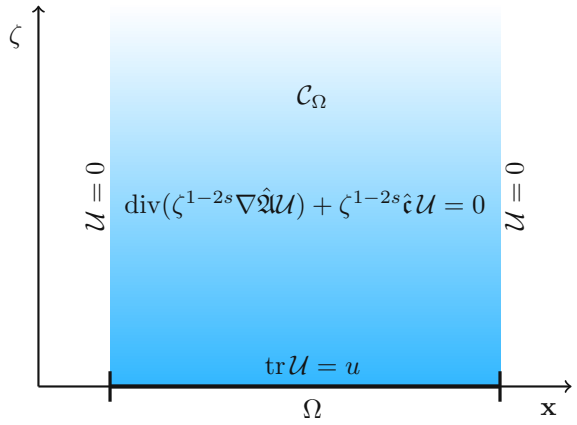

Figure 4.1: Harmonic extension problem in the semi-infinite cylinder $\mathcal{C}_\Omega = \Omega \times \mathbb{R}^+$.

Theorem 4.13. *Let $s \in (0,1)$, $f \in H^{-s}(\Omega)$, and $\mathcal{U}$ the unique solution to (4.23). Then there holds*

$$\mathcal{L}^{-s} f(\mathbf{x}) = (\operatorname{tr}_0 \mathcal{U})(\mathbf{x}) = \mathcal{U}(\mathbf{x}, 0).$$

4.2 The Fractional Diffusion Problem

We are interested in solutions to nonlocal PDEs of the form: Find $u \in H^s_0(\Omega)$ such that

$$\mathcal{L}^s u = f, \tag{4.24}$$

where $s \in (0,1)$ and $f \in H^{-s}(\Omega)$. To derive a variational formulation for this problem, we apply both sides of (4.24) to an arbitrary test function $v \in H^s_0(\Omega)$ and invoke (4.8) to deduce

$$\forall v \in H^s_0(\Omega): \quad (u, v)_{H^s_{\mathcal{L}}(\Omega)} = \langle f, v \rangle. \tag{4.25}$$

Existence and uniqueness of solutions to this problem are the subject of the following theorem.

Theorem 4.14. *Let $s \in (0,1)$ and $f \in H^{-s}(\Omega)$. Then there exists a unique solution $u \in H^s_0(\Omega)$ to (4.25) which is given by*

$$u = \sum_{j=1}^{\infty} \lambda_j^{-s} \langle f, \varphi_j \rangle \varphi_j \tag{4.26}$$

and there holds the stability estimate

$$\|u\|_{H^s_0(\Omega)} \preceq \|f\|_{H^{-s}(\Omega)}.$$

Proof. We plug (4.26) into (4.25), choose $v = \varphi_i$ for some $i \in \mathbb{N}$, and consult (4.9) to confirm that

$$(u, \varphi_i)_{H^s_{\mathcal{L}}(\Omega)} = \sum_{j=1}^{\infty} \lambda_j^{-s} \langle f, \varphi_j \rangle (\varphi_j, \varphi_i)_{H^1_{\mathcal{L}}(\Omega)} = \sum_{j=1}^{\infty} \langle f, \varphi_j \rangle (\varphi_j, \varphi_i)_{L^2(\Omega)} = \langle f, \varphi_i \rangle.$$

Since $(\varphi_j)_{j=1}^{\infty}$ is a basis of $H^s_0(\Omega)$, it follows that u indeed solves (4.25). Its uniqueness and the stability estimate are a direct consequence of the Lax-Milgram theorem. □

The previous theorem shows that the inverse fractional diffusion operator $\mathcal{L}^{-s} : H^{-s}(\Omega) \to H^s_0(\Omega)$ is an operator of order $2s$. This "shift property", however, holds only in a limited range. As shown in [Gru16] for the fractional Laplacian, see also [LPG+20], the expected shift is essentially valid until the regularity $H^{s+\frac{1}{2}}(\Omega)$ is reached. Here, we again make use of the extended definition of fractional Sobolev spaces for $s > 1$ in the sense of Remark 3.35.

Theorem 4.15. *Let $\Omega \subset \mathbb{R}^d$ be a C^∞-domain, $s \in (0,1)$, $\rho \in (\frac{1}{2}, 2 + \frac{1}{2})$, $f \in H^\rho(\Omega)$, and u the unique solution to (4.25) with $\mathcal{L} = -\Delta$. Then there holds for all $\varepsilon > 0$*

$$u \in \begin{cases} H_0^{\rho+2s}(\Omega), & \text{if } \rho \leq \frac{1}{2}, \\ H_0^{\frac{1}{2}-\varepsilon+2s}(\Omega), & \text{if } \rho \in (\frac{1}{2}, 2 + \frac{1}{2}). \end{cases}$$

Moreover, $\rho \in (\frac{1}{2}, 2 + \frac{1}{2})$ implies $u \in H_0^{\rho+2s}(\Omega)$ if and only if $f = 0$.

The statement of Theorem 4.15 stands in strong contrast to the classical shift theorem known from the integer-order case. The latter states that the smoothness of the domain and the data implies the smoothness of the solution itself. However, even if $\Omega = B_1(0)$ is the unit ball and $f \in C^\infty(\Omega)$, the solution to (4.24) is not expected to be more regular than $H_0^{\frac{1}{2}+2s}(\Omega)$. To see that this "$s$-gain" is indeed sharp, we consider the fractional Laplacian on $\Omega = (0, 2\pi)$ in which case the eigenfunctions and eigenvalues with homogeneous Dirichlet boundary conditions are given by $\varphi_j(x) = \sin(\pi x)$ and $\lambda_j = j^2$. Following [LPG+20], we define

$$f(x) := \sum_{j=1}^{\infty} \frac{1}{\sqrt{j} \ln(j+1)} \sin(jx). \tag{4.27}$$

By means of the substitution $y = \ln(x+1)$ it follows $f \in L^2(\Omega)$ since

$$\|f\|^2_{L^2(\Omega)} = \sum_{j=1}^{\infty} \frac{1}{j \ln^2(j+1)} \leq \int_1^{\infty} \frac{dx}{x \ln^2(x+1)} = \int_{\ln(2)}^{\infty} \frac{dy}{(e^y - 1) y^2} < \infty.$$

But $f \notin H^\rho_0(\Omega)$ for any $\rho \in \mathbb{R}^+$ since

$$\|f\|^2_{H^\rho_0(\Omega)} = \sum_{j=1}^{\infty} \frac{j^{2\rho}}{j \ln^2(j+1)} \geq \sum_{j=1}^{\infty} \frac{1}{j} = \infty.$$

The solution to (4.24) with f as in (4.27) is given by

$$u(x) = \sum_{j=1}^{\infty} \frac{j^{-2s}}{\sqrt{j}\ln(j+1)} \sin(jx)$$

and $\|u\|_{H_0^{2s+\rho}(\Omega)} = \|f\|_{H_0^{\rho}(\Omega)}$ for any $\rho \in \mathbb{R}^+$. The latter implies that $u \in H_0^{2s}(\Omega)$ and $u \notin H_0^{2s+\varepsilon}(\Omega)$ for any $\varepsilon > 0$, that is, the regularity estimate stated above cannot be improved.

The lack of regularity is essentially due to the behaviour of u close to the boundary of Ω. In the *interior* of Ω, it is known that the smoothness of the data implies smoothness of the solution. For $\mathbf{x}$ close to $\partial\Omega$, however, there holds

$$u(\mathbf{x}) \sim \begin{cases} \operatorname{dist}(\mathbf{x}, \partial\Omega)^{2s} + v(\mathbf{x}), & \text{if } s \in (0, \frac{1}{2}), \\ \operatorname{dist}(\mathbf{x}, \partial\Omega)|\ln(\operatorname{dist}(\mathbf{x}, \partial\Omega))| + v(\mathbf{x}), & \text{if } s = \frac{1}{2}, \\ \operatorname{dist}(\mathbf{x}, \partial\Omega) + v(\mathbf{x}), & \text{if } s \in (\frac{1}{2}, 1), \end{cases} \tag{4.28}$$

for some smooth function v defined on Ω; see [Gru16] for the case where Ω is a C^∞-domain and [CS16] when only limited regularity is available. The singular behaviour of u should be taken into account in the design of numerical schemes, e.g., in terms of a refined mesh towards the boundary of Ω.

We conclude this section with a comparison of solutions to (4.25) for different values of the fractional parameter on the L-shape domain $\Omega = (0,1)^2 \setminus ([0.5,1] \times [0,0.5])$ with $f = 1$. The singular behaviour of u predicted by (4.28) is mirrored in Figure 4.2. While u is essentially smooth in the interior of Ω, its limited regularity becomes visible in a neighborhood of $\partial\Omega$. For s close to zero, the fractional Laplacian $(-\Delta)^s$ is close to the identity operator, whence the solution is nearly constant in the interior. Towards the boundary, the zero trace is imposed which forces $(-\Delta)^s u(\mathbf{x})$ to decrease rapidly as $\mathbf{x}$ approaches $\partial\Omega$.

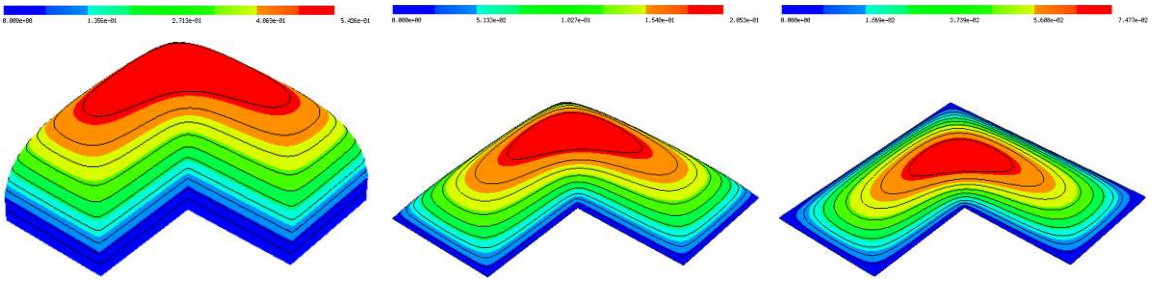

Figure 4.2: Solution u to (4.25) for $\mathcal{L} = -\Delta$ and $s = 0.2$ (left), $s = 0.5$ (middle), and $s = 0.8$ (right) on the L-shape domain.

4.3 Non-Equivalent Definitions of the Fractional Laplacian

The definition of $\mathcal{L}^s$ as interpolation operator is a natural one. There exist, however, several reasonable ways to give meaning to the fractional powers of differential operators that are

mathematically distinct from the one obtained by Definition 4.2. We mention here only two that frequently appear in the literature and restrict ourselves to $\mathcal{L} = -\Delta$. As preparation, we provide the following definition.

Definition 4.16. *Let $\mathbf{x} \in \mathbb{R}^d$ and u a function defined on $\mathbb{R}^d \setminus \{\mathbf{x}\}$ such that $u \in L_1(\mathbb{R}^d \setminus B_\varepsilon(\mathbf{x}))$ for all $\varepsilon > 0$. We define the Cauchy principal value of u by*

$$\text{P.V.} \int_{\mathbb{R}^d} u(\mathbf{y})\, d\mathbf{y} := \lim_{\varepsilon \to 0} \int_{\mathbb{R}^d \setminus B_\varepsilon(\mathbf{x})} u(\mathbf{y})\, d\mathbf{y}.$$

The Cauchy principal value allows one to assign values to certain integrals which do not exist in the classical sense of Lebesgue. Consider e.g., the scalar function $u(x) = x^{-2}$, which is not integrable on $\mathbb{R}$, i.e., the limits

$$\lim_{\varepsilon \to 0^-} \int_{-\infty}^{\varepsilon} \frac{1}{x^2}\, dx, \qquad \lim_{\varepsilon \to 0^+} \int_{\varepsilon}^{\infty} \frac{1}{x^2}\, dx,$$

and therefore

$$\int_{\mathbb{R}} \frac{1}{x^2}\, dx = \lim_{\varepsilon \to 0^-} \int_{-\infty}^{\varepsilon} \frac{1}{x^2}\, dx + \lim_{\varepsilon \to 0^+} \int_{\varepsilon}^{\infty} \frac{1}{x^2}\, dx$$

do not exist. Due to symmetry, however, the singularity averages out when the Cauchy principle value is employed

$$\text{P.V.} \int_{\mathbb{R}} \frac{1}{x^2}\, dx = \lim_{\varepsilon \to 0} \left(\int_{-\infty}^{-\varepsilon} \frac{1}{x^2}\, dx + \int_{\varepsilon}^{\infty} \frac{1}{x^2}\, dx \right) = 0.$$

Clearly, if an integral exists in the classical sense, it coincides with its principal value.

4.3.1 The Integral Fractional Laplace

We introduce the so-called *integral fractional Laplacian* as singular integral operator on Ω [NPV12, LPG+20], where the zero extension $\tilde{u}$ of u is defined by

$$\tilde{u}(\mathbf{x}) := \begin{cases} u(\mathbf{x}), & \mathbf{x} \in \Omega, \\ 0, & \mathbf{x} \in \mathbb{R}^d \setminus \Omega. \end{cases}$$

Definition 4.17. *For all $s \in (0,1)$ we define the integral fractional Laplacian by*

$$(-\Delta)_I^s u(\mathbf{x}) := C_{d,s}\, \text{P.V.} \int_{\mathbb{R}^d} \frac{\tilde{u}(\mathbf{x}) - \tilde{u}(\mathbf{y})}{\|\mathbf{x} - \mathbf{y}\|_2^{d+2s}}\, d\mathbf{y}, \qquad C_{d,s} := 2^{2s} \frac{s\Gamma(s + \frac{d}{2})}{\pi^{\frac{d}{2}} \Gamma(1-s)}. \tag{4.29}$$

The integral fractional Laplacian is well-defined for all $u \in C_0^\infty(\Omega)$ and extends to $H_0^s(\Omega)$ by density. If $s \in [\frac{1}{2}, 1)$, the integrand in (4.29) is not contained in $L^1(\mathbb{R}^d)$ due to the singularity at $\mathbf{y} = \mathbf{x}$. Thanks to the symmetrical structure of the integral, however, the singularity averages out, causing the limit in (4.29) to be finite. If $s \in (0, \frac{1}{2})$, the integral

$$\int_{\mathbb{R}^d} \frac{\tilde{u}(\mathbf{x}) - \tilde{u}(\mathbf{y})}{\|\mathbf{x} - \mathbf{y}\|_2^{d+2s}}\, d\mathbf{y}$$

exists as improper integral and the Cauchy principal value can be omitted. An equivalent representation of $(-\Delta)_I^s$ that works for all $s \in (0,1)$ without Cauchy's principal value reads

$$(-\Delta)_I^s u(\mathbf{x}) = -\frac{C_{d,s}}{2} \int_{\mathbb{R}^d} \frac{\tilde{u}(\mathbf{x}+\mathbf{y}) - 2\tilde{u}(\mathbf{x}) + \tilde{u}(\mathbf{x}-\mathbf{y})}{\|\mathbf{y}\|_2^{d+2s}} \, d\mathbf{y}. \tag{4.30}$$

The interested reader is directed to [NPV12] for further details. Just like for $(-\Delta)^s$, the value of $(-\Delta)_I^s u(\mathbf{x})$ depends not only on the behaviour of u near $\mathbf{x}$ but on the values of u in the entire domain, rendering the operator to be nonlocal. In particular, the function u might be identically zero in a neighborhood of $\mathbf{x}$ but $(-\Delta)^s u(\mathbf{x}) \neq 0$. Consider e.g., a nonnegative function $u \in C_0^\infty(\Omega)$ and $\mathbf{x} \in \mathbb{R}^d \setminus \operatorname{supp} u$. Then

$$\begin{aligned}(-\Delta)^s u(\mathbf{x}) &= C_{d,s} \, \mathrm{P.V.} \int_{\operatorname{supp} u} \frac{u(\mathbf{x}) - u(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}} \, d\mathbf{y} \\ &= C_{d,s} \, \mathrm{P.V.} \int_{\operatorname{supp} u} \frac{-u(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}} \, d\mathbf{y} < 0.\end{aligned}$$

While obvious for $(-\Delta)^s$, it is not clear at first sight why (4.29) should yield a reasonable generalization of the classical Laplacian. One prominent approach to motivate (4.29) as variant of the fractional Laplacian is based on a probabilistic model [Val09, BV16] which describes the random walk of a particle with jumps of arbitrary step size. To see this, let us start with the integer case $s = 1$, $\Omega = \mathbb{R}$, a prescribed time step $\tau \in \mathbb{R}^+$, and a step size $h \in \mathbb{R}^+$. Let $u(x,t)$ denote the probability of a particle being at point x at time t. If we assume that the particle jumps with probability $\frac{1}{2}$ either to its left or its right, we have

$$u(x, t+\tau) = \frac{1}{2}u(x+h,t) - \frac{1}{2}u(x-h,t).$$

Assuming the relation $2\tau = h^2$, we obtain

$$\frac{u(x,t+\tau) - u(x,t)}{\tau} = \frac{u(x+h,t) - 2(u,x,t) + u(x-h,t)}{h^2}.$$

Letting τ and h approach zero, we arrive at the well-known *heat equation*

$$\partial_t u = \Delta u.$$

The integral fractional Laplacian allows jumps of arbitrary step size to enter the model. For this purpose, let $s \in (0,1)$ and denote with P a probability measure on $\mathbb{N}$ defined by

$$\mathrm{P}(\mathcal{N}) := c_s \sum_{j \in \mathcal{N}} \frac{1}{j^{1+2s}}, \qquad c_s := \left(\sum_{j=1}^{\infty} \frac{1}{j^{1+2s}} \right)^{-1},$$

for any $\mathcal{N} \subset \mathbb{N}$. As illustrated in Figure 4.3, the motion of the particle in $\mathbb{R}^d$ at each time step τ is then defined by $jh\mathbf{n}$, where $\mathbf{n} \in \partial B_1(0)$ is a random direction chosen according to a uniform distribution, $j \in \mathbb{N}$ a random parameter chosen according to the probability

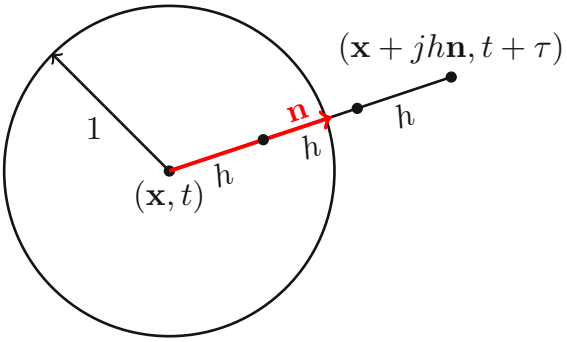

Figure 4.3: Random walk of a particle at time t and $t+\tau$.

distribution P, and $h \in \mathbb{R}$ the minimal step size. The probability of the particle being at point x at time $t+\tau$ now reads

$$u(\mathbf{x}, t+\tau) = \frac{c_s}{|\partial B_1(0)|} \int_{\partial B_1(0)} \sum_{j\in\mathbb{N}} \frac{u(\mathbf{x}+jh\mathbf{n}, t)}{j^{1+2s}} d\sigma, \tag{4.31}$$

where $\int_{\partial B_1(0)} \cdot\, d\sigma$ denotes the surface integral over the unit sphere. The right-hand side of (4.31) can be seen as the sum over all probabilities of the particle being at time t at any other location $\mathbf{x}+jh\mathbf{n}$, for some direction $\mathbf{n} \in B_1(0)$ and $j \in \mathbb{N}$, times the probability of jumping from there to $\mathbf{x}$. The change of variable $\mathbf{n} \mapsto -\mathbf{n}$ gives

$$u(\mathbf{x}, t+\tau) = \frac{c_s}{|\partial B_1(0)|} \int_{\partial B_1(0)} \sum_{j\in\mathbb{N}} \frac{u(\mathbf{x}-jh\mathbf{n}, t)}{j^{1+2s}} d\sigma,$$

so that

$$u(\mathbf{x}, t+\tau) = \frac{c_s}{2|\partial B_1(0)|} \int_{\partial B_1(0)} \sum_{j\in\mathbb{N}} \frac{u(\mathbf{x}+jh\mathbf{n}, t) + u(\mathbf{x}-jh\mathbf{n}, t)}{j^{1+2s}} d\sigma.$$

Since $c_s/(2|\partial B_1(0)|)$ normalizes the above integral, we derive

$$u(\mathbf{x}, t+\tau) - u(\mathbf{x}, t) = \frac{c_s}{2|\partial B_1(0)|} \int_{\partial B_1(0)} \sum_{j\in\mathbb{N}} \frac{u(\mathbf{x}+jh\mathbf{n}, t) + u(\mathbf{x}-jh\mathbf{n}, t) - 2u(\mathbf{x})}{j^{1+2s}} d\sigma.$$

Setting $\tau = h^{2s}$, we obtain

$$\frac{u(\mathbf{x}, t+\tau) - u(\mathbf{x}, t)}{\tau} = \frac{c_s}{2|\partial B_1(0)|} \int_{\partial B_1(0)} \sum_{j\in\mathbb{N}} h \frac{u(\mathbf{x}+jh\mathbf{n}, t) + u(\mathbf{x}-jh\mathbf{n}), t) - 2u(\mathbf{x})}{(jh)^{1+2s}} d\sigma.$$

For fixed $\mathbf{n}$, we recognize the sum as Riemann sum for an integral. Under the given assumptions on τ and h, we take the limit $h \to 0$ and use polar coordinates to finally arrive at

$$\begin{aligned}
\partial_t u(\mathbf{x}, t) &= \frac{c_s}{2|\partial B_1(0)|} \int_{\partial B_1(0)} \int_0^\infty \frac{u(\mathbf{x}+r\mathbf{n}, t) + u(\mathbf{x}-r\mathbf{n}, t) - 2u(\mathbf{x}, t)}{|r|^{1+2s}} \, dr \, d\mathbf{n} \\
&= \frac{c_s}{2|\partial B_1(0)|} \int_{\mathbb{R}^d} \frac{u(\mathbf{x}+\mathbf{y}, t) - 2u(\mathbf{x}, t) + u(\mathbf{x}-\mathbf{y}, t)}{\|\mathbf{y}\|_2^{d+2s}} \, d\mathbf{y}.
\end{aligned}$$

The constant in front of the integral evaluates to $c_s/(2\partial B_1(0)) = C_{d,s}$. In view of (4.30), the expression above is precisely what is defined in (4.29) and therefore provides a reasonable definition of the fractional Laplacian on $\Omega = \mathbb{R}^d$. If $\Omega \subsetneq \mathbb{R}^d$ is a bounded subset of $\mathbb{R}^d$, then the definition of $(-\Delta)_I^s$ is still coherent with this probabilistic interpretation with the modification that particles hitting the boundary of Ω are destroyed. The particular choice of the normalization in (4.29) ensures that, under reasonable assumptions on u, one has [DL21]

$$\lim_{s\to 0^+} (-\Delta)_I^s u(\mathbf{x}) = u(\mathbf{x}), \qquad \lim_{s\to 1^-} (-\Delta)_I^s u(\mathbf{x}) = -\Delta u(\mathbf{x}).$$

Throughout the last two decades, the integral fractional Laplacian has evoked a significant amount of research activity [WGP17, BLP19a, FKM20, ADS21, FMP21] and appears to be, together with our definition of $(-\Delta)^s$, the most popular choice to define the fractional powers of the Laplacian. Worth mentioning, however, the operators $(-\Delta)^s$ and $(-\Delta)_I^s$ are indeed inherently different. As shown in [SEV14], the smallest eigenvalue of $(-\Delta)_I^s$ is known to be strictly smaller than the one of $(-\Delta)^s$. Furthermore, the eigenfunctions of the integral fractional Laplace are only Hölder continuous while the eigenfunctions of $(-\Delta)^s$ coincide with those of $-\Delta$ and thus are smooth if Ω is smooth. Finally, we refer to [CS16, Gru16] to see that solutions to the fractional Poisson problem involving the integral fractional Laplacian behave like

$$u(\mathbf{x}) \sim \operatorname{dist}(\mathbf{x}, \partial\Omega)^s + v(\mathbf{x})$$

for $\mathbf{x}$ close to $\partial\Omega$, where v is a smooth function defined on Ω.

4.3.2 The Regional Fractional Laplace

Somewhat unnatural, the integral fractional Laplacian accesses values of u outside its original domain of definition Ω. One possibility to overcome this inconvenience is provided by the *regional fractional Laplacian* (sometimes also called restricted fractional Laplacian), which restricts the integration domain in (4.29) to Ω [GM05, GM06, Gua06].

Definition 4.18. *For all $s \in (0,1)$ we define the regional fractional Laplacian by*

$$(-\Delta)_R^s u(\mathbf{x}) := C_{d,s}\,\mathrm{P.V.} \int_\Omega \frac{u(\mathbf{x}) - u(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}}\, d\mathbf{y}.$$

Both $(-\Delta)_I^s$ and $(-\Delta)_R^s$ are distinct mathematical objects since

$$\begin{aligned}
(-\Delta)_I^s u(\mathbf{x}) &= C_{d,s}\,\mathrm{P.V.} \int_{\mathbb{R}^d} \frac{\tilde{u}(\mathbf{x}) - \tilde{u}(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}}\, d\mathbf{y} \\
&= C_{d,s}\,\mathrm{P.V.} \int_\Omega \frac{u(\mathbf{x}) + u(\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}}\, d\mathbf{y} + C_{d,s} \int_{\mathbb{R}^d\setminus\Omega} \frac{u(\mathbf{x})}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}}\, d\mathbf{y} \\
&= (-\Delta)_R^s u(\mathbf{x}) + C_{d,s} \int_{\mathbb{R}^d\setminus\Omega} \frac{u(\mathbf{x})}{\|\mathbf{x}-\mathbf{y}\|_2^{d+2s}}\, d\mathbf{y}.
\end{aligned}$$

Unlike the integral fractional Laplacian, the random motion of particles modeled by $(-\Delta)_R^s$ allows for jumps from the exterior into the domain Ω, but are either reflected in Ω or killed when reaching the boundary $\partial\Omega$, see for instance [BBC03, CKS10, GM05]. For further details and nice surveys on differences and similarities between various versions of the fractional Laplacian we refer to the expositions [DWZ17, LPG+20, DL21].

5 Fractional Evolution Equations

The previous chapter deals with the fractional powers of the spatial differential operator $\mathcal{L}$. The purpose of this chapter is to give a brief survey over the field of fractional calculus in order to introduce the time-fractional derivative ∂_t^α of order $\alpha \in (0,1]$ in the sense of Caputo. Provided the terminology, we investigate fractional evolution equations of the form

$$\partial_t^\alpha u + \mathcal{L}^s u = f, \quad \text{in } \Omega \times (0,T), \tag{5.1a}$$
$$u = 0, \quad \text{on } \partial\Omega \times (0,T), \tag{5.1b}$$
$$u = u_0, \quad \text{on } \Omega \times \{0\}, \tag{5.1c}$$

where $T \in \mathbb{R}^+$, $s \in [0,1]$, $f \in L^\infty(0,T;L^2(\Omega))$, and $u_0 \in L^2(\Omega)$. We derive a variational formulation for this problem, prove its unique solvability, and provide an explicit representation of the solution in terms of the eigenfunctions of $\mathcal{L}$.

5.1 Fractional Calculus

Provided a smooth function u and some end time $T \in \mathbb{R}^+ \cup \{\infty\}$, which we assume to be fixed throughout this section, the subject of interest in fractional calculus is to interpolate the sequence of differentials and integrals

$$\dots \partial_t^2 u(t),\ \partial_t u(t),\ u(t),\ \int_0^t u(\tau)\,d\tau,\ \int_a^t \int_a^{\tau_1} u(\tau)\,d\tau\,d\tau_1, \dots, \quad t \in (0,T),$$

where $u : (0,T) \to \mathcal{H}$ is a function with values in a Hilbert space $\mathcal{H}$. The interest in these problems has deepened throughout the last decades; see [Baz01, Ana18, LP12] for the abstract Bochner framework and [Die10, Pod99, MR93, OS74, SKM93, Mai10, BDST12, FS21] for $\mathcal{H} = \mathbb{R}$. Our main focus lies in the definition of fractional time-derivatives ∂_t^α. One possible way of doing this, which is also the one we pursue in this thesis, comes in two stages. First, one introduces the fractional integral J^α of order α, to define, in the second stage, ∂_t^α as "inverse" of J^α.

5.1.1 Fractional Integrals

The driving force in the development of fractional integrals is the so-called *Cauchy formula for repeated integration*

$$J^n u(t) := \int_0^t \int_0^{\tau_1} \cdots \int_0^{\tau_{n-1}} f(\tau_n)\,d\tau_n \dots d\tau_2\,d\tau_1 = \frac{1}{(n-1)!} \int_0^t (t-\tau)^{n-1} u(\tau)\,d\tau, \tag{5.2}$$

see [SKM93, eq. (2.16)] for the scalar case and [Mai10, eq. (1.1)] for the Bochner framework. The formula holds for all $n \in \mathbb{N}$, $u \in C([0,T];\mathcal{H})$, and follows from a standard induction argument. Due to Theorem 2.22, (5.2) can be written in terms of the Gamma function

$$J^n u(t) = \frac{1}{\Gamma(n)} \int_0^t (t-\tau)^{n-1} u(\tau)\, d\tau. \tag{5.3}$$

Observing that there is no reason to restrict (5.3) to integer values of n only, we arrive at the following classical definition.

Definition 5.1. *Let $\alpha \in \mathbb{R}^+$ and $u \in L^1(0,T;\mathcal{H})$. We define the Riemann-Liouville fractional integral by*

$$J^\alpha u(t) := \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} u(\tau)\, d\tau, \qquad t \in (0,T). \tag{5.4}$$

By convention, we set $J^0 u(t) := u(t)$.

More succinctly, the Riemann-Liouville fractional integral can be written as convolution operator

$$J^\alpha u(t) = (u * \mathcal{K}_\alpha)(t), \qquad \mathcal{K}_\alpha(t) := \frac{t^{\alpha-1}}{\Gamma(\alpha)} \mathrm{I}. \tag{5.5}$$

Both u and $\mathcal{K}_\alpha$ are contained in $L^1(0,T;\mathcal{H})$. Hence, according to the fourth property in Lemma 2.12, J^α is well-defined and $J^\alpha u \in L^1(0,T;\mathcal{H})$. Due to (5.3), J^α coincides with the classical repeated integral whenever $\alpha \in \mathbb{N}$. Several well-known properties from standard integration theory are preserved by our generalization.

Lemma 5.2. *Let $\alpha \in \mathbb{R}_0^+$ and $u \in L^1(0,T;\mathcal{H})$. Then there holds*

1. *J^α is a linear operator,*
2. *$J^\alpha J^\beta u(t) = J^\beta J^\alpha u(t) = J^{\alpha+\beta} u(t)$ for $\beta \in \mathbb{R}^+$ and almost all $t \in (0,T)$,*
3. *for all $n \in \mathbb{N}_0$ with $n < \alpha$ and almost all $t \in (0,T)$*

$$\partial_t^n J^\alpha u = J^{\alpha-n} u. \tag{5.6}$$

Proof. The first property is clear. The second is a direct consequence of (5.5) and the fact that the convolution is commutative, associative, and $\mathcal{K}_\alpha * \mathcal{K}_\beta = \mathcal{K}_{\alpha+\beta}$. The third one holds since, by the second property,

$$\partial_t^n J^\alpha u = \partial_t^n J^n J^{\alpha-n} u = J^{\alpha-n} u. \qquad \square$$

We mention that even for smooth functions, the order of differentiation and integration in (5.6) cannot be interchanged without further ado. To provide a more intuitive understanding of fractional integrals, we consider the following example, where we set $\mathcal{H} = \mathbb{R}$.

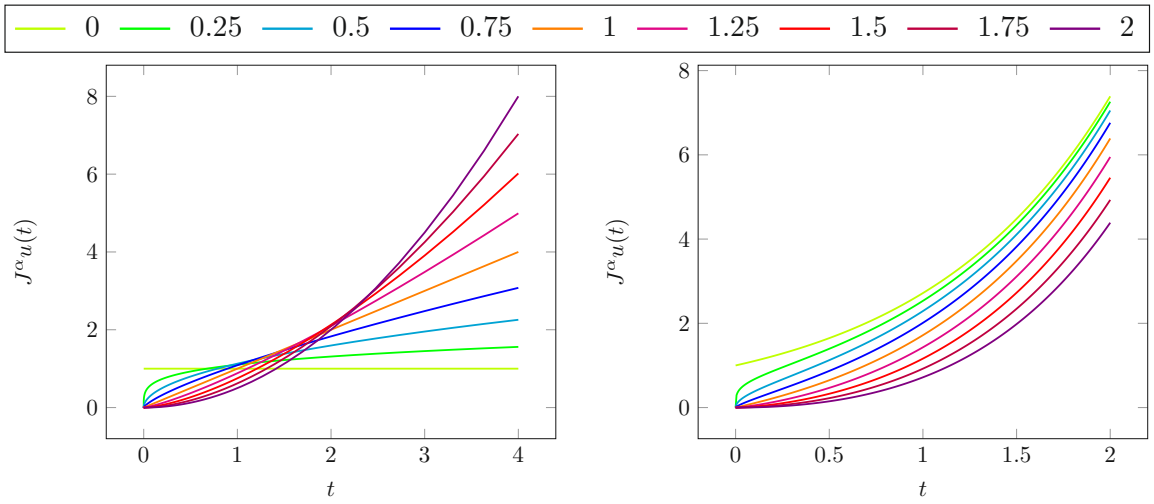

Figure 5.1: Riemann-Liouville fractional integral $J^\alpha u(t)$ for $u(t) = 1$ (left) and $u(t) = e^t$ (right) and different orders $\alpha \in [0, 2]$.

Example 5.3. *Let $s > -1$ and $u(t) = t^s$. Since $u \in L^1(0,T) := L^1(0,T;\mathbb{R})$, its Riemann-Liouville fractional integral $J^\alpha u$ is well-defined for all $\alpha \in \mathbb{R}_0^+$ and can be computed explicitly. After the substitution $\zeta = \frac{\tau}{t}$, we apply Lemma 2.24 to infer*

$$\begin{aligned} J^\alpha u(t) &= \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} \tau^s \, d\tau \\ &= \frac{1}{\Gamma(\alpha)} t^{s+\alpha} \int_0^1 (1-\zeta)^{\alpha-1} \zeta^s \, d\zeta = \frac{\Gamma(s+1)}{\Gamma(s+\alpha+1)} t^{s+\alpha}, \end{aligned}$$

which can be seen as a straightforward generalization of the integer-order case. We illustrate the fractional integrals of u for $s = 0$ and different orders α in Figure 5.1. In accordance with Definition 5.1, we observe that for all positive values of α the fractional integral is zero in $t = 0$. For small values, the function $J^\alpha u(t)$ exhibits an algebraic singularity in $t = 0$. Increasing values of the exponent, however, improve its regularity close to the origin.

Remark 5.4. *It can be shown that J^α indeed improves the smoothness properties of its arguments. Roughly spoken, $J^\alpha u(t)$ can be written as sum of two expressions one of which is better behaved than u, while the other one might be nonsmooth in 0 but infinitely differentiable elsewhere [Die10, Theorem 2.5].*

The Riemann-Liouville integral satisfies continuity properties with respect to both the fractional exponent and the input function u. Exemplarily, we give the following statement.

Proposition 5.5. *Let $\alpha \in \mathbb{R}_0^+$ and $(u_n)_{n\in\mathbb{N}} \subset C([0,T];\mathcal{H})$ a sequence of functions that converges uniformly to some $u \in C([0,T];\mathcal{H})$. Then there holds for almost all $t \in (0,T)$*

$$\lim_{n\to\infty} J^\alpha u_n(t) = J^\alpha u(t) \quad \text{in } \mathcal{H}.$$

Proof. Following [Die10, Theorem 2.7], we apply the first point in Lemma 2.12 to observe

$$\begin{aligned}\|J^\alpha u_n(t) - J^\alpha u(t)\|_{\mathcal{H}} &\leq \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} \|u_n(\tau) - u(\tau)\|_{\mathcal{H}} \, d\tau \\ &\leq \frac{1}{\Gamma(\alpha)} \sup_{\tau\in[0,T]} \|u_n(\tau) - u(\tau)\|_{\mathcal{H}} \int_0^t (t-\tau)^{\alpha-1} \, d\tau \\ &\leq \frac{T^\alpha}{\Gamma(\alpha)} \sup_{\tau\in[0,T]} \|u_n(\tau) - u(\tau)\|_{\mathcal{H}},\end{aligned}$$

which converges to zero as $n \to \infty$. □

This result allow us to look at another instructive example that shall be instrumental for further discussions.

Example 5.6. *Consider the exponential function $u(t) = e^{t\lambda}$ for some $\lambda > 0$. Thanks to Proposition 5.5 and Example 5.3 we have*

$$J^\alpha u(t) = J^\alpha \left(\sum_{j=0}^\infty \frac{t^j \lambda^j}{j!} \right) = \sum_{j=0}^\infty \frac{\lambda^j}{j!} J^\alpha(t^j) = \sum_{j=0}^\infty \frac{\lambda^j t^{j+\alpha}}{\Gamma(j+\alpha+1)} = t^\alpha E_{1,\alpha+1}(t\lambda)$$

for all $\alpha \in \mathbb{R}_0^+$. These fractional integrals of u are illustrated in Figure 5.1 for $\alpha \in [0,2]$ and $\lambda = 1$. The example shows that fractional integrals do not reproduce the exponential function in the same manner as the integer ones do.

5.1.2 Fractional Differentiation

The definition of the fractional integral is straightforward. More delicate is the situation for fractional derivatives. One cannot simply allow for negative values of α in (5.4) without suffering from substantial regularity limitations on u. To overcome this difficulty, several mathematically distinct definitions of time-fractional differential operators have been proposed. We refer to the monographs [SKM93, Pod99, Die10] for detailed expositions. Among the most prominent ones, there are *Riemann-Liouville* and *Caputo fractional derivatives*, which are defined as the left- and the right inverse of the fractional integral, respectively. On its natural domain of definition, the latter is not invertible which is why these two notions of fractional differentiation do not coincide in general. The Caputo fractional derivative, however, turns out to be the more natural choice when it comes to actual applications. We therefore introduce it as *the* fractional derivative in the following definition [Cap67, Cap69]. Here and throughout, we write $\lceil \cdot \rceil$ to denote the ceiling function.

Definition 5.7. *Let $u : (0,T) \to \mathcal{H}$, $\alpha \in \mathbb{R}_0^+$, and $n = \lceil \alpha \rceil$. The fractional derivative of order α is defined by*

$$\partial_t^\alpha u(t) := J^{n-\alpha} \partial_t^n u(t), \qquad t \in (0,T),$$

whenever $\partial_t^n u \in L^1(0,T;\mathcal{H})$.

The expression $\partial_t^\alpha u(t)$ can be seen as classical (weak) derivative of u that is perturbed by a fractional integral of suitable order. A more explicit representation of ∂_t^α for all $\alpha \notin \mathbb{N}_0$ follows by direct substitution

$$\partial_t^\alpha u(t) = \frac{1}{\Gamma(n-\alpha)} \int_0^t (t-\tau)^{n-\alpha-1} \partial_t^n u(\tau)\, d\tau.$$

This integral representation makes it clear that the whole evolution history of u needs to be taken into account to evaluate $\partial_t^\alpha u(t)$ for one $t \in (0,T)$, rendering the fractional derivative to be nonlocal. Note that all classical derivatives are recovered from ∂_t^α whenever $\alpha \in \mathbb{N}$. Hence, ∂_t^α is a *local* operator if and only if $\alpha \in \mathbb{N}_0$. The reader is encouraged to compare the following properties for fractional derivatives with the corresponding ones from classical calculus.

Lemma 5.8. *For all $\alpha \in \mathbb{R}_0^+$ the fractional derivative ∂_t^α is a linear operator. Moreover, there holds for almost every $t \in (0,T)$*

1. *$\partial_t^\alpha \partial_t^m u(t) = \partial_t^{\alpha+m} u(t)$ for all $m \in \mathbb{N}$,*
2. *$\partial_t^\alpha J^m u(t) = \partial_t^{\alpha-m} u(t)$ for all $m \in \mathbb{N}$ with $m \leq \alpha$,*
3. *∂_t^α is a right-inverse of J^α, that is,*

$$J^\alpha \partial_t^\alpha u(t) = u(t). \tag{5.7}$$

Proof. The linearity of ∂_t^α follows from the linearity of J^α and the integer-order derivative. To confirm property 1, we write $n = \lceil \alpha \rceil$ to deduce for almost all $t \in (0,T)$

$$\partial_t^\alpha \partial_t^m u(t) = J^{n-\alpha} \partial_t^{n+m} u(t) = J^{n+m-(\alpha+m)} \partial_t^{n+m} u(t) = \partial_t^{\alpha+m} u(t),$$

where we use that $n+m = \lceil \alpha + m \rceil$. Similarly, there holds

$$\partial_t^\alpha J^m u(t) = J^{n-\alpha} \partial_t^{n-m} u(t) = J^{n-m-(\alpha-m)} \partial_t^{n-m} u(t) = \partial_t^{\alpha-m} u(t),$$

proving property 2. The last one is a direct consequence of the second claim in Lemma 5.2, since

$$J^\alpha \partial_t^\alpha u(t) = J^\alpha J^{n-\alpha} \partial_t^n u(t) = J^n \partial_t^n u(t) = u(t). \qquad \square$$

Note that not every relation can be carried over from the classical setting in a direct fashion. In general, the properties 1 to 3 in Lemma 5.8 do not hold after interchanging the order of the operators even if u is smooth. Moreover, $\partial_t^\alpha \partial_t^\beta u(t) \neq \partial_t^{\alpha+\beta} u(t)$ for arbitrary values of $\alpha, \beta \in \mathbb{R}^+$.

The domain of ∂_t^α can be made more explicit by means of so-called absolutely continuous functions [Die10]. To simplify matters, we study the required regularity assumptions on u by means of the following instructive example.

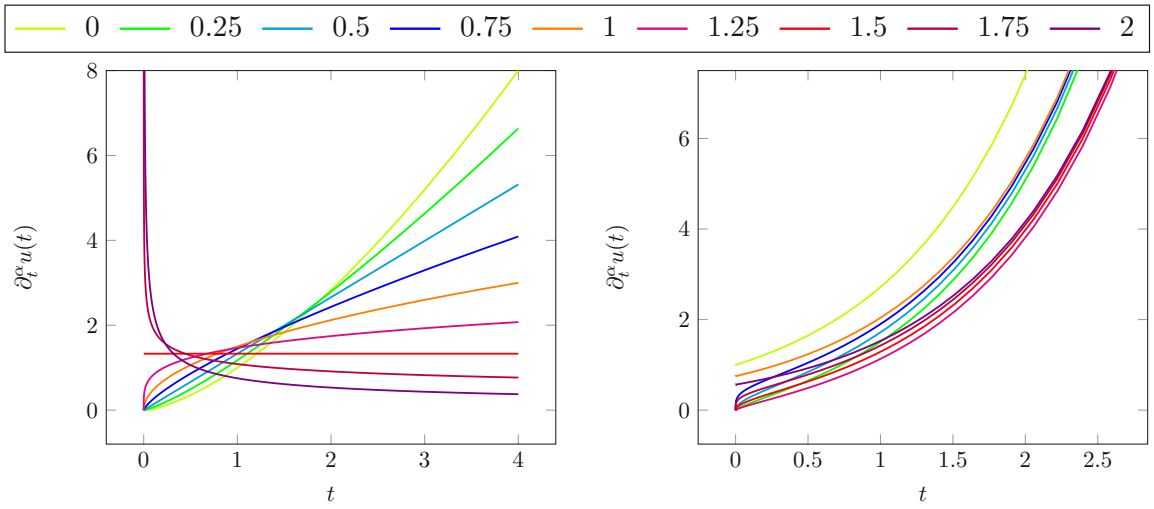

Figure 5.2: Fractional derivative $\partial_t^\alpha u(t)$ of $u(t) = t^{\frac{3}{2}}$ (left) and $u(t) = e^{\frac{3t}{4}}$ (right) for different orders $\alpha \in [0, 2]$.

Example 5.9. *Let $n = \lceil \alpha \rceil$, and $u(t) = t^s$ for some $s \in \mathbb{R}$. We compute*

$$\partial_t^n u(t) = \begin{cases} \frac{\Gamma(s+1)}{\Gamma(s-n+1)} t^{s-n}, & \text{if } s > n-1 \text{ or } s \in [n-1, -\infty) \setminus \{n-1, n-2, \cdots\}, \\ 0, & \text{if } s \in \{n-1, n-2, \dots, 0\}, \\ (-1)^n \frac{\Gamma(-s+n)}{\Gamma(-s)} t^{s-n}, & \text{if } s \in -\mathbb{N}. \end{cases} \tag{5.8}$$

To ensure $\partial_t^n u \in L^1(0,T)$, we thus require $s \in \{n-1, \dots, 0\}$ or $s > n-1 = \lfloor \alpha \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Under these assumptions on s, we infer from (5.8) and Example 5.3 that

$$\partial_t^\alpha u(t) = J^{n-\alpha} \partial_t^n u(t) = \begin{cases} \frac{\Gamma(s+1)}{\Gamma(s-\alpha+1)} t^{s-\alpha}, & \text{if } s > n-1, \\ 0, & \text{if } s \in \{0, \dots, n-1\}, \end{cases}$$

which can be seen as straightforward generalizations of the integer-order case. We choose $s = \frac{3}{2}$ and $\alpha \in [0, 2]$ to plot the fractional derivatives of $u(t)$ in Figure 5.2. For $\alpha = \frac{3}{2}$, the fractional derivative of u is identically one. For larger values of α, the derivatives possess a pole at $t = 0$ while smaller values of the exponent improve the smoothness of $\partial_t^\alpha u(t)$.

Another intriguing example is the following.

Example 5.10. *We compute the fractional derivative of order $\alpha \in \mathbb{R}_0^+$ of $u(t) = e^{t\lambda}$ for some $\lambda \in \mathbb{R}^+$ fixed. Thanks to the absolute convergence of the exponential series, Proposition 5.5, and Example 5.3, there holds with $n = \lceil \alpha \rceil$*

$$J^{n-\alpha} u(t) = \sum_{j=0}^{\infty} \frac{J^{n-\alpha}((\lambda t)^j)}{j!} = \sum_{j=0}^{\infty} \frac{\Gamma(j+1)\lambda^j t^{j+n-\alpha}}{j!\Gamma(j+n-\alpha+1)} = \sum_{j=0}^{\infty} \frac{\lambda^j t^{j+n-\alpha}}{\Gamma(j+n-\alpha+1)}. \tag{5.9}$$

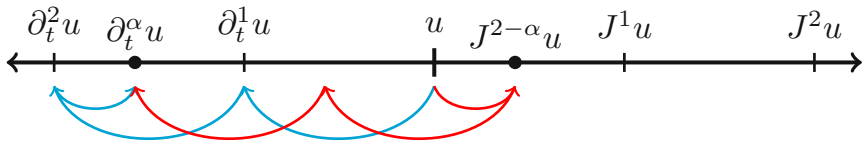

Figure 5.3: Illustration of ${}_R\partial_t^\alpha$ (red) and ∂_t^α (cyan) for $\alpha \in (1,2)$.

Since

$$\partial_t^n u(t) = \lambda^n e^{t\lambda},$$

we deduce from (5.9)

$$\partial_t^\alpha u(t) = \lambda^n J^{n-\alpha}(e^{t\lambda}) = \lambda^n \sum_{j=0}^{\infty} \frac{\lambda^j t^{j+n-\alpha}}{\Gamma(j+n-\alpha+1)} = \lambda^n t^{n-\alpha} E_{1,n-\alpha+1}(t\lambda),$$

which is depicted for $\lambda = \frac{3}{4}$ *and different orders* $\alpha \in [0,2]$ *in Figure 5.2.*

In its present form, the fractional derivative of order α is defined for all $\mathcal{H}$-valued functions that satisfy $\partial_t^n u \in L^1(0,T;\mathcal{H})$, where $n = \lceil \alpha \rceil$. If $\alpha \in (0,1)$, Example 5.9 shows that for $u(t) = t^s$ we require $s \geq 0$ to ensure that ∂_t^α is well-defined. This implies $\partial_t u \in L^2(0,T;\mathcal{H})$ so that $u \in H^1((0,T);\mathcal{H})$. This is rather prohibitive in view of the fact that the regularity of solutions to fractional PDEs is limited. To mitigate this problem, we introduce an alternative definition of fractional time-derivatives in the form of the Riemann-Liouville fractional derivative.

Definition 5.11. *Let* $u : (0,T) \to \mathcal{H}$, $\alpha \in \mathbb{R}_0^+$, *and* $n = \lceil \alpha \rceil$. *We define the Riemann-Liouville fractional derivative by*

$${}_R\partial_t^\alpha u(t) := \partial_t^n J^{n-\alpha} u(t), \qquad t \in (0,T).$$

Since $J^0 u = u$, Riemann-Liouville fractional derivatives coincide with all classical derivatives of u whenever α is an integer and

$${}_R\partial_t^\alpha u(t) = \frac{1}{\Gamma(n-\alpha)} \partial_t^n \int_0^t (t-\tau)^{n-\alpha-1} u(\tau)\, d\tau$$

if $\alpha \notin \mathbb{N}_0$. The differences between our definition of the fractional derivative and the one in the sense of Riemann-Liouville is made visible in Figure 5.3. Unlike ∂_t^α, ${}_R\partial_t^\alpha$ first applies the fractional integral of suitable order followed by an appropriate number of integer-order derivatives. This definition demands less regularity and can be applied to functions u where $\partial_t^\alpha u$ is not even defined. To emphasize this matter, we examine the Riemann-Liouville fractional derivative of the power function $u(t) = t^s$ to compare its regularity requirements, encoded in the exponent s, to the ones imposed by Example 5.9.

Example 5.12. *We are interested in the computation of the Riemann-Liouville fractional derivative of the function $u(t) = t^s$. Since $u \in L^1(0,T)$ is a necessary condition for $J^{n-\alpha}u(t)$, $n = \lceil \alpha \rceil$, and thus also for ${}_R\partial_t^\alpha u(t) = \partial_t^n J^{n-\alpha}u(t)$, to be defined, we require $s > -1$. Thanks to Example 5.3, we have*

$$ {}_R\partial_t^\alpha u(t) = \partial_t^n J^{n-\alpha}u(t) = \frac{\Gamma(s+1)}{\Gamma(s+n-\alpha+1)}\partial_t^n(t^{n-\alpha+s}). \tag{5.10} $$

If $\alpha - s \in \mathbb{N}$, then the right-hand side of (5.10) is the classical n^{th} derivative of a polynomial of degree $n-(\alpha-s) \in \{0,1,\dots,n-1\}$. We find

$$ {}_R\partial_t^\alpha u(t) = 0 $$

whenever $\alpha - s \in \mathbb{N}$. On the other hand, if $\alpha - s \notin \mathbb{N}$, there holds

$$ {}_R\partial_t^\alpha u(t) = \frac{\Gamma(s+1)}{\Gamma(s-\alpha+1)}t^{s-\alpha}. $$

Noting that $s > n-1$ implies $\alpha - s \notin \mathbb{N}$, we conclude, in view of Example 5.9, that $\partial_t^\alpha u(t) = {}_R\partial_t^\alpha u(t)$ whenever $u(t) = t^s$ and $s > n-1$. The Riemann-Liouville fractional derivative, however, is also meaningful if $s \in (-1, n-1]$. In particular, if $\alpha \in (0,1)$, then

$$ {}_R\partial_t^\alpha u(t) = \begin{cases} 0, & \text{if } s = \alpha - 1, \\ \frac{\Gamma(s+1)}{\Gamma(s-\alpha+1)}t^{s-\alpha}, & \text{if } s \in (-1,\infty) \setminus \{\alpha-1\}, \end{cases} \tag{5.11} $$

whereas $\partial_t u(t) = st^{s-1} \notin L^1(0,T)$ for all $s \in (-1,0)$ so that $\partial_t^\alpha u(t) = J^{n-\alpha}\partial_t u(t)$ is not well-defined in this case.

The power function is a showcase for the different regularity requirements imposed by ∂_t^α and ${}_R\partial_t^\alpha$. Indeed one can show that the domain of ${}_R\partial_t^\alpha$ is a super set of the one of ∂_t^α [Die10, Pod99]. Not only the domains, however, but also the operators themselves are inherently different as the following example shows.

Example 5.13. *We compute the Riemann-Liouville fractional derivative of $u(t) = e^{t\lambda}$ for some $\lambda \in \mathbb{R}^+$. Thanks to Proposition 5.5 and Example 5.3, we find that*

$$ J^{n-\alpha}u(t) = \sum_{j=0}^{\infty}\frac{J^{n-\alpha}((\lambda t)^j)}{j!} = \sum_{j=0}^{\infty}\frac{\Gamma(j+1)\lambda^j t^{j+n-\alpha}}{j!\Gamma(j+n-\alpha+1)} = \sum_{j=0}^{\infty}\frac{\lambda^j t^{j+n-\alpha}}{\Gamma(j+n-\alpha+1)}. $$

Hence,

$$ {}_R\partial_t^\alpha u(t) = \partial_t^n\left(\sum_{j=0}^{\infty}\frac{\lambda^j t^{j+n-\alpha}}{\Gamma(j+n-\alpha+1)}\right) = \sum_{j=0}^{\infty}\frac{\lambda^j t^{j-\alpha}}{\Gamma(j-\alpha+1)} = t^{-\alpha}E_{1,1-\alpha}(\lambda t). $$

Comparing these results with those obtained in Example 5.10, we see that ∂_t^α and ${}_R\partial_t^\alpha$ are indeed inherently different operators.

The following lemma follows from straightforward adaptions of Lemma 5.8 to the Riemann-Liouville setting.

Lemma 5.14. *For all $\alpha \in \mathbb{R}_0^+$ the Riemann-Liouville fractional derivative ${}_R\partial_t^\alpha$ is a linear operator. Moreover, there holds for almost every $t \in (0,T)$*

1. $\partial_t^m {}_R\partial_t^\alpha u(t) = {}_R\partial_t^{\alpha+m} u(t)$ *for all* $m \in \mathbb{N}$,
2. ${}_R\partial_t^\alpha u(t) = \partial_t^m J^{m-\alpha} u(t)$ *for all* $m \in \mathbb{N}$ *with* $m \geq \alpha$,
3. ${}_R\partial_t^\alpha$ *is a left-inverse of of* J^α, *that is,*

$$ {}_R\partial_t^\alpha J^\alpha u(t) = u(t). $$

Proof. The linearity of ∂_t^α follows from the linearity of J^α and the integer-order derivative. Property 1 holds since

$$ \partial_t^m {}_R\partial_t^\alpha u(t) = \partial_t^{m+n} J^{n-\alpha} u(t) = \partial_t^{m+n} J^{m+n-(m+\alpha)} u(t) $$

and $m+n = \lceil m+\alpha \rceil$. Similarly, there holds

$$ \partial_t^m J^{m-\alpha} u(t) = \partial_t^n \partial_t^{m-n} J^{m-n} J^{n-\alpha} u(t) = \partial_t^n J^{n-\alpha} u(t) = {}_R\partial_t^\alpha u(t), $$

proving property 2. To confirm the last conjecture, we apply the second point in Lemma 5.2 to see that

$$ {}_R\partial_t^\alpha J^\alpha u(t) = \partial_t^n J^{n-\alpha} J^\alpha u(t) = \partial_t^n J^n u(t) = u(t). $$ □

Despite being a driving force in the development of fractional calculus, the Riemann-Liouville derivative has yet taken only a minor role in applied modern science. One of the reasons for this issue is laid out in the following proposition; see e.g., [CM11] for a proof.

Proposition 5.15. *Let $\alpha \in \mathbb{R}^+ \setminus \mathbb{N}$, $n = \lceil \alpha \rceil$, and $u \in H^n((0,T);\mathcal{H})$. Then there holds*

$$ {}_R\partial_t^\alpha u(t) = \sum_{j=1}^{n-1} \frac{\partial_t^j u(0)}{\Gamma(j-\alpha+1)} t^{j-\alpha} + J^{n-\alpha} \partial_t^n u(t). \tag{5.12} $$

Equation (5.12) shows that Riemann-Liouville fractional derivatives have a singularity at the origin $t = 0$. Therefore, differential equations involving these derivatives do not allow for initial conditions of the form (5.1c) but require modifications that only have a limited physical meaning. Moreover, (5.11) shows that

$$ {}_R\partial_t^\alpha (1) = \frac{1}{\Gamma(1-\alpha)} t^{-\alpha}, \qquad \alpha \in (0,1). $$

Hence, the Riemann-Liouville fractional derivative of constant functions is not identically zero if $\alpha \notin \mathbb{N}$, which contradicts our natural understanding of derivatives. As the following theorem shows, ∂_t^α can be seen as regularization of ${}_R\partial_t^\alpha$ which "subtracts" the terms causing the singularity. For simplicity, we restrict ourselves to the case $\alpha \in [0,1]$ and refer to [Die10, Theorem 3.1] and [Baz01, Section 1.2] for a more general treatment of this subject.

Theorem 5.16. *For all $\alpha \in [0,1]$ there holds*

$$\partial_t^\alpha u(t) = {}_R\partial_t^\alpha (u(t) - u(0)). \tag{5.13}$$

In accordance with Example 5.9 and 5.12, this result shows that ∂_t^α and ${}_R\partial_t^\alpha$, $\alpha \in [0,1]$, coincide for all functions u that vanish at the origin, provided that both derivatives are defined. While the left-hand side of (5.13) exists if $\partial_t^\alpha u(t)$ exists, the right-hand side is meaningful if ${}_R\partial_t^\alpha u(t)$, which imposes less regularity assumptions on u, and $u(0)$ are well-defined. The latter condition is weaker than the previous one which allows us to extend the definition of ∂_t^α in the following manner, where we again limit ourselves to $\alpha \in [0,1]$.

Definition 5.17. *For all $\alpha \in [0,1]$ we define the fractional derivative of order α by*

$$\partial_t^\alpha u(t) := {}_R\partial_t^\alpha (u(t) - u(0)), \qquad t \in (0,T). \tag{5.14}$$

Remark 5.18. *Definition 5.17 is not the most general version of the Caputo fractional derivative that can be found in the literature. If ∂_t^α is the derivative of order $\alpha \in (0,1)$, one can expect a natural interpretation of $\partial_t^\alpha u(t)$ for functions being contained in the interpolation space $H^\alpha((0,T);\mathcal{H}) := [\mathcal{H}_0, \mathcal{H}_1]_\alpha$, where $\mathcal{H}_0 = L^2(0,T;\mathcal{H})$ and $\mathcal{H}_1 = H^1((0,T);\mathcal{H})$. Indeed, one can show that (5.14) is meaningful for all $u \in H^\alpha((0,T);\mathcal{H})$ if $\alpha > \frac{1}{2}$. In case of $0 < \alpha \leq \frac{1}{2}$, one has to resort to density arguments. In its present form, however, our definition of ∂_t^α is sufficient for the scope of this thesis and we direct the interested reader to [GLY15, Kar18, LS21] for a detailed investigation of this matter.*

Remark 5.19. *It goes without saying that ∂_t^α and ${}_R\partial_t^\alpha$ do not cover all possible definitions of fractional in-time differentiation operators that can be found in the literature. We name e.g., the Grünwald-Letnikov and the Miller-Ross sequential derivative [Pod99, Die10], which generally lead to operators that are distinct from the ones inspected here.*

A classical approach for solving scalar fractional differential equations is based on the Laplace transform. Therefore, we require one final technical result before building the bridge to fractional parabolic equations.

Lemma 5.20. *For all $\alpha \in (0,1]$ there holds*

$$\mathscr{L}[\partial_t^\alpha u](z) = z^\alpha \mathscr{L}[u](z) - z^{\alpha-1} u(0).$$

Proof. Recalling (5.5), we write

$$\partial_t^\alpha u(t) = (\mathcal{K}_{1-\alpha} * \partial_t u)(t), \qquad \mathcal{K}_{1-\alpha}(t) = \frac{t^{-\alpha}}{\Gamma(1-\alpha)}\,\mathrm{I}\,.$$

We successively apply (2.7), Lemma 2.23, and (2.8) to deduce

$$\mathscr{L}[\partial_t^\alpha u](z) = \mathscr{L}[\mathcal{K}_{1-\alpha}](z)\mathscr{L}[\partial_t u](z) = z^{\alpha-1}(z\mathscr{L}[u](z) - u(0)) = z^\alpha \mathscr{L}[u](z) - z^{\alpha-1}u(0). \qquad \square$$

5.2 Weak Formulation, Existence, and Uniqueness

We now come to the core of this chapter and consider the fully space-time fractional diffusion equation

$$\partial_t^\alpha u + \mathcal{L}^s u = f, \quad \text{in } \Omega \times (0,T), \tag{5.15a}$$
$$u = 0, \quad \text{on } \partial\Omega \times (0,T), \tag{5.15b}$$
$$u = u_0, \quad \text{on } \Omega \times \{0\}, \tag{5.15c}$$

where $\alpha \in (0,1]$, $T \in \mathbb{R}^+$, $s \in [0,1]$, $f \in L^\infty(0,T;L^2(\Omega))$, and $u_0 \in L^2(\Omega)$. To derive a weak formulation for this problem, we consider $u : (0,T) \to H_0^s(\Omega)$ as a function with Hilbert-valued range. Assuming $\partial_t^\alpha u(t) \in H^{-s}(\Omega)$, we may apply a test function $v \in H_0^s(\Omega)$ on both sides of (5.15a) to deduce, after identifying $L^2(\Omega)$ with its dual space, for almost all $t \in (0,T)$

$$\langle \partial_t^\alpha u(t), v\rangle + \langle \mathcal{L}^s u(t), v\rangle = (f(t), v)_{L^2(\Omega)}.$$

To give meaning to the initial condition (5.15c), we require $u \in C([0,T];L^2(\Omega))$. Invoking (4.7), we arrive at the weak formulation: Find $u \in L^2(0,T;H^s(\Omega)) \cap C([0,T];L^2(\Omega))$ with $\partial_t^\alpha u \in L^2(0,T;H^{-s}(\Omega))$ such that for almost all $t \in (0,T)$

$$\forall v \in H_0^s(\Omega): \quad \langle \partial_t^\alpha u(t), v\rangle + (u(t), v)_{H^s_{\mathcal{L}}} = (f(t), v)_{L^2(\Omega)}, \tag{5.16a}$$
$$u(0) = u_0. \tag{5.16b}$$

To find a solution to this problem, we make the ansatz

$$u(t) = \sum_{i=1}^{\infty} u_i(t)\varphi_i$$

and test (5.16a) with $v = \varphi_j$ for some $j \in \mathbb{N}$. This shows that for all $j \in \mathbb{N}$ there must hold

$$\partial_t^\alpha u_j(t) + \lambda_j^s u_j(t) = f_j(t), \tag{5.17a}$$
$$u_j(0) = u_{0,j}, \tag{5.17b}$$

where $f_j(t) = (\varphi_j, f(t))_{L^2(\Omega)}$ and $u_{0,j} = (\varphi_j, u_0)_{L^2(\Omega)}$. Following [Pod99, p. 140], we apply the Laplace transform on both sides of equation (5.17a) to arrive, after consulting Lemma 5.20, at the algebraic equation

$$z^\alpha \mathscr{L}[u_j](z) - z^{\alpha-1}u_{0,j} + \lambda_j^s \mathscr{L}[u_j](z) = \mathscr{L}[f_j](z).$$

Rearranging the terms reveals

$$\mathscr{L}[u_j](z) = \frac{z^{\alpha-1}u_{0,j}}{z^\alpha + \lambda_j^s} + \frac{\mathscr{L}[f_j](z)}{z^\alpha + \lambda_j^s}. \tag{5.18}$$

Invoking Lemma 2.31 with $k = 0$ and $\beta = 1$, we infer that

$$\mathscr{L}^{-1}\left[\frac{z^{\alpha-1}u_{0,j}}{z^\alpha + \lambda_j^s}\right](t) = E_{\alpha,1}(-t^\alpha \lambda_j^s)u_{0,j}.$$

To compute the inverse Laplace transform of the second term in (5.18), we apply the second property in Lemma 2.20 followed by Lemma 2.31 with $k = 0$ and $\beta = \alpha$ to find

$$\begin{aligned}\mathscr{L}^{-1}\left[\frac{\mathscr{L}[f_j](z)}{z^\alpha + \lambda_j^s}\right](t) &= f_j(t) * \mathscr{L}^{-1}\left[\frac{1}{z^\alpha + \lambda_j^s}\right](t) \\ &= f_j(t) * t^{\alpha-1}E_{\alpha,\alpha}(-t^\alpha\lambda_j^s) = \int_0^t (t-\tau)^{\alpha-1}E_{\alpha,\alpha}(-(t-\tau)^\alpha\lambda_j^s)f_j(\tau)\,d\tau.\end{aligned}$$

By the linearity of the Laplace transform, we conclude that the unique solution to (5.17) can be expressed as

$$u_j(t) = E_\alpha(-t^\alpha\lambda_j^s)u_{0,j} + \int_0^t (t-\tau)^{\alpha-1}E_{\alpha,\alpha}(-(t-\tau)^\alpha\lambda_j^s)f_j(\tau)\,d\tau.$$

Summation of these scalar solutions over every eigenmode yields the following result; see also [NOS16, BLP17b].

Theorem 5.21. *Let $\alpha \in (0,1]$, $s \in [0,1]$, $f \in L^\infty(0,T;L^2(\Omega))$, and $u_0 \in L^2(\Omega)$. Then (5.16) possesses a unique solution $u \in L^2(0,T;H^s(\Omega)) \cap C([0,T];L^2(\Omega))$ with $\partial_t^\alpha u \in L^2(0,T;H^{-s}(\Omega))$ that satisfies*

$$\begin{aligned}u(t) &= E_\alpha(-t^\alpha\mathcal{L}^s)u_0 + \int_0^t (t-\tau)^{\alpha-1}E_{\alpha,\alpha}(-(t-\tau)^\alpha\mathcal{L}^s)f(\tau)\,d\tau \\ &:= \sum_{j=1}^\infty \left(E_{\alpha,1}(-t^\alpha\lambda_j^s)u_{0,j} + \int_0^t (t-\tau)^{\alpha-1}E_{\alpha,\alpha}(-(t-\tau)^\alpha\lambda_j^s)f_j(\tau)\,d\tau\right)\varphi_j,\end{aligned} \tag{5.19}$$

where $u_{0,j} = (\varphi_j, u_0)_{L^2(\Omega)}$ and $f_j(t) = (\varphi_j, f(t))_{L^2(\Omega)}$.

For regularity estimates of solutions to (5.16) when $\mathcal{L}^s = (-\Delta)^s$ we direct the reader to [NOS16], see also [SY11]. Note that for the integer-order case $\alpha = s = 1$, the well-known variation-of-constants formula for local parabolic problems is recovered

$$u(t) = e^{-t\mathcal{L}}u_0 + \int_0^t e^{-(t-\tau)\mathcal{L}}f(\tau)\,d\tau.$$

For arbitrary right-hand side functions $f \in L^\infty(0,T;L^2(\Omega))$, the integral in (5.19) cannot be computed explicitly and requires numerical approximation. In a few particular scenarios, however, it can be expressed in terms of elementary functions.

Corollary 5.22. *Let $\alpha \in (0,1]$, $s \in [0,1]$, $f(t) = \sum_{i=0}^n t^i v_i$ for some $n \in \mathbb{N}$ and $v_i \in L^2(\Omega)$, and $u_0 \in L^2(\Omega)$. Then (5.16) possesses a unique solution $u \in L^2(0,T;H^s(\Omega)) \cap C([0,T];L^2(\Omega))$ with $\partial_t^\alpha u \in L^2(0,T;H^{-s}(\Omega))$ that satisfies*

$$\begin{aligned}u(t) &= E_\alpha(-t^\alpha\mathcal{L}^s)u_0 + \sum_{i=0}^n \Gamma(i+1)t^{\alpha+i}E_{\alpha,\alpha+i+1}(-t^\alpha\mathcal{L}^s)v_i \\ &:= \sum_{j=1}^\infty \left(E_\alpha(-t^\alpha\lambda_j^s)u_{0,j} + \sum_{i=0}^n \Gamma(i+1)t^{\alpha+i}E_{\alpha,\alpha+i+1}(-t^\alpha\lambda_j^s)v_{i,j}\right)\varphi_j,\end{aligned}$$

where $u_{0,j} = (\varphi_j, u_0)_{L^2(\Omega)}$ and $v_{i,j} = (\varphi_j, v_i)_{L^2(\Omega)}$.

Proof. Due to

$$(\varphi_j, f(t))_{L^2(\Omega)} = \sum_{i=0}^{n} t^i v_{i,j},$$

the series representation (5.19) evaluate to

$$u_j(t) = E_\alpha(-t^\alpha \lambda_j^s) u_{0,j} + \sum_{i=0}^{n} v_{i,j} \int_0^t (t-\tau)^{\alpha-1} E_{\alpha,\alpha}(-(t-\tau)^\alpha \lambda_j^s) \tau^i \, d\tau.$$

The conjecture now follows from Lemma 2.32. □

We mention that also for $\alpha = 0$ the equation (5.16a) has a unique solution that can be expressed in terms of the Mittag-Leffler function in the sense of Remark 2.26. In this regime, the fractional derivative degenerates to $\partial_t^\alpha = \partial_t^0 = \mathrm{I}$ in which case $f(t) = f$ is assumed to be constant and the initial condition (5.16b) is neglected.

Theorem 5.23. *Let $\alpha = 0$, $s \in [0,1]$, and $f \in L^2(\Omega)$. Then there exists a unique solution to (5.16a) which is given by*

$$u = E_0(-\mathcal{L}^s) := \sum_{j=1}^{\infty} \frac{f_j}{1+\lambda_j^s}, \qquad f_j = (\varphi_j, f)_{L^2(\Omega)}.$$

Proof. If $\alpha = 0$, (5.16a) can be written as

$$\forall v \in H_0^s(\Omega): \quad (u,v)_{H^s_{\tilde{\mathcal{L}}}} = (f,v)_{L^2(\Omega)},$$

where $\tilde{\mathcal{L}} := \mathcal{L} + \mathrm{I}$. The claim now follows from Theorem 4.14 and the observation that the eigenvalues $(\tilde{\lambda}_j)_{j=1}^\infty$ of $\tilde{\mathcal{L}}$ satisfy $\tilde{\lambda}_j = \lambda_j + 1$ for all $j \in \mathbb{N}$, where λ_j is the j^{th} eigenvalue of $\mathcal{L}$. □

Remark 5.24. *The situation is very similar if $\alpha \in (1,2)$, which is commonly referred to as fractional wave equation. Provided an additional condition on the first derivative of u, its solution can also be expressed in terms of the generalized Mittag-Leffler $E_{\alpha,\beta}(-t^\alpha \lambda^s)$ with $\alpha \in (1,2)$; see [LS21, OS18].*

The impact of the fractional parameters on solutions to (5.16) for $\mathcal{L} = -\Delta$, $\Omega = (0,1)^2$, $f \equiv 0$, and

$$u_0(\mathbf{x}) = \begin{cases} 1, & \text{if } (x-0.5)(y-0.5) \geq 0, \\ 0, & \text{else,} \end{cases} \qquad \mathbf{x} = (x,y) \in \Omega,$$

is reported in Figure 5.4. The example illustrates solutions to the fractional in-space, fractional in-time, and a fully space-time fractional diffusion equation. The diffusion process is faster when the value of s increases or α decreases.

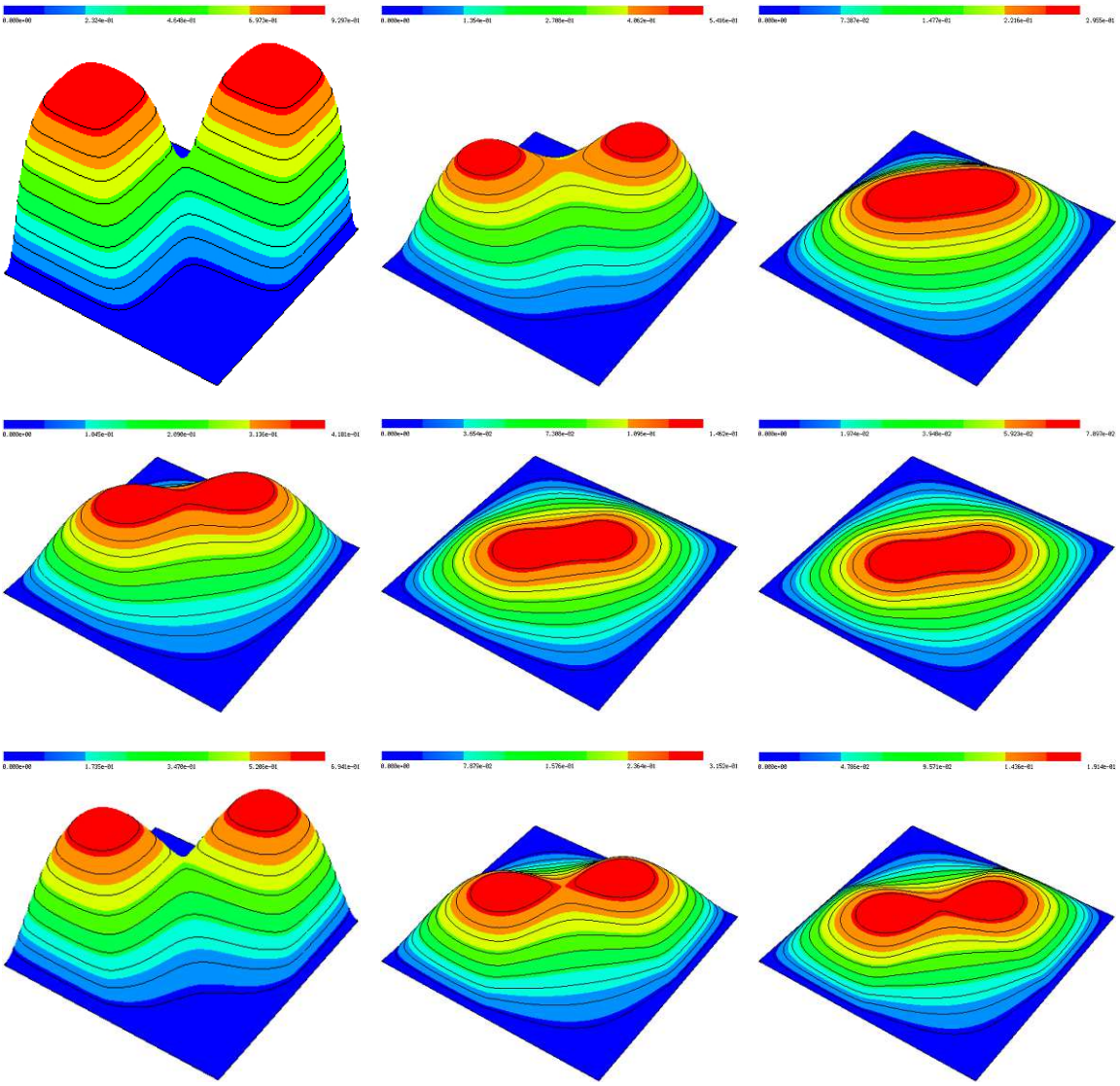

Figure 5.4: Evolution of the solutions u to the fractional heat equation (5.16) when $\mathcal{L} = -\Delta$ for $t = 0.01, 0.05, 0.1$ (from left to right) and $(\alpha, s) \in \{(1, 0.75), (0, 75, 1), (0, 75, 0.75)\}$ from top to bottom.

6 Numerical Approximation of Fractional Diffusion Problems

The literature provides an ample coverage on the numerical treatment of fractional elliptic PDEs [ILTA05, ILTA06, NOS15, BP15, Vab15, MN18, HLM+18, BLP19b, HMP21, DS21, DH21, HKL+21b, HKL+21a, Vab21a, Vab21b], space-fractional evolution equations [BLP17a, AM17, MR20b, Vab21c], time-fractional evolution equations [Lub88, JLZ15, KW21, FRW21], and fully space-time fractional PDEs [MN11, YTLI11, NOS16, BLP17b, Rie20, DHS21]. One way or another, any of the schemes listed above has to compensate for the nonlocality of the problem. In this chapter we present three different localization techniques to approximate fractional diffusion problems of elliptic and parabolic type.

1. The previous two chapters show that solutions to fractional PDEs can be expressed explicitly by means of the eigenfunctions of the integer-order differential operator $\mathcal{L}$. Hence, the solution can be made available by means of the local eigenvalue problems

$$(\varphi_j, v)_{H^1_{\mathcal{L}}(\Omega)} = \lambda_j (\varphi_j, v)_{L^2(\Omega)}, \qquad v \in H^1_0(\Omega),$$

   which can be computed using standard tools from finite element theory.

2. In Section 4.1.2, it is shown that the inverse fractional diffusion operator can be interpreted as Bochner integral

$$\mathcal{L}^{-s} f = \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-s} (\mathcal{L} + \zeta \,\mathrm{I})^{-1} f \, d\zeta, \qquad f \in L^2(\Omega).$$

   The integral can be discretized using a quadrature whose evaluation requires the solution to multiple standard reaction-diffusion problems.

3. According to Theorem 4.13, $\mathcal{L}^{-s} f$ can be recovered from the PDE

$$\begin{aligned} -\mathcal{L}\mathcal{U} + \frac{1-2s}{\zeta}\mathcal{U} + \partial_\zeta^2 \mathcal{U} &= 0, && \text{in } \mathcal{C}_\Omega, \\ \mathcal{U} &= 0, && \text{on } \partial\Omega \times \mathbb{R}^+, \\ \frac{\partial \mathcal{U}}{\partial \mathbf{n}_s} &= d_s f, && \text{on } \Omega \times \{0\}. \end{aligned}$$

   Therefore, the nonlocal problem can be localized at the cost of solving a degenerate PDE in a $d+1$-dimensional domain.

Clearly, the exact solutions to the respective local problems mentioned above are analytically not available and thus require numerical approximation. The discretization scheme of our choice is the *finite element method*, which we briefly introduce in the following section.

6.1 The Finite Element Method

We intend to gather a minimum of information to provide a superficial understanding of the material and refer the interested reader to the monographs [Cia02, Tho06] for an in-depth review on this topic. Throughout what follows, we assume $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, and denote with $\mathcal{T}_h$ a partition of Ω into intervals, triangles, and tetrahedrons in one, two, and three dimensions, respectively, whose components $T \in \mathcal{T}_h$ we call *elements*. The subscript h indicates that discretized objects are considered.

Definition 6.1. *A triangulation $\mathcal{T}_h$ of Ω is called regular if*

1. *the intersection of the interior of two distinct elements of $\mathcal{T}_h$ is empty,*
2. *the intersection of two elements of $\mathcal{T}_h$ is either empty or a common face, edge, or vertex of both,*
3. *the domain is covered by the elements $\overline{\Omega} = \cup_{T \in \mathcal{T}_h} \overline{T}$.*

Depending on the shape of the elements, a triangulation fulfills different regularity assumptions. Here and in what follows, $\operatorname{diam}(T)$ and $|T|$ denotes the diameter and the length/area/volume of T, respectively.

Definition 6.2. *A family of triangulations $(\mathcal{T}_h)_{h \in \mathbb{R}_0^+}$ is called*

- *shape-regular if there exists some constant $c \in \mathbb{R}^+$ such that*

$$\max_{T \in \mathcal{T}_h} \frac{\operatorname{diam}(T)^d}{|T|} \leq c,$$

- *quasi-uniform if $c \in \mathbb{R}^+$ exists such that*

$$\min_{T \in \mathcal{T}_h} \operatorname{diam}(T) \geq c \max_{T \in \mathcal{T}_h} \operatorname{diam}(T).$$

Shape regular triangulations have the convenient property that their elements do not degenerate in the sense that all angles are strictly bounded away from 0 and 180 degrees. In a quasi-uniform triangulation all elements have nearly the same size. Having ourselves familiarized with these concepts, we now introduce a finite element as follows [Cia02].

Definition 6.3. *The triplet (T, V_T, Ψ_T) is said to be a finite element if*

1. *$T \subset \mathbb{R}^d$ is a bounded and closed set with non-empty interior and piecewise smooth boundary,*
2. *V_T is a function space on T of finite dimension $n \in \mathbb{N}$,*
3. *$\Psi_T = \{\psi_{T,1}, \ldots, \psi_{T,n}\}$ is a basis of the dual space V_T' of V_T.*

In Definition 6.3, the space V_T is often referred to as space of *shape functions* and Ψ_T are the *degrees of freedom (dofs)*.

Definition 6.4. *Let $\mathcal{T}_h$ be a regular triangulation of Ω and (T, V_T, Ψ_T) a finite element for each $T \in \mathcal{T}_h$. We call the space of shape functions, where the dofs shared between two neighboring elements coincide,*

$$V_h := \{u \in \prod_{T\in\mathcal{T}_h} V_T : \forall \psi \in \Psi_{T_i} \cap \Psi_{T_j} : \psi(u|_{T_i}) = \psi(u|_{T_j})\},$$

a finite element space.

A large variety of finite element spaces exist to discretize $H_0^1(\Omega)$ efficiently. We provide here the description of a standard one that shall be used in all our numerical implementations. Provided a triangulation $\mathcal{T}_h$, we set $V_T = \mathcal{P}_1(T)$ for all $T \in \mathcal{T}_h$, where $\mathcal{P}_k(T)$ denotes the space of polynomials of degree k on T. Noting that $\dim V_T = d + 1$, we define the dofs $\Psi_T := \{\psi_{T,1}, \dots, \psi_{T,d+1}\}$ by

$$\psi_{T,1}(u) := u(V_i), \qquad i = 1, \dots, d+1,$$

for all $u \in V_T$, where $\{V_1, \dots, V_{d+1}\}$ is the set of vertices of T. Imposing $u(V_i) = 0$ if $V_i \in \partial\Omega$, we obtain the so-called *Lagrange finite element space of order one with vanishing trace*

$$\mathcal{P}_1^0(\mathcal{T}_h) := \{u \in C(\overline{\Omega}) : u|_T \in \mathcal{P}_1(T)\, \forall T \in \mathcal{T}_h, u|_{\partial\Omega} = 0 \text{ on } \partial\Omega\}.$$

Before we proceed with the first of three approximation schemes, we fix some further notation. For a fixed finite element space V_h with basis $(b_{h,j})_{j=1}^N$, $N \in \mathbb{N}$, we introduce the *mass matrix* $\mathbf{M} = (\mathbf{M})_{i,j=1}^N \in \mathbb{R}^{N\times N}$ of V_h as

$$\mathbf{M}_{ij} := (b_{h,j}, b_{h,i})_{L^2(\Omega)}, \qquad i, j = 1, \dots, N.$$

In dependency of the differential operator $\mathcal{L}$, we introduce, recalling (4.5), the *stiffness matrix* $\mathbf{A} = (\mathbf{A})_{i,j=1}^N \in \mathbb{R}^{N\times N}$ by

$$\mathbf{A}_{ij} := (b_{h,j}, b_{h,i})_{H^1_{\mathcal{L}}(\Omega)}, \qquad i, j = 1, \dots, N.$$

Both $\mathbf{M}$ and $\mathbf{A}$ are symmetric and positive definite by construction. For any $v_h \in V_h$ we denote with $\mathbf{v}_h = (v_{h,1}, \dots, v_{h,N})^T \in \mathbb{R}^{N\times 1}$ the coefficient vector of v_h such that

$$\mathbf{v}_h = \sum_{j=1}^N v_{h,j} b_{h,j}.$$

6.2 The Discrete Eigenfunction Method

In this section, we make use of the so-called *discrete eigenfunction method (DEM)* [BP15, BLP17b, LPG+20, Hof20], see also [ILTA05, ILTA06, YTLI11], to approximate solutions to fractional PDEs generated by the elliptic and self-adjoint differential operator $\mathcal{L}$. For this

purpose, we choose $V_h \subset H_0^1(\Omega)$ to be a finite element space and $(\varphi_{h,j})_{j=1}^N \subset V_h$ its basis consisting of discrete eigenfunctions with the property

$$\forall v_h \in V_h : \quad (\varphi_{h,j}, v_h)_{H^s_{\mathcal{L}}(\Omega)} = \lambda_{h,j}(\varphi_{h,j}, v_h)_{L^2(\Omega)}, \qquad (\varphi_{h,i}, \varphi_{h,j})_{L^2(\Omega)} = \delta_{ij}, \tag{6.2}$$

for all $i, j = 1, \dots, N$, where δ_{ij} denotes the Kronecker delta. The quantities $(\lambda_{h,j})_{j=1}^N \subset \mathbb{R}^+$ are the discrete eigenvalues.

The DEM presumes that the solution u is given analytically in terms of the continuous eigenvalues and eigenfunctions of the differential operator $\mathcal{L}$ as

$$u = \sum_{i=0}^{m} u^i, \qquad u^i = \sum_{j=1}^{\infty} f_i^{\boldsymbol{\tau}}(\lambda_j)(\varphi_j, b_i)_{L^2(\Omega)}\varphi_j, \tag{6.3}$$

where $m \in \mathbb{N}$, $f_i^{\boldsymbol{\tau}}$ is a problem-specific function that depends on a collection of parameters encoded in the vector $\boldsymbol{\tau} \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$, and b_i corresponds to the given data of the PDE, which is assumed to be in $L^2(\Omega)$ for simplicity. In light of the results presented in Section 4.2 and 5.2, the following configurations are of interest for us.

- In stationary fractional diffusion problems like the ones treated in Theorem 4.14, one has $m = 0$, the parametric function $f_0^{\boldsymbol{\tau}}(\lambda) = f^s(\lambda) = \lambda^{-s}$ is a power function with exponent $\boldsymbol{\tau} = s \in \Theta = (0, 1)$, and $b_0 = b \in L^2(\Omega)$ corresponds to the source term f in (4.25).
- In homogeneous fractional evolution equations, we have, in view of Theorem 5.21, a representation formula of the form (6.3) with $m = 0$, $f_0^{\boldsymbol{\tau}}(\lambda) = f^{\boldsymbol{\tau}}(\lambda) = E_\alpha(-t^\alpha \lambda^s)$, $\boldsymbol{\tau} = (\alpha, t, s) \in \Theta = (0, 1] \times \mathbb{R}_0^+ \times [0, 1]$, and $b_0 = b \in L^2(\Omega)$ being the initial condition u_0 of the PDE (5.16).
- If we include a source term of the form $f(t) = \sum_{i=0}^n t^i v_i$ in (5.16) for some $n \in \mathbb{N}$ and $v_i \in L^2(\Omega)$, Corollary 5.22 shows that we require $m = n$, $f_i^{\boldsymbol{\tau}}(\lambda) = t^{\alpha+i} E_{\alpha,\alpha+i+1}(-t^\alpha \lambda^s)$, $\boldsymbol{\tau} = (\alpha, t, s) \in \Theta = (0, 1] \times \mathbb{R}^+ \times [0, 1]$, and $b_i = v_i$ for $i = 0, \dots, n$.
- In the degenerate parabolic case $\alpha = 0$, mentioned in Theorem 5.23, we are interested in the configuration $m = 0$, $f_0^{\boldsymbol{\tau}}(\lambda) = f^{\boldsymbol{\tau}}(\lambda) = (1 + \lambda^s)^{-1}$, $\boldsymbol{\tau} = s \in \Theta = (0, 1)$, and $b \in L^2(\Omega)$ being the time-independent right-hand side function.

The idea of the DEM is to replace the infinite series in (6.3) by means of a finite sum over the discrete eigenvalues and eigenfunctions, that is,

$$u_h^{\text{DEM}} := \sum_{i=0}^{m} u_h^{\text{DEM},i}, \qquad u_h^{\text{DEM},i} := \sum_{j=1}^{N} f_i^{\boldsymbol{\tau}}(\lambda_{h,j})(\varphi_{h,j}, b_i)_{L^2(\Omega)}\varphi_{h,j}. \tag{6.4}$$

More succinctly, the DEM approximation can be written in terms of the matrix

$$\mathbf{L} := \mathbf{M}^{-1}\mathbf{A} \in \mathbb{R}^{N \times N},$$

where $\mathbf{M}$ and $\mathbf{A}$ denote the mass and stiffness matrix, respectively. Although elementary, the following observation gathers crucial ingredients for our further discussion.

Lemma 6.5. *The matrix* $\mathbf{L}$ *is positive definite, diagonalizable, and its eigenvectors are exactly the coefficient vectors of* $(\varphi_{h,j})_{j=1}^N$*, i.e.,*

$$\mathbf{L}\boldsymbol{\varphi}_{h,j} = \lambda_{h,j}\boldsymbol{\varphi}_{h,j}, \qquad j = 1, \dots, N. \tag{6.5}$$

Proof. The fact that $\mathbf{L}$ is positive definite and diagonalizable follows from the observation that $\mathbf{L} = \mathbf{M}^{-1}\mathbf{A}$ is similar to the matrix $\mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{M}^{-\frac{1}{2}}$, which is positive definite and diagonalizable itself. To proof the remainder of the statement, we deduce from the first identity in (6.2) that

$$\mathbf{A}\boldsymbol{\varphi}_{h,j} = \lambda_{h,j}\mathbf{M}\boldsymbol{\varphi}_{h,j},$$

whence (6.5) holds after multiplication with $\mathbf{M}^{-1}$ from the left. □

Remark 6.6. *Even though both* $\mathbf{M}$ *and* $\mathbf{A}$ *are sparse and symmetric, the matrix* $\mathbf{L}$ *itself is neither sparse nor symmetric.*

The matrix $\mathbf{L}$ can be seen as matrix approximation of the integer-order differential operator $\mathcal{L}$. If we collect its eigenvectors columnwise in the matrix $\mathbf{U} = [\boldsymbol{\varphi}_{h,1}, \dots, \boldsymbol{\varphi}_{h,N}] \in \mathbb{R}^{N\times N}$, it follows from the second identity in (6.2) that $\mathbf{U}^T\mathbf{M}\mathbf{U} = \mathbf{I}$. Hence

$$\mathbf{M} = \mathbf{U}^{-T}\mathbf{U}^{-1}, \tag{6.6}$$

where $\mathbf{U}^{-T} := (\mathbf{U}^T)^{-1}$. Two elementary consequences of (6.6) that shall be useful in the further course of this manuscript are stated below, where for any $b \in L^2(\Omega)$ we write $\mathbf{b} = (b_1, \dots, b_N)^T \in \mathbb{R}^{N\times 1}$ to label the coefficient vector of the L^2-orthogonal projection of b onto V_h, i.e.,

$$b_j = (\varphi_{h,j}, b)_{L^2(\Omega)}, \qquad j = 1, \dots, N.$$

Lemma 6.7. *Let* $\mathbf{U} = [\boldsymbol{\varphi}_{h,1}, \dots, \boldsymbol{\varphi}_{h,N}] \in \mathbb{R}^{N\times N}$ *denote the matrix whose columns contain the eigenvectors of* $\mathbf{L}$*.*

1. *If* $u_h \in V_h$*, then*
$$\|u_h\|_{L^2(\Omega)} = \|\mathbf{U}^{-1}\mathbf{u}_h\|_2.$$
2. *Let* $b \in L^2(\Omega)$ *and* $\mathbf{e}_j \in \mathbb{R}^N$ *be the* j^{th} *unit vector for any* $j \in \{1, \dots, N\}$*. Then*
$$(\varphi_{h,j}, b)_{L^2(\Omega)} = (\mathbf{e}_j, \mathbf{U}^{-1}\mathbf{b})_2.$$

Proof. This is a direct consequence of (6.6) since

$$\|u_h\|^2_{L^2(\Omega)} = (\mathbf{M}\mathbf{u}_h, \mathbf{u}_h)_2 = (\mathbf{U}^{-T}\mathbf{U}^{-1}\mathbf{u}_h, \mathbf{u}_h)_2 = (\mathbf{U}^{-1}\mathbf{u}_h, \mathbf{U}^{-1}\mathbf{u}_h)_2 = \|\mathbf{U}^{-1}\mathbf{u}_h\|_2^2$$

proves the first conjecture and

$$(\varphi_{h,j}, b)_{L^2(\Omega)} = (\mathbf{U}\mathbf{e}_j, \mathbf{M}\mathbf{b})_2 = (\mathbf{e}_j, \mathbf{U}^T\mathbf{M}\mathbf{b})_2 = (\mathbf{e}_j, \mathbf{U}^{-1}\mathbf{b})_2$$

the latter. □

We are now in position to write the coefficient vector of the DEM approximation in the following compact form.

Theorem 6.8. *The coefficient vector* $\mathbf{u}_h^{\mathrm{DEM}}$ *of the DEM approximation* (6.4) *satisfies*

$$\mathbf{u}_h^{\mathrm{DEM}} = \sum_{i=0}^{m} f_i^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}_i.$$

Proof. W.l.o.g. we assume $m = 0$ in (6.4) so that, after omitting the index i,

$$\mathbf{u}_h^{\mathrm{DEM}} = \sum_{j=1}^{N} f^{\boldsymbol{\tau}}(\lambda_{h,j})(\varphi_{h,j}, b)_{L^2(\Omega)}\varphi_{h,j} = \sum_{j=1}^{N} f^{\boldsymbol{\tau}}(\lambda_{h,j})(\mathbf{e}_j, \mathbf{U}^{-1}\mathbf{b})_2\varphi_{h,j}, \tag{6.7}$$

where the second identity follows from Lemma 6.7. Noting that the coefficient vector of $\varphi_{h,j}$ is exactly the j^{th} column of $\mathbf{U}$, the right-hand side of (6.7) can be written in matrix language

$$\mathbf{u}_h^{\mathrm{DEM}} = \mathbf{U}f^{\boldsymbol{\tau}}(\mathbf{D})\mathbf{U}^{-1}\mathbf{b},$$

where $\mathbf{D} = \mathrm{diag}(\lambda_{h,1}, \dots, \lambda_{h,N})$. This implies the claim. □

The discrete eigenfunction method is a popular scheme to approximate fractional diffusion problems [ILTA05, ILTA06, YTLI11, BP15, BP16, BLP17a, BLP17b, DS19, LPG+20, BGZ20, Hof20, DS21, DH21, DAC+21, DHS21]. For the lowest order Lagrangian finite element spaces, quasi-optimal convergence rates have been shown in [BP15, Theorem 4.3] for the stationary problem. These results have been generalized in [BP16] to a large class of differential operators that includes the ones we are interested in. For simplicity, we state the corresponding result here for the special case where Ω is convex and refer to [BP15, BP16, Lei18] for the general framework.

Theorem 6.9 (Convergence of the DEM - elliptic case)**.** *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded and convex domain,* $\mathcal{T}_h$ *a quasi-uniform triangulation on* Ω*,* $V_h = \mathcal{P}_1^0(\mathcal{T}_h)$ *the Lagrange finite element space of order one with vanishing trace,* $s \in (0,1)$*,* $f \in H_0^{2-2s}(\Omega)$ *in the sense of Remark 3.35, and* u_h^{DEM} *the DEM approximation of* $u = \mathcal{L}^{-s}f$*. Then there holds*

$$\|u - u_h^{\mathrm{DEM}}\|_{L^2(\Omega)} \preceq \ln(h^{-1})h^2\|f\|_{H_0^{2-2s}(\Omega)}.$$

We emphasize that the assumptions in Theorem 6.9 can be essentially relaxed at the cost of slower convergence rates. Roughly spoken, the rate of convergence depends on the smoothness of Ω such that e.g., for the L-shape domain, one can show a decay of the error like $\mathcal{O}(h^{\frac{4}{3}-\varepsilon})$ for any $\varepsilon > 0$.

In classical PDEs, the quality of approximation can significantly benefit from high-order schemes, where the function space $V_h = \mathcal{P}_1^0(\mathcal{T}_h)$ is enriched with polynomials of higher degree. Analogous results apply to nonlocal problems under a suitable refinement of the triangulation. The interested reader is referred to [BMS20] for a detailed discussion on these so-called *hp*-finite element methods.

The performance of the DEM in the time-dependent regime has been systematically studied in [BLP17a] for the lowest order case and $\alpha = 1$. These results have been generalized in [BLP17b] for arbitrary fractional exponents $\alpha \in (0, 1]$. Exemplarily, we state the corresponding result for the homogeneous space-time fractional diffusion problem when Ω is convex and refer to [BLP17b, Theorem 3.3], see also [Lei18, Theorem III.3], for the more general setting.

Theorem 6.10 (Convergence of the DEM - parabolic case)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded and convex domain, $\mathcal{T}_h$ a quasi-uniform triangulation on Ω, $V_h = \mathcal{P}_1^0(\mathcal{T}_h)$, $\alpha \in (0, 1]$, $\beta \in \mathbb{R}$, $s \in (0, 1)$, $u_0 \in H_0^2(\Omega)$, and u_h^{DEM} the DEM approximation of $u = E_{\alpha,\beta}(-t^\alpha \mathcal{L}^s)u_0$. Then there holds for all $t \in \mathbb{R}^+$*

$$\|u(t) - u_h^{\mathrm{DEM}}(t)\|_{L^2(\Omega)} \preceq \max\{1, \ln(t^{-\alpha})\} h^2 \|u_0\|_{H_0^2(\Omega)}.$$

6.3 Quadrature Approximations

Another approach for computing solutions to fractional PDEs are quadrature approximations [BP15, BP16, BLP17a, BLP17b, Lei18, BLP19b, BGZ20, Rie20, DAC+21, AN21, DZ21], see also [DH21]. Thanks to Theorem 4.10, solutions to stationary fractional diffusion problems might be written in integral form

$$u = \mathcal{L}^{-s} f = \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-s} (\mathcal{L} + \zeta \, \mathrm{I})^{-1} f \, d\zeta, \qquad f \in L^2(\Omega).$$

Upon replacing the integrand with our finite element approximations, we obtain an integral representation of the DEM approximation as the following lemma shows.

Lemma 6.11. *Let $\mathbf{u}_h^{\mathrm{DEM}} = \mathbf{L}^{-s}\mathbf{f}$ be the DEM approximation of $u = \mathcal{L}^{-s} f$. Then there holds*

$$\mathbf{u}_h^{\mathrm{DEM}} = \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-s} (\mathbf{L} + \zeta \mathbf{I})^{-1} \mathbf{f} \, d\zeta. \tag{6.8}$$

Proof. This is a direct consequence of Theorem 2.37 and the scalar version of Balakrishnan's formula

$$\lambda^{-s} = \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-s} (\lambda + \zeta)^{-1} \, d\zeta. \qquad \square$$

The idea is now to approximate the integral in (6.8) by a suitable quadrature

$$\mathbf{u}_{h,m}^{\mathrm{QUAD}} := \sum_{j=1}^m \omega_j \eta_j^{-s} (\mathbf{L} + \eta_j \mathbf{I})^{-1} \mathbf{f} \approx \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{-s} (\mathbf{L} + \zeta \mathbf{I})^{-1} \mathbf{f} \, d\zeta = \mathbf{L}^{-s} \mathbf{f}, \tag{6.9}$$

where $(\omega_j)_{j=1}^m \subset \mathbb{R}$ and $(\eta_j)_{j=1}^m \subset \mathbb{R}_0^+$ label the *quadrature weights and nodes*, respectively. Schemes of this form have been proposed in [BP15, BP16, Lei18, BLP19b, Rie20, DAC+21]. One that we shall mention here explicitly is the so-called *sinc quadrature* method advocated in [BP15, BP16, Lei18, BLP19b]; see [LB92, Ste12] for a nice survey on sinc methods.

Provided the integers $n_-, n_+ \in \mathbb{N}$ and the sinc parameter $\mathrm{q} \in \mathbb{R}^+$, the weights and nodes of this quadrature are given by

$$\omega_j = \frac{\mathrm{q}\sin(\pi s)}{\pi} e^{j\mathrm{q}}, \qquad \eta_j = e^{j\mathrm{q}},$$

for all $j = -n_-, \dots, n_+$, which is obviously an approximation of the form (6.9) after a suitable index shift. Hence, we define

$$\mathbf{u}_{h,\mathrm{q}}^{\mathrm{Sinc}} := \frac{\mathrm{q}\sin(\pi s)}{\pi} \sum_{j=-n_-}^{n_+} e^{(1-s)j\mathrm{q}} (\mathbf{L} + e^{j\mathrm{q}}\mathbf{I})^{-1}\mathbf{f}. \tag{6.10}$$

The approximation so obtained has the benefit that it requires multiple solutions to standard reaction-diffusion problems which can be computed efficiently in parallel. Its quality of approximation essentially hinges on the performance of the scalar quadrature and can be quantified in the following manner [BLP19b, Theorem 3.2].

Theorem 6.12. *Let $s \in (0,1)$ and $\mathrm{q} \in \mathbb{R}^+$. Then there holds*

$$\|\mathbf{u}_h^{\mathrm{Dem}} - \mathbf{u}_{h,\mathrm{q}}^{\mathrm{Sinc}}\|_{\mathbf{M}} \preceq \left(\frac{e^{-\frac{\pi^2}{2\mathrm{q}}}}{\sinh(\frac{\pi^2}{2\mathrm{q}})} + \frac{e^{-(1-s)\mathrm{q}n_-}}{1-s} + \frac{e^{-s\mathrm{q}n_+}}{s} \right) \|\mathbf{f}\|_{\mathbf{M}}. \tag{6.11}$$

The first term on the right-hand side of (6.11) is the contribution of the quadrature error over the *bounded* integration domain $[e^{-\mathrm{q}n_-}, e^{\mathrm{q}n_+}]$, while the latter can be seen as truncation errors. In practice, it is desirable to balance the three exponentials on the right-hand side of (6.11), which can be achieved by

$$\frac{\pi^2}{2\mathrm{q}} \approx s\mathrm{q}n_+ \approx (1-s)\mathrm{q}n_-.$$

As suggested in [BLP19b], we therefore impose

$$n_+ = \left\lceil \frac{\pi^2}{2s\mathrm{q}^2} \right\rceil, \qquad n_- = \left\lceil \frac{\pi^2}{2(1-s)\mathrm{q}^2} \right\rceil, \tag{6.12}$$

such that only the sinc parameter $\mathrm{q} \in \mathbb{R}^+$ needs to be determined a priori. The surrogate so obtained satisfies

$$\|\mathbf{u}_h^{\mathrm{Dem}} - \mathbf{u}_{h,\mathrm{q}}^{\mathrm{Sinc}}\|_{\mathbf{M}} \preceq \left(\frac{1}{s} + \frac{1}{1-s} \right) \left(\frac{e^{-\frac{\pi^2}{2\mathrm{q}}}}{\sinh(\frac{\pi^2}{2\mathrm{q}})} + e^{-\frac{\pi^2}{2\mathrm{q}}} \right) \|\mathbf{f}\|_{\mathbf{M}}. \tag{6.13}$$

Asymptotically, the upper bound behaves like

$$\left(\frac{1}{s} + \frac{1}{1-s} \right) e^{-\frac{\pi^2}{2\mathrm{q}}} \|\mathbf{f}\|_{\mathbf{M}}, \quad \text{as } \mathrm{q} \to 0^+.$$

Remark 6.13. *Combining Theorem 6.9 with (6.13) allows one to bound the total discrepancy between to the exact solution $u = \mathcal{L}^{-s} f$ and the sinc approximation (6.10).*

Our discussions from above show that quadrature approximations of fractional diffusion problems come in two stages. First, a finite element method is applied to replace the exact solution u with the DEM approximation u_h^{DEM}, which, in the second stage, is approximated by a quadrature scheme. This second layer of approximation is reflected as rational approximation of the power function $f^\tau(\lambda) = \lambda^{-s}$ in the following theorem; cf. [Hof20].

Theorem 6.14. *Let $s \in (0,1)$, $\mathrm{q} \in \mathbb{R}^+$, and*

$$r^{\mathrm{SINC}}(\lambda) := \frac{\mathrm{q}\sin(\pi s)}{\pi} \sum_{j=-n_-}^{n_+} e^{(1-s)j\mathrm{q}}(\lambda + e^{j\mathrm{q}})^{-1}.$$

Then there holds

$$\mathbf{u}_{h,\mathrm{q}}^{\mathrm{SINC}} = r^{\mathrm{SINC}}(\mathbf{L})\mathbf{f}.$$

Proof. This is a direct consequence of (6.10) and the first property in Lemma 2.34. □

In view of these results, one cannot expect $\mathbf{u}_{h,\mathrm{q}}^{\mathrm{SINC}}$ to be more accurate than $\mathbf{u}_h^{\mathrm{DEM}}$. The latter, however, requires diagonalization of the matrix $\mathbf{L}$, which has $\mathcal{O}(N^3)$ complexity, whereas the dominant computational effort in the evaluation of $\mathbf{u}_{h,\mathrm{q}}^{\mathrm{SINC}}$ amounts to $n := n_- + n_+ + 1$ linear solves. Since typically $n \ll N$, the sinc method is computationally considerably cheaper than the DEM approximation.

Also for time-dependent problems, quadrature schemes provide an attractive tool to approximate the accurate but expensive DEM solution [BLP17a, BLP17b, Lei18, Rie20]. Provided an integration contour $\mathcal{C}$ that encloses the spectrum of $\mathbf{L}$, Cauchy's integral theorem states that we may write the Mittag-Leffler function as

$$E_\alpha(-t^\alpha \lambda^s) = \frac{1}{2\pi i}\int_{\mathcal{C}} \frac{E_\alpha(-t^\alpha z)}{\lambda - z}\, dz.$$

Due to Theorem 2.36, it follows

$$E_\alpha(-t^\alpha \mathbf{L}^s) = \frac{1}{2\pi i}\int_{\mathcal{C}} E_\alpha(-t^\alpha z)(\mathbf{L} - z\mathbf{I})^{-1}\, dz. \tag{6.14}$$

The authors of [BLP17a, BLP17b, Lei18] make use of the hyperbolic contour

$$\mathcal{C} = \{z(\zeta) : \zeta \in \mathbb{R}\}, \qquad z(\zeta) := b(\cosh(\zeta) + i\sinh(\zeta)),$$

where $0 < b < \lambda_{h,1}/\sqrt{2}$ is a parameter. Along with this choice, the integral in (6.14) can be written as

$$E_\alpha(-t^\alpha \mathbf{L}^s) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} E_\alpha(-t^\alpha z(\zeta)^s) z'(\zeta)(\mathbf{L} - \zeta\mathbf{I})^{-1}\, d\zeta.$$

Given a suitable choice of $n_-, n_+ \in \mathbb{N}$ and the quadrature spacing $\mathrm{q} \in \mathbb{R}^+$, a sinc approximation to the homogeneous fractional evolution equation is given by

$$\mathbf{u}_{h,\mathrm{q}}^{\mathrm{SINC}}(t) := \frac{\mathrm{q}}{2\pi i}\sum_{j=-n_-}^{n_+} E_\alpha(-t^\alpha z(\zeta_j)^s) z'(\zeta_j)(\mathbf{L} - \zeta_j\mathbf{I})^{-1}\mathbf{u}_0, \qquad \zeta_j := j\mathrm{q}. \tag{6.15}$$

Using similar techniques as for the elliptic case, it can be shown that the approximation error decays exponentially in q, see [BLP17a] for $\alpha = 1$ and [BLP17b] for the fully space-time fractional regime. Unlike (6.10), however, the computation of (6.15) requires the solution to complex-valued problems even if both $\mathbf{L}$ and $\mathbf{u}_0$ are real.

6.4 Tensor FEM for the Extension Method

One final approach that we mention here is the harmonic extension method that interprets the fractional diffusion operator as Dirichlet-to-Neumann map [CS07, ST10, CT10, CDDS11, BCdPS13]. According to Theorem 4.13, $\mathcal{L}^s$ can be localized by means of the degenerate elliptic PDE

$$\operatorname{div}(\zeta^{1-2s}\hat{\mathfrak{A}}\nabla\mathcal{U}) + \zeta^{1-2s}\hat{\mathfrak{c}}\,\mathcal{U} = 0, \qquad \text{in } \mathcal{C}_\Omega, \tag{6.16a}$$

$$\mathcal{U} = 0, \qquad \text{on } \partial\Omega \times \mathbb{R}^+, \tag{6.16b}$$

$$\frac{\partial \mathcal{U}}{\partial \mathbf{n}_s} = d_s f, \qquad \text{on } \Omega \times \{0\}, \tag{6.16c}$$

where $\mathcal{C}_\Omega = \Omega \times \mathbb{R}^+$ and $\hat{\mathfrak{A}}$ and $\hat{\mathfrak{c}}$ are defined by (4.18). To derive a variational formulation for (6.16), we multiply (6.16a) with a test function $v \in \mathring{H}^1_s(\mathcal{C}_\Omega)$, integrate over the cylinder, and apply integration by parts to deduce

$$\forall v \in \mathring{H}^1_s(\mathcal{C}_\Omega): \quad \int_{\mathcal{C}_\Omega} \zeta^{1-2s}\left(\hat{\mathfrak{A}}\nabla\mathcal{U}\cdot\nabla v + \hat{\mathfrak{c}}\,\mathcal{U}v\right) d(\mathbf{x},\zeta) = d_s(f, v(\mathbf{x},0))_{L^2(\Omega)}. \tag{6.17}$$

Theorem 6.15. *Let $s \in (0,1)$ and $f \in L^2(\Omega)$. Then there exists a unique solution $\mathcal{U} \in \mathring{H}^1_s(\mathcal{C}_\Omega)$ to (6.17).*

Proof. This is a direct consequence of the Lax-Milgram theorem. □

In its present form, the solution to (6.17) is computationally out of reach since it is formulated on the unbounded domain $\mathcal{C}_\Omega = \Omega \times \mathbb{R}^+$. To make the variational formulation amenable to finite element discretization, we proceed as in [NOS15] and pose the problem on the truncated domain $\mathcal{C}^{\text{cut}}_\Omega := \Omega \times (0, \zeta_{\text{cut}})$ for some $\zeta_{\text{cut}} \in \mathbb{R}^+$: Find $\mathcal{U}_{\text{cut}} \in \mathring{H}^1_s(\mathcal{C}^{\text{cut}}_\Omega)$ such that

$$\forall v \in \mathring{H}^1_s(\mathcal{C}^{\text{cut}}_\Omega): \quad \int_{\mathcal{C}^{\text{cut}}_\Omega} \zeta^{1-2s}\left(\hat{\mathfrak{A}}\nabla\mathcal{U}_{\text{cut}}\cdot\nabla v + \hat{\mathfrak{c}}\,\mathcal{U}_{\text{cut}}v\right) d(\mathbf{x},\zeta) = d_s(f, v(\mathbf{x},0))_{L^2(\Omega)}, \tag{6.18}$$

where $\mathring{H}^1_s(\mathcal{C}^{\text{cut}}_\Omega) := \{v \in L^2_s(\mathcal{C}^{\text{cut}}_\Omega) : \nabla v \in L^2_s(\mathcal{C}^{\text{cut}}_\Omega), v = 0 \text{ on } \partial\Omega \times (0,\zeta_{\text{cut}}) \cup \Omega \times \{\zeta_{\text{cut}}\}\}$. The following proposition justifies this approach.

Proposition 6.16. *Let $\lambda_{\min}$ denote the smallest eigenvalue of $\mathcal{L}$, $\mathcal{U}$ the solution to (6.17), and $\tilde{\mathcal{U}}_{\text{cut}}$ the zero extension of the solutions to (6.18) with $\zeta_{\text{cut}} \geq 1$. Then there holds*

$$\|\nabla(\mathcal{U} - \tilde{\mathcal{U}}_{\text{cut}})\|_{L^2_s(\mathcal{C}_\Omega)} \preceq e^{-\frac{\lambda_{\min}\zeta_{\text{cut}}}{4}}\|f\|_{L^2(\Omega)}.$$

Proof. A key ingredient in the proof is the representation formula (3.28) and the exponential decay of ϕ_j in the extended direction. We refer to [NOS15, Lemma 3.3] for details. □

To approximate (6.18), a tensor finite element method is proposed in [NOS15], see also [BMN+18]. To make matters precise, let $\mathcal{T}_h$ denote a triangulation on Ω, $0 = \zeta_0 < \cdots < \zeta_m = \zeta_{\text{cut}}$ a partition of $[0, \zeta_{\text{cut}}]$, $I_j := [\zeta_{j-1}, \zeta_j]$ for all $j = 1, \dots, m$, and $I_{\text{cut}} := \{I_j : j = 1, \dots, m\}$. A mesh on $\mathcal{C}_\Omega^{\text{cut}}$ can then be defined in a tensor product fashion

$$\mathcal{T}_h^{\mathcal{C}_\Omega^{\text{cut}}} := \mathcal{T}_h \otimes I_{\text{cut}} := \{T \times I : T \in \mathcal{T}_h, I \in I_{\text{cut}}\}.$$

On this grid, a tensor finite element space is obtained by

$$\begin{aligned} V_h^{\mathcal{C}_\Omega^{\text{cut}}} := \{w \in C(\overline{\mathcal{C}_\Omega^{\text{cut}}}) : w|_{T\times J} \in \mathcal{P}_1(T) \otimes \mathcal{P}_1(J) \,\forall\, T \times J \in \mathcal{T}_h^{\mathcal{C}_\Omega^{\text{cut}}}, \\ w(\cdot, \zeta_{\text{cut}}) = 0, u(\mathbf{x}, \zeta) = 0 \text{ on } \partial\Omega \times (0, \zeta_{\text{cut}})\}, \end{aligned} \tag{6.19}$$

where $\mathcal{P}_1(T) \otimes \mathcal{P}_1(J) := \{p(\mathbf{x})q(\zeta) : p \in \mathcal{P}_1(T), q \in \mathcal{P}_1(J)\}$. Alternatively, the finite element space (6.19) can be written as

$$V_h^{\mathcal{C}_\Omega^{\text{cut}}} = \text{span}\{b_i^\Omega(\mathbf{x}) b_j^{\text{cut}}(\zeta) : i = 1, \dots, N, j = 0, \dots, m\}, \tag{6.20}$$

where $(b_j^\Omega)_{j=1}^N$ and $(b_j^{\text{cut}})_{j=0}^m$ denote a basis of the finite element space in $\mathbf{x}$- and ζ-direction, respectively. With this at hand, the discrete Galerkin formulation to (6.18) now reads: Find $\mathcal{U}_h \in V_h^{\mathcal{C}_\Omega^{\text{cut}}}$ such that

$$\forall v_h \in V_h^{\mathcal{C}_\Omega^{\text{cut}}} : \quad \int_{\mathcal{C}_\Omega^{\text{cut}}} \zeta^{1-2s} \left(\hat{\mathfrak{A}} \nabla \mathcal{U}_h \cdot \nabla v_h + \hat{\mathfrak{c}} \mathcal{U}_h v_h\right) d(\mathbf{x}, \zeta) = d_s(f, v_h(\mathbf{x}, 0))_{L^2(\Omega)}. \tag{6.21}$$

In accordance with the continuous problem, (6.21) has a unique solution by the Lax-Milgram theorem. In light of Theorem 4.13, the finite element approximation to the stationary fractional diffusion problem is then obtained by

$$u_{h,\zeta_{\text{cut}}}^{\text{Ext}}(\mathbf{x}) := \mathcal{U}_h(\mathbf{x}, 0). \tag{6.22}$$

A deeper analysis of the harmonic extension problem shows that the solution behaves reasonably well in $\mathbf{x}$-direction but has an algebraic singularity at $\zeta = 0$. This lack of regularity can be compensated by choosing the partition $(\zeta_j)_{j=0}^m$ of $(0, \zeta_{\text{cut}})$ in a geometrically refined manner towards 0, that is,

$$\zeta_j = \left(\frac{j}{m}\right)^{\text{g}} \zeta_{\text{cut}}, \quad j = 0, \dots, m, \tag{6.23}$$

for some grading parameter $\text{g} > 1$. Assuming that the finite element space $V_h^{\mathcal{C}_\Omega^{\text{cut}}}$ is constructed in this graded manner, one can prove the following result [NOS15, Theorem 5.4].

Theorem 6.17. *Let Ω be convex, $s \in (0,1)$, $\text{g} > \frac{3}{2s}$ in (6.23), $\zeta_{\text{cut}} > 1$, $f \in H_0^{1-s}(\Omega)$, $\mathcal{N}$ the number of elements contained in $\mathcal{T}_h^{\mathcal{C}_\Omega^{cut}}$, $\lambda_{\min}$ the smallest eigenvalue of $\mathcal{L}$, and $u = \mathcal{L}^{-s} f$. Then there holds*

$$\|u - u_{h,\zeta_{\text{cut}}}^{\text{Ext}}\|_{L^2(\Omega)} \preceq \left(e^{-\frac{\sqrt{\lambda_{\min}}\zeta_{\text{cut}}}{4}} + \zeta_{\text{cut}}^s \mathcal{N}^{-\frac{1}{d+1}}\right) \|f\|_{H_0^{1-s}(\Omega)}. \tag{6.24}$$

Choosing $\zeta_{\text{cut}} \approx \log(\mathcal{N})$ balances both terms in (6.24) and yields

$$\|u - u^{\text{Ext}}_{h,\zeta_{\text{cut}}}\|_{L^2(\Omega)} \preceq |\ln(\mathcal{N})|^s \mathcal{N}^{-\frac{1}{d+1}} \|f\|_{H^{1-s}_0(\Omega)},$$

which resembles, up to the logarithmic factor, the classical a priori estimate for the finite element method applied to $\mathcal{L}$ in $d+1$ dimensions. However, the original problem is posed on $\Omega \subset \mathbb{R}^d$ which is why the convergence rate is still sub-optimal due to the presence of the artificial dimension in ζ-direction. The implementation of hybrid *hp*-finite element methods allows one to overcome these limitations [BMN+18].

Remark 6.18. *The truncation of the cylinder can be avoided using a spectral method in the extended direction. The interested reader is directed to [AG18] for details.*

In Theorem 6.14 it is shown that the sinc quadrature approximation can be interpreted as matrix-vector product of the form $r^{\text{Sinc}}(\mathbf{L})\mathbf{b}$ where r^{Sinc} represents a rational approximation of the power function λ^{-s}. As noted in [Hof20, Theorem 2], the same also applies to the extension method if $\mathcal{L} = -\Delta$.

Theorem 6.19. *Let d_s be defined by Lemma 3.24, $u^{\text{Ext}}_{h,\zeta_{\text{cut}}}$ the extension approximation (6.22), $V^{\zeta_{\text{cut}}} = \operatorname{span}\{b_0^{\zeta_{\text{cut}}}, \dots, b_m^{\zeta_{\text{cut}}}\}$ with $b_j^{\zeta_{\text{cut}}}$ as in (6.20), and (ψ_j, μ_j) the eigenfunctions and eigenvalues of the one-dimensional eigenvalue problem*

$$\forall v \in V^{\zeta_{\text{cut}}} : \quad (\zeta^{1-2s}\psi_j', v')_{L^2((0,\zeta_{\text{cut}}))} = \mu_j (\zeta^{1-2s}\psi_j, v)_{L^2((0,\zeta_{\text{cut}}))}, \qquad j = 0, \dots, m.$$

Then there holds

$$\mathbf{u}^{\text{Ext}}_{h,\zeta_{\text{cut}}} = r^{\text{Ext}}(\mathbf{L})\mathbf{f},$$

where

$$r^{\text{Ext}}(\lambda) := d_s \sum_{j=0}^{m} \frac{\psi_j(0)^2}{\lambda + \mu_j}.$$

Although one can expect the DEM approximation to be closer to the exact solution, $\mathbf{u}^{\text{Ext}}_{h,\zeta_{\text{cut}}}$ is computationally more efficient since m is typically significantly smaller than N.

The extension technique developed in [NOS15] can be adapted to space-time fractional diffusion problems [NOS16]. The idea is to rewrite the fractional parabolic equation (5.15) as quasi-stationary elliptic problem with dynamic boundary conditions

$$\operatorname{div}(\zeta^{1-2s}\hat{\mathfrak{A}}\nabla\mathcal{U}) + \zeta^{1-2s}\hat{\mathfrak{c}}\mathcal{U} = 0, \qquad \text{in } \mathcal{C}_\Omega \times (0,T), \tag{6.25a}$$

$$\mathcal{U} = 0, \qquad \text{on } (\partial\Omega \times \mathbb{R}^+) \times (0,T), \tag{6.25b}$$

$$d_s \partial_t^\alpha \mathcal{U} + \frac{\partial \mathcal{U}}{\partial \mathbf{n}_s} = d_s f, \qquad \text{on } (\Omega \times \{0\}) \times (0,T), \tag{6.25c}$$

$$\mathcal{U} = u_0, \qquad \text{on } (\Omega \times \{0\}) \times \{0\}. \tag{6.25d}$$

Similarly to the elliptic case, one derives a variational formulation for (6.25) on the truncated domain $\mathcal{C}_\Omega^{\text{cut}} \times (0,T)$. An in-depth analysis on this matter can be found in [NOS16] for $\alpha \in (0,1]$ and [MR20b] for $\alpha = 1$ using *hp*-finite element methods.

7 The Rational Krylov Method

The previous chapter provides the description of three algorithms to approximate solutions of fractional PDEs. These sulutions $u = u(\boldsymbol{\tau})$ depend on multiple parameters, such as the spatial fractional order $s \in (0,1)$ or the order of the fractional time derivative $\alpha \in (0,1]$, which we collect in the parameter vector $\boldsymbol{\tau} \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$. In the further course of this thesis, we choose the discrete eigenfunction method as a starting point to approximate the entire solution manifold $\{u(\boldsymbol{\tau}) : \boldsymbol{\tau} \in \Theta\}$. The dominant computational costs of the DEM boil down to the evaluation of a parametric matrix-vector product $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$, where

- $\mathbf{L} = \mathbf{M}^{-1}\mathbf{A} \in \mathbb{R}^{N\times N}$ is a diagonalizable positive definite (but not necessarily symmetric) matrix-approximation of the integer-order differential operator $\mathcal{L}$,
- $\mathbf{b} \in \mathbb{R}^N$ some coefficient vector stemming from the user-provided data,
- $f^{\boldsymbol{\tau}}$ a matrix function that corresponds to the particular problem at hand and depends on the parameter vector $\boldsymbol{\tau} \in \Theta$.

The matrix $f^{\boldsymbol{\tau}}(\mathbf{L})$ is typically dense. Its evaluation requires the knowledge of all eigenvalues and eigenvectors of $\mathbf{L}$ which is a task of $\mathcal{O}(N^3)$ complexity. Due to limitations in computational resources, one is obliged to resort to *model order reduction strategies* if N is large. These schemes strive to reduce the computational costs by a significant margin while keeping the discretization error to a tolerable level. While the sinc and the extension approximation presented in the previous chapter already diminish the costs of conventional DEMs, one might still need $\mathcal{O}(100)$ classical PDE solves[1] to ensure the accuracy for a single solve of the fractional PDE. The purpose of this chapter is to alleviate the expenses of computing $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ using a standard model order reduction strategy in the form of the *rational Krylov method (RKM)* [Güt10, Ruh84, Saa81, HL97]. To this end, we take the point of view that the mesh parameter h in the DEM is small, causing N to be large, to ensure an accurate approximation of the continuous solution. We take the surrogate so obtained as our underlying *truth solution*. To emphasize this novel perspective, we therefore omit the h-dependency in our notation for the remainder of this thesis and assume $\mathbf{L} \in \mathbb{R}^{N\times N}$ to be a fixed positive definite matrix of dimension $N \gg 1$.

The key idea of each RKM is the extraction of a surrogate $\mathbf{u}_{k+1}$ from a search space $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$ of dimension $k+1 \ll N$ with the property $\mathbf{u}_{k+1} \approx f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$. Essential questions are:

1. How to choose the low-dimensional search space $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$?
2. How to extract $\mathbf{u}_{k+1}$ from $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$?

The first point is addressed in the following section.

[1] In case of the extension method, this is understood in the sense of Theorem 6.19.

7.1 The Rational Krylov Space

Throughout the remainder of this thesis, we designate the smallest and largest eigenvalues of $\mathbf{L}$ with $\lambda_{\min}$ and $\lambda_{\max}$, respectively, and introduce the spectral interval of $\mathbf{L}$ as $\Sigma := [\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}^+$. One of the central definitions of this chapter is stated below, where we write $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$.

Definition 7.1. *In dependence of the parameter set $\Xi = \{\xi_0, \dots, \xi_k\} \subset \overline{\mathbb{R}} \setminus \Sigma$, whose elements are called poles, we introduce the polynomial*

$$q_\Xi(\lambda) := \prod_{\substack{j=0 \\ \xi_j \neq \infty}}^{k} (\lambda - \xi_j). \tag{7.1}$$

For all $k \in \mathbb{N}_0$ we define the rational Krylov space of $\mathbf{L}$ and $\mathbf{b}$ with pole set Ξ by

$$\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b}) := \operatorname{span}\{q_\Xi(\mathbf{L})^{-1}\mathbf{b}, q_\Xi(\mathbf{L})^{-1}\mathbf{L}\mathbf{b}, \dots, q_\Xi(\mathbf{L})^{-1}\mathbf{L}^k\mathbf{b}\}.$$

Remark 7.2. *Although not explicit in our notation, the reader is encouraged to always associate the index $k+1$ in the expression $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ to the cardinality of the pole set Ξ. In case we compare the pole sets of two rational Krylov spaces $\mathcal{Q}_j^{\Xi}(\mathbf{L}, \mathbf{b})$ and $\mathcal{Q}_k^{\Xi}(\mathbf{L}, \mathbf{b})$ for some $j, k \in \mathbb{N}$, we write Ξ_j and Ξ_k, respectively, for more clarity in exposition.*

Note that

- $\Xi \subset \overline{\mathbb{R}} \setminus \Sigma$ implies that the inverse of $q_\Xi(\mathbf{L})$ exists such that $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ is well-defined,
- the rational Krylov space is independent of the particular ordering of the poles,
- if $\xi_j = \infty$ for one $j = 0, \dots, k$, then $\mathbf{b} \in \mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$.

In its original form, Definition 7.1 goes back to the Russian mathematician Nikolay Krylov [Kry31], who investigated subspace approximations based on the *polynomial Krylov space*

$$\mathcal{K}_{k+1}(\mathbf{L}, \mathbf{b}) := \operatorname{span}\{\mathbf{b}, \mathbf{L}\mathbf{b}, \dots, \mathbf{L}^k\mathbf{b}\}.$$

Clearly, the polynomial case is recovered from $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ if one sets $\Xi = \{\infty, \dots, \infty\}$ such that $q_\Xi \equiv 1$. Other commonly used configurations are

- $\Xi = \{\xi, \dots, \xi\}$ for some $\xi \in \mathbb{R} \setminus \Sigma$, in which case $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ is known as *shift-and-invert Krylov space* [MN04, EH06],
- $\Xi = \{\infty, 0, \infty, 0, \dots\}$, which yields the so-called *extended Krylov space* [DK98, KS10].

In order to provide more information about the particular structure of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, we introduce the set of polynomials $\mathcal{P}_N$ of maximal degree N. The minimal polynomial $p_{\mathbf{L},\mathbf{b}}^{\min} \in \mathcal{P}_N$ of $\mathbf{b}$ with respect to $\mathbf{L}$ is defined as the unique monic polynomial of lowest degree such that $p_{\mathbf{L},\mathbf{b}}^{\min}(\mathbf{L})\mathbf{b} = 0$, see [Wei60]. This allows us to introduce the invariance index of $\mathbf{L}$ and $\mathbf{b}$.

Definition 7.3. *The invariance index* $\mathcal{I} \in \mathbb{N}$ *of* $\mathbf{L}$ *and* $\mathbf{b}$ *is defined by*

$$\mathcal{I} := \deg(p^{\min}_{\mathbf{L},\mathbf{b}}),$$

where $\deg(p^{\min}_{\mathbf{L},\mathbf{b}})$ *is the degree of the minimal polynomial* $p^{\min}_{\mathbf{L},\mathbf{b}}$.

As shown in [Güt10], the invariance index $\mathcal{I}$ of $\mathbf{L}$ and $\mathbf{b}$ is related to the rational Krylov space of $\mathbf{L}$ and $\mathbf{b}$ in the following manner.

Proposition 7.4. *Let* $\mathcal{I} \in \mathbb{N}$ *denote the invariance index of* $\mathbf{L}$ *and* $\mathbf{b}$. *Then there holds*

$$\dim \mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b}) = \min\{k+1, \mathcal{I}\}.$$

Whenever we have a sequence of pole sets satisfying $\Xi_1 \subset \Xi_2 \subset \dots$ with $|\Xi_k| = k$, the denominator polynomials q_{Ξ_k} of two consecutive spaces differ only by a linear factor such that the rational Krylov spaces are nested

$$\mathcal{Q}^{\Xi_1}_1(\mathbf{L}, \mathbf{b}) \subset \cdots \subset \mathcal{Q}^{\Xi_{\mathcal{I}}}_{\mathcal{I}}(\mathbf{L}, \mathbf{b}) = \mathcal{Q}^{\Xi_{\mathcal{I}+1}}_{\mathcal{I}+1}(\mathbf{L}, \mathbf{b}) = \cdots.$$

Note that $\mathcal{I}$ is independent of the poles, hence $\mathcal{Q}^{\Xi}_{\mathcal{I}}(\mathbf{L}, \mathbf{b}) = \mathcal{K}_{\mathcal{I}}(\mathbf{L}, \mathbf{b})$ for any $\Xi \subset \overline{\mathbb{R}} \setminus \Sigma$. Recognizing this fact, we now focus on a more specific characterization of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$, which, among others, justifies its nomenclature.

Lemma 7.5. *Let* $k \in \mathbb{N}_0$.

1. *The rational Krylov space of* $\mathbf{L}$ *and* $\mathbf{b}$ *with poles in* Ξ *is a polynomial Krylov space of* $\mathbf{L}$ *and* $q_{\Xi}(\mathbf{L})^{-1}\mathbf{b}$, *i.e.,*

$$\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b}) = \mathcal{K}_{k+1}(\mathbf{L}, q_{\Xi}(\mathbf{L})^{-1}\mathbf{b}).$$

2. *There holds*

$$\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b}) = \operatorname{span}\{r_k(\mathbf{L})\mathbf{b} : r_k(\lambda) = p_k(\lambda)/q_{\Xi}(\lambda), p_k \in \mathcal{P}_k\}.$$

3. *If* $\Xi = \{\infty, \xi, \dots, \xi\}$ *for some* $\xi \in \mathbb{R} \setminus \Sigma$, *then the rational Krylov space of* $\mathbf{L}$ *and* $\mathbf{b}$ *with poles in* Ξ *is a polynomial Krylov space of* $(\mathbf{L} - \xi\mathbf{I})^{-1}$ *and* $\mathbf{b}$, *i.e.,*

$$\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b}) = \mathcal{K}_{k+1}((\mathbf{L} - \xi\mathbf{I})^{-1}, \mathbf{b}) = \operatorname{span}\{\mathbf{b}, (\mathbf{L} - \xi\mathbf{I})^{-1}\mathbf{b}, \dots, (\mathbf{L} - \xi\mathbf{I})^{-k}\mathbf{b}\}.$$

4. *If all poles are finite and pairwise distinct, then*

$$\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b}) = \operatorname{span}\{(\mathbf{L} - \xi_0\mathbf{I})^{-1}\mathbf{b}, \dots, (\mathbf{L} - \xi_k\mathbf{I})^{-1}\mathbf{b}\}.$$

Proof. See [Güt10]. □

The first property allows one to transfer several properties that apply to polynomial Krylov spaces to the rational Krylov setting. While the second conjecture is useful for analytical considerations, the third and fourth property are of interest for computational purposes.

7.2 Rayleigh-Ritz Extraction

Throughout this section, we assume f to be a generic function defined on the spectral interval Σ of $\mathbf{L}$. With the specification of the rational Krylov space at hand, we now obtain the rational Krylov approximation of $f(\mathbf{L})\mathbf{b}$ via so-called *Rayleigh-Ritz extraction*, that is

$$\mathbf{u}_{k+1} := \mathbf{V} f(\mathbf{L}_{k+1}) \mathbf{V}^T \mathbf{b}, \qquad \mathbf{L}_{k+1} := \mathbf{V}^T \mathbf{L} \mathbf{V}, \tag{7.2}$$

where $\mathbf{V} \in \mathbb{R}^{N \times (k+1)}$ is a matrix whose columns form an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ with respect to the Euclidean inner product. For the sake of brevity, we call each such matrix a $(\cdot,\cdot)_2$-orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$. The matrix $\mathbf{P} := \mathbf{V}\mathbf{V}^T \in \mathbb{R}^{N \times N}$ is the orthogonal projector of $\mathbb{R}^N$ onto $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, i.e.,

$$\mathbf{P}^2 = \mathbf{P}, \qquad \operatorname{range}(\mathbf{P}) = \mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b}), \qquad \mathbf{P} = \mathbf{P}^T.$$

Hence, $\mathbf{u}_{k+1}$ allows the interpretation as matrix-vector product of the projected matrix $\mathbf{L}_{k+1}$ and $\mathbf{V}^T\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, in the coordinate space $\mathbb{R}^{k+1}$. If $k \ll N$, the rational Krylov approximation can be computed with standard algorithms for dense matrices.

In practical scenarios, $\mathbf{V}$ is orthonormal with respect to some generic inner product $(\cdot,\cdot)$ customized for the current application. E.g., in finite element problems, $(\cdot,\cdot)$ is typically the discrete L^2-inner product $(\cdot,\cdot)_{\mathbf{M}}$, where $\mathbf{M}$ denotes the mass matrix. To generalize (7.2) for these applications, we introduce the concept of pseudo-inverse matrices [BIG03].

Definition 7.6. *Let $k \in \mathbb{N}$ with $k+1 \leq N$, $\mathbf{V} \in \mathbb{R}^{N \times (k+1)}$ a matrix with linear independent columns, and $(\cdot,\cdot)$ an inner product on $\mathbb{R}^N$. We introduce the Moore-Penrose inverse $\mathbf{V}^\dagger \in \mathbb{R}^{(k+1)\times N}$ of $\mathbf{V}$ as the unique solution of the systems of linear equations*

$$\begin{aligned} \mathbf{V}\mathbf{V}^\dagger\mathbf{V} &= \mathbf{V}, & (\mathbf{V}^\dagger\mathbf{V})^* &= \mathbf{V}^\dagger\mathbf{V}, \\ \mathbf{V}^\dagger\mathbf{V}\mathbf{V}^\dagger &= \mathbf{V}^\dagger, & (\mathbf{V}\mathbf{V}^\dagger)^* &= \mathbf{V}\mathbf{V}^\dagger, \end{aligned} \tag{7.3}$$

where for any $\mathbf{P} \in \mathbb{R}^{N\times N}$, $\mathbf{P}^*$ is the adjoint of $\mathbf{P}$ with respect to the inner product $(\cdot,\cdot)$, i.e.,

$$(\mathbf{P}\mathbf{v}, \mathbf{w}) = (\mathbf{v}, \mathbf{P}^*\mathbf{w}), \qquad \mathbf{v}, \mathbf{w} \in \mathbb{R}^N.$$

By construction, there holds

$$(\mathbf{V}\mathbf{V}^\dagger)^2 = \mathbf{V}\mathbf{V}^\dagger, \qquad (\mathbf{V}\mathbf{V}^\dagger)^* = \mathbf{V}\mathbf{V}^\dagger.$$

Therefore, $\mathbf{P} := \mathbf{V}\mathbf{V}^\dagger$ is the orthogonal projector onto the span of $\mathbf{V}$ with respect to the inner product $(\cdot,\cdot)$. Other useful properties of $\mathbf{V}^\dagger$ are collected in the following lemma, where we write $\mathbf{I}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ to denote the unit matrix of dimension $k+1$.

Lemma 7.7. *There holds*

1. $\mathbf{V}^\dagger\mathbf{V} = \mathbf{I}_{k+1}$,

2. $(\mathbf{VT})^\dagger = \mathbf{T}^{-1}\mathbf{V}^\dagger$ *for any* $\mathbf{T} \in \mathbb{R}^{(k+1)\times(k+1)}$ *invertible.*

Proof. See [Güt10, p. 16]. □

The presence of $\mathbf{V}^\dagger$ gives rise to the following generalization of (7.2).

Definition 7.8. *Let* $\mathbf{V} \in \mathbb{R}^{N\times(k+1)}$ *be a basis of* $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$*. We define the rational Krylov approximation of* $f(\mathbf{L})\mathbf{b}$ *as*

$$\mathbf{u}_{k+1} := \mathbf{V} f(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b} \in \mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b}), \qquad \mathbf{L}_{k+1} := \mathbf{V}^\dagger \mathbf{L}\mathbf{V} \in \mathbb{R}^{(k+1)\times(k+1)}.$$

The matrix $\mathbf{L}_{k+1}$ is often referred to as *compression of* $\mathbf{L}$ *on* $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$ and requires the choice of a particular basis $\mathbf{V}$. For any other basis $\mathbf{W} \in \mathbb{R}^{N\times(k+1)}$ of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$, however, there exists some transformation matrix $\mathbf{T} \in \mathbb{R}^{(k+1)\times(k+1)}$ such that $\mathbf{W} = \mathbf{VT}$. It follows from Lemma 7.7 that

$$\hat{\mathbf{L}}_{k+1} := \mathbf{W}^\dagger \mathbf{L}\mathbf{W} = \mathbf{T}^{-1}\mathbf{V}^\dagger \mathbf{L}\mathbf{V}\mathbf{T} = \mathbf{T}^{-1}\mathbf{L}_{k+1}\mathbf{T}. \tag{7.4}$$

Hence, the compression $\hat{\mathbf{L}}_{k+1}$ obtained by $\mathbf{W}$ is similar to the compression $\mathbf{L}_{k+1}$ obtained by $\mathbf{V}$. The rational Krylov approximation $\mathbf{u}_{k+1}$ is entirely independent of the particular basis and thus well-defined.

Lemma 7.9. *Let* $\mathbf{V}$ *and* $\mathbf{W}$ *denote two bases of* $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$ *with compressions*

$$\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L}\mathbf{V}, \qquad \hat{\mathbf{L}}_{k+1} = \mathbf{W}^\dagger \mathbf{L}\mathbf{W}.$$

Then there holds

$$\mathbf{V} f(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b} = \mathbf{W} f(\hat{\mathbf{L}}_{k+1})\mathbf{W}^\dagger \mathbf{b}.$$

Proof. This is a direct consequence of (7.4) and the second property in Lemma 7.7 since

$$\mathbf{W} f(\hat{\mathbf{L}}_{k+1})\mathbf{W}^\dagger \mathbf{b} = \mathbf{VT} f(\mathbf{T}^{-1}\mathbf{L}_{k+1}\mathbf{T})\mathbf{T}^{-1}\mathbf{V}^\dagger \mathbf{b} = \mathbf{V} f(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}.$$ □

Our matrix $\mathbf{L}$ is of the form $\mathbf{L} = \mathbf{M}^{-1}\mathbf{A}$. A straightforward implementation of $\mathbf{L}_{k+1}$ would thus require the inversion of the mass matrix $\mathbf{M}$. Orthonormalizing $\mathbf{V}$ with respect to the discrete L^2-inner product $(\cdot,\cdot)_{\mathbf{M}}$ allows one to overcome this inconvenience.

Lemma 7.10. *Let* $\mathbf{V}$ *be an* $(\cdot,\cdot)_{\mathbf{M}}$*-orthonormal basis of* $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$*,* $\mathbf{L} = \mathbf{M}^{-1}\mathbf{A}$*, and* $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L}\mathbf{V}$*. Then there holds*

1. $\mathbf{V}^\dagger = \mathbf{V}^T\mathbf{M}$,
2. $\mathbf{L}_{k+1} = \mathbf{V}^T\mathbf{A}\mathbf{V}$.

Proof. By direct substitution, one verifies that $\mathbf{V}^T\mathbf{M}$ satisfies (7.3) so that $\mathbf{V}^\dagger = \mathbf{V}^T\mathbf{M}$. The second claim follows from the first one since $\mathbf{V}^\dagger\mathbf{L}\mathbf{V} = \mathbf{V}^T\mathbf{M}\mathbf{L}\mathbf{V} = \mathbf{V}^T\mathbf{M}\mathbf{M}^{-1}\mathbf{A}\mathbf{V} = \mathbf{V}^T\mathbf{A}\mathbf{V}$. □

Provided that $\mathbf{V}$ is orthonormal with respect to $(\cdot,\cdot)_{\mathbf{M}}$, Lemma 7.10 shows that the rational Krylov approximation can be evaluated via

$$\mathbf{u}_{k+1} = \mathbf{V} f(\mathbf{A}_{k+1}) \mathbf{V}^T \mathbf{M} \mathbf{b}, \qquad \mathbf{A}_{k+1} = \mathbf{V}^T \mathbf{A} \mathbf{V},$$

without explicitly computing $\mathbf{M}^{-1}$. Recognizing this fact, we choose

$$(\cdot,\cdot) := (\cdot,\cdot)_{\mathbf{M}}, \qquad \|\cdot\| := \|\cdot\|_{\mathbf{M}} \tag{7.5}$$

as our standard inner product and norm henceforth and say that a basis $\mathbf{V} \in \mathbb{R}^{N\times(k+1)}$ is orthonormal if it is orthonormal with respect to $(\cdot,\cdot)$. Our analysis is carried out in this discrete L^2-setting which appears to be a natural one to study the finite element problems we are interested in.

7.2.1 Properties of the Rational Krylov Method

In this section we collect several remarkable properties of the rational Krylov method which provide the corner stones of our analysis. At first, we study under which assumptions the rational Krylov approximation is exact.

Proposition 7.11. *Let $\mathcal{I}$ be the invariance index of $\mathbf{L}$ and $\mathbf{b}$, $\mathbf{V}$ a basis of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$, and $\mathbf{u}_{k+1} = \mathbf{V} f(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}$. If $k+1 \geq \mathcal{I}$, then there holds $\mathbf{u}_{k+1} = f(\mathbf{L})\mathbf{b}$.*

Proof. [Güt10, Lemma 3.11]. □

It is worth mentioning that $f(\mathbf{L})\mathbf{b} \in \mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$ for some $k+1 < \mathcal{I}$ does not imply exactness of the rational Krylov approximation, cf. [Güt10, Remark 3.12]. However, if

$$f \in \mathcal{P}_k/q_\Xi := \{p_k/q_\Xi : p_k \in \mathcal{P}_k\},$$

the rational Krylov approximation is exact even if $k+1 < \mathcal{I}$.

Lemma 7.12. *Let $\mathbf{V}$ be a basis of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, and $r_k \in \mathcal{P}_k/q_\Xi$. Then the rational Krylov approximation of $r_k(\mathbf{L})\mathbf{b}$ is exact, i.e.,*

$$r_k(\mathbf{L})\mathbf{b} = \mathbf{V} r_k(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}.$$

Proof. See [Güt10, Lemma 4.6]. □

A fundamental ingredient in the analysis of RKMs is the following definition which, due to (7.4) and the fact that similar matrices share the same eigenvalues, is well-defined.

Definition 7.13. *Let $\mathbf{V}$ be a basis of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$ and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$ its compression. The eigenvalues $(\mu_j^{(k)})_{j=0}^k$ of $\mathbf{L}_{k+1}$ are called rational Ritz values.*

The nomenclature $\mu_j^{(k)}$, $j = 0,\dots,k$, is intended remind the reader that the rational Ritz values of $\mathbf{L}$ on $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$ are typically not contained in the rational Ritz values on any larger space. They are contained, however, in the spectral interval of $\mathbf{L}$ and play a fundamental role in Rayleigh-Ritz approximations. Their relation to rational Krylov approximations is laid out in the following theorem and can be found in [Güt10, Theorem 4.8].

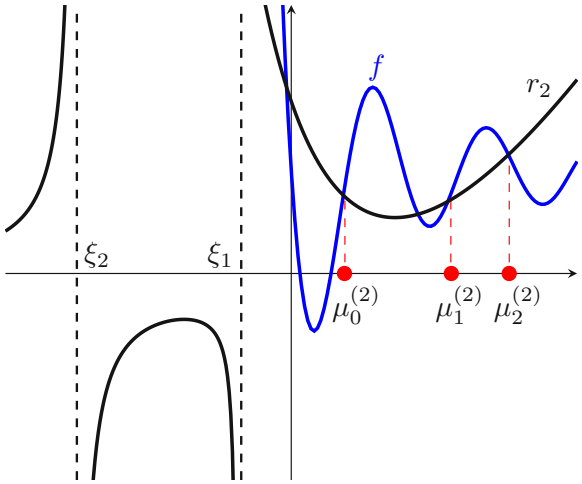

Figure 7.1: Rational function r_2 from Theorem 7.14 with poles in $\Xi = \{\infty, \xi_1, \xi_2\}$ that interpolates f in the rational Ritz values $\Lambda = \{\mu_0^{(2)}, \mu_1^{(2)}, \mu_2^{(2)}\}$ of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$.

Theorem 7.14. *Let* $\mathbf{V}$ *be a basis of* $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, *and* $\mathbf{u}_{k+1} = \mathbf{V} f(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$. *Then there holds*

$$\mathbf{u}_{k+1} = r_k(\mathbf{L})\mathbf{b},$$

where $r_k \in \mathcal{P}_k / q_\Xi$ *is a rational function that interpolates* f *in the rational Ritz values* $\Lambda = \{\mu_0^{(k)}, \dots, \mu_k^{(k)}\}$ *of* $\mathbf{L}$ *on* $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$.

To clarify the situation, we depict the rational interpolant from Theorem 7.14 for $k = 2$ and $\Xi = \{\infty, \xi_1, \xi_2\}$ in Figure 7.1. The rational function $r_2 = p_2 / q_\Xi$ has its poles in $\{\xi_1, \xi_2\}$ and interpolates f in the rational Ritz values $\mu_0^{(2)}, \mu_1^{(2)}$, and $\mu_2^{(2)}$. Since the latter are pairwise distinct, the numerator polynomial p_2 is uniquely determined by the interpolation property $p_2(\mu_j^{(2)}) = f(\mu_j^{(2)})$ for all $j = 0, 1, 2$. Recalling (7.5), another remarkable property we cite here is the following variant of [Güt10, Lemma 4.5].

Proposition 7.15. *Let* $\mathbf{V}$ *be a basis of* $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, $\chi_{k+1} \in \mathcal{P}_{k+1}$ *the characteristic polynomial of* $\mathbf{L}_{k+1}$, *and* $r_{k+1}^* = \chi_{k+1} / q_\Xi$. *Then there holds*

$$\|r_{k+1}^*(\mathbf{L})\mathbf{b}\| = \min_{r_{k+1} \in \mathcal{P}_{k+1}^{\infty} / q_\Xi} \|r_{k+1}(\mathbf{L})\mathbf{b}\|,$$

where $\mathcal{P}_{k+1}^{\infty} / q_\Xi$ *denotes the set of all rational functions* $r_{k+1} \in \mathcal{P}_{k+1} / q_\Xi$ *with monic numerator polynomial.*

The quality of the rational Krylov approximation clearly depends on the rational Krylov space and the way it is extracted from it. As shown in [Güt10, Theorem 4.10], an extraction according to Definition 7.8 yields a quasi-optimal surrogate. We adapt this result in the following theorem to the discrete L^2-norm (7.5), which better suits the study of our problem, and write

$$\|f\|_E := \sup_{\lambda \in E} |f(\lambda)|$$

to denote the supremum norm on $E \subset \mathbb{C}$.

Theorem 7.16. *Let* $\mathbf{V}$ *be an orthonormal basis of* $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, *and* $\mathbf{u}_{k+1} = \mathbf{V} f(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$. *Then there holds*

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k / q_\Xi} \|f - r_k\|_\Sigma. \tag{7.6}$$

Proof. Let $p_k \in \mathcal{P}_k$ be arbitrary and $r_k = p_k / q_\Xi$. Due to Lemma 7.12 there holds $r_k(\mathbf{L})\mathbf{b} = \mathbf{V} r_k(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$ whence

$$\begin{aligned} \|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| &= \|f(\mathbf{L})\mathbf{b} - r_k(\mathbf{L})\mathbf{b} + \mathbf{V} r_k(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b} - \mathbf{V} f(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}\| \\ &\leq \|f(\mathbf{L})\mathbf{b} - r_k(\mathbf{L})\mathbf{b}\| + \|\mathbf{V}(r_k - f)(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}\|. \end{aligned} \tag{7.7}$$

Let now $\mathbf{U}$ denote the matrix of orthonormalized eigenvectors of $\mathbf{L}$, then

$$f(\mathbf{L}) - r_k(\mathbf{L}) = \mathbf{U}(f(\mathbf{D}) - r_k(\mathbf{D}))\mathbf{U}^{-1},$$

where $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix containing the eigenvalues of $\mathbf{L}$. Applying the first property in Lemma 6.7 twice we observe

$$\begin{aligned} \|f(\mathbf{L})\mathbf{b} - r_k(\mathbf{L})\mathbf{b}\| &= \|(f(\mathbf{D}) - r_k(\mathbf{D}))\mathbf{U}^{-1}\mathbf{b}\|_2 \\ &\leq \max_{j=1,\ldots,N} |f(\lambda_j) - r_k(\lambda_j)| \|\mathbf{U}^{-1}\mathbf{b}\|_2 \leq \|f - r_k\|_\Sigma \|\mathbf{b}\|, \end{aligned}$$

bounding the first term on the right-hand side of (7.7). To estimate the latter, we apply the second claim in Lemma 7.10 to see that $\mathbf{L}_{k+1}$ is symmetric and positive definite. Hence, there exists some $(\cdot,\cdot)_2$-orthonormal matrix $\mathbf{U}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$ such that $\mathbf{L}_{k+1} = \mathbf{U}_{k+1} \mathbf{D}_{k+1} \mathbf{U}_{k+1}^T$, where $\mathbf{D}_{k+1} = \operatorname{diag}(\mu_0^{(k)}, \ldots, \mu_k^{(k)})$ is the diagonal matrix containing the rational Ritz values of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$. Since the latter are contained in Σ, it follows from orthonormal properties of $\mathbf{V}$ and $\mathbf{U}_{k+1}$ that

$$\begin{aligned} \|\mathbf{V}(r_k - f)(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}\| &= \|(r_k - f)(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}\|_2 \\ &= \|\mathbf{U}_{k+1}(r_k - f)(\mathbf{D}_{k+1}) \mathbf{U}_{k+1}^T \mathbf{V}^\dagger \mathbf{b}\|_2 \\ &\leq \max_{j=0,\ldots,k} |r_k(\mu_j^{(k)}) - f(\mu_j^{(k)})| \|\mathbf{U}_{k+1} \mathbf{V}^\dagger \mathbf{b}\|_2 \leq \|r_k - f\|_\Sigma \|\mathbf{V}^\dagger \mathbf{b}\|_2. \end{aligned}$$

Since $\mathbf{V}$ is orthonormal, there holds $\|\mathbf{V}^\dagger \mathbf{b}\|_2 = \|\mathbf{V}\mathbf{V}^\dagger \mathbf{b}\|$. Noting that $\mathbf{P} := \mathbf{V}\mathbf{V}^\dagger$ is the orthonormal projector onto $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, we find

$$\|\mathbf{V}(r_k - f)(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}\| \leq \|r_k - f\|_\Sigma \|\mathbf{b}\|.$$

Combining our findings from above yields

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2\|f - r_k\|_\Sigma \|\mathbf{b}\|. \tag{7.8}$$

The conjecture now follows from the observation that the numerator p_k of $r_k = p_k / q_\Xi$ was chosen arbitrary whence (7.8) remains valid if we take the minimum over all $p_k \in \mathcal{P}_k$. □

Remark 7.17. *Inspection of the proof above shows that the upper bound (7.6) can be improved to*

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_\Xi} \|f - r_k\|_E, \qquad E := \sigma(\mathbf{L}) \cup \sigma(\mathbf{L}_{k+1}),$$

where $\sigma(\mathbf{L})$ and $\sigma(\mathbf{L}_{k+1})$ denotes the (discrete) spectra of the matrices $\mathbf{L}$ and $\mathbf{L}_{k+1}$, respectively. This is a min-max problem on a discrete set which might be considerably smaller than the continuous one, especially if the eigenvalues of $\mathbf{L}$ are not uniformly distributed across Σ. Since the eigenvalues of $\mathbf{L}_{k+1}$ are analytically not available, however, one is often obliged to resorts to (7.6).

Remark 7.18. *In Definition 7.1 we require the pole set to be contained in the real line. We mention that the results presented above also hold in the more general case where $\Xi \subset \overline{\mathbb{C}} \setminus \Sigma$, $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. The problems that we are interested in, however, are real whence it is computationally convenient to restrict the pole set to the real line in order to avoid complex arithmetic.*

7.2.2 Computational Aspects

We dedicate our attention to the numerical implementation of the rational Krylov surrogate $\mathbf{u}_{k+1} = \mathbf{V} f(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$. Since its dominant computational costs come from the evaluation of the basis, we focus in this section on the efficient numerical realization of $\mathbf{V}$. Ideally, the construction of $\mathbf{V}$ is both efficient and numerically stable. A procedure that satisfies each of these demands is the *rational Arnoldi algorithm*. Given an orthonormal basis of $\mathcal{Q}^\Xi_{j+1}(\mathbf{L}, \mathbf{b})$, the idea is to choose a vector $\mathbf{u}_j \in \mathcal{Q}^\Xi_{j+1}(\mathbf{L}, \mathbf{b})$ such that $\mathbf{w}_{j+1} := (\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{u}_j \in \mathcal{Q}^\Xi_{j+2}(\mathbf{L}, \mathbf{b}) \setminus \mathcal{Q}^\Xi_{j+1}(\mathbf{L}, \mathbf{b})$. By a standard Gram-Schmidt procedure, this allows us to orthogonalize $\mathbf{w}_{j+1}$ against all previous basis vectors in order to incrementally construct a basis for a rational Krylov space $\mathcal{Q}^\Xi_{j+2}(\mathbf{L}, \mathbf{b})$ given a basis for $\mathcal{Q}^\Xi_{j+1}(\mathbf{L}, \mathbf{b})$. The theoretical justification for this approach is provided in the following lemma, where, for the moment, we restrict ourselves to poles that are nonzero; see [Güt10, Lemma 5.1] for a proof.

Lemma 7.19. *Let $\Xi \subset \overline{\mathbb{R}} \setminus (\Sigma \cup \{0\})$. Then there exists a vector $\mathbf{u}_j \in \mathcal{Q}^\Xi_{j+1}(\mathbf{L}, \mathbf{b})$ with the property $(\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{u}_j \in \mathcal{Q}^\Xi_{j+2}(\mathbf{L}, \mathbf{b}) \setminus \mathcal{Q}^\Xi_{j+1}(\mathbf{L}, \mathbf{b})$ if and only if $j + 1 < \mathcal{I}$.*

Remark 7.20. *If all poles are finite, Lemma 7.19 remains in force if we replace $(\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{u}_j$ by $(\mathbf{L} - \xi_{j+1}\mathbf{I})^{-1}\mathbf{u}_j$ which appears to be more natural in light of the original definition of the rational Krylov space. The former, however, is meaningful also for infinite poles in which case $(\mathbf{I} - \frac{1}{\infty}\mathbf{L})^{-1}\mathbf{L}\mathbf{u}_j = \mathbf{L}\mathbf{u}_j$ designates a polynomial Krylov step.*

The vector $\mathbf{u}_j$ provided by Lemma 7.19 is called *continuation vector*. To simplify matters, we choose this vector, given an orthonormal basis $[\mathbf{v}_0, \ldots, \mathbf{v}_j]$ of $\mathcal{Q}^\Xi_{j+1}(\mathbf{L}, \mathbf{b})$, as $\mathbf{u}_j = \mathbf{v}_j$ henceforth (c.f. Remark 7.21) and refer to [Ruh84, Güt10] for a more general treatment of this matter. Provided a set of poles $\Xi \subset \overline{\mathbb{R}} \setminus (\Sigma \cup \{0\})$, the rational Arnoldi algorithm now proceeds as follows: Starting with $\mathbf{v}_0 := \tilde{\mathbf{v}}_0/\|\tilde{\mathbf{v}}_0\|$, $\tilde{\mathbf{v}}_0 := (\mathbf{I} - \xi_0^{-1}\mathbf{L})^{-1}\mathbf{b}$, one incrementally

generates the new basis vector according to

$$\tilde{\mathbf{v}}_{j+1} := (\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}_j - \sum_{i=0}^{j}(\mathbf{v}_i, (\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}_j)\mathbf{v}_i, \qquad \mathbf{v}_{j+1} := \frac{\tilde{\mathbf{v}}_{j+1}}{\|\tilde{\mathbf{v}}_{j+1}\|},$$

for $j = 0, \dots, k-1$. Worth mentioning, the orthonormalization coefficients $h_{i,j} := (\mathbf{v}_i, (\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}_j)$, $i \leq j$, can be used to obtain a compression of $\mathbf{L}$ without explicitly computing $\mathbf{V}^\dagger\mathbf{L}\mathbf{V}$; see e.g., [Güt10, Chapter 5] for details. In Algorithm 1, we therefore store these coefficients in the Hessenberg matrix

$$\mathbf{H}_{k+1} := \begin{pmatrix} h_{0,0} & \dots & \dots & h_{0,k} \\ h_{1,0} & & & \vdots \\ & \ddots & & \vdots \\ 0 & & h_{k,k-1} & h_{k,k} \end{pmatrix} \in \mathbb{R}^{(k+1)\times(k+1)},$$

where we set $h_{j+1,j} := \|\tilde{\mathbf{v}}_{j+1}\|$.

Algorithm 1 Rational Arnoldi Algorithm

Input: $\mathbf{L} \in \mathbb{R}^{N\times N}$, $\mathbf{b} \in \mathbb{R}^{N\times N}$, $\Xi = \{\xi_0, \dots, \xi_k\} \subset \overline{\mathbb{R}} \setminus (\Sigma \cup \{0\})$

function RationalArnoldi($\mathbf{L}$, $\mathbf{b}$, Ξ)

$\tilde{\mathbf{v}}_0 = (\mathbf{I} - \xi_0^{-1}\mathbf{L})^{-1}\mathbf{b}$

$h_{1,0} = \|\tilde{\mathbf{v}}_0\|$

$\mathbf{v}_0 = h_{1,0}^{-1}\tilde{\mathbf{v}}_0$

for $j = 0, \dots, k-1$ **do**

$\mathbf{w}_{j+1} = (\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}_j$

for $i = 0, \dots, j$ **do**

$h_{i,j} = (\mathbf{v}_i, \mathbf{w}_{j+1})$

end for

$\tilde{\mathbf{v}}_{j+1} = \mathbf{w}_{j+1} - \sum_{i=0}^{j} h_{i,j}\mathbf{v}_i$

$h_{j+1,j} = \|\tilde{\mathbf{v}}_{j+1}\|$

$\mathbf{v}_{j+1} = h_{j+1,j}^{-1}\tilde{\mathbf{v}}_{j+1}$

end for

end function

Output: Basis $\mathbf{V} = [\mathbf{v}_0, \dots, \mathbf{v}_k]$ of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ and matrix $H_{k+1} = (h_{i,j})_{i,j=0}^{k}$.

In its present form, the rational Arnoldi algorithm does not allow for poles at zero. This inconvenience, however, can be overcome if one applies the rational Arnoldi algorithm to the preprocessed matrix $\tilde{\mathbf{L}} := \mathbf{L} - \sigma\mathbf{I}$ and poles $\tilde{\Xi} := \{\xi_0 - \sigma, \dots, \xi_k - \sigma\}$ for some suitable shift $\sigma \in \mathbb{R} \setminus \Xi$. Since $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b}) = \mathcal{Q}_{k+1}^{\tilde{\Xi}}(\tilde{\mathbf{L}}, \mathbf{b})$, one might equally well build a basis for the latter Krylov space having nonzero poles.

Remark 7.21. *There are two possible scenarios in which case Algorithm 1 breaks down in the sense that* $\mathbf{v}_{j+1} = \mathbf{0}$. *The first one eventuates as soon as the invariance index* $j + 1 = \mathcal{I}$ *is reached and is often referred to as lucky breakdown, since in this case the*

rational Krylov approximation is already exact (c.f. Proposition 7.11). Conversely, a so-called unlucky breakdown occurs whenever the enrichment of the basis terminates before the invariance index is reached. In this case, $\mathbf{u}_j = \mathbf{v}_j$ *does not satisfy the desired property from Lemma 7.19 such that* $(\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}_j \in \operatorname{span}\{\mathbf{v}_0, \dots, \mathbf{v}_j\}$*. However, our choice of the continuation vector* $\mathbf{u}_j$ *is fairly standard in the literature and we have never observed such an unlucky breakdown in any of our experiments.*

The bottle neck of each rational Arnoldi algorithm is the computation of solutions to the shifted linear systems of equations $(\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}_j$. The latter reduces to a matrix-vector product if $\xi_{j+1} = \infty$. Therefore, a legitimate question is whether the choice of finite poles pays off taking into account the additional complexity compared to polynomial Krylov methods. The following example addresses this matter.

Example 7.22. *Consider the heat equation*

$$\begin{aligned} \partial_t u - \Delta u &= 0, && \text{in } \Omega \times \mathbb{R}^+, \\ u &= 0, && \text{on } \partial\Omega \times \mathbb{R}^+, \\ u &= u_0, && \text{on } \Omega \times \{0\}, \end{aligned}$$

on the unit square $\Omega = (0,1)^2$ *with initial condition*

$$u_0(\mathbf{x}) = xy(1-x)(1-y), \qquad \mathbf{x} = (x,y) \in \Omega.$$

We apply the discrete eigenfunction method using $V_h = \mathcal{P}_1^0(\mathcal{T}_h)$ *as finite element space over a quasi-uniform triangular mesh* $\mathcal{T}_h$*. The DEM surrogate can be written as* $\mathbf{u}^{\text{DEM}} = e^{-t\mathbf{L}}\mathbf{u}_0$*. The latter is approximated using a rational Krylov space* $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ *with poles in*

- $\Xi_{\text{inf}} = \{\infty, \dots, \infty\}$*, which yields a polynomial Krylov method,*
- $\Xi_{\text{si}} = \{1, \dots, 1\}$*, which yields a shift-and-invert Krylov method.*

In dependency of the mesh parameter h*, we introduce the discrete* L^2*-error*

$$E(k, \Xi, t, h) := \|e^{-t\mathbf{L}}\mathbf{u}_0 - \mathbf{V}e^{-t\mathbf{L}_{k+1}}\mathbf{V}^\dagger\mathbf{u}_0\|,$$

where $\mathbf{V}$ *is an orthonormal basis of* $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ *and* $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$*. We depict the quantity* $E(k, \Xi, 0.1, h)$ *in Figure 7.2 for decreasing mesh sizes* $h = 0.04, 0.02, 0.01$*. The error of the orthogonal projection* $\mathbf{V}\mathbf{V}^\dagger e^{-0.1\mathbf{L}}\mathbf{u}_0$ *of* $e^{-0.1\mathbf{L}}\mathbf{u}_0$ *onto the respective Krylov space is indicated by the triangles and serves as a benchmark for the best possible approximation within the search space. In accordance with Theorem 7.16, the latter almost perfectly matches the actual Krylov approximation error.*

We further observe that the number k *required by the polynomial Krylov method to achieve a desired accuracy is almost proportional to the mesh parameter* h*. In contrast, the accuracy of the RKM using* Ξ_{si} *as pole set appears to be independent of* h*. For all values of the mesh size, the latter reaches machine precision whenever* $k \approx 15$ *irrespectively of the problem size. Under the assumption that each iteration involves a linear system of dimension* $N \times N$ *that can be solved in* $\mathcal{O}(N)$ *operations, the benefits of finite poles quickly justify the additional computational complexity.*

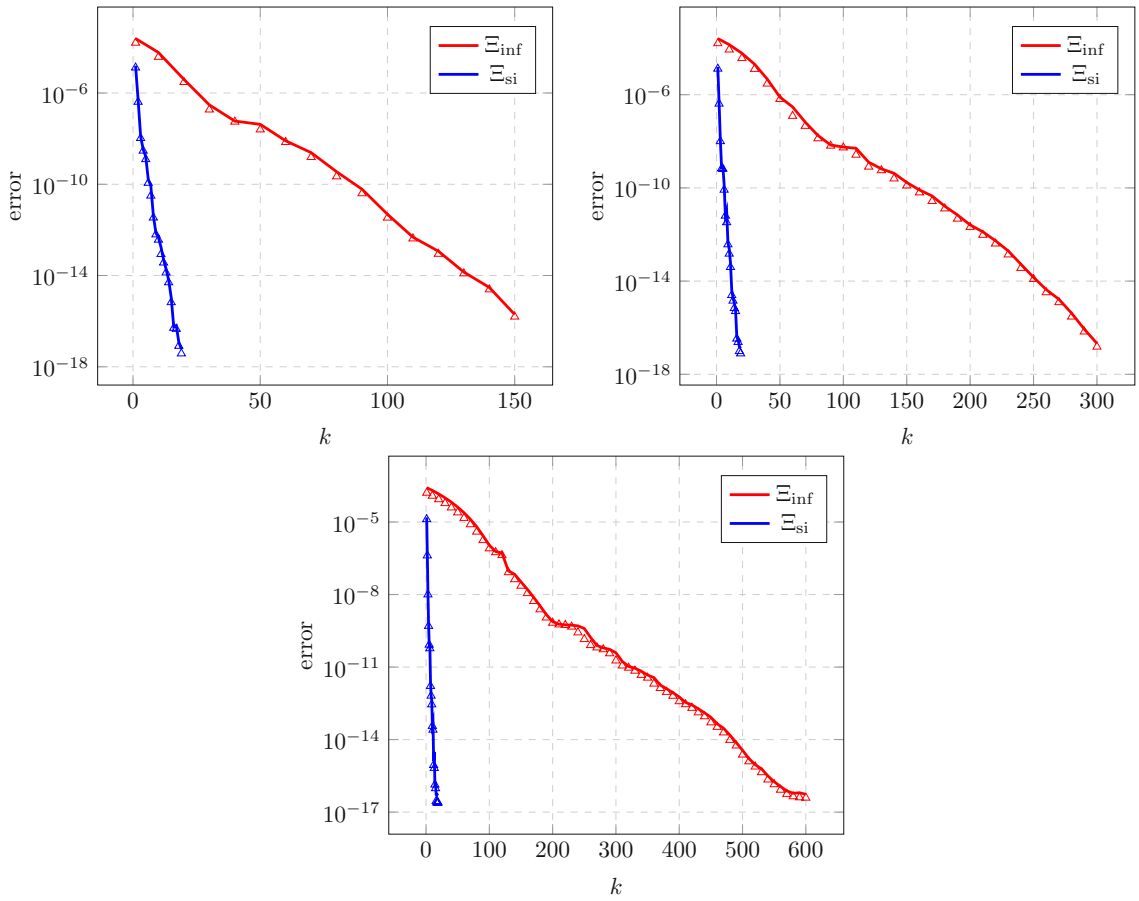

Figure 7.2: Error $E(k, \Xi, 0.1, h)$ for $h = 0.04$ (top left), $h = 0.02$ (top right), and $h = 0.01$ (bottom) with $\Xi \in \{\Xi_{\text{inf}}, \Xi_{\text{si}}\}$. The triangles indicate the orthogonal projection of the DEM surrogate $e^{-0.1\mathbf{L}}\mathbf{u}_0$ onto the respective Krylov space.

As the previous example shows, the challenge with Krylov subspace methods is to find the subtle trade-off between finite poles, which typically yield better search spaces, and poles at infinity, which are computationally less expensive. Since this is still a matter of ongoing research [GS21] and far beyond the scope of this thesis, we only want to provide a superficial overview of some aspects which should be included in the decision of the particular selection of poles and the computation of the basis.

– A general rule of thumb is that for small N the higher computational effort of rational Krylov methods does not pay off compared to their polynomial counterparts.

– If N is of moderate size that allows the systems of equations to be solved using a direct solver, it is convenient to choose Ξ in a way such that the poles do not vary often. In this case, one can compute a factorization for the matrix $\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L}$ once and for all in order to efficiently query $\mathbf{v} \mapsto (\mathbf{I} - \xi_{j+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}$ for all poles that share the same value.

– If N is large, direct solvers entail extensive memory requirements whence one is typically obliged to resort to iterative methods. In this regime, there is no computational advantage in solving multiple linear systems with the same poles.

– Typically, one does not solve the linear system involving $\mathbf{L} = \mathbf{M}^{-1}\mathbf{A}$ but resorts to the equivalent problem

$$(\mathbf{M} - \xi_{j+1}^{-1}\mathbf{A})\tilde{\mathbf{v}}_{j+1} = \mathbf{M}\mathbf{v}_j, \tag{7.10}$$

which is computationally convenient since both $\mathbf{M}$ and $\mathbf{A}$ are symmetric and sparse.

– The performance of iterative methods hinges on the condition number of $\mathbf{L}$. The latter is typically large and thus requires the implementation of a preconditioner. If the matrix comes from a differential operator of the form (4.1), efficient multigrid preconditioner are available whose convergence rates are independent of the respective pole and the mesh size.

– If all poles are pairwise distinct, then by Lemma 7.5

$$\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b}) = \operatorname{span}\{(\mathbf{L} - \xi_0\mathbf{I})^{-1}\mathbf{b}, \dots, (\mathbf{L} - \xi_k\mathbf{I})^{-1}\mathbf{b}\}.$$

This allows one to distribute each task $(\mathbf{L} - \xi_j\mathbf{I})^{-1}\mathbf{b}$, $j = 0, \dots, k$, to one processor in order to solve the systems of equations efficiently in parallel. In favour of numerical stability, the resulting vectors are then orthonormalized to obtain the desired basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$. Compared to the sequentially executed rational Arnoldi algorithm, however, the parallel approach is more prone to numerical instabilities which needs to be counteracted by repeated orthogonalization runs in each step. A more general point of view on parallel Arnoldi algorithms is given in [Güt10, BG17].

– If the linear systems of equations are tackled by iterative methods, one possibility to accelerate the computations even further is to terminate the iteration before machine precision is attained. These so-called *inexact solves* give rise to a perturbed rational Krylov basis and have been systematically studied in [LM98, Güt10, GS21].

– A rather recent tool that can be applied to build the rational Krylov space are so-called $\mathcal{H}$-matrices [Hac99]. $\mathcal{H}$-matrices are blockwise low-rank matrices that can be stored at logarithmic-linear complexity and allow for efficient approximations of inverse FEM matrices [FMP13, AFM21]. The latter can be employed to efficiently compute solutions to (7.10).

8 A Unified Analysis of Rational Krylov Methods in Fractional Diffusion

The key ingredient of each rational Krylov method is the particular selection of its poles. Typically, these parameters are chosen based on some upper bound η of the rational Krylov error

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq \eta(f, \mathbf{L}, \mathbf{b}, \Xi)$$

that quantifies the performance of the RKM in dependence of Ξ. One possible choice of η is provided by Theorem 7.16 via

$$\eta(f, \mathbf{L}, \mathbf{b}, \Xi) = 2\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_\Xi} \|f - r_k\|_\Sigma.$$

Whenever $r_k \approx f$ in Σ, the poles of r_k should constitute good poles for building the rational Krylov space. In fractional diffusion problems, however, the function $f = f^{\boldsymbol{\tau}}$ that we are interested in typically depends on a parameter vector $\boldsymbol{\tau} \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$, such as

- the power function $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ with $\boldsymbol{\tau} = s \in \Theta = [0, 1]$,
- functions of Mittag-Leffler type $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s)$ with $\boldsymbol{\tau} \in \{(\alpha, \beta, t, s) \in [0, 1] \times \mathbb{R}^+ \times \mathbb{R}^+ \times [0, 1] : \beta \geq \alpha\}$.
- We also aim to incorporate the power function $f^{\boldsymbol{\tau}}(\lambda) = \lambda^s$, $\boldsymbol{\tau} = s \in \Theta = [0, 1]$, with positive exponent in our analysis, which is of interest in the application of forward fractional diffusion operators, the computation of interpolation norms, and the implementation of time-stepping schemes for time-dependent problems generated by a fractional-in-space differential operator.

In all these cases, any possible rational approximation $r_k^{\boldsymbol{\tau}} \approx f^{\boldsymbol{\tau}}$ heavily depends on the problem-specific parameters and is thus unfeasible whenever $\boldsymbol{\tau} \mapsto \mathbf{u}_{k+1} \approx f^{\boldsymbol{\tau}}(L)\mathbf{b}$ is queried for multiple instances of the parameter. While the precise choice of the poles is discussed systematically in Chapter 10, we provide here the necessary preparations and quantify the rational Krylov error by an upper bound of the form $\eta(f^{\boldsymbol{\tau}}, \mathbf{L}, \mathbf{b}, \Xi) = \eta_1(f^{\boldsymbol{\tau}}, \mathbf{L}, \mathbf{b})\eta_2(\mathbf{L}, \Xi)$ which allows us to choose Ξ independently of $\boldsymbol{\tau}$. Following [DHS21], we proceed in three steps.

1. In the first step, we show that the functions we are interested in are either of Cauchy-Stieltjes, complete Bernstein, or Laplace-Stieltjes type, i.e., they admit a representation of the form

$$f^{\boldsymbol{\tau}}(\lambda) = \int_0^\infty \mu^{\boldsymbol{\tau}}(\zeta) g(\lambda, \zeta)\, d\zeta, \qquad g(\lambda, \zeta) \in \{(\lambda + \zeta)^{-1}, \lambda/(\lambda + \zeta), e^{-\zeta\lambda}\}, \tag{8.1}$$

   where $\mu^{\boldsymbol{\tau}}$ is a real-valued function such that the integral is absolutely convergent.

2. Leveraging our knowledge gained, we apply Theorem 2.37 to write $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ and its rational Krylov surrogate as

$$f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} = \int_0^\infty \mu^{\boldsymbol{\tau}}(\zeta) g(\mathbf{L},\zeta)\, d\zeta, \qquad \mathbf{V} f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b} = \int_0^\infty \mu^{\boldsymbol{\tau}}(\zeta)\mathbf{V} g(\mathbf{L}_{k+1},\zeta)\mathbf{V}^\dagger \mathbf{b}\, d\zeta,$$

for any basis $\mathbf{V}$ of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$ with $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L}\mathbf{V}$. Since

$$\begin{aligned} \|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{V} f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}\| &= \left\| \int_0^\infty \mu^{\boldsymbol{\tau}}(\zeta) \left(g(\mathbf{L},\zeta) - \mathbf{V} g(\mathbf{L}_{k+1},\zeta)\mathbf{V}^\dagger \mathbf{b} \right) d\zeta \right\| \\ &\leq \int_0^\infty \mu^{\boldsymbol{\tau}}(\zeta) \| g(\mathbf{L},\zeta) - \mathbf{V} g(\mathbf{L}_{k+1},\zeta)\mathbf{V}^\dagger \mathbf{b}\|\, d\zeta, \end{aligned} \tag{8.2}$$

the main objective of the second part of this chapter is to quantify the rational Krylov error of either of the three kernel functions

$$\|g(\mathbf{L},\zeta) - \mathbf{V} g(\mathbf{L}_{k+1},\zeta)\mathbf{V}^\dagger \mathbf{b}\| \leq \eta_g(\Xi,\zeta) \tag{8.3}$$

by some suitable upper bound η_g which is entirely independent of the parameter $\boldsymbol{\tau}$.

3. Finally, in the remainder of this chapter, we combine (8.2) and (8.3) to derive estimates

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq \eta_1(f^{\boldsymbol{\tau}},\mathbf{L},\mathbf{b})\eta_2(\mathbf{L},\Xi)$$

for all functions of the form (8.1).

8.1 Stieltjes and Complete Bernstein Functions in Fractional Diffusion Problems

8.1.1 Cauchy-Stieltjes Functions

One possible definition of Cauchy-Stieltjes functions is obtained by setting $g(\lambda,\zeta) = 1/(\lambda+\zeta)$ in (8.1). In slightly more general form, the definition we shall use throughout this thesis is the following [SSV12, Ber08].

Definition 8.1. *A function $f : \mathbb{R}^+ \to \mathbb{R}$ is said to be a Cauchy-Stieltjes function if*

$$f(\lambda) = \frac{\omega}{\lambda} + \theta + \int_0^\infty \frac{\mu_C(\zeta)}{\lambda+\zeta}\, d\zeta \tag{8.4}$$

for $\omega,\theta \in \mathbb{R}_0^+$ and some nonnegative real-valued function μ_C such that the integral is absolutely convergent. The function μ_C is called Cauchy-Stieltjes density and (ω,θ,μ_C) is the Cauchy-Stieltjes triple of f. We denote the set of all Cauchy-Stieltjes functions by $\mathcal{CS}$.

There is a one-to-one relation between $f \in \mathcal{CS}$ and its respective Cauchy-Stieltjes triple (ω,θ,μ_C), whence the map $\mu_C \mapsto f$ is commonly referred to as *Cauchy-Stieltjes transform* [EMOT54]. Typical examples of Cauchy-Stieltjes functions are [SSV12, p. 13-14]

$$\frac{\ln(1+\lambda)}{\lambda} = \int_1^\infty \frac{\zeta^{-1}}{\lambda+\zeta}\, d\zeta, \qquad \frac{\arctan\left(\sqrt{\lambda}\right)}{\sqrt{\lambda}} = \int_0^1 \frac{(4\zeta)^{-\frac{1}{2}}}{\lambda+\zeta}\, d\zeta, \tag{8.5}$$

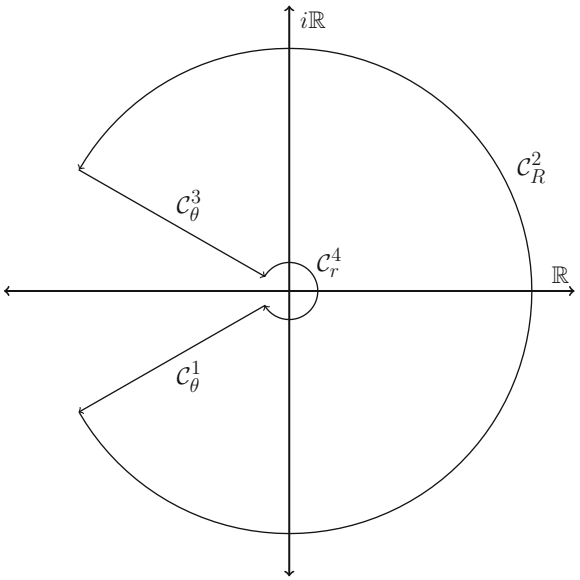

Figure 8.1: Integration contour $\mathcal{C}(\theta, r, R)$ defined by (8.6).

where $\omega = \theta = 0$ in (8.4). For some $f \in \mathcal{CS}$, the corresponding density function can be computed explicitly. To see this, assume f to be analytic in $\mathbb{C} \setminus \mathbb{R}_0^-$ and define the contour $\mathcal{C}(\theta, r, R) := \mathcal{C}_\theta^1 + \mathcal{C}_R^2 - \mathcal{C}_\theta^3 - \mathcal{C}_r^4$, depicted in Figure 8.1, where

$$\left.\begin{aligned} \mathcal{C}_\theta^1(\zeta) &:= \zeta e^{-i\theta}, & \mathcal{C}_R^2(\phi) &:= Re^{i\phi} \\ \mathcal{C}_\theta^3(\zeta) &:= \zeta e^{i\theta}, & \mathcal{C}_r^4(\phi) &:= re^{-i\phi} \end{aligned}\right\} (\zeta, \phi) \in [r, R] \times (-\pi, \pi), \tag{8.6}$$

for some $\theta \in (0, \pi)$ and $0 < r < R$. By Cauchy's integral formula, cf. Theorem 2.35, we find for all $\lambda \in (r, R)$

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi i} \int_{\mathcal{C}(\theta,r,R)} \frac{f(z)}{\lambda - z}\, dz \\ &= \frac{1}{2\pi i} \int_r^R \frac{f(\zeta e^{-i\theta})e^{-i\theta}}{\lambda - \zeta e^{-i\theta}} - \frac{f(\zeta e^{i\theta})e^{i\theta}}{\lambda - \zeta e^{i\theta}}\, d\zeta + \frac{1}{2\pi} \int_{-\theta}^{\theta} \frac{Rf(Re^{i\phi})e^{i\phi}}{\lambda - Re^{i\phi}} - r\frac{f(re^{-i\phi})e^{-i\phi}}{\lambda - re^{-i\phi}}\, d\phi. \end{aligned}$$

If we assume

$$\lim_{|z|\to\infty} |f(z)| = 0, \qquad \lim_{|z|\to 0} |zf(z)| = 0, \tag{8.7}$$

uniformly in $\arg(z) \in (-\pi, \pi)$, we can send $R \to \infty$ and $r \to 0$, causing the latter integral to vanish. In combination with $\theta \to \pi$, we thus obtain, after the transformation $\zeta \mapsto -\zeta$,

$$f(\lambda) = \int_0^\infty \frac{\mu_C(\zeta)}{\lambda + \zeta}\, d\zeta, \quad \mu_C(\zeta) = \frac{1}{2\pi i} \lim_{\phi\to\pi} \left(f(\zeta e^{i\phi}) - f(\zeta e^{-i\phi}) \right). \tag{8.8}$$

We arrive at the following result.

Lemma 8.2. *Let f be analytic in $\mathbb{C}\setminus\mathbb{R}_0^-$ so that (8.7) holds. Then $f\in\mathcal{CS}$ and its Cauchy-Stieltjes density μ_C is given by*

$$\mu_C(\zeta)=\frac{1}{2\pi i}\lim_{\phi\to\pi}\left(f(\zeta e^{i\phi})-f(\zeta e^{-i\phi})\right). \tag{8.9}$$

Although (8.9) provides a convenient tool to determine Cauchy-Stieltjes densities for some $f\in\mathcal{CS}$, we emphasize that the analysis provided in the sequel can be applied to any $f\in\mathcal{CS}$ without the explicit knowledge of μ_C. For this purpose, the following proposition is often useful, where we use the notation

$$\begin{aligned}\mathbb{H}_{\Im>0}&:=\{z\in\mathbb{C}:\Im z>0\},&\qquad \mathbb{H}_{\Im<0}&:=\{z\in\mathbb{C}:\Im z<0\},\\ \mathbb{H}_{\Im\geq 0}&:=\{z\in\mathbb{C}:\Im z\geq 0\},&\qquad \mathbb{H}_{\Im\leq 0}&:=\{z\in\mathbb{C}:\Im z\leq 0\},\end{aligned}$$

for the (closed) upper and lower complex half plane, respectively.

Proposition 8.3. *A function $f:\mathbb{R}^+\to\mathbb{R}$ is a Cauchy-Stieltjes function if and only if*

1. *$f(\lambda)\in\mathbb{R}_0^+$ for all $\lambda\in\mathbb{R}^+$,*
2. *f has an analytic extension to the slit plane $\mathbb{C}\setminus\mathbb{R}_0^-$, that we call f again, such that $f(z)\in\mathbb{H}_{\Im\leq 0}$ for all $z\in\mathbb{H}_{\Im>0}$.*

Proof. See [Ber08, Theorem 3.2]. □

Example 8.4. *Since the function*

$$f(z):=\frac{1}{z(1+z^2)}$$

possesses poles at $\pm i$, we find that f cannot be extended analytically to the slit plane $\mathbb{C}\setminus\mathbb{R}_0^-$. By Proposition 8.3, there holds $f\notin\mathcal{CS}$. On the other hand, the function

$$h(\lambda):=\frac{\sqrt{\lambda+c}}{\lambda}\in\mathcal{CS}$$

for all $c\in\mathbb{R}^+$, since it is nonnegative on $\mathbb{R}^+$ and any $z\in\mathbb{H}_{\Im>0}$ satisfies

$$\arg(h(z))=\arg(\sqrt{z+c})-\arg(z)=\frac{1}{2}\arg(z+c)-\arg(z)\in(-\pi,0).$$

Provided that f is a real-valued function defined on $\mathbb{R}^+$, it follows by the Schwarz reflection principle [Hen93] that any possible analytic continuation of f must satisfy $f(\overline{z})=\overline{f(z)}$. Hence, Proposition 8.3 shows that f maps the upper half plane to the lower half plane and vice versa. Several other remarkable properties of Cauchy-Stieltjes functions are listed in the following proposition, which are often useful to derive the Cauchy-Stieltjes membership of some function from available $f,g\in\mathcal{CS}$.

Proposition 8.5.

1. *The set $\mathcal{CS}$ is a convex cone: $af+bg\in\mathcal{CS}$ for all $a,b\in\mathbb{R}_0^+$ and $f,g\in\mathcal{CS}$.*

2. *The set $\mathcal{CS}$ is closed under pointwise limits: if $(f_n)_{n\in\mathbb{N}} \subset \mathcal{CS}$ converges pointwise to some f, then $f \in \mathcal{CS}$.*

3. *Let $f, g \in \mathcal{CS}$, $f \neq 0$, $c \in \mathbb{R}^+$, and $s \in [0,1]$. Then there holds*

$$\frac{1}{f(\frac{1}{\lambda})} \in \mathcal{CS}, \qquad \frac{1}{\lambda f(\lambda)} \in \mathcal{CS}, \qquad \frac{f}{cf+1} \in \mathcal{CS}, \qquad g \circ \frac{1}{f} \in \mathcal{CS}, \qquad f^s g^{1-s} \in \mathcal{CS}.$$

Proof. See [SSV12, Theorem 2.2] for the first two claims and [Ber08, p. 9] for the latter. □

The importance of Cauchy-Stieltjes functions in fractional diffusion problems is due to Balakrishnan's formula, which shows that $f(\lambda) = \lambda^{-s} \in \mathcal{CS}$ for all $s \in [0,1]$. This membership also easily follows from Proposition 8.2 since f satisfies (8.8) and

$$\mu_C(\zeta) = \frac{1}{2\pi i} \lim_{\phi\to\pi} \left(f(\zeta e^{i\phi}) - f(\zeta e^{-i\phi}) \right) = \frac{1}{2\pi i} \lim_{\phi\to\pi} \left(\zeta^{-s} e^{i\phi s} - \zeta^{-s} e^{-i\phi s} \right) = \zeta^{-s} \frac{\sin(\pi s)}{\pi}.$$

We manifest this important observation in the following theorem.

Theorem 8.6. *For all $s \in [0,1]$ there holds $\lambda^{-s} \in \mathcal{CS}$. If $s \in (0,1)$, then*

$$\lambda^{-s} = \frac{\sin(\pi s)}{\pi} \int_0^\infty \frac{\zeta^{-s}}{\lambda + \zeta}\, d\zeta, \qquad \lambda \in \mathbb{R}^+. \tag{8.10}$$

The fractional power function $f(\lambda) = \lambda^{-s}$, $s \in [0,1]$, shall serve us as paradigm of a Cauchy-Stieltjes function arising from elliptic fractional diffusion problems. We stress, however, that the results provided in the sequel also apply to other possibly interesting functions of fractional diffusion type, e.g.,

$$\frac{1}{\lambda^s + c} \in \mathcal{CS}, \qquad \frac{1}{(\lambda + c)^s} \in \mathcal{CS}, \qquad c \in \mathbb{R}_0^+.$$

Note that not all functions we are interested in have Cauchy-Stieltjes membership. Consider e.g., the fractional power function $f(z) = z^s$ with positive exponent $s \in (0,1)$. Then

$$f(i) = e^{i\pi s} = \cos(\pi s) + i \sin(\pi s) \in \mathbb{H}_{\Im>0}.$$

Since $i \in \mathbb{H}_{\Im>0}$, it follows from Proposition 8.3 that $f \notin \mathcal{CS}$. Nevertheless, we can exploit (8.10) to deduce

$$\frac{\lambda^s}{\lambda} = \lambda^{s-1} = \frac{\sin(\pi(s-1))}{\pi} \int_0^\infty \frac{\zeta^{s-1}}{\lambda + \zeta}\, d\zeta.$$

Multiplication with λ combined with the symmetry of sin around $\frac{\pi}{2}$ gives

$$\lambda^s = \frac{\sin(\pi s)}{\pi} \int_0^\infty \zeta^{s-1} \frac{\lambda}{\lambda + \zeta}\, d\zeta, \tag{8.11}$$

which has already been noted in Theorem 4.7. We conclude that λ^s can be written as (8.1) with resolvent-type kernel $g(\lambda, \zeta) = \lambda/(\lambda + \zeta)$. Functions of this form are called *complete Bernstein functions* and are the matter of the following section.

8.1.2 Complete Bernstein Functions

Definition 8.7. *A function $f : \mathbb{R}^+ \to \mathbb{R}$ is said to be a complete Bernstein function if*

$$f(\lambda) = \omega\lambda + \theta + \int_0^\infty \mu_B(\zeta)\frac{\lambda}{\lambda+\zeta}\,d\zeta \tag{8.12}$$

for $\omega, \theta \in \mathbb{R}_0^+$ and some nonnegative real-valued function μ_B such that the integral is absolutely convergent. The function μ_B is called Bernstein density and (ω, θ, μ_B) is the complete Bernstein triple of f. We denote the set of all complete Bernstein function by $\mathcal{CB}$.

Similarly to Cauchy-Stieltjes functions, the triplet (ω, θ, μ_B) is uniquely determined by the function f and vice versa. Typical examples of complete Bernstein functions are [SSV12]

$$\ln(1+\lambda) = \int_0^1 \zeta^{-1}\frac{\lambda}{\lambda+\zeta}\,d\zeta, \qquad \sqrt{\lambda}\left(1+e^{-2\sqrt{\lambda}}\right) = \frac{2}{\pi}\int_0^\infty \frac{\cos^2(\sqrt{\zeta})}{\sqrt{\zeta}}\frac{\lambda}{\lambda+\zeta}\,d\zeta.$$

A systematic list of complete Bernstein functions combined with their respective Bernstein densities, if available, can be found in [SSV12]. The one that is of utmost interest for us is the matter of the following theorem.

Theorem 8.8. *For all $s \in [0,1]$ there holds $\lambda^s \in \mathcal{CB}$.*

Proof. The conjecture follows from (8.11). □

The first equivalence in the following proposition is a direct consequence of the integral representation of Cauchy-Stieltjes and complete Bernstein functions. For the remaining ones we refer to Proposition 7.1 and Theorem 7.3 in [SSV12].

Proposition 8.9. *There holds $f \in \mathcal{CB}$ if and only if $f(\lambda)/\lambda \in \mathcal{CS}$. Assuming $f \not\equiv 0$, then*

$$f \in \mathcal{CB} \iff \frac{1}{f} \in \mathcal{CS} \iff \frac{\lambda}{f(\lambda)} \in \mathcal{CB}. \tag{8.13}$$

Since $f(z) = 1/z = \overline{z}/|z|^2$ maps $\mathbb{H}_{\Im\geq 0}$ onto $\mathbb{H}_{\Im\leq 0}$ and vice versa, we infer from the first equivalence in (8.13) and Proposition 8.3 the following characterization of $\mathcal{CB}$.

Proposition 8.10. *A function $f : \mathbb{R}^+ \to \mathbb{R}$ is a complete Bernstein function if and only if*

1. *$f(\lambda) \in \mathbb{R}_0^+$ for all $\lambda \in \mathbb{R}^+$,*
2. *f has an analytic continuation to the complex slit plane $\mathbb{C}\setminus\mathbb{R}_0^-$ such that $f(z) \in \mathbb{H}_{\Im\geq 0}$ for all $z \in \mathbb{H}_{\Im>0}$.*

The different characterizations of $\mathcal{CS}$ and $\mathcal{CB}$, stated in Proposition 8.3 and 8.10, are illustrated in Figure 8.2. We sample several points along the unit circle and plot their image under the Cauchy-Stieltjes and complete Bernstein map $z^{-\frac{1}{2}}$ and $z^{\frac{1}{2}}$, respectively. In both cases, the nodes are mapped to the right-half plane, however, in the Cauchy-Stieltjes case the nodes contained in $\mathbb{H}_{\Im>0}$ are mapped to the lower-half plane whereas z^s preserves the sign of their imaginary part. In view of this close relation between $\mathcal{CB}$ and $\mathcal{CS}$, the following properties should come as no surprise.

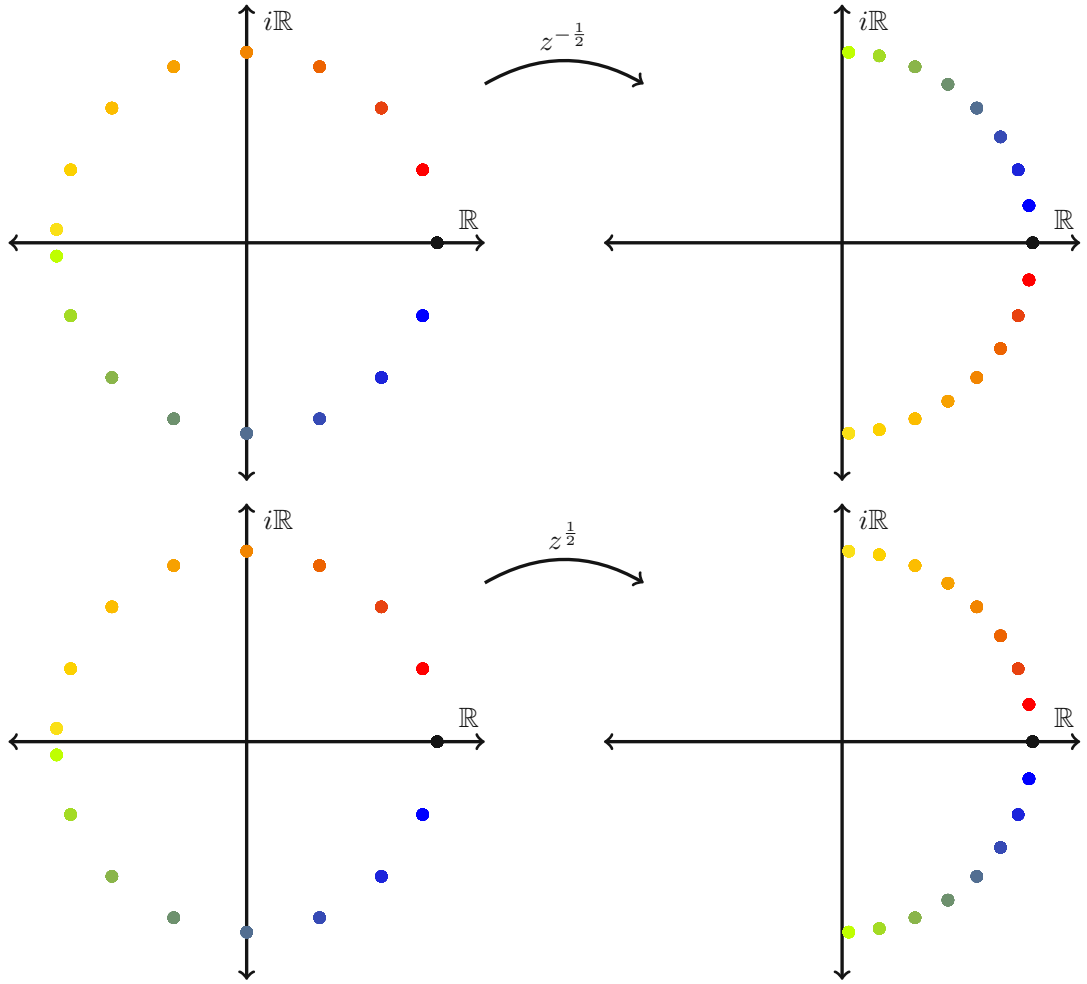

Figure 8.2: Action of $z^{-\frac{1}{2}} \in \mathcal{CS}$ and $z^{\frac{1}{2}} \in \mathcal{CB}$ on arguments on the unit circle.

Lemma 8.11.

1. *The set $\mathcal{CB}$ is a convex cone: $af + bg \in \mathcal{CB}$ for all $a, b \in \mathbb{R}_0^+$ and $f, g \in \mathcal{CB}$.*
2. *The set $\mathcal{CB}$ is closed under pointwise limits: if $(f_n)_{n\in\mathbb{N}} \subset \mathcal{CB}$ converges pointwise to some f, then $f \in \mathcal{CB}$.*
3. *Let $f, g \in \mathcal{CB}$ with $f \neq 0$, $c \in \mathbb{R}^+$, and $s \in [0, 1]$. Then there holds*

$$\frac{1}{f(\frac{1}{\lambda})} \in \mathcal{CB}, \quad \lambda f\Big(\frac{1}{\lambda}\Big) \in \mathcal{CB}, \quad \frac{f(\lambda)}{c\lambda + 1} \in \mathcal{CB}, \quad \frac{\lambda}{f(\lambda)} \in \mathcal{CB}, \quad f^s g^{1-s} \in \mathcal{CB}.$$

Proof. See [SSV12, Section 7]. □

The following result comes in handy if one wants to construct new Cauchy-Stieltjes or complete Bernstein functions from given functions in $\mathcal{CS}$ or $\mathcal{CB}$. For the sake of brevity we

use $A \circ B$ as a shorthand for $\{a \circ b : a \in A, b \in B\}$. Its proof is a direct consequence of Proposition 8.3 and 8.10.

Corollary 8.12. *There holds*

1. $\mathcal{CB} \circ \mathcal{CS} \subset \mathcal{CS}$,
2. $\mathcal{CS} \circ \mathcal{CB} \subset \mathcal{CS}$,
3. $\mathcal{CB} \circ \mathcal{CB} \subset \mathcal{CB}$,
4. $\mathcal{CS} \circ \mathcal{CS} \subset \mathcal{CB}$.

Apart from λ^s, the set of complete Bernstein functions contains several interesting functions of fractional type, such as

$$\lambda^s + c \in \mathcal{CB}, \qquad \frac{\lambda^s}{c\lambda^s + 1} \in \mathcal{CB}, \qquad c \in \mathbb{R}_0^+,$$

for all $s \in [0, 1]$. In combination with $\mathcal{CS}$, they cover a majority of the most important functions that arise from stationary fractional diffusion problems. In fractional evolution equations, however, we are interested in functions of the form $f^\tau(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s)$. Even in the integer-order case $\alpha = \beta = s = 1$, the latter is not contained in $\mathcal{CS} \cup \mathcal{CB}$ since for any $z = re^{i\phi}$ we have

$$e^{-z} = e^{-r\cos\phi} e^{-ir\sin\phi} = e^{-r\cos\phi} \left(\cos(r\sin\phi) - i\sin(r\sin\phi)\right). \tag{8.14}$$

Provided that $\phi \in (0, \frac{\pi}{2})$ there holds $z \in \mathbb{H}_{\Im>0}$ but for e.g., $r = 1$, we have $\Im(e^{-z}) < 0$ whence $e^{-\lambda} \notin \mathcal{CB}$. On the other hand, if we choose $r = 3\pi/(2\sin\phi)$ there holds

$$\Im(e^{-z}) = -e^{-\frac{3\pi\cot\phi}{2}} \sin\left(\frac{3\pi}{2}\right) = e^{-\frac{3\pi\cot\phi}{2}} > 0,$$

whence $e^{-\lambda} \notin \mathcal{CS}$ by Proposition 8.3. To complete our systematic classification, we introduce one final class of functions which turns out to be the missing ingredient to unify the problems that we are interested in.

8.1.3 Laplace-Stieltjes Functions

Definition 8.13. *A function $f : \mathbb{R}^+ \to \mathbb{R}$ is said to be a Laplace-Stieltjes function if there exists a nonnegative real-valued function μ_L such that*

$$f(\lambda) = \int_0^\infty \mu_L(\zeta) e^{-\zeta\lambda}\, d\zeta. \tag{8.15}$$

The function μ_L is called Laplace-Stieltjes density of f. We denote the set of all Laplace-Stieltjes functions by $\mathcal{LS}$.

More succinctly, (8.15) can be written as $f(\lambda) = \mathscr{L}[\mu_L](\lambda)$ whence there is a one-to-one correspondence between the function f and its Laplace-Stieltjes density μ_L [Wid43]. Well-known examples of Laplace-Stieltjes functions include

$$\frac{1-e^{-\lambda}}{\lambda} = \int_0^1 e^{-\zeta\lambda}\,d\zeta, \qquad \frac{1}{\lambda+c} = \int_0^\infty e^{-c\zeta}e^{-\zeta\lambda}\,d\zeta, \quad c \in \mathbb{R}_0^+. \tag{8.16}$$

Given some $f \in \mathcal{LS}$, the respective Laplace-Stieltjes density is obtained by applying the inverse Laplace transform, that is, $\mu_L = \mathscr{L}^{-1}[f]$. For a large class of functions, the latter is known explicitly. To apply the results provided in the sequel, however, it suffices to know whether a function is contained in $\mathcal{LS}$ without the explicit knowledge of μ_L. A convenient tool for this purpose is the well-known characterization of $\mathcal{LS}$ as the set of completely monotonic functions.

Definition 8.14. *A function $f \in C^\infty(\mathbb{R}^+)$ is said to be completely monotonic if*

$$(-1)^n f^{(n)}(\lambda) \geq 0, \qquad \lambda \in \mathbb{R}^+,$$

for all $n \in \mathbb{N}_0$. We denote the set of all completely monotonic functions with $\mathcal{CM}$.

As shown in [Ber29], the set $\mathcal{CM}$ coincides with $\mathcal{LS}$.

Theorem 8.15 (Bernstein)**.** *If $f \in \mathcal{CM}$, then f is the Laplace transform of a unique function $\mu_L : \mathbb{R}^+ \to \mathbb{R}_0^+$ such that*

$$f(\lambda) = \int_0^\infty \mu_L(\zeta)e^{-\zeta\lambda}\,d\zeta. \tag{8.17}$$

Conversely, whenever $\mu_L : \mathbb{R}^+ \to \mathbb{R}_0^+$ is a function with the property $\mathscr{L}[\mu_L](\lambda) < \infty$ for all $\lambda \in \mathbb{R}^+$, then $\lambda \mapsto \mathscr{L}[\mu_L](\lambda)$ is completely monotonic. In particular, we have $\mathcal{LS} = \mathcal{CM}$.

As a direct consequence, we find that any $f \in \mathcal{LS}$ is a nonnegative, decreasing, and convex function. As such, the limit

$$f(0^+) := \lim_{\lambda \to 0^+} f(\lambda) \tag{8.18}$$

exists. The function $f(\lambda) = \lambda^{-1} \in \mathcal{LS}$ shows, however, that the value of $f(0^+)$ is not necessarily finite. Another well-known property that follows from (8.17) is the following result, see [Wid43].

Lemma 8.16. *Any $f \in \mathcal{LS}$ allows for an analytic continuation to the right half plane $\{z \in \mathbb{C} : \Re z > 0\}$.*

To establish a connection between $\mathcal{LS}$ and the function classes introduced in the previous sections, we apply the second identity in (8.16) and Fubini's theorem to conclude for all $f \in \mathcal{CS}$

$$f(\lambda) = \int_0^\infty \frac{\mu_C(\zeta)}{\lambda+\zeta}\,d\zeta = \int_0^\infty \int_0^\infty \mu_C(\zeta)e^{-\zeta s}e^{-s\lambda}\,ds\,d\zeta = \int_0^\infty e^{-s\lambda}\int_0^\infty \mu_C(\zeta)e^{-\zeta s}\,d\zeta\,ds.$$

This proves the following result.

Lemma 8.17. *There holds $\mathcal{CS} \subset \mathcal{LS}$. Moreover, for any function $f \in \mathcal{CS}$ with Cauchy-Stieltjes density μ_C the Laplace-Stieltjes density μ_L of f is given by*

$$\mu_L(\zeta) = \int_0^\infty \mu_C(s) e^{-\zeta s}\, ds.$$

Note that the inclusion is indeed strict since for $f(\lambda) = e^{-\lambda}$ there holds

$$(-1)^n f^{(n)}(\lambda) = (-1)^n (-1)^n e^{-\lambda} = e^{-\lambda} \geq 0$$

for all $\lambda \in \mathbb{R}^+$. Hence, by Bernstein's theorem, we find $e^{-\lambda} \in \mathcal{LS}$ but $e^{-\lambda} \notin \mathcal{CS}$.

Remark 8.18. *Due to $\mathcal{CS} \subset \mathcal{LS}$ any Cauchy-Stieltjes function is completely monotonic. The connection between $\mathcal{LS} = \mathcal{CM}$ and $\mathcal{CB}$ is that any $f \in \mathcal{CB}$ satisfies $f' \in \mathcal{CM}$, see [SSV12, Theorem 3.2]. However, not every function with this property is a complete Bernstein function. The collection of all functions $f : \mathbb{R}^+ \to \mathbb{R}$ that satisfy $f' \in \mathcal{CM}$ is the set of Bernstein functions. This is the superset of $\mathcal{CB}$ whose elements can be written as*

$$f(\lambda) = \omega + \theta\lambda + \int_0^\infty \mu(\zeta)(1 - e^{-\zeta\lambda})\, d\zeta \tag{8.19}$$

for $\omega, \theta \in \mathbb{R}_0^+$ and some nonnegative real-valued function μ such that the integral is absolutely convergent. There holds $f \in \mathcal{CB}$ if and only if μ in (8.19) is completely monotonic.

A few remarkable properties of Laplace-Stieltjes functions are listed in the following lemma.

Lemma 8.19.

1. *The set $\mathcal{LS}$ is a convex cone: $af + bg \in \mathcal{LS}$ for all $a, b \in \mathbb{R}_0^+$ and $f, g \in \mathcal{LS}$.*
2. *The set $\mathcal{LS}$ is closed under pointwise convergence: if $(f_n)_{n\in\mathbb{N}} \subset \mathcal{LS}$ converges pointwise to some f, then $f \in \mathcal{LS}$.*
3. *The set $\mathcal{LS}$ is closed under multiplication: $fg \in \mathcal{LS}$ for all $f, g \in \mathcal{LS}$.*
4. *There holds $\mathcal{LS} \circ \mathcal{CB} \subset \mathcal{LS}$.*
5. *If $f \in \mathcal{LS}$ and g is positive on $\mathbb{R}^+$, then $f \circ g \in \mathcal{LS}$. In particular, $f(c\lambda) \in \mathcal{LS}$ for any $c \in \mathbb{R}^+$.*
6. *If $f : \mathbb{R}^+ \to \mathbb{R}_0^+$ with $\ln \circ f \in \mathcal{LS}$, then $f \in \mathcal{LS}$.*

Proof. See Corollary 1.6 and Theorem 3.6 in [SSV12] for the first four assertions and [Mer12, Lemma 3.4] for the latter. □

Provided these tools, our goal is to prove that functions arising from time-dependent fractional diffusion problems have Laplace-Stieltjes membership, i.e.,

$$f(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s) \in \mathcal{LS} \tag{8.20}$$

for all $\alpha \in (0,1]$, $\beta \geq \alpha$, $t \in \mathbb{R}_0^+$, and $s \in [0,1]$. Recalling Remark 2.26, we also want to include the extremal case $\alpha = 0$, whence it is fruitful to define

$$\mathrm{e}_{\alpha,\beta}(t,\lambda) := \begin{cases} E_{\alpha,\beta}(t\lambda), & \text{if } \alpha > 0, \\ \frac{1}{\Gamma(\beta)}(1+\lambda)^{-1}, & \text{if } \alpha = 0. \end{cases} \tag{8.21}$$

Theorem 8.20. *If $(\alpha,\beta,t,s) \in \Theta_{\mathrm{L}} := \{(\alpha,\beta,t,s) \in [0,1]\times\mathbb{R}^+\times\mathbb{R}_0^+\times[0,1] : \beta \geq \alpha\}$, then there holds $e_{\alpha,\beta}(-t^\alpha,\lambda^s) \in \mathcal{LS}$.*

Since $(1+\lambda)^{-1} \in \mathcal{CM}$, it follows from Theorem 8.15 that the proof holds if $\alpha = 0$. To confirm the conjecture for $\alpha \in (0,1]$ we require the following intermediate result.

Lemma 8.21. *Let $\alpha,\beta \in \mathbb{R}^+$. Then $E_{\alpha,\beta}(-\lambda) \in \mathcal{LS}$ if and only if $\alpha \in (0,1]$ and $\beta \geq \alpha$.*

Proof. See [Sch96] for the explicit derivation of the Laplace-Stieltjes density and [Mil99] for a shorter proof validating only the membership itself. □

Proof of Theorem 8.20. Under the given assumptions on α and β, the fifth property in Lemma 8.19 reveals that $E_{\alpha,\beta}(-t^\alpha\lambda) \in \mathcal{LS}$ for any $t \in \mathbb{R}_0^+$. By the fourth property in Lemma 8.19, the composition of the latter with $\lambda^s \in \mathcal{CB}$, $s \in [0,1]$, is a Laplace-Stieltjes function itself and Theorem 8.20 is proved. □

For some values of the parameters, the results of Theorem 8.20 can be confined. For this purpose, let us consider the integer-order case $\alpha = 1$. In analogy to (8.14), we find that for all $z = re^{i\phi}$

$$e^{-z^s} = e^{-r^s e^{i\phi s}} = e^{-r^s\cos(\phi s)}e^{-ir^s\sin(\phi s)}. \tag{8.22}$$

Provided that $s \in (0,\frac{1}{2})$ and $\phi \in (-\pi,\pi)$, (8.22) converges uniformly to zero as $r \to \infty$. Furthermore, $ze^{-z} \to 0$ as $z \to 0$. Applying Lemma 8.2, we find that $e^{-\lambda^s} \in \mathcal{CS}$ if $s \in (0,\frac{1}{2})$ and its Cauchy-Stieltjes density evaluates to

$$\begin{aligned} \mu_C(\zeta) &= \frac{1}{2\pi i}\lim_{\phi\to\pi}\left(e^{-\zeta^s e^{i\phi s}} - e^{-\zeta^s e^{-i\phi s}}\right) \\ &= \frac{1}{2\pi i}\left(e^{-\zeta^s e^{i\pi s}} - e^{-\zeta^s e^{-i\pi s}}\right) = \frac{1}{\pi}e^{-\zeta^s\cos(\pi s)}\sin(\zeta^s\sin(\pi s)). \end{aligned}$$

For arbitrary $\alpha \in [0,1]$, the situation can be generalized as follows (cf. [YTLI11, Proposition 4.4] and [MN18, Example 4.3]).

Proposition 8.22. *If $(\alpha,\beta,t,s) \in \Theta_{\mathrm{C}} := \{(\alpha,\beta,t,s) \in \Theta_{\mathrm{L}} : \frac{\alpha}{2}+s < 1\}$, then $e_{\alpha,\beta}(-t^\alpha,\lambda^s) \in \mathcal{CS}$ and the Cauchy-Stieltjes density is given by*

$$\mu_C(\zeta) = -\frac{1}{\pi}\Im\left(E_{\alpha,\beta}(-t^\alpha\zeta^s e^{i\pi s})\right).$$

Proof. The cases where either of the parameters α, t, or s are zero follow from the fact that constant functions as well as $(1+\lambda^s)^{-1}$, $s \in [0,1]$, are contained in $\mathcal{CS}$. For the remainder of this proof, let us therefore assume that α, t, and s are strictly positive such

that $\mathrm{e}_{\alpha,\beta}(-t^\alpha,\lambda^s) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$. We check under which restrictions on the parameters the requirements (8.7) are satisfied such that Lemma 8.2 can be applied. Clearly, the second property in (8.7) holds for all admissible values of α, β, t, and s. To investigate the behaviour of $|f(z)|$ as $z \to \infty$, we consult Theorem 2.29 to infer

$$\arg(-t^\alpha z^s) \in \left(\frac{\alpha\pi}{2}, \pi\right) \quad \Longrightarrow \quad |E_{\alpha,\beta}(-t^\alpha z^s)| \leq \frac{c_{\alpha,\beta}}{1+t^\alpha|z|^s}. \tag{8.23}$$

Utilizing polar coordinates $z = re^{i\phi}$, the left-hand side of (8.23) reads

$$\arg(-t^\alpha z^s) = \arg(-e^{i\phi s}) = \pi - \phi s \in \left(\frac{\alpha\pi}{2}, \pi\right).$$

In terms of ϕ, the condition can be reformulated as $\phi \in (0, \frac{(2-\alpha)\pi}{2s})$. To guarantee

$$\lim_{|z|\to\infty} |E_{\alpha,\beta}(-t^\alpha z^s)| = 0$$

uniformly in $\phi \in (0,\pi)$, we thus require

$$\frac{(2-\alpha)\pi}{2s} > \pi,$$

or equivalently, $\frac{\alpha}{2} + s < 1$. Since $|E_{\alpha,\beta}(-t^\alpha\lambda^s)| \to 0$ for increasing values of $\lambda \in \mathbb{R}$ and $|E_{\alpha,\beta}(-t^\alpha\overline{z}^s)| = |\overline{E_{\alpha,\beta}(-t^\alpha z^s)}| = |E_{\alpha,\beta}(-t^\alpha z^s)|$, we deduce that $|E_{\alpha,\beta}(-t^\alpha z^s)|$ converges uniformly to 0 as $|z| \to \infty$ if $\arg(z) \in (-\pi,\pi)$ and $\frac{\alpha}{2} + s < 1$. Thanks to Lemma 8.2, we deduce $E_{\alpha,\beta}(-t^\alpha\lambda^s) \in \mathcal{CS}$ with Cauchy-Stieltjes density

$$\begin{aligned}\mu_C(\zeta) &= \frac{i}{2\pi}\left(E_{\alpha,\beta}(-t^\alpha\zeta^s e^{i\pi s}) - E_{\alpha,\beta}(-t^\alpha\zeta^s e^{-i\pi s})\right)\\ &= \frac{i}{2\pi}\sum_{j=0}^{\infty}\frac{(-t^\alpha\zeta^s)^k\cdot 2i\sin(j\pi s)}{\Gamma(\alpha j+\beta)} = -\frac{1}{\pi}\Im\sum_{j=0}^{\infty}\frac{\left(-t^\alpha\zeta^s e^{i\pi s}\right)^j}{\Gamma(\alpha j+\beta)} = -\frac{1}{\pi}\Im E_{\alpha,\beta}(-t^\alpha\zeta^s e^{i\pi s}).\end{aligned}$$

□

For the reader's convenience, we conclude this section with an illustration of the systematic classification of the fractional power function and $f(\lambda) = \mathrm{e}_{\alpha,\beta}(-t^\alpha,\lambda^s)$ in Figure 8.3. For $f(\lambda) = \lambda^s$, we see that in the extremal case $s = 0$ the function is in the intersection of $\mathcal{CS}$ and $\mathcal{CB}$. For $f^\tau(\lambda) = \mathrm{e}_{\alpha,\beta}(-t^\alpha,\lambda^s)$ the critical case $\frac{\alpha}{2} + s = 1$ only guarantees membership in the Laplace-Stieltjes set.

8.2 Approximability of the Matrix Kernels

The previous section allows us to take the unified point of view that the DEM approximation of fractional diffusion problems is obtained by evaluating a matrix-vector product of the form $f^\tau(\mathbf{L})\mathbf{b}$, where f^τ is either of Stieltjes or complete Bernstein type. The computations in (8.2) show that the approximability of these functions hinges on the approximability of

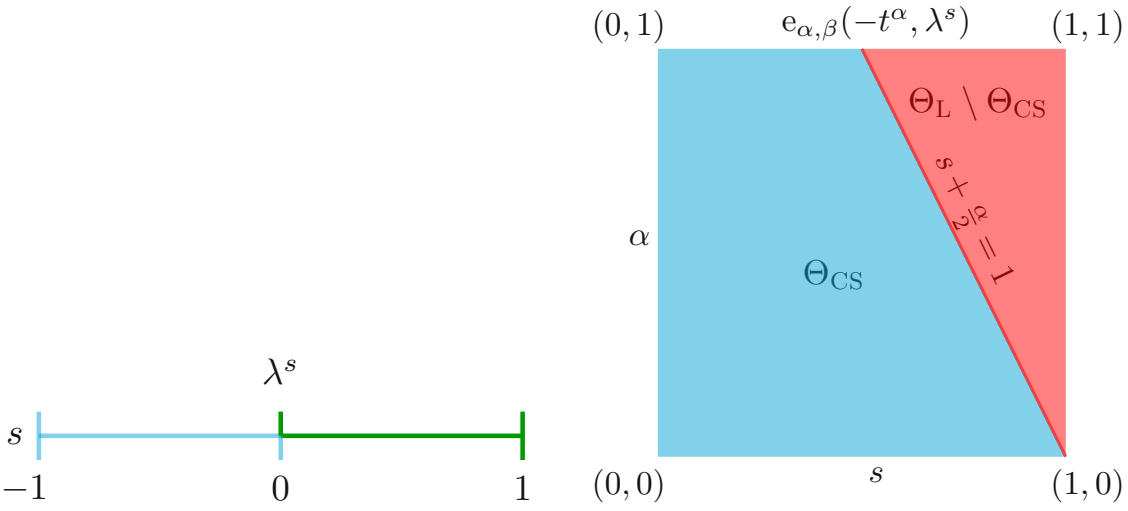

Figure 8.3: Classification of the fractional power function λ^s, $s \in [-1,1]$, and the Mittag-Leffler type function $\mathrm{e}_{\alpha,\beta}(-t^\alpha,\lambda^s)$ for $(\alpha,s) \in [0,1]^2$, fixed $t > 0$, and $\beta \geq \alpha$. The cyan, green, and red areas indicate Cauchy-Stieltjes, complete Bernstein, and Laplace-Stieltjes membership of the respective function.

the respective matrix kernels. Recognizing this fact, the main purpose of this section is to provide an upper bound for the rational Krylov approximation error in the form of

$$\|g(\mathbf{L},\zeta) - \mathbf{V}g(\mathbf{L}_{k+1},\zeta)\mathbf{V}^\dagger\mathbf{b}\| \leq \eta_g(\Xi,\zeta), \tag{8.24}$$

where $\mathbf{V}$ is an orthonormal basis of $\mathcal{Q}^\Xi_{k+1}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$ its compression, $g(\lambda,\zeta) \in \{(\lambda+\zeta)^{-1}, \lambda/(\lambda+\zeta), e^{-\zeta\lambda}\}$, and $\eta_g(\Xi,\zeta)$ an upper bound depending on Ξ and ζ only. We start our investigation with the Cauchy-Stieltjes kernel.

8.2.1 The Resolvent Kernel

Let $\mathbf{V}$ be an orthonormal basis of the rational Krylov space $\mathcal{Q}^\Xi_{k+1}(\mathbf{L},\mathbf{b})$ and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$. Due to Theorem 7.16, we may bound the resolvent error by

$$\|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| \leq 2\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_\Xi} \|(\lambda+\zeta)^{-1} - r_k(\lambda)\|_\Sigma \tag{8.25}$$

for any $\zeta \in \mathbb{R}^+_0$, where Σ is the spectral interval of $\mathbf{L}$. To bound the right-hand side of (8.25), we introduce, recalling (7.1), the following rational interpolant of $(\lambda+\zeta)^{-1}$ which is an integral part of our analysis; cf. [Güt10, Section 7.5.2].

Definition 8.23. *Provided a set of nodes $\Lambda = \{\sigma_0,\dots,\sigma_l\} \subset \overline{\mathbb{C}}$, $l \in \mathbb{N}$, we define for any $\zeta \in \mathbb{C}$ the rational interpolant $r^\zeta_{\Lambda,\Xi}$ by*

$$r^\zeta_{\Lambda,\Xi}(\lambda) := \frac{1 - \frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)}}{\lambda+\zeta}, \qquad r_{\Lambda,\Xi}(\lambda) := \frac{q_\Lambda(\lambda)}{q_\Xi(\lambda)} \in \mathcal{R}_{l+1,k+1}. \tag{8.26}$$

Recalling $\deg(q_\Xi) = |\{\xi \in \Xi : \xi \neq \infty\}|$ for any $\Xi \subset \overline{\mathbb{C}}$, we collect in the following Lemma several properties of the rational interpolant, which, among others, justify its nomenclature; cf. [Güt10, DHS21].

Lemma 8.24. *Let $\Lambda = \{\sigma_0, \dots, \sigma_l\} \subset \overline{\mathbb{C}}$ and $\zeta \in \mathbb{C}$.*

1. *Let $n = \deg(q_\Lambda)$, $m = \deg(q_\Xi)$, and $j = \max\{n, m\}$. Then $r^\zeta_{\Lambda,\Xi} \in \mathcal{P}_{j-1}/q_\Xi$.*
2. *$r^\zeta_{\Lambda,\Xi}$ interpolates the resolvent in $\Lambda_{\text{fin}} := \{\sigma \in \Lambda : \sigma \neq \infty\}$, i.e.,*
$$\forall \sigma \in \Lambda_{\text{fin}} : \quad r^\zeta_{\Lambda,\Xi}(\sigma) = \frac{1}{\sigma + \zeta}.$$
3. *The absolute error is given by*
$$\frac{1}{\lambda + \zeta} - r^\zeta_{\Lambda,\Xi}(\lambda) = \frac{1}{\lambda + \zeta} \frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)}. \tag{8.27}$$

Proof. Let $c = r_{\Lambda,\Xi}(-\zeta)$, then there holds
$$r^\zeta_{\Lambda,\Xi}(\lambda) = \frac{\frac{cq_\Xi(\lambda) - q_\Lambda(\lambda)}{cq_\Xi(\lambda)}}{\lambda + \zeta} = \frac{cq_\Xi(\lambda) - q_\Lambda(\lambda)}{cq_\Xi(\lambda)(\lambda + \zeta)} \in \mathcal{R}_{j,m+1}.$$
The original definition of $r_{\Lambda,\Xi}$ in (8.26) shows, however, that $-\zeta$ is a root of both the denominator and the numerator polynomial, whence in fact $r^\zeta_{\Lambda,\Xi} \in \mathcal{P}_{j-1}/q_\Xi$ as claimed. The second property follows from $r_{\Lambda,\Xi}(\sigma) = 0$ for all $\sigma \in \Lambda_{\text{fin}}$. The third one holds since
$$\frac{1}{\lambda + \zeta} - r^\zeta_{\Lambda,\Xi}(\lambda) = \frac{\frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)}}{\lambda + \zeta} = \frac{1}{\lambda + \zeta} \frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)}.$$
□

Since $\deg(q_\Xi) \leq k + 1$, it follows from the first property of the previous lemma that $r^\zeta_{\Lambda,\Xi} \in \mathcal{P}_k/q_\Xi$ if $|\Lambda| \leq k + 1$. In this case, the rational interpolant is an admissible choice to bound the right-hand side of (8.25). Together with the third property in Lemma 8.24, we deduce
$$\begin{aligned} \|(\mathbf{L} + \zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1} + \zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| &\leq 2\|\mathbf{b}\|\|(\lambda + \zeta)^{-1} - r^\zeta_{\Lambda,\Xi}(\lambda)\|_\Sigma \\ &\leq 2\|\mathbf{b}\|\|(\lambda + \zeta)^{-1}\|_\Sigma \frac{\|r_{\Lambda,\Xi}(\lambda)\|_\Sigma}{|r_{\Lambda,\Xi}(-\zeta)|}. \end{aligned} \tag{8.28}$$
These observations allow us to derive an upper bound of the form (8.24) when $g(\mathbf{L}, \zeta)$ is the resolvent function; cf. [MR20a, DHS21].

Theorem 8.25. *Let $\zeta \in \mathbb{C}$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}^\Xi_{k+1}(\mathbf{L}, \mathbf{b})$, and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$. Then there holds*
$$\|(\mathbf{L} + \zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1} + \zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| \leq 2\|\mathbf{b}\|\|(\lambda + \zeta)^{-1}\|_\Sigma \min_{r_{k+1} \in \mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{|r_{k+1}(-\zeta)|}.$$

Proof. This follows from (8.28) after taking the minimum over all Λ with $|\Lambda| \leq k + 1$. □

The proof of Theorem 8.25 holds without any restrictions on $\Xi \subset \mathbb{R} \setminus \Sigma$. As shown below, the situation is different in the complete Bernstein case, where we require $\infty \in \Xi$ to bound the rational Krylov approximation error.

8.2.2 The Complete Bernstein Kernel

The complete Bernstein kernel $g(\mathbf{L},\zeta) = \mathbf{L}(\mathbf{L}+\zeta\mathbf{I})^{-1}$ bears a close relation to the Cauchy-Steiltjes kernel. It is thus not surprising that the derivation of its upper bound follows in a similar fashion. A subtle modification of Definition 8.23 reads as follows.

Definition 8.26. *Provided a set of nodes $\Lambda = \{\sigma_0,\dots,\sigma_l\} \subset \overline{\mathbb{C}}$, $l \in \mathbb{N}$, we define for any $\zeta \in \mathbb{C}$ the rational interpolant $\hat{r}^{\zeta}_{\Lambda,\Xi}$ by*

$$\hat{r}^{\zeta}_{\Lambda,\Xi}(\lambda) := \lambda\, r^{\zeta}_{\Lambda,\Xi}(\lambda).$$

The following observations are a direct consequence of Lemma 8.24.

Lemma 8.27. *Let $\Lambda = \{\sigma_0,\dots,\sigma_l\} \subset \overline{\mathbb{C}}$ and $\zeta \in \mathbb{C}$.*

1. *Let $n = \deg(q_\Lambda)$, $m = \deg(q_\Xi)$, and $j = \max\{n,m\}$. Then $\hat{r}^{\zeta}_{\Lambda,\Xi} \in \mathcal{P}_j/q_\Xi$.*
2. *$\hat{r}^{\zeta}_{\Lambda,\Xi}$ satisfies the interpolation property*

$$\hat{r}^{\zeta}_{\Lambda,\Xi}(\sigma) = \frac{\sigma}{\sigma+\zeta}$$

   for all $\sigma \in \Lambda_{\mathrm{fin}} = \{\sigma \in \Lambda : \sigma \neq \infty\}$.
3. *The absolute error is given by*

$$\frac{\lambda}{\lambda+\zeta} - \hat{r}^{\zeta}_{\Lambda,\Xi}(\lambda) = \frac{\lambda}{\lambda+\zeta}\frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)}.$$

The main result of this section is stated in the following theorem; cf. [DHS21, Lemma 4].

Theorem 8.28. *Let $\zeta \in \mathbb{C}$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$, and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$. Assume that $\infty \in \Xi$. Then there holds*

$$\|\mathbf{L}(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}\mathbf{L}_{k+1}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| \le 2\|\mathbf{b}\|\|\lambda/(\lambda+\zeta)\|_\Sigma \min_{r_k \in \mathcal{P}_k/q_\Xi} \frac{\|r_k\|_\Sigma}{|r_k(-\zeta)|}.$$

Proof. We apply Theorem 7.16 to observe

$$\|\mathbf{L}(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}\mathbf{L}_{k+1}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| \le 2\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_\Xi} \|\lambda/(\lambda+\zeta) - r_k(\lambda)\|_\Sigma.$$

Since $\infty \in \Xi$, there holds $\deg(q_\Xi) \le k$. The first property in Lemma 8.27 reveals that $\hat{r}^{\zeta}_{\Lambda,\Xi} \in \mathcal{P}_k/q_\Xi$ is an admissible choice to bound the minimum above if $|\Lambda| \le k$. If this is the case, we may consult the third property in Lemma 8.27 to deduce

$$\begin{aligned}\|\mathbf{L}(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}\mathbf{L}_{k+1}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| &\le 2\|\mathbf{b}\|\|\lambda/(\lambda+\zeta) - \hat{r}^{\zeta}_{\Lambda,\Xi}(\lambda)\|_\Sigma \\ &\le 2\|\mathbf{b}\|\|\lambda/(\lambda+\zeta)\|_\Sigma \frac{\|r_{\Lambda,\Xi}(\lambda)\|_\Sigma}{|r_{\Lambda,\Xi}(-\zeta)|}.\end{aligned}$$

Since Λ is arbitrary, we may take the minimum over all Λ with $|\Lambda| \le k$ to confirm that the claim is valid. □

8.2.3 The Exponential Kernel

The treatment of the exponential kernel $g(\mathbf{L},\zeta) = e^{-\zeta\mathbf{L}}$ is more delicate. In line with [DKZ09, MR20a, DHS21], we apply the second identity in (8.16) to rewrite $e^{-\zeta\lambda}$ via inverse Laplace transform

$$e^{-\zeta\lambda} = \mathscr{L}^{-1}[(\lambda+\cdot)^{-1}](\zeta) = \frac{1}{2\pi i}\int_{i\mathbb{R}} \frac{e^{\zeta z}}{\lambda+z}\,dz, \tag{8.29}$$

where $i\mathbb{R}$ denotes the integration path starting at $-i\infty$ and ending in $i\infty$. Given an orthonormal basis $\mathbf{V}$ of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L},\mathbf{b})$ and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, we apply Theorem 2.36, after the transformation $z \mapsto -z$, to deduce

$$e^{-\zeta\mathbf{L}}\mathbf{b} - \mathbf{V}e^{-\zeta\mathbf{L}_{k+1}}\mathbf{V}^\dagger\mathbf{b} = \frac{1}{2\pi i}\int_{i\mathbb{R}} e^{\zeta z}\left((\mathbf{L}+z\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1}+z\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\right)dz. \tag{8.30}$$

These computations show that the approximability of the exponential kernel $e^{-\zeta\mathbf{L}}$, $\zeta\in\mathbb{R}^+_0$, is closely related to the approximability of the resolvent $(\mathbf{L}+z\mathbf{I})^{-1}$ with $z\in i\mathbb{R}$. An intuitive approach to bound (8.30) would be to apply Theorem 8.25 so that

$$\begin{aligned}\|e^{-\zeta\mathbf{L}}\mathbf{b} - \mathbf{V}e^{-\zeta\mathbf{L}_{k+1}}\mathbf{V}^\dagger\mathbf{b}\| &\le \frac{1}{2\pi}\int_{i\mathbb{R}} \|(\lambda+z)^{-1}\|_\Sigma \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{|r_{k+1}(-z)|}\,dz\\ &\le \frac{1}{2\pi}\int_{i\mathbb{R}} \frac{1}{|\lambda_{\min}+z|}\,dz \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(z)| : z\in i\mathbb{R}\}},\end{aligned}$$

which bears a close resemblance to the estimates derived in Theorem 8.25 and 8.28. Unfortunately, the upper bound so obtained is not meaningful since

$$\int_{i\mathbb{R}} \frac{1}{|\lambda_{\min}+z|}\,dz = \infty.$$

More careful computations allow one to overcome this difficulty. For convenience, we introduce the following definition.

Definition 8.29. *For all $\Xi = \{\xi_0,\dots,\xi_k\} \subset \overline{\mathbb{C}}$ we define the rational function $r_\Xi \in \mathcal{R}_{k+1,k+1}$ by*

$$r_\Xi(z) := r_{-\Xi,\Xi}(z) = \prod_{\substack{j=0\\ \xi_j\neq\infty}}^{k} \frac{z+\xi_j}{z-\xi_j}.$$

The following technical lemma can be found in the proof of [MR20a, Theorem 2] and is instrumental for our approach.

Lemma 8.30. *Let $\Xi = \{\xi_0,\dots,\xi_k\} \subset -\Sigma\cup\{\infty\}$, $m = \deg(q_\Xi)$, and*

$$h(\lambda,\zeta) := \frac{1}{2\pi i}\int_{i\mathbb{R}} e^{\zeta z}(\lambda+z)^{-1} r_\Xi(\lambda) r_\Xi(-z)^{-1}\,dz.$$

Then there holds for all $\lambda\in\Sigma$ and $\zeta\in\mathbb{R}^+_0$

$$|h(\lambda,\zeta)| \le 2\gamma_m |r_\Xi(\lambda)|, \qquad \gamma_m := 2.23 + \frac{2}{\pi}\ln\left(4m\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}\pi}}\right).$$

If we restrict all finite poles to the negative spectral interval, we can bound the rational Krylov approximation error of the matrix kernel $g(\mathbf{L}, \zeta) = e^{-\zeta \mathbf{L}}$ as follows; see [MR20a, Theorem 2] for the case where $\mathbf{L}$ is symmetric.

Theorem 8.31. *Let $\zeta \in \mathbb{R}_0^+$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ with poles $\Xi \subset -\Sigma \cup \{\infty\}$, $m = \deg(q_\Xi)$, γ_m as in Lemma 8.30, and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$. Then there holds*

$$\|e^{-\zeta \mathbf{L}}\mathbf{b} - \mathbf{V} e^{-\zeta \mathbf{L}_{k+1}} \mathbf{V}^\dagger \mathbf{b}\| \leq 4\gamma_m \|\mathbf{b}\| \|r_\Xi\|_\Sigma. \tag{8.31}$$

Proof. Due to the first property in Lemma 8.24, there holds $r^\zeta_{-\Xi,\Xi} \in \mathcal{P}_k/q_\Xi$. By Lemma 7.12, we find that $r^\zeta_{-\Xi,\Xi}(\mathbf{L})\mathbf{b} = \mathbf{V} r^\zeta_{-\Xi,\Xi}(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}$. Subtracting $e^{\zeta z} r^\zeta_{-\Xi,\Xi}(\mathbf{L})\mathbf{b}$ and adding $e^{\zeta z}\mathbf{V} r^\zeta_{-\Xi,\Xi}(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}$ inside the integral of (8.30) combined with (8.27) reveals

$$\begin{aligned} e^{-\zeta \mathbf{L}}\mathbf{b} - \mathbf{V} e^{-\zeta \mathbf{L}_{k+1}} \mathbf{V}^\dagger \mathbf{b} &= \frac{1}{2\pi i} \int_{i\mathbb{R}} e^{\zeta z} (\mathbf{L} + z\mathbf{I})^{-1} r_\Xi(\mathbf{L}) r_\Xi(-z)^{-1} \mathbf{b}\, dz \\ &\quad - \frac{1}{2\pi i} \int_{i\mathbb{R}} e^{\zeta z} \mathbf{V} (\mathbf{L}_{k+1} + z\mathbf{I}_{k+1})^{-1} r_\Xi(\mathbf{L}_{k+1}) r_\Xi(-z)^{-1} \mathbf{V}^\dagger \mathbf{b}\, dz \\ &= h(\mathbf{L}, \zeta)\mathbf{b} - \mathbf{V} h(\mathbf{L}_{k+1}, \zeta) \mathbf{V}^\dagger \mathbf{b}, \end{aligned}$$

with $h(\lambda, \zeta)$ as in Lemma 8.30. We apply Theorem 7.16 to find that

$$\|e^{-\zeta \mathbf{L}}\mathbf{b} - \mathbf{V} e^{-\zeta \mathbf{L}_{k+1}} \mathbf{V}^\dagger \mathbf{b}\| \leq 2\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_\Xi} \|h(\lambda, \zeta) - r_k(\lambda)\|_\Sigma \leq 2\|\mathbf{b}\| \|h(\lambda, \zeta)\|_\Sigma.$$

The inequality (8.31) now follows directly from Lemma 8.30. □

8.3 Approximability of Stieltjes and Complete Bernstein Functions

In this section, we leverage our insights gained to bound the rational Krylov approximation error for Stieltjes and complete Bernstein functions. Recalling (8.18), the central statement of this chapter is stated in the theorem below.

Theorem 8.32. *Let $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V} f(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$.*

1. *Let $f \in \mathcal{CS}$ with ω and θ as in (8.4). If $\theta = 0$, then*

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2f(\lambda_{\min})\|\mathbf{b}\| \min_{r_{k+1} \in \mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(\lambda)| : \lambda \in \mathbb{R}_0^-\}}. \tag{8.32}$$

   Assuming $\infty \in \Xi$, then (8.32) holds even if $\theta \neq 0$.

2. *Let $f \in \mathcal{CB}$ with ω and θ as in (8.12) and assume $\infty \in \Xi$. If $\omega = 0$, then there holds*

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2f(\lambda_{\max})\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_\Xi} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)| : \lambda \in \mathbb{R}_0^-\}}. \tag{8.33}$$

   Assuming $\{\infty, \infty\} \subset \Xi$, then (8.33) holds even if $\omega \neq 0$.

3. Assume $\Xi \subset -\Sigma \cup \{\infty\}$, $m = \deg(q_\Xi)$, and γ_m as in Lemma 8.30. If $f \in \mathcal{LS}$, then

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 4\gamma_m f(0^+)\|\mathbf{b}\|\|r_\Xi\|_\Sigma. \tag{8.34}$$

Proof. Let $f \in \mathcal{CS}$ and (ω, θ, μ_C) its Cauchy-Stieltjes triple. If $\theta = 0$, then

$$\begin{aligned} f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1} = \omega \left(\mathbf{L}^{-1}\mathbf{b} - \mathbf{V}\mathbf{L}_{k+1}^{-1}\mathbf{V}^\dagger\mathbf{b}\right) \\ + \int_0^\infty \mu_C(\zeta)\left((\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\right) d\zeta. \end{aligned}$$

We apply triangle inequality to deduce

$$\begin{aligned} \|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq \omega\|\mathbf{L}^{-1}\mathbf{b} - \mathbf{V}\mathbf{L}_{k+1}^{-1}\mathbf{V}^\dagger\mathbf{b}\| \\ + \int_0^\infty \mu_C(\zeta)\|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| d\zeta. \end{aligned}$$

Invoking Theorem 8.25 with $\zeta = 0$ and $\zeta \in \mathbb{R}^+$, respectively, we arrive at

$$\begin{aligned} \|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| &\leq 2\left(\frac{\omega}{\lambda_{\min}} + \int_0^\infty \frac{\mu_C(\zeta)}{\lambda_{\min}+\zeta} d\zeta\right)\|\mathbf{b}\| \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \\ &= 2f(\lambda_{\min})\|\mathbf{b}\| \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(\lambda)| : \lambda \in \mathbb{R}_0^-\}}. \end{aligned}$$

Clearly, if $\infty \in \Xi$, then $\mathbf{b} \in \mathcal{Q}_{k+1}^\Xi(\mathbf{L}, \mathbf{b})$ so that any additional contribution in the form of $\theta\mathbf{b} = \theta\mathbf{V}\mathbf{V}^\dagger\mathbf{b}$ is computed exactly.

To prove the second claim, let $f \in \mathcal{CB}$ and (ω, θ, μ_B) its complete Bernstein triple. Since $\infty \in \Xi$ and $\omega = 0$, it follows from Theorem 8.28

$$\begin{aligned} \|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| &\leq \int_0^\infty \mu_B(\zeta)\|\mathbf{L}(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}\mathbf{L}_{k+1}(\zeta\mathbf{I}_{k+1}+\mathbf{L}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\| d\zeta \\ &\leq 2\|\mathbf{b}\| \int_0^\infty \mu_B(\zeta)\frac{\lambda_{\max}}{\zeta+\lambda_{\max}} d\zeta \min_{r_k\in\mathcal{P}_k/q_\Xi} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \\ &= 2\left(f(\lambda_{\max}) - \theta\right)\|\mathbf{b}\| \min_{r_k\in\mathcal{P}_k/q_\Xi} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)| : \lambda \in \mathbb{R}_0^-\}}, \end{aligned}$$

which directly implies (8.33). Assume now $\{\infty, \infty\} \subset \Xi$. Noting that $q_\Xi \in \mathcal{P}_{k-1}$, it follows from the second property in Lemma 7.5 that $\mathbf{L}\mathbf{b} = r_k(\mathbf{L})\mathbf{b}$, $r_k(\lambda) = \lambda q_\Xi(\lambda)/q_\Xi(\lambda) \in \mathcal{P}_k/q_\Xi$, is contained in $\mathcal{Q}_{k+1}^\Xi(\mathbf{L}, \mathbf{b})$ so that any additional contribution in the form of $\omega\mathbf{L}\mathbf{b} = \omega\mathbf{V}\mathbf{V}^\dagger\mathbf{L}\mathbf{b}$ is computed exactly.

Finally, if $f \in \mathcal{LS}$ with Laplace-Stieltjes density μ_L, we apply Theorem 8.31 to deduce

$$\begin{aligned} \|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| &\leq \int_0^\infty \mu_L(\zeta)\|e^{-\zeta\mathbf{L}}\mathbf{b} - \mathbf{V}e^{-\zeta\mathbf{L}}\mathbf{V}^\dagger\mathbf{b}\| d\zeta \\ &\leq 4\gamma_m\|\mathbf{b}\|\|r_\Xi\|_\Sigma \int_0^\infty \mu_L(\zeta)\, d\zeta = 4\gamma_m f(0^+)\|\mathbf{b}\|\|r_\Xi\|_\Sigma. \end{aligned}$$

□

Remark 8.33. *The very same upper bounds as in Theorem 8.32 can be obtained using so-called skeleton approximations [DKZ09, MR20a, Ose07] or the Hermite-Walsh formula for rational interpolants [Wal60, Theorem VIII.2], [BR09, p. 24], [BG12].*

Theorem 8.32 is an integral part of our analysis and thus deserves some further discussions. The inequalities (8.32) and (8.33) show that the choice of a "good" pole set Ξ is closely related to rational functions that are uniformly as small as possible on Σ and as large as possible on $\mathbb{R}_0^-$ if $f \in \mathcal{CS} \cup \mathcal{CB}$. The involved constants depend on the function f evaluated at the extremal eigenvalues of $\mathbf{L}$. Our results suggest that one should ensure

- $\infty \in \Xi$ if $\theta \neq 0$ and $f \in \mathcal{CS}$,
- $\{\infty, \infty\} \subset \Xi$ if $\theta, \omega \neq 0$ and $f \in \mathcal{CB}$,

which is computationally inexpensive since it involves polynomial Krylov steps only. The estimate derived for Laplace-Stieltjes functions is seemingly different from the ones obtained in (8.32) and (8.33). The quality of the poles is related to the maximal deviation of r_Ξ in Σ and the constant in (8.34) involves the value of f at 0. The latter is not necessarily finite as the function $\lambda^{-1} \in \mathcal{LS}$ shows. As noted in [MR20a, Remark 4], however, this inconvenience can be tackled in the following manner: Given $f \in \mathcal{LS}$ it follows from the fifth property in Lemma 8.19 that $\tilde{f}(\lambda) := f(\lambda + c) \in \mathcal{LS}$ for any $c \in \mathbb{R}^+$. Provided that $c \in (0, \lambda_{\min})$, we may thus define $\tilde{\mathbf{L}} := \mathbf{L} - c\mathbf{I}$ such that $\tilde{f}(\tilde{\mathbf{L}}) = f(\mathbf{L})$ and $\tilde{f}(0) = f(c) < \infty$ by construction. The third claim in Theorem 8.32 reveals

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| = \|\tilde{f}(\tilde{\mathbf{L}})\mathbf{b} - \mathbf{V}\tilde{f}(\tilde{\mathbf{L}}_{k+1})\mathbf{V}^\dagger\mathbf{b}\| \leq 4\gamma_m f(c)\|\mathbf{b}\|\|r_\Xi\|_{\tilde{\Sigma}},$$

where $\tilde{\mathbf{L}}_{k+1} := \mathbf{L}_{k+1} - c\mathbf{I}_{k+1}$ and $\tilde{\Sigma} := [\lambda_{\min} - c, \lambda_{\max} + c] \subset \mathbb{R}^+$. This allows one to recover a meaningful upper bound on the basis of the modified spectral interval $\tilde{\Sigma}$. Since the ratio of $\lambda_{\min} - c$ and $\lambda_{\max} - c$ is smaller than the one of the unchanged spectral interval, values of c close to $\lambda_{\min}$ lead to worse condition numbers but to smaller constants $f(c)$. On the other hand, if $\tilde{f}$ happens to be completely monotonic even for some $c < 0$, a smaller condition number at the cost of a larger constant $f(c)$ might be obtained.

In all our experiments, we find that the logarithmically increasing factor γ_m in (8.34) is somewhat pessimistic. If $f \in \mathcal{LS}$ extends continuously to the imaginary axis and satisfies a certain decay condition, it is possible to replace γ_m with an absolute constant. In view of Lemma 8.16, the key idea is to bring, instead of its kernel, the function f itself via Cauchy's integral formula in the form of

$$f(z) = \frac{1}{2\pi i} \int_{i\mathbb{R}} \frac{f(z)}{\lambda - z}\, dz,$$

cf. [MN18]. The upper bound so obtained is very similar to the minimization problems in (8.32) and (8.33) but involves the imaginary axis instead of $\mathbb{R}_0^-$.

Theorem 8.34. *Let $f \in \mathcal{LS}$, $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^\Xi(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}f(\mathbf{L}_{k+1})\mathbf{V}^\dagger\mathbf{b}$. Assume that f extends continuously to the imaginary axis such that $|f(z)| \to 0$ as $|z| \to \infty$ for $\Re z \geq 0$ and*

$$c_f := \int_{i\mathbb{R}} \left| \frac{f(z)}{\lambda_{\min} + z} \right| dz < \infty. \tag{8.35}$$

Then there holds

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq \frac{c_f}{\pi}\|\mathbf{b}\| \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(z)| : z \in i\mathbb{R}\}}.$$

Proof. Due to the assumptions on f, we may use the imaginary axis in Cauchy's integral theorem to write

$$f(\mathbf{L})\mathbf{b} = \frac{1}{2\pi i}\int_{i\mathbb{R}} f(z)(\mathbf{L} - z\mathbf{I})^{-1}\mathbf{b}\, dz,$$
$$\mathbf{u}_{k+1} = \mathbf{V}f(\mathbf{L}_{k+1})\mathbf{V}^\dagger\mathbf{b} = \frac{1}{2\pi i}\int_{i\mathbb{R}} f(z)\mathbf{V}(\mathbf{L}_{k+1} - z\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\, dz.$$

Since f is a real-valued function defined on $\mathbb{R}^+$, it follows by the Schwarz reflection principle [Hen93] that $f(\overline{z}) = \overline{f(z)}$. After the transformation $z \mapsto -z$, we thus obtain

$$\begin{aligned} f(\mathbf{L})\mathbf{b} &= \frac{1}{2\pi i}\int_{i\mathbb{R}} f(-z)(\mathbf{L} + z\mathbf{I})^{-1}\mathbf{b}\, dz \\ &= \frac{1}{2\pi i}\int_{i\mathbb{R}} f(\overline{z})(\mathbf{L} + z\mathbf{I})^{-1}\mathbf{b}\, dz = \frac{1}{2\pi i}\int_{i\mathbb{R}} \overline{f(z)}(\mathbf{L} + z\mathbf{I})^{-1}\mathbf{b}\, dz, \end{aligned}$$

and analogously

$$\mathbf{u}_{k+1} = \frac{1}{2\pi i}\int_{i\mathbb{R}} \overline{f(z)}(\mathbf{L}_{k+1} + z\mathbf{I}_{k+1})^{-1}\mathbf{b}\, dz.$$

Triangle inequality combined with Corollary 8.25 reveals

$$\begin{aligned} \|f(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| &\leq \frac{1}{2\pi}\int_{i\mathbb{R}} |f(z)|\|(\mathbf{L} + z\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1} + z\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\|\, dz \\ &\leq \frac{1}{\pi}\|\mathbf{b}\|\int_{i\mathbb{R}} \left|\frac{f(z)}{\lambda_{\min} + z}\right| dz \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(z)| : z \in i\mathbb{R}\}}, \end{aligned}$$

as to be proved. □

9 Zolotarëv's Rational Approximation Problems

The previous chapter shows that for a large class of fractional diffusion problems the rational Krylov error can be bounded in terms of a rational approximation problem. In view of Theorem 8.32, it is desirable to extract surrogate from a rational Krylov space $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ with poles $\Xi = \{\xi_0, \dots, \xi_k\}$ choosen according to

$$\min_{\substack{\Xi \subset \overline{\mathbb{R}} \\ |\Xi| = k+1}} \min_{r_{k+1} \in \mathcal{P}_{k+1}/q_\Xi} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(z)| : z \in \mathbb{R}_0^-\}} \tag{9.1}$$

if $f^\tau \in \mathcal{CS}$. If $f \in \mathcal{CB}$, we fix $\xi_0 = \infty$ and are interested in (9.1) with $\mathcal{P}_{k+1}$ replaced by $\mathcal{P}_k$. If $f \in \mathcal{LS}$ satisfies (8.35), Theorem 8.34 motivates us to minimize (9.1) with $i\mathbb{R}$ in place of $\mathbb{R}_0^-$. Either of these configurations can be reduced to the so-called *generalized third Zolotarëv problem:* Find $r_k^* \in \mathcal{R}_{k,k}$ such that

$$\frac{\|r_k^*\|_\Sigma}{\inf\{|r_k^*(z)| : z \in \mathbb{B}\}} = \inf_{r_k \in \mathcal{R}_{k,k}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(z)| : z \in \mathbb{B}\}}, \quad \mathbb{B} \in \{\mathbb{R}_0^-, i\mathbb{R}\}. \tag{9.2}$$

Roughly spoken, the set Ξ should be chosen according to the poles of a rational function that is uniformly small on Σ and uniformly large on $\mathbb{B}$. In view of (8.34), another rational approximation problem that we are interested in is encoded in the quantity

$$\min_{\substack{\Xi \subset -\Sigma \\ |\Xi| = k}} \|r_\Xi\|_\Sigma, \tag{9.3}$$

which involves rational functions that are uniformly small on the spectral interval. This minimization problem is known as *Zolotarëv's minimal deviation problem* and turns out to be a special case of the generalized third Zolotarëv problem. In this chapter, we analyze (9.2) and (9.3) to provide the foundation for various pole selection strategies presented in Chapter 10.

9.1 The Third Zolotarëv Problem

We consider (9.2) in the following more general form: Given two nonempty and disjoint subsets of the complex plane $\mathbb{A}$, $\mathbb{B} \subset \mathbb{C}$, we search for a rational function $r_k^* \in \mathcal{R}_{k,k}$ with the property

$$\frac{\sup\{|r_k^*(z)| : z \in \mathbb{A}\}}{\inf\{|r_k^*(z)| : z \in \mathbb{B}\}} = \inf_{r_k \in \mathcal{R}_{k,k}} \frac{\sup\{|r_k(z)| : z \in \mathbb{A}\}}{\inf\{|r_k(z)| : z \in \mathbb{B}\}}. \tag{9.4}$$

The problem (9.4) is known as the *generalized third Zolotarëv problem* on $(\mathbb{A}, \mathbb{B})$ [Gon69, Tod84, PP88, Ach92, Wac13, RTW20]. Its systematical study goes back to the Russian mathematician Jegor Zolotarëv in the second half of the 19^{th} century [Zol77] and was originally posed on $\mathbb{A} = \{\lambda \in \mathbb{R} : |\lambda| \leq 1\}$ and $\mathbb{B} = \{\lambda \in \mathbb{R} : |\lambda| \geq c\}$ for some $c > 1$. The quantity

$$Z_k(\mathbb{A}, \mathbb{B}) := \inf_{r_k \in \mathcal{R}_{k,k}} \frac{\sup\{|r_k(z)| : z \in \mathbb{A}\}}{\inf\{|r_k(z)| : z \in \mathbb{B}\}}$$

is called *Zolotarëv number of* $(\mathbb{A}, \mathbb{B})$. An equivalent representation which is frequently used in the literature reads [Gon69, IT95]

$$Z_k(\mathbb{A}, \mathbb{B}) = \inf_{r_k \in \mathcal{R}_{k,k}^{\mathbb{B}}} \sup_{z \in \mathbb{A}} |r_k(z)|, \qquad \mathcal{R}_{k,k}^{\mathbb{B}} := \{r_k \in \mathcal{R}_{k,k} : \inf_{z \in \mathbb{B}} |r_k(z)| = 1\}.$$

A few immediate observations are the following.

Lemma 9.1. *There holds*

1. $Z_0(\mathbb{A}, \mathbb{B}) = 1$,
2. $(Z_k(\mathbb{A}, \mathbb{B}))_{k \in \mathbb{N}_0}$ *is a decreasing sequence in* $\mathbb{R}_0^+$,
3. $\mathbb{A}_1 \subset \mathbb{A}_2$ *and* $\mathbb{B}_1 \subset \mathbb{B}_2$ *implies* $Z_k(\mathbb{A}_1, \mathbb{B}_1) \leq Z_k(\mathbb{A}_2, \mathbb{B}_2)$,
4. $Z_{k+j}(\mathbb{A}, \mathbb{B}) \leq Z_k(\mathbb{A}, \mathbb{B}) Z_j(\mathbb{A}, \mathbb{B})$,
5. $Z_k(\mathbb{A}, \mathbb{B}) = Z_k(\mathbb{B}, \mathbb{A})$ *and* r_k^* *minimizes* $Z_k(\mathbb{A}, \mathbb{B})$ *if and only if* $1/r_k^*$ *minimizes* $Z_k(\mathbb{B}, \mathbb{A})$.

Proof. The first statement is clear. The properties 2 to 4 follow from the observation that the infimum/supremum on a subset is larger/smaller than the one on the original set. The last claim follows from

$$Z_k(\mathbb{A}, \mathbb{B}) = \inf_{r_k \in \mathcal{R}_{k,k}} \frac{\sup\{|r_k(z)| : z \in \mathbb{A}\}}{\inf\{|r_k(z)| : z \in \mathbb{B}\}} = \inf_{r_k \in \mathcal{R}_{k,k}} \frac{\sup\{1/|r_k(z)| : z \in \mathbb{B}\}}{\inf\{1/|r_k(z)| : z \in \mathbb{A}\}} = Z_k(\mathbb{B}, \mathbb{A}).$$

Therefore, r_k^* minimizes $Z_k(\mathbb{A}, \mathbb{B})$ if and only if

$$\frac{\sup\{1/|r_k^*(z)| : z \in \mathbb{B}\}}{\inf\{1/|r_k^*(z)| : z \in \mathbb{A}\}} = Z_k(\mathbb{B}, \mathbb{A}). \qquad \square$$

An equivalent formulation of (9.4) is obtained by the *generalized fourth Zolotarëv problem*: Find $\hat{r}_k \in \mathcal{R}_{k,k}$ such that

$$\sup_{z \in \mathbb{A} \cup \mathbb{B}} |\hat{r}_k(z) - \mathbb{1}_{\mathbb{A},\mathbb{B}}(z)| = \inf_{r_k \in \mathcal{R}_{k,k}} \sup_{z \in \mathbb{A} \cup \mathbb{B}} |r_k(z) - \mathbb{1}_{\mathbb{A},\mathbb{B}}(z)|, \tag{9.5}$$

where

$$\mathbb{1}_{\mathbb{A},\mathbb{B}}(z) := \begin{cases} 1, & z \in \mathbb{A}, \\ -1, & z \in \mathbb{B}. \end{cases}$$

For real intervals it was shown in [Ach92, Chapter 9] that the third and fourth Zolotarëv problem is equivalent in the sense that every solution r_k^* of (9.4) is related to $\hat{r}_k$ satisfying (9.5) via

$$\hat{r}_k(z) = \frac{1 - Z_k(\mathbb{A},\mathbb{B})}{1 + Z_k(\mathbb{A},\mathbb{B})} \frac{r_k^*(z) - \sqrt{Z_k(\mathbb{A},\mathbb{B})}}{r_k^*(z) + \sqrt{Z_k(\mathbb{A},\mathbb{B})}}.$$

As shown in [IT95], this result remains in force if $\mathbb{A}$ and $\mathbb{B}$ are subsets of the complex plane.

Remark 9.2. *Inspired by the close collaboration with his fellow mathematician Pafnuty Chebyschev, Zolotarëv posed in total four fundamental best approximation problems which are known as the four Zolotarëv problems. The first and the second one deal with best-approximation properties of polynomials.*

Remark 9.3. *Zolotarëv problems have been studied in more general form in [LS94, LS01], where the infimum is taken over $\mathcal{R}_{k,m}$, $m \in \mathbb{N}$, instead of $\mathcal{R}_{k,k}$. This theory could be used to optimize the Krylov space under the condition that a fraction of the poles must be infinite.*

The third Zolotarëv problem finds applications in many different branches of modern science, e.g., Alternating Directional Implicit (ADI) methods for solving Sylvester matrix equations [BVY62, Leb77, Bec11, Wac13], singular value decompositions [NF16, BT17], and generalized eigenvalue problems [GPTV15]. In view of Theorem 8.32 and 8.34, our main focus lies in the discussion of the following questions:

1. Can the extremal rational function r_k^* in (9.4) be determined explicitly?
2. How fast does $Z_k(\mathbb{A},\mathbb{B})$ decay for increasing values of k?

To answer these questions, we avail ourselves of well-known tools from logarithmic potential theory which allow us to characterize the distribution of the roots and poles of r_k^* in terms of the so-called *equilibrium measure*. A closely related concept, the so-called *condenser capacity*, then provides the necessary information to quantify the rate of convergence of the Zolotarëv number as $k \to \infty$.

9.2 Preliminaries from Logarithmic Potential Theory

This section is intended for those who want to acquire the basics of modern logarithmic potential theory as quickly as possible. In-depth reviews of this matter can be found in [Lan72, Ran95, ST97]. For an excellent beginner's guide in the context of rational approximations we refer to [LS06, Saf10].

9.2.1 The Classical Case

One of the fundamental problems that have contributed to the development of this field of research is the following *electrostatic problem:* Provided a compact set $\mathbb{A}$ in the complex plane, we want to distribute a collection of like-charged particles on $\mathbb{A}$ such that a minimal energy configuration is attained. Here, the distribution of charges is represented by the

set of probability measures on $\mathbb{A}$ and the energy between two particles is modeled to be proportional to the reciprocal of the distance between them. We formalize these ideas in the following fundamental definition.

Definition 9.4. *Let $\mathbb{A} \subset \mathbb{C}$ be a compact set. We introduce $\mathcal{M}(\mathbb{A})$ as the set of all Borel probability measures supported on $\mathbb{A}$. The logarithmic potential of $\nu \in \mathcal{M}(\mathbb{A})$ is defined by*

$$U^{\nu}(z) := \int \ln \frac{1}{|z-t|} d\nu(t)$$

for all $z \in \mathbb{A}$. The quantity

$$I(\nu) := \int U^{\nu}\, d\nu = \int \int \ln \frac{1}{|z-t|}\, d\nu(t)\, d\nu(z)$$

is said to be the logarithmic energy of ν. The logarithmic energy of $\mathbb{A}$ is defined by

$$V_{\mathbb{A}} := \inf_{\nu \in \mathcal{M}(\mathbb{A})} I(\nu). \tag{9.6}$$

The goal of the electrostatic problem is to find a measure $\nu \in \mathcal{M}(\mathbb{A})$ such that the infimum in (9.6) is attained. According to the *fundamental theorem of Frostman*, this problem has a unique solution [Ran95, Theorem 3.3.4].

Theorem 9.5 (Frostman)**.** *Let $\mathbb{A} \subset \mathbb{C}$ be compact with finite logarithmic energy $V_{\mathbb{A}} < \infty$. Then there exists a unique measure $\nu_{\mathbb{A}} \in \mathcal{M}(\mathbb{A})$ with the property*

$$I(\nu_{\mathbb{A}}) = V_{\mathbb{A}}.$$

The unique measure $\nu_{\mathbb{A}}$ provided by Theorem 9.5 is called *equilibrium measure of* $\mathbb{A}$ and satisfies the following property, where we refer to the outer boundary $\partial_{\infty}\mathbb{A}$ of $\mathbb{A} \subset \mathbb{C}$ as the boundary of the unbounded component of the complement of $\mathbb{A}$.

Lemma 9.6. *Let $\mathbb{A} \subset \mathbb{C}$ be compact with $V_{\mathbb{A}} < \infty$ and $\nu_{\mathbb{A}}$ the equilibrium measure of $\mathbb{A}$. Then there holds* $\operatorname{supp} \nu_{\mathbb{A}} \subset \partial_{\infty}\mathbb{A}$.

Of particular importance in logarithmic potential theory is the following quantity that is intrinsically linked to the equilibrium measure of $\mathbb{A}$.

Definition 9.7. *Let $\mathbb{A} \subset \mathbb{C}$ be compact. Then the logarithmic capacity of $\mathbb{A}$ is defined by*

$$\operatorname{cap}(\mathbb{A}) := e^{-V_{\mathbb{A}}}. \tag{9.7}$$

If $V_{\mathbb{A}} = \infty$ in (9.7), then $\operatorname{cap}(\mathbb{A}) = 0$ by definition. Such sets are called *polar sets*. From the electrostatic point of view, these are the sets that are "too small" to hold a charge. Any subset of a polar set is a polar set itself. Moreover, any polar set has Lebesgue measure zero. The converse is not true. Lemma 9.6, $\mathcal{M}(\partial_{\infty}\mathbb{A}) \subset \mathcal{M}(\mathbb{A})$, and the uniqueness of the equilibrium measure show that one can "fill up" the holes of a set without changing its logarithmic capacity. More precisely, there holds the following result.

Proposition 9.8. *Let $\mathbb{A} \subset \mathbb{C}$ be compact with positive logarithmic capacity and $\nu_{\mathbb{A}}$ its equilibrium measure. Then there holds*

$$\operatorname{cap}(\partial_\infty \mathbb{A}) = \operatorname{cap}(\mathbb{A}).$$

Further remarkable properties of the logarithmic capacity are listed in the following lemma.

Lemma 9.9. *Let $\mathbb{A}$ and $\mathbb{B}$ denote two compact subsets of the complex plane.*

1. *The logarithmic capacity is monotone, i.e., $\mathbb{A} \subset \mathbb{B}$ implies that $\operatorname{cap}(\mathbb{A}) \leq \operatorname{cap}(\mathbb{B})$.*
2. *For all $a, b \in \mathbb{C}$ there holds $\operatorname{cap}(a\mathbb{A} + b) = |a| \operatorname{cap}(\mathbb{A})$.*
3. *If $\mathbb{A}$ has area A, then $\operatorname{cap}(\mathbb{A}) \geq \sqrt{\frac{A}{\pi}}$.*
4. *If $\mathbb{A}$ has diameter D, then $\operatorname{cap}(\mathbb{A}) \leq \frac{D}{2}$.*

Proof. See [Ros97, Theorem 1.2]. □

The logarithmic capacity is a nonnegative set function that in a sense determines the geometry of its arguments. There holds $\operatorname{cap}(\emptyset) = 0$. Moreover, by the second property in Lemma 9.9, we have that cap is invariant under rotations and translations whence it bears a close resemblance to the Lebesgue measures. Unlike the latter, however, it is not additive which makes the evaluation of $\operatorname{cap}(\mathbb{A})$ a highly nontrivial task. For some geometries, well-known tools from complex analysis can be consulted to determine $\operatorname{cap}(\mathbb{A})$ analytically. Recalling that every simply connected set has "no holes", that is, it is path-connected such that every path between two points can be continuously transformed into any other such path while preserving the two endpoints in question, we state the following result which is a consequence of the Riemann mapping theorem.

Lemma 9.10. *Let $\mathbb{A} \subset \mathbb{C}$ be simply connected and compact. Then there exists a unique real number $\rho \in \mathbb{R}_0^+$ and a holomorphic function $\mathfrak{R}$ with nonvanishing derivative which maps the complement of $\mathbb{A}$ onto $\overline{B_\rho(0)}^c$.*

Proof. See [Hen93, Corollary 5.10d] and [Saf10, Theorem 3.5]. □

The map $\mathfrak{R}$ in Lemma 9.10 is commonly referred to as *Riemann map of* $\mathbb{A}$ and is illustrated in Figure 9.1. We call $\operatorname{mod}(\mathbb{A}) := \rho$ the *Riemann modulus of* $\mathbb{A}$. For some simple geometries, $\operatorname{mod}(\mathbb{A})$ is known analytically, in which case the capacity of $\mathbb{A}$ can be determined in the following convenient manner [Saf10, p. 20].

Theorem 9.11. *Let $\mathbb{A} \subset \mathbb{C}$ be simply connected and compact. Then there holds*

$$\operatorname{cap}(\mathbb{A}) = \operatorname{mod}(\mathbb{A}).$$

Typical examples of complex subsets whose capacity can be computed using Theorem 9.11 include [Lan72, Ran95]

- closed discs $\mathbb{A} = \overline{B_r(\mathbf{x})}$ with $\operatorname{cap}(\overline{B_r(\mathbf{x})}) = r$,
- ellipses E with semi-axes a and b, satisfying $\operatorname{cap}(E) = (a+b)/2$,
- the union of two disjoint intervals $\operatorname{cap}([-b,-a] \cup [a,b]) = \sqrt{b^2 - a^2}/2$.

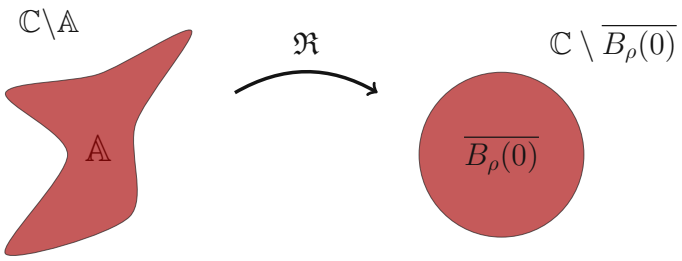

Figure 9.1: Illustration of the Riemann map $\mathfrak{R}$ that maps the complement of $\mathbb{A}$ to the complement of $\overline{B_\rho(0)}$, where $\rho = \mathrm{mod}(\mathbb{A})$ is the Riemann modulus of $\mathbb{A}$.

9.2.2 Generalizations to Signed Measures

For the later use, it turns out to be fruitful to generalize the above concepts for *signed measures* $\nu \in \mathcal{M}(\mathbb{A}, \mathbb{B})$ defined by

$$\mathcal{M}(\mathbb{A}, \mathbb{B}) := \{\nu = \nu_{\mathbb{A}} - \nu_{\mathbb{B}} : \nu_{\mathbb{A}} \in \mathcal{M}(\mathbb{A}), \nu_{\mathbb{B}} \in \mathcal{M}(\mathbb{B})\}.$$

The following definition is now essential [Gon69].

Definition 9.12. *Let $\mathbb{A} \subset \mathbb{C}$ be compact and $\mathbb{B} \subset \mathbb{C}$ closed, having positive logarithmic capacity each, with* $\mathrm{dist}(\mathbb{A}, \mathbb{B}) > 0$. *Then the pair $(\mathbb{A}, \mathbb{B})$ is called a condenser and the sets $\mathbb{A}$ and $\mathbb{B}$ are called plates.*

A straightforward generalization of Definition 9.4 reads as follows.

Definition 9.13. *Let $(\mathbb{A}, \mathbb{B})$ be a condenser. The logarithmic potential of $\nu \in \mathcal{M}(\mathbb{A}, \mathbb{B})$ is defined by*

$$U^\nu(z) := \int \ln \frac{1}{|z-t|} d\nu(t)$$

for all $z \in \mathbb{A}$. The quantity

$$I(\nu) := \int U^\nu \, d\nu = \int \int \ln \frac{1}{|z-t|} \, d\nu(t) \, d\nu(z)$$

is said to be the logarithmic energy of ν. The logarithmic energy of the condenser $(\mathbb{A}, \mathbb{B})$ is defined by

$$V_{(\mathbb{A},\mathbb{B})} := \inf_{\nu \in \mathcal{M}(\mathbb{A},\mathbb{B})} I(\nu). \tag{9.8}$$

In line with Theorem 9.5, for each condenser $(\mathbb{A}, \mathbb{B})$ there exists a unique solution $\nu_{(\mathbb{A},\mathbb{B})}$ to the energy problem (9.8) that we call *equilibrium measure of the condenser*. Its support is contained in $\partial(\mathbb{A} \cup \mathbb{B})$, but not necessarily in $\partial_\infty(\mathbb{A} \cup \mathbb{B})$. The equilibrium measure is the key instrument for providing solutions to the third Zolotarëv problem. To quantify the rate of decay of $Z_k(\mathbb{A}, \mathbb{B})$, we require the following terminology.

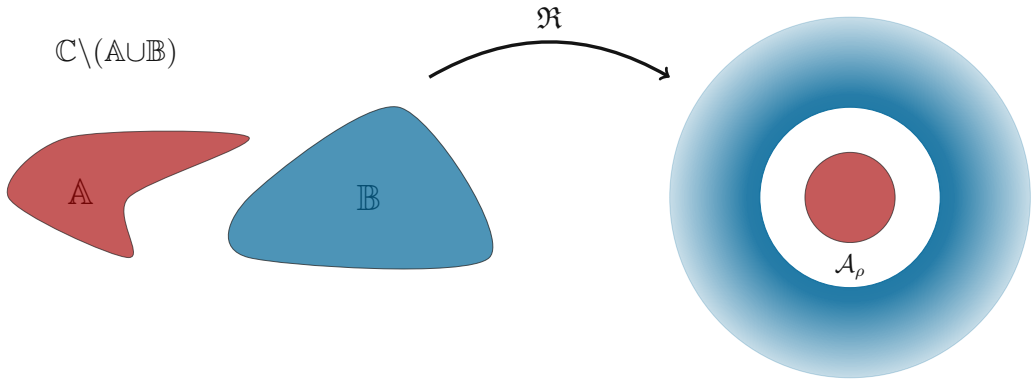

Figure 9.2: Illustration of the Riemann map $\mathfrak{R}$ that maps the complement of $\mathbb{A}\cup\mathbb{B}$ to the annulus $\mathcal{A}_\rho$, where $\rho = \mathrm{mod}(\mathbb{A},\mathbb{B})$ is the Riemann modulus of the condenser $(\mathbb{A},\mathbb{B})$.

Definition 9.14. *The capacity of the condenser* $(\mathbb{A},\mathbb{B})$ *is defined by*

$$\mathrm{cap}(\mathbb{A},\mathbb{B}) := \frac{1}{V_{(\mathbb{A},\mathbb{B})}}.$$

A few remarkable properties of condenser capacities are listed in the following lemma [Lan72, Güt10].

Lemma 9.15. *Let* $(\mathbb{A},\mathbb{B})$ *be a condenser. Then there holds*

1. $\mathrm{cap}(\partial\mathbb{A},\partial\mathbb{B}) = \mathrm{cap}(\mathbb{A},\mathbb{B})$,
2. $\mathrm{cap}(T(\mathbb{A}),T(\mathbb{B}))$ *for any analytic map* $T:\mathbb{C}\to\mathbb{C}$ *with nonvanishing derivative,*
3. $\mathbb{A}_1\subset\mathbb{A}_2$ *and* $\mathbb{B}_1\subset\mathbb{B}_2$ *imply that* $\mathrm{cap}(\mathbb{A}_1,\mathbb{B}_1)\le\mathrm{cap}(\mathbb{A}_2,\mathbb{B}_2)$.

The computation of condenser capacities is a difficult task. Similar to the classical case, however, standard tools from complex analysis can be consulted to compute $\mathrm{cap}(\mathbb{A},\mathbb{B})$ for some particular plate configurations. If $\mathbb{A}$ and $\mathbb{B}$ are connected closed sets, not single points, and do not separate the plane, then their complement $\Omega = (\mathbb{A}\cup\mathbb{B})^c$ is a nonempty, open, and doubly connected set [Hen93]. The latter means that for any two points in Ω there are two different paths that cannot be smoothly deformed into each other. By the Riemann mapping theorem for doubly connected regions [Hen93, Theorem 5.10h], Ω can be mapped conformally to the annulus $\mathcal{A}_\rho := \{z\in\mathbb{C} : 1<|z|<\rho\}$ for some $\rho>1$. We summarize these results in a form that is suitable for the study of condenser capacities.

Lemma 9.16. *Let* $(\mathbb{A},\mathbb{B})$ *be a condenser whose plates are connected, not single points, and do not separate the plane. Then there exists a unique real number* $\rho>1$ *and a holomorphic function* $\mathfrak{R}$ *with nonvanishing derivative which maps the complement of* $\mathbb{A}\cup\mathbb{B}$ *onto the annulus* $\mathcal{A}_\rho := \{z\in\mathbb{C} : 1<|z|<\rho\}$.

The quantity ρ in Lemma 9.16 is called *Riemann modulus of* $\Omega = (\mathbb{A} \cup \mathbb{B})^c$. For our purpose, it is more natural to associate ρ directly to the condenser $(\mathbb{A}, \mathbb{B})$ rather than to Ω, whence we introduce the *Riemann modulus* $\operatorname{mod}(\mathbb{A}, \mathbb{B})$ *of the condenser* $(\mathbb{A}, \mathbb{B})$ as $\operatorname{mod}(\mathbb{A}, \mathbb{B}) := \rho$. The latter is depicted in Figure 9.2. If $\operatorname{mod}(\mathbb{A}, \mathbb{B})$ is available, the capacity of $(\mathbb{A}, \mathbb{B})$ can be computed as follows.

Theorem 9.17. *Let* $(\mathbb{A}, \mathbb{B})$ *be a condenser whose plates are connected, not single points, and do not separate the plane. Then there holds*

$$\operatorname{cap}(\mathbb{A}, \mathbb{B}) = \frac{1}{\ln\left(\operatorname{mod}(\mathbb{A}, \mathbb{B})\right)}.$$

Example 9.18. *We want to compute the capacity of the condenser* $(\mathbb{A}, \mathbb{B})$*, where* $\mathbb{A} = \overline{B_{R_1}(0)}$ *is compact and* $\mathbb{B} = B_{R_2}(0)^c$ *is closed with* $R_1 < R_2$*. Since* $\mathfrak{R}_1(z) = z/R_1$ *is holomorphic with nonvanishing derivative in* $\Omega = (\mathbb{A} \cup \mathbb{B})^c$ *and maps* Ω *onto the annulus* $\mathcal{A}_{R_2/R_1}$*, it follows from Lemma 9.16 that* R_2/R_1 *is the Riemann modulus of* $(\mathbb{A}, \mathbb{B})$ *and thus, by Theorem 9.17,*

$$\operatorname{cap}(\mathbb{A}, \mathbb{B}) = \frac{1}{\ln\left(\frac{R_2}{R_1}\right)}.$$

9.3 Solutions and Upper Bounds to the Third Zolotarëv Problem

We return to the third Zolotarëv problem and establish a connection to logarithmic potential theory. For this purpose, we introduce the following notion of convergence in $\mathcal{M}(\mathbb{A}, \mathbb{B})$.

Definition 9.19. *Let* $\mathbb{A} \subset \mathbb{C}$ *be compact. A sequence of measures* $(\nu_k)_{k\in\mathbb{N}} \subset \mathcal{M}(\mathbb{A})$ *is said to be weak-star convergent to some* $\nu \in \mathcal{M}(\mathbb{A})$ *if*

$$\lim_{k\to\infty} \int f \, d\nu_k = \int f \, d\nu \tag{9.9}$$

for all $f \in C(\mathbb{A})$*. For a condenser* $(\mathbb{A}, \mathbb{B})$*, a sequence of signed measures* $(\nu_k)_{k\in\mathbb{N}} \subset \mathcal{M}(\mathbb{A}, \mathbb{B})$ *is said to be weak-star convergent to some* $\nu \in \mathcal{M}(\mathbb{A}, \mathbb{B})$ *if* (9.9) *holds for all* $f \in C(\mathbb{A} \cup \mathbb{B})$*. In both cases, we write* $\nu_k \overset{*}{\to} \nu$*.*

For a complex subset $\Lambda := \{\sigma_1^{(k)}, \ldots, \sigma_k^{(k)}\}$ we define the *associated normalized counting measure* by

$$\nu_k := \frac{1}{k} \sum_{j=1}^{k} \delta_{\sigma_j^{(k)}},$$

where δ_z denotes the Dirac unit measure in the point $z \in \mathbb{C}$. For the pairing (Λ, Ξ), where $\Xi = \{\xi_1, \ldots, \xi_k\} = \{\xi_1^{(k)}, \ldots, \xi_k^{(k)}\} \subset \mathbb{C}$ is another subset of the complex plane, the *associated normalized signed counting measure* is defined by

$$\nu_k := \frac{1}{k} \sum_{j=1}^{k} \delta_{\sigma_j^{(k)}} - \frac{1}{k} \sum_{j=1}^{k} \delta_{\xi_j^{(k)}}. \tag{9.10}$$

Defining $r_{\Lambda,\Xi}$ as in (8.26)[1], the logarithmic potential of the normalized signed counting measure ν_k associated to Λ and Ξ relates to the absolute value of $r_{\Lambda,\Xi}$ via

$$\begin{aligned} U^{\nu_k}(z) &= \frac{1}{k}\sum_{j=1}^{k} \ln\left(\frac{1}{|z-\sigma_j^{(k)}|}\right) - \frac{1}{k}\sum_{j=1}^{k} \ln\left(\frac{1}{|z-\xi_j^{(k)}|}\right) \\ &= -\frac{1}{k}\ln\left(\prod_{j=1}^{k}|z-\sigma_j^{(k)}|\right) + \frac{1}{k}\ln\left(\prod_{j=1}^{k}|z-\xi_j^{(k)}|\right) \\ &= -\frac{1}{k}\ln\left(\prod_{j=1}^{k}\frac{|z-\sigma_j^{(k)}|}{|z-\xi_j^{(k)}|}\right) = -\ln|r_{\Lambda,\Xi}(z)|^{\frac{1}{k}}, \end{aligned}$$

see also [Saf10, Güt10]. This observation is a key ingredient in the proof of the following result [Gon67, Gon69, LS94].

Theorem 9.20. *Let $(\mathbb{A},\mathbb{B})$ be a condenser. Then there holds*

$$\lim_{k\to\infty}(Z_k(\mathbb{A},\mathbb{B}))^{\frac{1}{k}} = e^{-\frac{1}{\operatorname{cap}(\mathbb{A},\mathbb{B})}}. \tag{9.11}$$

Let $\nu_{(\mathbb{A},\mathbb{B})}$ be the equilibrium measure of the condenser, $\Lambda = \{\sigma_1^{(k)},\dots,\sigma_k^{(k)}\} \subset \mathbb{A}$, $\Xi = \{\xi_1^{(k)},\dots,\xi_k^{(k)}\} \subset \mathbb{B}$, and ν_k the associated normalized signed counting measure defined by (9.10). If $\nu_k \xrightarrow{*} \nu_{(\mathbb{A},\mathbb{B})}$, then there holds

$$\lim_{k\to\infty}\left(\frac{\sup\{|r_{\Lambda,\Xi}(z)| : z\in\mathbb{A}\}}{\inf\{|r_{\Lambda,\Xi}(z)| : z\in\mathbb{B}\}}\right)^{\frac{1}{k}} = e^{-\frac{1}{\operatorname{cap}(\mathbb{A},\mathbb{B})}}. \tag{9.12}$$

The importance of Theorem 9.20 is twofold. First, (9.11) shows that the Zolotarëv number converges geometrically to zero as $k\to\infty$, that is,

$$Z_k(\mathbb{A},\mathbb{B}) \le C_{(\mathbb{A},\mathbb{B})}e^{-\frac{k}{\operatorname{cap}(\mathbb{A},\mathbb{B})}} \tag{9.13}$$

for some positive constant $C_{(\mathbb{A},\mathbb{B})} \in \mathbb{R}^+$. Note that $\operatorname{cap}(\mathbb{A},\mathbb{B})$ cannot be replaced by any smaller number and, due to Theorem 9.17,

$$Z_k(\mathbb{A},\mathbb{B}) \le C_{(\mathbb{A},\mathbb{B})}\rho^{-k}, \qquad \rho := \operatorname{mod}(\mathbb{A},\mathbb{B}), \tag{9.14}$$

if $\mathbb{A}$ and $\mathbb{B}$ are connected, not single points, and do not separate the plane. On the other hand, (9.12) shows that an asymptotically optimal rational function can be obtained if one distributes the zeros and poles according to the equilibrium measure of the condenser $(\mathbb{A},\mathbb{B})$. The literature provides a variety of different possibilities to construct such sequences so that $\nu_k \xrightarrow{*} \nu_{(\mathbb{A},\mathbb{B})}$ holds [WR66, Wac88, EW91, Sta91, Sta92, Sta93]. We mention here only two.

[1] Here we set $\sigma_0^{(k)} = \xi_0^{(k)} = \infty$ and identify Σ and Λ with $\{\infty,\sigma_1^{(k)}\dots,\sigma_k^{(k)}\}$ and $\{\infty,\xi_1^{(k)}\dots,\xi_k^{(k)}\}$, respectively.

1. If the plates of the condenser $(\mathbb{A}, \mathbb{B})$ are connected, not single points, and do not separate the plane, then by Lemma 9.16 there exists a Riemann map $\mathfrak{R}$ that transplants the complement of $\mathbb{A} \cup \mathbb{B}$ to the annulus $\mathcal{A}_\rho$, where $\rho = \operatorname{mod}(\mathbb{A}, \mathbb{B})$ is the modulus of the condenser. The *generalized Fejér points of order* k are defined by [Wal65, Sta91]

$$\begin{aligned} \sigma_j^{(k)} &:= \mathfrak{R}^{-1}\left(e^{\frac{2\pi i j}{k}}\right), \quad j = 1, \dots, k, \\ \xi_j^{(k)} &:= \mathfrak{R}^{-1}\left(\rho e^{\frac{2\pi i j}{k}}\right), \quad j = 1, \dots, k, \end{aligned} \tag{9.15}$$

   and provide asymptotically minimal rational functions for the third Zolotarëv problem. Clearly, the availability of $(\xi_j^{(k)})_{j=1}^k$ hinges on the knowledge of the mapping $\mathfrak{R}$. If $\mathbb{A}$ and $\mathbb{B}$ are e.g., bounded polygons, then $\mathfrak{R}$ can be constructed as a doubly-connected Schwarz–Christoffel mapping [Hu98].

2. Another approach has been presented in [Bag69]. Provided two starting points $\sigma_1 \in \mathbb{A}$ and $\xi_1 \in \mathbb{B}$ with $\Lambda := \{\sigma_1\}$ and $\Xi := \{\xi_1\}$, the *generalized Leja points* are defined inductively by

$$\begin{aligned} \sigma_{j+1}^{(k)} &:= \sigma_{j+1} := \arg\max_{z \in \mathbb{A}} |r_{\Lambda,\Xi}(z)|, \\ \xi_{j+1}^{(k)} &:= \xi_{j+1} := \arg\min_{z \in \mathbb{B}} |r_{\Lambda,\Xi}(z)|. \end{aligned} \tag{9.16}$$

   Unlike the generalized Fejér points, this procedure generates a nested sequence of roots and nodes.

Any sequence of rational functions generated by either of the two methods mentioned above satisfies

$$\lim_{k\to\infty} \left(\frac{\sup\{|r_{\Lambda,\Xi}(z)| : z \in \mathbb{A}\}}{\inf\{|r_{\Lambda,\Xi}(z)| : z \in \mathbb{B}\}} \right)^{\frac{1}{k}} = e^{-\frac{1}{\operatorname{cap}(\mathbb{A},\mathbb{B})}}. \tag{9.17}$$

Although useful in practice, (9.17) provides no information about the optimality of $r_{\Lambda,\Xi}$ for finite k. Indeed, both (9.15) and (9.16) generally fail to recover solutions to the third Zolotarëv problem whenever $k < \infty$. Likewise, (9.13) only reveals the *rate of convergence* but does not say anything about the constant $C_{(\mathbb{A},\mathbb{B})}$. In practice, however, one is often interested in quantifying the value of $C_{(\mathbb{A},\mathbb{B})}$ to estimate the smallest $k \in \mathbb{N}$ such that the value of $Z_k(\mathbb{A}, \mathbb{B})$ falls below a user-defined threshold. The derivation of rigorous estimates for $Z_k(\mathbb{A}, \mathbb{B})$ involving arbitrary condenser is still a subject of ongoing research. Under rather general assumptions on $\mathbb{A}$ and $\mathbb{B}$, the value of $C_{(\mathbb{A},\mathbb{B})}$ can be estimated by means of the so-called *total rotation* of the plates [RTW20]. In the following, we investigate a few particular geometries for which an optimal rational function and the value of $C_{(\mathbb{A},\mathbb{B})}$ are known explicitly.

9.3.1 Real, Symmetric, and Normalized Intervals

A solution to (9.4) for the condenser $(\mathbb{A}, -\mathbb{A})$, $\mathbb{A} = [\delta, 1]$, $\delta \in (0, 1)$, was derived almost 150 years ago by Zolotarëv [Zol77]. The explicit representation of r_k^* makes heavy use of

elliptic integrals and Jacobi elliptic functions, which we briefly recall here. To this end, we introduce the *incomplete elliptic integral of first kind* $\mathcal{K} : [0, \frac{\pi}{2}] \times [0, 1) \to \mathbb{R}^+$ defined by [AS64, Section 17], [OLBC10, Section 19]

$$\mathcal{K}(\phi, \mathrm{k}) := \int_0^{\phi} \frac{d\theta}{\sqrt{1 - \mathrm{k}^2 \sin^2 \theta}}.$$

The parameter k is called *elliptic modulus*. If $\phi = \frac{\pi}{2}$, one sets $\mathcal{K}(\mathrm{k}) := \mathcal{K}(\frac{\pi}{2}, \mathrm{k})$ which is named *complete elliptic integral of first kind*. There holds [AS64, Section 17]

$$\mathcal{K}(0) = \frac{\pi}{2}, \qquad \mathcal{K}(\mathrm{k}) \approx \frac{1}{2} \ln\left(\frac{16}{1 - \mathrm{k}^2}\right) \text{ as } \mathrm{k} \to 1^-. \tag{9.18}$$

Remark 9.21. *The definition of $\mathcal{K}$ is not unique in the literature. In many textbooks, the elliptic modulus* k *is replaced by the parameter* $\mathrm{m} = \mathrm{k}^2$.

For the later use, we further introduce the *Grötzsch ring function* $\mu : (0, 1) \to \mathbb{R}_0^+$ as

$$\mu(\mathrm{k}) := \frac{\pi}{2} \frac{\mathcal{K}(\sqrt{1 - \mathrm{k}^2})}{\mathcal{K}(\mathrm{k})}.$$

Together with the complete elliptic integral of first kind, the function μ is depicted in Figure 9.3. Unlike $\mathcal{K}$, the Grötzsch ring function is strictly decreasing and can by bounded by [OLBC10, eq. (19.2.8)]

$$\ln\left(\frac{(1 + \sqrt[4]{1 - \mathrm{k}^2})^2}{\mathrm{k}}\right) \le \mu(\mathrm{k}) \le \ln\left(\frac{2(1 + \sqrt{1 - \mathrm{k}^2})}{\mathrm{k}}\right) \le \ln\left(\frac{4}{\mathrm{k}}\right), \tag{9.19}$$

which shows

$$\mu(\mathrm{k}) \approx \ln\left(\frac{4}{\mathrm{k}}\right) \text{ as } \mathrm{k} \to 0^+.$$

Moreover, it follows from (9.18) that

$$\lim_{\mathrm{k} \to 1^-} \mu(\mathrm{k}) = 0.$$

Finally, we also introduce the *Jacobi elliptic function* dn which is "inversely" defined by [AS64, Section 16], [OLBC10, Section 22]

$$\mathrm{dn}(z, \mathrm{k}) := \sqrt{1 - \mathrm{k}^2 \sin^2 \phi}, \qquad z = \mathcal{K}(\phi, \mathrm{k}),$$

for all $\mathrm{k} \in [0, 1)$ and $z \in \mathbb{C}$. As a function of z with fixed k, dn is even, meromorphic, and doubly periodic with real period $2\mathcal{K}(\mathrm{k})$ and imaginary period $4i\mathcal{K}(\sqrt{1 - \mathrm{k}^2})$. For real arguments, the function is illustrated in Figure 9.3.

Given the terminology provided above, we are now in position to state the central definition of this chapter.

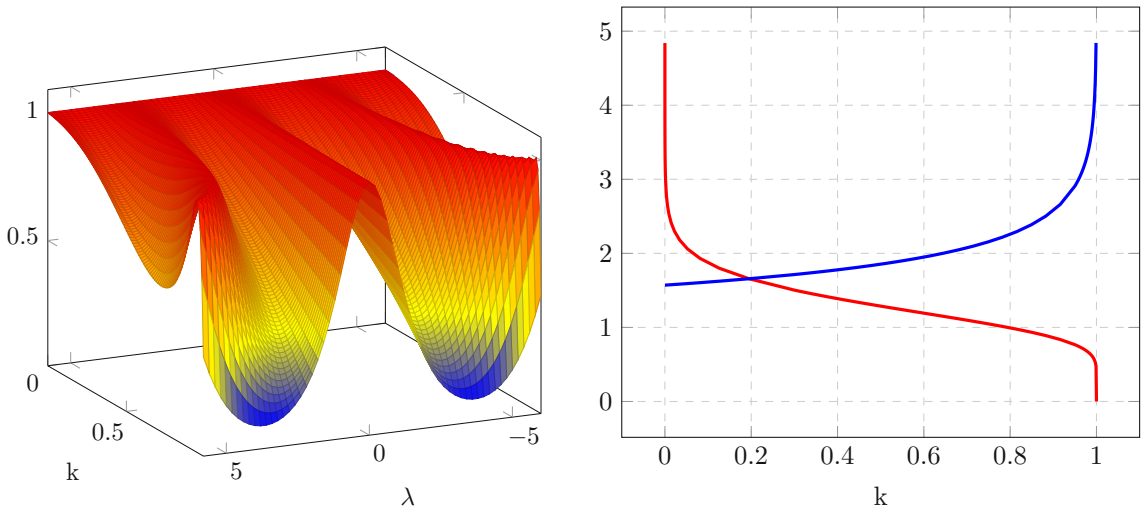

Figure 9.3: Illustration of the Jacobi elliptic function $\mathrm{dn}(\lambda, \mathrm{k})$ (left) with real argument $\lambda \in [-5, 5]$ and elliptic modulus $\mathrm{k} \in [0, 1]$. On the right, we see the complete elliptic integral $\mathcal{K}(\mathrm{k})$ of first kind (blue) and the Grötzsch ring function $\mu(\mathrm{k})$ (red) on $(0, 1)$.

Definition 9.22. *Let $k \in \mathbb{N}$ and $\delta \in (0, 1)$. We define the Zolotarëv points of order k on $[\delta, 1]$ as*

$$\mathcal{Z}_j^{(k)} := \mathrm{dn}\left(\frac{2(k-j)+1}{2k}\mathcal{K}(\delta'), \delta'\right), \qquad \delta' := \sqrt{1-\delta^2}, \tag{9.20}$$

for all $j = 1, \ldots, k$.

Our interest in these points is due to the following theorem [Tod84, Ach92, IT95].

Theorem 9.23. *Let $\delta \in (0, 1)$, $\mathbb{A} = [\delta, 1]$, and $\left(\mathcal{Z}_j^{(k)}\right)_{j=1}^k$ the Zolotarëv points of order k on $\mathbb{A}$. Then*

$$r_k^*(\lambda) = \prod_{j=1}^{k} \frac{\lambda - \mathcal{Z}_j^{(k)}}{\lambda + \mathcal{Z}_j^{(k)}} \tag{9.21}$$

solves the third Zolotarëv problem (9.4) on the condenser $(\mathbb{A}, -\mathbb{A})$.

The Zolotarëv points on $\mathbb{A} = [\delta, 1]$ are depicted in Figure 9.4 for $\delta = 0.01$ as roots and poles of the solution (9.21) to the third Zolotarëv problem on $(\mathbb{A}, -\mathbb{A})$. The nodes are roughly geometrically distributed across the interval. In accordance with the theory, r_k^* is uniformly small on $\mathbb{A}$ and uniformly large on $-\mathbb{A}$.

Remark 9.24. *In practice, one is often interested in the evaluation of (9.20) for small values of δ. In this regime, the evaluation of δ' is prone to cancellation errors. To counteract*

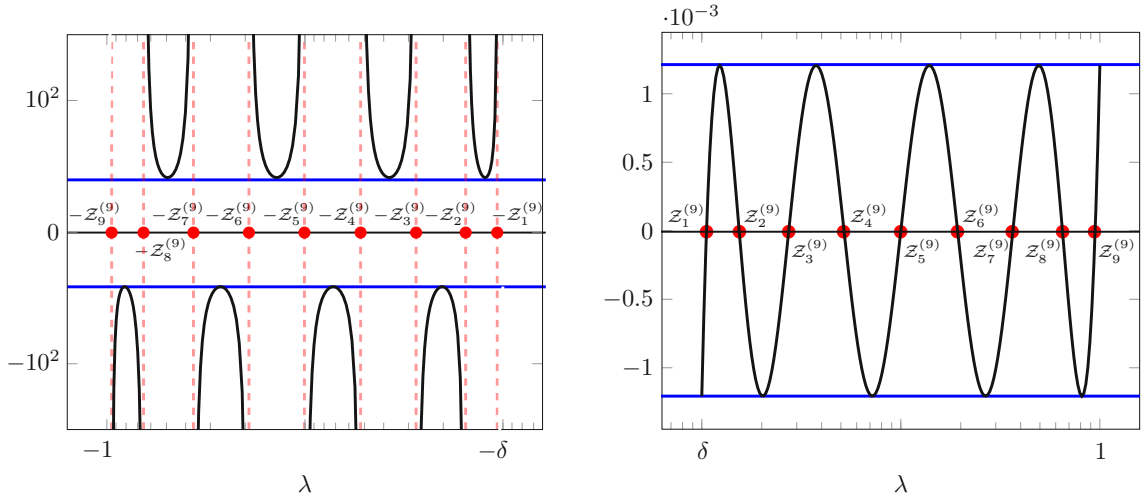

Figure 9.4: Solution (9.21) to the third Zolotarëv problem on the condenser $(\mathbb{A}, -\mathbb{A})$ with $\mathbb{A} = [\delta, 1]$ and $\delta = 0.01$ on $-\mathbb{A}$ (left) and $\mathbb{A}$ (right). The red dots indicate the poles and zeros of r_k^* which coincide with the negative and positive Zolotarëv points on $\mathbb{A}$, respectively.

this, it is advisable to resort to the asymptotic formulas [AS64, (16.15.3) and (17.3.26)]

$$\operatorname{dn}(z, \sqrt{1-\delta^2}) \approx \frac{1}{\cosh(z)} + \frac{\delta^2}{4}\left(\sinh(z)\cosh(z) + z\right)\frac{\tanh(z)}{\cosh(z)},$$
$$\mathcal{K}(\sqrt{1-\delta^2}) \approx \frac{1}{2}\ln\left(\frac{16}{\delta^2}\right),$$

for small values of δ.

Given an explicit solution to the third Zolotarëv problem on $(\mathbb{A}, -\mathbb{A})$, our main focus now lies in the derivation of explicit bounds for the Zolotarëv number $Z_k(\mathbb{A}, -\mathbb{A})$. Whenever $\mathbb{A} = [\delta, 1] \subset \mathbb{R}^+$, the Riemann modulus $\rho_\delta := \operatorname{mod}(\mathbb{A}, -\mathbb{A})$ of the condenser is known explicitly and reads [BT17]

$$\rho_\delta = e^{\frac{\pi^2}{\mu(\delta)}},$$

where μ is the Grötzsch ring function. Thanks to (9.14) we obtain

$$Z_k(\mathbb{A}, -\mathbb{A}) \leq C_{\mathbb{A}} \rho_\delta^{-k},$$

where $C_{\mathbb{A}} = C_{\mathbb{A},-\mathbb{A}} \in \mathbb{R}^+$. To quantify the latter, a product formula for $Z_k(\mathbb{A}, -\mathbb{A})$ has been derived[2] in [BT17] and reads

$$Z_k(\mathbb{A}, -\mathbb{A}) = 4\rho_\delta^{-k} \prod_{j=1}^{\infty} \frac{(1+\rho_\delta^{-4jk})^4}{(1+\rho_\delta^{2k}\rho_\delta^{-4jk})^4}.$$

[2] More precisely, the authors of [BT17] corrected an erroneous formula that was originally derived in [Leb77, eq. (1.11)] but contained some typos.

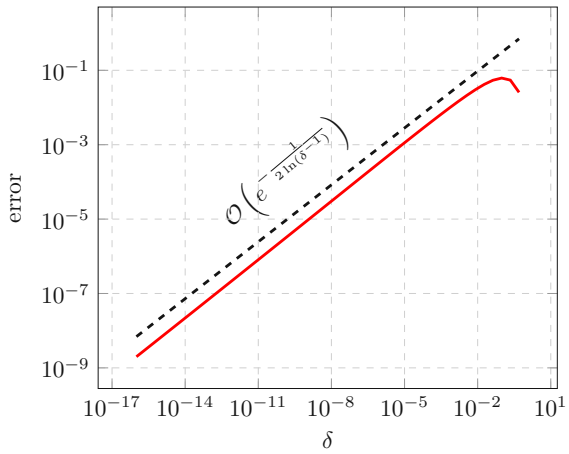

Figure 9.5: Absolute error $|4e^{-\frac{\pi^2}{\mu(\delta)}} - 4e^{-\frac{\pi^2}{\ln(4\delta^{-1})}}|$ for $\delta \in [10^{-16}, \frac{1}{2}]$.

This is the key ingredient in the derivation of a rigorous upper bound of the Zolotarëv number for real, symmetric, and normalized intervals $\mathbb{A} = [\delta, 1]$ with $\delta \in (0, 1)$.

Theorem 9.25. *Let $\delta \in (0, 1)$ and $\mathbb{A} = [\delta, 1]$. Then there holds for all $k \in \mathbb{N}$*

$$Z_k(\mathbb{A}, -\mathbb{A}) \leq 4\rho_\delta^{-k}, \qquad \rho_\delta = \frac{e^{\pi^2}}{\mu(\delta)}, \tag{9.22}$$

where μ is the Grötzsch ring function.

Proof. See [BT17, Corollary 3.2]. □

Remark 9.26. *The estimate is sharp in a sense that neither 4 nor ρ_δ can be replaced by any smaller/larger number, respectively.*

Assuming that δ is reasonably small, we can make the upper bound in (9.22) more intuitive. Invoking $\mu(\mathrm{k}) \leq \ln(\frac{4}{\mathrm{k}})$ from (9.19), we deduce the slightly weaker bound

$$Z_k(\mathbb{A}, -\mathbb{A}) \leq 4e^{-\frac{\pi^2 k}{\ln(4\delta^{-1})}}, \tag{9.23}$$

which does not involve the Grötzsch ring function. Numerically, we observe in Figure 9.5 that already for $\delta = 10^{-5}$ the error $|4e^{-\frac{\pi^2}{\mu(\delta)}} - 4e^{-\frac{\pi^2}{\ln(4\delta^{-1})}}|$ is in the range of 10^{-3}. Since the problems that we are interested in typically have large condition numbers, causing $\delta \ll 1$, we observe that the sub-optimality of (9.23) compared to (9.22) is negligible. For the reader's convenience, we therefore use (9.23) in our final estimates to make our analysis more amenable to those who are less familiar with the Grötzsch ring function.

9.3.2 Arbitrary Real Intervals

The previous section provides solutions and upper bounds to the third Zolotaëv problem on $(\mathbb{A}, \mathbb{B})$ when $\mathbb{A} = [\delta, 1]$ for some $\delta \in (0, 1)$ and $\mathbb{B} = -\mathbb{A}$. In view of Theorem 8.32 and 8.34,

however, our main interest lies in the treatment of condensers where $\mathbb{A} = \Sigma \subset \mathbb{R}^+$ is the spectral interval of $\mathbf{L}$ and $\mathbb{B}$ either the negative real line or the imaginary axis. In this section, we generalize the results from Section 9.3.1 to arbitrary disjoint real intervals $\mathbb{A} \subset \mathbb{R}^+$ and $\mathbb{B} \subset \mathbb{R}_0^-$, which includes $(\mathbb{A}, \mathbb{B}) = (\Sigma, \mathbb{R}_0^-)$ as special case. The main idea in the treatment of these problems is to transform $(\mathbb{A}, \mathbb{B})$ to the symmetric condenser $([\delta, 1], [-1, -\delta])$ for some suitable $\delta \in \mathbb{R}^+$, apply available results from the previous section, and transform the obtained solution back to the original configuration. The success of this approach hinges on the underlying transformation which in some sense needs to preserve the optimality property of solutions to the third Zolotarëv problem. The right transformations to do this are *Möbius transformations* which we briefly recall here.

A function $T : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is said to be a *Möbius transformation* if

$$T(z) = \frac{az + b}{cz + d} \tag{9.24}$$

for a quadruple of complex parameters $a, b, c, d \in \mathbb{C}$ satisfying $ad - bc \neq 0$. There holds $T(\infty) = \frac{a}{c}$ and $T(-\frac{d}{c}) = \infty$. As such, any Möbius transformation can be seen as a biholomorphic function on the extended complex plane. After resolving $w = (az+b)/(cz+d)$ for z, the inverse of (9.24) is given by

$$T^{-1}(z) = \frac{-dz + b}{cz - a}, \tag{9.25}$$

which makes T^{-1} a Möbius transformation itself. Any Möbius transformation is uniquely determined by its action on three points: Provided the triples $(z_1, z_2, z_3) \in \overline{\mathbb{C}}^3$ and $(w_1, w_2, w_3) \in \overline{\mathbb{C}}^3$, each pairwise distinct, there exists exactly one Möbius transformation T with the property $T(z_i) = w_i$ for all $i = 1, 2, 3$. To compute the latter, we note that

$$T(z) := \frac{(z_2 - z)(z_3 - z_1)}{(z_2 - z_1)(z_3 - z)} \tag{9.26}$$

is a Möbius transformation which maps z_1 to 1, z_2 to 0, z_3 to ∞. Accordingly, the same applies to the w_i if we replace z_i by their prescribed images in (9.26). This proves the following result.

Lemma 9.27. *Let (z_1, z_2, z_3) and (w_1, w_2, w_3) be pairwise distinct triples contained in the extended complex plane. Then there exists exactly one Möbius transformation T with the property $T(z_i) = w_i$ for $i = 1, 2, 3$, which is obtained by resolving*

$$\frac{(w_2 - w)(w_3 - w_1)}{(w_2 - w_1)(w_3 - w)} = \frac{(z_2 - z)(z_3 - z_1)}{(z_2 - z_1)(z_3 - z)} \tag{9.27}$$

for w.

If one of the z_i or w_i is infinite, it needs to be canceled before resolving for w. E.g., if $w_3 = \infty$, then the left-hand side of (9.27) evaluates to

$$\frac{(w_2 - w)(1 - \frac{w_1}{w_3})}{(w_2 - w_1)(1 - \frac{w}{w_3})} = \frac{w_2 - w}{w_2 - w_1}.$$

Since (9.24) is the composition of translations, rotations, dilations, and inversions along the unit circle, it holds that any Möbius transformation maps circles to circles if we interpret straight lines as circles that pass through infinity.

Of profound importance in the study of Möbius transformations is the *cross-ratio* of four pairwise distinct points $z_1, z_2, z_3, z_4 \in \overline{\mathbb{C}}$ defined by

$$(z_1, z_2; z_3, z_4) := \frac{(z_3 - z_1)(z_4 - z_2)}{(z_3 - z_2)(z_4 - z_1)}.$$

Note that the cross-ratio is the image of z_1 under the Möbius transformation T that maps z_2 to 1, z_3 to 0, and z_4 to ∞. Due to (9.27), it is invariant under the action of the latter, i.e.,

$$(z_1, z_2; z_3, z_4) = (T(z_1), T(z_2); T(z_3), T(z_4)). \tag{9.28}$$

Moreover, it provides a convenient criterion to check whether four points lie on the same circle. Since (9.26) maps the circle C, uniquely defined by the triple (z_2, z_3, z_4), to the real axis, there holds $z_1 \in \mathrm{C}$ if and only if $T(z_1) = (z_1, z_2; z_3, z_4) \in \mathbb{R}$.

Coming back to the third Zolotarëv problem, the following lemma shows that the Zolotarëv number is invariant under Möbius transformations; see also [BT17, MR20a].

Lemma 9.28. *Let T be a Möbius transformation and $(\mathbb{A}, \mathbb{B})$ a condenser. Then there holds*

$$Z_k(\mathbb{A}, \mathbb{B}) = Z_k(T(\mathbb{A}), T(\mathbb{B})).$$

Proof. Let T denote an arbitrary Möbius transformation. Then there holds

$$Z_k(\mathbb{A}, \mathbb{B}) = \inf_{r \in \mathcal{R}_{k,k}} \frac{\sup\{|r_k(z)| : z \in \mathbb{A}\}}{\inf\{|r_k(z)| : z \in \mathbb{B}\}} = \inf_{r \in \mathcal{R}_{k,k}} \frac{\sup\{|r_k(T^{-1}(z))| : z \in T(\mathbb{A})\}}{\inf\{|r_k(T^{-1}(z))| : z \in T(\mathbb{B})\}}.$$

Since $r_k \circ T^{-1} \in \mathcal{R}_{k,k}$, we find

$$Z_k(\mathbb{A}, \mathbb{B}) = \inf_{r \in \mathcal{R}_{k,k}} \frac{\sup\{|r_k(z)| : z \in T(\mathbb{A})\}}{\inf\{|r_k(z)| : z \in T(\mathbb{B})\}} = Z_k(T(\mathbb{A}), T(\mathbb{B})). \qquad \square$$

Following [BT17, MR20a], our plan is to apply Lemma 9.28 to a Möbius transformation that transplants $(\mathbb{A}, \mathbb{B})$ to the symmetric condenser $([\delta, 1], [-1, -\delta])$ for some $\delta < 1$ suitably chosen. The description of this transformation is provided in the following statement.

Lemma 9.29. *Let $-\infty \le a < b \le 0 < c < d$ and*

$$\delta_{[a,b;c,d]} := \frac{(\sqrt{\gamma} - 1)^2}{\gamma - 1} < 1, \tag{9.29}$$

where $\gamma := (d, a; b, c)$ is the cross-ratio of d, a, b, and c. Then the Möbius transformation $T_{[a,b;c,d]}$ defined by any three of the following conditions

$$T_{[a,b;c,d]}(a) = -1, \quad T_{[a,b;c,d]}(b) = -\delta_{[a,b;c,d]}, \quad T_{[a,b;c,d]}(c) = \delta_{[a,b;c,d]}, \quad T_{[a,b;c,d]}(d) = 1, \tag{9.30}$$

satisfies

$$T_{[a,b;c,d]}([a, b]) = [-1, -\delta_{[a,b;c,d]}], \qquad T_{[a,b;c,d]}([c, d]) = [\delta_{[a,b;c,d]}, 1].$$

Proof. Let $\delta < 1$ be arbitrary but fixed for the moment and define $T_{[a,b;c,d]}$ to be the Möbius transformation uniquely determined by the first three conditions in (9.30) after replacing $\delta_{[a,b;c,d]}$ with δ. We have to determine δ such that $T_{[a,b;c,d]} = 1$ holds. Due to (9.28), the latter is equivalent to

$$(d, a; b, c) = (1, T_{[a,b;c,d]}(a); T_{[a,b;c,d]}(b), T_{[a,b;c,d]}(c)). \tag{9.31}$$

By direct substitution we find that (9.31) is equivalent to

$$\gamma = \frac{(1+\delta)^2}{(1-\delta)^2}.$$

Rearranging the terms, we arrive at the quadratic equation

$$\delta^2 + \delta\frac{2(1+\gamma)}{(1-\gamma)} + 1 = 0.$$

Recalling $\delta < 1$ and $\gamma > 1$, we finally deduce

$$\delta = \frac{(\sqrt{\gamma}-1)^2}{\gamma - 1} = \delta_{[a,b;c,d]}$$

and the conjecture is valid. □

Upon defining

$$\rho_{[a,b]} := e^{-\frac{\pi^2}{\ln(4b/a)}}, \qquad 0 < a < b,$$

we leverage our knowledge about normalized and symmetric condensers to obtain a solution to the third Zolotarëv problem on arbitrary real condensers whose plates are intervals; see also [Leb77, Akh90, BT17, MR20a].

Theorem 9.30. *Let* $\mathbb{A} = [c, d] \subset \mathbb{R}^+$, $\mathbb{B} = [a, b] \subset \mathbb{R}_0^- \cup \{-\infty\}$, $T = T_{[a,b;c,d]}$, $\delta = \delta_{[a,b;c,d]}$, *and* $\left(\mathcal{Z}_j^{(k)}\right)_{j=1}^k$ *the Zolotarëv points on* $[\delta, 1]$. *Then*

$$r_k^*(\lambda) = \prod_{j=1}^k \frac{\lambda - T^{-1}(\mathcal{Z}_j^{(k)})}{\lambda - T^{-1}(-\mathcal{Z}_j^{(k)})} \tag{9.32}$$

solves the third Zolotarëv problem (9.4) *on the condenser* $(\mathbb{A}, \mathbb{B})$ *and*

$$Z_k(\mathbb{A}, \mathbb{B}) \le 4\rho_{[\delta,1]}^{-k}. \tag{9.33}$$

Proof. The inequality (9.33) follows from Lemma 9.28 and (9.23) since

$$Z_k(\mathbb{A}, \mathbb{B}) = Z_k(T(\mathbb{A}), T(\mathbb{B})) = Z_k([\delta, 1], [-1, -\delta]) \le 4e^{-\frac{\pi^2 k}{\ln(4\delta^{-1})}} = 4\rho_{[\delta,1]}^{-k}.$$

To confirm that (9.32) actually minimizes $Z_k(\mathbb{A},\mathbb{B})$, we note that r_k^* has its roots in $\{T^{-1}(\mathcal{Z}_1^{(k)}),\dots,T^{-1}(\mathcal{Z}_k^{(k)})\}$ and its poles in $\{T^{-1}(-\mathcal{Z}_1^{(k)}),\dots,T^{-1}(-\mathcal{Z}_k^{(k)})\}$. The same applies to the rational function

$$\prod_{j=1}^{k}\frac{T(\lambda)-\mathcal{Z}_j}{T(\lambda)+\mathcal{Z}_j}=(\bar{r}_k^*\circ T)(\lambda),$$

where $\bar{r}_k^*$ is the solution to the third Zolotarëv problem on $([\delta,1],[-1,-\delta])$ obtained by Theorem 9.23. Since $\bar{r}_k^*\circ T\in\mathcal{R}_{k,k}$, it follows that $r_k^*=c(\bar{r}_k^*\circ T)$ for some $c\in\mathbb{R}$. This reveals

$$\begin{aligned}\frac{\sup\{|r_k^*(z)|:z\in\mathbb{A}\}}{\inf\{|r_k^*(z)|:z\in\mathbb{B}\}}&=\frac{\sup\{|r_k^*(T^{-1}(z))|:z\in T(\mathbb{A})\}}{\inf\{|r_k^*(T^{-1}(z))|:z\in T(\mathbb{B})\}}\\&=\frac{\sup\{|\bar{r}_k^*(\lambda)|:\lambda\in[\delta,1]\}}{\inf\{|\bar{r}_k^*(\lambda)|:\lambda\in[-1,-\delta]\}}=Z_k([\delta,1],[-1,-\delta])=Z_k(\mathbb{A},\mathbb{B}),\end{aligned}$$

where the last equality follows from Lemma 9.28. This proves the claim. □

Theorem 9.30 allows us to derive explicit minimizer for the problem (9.2) when $\mathbb{B}=\mathbb{R}_0^-$. Our proof closely follows [MR20a, Lemma 4].

Theorem 9.31. *Let* $\kappa=\lambda_{\min}/\lambda_{\max}$, $\delta_{[\lambda_{\min},\lambda_{\max}]}:=2\kappa-1-2\sqrt{\kappa^2-\kappa}$, *and*

$$T_{[\lambda_{\min},\lambda_{\max}]}(z):=\frac{-z(\delta_{[\lambda_{\min},\lambda_{\max}]}+1)+2\lambda_{\max}\delta_{[\lambda_{\min},\lambda_{\max}]}}{z(\delta_{[\lambda_{\min},\lambda_{\max}]}+1)-2\lambda_{\max}}.$$

Then there holds $\delta_{[\lambda_{\min},\lambda_{\max}]}=\delta_{[-\infty,0;\lambda_{\min},\lambda_{\max}]}$, $T_{[\lambda_{\min},\lambda_{\max}]}(z)=T_{[-\infty,0;\lambda_{\min},\lambda_{\max}]}(z)$, *and*

$$T^{-1}_{[\lambda_{\min},\lambda_{\max}]}(z)=\frac{2\lambda_{\max}}{\delta_{[\lambda_{\min},\lambda_{\max}]}+1}\frac{z+\delta_{[\lambda_{\min},\lambda_{\max}]}}{z+1}.$$

In particular,

$$r_k^*(\lambda)=\prod_{j=1}^{k}\frac{\lambda-T^{-1}_{[\lambda_{\min},\lambda_{\max}]}(\mathcal{Z}_j^{(k)})}{\lambda-T^{-1}_{[\lambda_{\min},\lambda_{\max}]}(-\mathcal{Z}_j^{(k)})}\tag{9.34}$$

solves the third Zolotarëv problem on the condenser $(\Sigma,\mathbb{R}_0^-)$ *whenever* $\mathcal{Z}_1^{(k)},\dots,\mathcal{Z}_k^{(k)}$ *are the Zolotarëv points on* $[\delta_{[\lambda_{\min},\lambda_{\max}]},1]$. *There holds*

$$Z_k(\Sigma,\mathbb{R}_0^-)\le 4\rho^{-k}_{[\delta_{[\lambda_{\min},\lambda_{\max}]},1]}\le 4\rho^{-k}_{[\lambda_{\min},4\lambda_{\max}]}.\tag{9.35}$$

Proof. We compute

$$\gamma:=(\lambda_{\max},-\infty;0,\lambda_{\min})=\frac{\lambda_{\max}}{\lambda_{\max}-\lambda_{\min}}.$$

By direct substitution, we find

$$\delta_{[-\infty,0;\lambda_{\min},\lambda_{\max}]} = \frac{(\sqrt{\gamma}-1)^2}{\gamma-1} = \delta_{[\lambda_{\min},\lambda_{\max}]}.$$

Straightforward computations confirm that $T_{[\lambda_{\min},\lambda_{\max}]}$ maps $-\infty$, 0, and $\lambda_{\max}$ to -1, $-\delta_{[\lambda_{\min},\lambda_{\max}]}$, and 1, respectively. Thanks to Lemma, 9.27 we deduce $T_{[\lambda_{\min},\lambda_{\max}]}(z) = T_{[-\infty,0;\lambda_{\min},\lambda_{\max}]}(z)$. The representation formula for $T^{-1}_{[\lambda_{\min},\lambda_{\max}]}$ follows from (9.25). The fact that (9.34) solves the third Zolotarëv problem on $(\Sigma, \mathbb{R}_0^-)$ is a direct consequence of Theorem 9.30. The same applies to the first inequality in (9.35). To prove the latter, we write $\delta_{[\lambda_{\min},\lambda_{\max}]}$ in more convenient form (cf. [MR20a, Lemma 4])

$$\delta_{[\lambda_{\min},\lambda_{\max}]} = \frac{\lambda_{\max} - \lambda_{\max}\sqrt{1-\frac{\lambda_{\min}}{\lambda_{\max}}}}{\lambda_{\max} + \lambda_{\max}\sqrt{1-\frac{\lambda_{\min}}{\lambda_{\max}}}},$$

which follows from elementary computations. Due to $\sqrt{1-x} \le 1 - \frac{x}{2}$ for all $x \in [0,1]$ we deduce

$$\delta_{[\lambda_{\min},\lambda_{\max}]} \ge \frac{\lambda_{\max} - \lambda_{\max}(1-\frac{\lambda_{\min}}{2\lambda_{\max}})}{\lambda_{\max} + \lambda_{\max}(1-\frac{\lambda_{\min}}{2\lambda_{\max}})} \ge \frac{\frac{\lambda_{\min}}{2}}{2\lambda_{\max} - \frac{\lambda_{\min}}{2}} \ge \frac{\lambda_{\min}}{4\lambda_{\max}} = \frac{1}{4\kappa}.$$

The second inequality in (9.35) now follows from $\delta_{[\frac{1}{4\kappa},1]} = \delta_{[\lambda_{\min},4\lambda_{\max}]}$ and the fact that $\rho_{[a,b]}$ decreases as the ratio $\frac{a}{b}$ decreases. □

Remark 9.32. *In its present form, $\delta_{[\lambda_{\min},\lambda_{\max}]}$ is prone to cancellation errors and one should resort to the equivalent representation [MR20a, Lemma 4]*

$$\delta_{[\lambda_{\min},\lambda_{\max}]} = \frac{\lambda_{\max} - \lambda_{\max}\sqrt{1-\frac{\lambda_{\min}}{\lambda_{\max}}}}{\lambda_{\max} + \lambda_{\max}\sqrt{1-\frac{\lambda_{\min}}{\lambda_{\max}}}}.$$

To derive a solution to (9.3), we first compute the solution of the third Zolotarëv problem on the condenser $(\Sigma, -\Sigma)$. To this end, let γ be the cross-ratio of $(\lambda_{\max}, -\lambda_{\max}, -\lambda_{\min}, \lambda_{\min})$. Then by direct substitution we find

$$\gamma = \frac{(\lambda_{\max}+\lambda_{\min})^2}{(\lambda_{\max}-\lambda_{\min})^2}$$

and (9.29) evaluates to

$$\delta_{[-\lambda_{\max},-\lambda_{\min};\lambda_{\min},\lambda_{\max}]} = \frac{\lambda_{\min}}{\lambda_{\max}}.$$

The Möbius transformation $T = T_{[-\lambda_{\max},-\lambda_{\min};\lambda_{\min},\lambda_{\max}]}$ is uniquely determined by three of the four conditions in (9.30) and reads

$$T(z) = \frac{z}{\lambda_{\max}}.$$

Recalling Definition 8.29, we thus derive from Theorem 9.30 that

$$r_k^*(\lambda) = \prod_{j=1}^{k} \frac{\lambda - \lambda_{\max}\mathcal{Z}_j^{(k)}}{\lambda + \lambda_{\max}\mathcal{Z}_j^{(k)}} = r_{\mathcal{Z}}(\lambda), \qquad \mathcal{Z} = \{-\lambda_{\max}\mathcal{Z}_1^{(k)}, \dots, -\lambda_{\max}\mathcal{Z}_k^{(k)}\},$$

solves the third Zolotarëv problem on the condenser $(\Sigma, -\Sigma)$ whenever $\mathcal{Z}_1^{(k)}, \dots, \mathcal{Z}_k^{(k)}$ are the Zolotarëv points on $[\lambda_{\min}/\lambda_{\max}, 1]$. Since $r_{\mathcal{Z}}(-\lambda) = 1/r_{\mathcal{Z}}(\lambda)$, there holds

$$\begin{aligned} Z_k(\Sigma, -\Sigma) &= \frac{\sup\{|r_{\mathcal{Z}}(\lambda)| : \lambda \in \Sigma\}}{\inf\{|r_{\mathcal{Z}}(-\lambda)| : \lambda \in \Sigma\}} = \frac{\sup\{|r_{\mathcal{Z}}(\lambda)| : \lambda \in \Sigma\}}{\inf\{1/|r_{\mathcal{Z}}(\lambda)| : \lambda \in \Sigma\}} \\ &= \sup\{|r_{\mathcal{Z}}(\lambda)|^2 : \lambda \in \Sigma\} = \|r_{\mathcal{Z}}\|_\Sigma^2. \end{aligned}$$

Hence

$$\min_{\substack{\Xi \subset -\Sigma \\ |\Xi| = k}} \|r_\Xi\|_\Sigma \le \|r_{\mathcal{Z}}\|_\Sigma = \sqrt{Z_k(\Sigma, -\Sigma)}. \tag{9.36}$$

Even more, one can show that $r_{\mathcal{Z}}$ satisfies a Chebyshev type alternance property on Σ which allows one to prove that $\mathcal{Z}$ indeed minimizes *Zolotarëv's minimal deviation problem*: Find $\Psi \subset -\Sigma$, $|\Psi| = k$, such that

$$\|r_\Psi\|_\Sigma = \min_{\substack{\Xi \subset -\Sigma \\ |\Xi| = k}} \|r_\Xi\|_\Sigma. \tag{9.37}$$

We collect these findings in one of the main theorems of this section and refer to e.g., [Wac13] for that fact that (9.36) actually holds with equality.

Theorem 9.33. *Let $\delta = \lambda_{\min}/\lambda_{\max}$, $\mathcal{Z}_1^{(k)}, \dots, \mathcal{Z}_k^{(k)}$ the Zolotarëv points on $[\delta, 1]$, and $\mathcal{Z} := \{-\lambda_{\max}\mathcal{Z}_1^{(k)}, \dots, -\lambda_{\max}\mathcal{Z}_k^{(k)}\}$. Then there holds*

$$\|r_{\mathcal{Z}}\|_\Sigma = \min_{\substack{\Xi \subset -\Sigma \\ |\Xi| = k}} \|r_\Xi\|_\Sigma \le 2\rho_{[\lambda_{\min}, \lambda_{\max}]}^{-\frac{k}{2}}.$$

9.3.3 Perpendicular Intervals Parallel to the Axes

The final ingredient to bound the rational Krylov approximation error for functions of fractional diffusion type is the third Zolotarëv problem on the condenser $(\Sigma, i\mathbb{R})$. Such configurations have been studied in [BT00]. More specifically, the authors deal with geometries where the second plate is a line segment parallel to the imaginary axis. After a linear transformation, one can assume that the latter is given by $\mathbb{B} = [-il, il]$ for some $l \in \mathbb{R}^+$ suitably chosen. Let now $r_{\mathcal{Z}}$, as in Theorem 9.33, be the solution to Zolotarëv's minimal deviation problem on Σ. Then $|r_{\mathcal{Z}}(z)| = 1$ for all $z \in i\mathbb{R}$ such that

$$Z_k(\Sigma, \mathbb{B}) \le \frac{\sup\{|r_{\mathcal{Z}}(z)| : z \in \Sigma\}}{\sup\{|r_{\mathcal{Z}}(z)| : z \in \mathbb{B}\}} = \|r_{\mathcal{Z}}\|_\Sigma.$$

In other words, the third Zolotarëv problem of the condenser $(\Sigma, \mathbb{B})$, $\mathbb{B} = [-il, il]$ for some $l \in \mathbb{R}^+$, can be bounded by Zolotarëv's minimal deviation problem. It was discovered in [BT00]

that $r_{\mathcal{Z}}$ satisfies at least necessary optimality conditions, provided that l is sufficiently large. As of yet, it is not known whether $r_{\mathcal{Z}}$ yields the true global minimum of $Z_k(\Sigma, \mathbb{B})$. If $l = \infty$, however, any other rational function with possibly complex poles yields at most a twofold decrease of the error. We summarize these observations in a form that is suitable for the study of our problems and direct the interested reader to [DKZ09, Theorem 4.3] for a proof.

Proposition 9.34. *Let $r_{\mathcal{Z}}$ denote the solution to Zolotarëv's minimal deviation problem on Σ as in Theorem 9.33. Then there holds*

$$\frac{\sqrt{Z_k(\Sigma, -\Sigma)}}{2} \leq Z_k(\Sigma, i\mathbb{R}) \leq \sqrt{Z_k(\Sigma, -\Sigma)}.$$

In particular,

$$Z_k(\Sigma, i\mathbb{R}) \leq 2\rho_{[\lambda_{\min}, \lambda_{\max}]}^{-\frac{k}{2}}.$$

10 Pole Selection Strategies

In Chapter 6 it is shown that a large class of fractional diffusion problems can be approximated by matrix-vector products of the form $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$, where $\mathbf{L} \in \mathbb{R}^{N\times N}$ is the finite element matrix approximation of the differential operator, $\mathbf{b} \in \mathbb{R}^{N\times N}$ a vector, and $f^{\boldsymbol{\tau}}$ a parametric matrix function, such as the fractional power or the generalized Mittag-Leffler function. In Chapter 7, a model order reduction strategy in the form of a rational Krylov method is employed to approximate $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ efficiently. The quality of the surrogate depends on the rational Krylov space $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$ and the way it is extracted from it. Thanks to Theorem 7.16, the latter is quasi-optimal such that the performance of the method effectively hinges on a good choice of the search space, or equivalently, of its poles $\Xi = \{\xi_0, \dots, \xi_k\} \subset \overline{\mathbb{R}} \setminus \Sigma$. This chapter is devoted to the choice of these poles and consists of four parts.

In the first part, we avail ourselves of some pole distributions that are mostly familiar to the experts of the model order reduction community and show their suitability to the approximation of fractional PDEs. For some of these configurations, we rigorously prove exponential convergence rates that are uniform in the parameter $\boldsymbol{\tau}$. For those pole distributions that do not allow for such error bounds, we present, in the course of Section 10.2, a computable *guaranteed upper bound* for the rational Krylov error to assess their quality. Based on the insights gained we propose, in the third part, two novel pole selection algorithms that are competitive with or superior to many of the aforementioned pole sets. The remainder of this chapter underpins our analytical findings by a variety of numerical examples. We perform a detailed parameter study to illuminate the impact of changing values of $\boldsymbol{\tau}$ on the Krylov approximation and provide a systematical comparison of the pole distributions.

10.1 Analysis of Selected Pole Configurations

The precise choice of the pole set Ξ may depend on the matrix $\mathbf{L}$, the vector $\mathbf{b}$, and the function $f^{\boldsymbol{\tau}}$. Ideally, one would like to choose the poles in a way such that the error is small for all possible configurations of $\mathbf{L}$, $\mathbf{b}$, and $f^{\boldsymbol{\tau}}$. This, however, is a difficult or even impossible task. Hence, one is typically obliged to trade off various aspects against each other. The following properties should be incorporated in the selection process of Ξ.

– The dependence of the pole set on the parameter $\boldsymbol{\tau}$: In many practical applications [BOKG+14, SV16, AR19, ACR21], one is interested in querying the solution map $\boldsymbol{\tau} \mapsto f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ for multiple instances of the parameter. If the poles are independent of $\boldsymbol{\tau}$, one can compute the basis $\mathbf{V}$ of $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$, its compression $\mathbf{L}_{k+1}$, and the coordinate vector $\mathbf{V}^\dagger\mathbf{b}$ once and for all in the so-called *offline stage*. During the *online stage*, for each new parameter queried the surrogate is found in the coordinate space by computing $f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1})\mathbf{V}^\dagger\mathbf{b}$ at negligible extra costs.

– The dependence on $\mathbf{L}$ and $\mathbf{b}$: What information must be provided by the user to compute the pole set? Do the poles depend on $\mathbf{L}$ or $\mathbf{b}$?

– The presence of analytical results: In practice, one wishes to choose the smallest integer $k \in \mathbb{N}$ such that the approximation error falls below a user-defined tolerance. This in turn requires the availability of guaranteed upper bounds of the rational Krylov error.

– The performance of the rational Krylov approximation in the limit case: If $\tau \mapsto f^{\tau}(\mathbf{L})\mathbf{b}$ is a multi-query problem with $\boldsymbol{\tau} \in \Theta \subset \mathbb{R}^p$, $p \in \mathbb{N}$, it is desirable to choose the poles in such a way that the approximation error remains uniformly bounded for all values of $\boldsymbol{\tau} \in \Theta$ and does not degenerate e.g., when the fractional parameters approach an integer.

– The presence of hierarchical structures: From a computational point of view, it is desirable to adaptively enrich the rational Krylov space until the sought accuracy is obtained. This, however, is only feasible if the poles form an infinite parameter sequence such that $\mathcal{Q}_k^{\Xi}(\mathbf{L}, \mathbf{b}) \subset \mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ for all $k \in \mathbb{N}$.

Under consideration of the above listed properties, we analyze and compare several pole selection strategies and demonstrate their suitability to the fractional diffusion framework. We emphasize that the originality of this section does not lie so much in the presentation of these poles, for which we have partially drawn inspiration from existing literature, as in their systematic study and rigorous analysis in nonlocal diffusion processes.

10.1.1 Zolotarëv Poles

We start our presentation with poles chosen according to the third Zolotarëv problem. Provided a suitable choice of the condenser, these parameters have proven themselves as excellent poles for RKMs [DKZ09, Güt10, Güt13, DS19, MR20a, DS21, DH21, DHS21]. A selection of plates that perfectly fits the analytical framework presented in Chapter 8 is $\mathbb{A} = \Sigma$ and $\mathbb{B} \in \{-\Sigma, i\mathbb{R}\}$, where $\Sigma = [\lambda_{\min}, \lambda_{\max}]$ denotes the spectral interval of the matrix $\mathbf{L}$. In light of these results, we state one of the integral definitions of this chapter.

Definition 10.1. *Let $k \in \mathbb{N}$, $\delta = \lambda_{\min}/\lambda_{\max}$, and $\mathcal{Z}_1^{(k)}, \dots, \mathcal{Z}_k^{(k)}$ the Zolotarëv points of order k on $[\delta, 1]$. The Zolotarëv poles of order k on $-\Sigma$ are defined by*

$$\mathcal{Z} := \{-\lambda_{\max}\mathcal{Z}_1^{(k)}, \dots, -\lambda_{\max}\mathcal{Z}_k^{(k)}\}.$$

We set $\mathcal{Z}_\infty := \mathcal{Z} \cup \{\infty\}$. Further, let $\delta_{[\lambda_{\min},\lambda_{\max}]}$ and $T_{[\lambda_{\min},\lambda_{\max}]}$ be as in Theorem 9.31 and $\hat{\mathcal{Z}}_1^{(k)}, \dots, \hat{\mathcal{Z}}_k^{(k)}$ the Zolotarëv points of order k on $[\delta_{[\lambda_{\min},\lambda_{\max}]}, 1]$. The Zolotarëv poles of order k on $\mathbb{R}_0^-$ are defined by

$$\hat{\mathcal{Z}} := \{T^{-1}_{[\lambda_{\min},\lambda_{\max}]}(-\hat{\mathcal{Z}}_1^{(k)}), \dots, T^{-1}_{[\lambda_{\min},\lambda_{\max}]}(-\hat{\mathcal{Z}}_k^{(k)})\}$$

and $\hat{\mathcal{Z}}_\infty := \hat{\mathcal{Z}} \cup \{\infty\}$.

The pole set $\mathcal{Z}$ has been analyzed in [DKZ09] for computing matrix exponentials in evolutionary problems, in [DS19] for applying positive fractional powers of differential operators, and in [DS21, DH21] for solving stationary fractional diffusion problems. In a more general framework, $\mathcal{Z}$ and $\hat{\mathcal{Z}}$ have been studied in [MR20a] for Cauchy- and Laplace-Stieltjes functions under the assumption that $\mathbf{L}$ is symmetric. Generalization to positive definite but not necessarily symmetric matrices as well as complete Bernstein functions have been conducted in [DHS21].

What makes $\mathcal{Z}$ and $\hat{\mathcal{Z}}$ attractive choices for functions of fractional diffusion type is the fact that they are completely independent of the parameter $\boldsymbol{\tau}$ and solely require the knowledge of the extremal eigenvalues of $\mathbf{L}$. Thanks to our careful preparations provided in Chapter 8 and 9, we are in position to quantify their performance analytically. Starting with the Zolotarëv poles on $\mathbb{R}_0^-$, we state one of the central results of this chapter.

Theorem 10.2 ($\hat{\mathcal{Z}}_\infty$-pointwise convergence)**.** *Let Θ_{C} be defined as in Proposition 8.22, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+1}^{\hat{\mathcal{Z}}_\infty}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1}=\mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1}=\mathbf{V}f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1})\mathbf{V}^\dagger\mathbf{b}$. Then*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}-\mathbf{u}_{k+1}\| \le 8C_{\boldsymbol{\tau}}\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\|,$$

where

$$C_{\boldsymbol{\tau}} := \begin{cases} \lambda_{\min}^{-s}, & \text{if } f^{\boldsymbol{\tau}}(\lambda)=\lambda^{-s} \text{ and } s\in(0,1),\\ \lambda_{\max}^{s}, & \text{if } f^{\boldsymbol{\tau}}(\lambda)=\lambda^{s} \text{ and } s\in(0,1),\\ E_{\alpha,\beta}(-t^\alpha\lambda_{\min}^s), & \text{if } f^{\boldsymbol{\tau}}(\lambda)=E_{\alpha,\beta}(-t^\alpha\lambda^s) \text{ and } (\alpha,\beta,t,s)\in\Theta_{\mathrm{C}}. \end{cases}$$

Proof. Starting with the case $f^{\boldsymbol{\tau}}(\lambda)=\lambda^{-s}$, we make use of Theorem 8.6 to recall that $f^{\boldsymbol{\tau}}\in\mathcal{CS}$ for all $s\in(0,1)$. As such, we may apply Theorem 8.32 and Theorem 9.31 to find

$$\begin{aligned}\|\mathbf{L}^{-s}\mathbf{b}-\mathbf{u}_{k+1}\| &\le 2\lambda_{\min}^{-s}\|\mathbf{b}\| \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_{\hat{\mathcal{Z}}_\infty}} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(\lambda)|:\lambda\in\mathbb{R}_0^-\}}\\ &\le 2\lambda_{\min}^{-s}\|\mathbf{b}\| \min_{r_k\in\mathcal{P}_k/q_{\hat{\mathcal{Z}}}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)|:\lambda\in\mathbb{R}_0^-\}}\\ &= 2\lambda_{\min}^{-s}Z_k(\Sigma,\mathbb{R}_0^-)\|\mathbf{b}\| \le 8\lambda_{\min}^{-s}\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\|.\end{aligned}$$

If $f^{\boldsymbol{\tau}}(\lambda)=\lambda^s$, we apply Theorem 8.8 to recall $f^{\boldsymbol{\tau}}\in\mathcal{CB}$ for all $s\in(0,1)$. The complete Bernstein part of Theorem 8.32 combined with Theorem 9.31 now reveals

$$\begin{aligned}\|\mathbf{L}^{s}\mathbf{b}-\mathbf{u}_{k+1}\| &\le 2\lambda_{\max}^{s}\|\mathbf{b}\| \min_{r_k\in\mathcal{P}_k/q_{\hat{\mathcal{Z}}}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)|:\lambda\in\mathbb{R}_0^-\}}\\ &= 2\lambda_{\max}^{s}Z_k(\Sigma,\mathbb{R}_0^-)\|\mathbf{b}\| \le 8\lambda_{\max}^{s}\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\|.\end{aligned}$$

To complete the proof of this theorem, we apply Proposition 8.22 to see that $E_{\alpha,\beta}(-t^\alpha\lambda^s)\in\mathcal{CS}$ if $(\alpha,\beta,t,s)\in\Theta_{\mathrm{C}}$. Hence, we may use the first claim in Theorem 8.32 combined with

Theorem 9.31 to finally arrive at

$$\begin{aligned}\|E_{\alpha,\beta}(-t^\alpha \mathbf{L}^s)\mathbf{b} - \mathbf{u}_{k+1}\| &\leq 2E_{\alpha,\beta}(-t^\alpha \lambda_{\min}^s)\|\mathbf{b}\| \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_{\hat{\mathcal{Z}}_\infty}} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \\ &\leq 2E_{\alpha,\beta}(-t^\alpha \lambda_{\min}^s)\|\mathbf{b}\| \min_{r_k\in\mathcal{P}_k/q_{\hat{\mathcal{Z}}}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \\ &\leq 2E_{\alpha,\beta}(-t^\alpha \lambda_{\min}^s) Z_k(\Sigma, \mathbb{R}_0^-)\|\mathbf{b}\| \\ &\leq 8E_{\alpha,\beta}(-t^\alpha \lambda_{\min}^s)\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\|. \end{aligned}$$

□

Remark 10.3. *In the case of $f^\tau(\lambda) \in \{\lambda^{-s}, E_{\alpha,\beta}(-t^\alpha\lambda^s)\}$, Theorem 10.2 remains valid if we extract the rational Krylov surrogate from $\mathcal{Q}_k^{\mathcal{Z}}(\mathbf{L},\mathbf{b})$ instead of $\mathcal{Q}_{k+1}^{\mathcal{Z}_\infty}(\mathbf{L},\mathbf{b})$. For $f^\tau(\lambda) = \lambda^s$, however, the presence of $\mathbf{b}$ in the basis portfolio is essential.*

Including an additional pole at infinity grants uniform convergence rates in all admissible configurations of the parameter when $f^\tau(\lambda) \in \{\lambda^{-s}, \lambda^s\}$.

Theorem 10.4 ($\hat{\mathcal{Z}}_\infty$-uniform convergence)**.** *Assume that $\lambda_{\min} \geq 1$.*

1. *Let $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\hat{\mathcal{Z}}_\infty}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}\mathbf{L}_{k+1}^{-s}\mathbf{V}^\dagger\mathbf{b}$. Then there holds*

$$\max_{s\in[0,1]} \|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \leq 8\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\|.$$

2. *Let $\hat{\mathcal{Z}}_\infty^\infty := \hat{\mathcal{Z}}_\infty \cup \{\infty\}$, $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+2}^{\hat{\mathcal{Z}}_\infty^\infty}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}\mathbf{L}_{k+1}^{s}\mathbf{V}^\dagger\mathbf{b}$. Then there holds*

$$\max_{s\in[0,1]} \|\mathbf{L}^{s}\mathbf{b} - \mathbf{u}_{k+2}\| \leq 8\lambda_{\max}\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\|.$$

Proof. Revisiting Theorem 10.2, we see that, due to $\lambda_{\min}^{-s} \leq 1$, it suffices to prove that

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \leq 8\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\| \tag{10.1}$$

holds in the extremal cases $s = 0, 1$. If $s = 0$, then $\mathbf{L}^{-s}\mathbf{b} = \mathbf{b} = \mathbf{V}\mathbf{V}^\dagger\mathbf{b} = \mathbf{u}_{k+1}$ and the inequality trivially holds true. For $s = 1$, (10.1) directly follows from Theorem 8.25 with $\zeta = 0$ and Theorem 9.31 since

$$\begin{aligned}\|\mathbf{L}^{-1}\mathbf{b} - \mathbf{u}_{k+1}\| &\leq \frac{2}{\lambda_{\min}}\|\mathbf{b}\| \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_{\hat{\mathcal{Z}}_\infty}} \frac{\|r_{k+1}\|_\Sigma}{|r_{k+1}(0)|} \\ &\leq \frac{2}{\lambda_{\min}}\|\mathbf{b}\| \min_{r_k\in\mathcal{P}_k/q_{\hat{\mathcal{Z}}}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \\ &= \frac{2}{\lambda_{\min}} Z_k(\Sigma, \mathbb{R}_0^-)\|\mathbf{b}\| \leq \frac{8}{\lambda_{\min}}\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-k}\|\mathbf{b}\|. \end{aligned}$$

Similarly, $\{\infty,\infty\} \subset \hat{\mathcal{Z}}_\infty^\infty$ implies that the rational Krylov approximation is exact for both $\mathbf{L}^0\mathbf{b} = \mathbf{b}$ and $\mathbf{L}^1\mathbf{b} = \mathbf{L}\mathbf{b}$. Recognizing this fact, the second conjecture now follows from Theorem 10.2. □

The previous two theorems deserve some further discussions. Our analysis shows that $\hat{\mathcal{Z}}_\infty$ provides an excellent pole set for approximating positive and negative fractional powers of differential operators. If $\hat{\mathcal{Z}}_\infty$ is enriched with an additional pole at infinity, which is computationally inexpensive, the error remains uniformly bounded for all admissible values of s. Likewise, the very same rational Krylov space can be employed for approximating space-time fractional diffusion problems as long as $(\alpha, \beta, t, s) \in \Theta_{\mathrm{C}}$. Although we observe numerically that the error bound remains in force if $(\alpha, \beta, t, s) \in \Theta_{\mathrm{L}} \supset \Theta_{\mathrm{C}}$, our analytical tools do not include convergence results in this regime. This inconvenience can be overcome by means of the pole set $\mathcal{Z}$. As preparation, we make the following elementary but crucial observation.

Lemma 10.5. *Let $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V} f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$. Assume that $\Xi \subset \mathbb{R}_0^- \cup \{\infty\}$ contains at least one pole at infinity.*

1. *If $f^{\boldsymbol{\tau}} \in \mathcal{CS}$, then there holds*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2 f^{\boldsymbol{\tau}}(\lambda_{\min}) \|\mathbf{b}\| \|r_\Xi\|_\Sigma. \tag{10.2}$$

2. *Let $f^{\boldsymbol{\tau}} \in \mathcal{CB}$ with ω as in (8.12). If $\omega = 0$, then there holds*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2 f^{\boldsymbol{\tau}}(\lambda_{\max}) \|\mathbf{b}\| \|r_\Xi\|_\Sigma. \tag{10.3}$$

   Assuming $\{\infty, \infty\} \subset \Xi$, then (10.3) holds even if $\omega \neq 0$.

Proof. Since $\infty \in \Xi$ implies $\deg(q_\Xi) \leq k$, there holds $r_\Xi \in \mathcal{P}_k / q_\Xi$. Therefore,

$$\min_{r_k \in \mathcal{P}_k / q_\Xi} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \leq \frac{\|r_\Xi\|_\Sigma}{\inf\{|r_\Xi(\lambda)| : \lambda \in \mathbb{R}_0^-\}} = \|r_\Xi\|_\Sigma,$$

where the equality follows from $|r_\Xi(\lambda)| \geq 1$ for $\lambda \in \mathbb{R}_0^-$. The claim now follows from Theorem 8.32. □

The previous lemma combined with the third claim in Theorem 8.32 shows that the RKM error can be bounded by Zolotarëv's minimal deviation problem irrespectively of $f^{\boldsymbol{\tau}} \in \mathcal{CS} \cup \mathcal{CB} \cup \mathcal{LS}$. This allows us to *simultaneously* approximate solutions to fractional diffusion problems of elliptic and parabolic type.

Theorem 10.6 ($\mathcal{Z}_\infty$-pointwise convergence)**.** *Let Θ_{L} be defined as in Theorem 8.20, $c_{\boldsymbol{\tau}}$ as in Lemma 2.30 with $a = \lambda_{\min}$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+1}^{\mathcal{Z}_\infty}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V} f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$. Then there holds*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 4 C_{\boldsymbol{\tau}} \rho_{[\lambda_{\min}, \lambda_{\max}]}^{-\frac{k}{2}} \|\mathbf{b}\|,$$

where

$$C_{\boldsymbol{\tau}} := \begin{cases} \lambda_{\min}^{-s}, & \text{if } f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s} \text{ and } s \in (0,1), \\ \lambda_{\max}^{s}, & \text{if } f^{\boldsymbol{\tau}}(\lambda) = \lambda^{s} \text{ and } s \in (0,1), \\ \frac{c_{\boldsymbol{\tau}}}{\pi}, & \text{if } f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s) \text{ and } (\alpha, \beta, t, s) \in \Theta_{\mathrm{L}}. \end{cases} \tag{10.4}$$

If $(\alpha, \beta, t, s) \in \Theta_{\mathrm{C}}$, then (10.4) remains valid if we replace $\frac{c_{\boldsymbol{\tau}}}{\pi}$ by $E_{\alpha,\beta}(-t^\alpha \lambda_{\min}^s)$.

Proof. Since $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s} \in \mathcal{CS}$, we may apply the first property in Lemma 10.5 to find

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2\lambda_{\min}^{-s}\|\mathbf{b}\|\|r_{\mathcal{Z}_\infty}\|_\Sigma = 2\lambda_{\min}^{-s}\|\mathbf{b}\|\|r_{\mathcal{Z}}\|_\Sigma.$$

In this case, the claim directly follows from Theorem 9.33. The case $f^{\boldsymbol{\tau}}(\lambda) = \lambda^s$ can be verified in analogously. To complete the proof, let now $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$ and $(\alpha,\beta,t,s) \in \Theta_{\mathrm{L}}$. According to Theorem 8.20 we have $f^{\boldsymbol{\tau}} \in \mathcal{LS}$. Since $f^{\boldsymbol{\tau}}$ extends continuously to the imaginary axis, Theorem 8.34 is applicable, whence we deduce with Lemma 2.30

$$\begin{aligned}
\|E_{\alpha,\beta}(-t^\alpha\mathbf{L}^s)\mathbf{b} - \mathbf{u}_{k+1}\| &\leq \frac{c_{\boldsymbol{\tau}}}{\pi}\|\mathbf{b}\| \min_{r_{k+1}\in\mathcal{P}_{k+1}/q_{\mathcal{Z}_\infty}} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(z)| : z \in i\mathbb{R}\}} \\
&\leq \frac{c_{\boldsymbol{\tau}}}{\pi}\|\mathbf{b}\| \min_{r_k\in\mathcal{P}_k/q_{\mathcal{Z}}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(z)| : z \in i\mathbb{R}\}} \\
&\leq \frac{c_{\boldsymbol{\tau}}}{\pi}\|\mathbf{b}\| \frac{\|r_{\mathcal{Z}}\|_\Sigma}{\inf\{|r_{\mathcal{Z}}(z)| : z \in i\mathbb{R}\}}.
\end{aligned}$$

Since $|r_{\mathcal{Z}}|$ is identically one on the imaginary axis, it follows from Theorem 9.33

$$\|E_{\alpha,\beta}(-t^\alpha\mathbf{L}^s)\mathbf{b} - \mathbf{u}_{k+1}\| \leq \frac{c_{\boldsymbol{\tau}}}{\pi}\|r_{\mathcal{Z}}\|_\Sigma\|\mathbf{b}\| \leq \frac{2c_{\boldsymbol{\tau}}}{\pi}\rho_{[\lambda_{\min},\lambda_{\max}]}^{-\frac{k}{2}}\|\mathbf{b}\|. \qquad \square$$

If $(\alpha,\beta,t,s) \in \Theta_{\mathrm{C}}$, then $E_{\alpha,\beta}(-t^\alpha\lambda^s) \in \mathcal{CS}$ according to Proposition 8.22 and the claim follows in analogy to the case where $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$.

Remark 10.7. *If $f^{\boldsymbol{\tau}}(\lambda) \in \{\lambda^{-s}, E_{\alpha,\beta}(-t^\alpha\lambda^s)\}$, Theorem 10.6 remains valid if the rational Krylov approximation is extracted from $\mathcal{Q}_k^{\mathcal{Z}}(\mathbf{L},\mathbf{b})$ instead of $\mathcal{Q}_{k+1}^{\mathcal{Z}_\infty}(\mathbf{L},\mathbf{b})$; cf. Remark 10.3.*

As the following theorem shows, $\mathcal{Z}$ is the key ingredient to approximate the DEM approximation uniformly in $\boldsymbol{\tau}$ for *all* admissible values of the parameter. This comes at the cost of slightly weakened convergence rates.

Theorem 10.8 ($\mathcal{Z}_\infty$-uniform convergence)**.** *Assume that $\lambda_{\min} \geq 1$.*

1. *Let $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\mathcal{Z}_\infty}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}\mathbf{L}_{k+1}^{-s}\mathbf{V}^\dagger\mathbf{b}$. Then there holds*

$$\max_{s\in[0,1]} \|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \leq 4\rho_{[\lambda_{\min},\lambda_{\max}]}^{-\frac{k}{2}}\|\mathbf{b}\|.$$

2. *Let $\mathcal{Z}_\infty^\infty := \mathcal{Z}_\infty \cup \{\infty\}$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+2}^{\mathcal{Z}_\infty^\infty}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}\mathbf{L}_{k+1}^{s}\mathbf{V}^\dagger\mathbf{b}$. Then there holds*

$$\max_{s\in[0,1]} \|\mathbf{L}^{s}\mathbf{b} - \mathbf{u}_{k+2}\| \leq 4\lambda_{\max}\rho_{[\lambda_{\min},\lambda_{\max}]}^{-\frac{k}{2}}\|\mathbf{b}\|.$$

3. *Let γ_k be defined by Lemma 8.30, $e_{\alpha,\beta}(t,\lambda)$ as in (8.21), $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+1}^{\mathcal{Z}_\infty}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, $\mathbf{u}_{k+1} = \mathbf{V}\,e_{\alpha,\beta}(-t^\alpha,\mathbf{L}_{k+1}^s)\mathbf{V}^\dagger\mathbf{b}$, $\beta_{\min} \in \mathbb{R}^+$, and $\Theta_{\beta_{\min}} := \{(\alpha,\beta,t,s) \in \Theta_{\mathrm{L}} : \beta \geq \beta_{\min}\}$. Then there holds*

$$\max_{\boldsymbol{\tau}\in\Theta_{\beta_{\min}}} \|e_{\alpha,\beta}(-t^\alpha,\mathbf{L}^s)\mathbf{b} - \mathbf{u}_{k+1}\| \leq 8\gamma_k \max\left\{\frac{1}{\Gamma(\beta_{\min})}, 1\right\}\rho_{[\lambda_{\min},\lambda_{\max}]}^{-\frac{k}{2}}\|\mathbf{b}\|. \tag{10.5}$$

Proof. Arguing as in the proof of Theorem 10.4, it suffices to show that

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \leq 4\rho_{[\lambda_{\min},\lambda_{\max}]}^{-\frac{k}{2}}\|\mathbf{b}\| \tag{10.6}$$

for all $s = 0, 1$. If $s = 0$, then the rational Krylov approximation is exact and the inequality trivially holds true. For $s = 1$, (10.6) follows from Theorem 8.25 with $\zeta = 0$ since

$$\|\mathbf{L}^{-1}\mathbf{b} - \mathbf{u}_{k+1}\| \leq \frac{2}{\lambda_{\min}}\frac{\|r_{\mathcal{Z}}\|_{\Sigma}}{|r_{\mathcal{Z}}(0)|}\|\mathbf{b}\| = \frac{2}{\lambda_{\min}}\|r_{\mathcal{Z}}\|_{\Sigma}\|\mathbf{b}\| \leq 4\rho_{[\lambda_{\min},\lambda_{\max}]}^{-\frac{k}{2}}\|\mathbf{b}\|,$$

where the last inequality holds due to Theorem 9.33. The second conjecture follows from the observation that $\{\infty, \infty\} \subset \mathcal{Z}_{\infty}^{\infty}$ implies that the rational Krylov approximation is exact for both $\mathbf{L}^0\mathbf{b} = \mathbf{b}$ and $\mathbf{L}^1\mathbf{b} = \mathbf{L}\mathbf{b}$.

To complete the proof, let $f^{\boldsymbol{\tau}}(\lambda) = \mathrm{e}_{\alpha,\beta}(-t^{\alpha}, \lambda^s)$ with $\boldsymbol{\tau} = (\alpha, \beta, t, s) \in \Theta_{\beta_{\min}}$. According to Theorem 8.20 we have $\mathrm{e}_{\alpha,\beta}(-t^{\alpha}, \lambda^s) \in \mathcal{LS}$ and thus by the third point in Theorem 8.32

$$\|\mathrm{e}_{\alpha,\beta}(-t^{\alpha}, \mathbf{L}^s) - \mathbf{u}_{k+1}\| \leq 4\gamma_k f^{\boldsymbol{\tau}}(0)\|\mathbf{b}\|\|r_{\mathcal{Z}}\|_{\Sigma}.$$

The uniform error bound is now a direct consequence of Theorem 9.33 and the observation that

$$\max_{\boldsymbol{\tau}\in\Theta_{\beta_{\min}}} f^{\boldsymbol{\tau}}(0) = \max\left\{\frac{1}{\Gamma(\beta_{\min})}, 1\right\}. \qquad \square$$

The decay rate obtained by the Zolotarëv poles on $-\Sigma$ entail the factor $\frac{1}{2}$ in the exponent but lead to a better constant $\rho_{[\lambda_{\min},\lambda_{\max}]} \geq \rho_{[\lambda_{\min},4\lambda_{\max}]}$. Comparing these results to the ones obtained by $\hat{\mathcal{Z}}$, we find with $\kappa = \lambda_{\max}/\lambda_{\min}$

$$\begin{aligned}\rho_{[\lambda_{\min},4\lambda_{\max}]} = e^{\frac{\pi^2}{\ln(16\kappa)}} &= e^{\frac{\pi^2}{\ln(16)+\ln(\kappa)}} \\ &> e^{\frac{\pi^2}{\ln(16)+2\ln(\kappa)}} = e^{\frac{\pi^2}{2\ln(4)+2\ln(\kappa)}} = e^{\frac{\pi^2}{2\ln(4\kappa)}} = \rho_{[\lambda_{\min},\lambda_{\max}]}^{\frac{1}{2}}.\end{aligned} \tag{10.7}$$

Therefore, for large condition numbers κ, we have that $\rho_{[\lambda_{\min},4\lambda_{\max}]} \approx \rho_{[\lambda_{\min},\lambda_{\max}]}^2$.

For fixed values of $(\alpha, \beta, t, s) \in \Theta_{\mathrm{L}}$, Theorem 10.6 shows that the rational Krylov approximation error decreases with purely exponential convergence rates when $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^{\alpha}\lambda^s)$. The involved constant $c_{\boldsymbol{\tau}}$, however, deteriorates as either of the parameter approaches zero. Provided that the parameter β is bounded away from zero, (10.5) attests the rational Krylov surrogate uniform convergence in the parameters at the cost of the additional logarithmic factor encoded in γ_k. The degeneration of our upper bound as $\beta_{\min}$ approaches zero, however, is in a sense reasonable since the series representation of $E_{\alpha,\beta}$ diverges in this case.

Remark 10.9. *In order to implement the poles based on Zolotarëv's rational approximation problems, one requires the knowledge of the extremal eigenvalues of* $\mathbf{L}$*. The latter are typically not available, in which case one replaces them by some numerical approximations* $0 < \lambda_L \leq \lambda_{\min} < \lambda_{\max} \leq \lambda_U$ *to build the pole set thereupon. The performance of the surrogate* $\mathbf{u}_{k+1}$ *so obtained deteriorates only logarithmically if the approximations of* $\lambda_{\min}$ *and* $\lambda_{\max}$ *become worse.*

We complete this section with an illustration of the Zolotarëv poles on $-\Sigma$ and $\mathbb{R}_0^-$ with $\Sigma = [1, 1000]$ in Figure 10.1. The elements of $\mathcal{Z}$ are roughly geometrically distributed across the negative spectral interval and accumulate at $-\lambda_{\min}$ and $-\lambda_{\max}$. On the other hand, the Zolotarëv poles on $\mathbb{R}_0^-$ are sampled over the entire negative real line but cluster around $-\Sigma$.

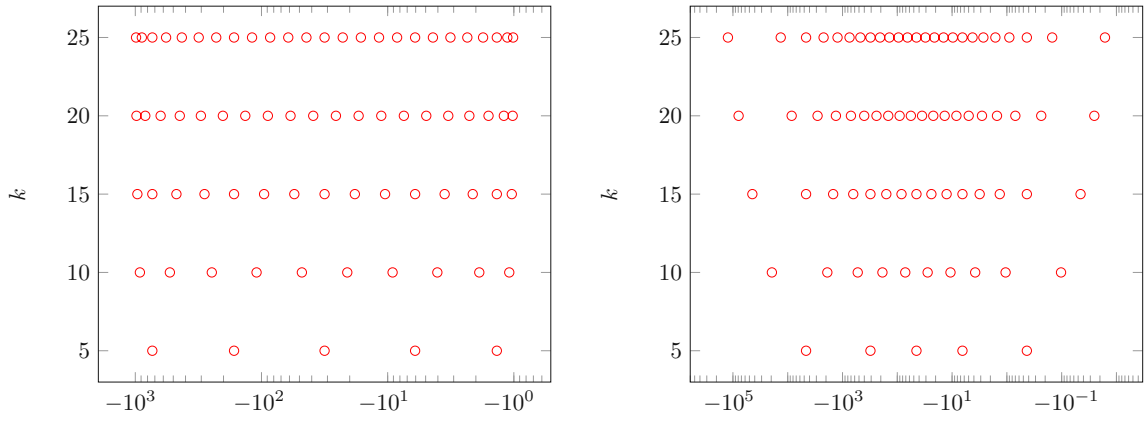

Figure 10.1: Zolotarëv poles on $-\Sigma$ (left) and $\mathbb{R}_0^-$ (right) for $\Sigma = [1, 1000]$ and different orders k.

10.1.2 EDS Poles

In general, Zolotarëv poles of order k have no common element with the Zolotarëv poles of order $k+1$ which is inconvenient if one wishes to incrementally build the rational Krylov space until the desired accuracy is reached. A nested counterpart of the poles provided by $\hat{\mathcal{Z}}$ and $\mathcal{Z}$ has been presented in [DKZ09, MR20a] and is based on the theory of *equidistributed sequences (EDS)*. The idea is to mimic the optimality property of Zolotarëv's poles in an asymptotic sense. More precisely, one makes use of the fact that the squared Zolotarëv points on $[\delta, 1]$, $\delta \in (0, 1)$, are asymptotically distributed according to the measure [DKZ09]

$$\nu_\delta(t) = \frac{1}{2\mathcal{K}(\sqrt{1-\delta^2})} \int_{\delta^2}^{t} \frac{dy}{\sqrt{(y-\delta^2)y(1-y)}} \tag{10.8}$$

with

$$\nu_\delta\left(\left(\mathcal{Z}_j^{(k)}\right)^2\right) = \frac{2(k-j)+1}{2k}, \qquad j = 1, \ldots, k. \tag{10.9}$$

One can try now to replace the right-hand side of (10.9) by an infinite sequence of nodes $(s_j)_{j\in\mathbb{N}} \subset [0, 1]$, which fill the interval in an almost uniform manner, in order to "inversely" describe an approximation for the Zolotarëv points. The weak-star limit of the normalized counting measure associated to $(s_j)_{j\in\mathbb{N}}$ should then be given by the Lebesgue measure on $[0, 1]$ and coincides with the one generated by the right-hand side of (10.9). To create a mathematical framework for this problem, we give the following definition [Cha68].

Definition 10.10. *A sequence $(s_j)_{j\in\mathbb{N}} \subset \mathbb{R}$ is said to be equidistributed on $[a,b] \subset \mathbb{R}$ if*

$$\lim_{j\to\infty} \frac{|\{s_1,\dots,s_j\} \cap [c,d]|}{j} = \frac{d-c}{b-a}$$

for all subintervals $[c,d] \subset [a,b]$.

Provided an EDS $(s_j)_{j\in\mathbb{N}}$, the normalized counting measure associated to the sequence $(\xi_j^2)_{j\in\mathbb{N}} \subset \mathbb{R}$, which we define by the relation

$$\nu_\delta(\xi_j^2) = s_j, \qquad j \in \mathbb{N},$$

is weak-star convergent to (10.8). In other words, the poles $\mathcal{E} := \{-\xi_1,\dots,-\xi_k\}$ are asymptotically distributed like the Zolotarëv poles on $[-1,-\delta]$. This allows one to proof the following result which we have essentially taken from [DKZ09, Theorem 4.4].

Theorem 10.11. *Let $\delta \in (0,1)$, $\mathbb{A} = [\delta,1]$, ν_δ defined by (10.8), and $(s_j)_{j\in\mathbb{N}} \subset [0,1)$ an equidistributed sequence on $[0,1]$. Set*

$$\xi_j := \sqrt{y_j}, \qquad \nu_\delta(y_j) = s_j, \qquad j \in \mathbb{N}, \tag{10.10}$$

and $\mathcal{E} = \{-\xi_1,\dots,-\xi_k\}$. Then the rational function $r_\mathcal{E}$ satisfies

$$\lim_{k\to\infty} \|r_\mathcal{E}\|_\mathbb{A}^{\frac{1}{k}} = \rho_{[\delta,1]}^{-\frac{1}{2}}. \tag{10.11}$$

We call $\xi_1,\dots,\xi_k$ the *EDS points on* $\mathbb{A}$. Note that (10.11) implies

$$\|r_\mathcal{E}\|_\mathbb{A} \preceq \rho_{[\delta,1]}^{-\frac{k}{2}}. \tag{10.12}$$

The roots $y_1,\dots,y_k$ in (10.10) can be computed numerically, e.g., by Newton's method. Asymptotically optimal poles on $-\Sigma$ or $\mathbb{R}_0^-$ are obtained by applying either a scaling or a Möbius transformation to the elements of $\mathcal{E}$, respectively. To complete the description of these poles, it remains to be clarified how one can construct equidistributed sequences explicitly. The following result addresses this matter [Cha68].

Lemma 10.12. *For all $r \in \mathbb{Q}$ the sequence $(s_j)_{j\in\mathbb{N}}$ defined by*

$$s_j := jr - \lfloor jr \rfloor$$

is an equidistributed sequence on $[0,1]$.

In all our experiments we choose $r = 1/\sqrt{2}$ which is why we make it part of the following definition.

Definition 10.13. *Let $\delta = \lambda_{\min}/\lambda_{\max}$ and $\xi_1,\dots,\xi_k$ the EDS points on $[\delta,1]$ generated by the sequence $s_j := j/\sqrt{2} - \lfloor j/\sqrt{2} \rfloor$. We define the EDS poles on $-\Sigma$ as*

$$\mathcal{E} := \{-\lambda_{\max}\xi_1,\dots,-\lambda_{\max}\xi_k\},$$

and set $\mathcal{E}_\infty := \mathcal{E} \cup \{\infty\}$. *Accordingly, let* $\delta_{[\lambda_{\min},\lambda_{\max}]}$ *and* $T_{[\lambda_{\min},\lambda_{\max}]}$ *be as in Theorem 9.31 and* $\hat{\xi}_1, \dots, \hat{\xi}_k$ *the EDS points on* $[\delta_{[\lambda_{\min},\lambda_{\max}]}, 1]$ *generated by* $(s_j)_{j\in\mathbb{N}}$. *We define the EDS poles on* $\mathbb{R}_0^-$ *as*

$$\hat{\mathcal{E}} := \{T^{-1}_{[\lambda_{\max},\lambda_{\max}]}(-\hat{\xi}_1), \dots, T^{-1}_{[\lambda_{\min},\lambda_{\max}]}(-\hat{\xi}_k)\}$$

and set $\hat{\mathcal{E}}_\infty := \hat{\mathcal{E}} \cup \{\infty\}$.

Only one new pole is added at each stage to the k previously selected parameters which are left unchanged. This makes $\mathcal{E}$ and $\hat{\mathcal{E}}$ an attractive *nested* alternative to $\mathcal{Z}$ and $\hat{\mathcal{Z}}$, respectively. Similarly to their competitors, pole sets based on EDS are independent of the parameter $\boldsymbol{\tau}$, the vector $\mathbf{b}$, and solely require some rough spectral bounds for the extremal eigenvalues of $\mathbf{L}$. Thanks to Theorem 10.11, their performance can be quantified as follows.

Theorem 10.14. *Let* $\mathbf{V}$ *be an orthonormal basis of* $\mathcal{Q}^{\Xi}_{k+1}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L}\mathbf{V}$, $\mathbf{u}_{k+1} = \mathbf{V} f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}$, $f^{\boldsymbol{\tau}} \in \{\lambda^{-s}, \lambda^s\}$ *with* $s \in (0,1)$ *or* $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s)$ *with* $(\alpha, \beta, t, s) \in \Theta_{\mathrm{L}}$.

1. *If* $\Xi = \mathcal{E}_\infty$, *then there holds*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \rho^{-\frac{k}{2}}_{[\lambda_{\min},\lambda_{\max}]} \|\mathbf{b}\|.$$

2. *If* $\Xi = \hat{\mathcal{E}}_\infty$ *and* $f^{\boldsymbol{\tau}}(\lambda) \in \{\lambda^{-s}, \lambda^s\}$ *then*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \rho^{-k}_{[\lambda_{\min},4\lambda_{\max}]} \|\mathbf{b}\|. \tag{10.13}$$

Moreover, (10.13) *remains valid for* $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s)$ *if* $(\alpha, \beta, t, s) \in \Theta_{\mathrm{C}}$.

Proof. We start with $\Xi = \mathcal{E}_\infty$. If $f^{\boldsymbol{\tau}}(\lambda) \in \{\lambda^{-s}, \lambda^s\}$, it follows from Lemma 10.5 that

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \|r_{\mathcal{E}}\|_\Sigma \|\mathbf{b}\|.$$

Let now $\bar{\mathcal{E}}$ denote the EDS poles on $-\mathbb{A} = [-1, -\delta]$ with $\delta = \lambda_{\min}/\lambda_{\max}$. Observing that $r_{\bar{\mathcal{E}}} \circ T = r_{\mathcal{E}}$ if $T(z) = z/\lambda_{\max}$, it follows from (10.12)

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \|r_{\bar{\mathcal{E}}} \circ T\|_\Sigma \|\mathbf{b}\| = \|r_{\bar{\mathcal{E}}}\|_{\mathbb{A}} \|\mathbf{b}\| \preceq \rho^{-\frac{k}{2}}_{[\delta,1]} \|\mathbf{b}\| = \rho^{-\frac{k}{2}}_{[\lambda_{\min},\lambda_{\max}]} \|\mathbf{b}\|.$$

If $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s) \in \mathcal{LS}$, we may apply Theorem 8.34 and Lemma 2.30 to find

$$\begin{aligned}
\|E_{\alpha,\beta}(-t^\alpha \mathbf{L}^s)\mathbf{b} - \mathbf{u}_{k+1}\| &\preceq \|\mathbf{b}\| \min_{r_{k+1} \in \mathcal{P}_{k+1}/q_{\mathcal{E}_\infty}} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(z)| : z \in i\mathbb{R}\}} \\
&\leq \|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_{\mathcal{E}}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(z)| : z \in i\mathbb{R}\}} \\
&\leq \|\mathbf{b}\| \frac{\|r_{\mathcal{E}}\|_\Sigma}{\inf\{|r_{\mathcal{E}}(z)| : z \in i\mathbb{R}\}} \leq \|\mathbf{b}\| \|r_{\mathcal{E}}\|_\Sigma
\end{aligned}$$

since $|r_{\mathcal{E}}(z)| = 1$ for all $z \in i\mathbb{R}$ and the proof follows just like in the previous case. If $\Xi = \hat{\mathcal{E}}_\infty$, then $f^\tau \in \mathcal{CS} \cup \mathcal{CB}$ under the given restrictions on the parameters. We may thus consult Theorem 8.32 to find

$$\begin{aligned}\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| &\preceq \|\mathbf{b}\| \min_{r_{k+1} \in \mathcal{P}_{k+1}/q_{\hat{\mathcal{E}}_\infty}} \frac{\|r_{k+1}\|_\Sigma}{\inf\{|r_{k+1}(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \\ &\leq \|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_{\hat{\mathcal{E}}}} \frac{\|r_k\|_\Sigma}{\inf\{|r_k(\lambda)| : \lambda \in \mathbb{R}_0^-\}}.\end{aligned}$$

Let now $\tilde{\mathcal{E}}$ denote the EDS poles on $-\mathbb{A} = [-1, -\delta_{[\lambda_{\min},\lambda_{\max}]}]$. Noting that $r_{\tilde{\mathcal{E}}} \circ T_{[\lambda_{\min},\lambda_{\max}]} \in \mathcal{P}_k/q_{\hat{\mathcal{E}}}$, we deduce from the mapping properties of $T_{[\lambda_{\min},\lambda_{\max}]}$

$$\begin{aligned}\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| &\preceq \|\mathbf{b}\| \frac{\|r_{\tilde{\mathcal{E}}} \circ T_{[\lambda_{\min},\lambda_{\max}]}\|_\Sigma}{\inf\{|(r_{\tilde{\mathcal{E}}} \circ T_{[\lambda_{\min},\lambda_{\max}]})(\lambda)| : \lambda \in \mathbb{R}_0^-\}} \\ &= \frac{\|r_{\tilde{\mathcal{E}}}\|_{\mathbb{A}}}{\inf\{|r_{\tilde{\mathcal{E}}}(\lambda)| : \lambda \in -\mathbb{A}\}} = \|r_{\tilde{\mathcal{E}}}\|_{\mathbb{A}},\end{aligned}$$

since $|r_{\tilde{\mathcal{E}}}(\lambda)| \geq 1$ for all $\lambda \in \mathbb{R}_0^-$. The claim now follows from (10.12) and the second inequality in (9.35). □

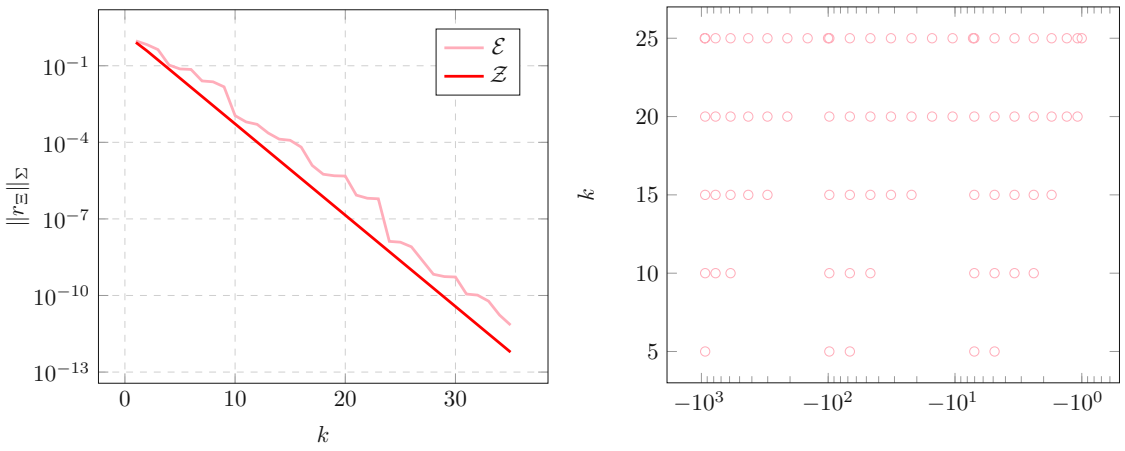

Figure 10.2: Maximum norm $\|r_\Xi\|_\Sigma$ on $\Sigma = [1, 1000]$ for $\Xi \in \{\mathcal{E}, \mathcal{Z}\}$ (left) and EDS poles on $-\Sigma$ for various k (right).

The price one has to pay for the nested structure of the EDS poles is that explicit constants in their estimates are not available. We observe experimentally, however, that already for small values of k, $\|r_{\mathcal{E}}\|_\Sigma$ provides a decent approximation to $\|r_{\mathcal{Z}}\|_\Sigma$. To see this, we plot the maximal deviation of r_Ξ on $\Sigma = [1, 1000]$ for $\Xi \in \{\mathcal{E}, \mathcal{Z}\}$ in Figure 10.2. From the very beginning, $\|r_{\mathcal{E}}\|_\Sigma$ behaves qualitatively very similar to the true minimizer $\|r_{\mathcal{Z}}\|_\Sigma$ and decays like $\mathcal{O}(\rho_{[1,1000]}^{-k/2})$. One can thus expect $\mathcal{E}$ and $\hat{\mathcal{E}}$ to provide a competitive *nested* alternative to $\mathcal{Z}$ and $\hat{\mathcal{Z}}$, respectively.

Before we proceed with the so-called spectral poles, we plot the point set $\mathcal{E}$ for $\Sigma = [1, 1000]$ in Figure 10.2. The EDS poles exhibit a repeated pattern that, after a few iterations, is qualitatively similar to the one obtained by the Zolotarëv poles.

10.1.3 Spectral Poles

In some sense, pole sets based on the third Zolotarëv problem are built on the assumption that the eigenvalues of $\mathbf{L}$ are uniformly distributed across Σ. If the spectral density of the operator is nonuniform, e.g., when the eigenvalues accumulate at an end point of Σ, more refined pole configurations might exist. In this section, we discuss two pole generation algorithms that include spectral information about the particular matrix $\mathbf{L}$. The first one has been proposed in [GK13] and is closely related to the method described in [DLZ10, DS11]. Its main idea is to quantify the error in terms of a rational function involving the poles and rational Ritz values of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$. Instrumental for this approach is the so-called *Hermite-Walsh formula for rational interpolants*; see [Wal60, Theorem VIII.2], [BR09, BG12, Güt13, Tre19].

Theorem 10.15. *Let $\mathcal{C} \subset \mathbb{C}$ be an integration contour, f a function that is analytic in $\mathrm{Int}(\mathcal{C})$ and extends continuously to $\mathcal{C}$, $r_k^f \in \mathcal{R}_{n,k}$ a rational function with poles in $\Xi = \{\xi_0, \dots, \xi_k\} \subset \overline{\mathbb{C}}$ that interpolates f in the nodes $\Lambda = \{\sigma_0, \dots, \sigma_n\} \subset \overline{\mathbb{C}}$, and $r_{\Lambda,\Xi}$ defined by (8.26). Then there holds for all $z \in \mathrm{Int}(\mathcal{C})$*

$$f(z) - r_k^f(z) = \frac{1}{2\pi i} \int_{\mathcal{C}} \frac{r_{\Lambda,\Xi}(z)}{r_{\Lambda,\Xi}(\zeta)} \frac{f(\zeta)}{z - \zeta} \, d\zeta. \tag{10.14}$$

Provided that $f^{\boldsymbol{\tau}} \in \mathcal{CS}$ with Cauchy-Stieltjes triple $(0, 0, \mu_C)$, formula (10.14) remains valid if we bend the contour to the negative real axis such that, after the transformation $\zeta \mapsto -\zeta$, we obtain

$$f^{\boldsymbol{\tau}}(\lambda) - r_k^{f^{\boldsymbol{\tau}}}(\lambda) = \int_0^{\infty} \frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)} \frac{\mu_C^{\boldsymbol{\tau}}(\zeta)}{\lambda + \zeta} \, d\zeta, \qquad \lambda \in \mathbb{R}^+. \tag{10.15}$$

Let now $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ and $\mathbf{L}_{k+1}$ its compression. Then by Theorem 7.14, there holds for $\mathbf{u}_{k+1} = \mathbf{V} f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1}) \mathbf{V}^{\dagger} \mathbf{b}$

$$\mathbf{u}_{k+1} = r_k^{\boldsymbol{\tau}}(\mathbf{L}) \mathbf{b},$$

where $r_k^{\boldsymbol{\tau}} \in \mathcal{P}_k / q_{\Xi}$ interpolates $f^{\boldsymbol{\tau}}$ in the rational Ritz values $\Lambda = \{\mu_0^{(k)}, \dots, \mu_k^{(k)}\}$ of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$. We apply (10.15) to deduce

$$\begin{aligned} \| f^{\boldsymbol{\tau}}(\mathbf{L}) \mathbf{b} - \mathbf{u}_{k+1} \| &= \left\| \int_0^{\infty} \mu_C^{\boldsymbol{\tau}}(\zeta) (\mathbf{L} + \zeta \mathbf{I})^{-1} \frac{r_{\Lambda,\Xi}(\mathbf{L}) \mathbf{b}}{r_{\Lambda,\Xi}(-\zeta)} \, d\zeta \right\| \\ &\leq \| r_{\Lambda,\Xi}(\mathbf{L}) \mathbf{b} \| \left\| \int_0^{\infty} \frac{\mu_C^{\boldsymbol{\tau}}(\zeta)}{|r_{\Lambda,\Xi}(-\zeta)|} (\mathbf{L} + \zeta \mathbf{I})^{-1} \, d\zeta \right\|. \end{aligned} \tag{10.16}$$

The goal is now to choose the next pole ξ_{k+1} in a way such that the upper bound provided by (10.16) is minimized. According to Proposition 7.15, $r_{\Lambda,\Xi}$ already minimizes $\| r_k(\mathbf{L}) \mathbf{b} \|$ among all monic rational functions with prescribed denominator q_{Ξ}. Hence, it suffices to solely focus on the integrand and we wish to choose the next pole $\xi_{k+1} \in \mathbb{R}_0^-$ such that $|r_{\Lambda,\Xi}(-\zeta)|$ becomes as large as possible on $\mathbb{R}^+$. For this purpose, the authors of [GK13]

proposed to place the new pole at the location of a global minimizer of $|r_{\Lambda,\Xi}(-\zeta)|$, i.e.,

$$\xi_{k+1} := -\operatorname*{arg\,min}_{\zeta\in\mathbb{R}_0^+} |r_{\Lambda,\Xi}(-\zeta)| = \operatorname*{arg\,min}_{\zeta\in\mathbb{R}_0^-} |r_{\Lambda,\Xi}(\zeta)| = \operatorname*{arg\,min}_{\zeta\in\mathbb{R}_0^-} \prod_{\substack{j=0\\ \xi_j\neq\infty}}^{k} \left|\frac{\zeta-\mu_j^{(k)}}{\zeta-\xi_j}\right|. \tag{10.17}$$

Typically, one does not compute ξ_{k+1} by minimizing (10.17) over the parameter continuum but instead over a discrete training set. For bounded parameter domains, it is common to use a uniform grid which is adaptively refined to ensure that no local minima are skipped. Over unbounded domains, as it is the case here, the situation is more delicate. One possibility is to introduces two cut-off parameters $\mathrm{n_c^-}, \mathrm{n_c^+} \in \mathbb{R}^+$ and discretize the bounded domain $[-\mathrm{n_c^-}, \mathrm{n_c^+}]$ under a uniform grid. A discretization for $\mathbb{R}_0^-$ is then obtained by the exponential map $\zeta \mapsto -e^{\zeta}$. Since we use such a discretization of $\mathbb{R}_0^-$ in any of our numerical experiments, we incorporate it in the following definition.

Definition 10.16. *Let* $\mathrm{n_c^-}, \mathrm{n_c^+} \in \mathbb{R}^+$ *and* $\mathcal{T}_{\mathrm{train}}^{\mathrm{n_c^\pm}} \subset \mathbb{R}_0^-$ *a training set of* $-[e^{-\mathrm{n_c^-}}, e^{\mathrm{n_c^+}}]$. *The pole set* $\hat{\mathcal{S}}_\infty = \{\xi_0, \dots, \xi_k\}$ *is defined inductively by* $\xi_0 := \infty$ *and*

$$\xi_{k+1} := \operatorname*{arg\,min}_{\zeta\in\mathcal{T}_{\mathrm{train}}^{\mathrm{n_c^\pm}}} |r_{\Lambda,\Xi}(\zeta)|, \tag{10.18}$$

where $\Lambda = \{\mu_0^{(k)}, \dots, \mu_k^{(k)}\}$ *are the rational Ritz values of* $\mathbf{L}$ *on* $\mathcal{Q}_{k+1}^{\hat{\mathcal{S}}_\infty}(\mathbf{L}, \mathbf{b})$. *We call* $\hat{\mathcal{S}} := \{\xi_1, \dots, \xi_k\}$ *the spectral poles on* $\mathbb{R}_0^-$.

Definition 10.16 provides a promising pole candidate whenever $f^{\boldsymbol{\tau}} \in \mathcal{CS}$. Let now $f^{\boldsymbol{\tau}} \in \mathcal{CB}$ and $r_k^{f^{\boldsymbol{\tau}}} \in \mathcal{R}_{n,k}$ a rational function with poles in $\Xi = \{\xi_0, \dots, \xi_k\}$ that interpolates $f^{\boldsymbol{\tau}}$ in $\Lambda = \{\sigma_0, \dots, \sigma_n\}$. Then $r_k^{f^{\boldsymbol{\tau}}}(\lambda)/\lambda$ interpolates $f^{\boldsymbol{\tau}}(\lambda)/\lambda$ in Λ. Since $f^{\boldsymbol{\tau}}(\lambda)/\lambda \in \mathcal{CS}$ by Proposition 8.9, we may apply (10.15) to confirm that

$$\frac{f^{\boldsymbol{\tau}}(\lambda)}{\lambda} - \frac{r_k^{f^{\boldsymbol{\tau}}}(\lambda)}{\lambda} = \int_0^\infty \frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)} \frac{\mu_C^{\boldsymbol{\tau}}(\zeta)}{\lambda+\zeta}\, d\zeta, \tag{10.19}$$

where $\mu_C^{\boldsymbol{\tau}}$ is the Cauchy-Stieltjes density of $f^{\boldsymbol{\tau}}(\lambda)/\lambda$. Multiplying (10.19) by λ reveals

$$f^{\boldsymbol{\tau}}(\lambda) - r_k^{f^{\boldsymbol{\tau}}}(\lambda) = \int_0^\infty \frac{r_{\Lambda,\Xi}(\lambda)}{r_{\Lambda,\Xi}(-\zeta)} \mu_C^{\boldsymbol{\tau}}(\zeta) \frac{\lambda}{\lambda+\zeta}\, d\zeta, \tag{10.20}$$

which can be seen as a variant of the Hermite-Walsh formula for complete Bernstein functions. Therefore, the same arguments as in the Cauchy-Stieltjes case can be applied to conclude that $\hat{\mathcal{S}}$ provides a promising pole candidate for all $f^{\boldsymbol{\tau}} \in \mathcal{CB}$.

In accordance with poles based on the third Zolotarëv problem, a sampling of Ξ over the entire negative real line does not seem natural when Laplace-Stieltjes functions are involved. In view of (8.29) a more intuitive selection of the training set could be $\mathcal{T}_{\mathrm{train}}^{\mathrm{n_c^\pm}} \subset i\mathbb{R}$ in (10.18) which has the disadvantage of complex poles even though both $\mathbf{L}$ and $\mathbf{b}$ are real. A different approach which circumvents this difficulty has been presented in [DLZ10]. Originally proposed for the exponential function, the authors advocate to sample the poles

over $-\Sigma$, which requires the discretization of a bounded domain only. This restriction can be justified for all functions of Stieltjes and complete Bernstein type since Lemma 10.5 and the third point in Theorem 8.32 show that the rational Krylov error can be bounded by $\|r_\Xi\|_\Sigma$, whose minimum is attained by the Zolotarëv poles $\mathcal{Z}$ on $-\Sigma$. The latter are contained in $-\Sigma$ and thus justify the restriction to this bounded domain. These arguments provide the necessary motivation to introduce the following variant of Definition 10.16.

Definition 10.17. *Let $\mathcal{T}_{\text{train}}$ be a training set of $-\Sigma$. The pole set $\mathcal{S}_\infty = \{\xi_0, \dots, \xi_k\}$ is defined inductively by $\xi_0 := \infty$ and*

$$\xi_{k+1} := \operatorname*{arg\,min}_{\zeta \in \mathcal{T}_{\text{train}}} |r_{\Lambda,\Xi}(\zeta)|.$$

We call $\mathcal{S} := \{\xi_1, \dots, \xi_k\}$ the spectral poles on $-\Sigma$.

In view of (10.17), $\mathcal{S}$ allows the interpretation as greedy approximation of the problem: Find $\Psi \subset -\Sigma$ with $|\Psi| = k$ such that

$$\|r_{\Lambda,\Psi}\|_\Sigma = \min_{\substack{\Xi \subset -\Sigma \\ |\Xi| = k}} \|r_{\Lambda,\Xi}\|_\Sigma, \tag{10.21}$$

where $\Lambda = \{\mu_0^{(k)}, \dots, \mu_k^{(k)}\}$ are the rational Ritz values of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^\Xi(\mathbf{L}, \mathbf{b})$. This can be viewed as a variation of Zolotarëv's minimal deviation problem (9.37) where the numerator of r_Ξ is replace by the characteristic polynomial of $\mathbf{L}_{k+1}$ for better adjustments towards the true spectrum of $\mathbf{L}$. In view of the fact that the Zolotarëv poles are build on the assumption that the spectral density of $\mathbf{L}$ is uniform, it can be expected that the minimizer of (10.21) does not significantly differ from the solution to Zolotarëv's original minimal deviation problem whenever $\mathbf{L}$ satisfies such a property.

Remark 10.18. *Due to the close relation between $\mathcal{S}$ and $\mathcal{Z}$, an alternative strategy to define $\hat{\mathcal{S}}$ could be to greedily minimize (10.21) over $\mathcal{T}_{\text{train}} \subset [-1, -\delta_{[\lambda_{\min},\lambda_{\max}]}]$, with $\delta_{[\lambda_{\min},\lambda_{\max}]}$ as in Theorem 9.31, and transplant the poles so obtained to the entire negative real line using the Möbius transformation $T_{[\lambda_{\min},\lambda_{\max}]}$. We do not further pursue this approach.*

To confirm experimentally that the pole set $\mathcal{S}$ may provide a reasonable approximation to Zolotarëv's minimal deviation problem, we compare the maximal deviation of $r_\mathcal{S}$ on $\Sigma = [1, 1000]$ to the one obtained by the minimizer $r_\mathcal{Z}$ for two different matrices of dimension $N = 1000$. The first matrix is defined by $\mathbf{L}_1 := \operatorname{diag}(1, \dots, N) \in \mathbb{R}^{N \times N}$ such that its spectrum is uniformly distributed across Σ. In this case, Figure 10.3 shows that the pole set $\mathcal{S}$ provides a decent approximation to the true solution $\mathcal{Z}$. The situation is rather different if the eigenvalues do not satisfy such a uniform pattern. The right plot in Figure 10.3 depicts $\|r_\mathcal{S}\|_\Sigma$ for the diagonal matrix $\mathbf{L}_2 \in \mathbb{R}^{N \times N}$ with $\lambda_{\min} = 1$, $\lambda_{\max} = 1000$, and 998 towards 1 geometrically refined eigenvalues on $[1, 2]$. In this regime, the rational function $r_\mathcal{S}$ does not become uniformly small on Σ since the spectrum of $\mathbf{L}_2$ biases the selection of the poles.

To visualize the impact of the spectral density on the parameter selection, we plot the spectral poles on $-\Sigma$ for different orders in Figure 10.4 for the matrices $\mathbf{L}_1$ and $\mathbf{L}_2$. In the

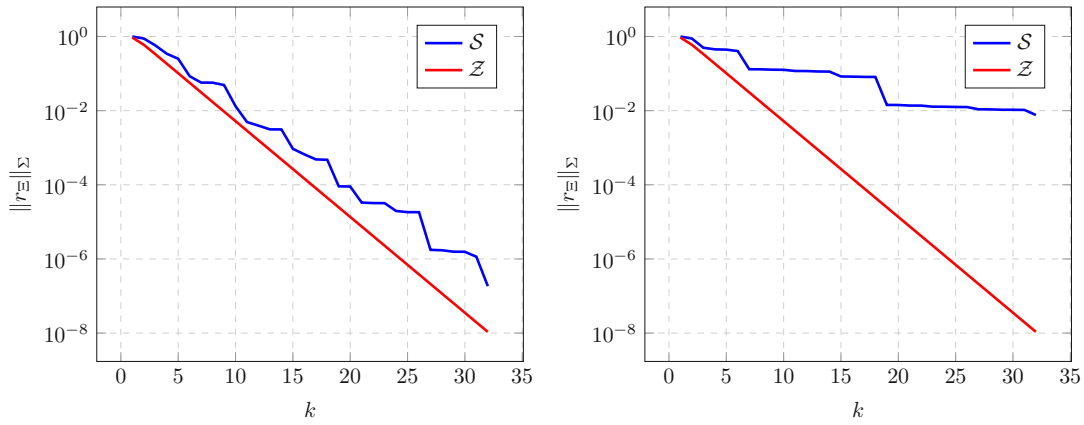

Figure 10.3: Maximum norm $\|r_\Xi\|_\Sigma$ on $\Sigma = [1, 1000]$ with $\Xi \in \{\mathcal{S}, \mathcal{Z}\}$ for the matrices $\mathbf{L}_1$ (left) and $\mathbf{L}_2$ (right).

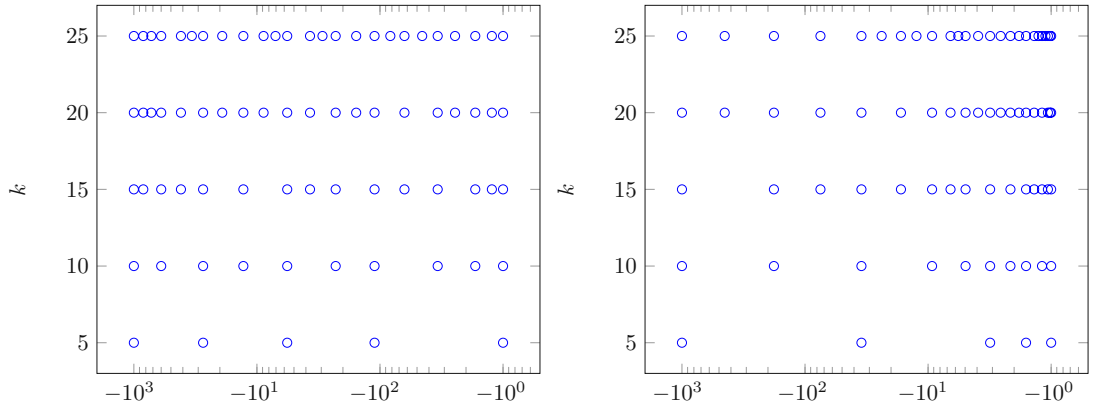

Figure 10.4: Spectral poles $\mathcal{S}$ on $-\Sigma = [-1000, -1]$ for different orders k and the matrices $\mathbf{L}_1$ (left) and $\mathbf{L}_2$ (right).

uniform case, we observe that $\mathcal{S}$ exhibits a similar pattern to the one depicted in Figure 10.1 for the Zolotarëv poles on $-\Sigma$. In contrast, the poles computed for the matrix $\mathbf{L}_2$ accumulate at $-\lambda_{\min}$ which can be traced back to the cluster of eigenvalues of $\mathbf{L}_2$ which is located there.

From a computational point of view, both $\mathcal{S}$ and $\hat{\mathcal{S}}$ provide an attractive choice of poles as they are independent of the parameter $\boldsymbol{\tau}$ and the vector $\mathbf{b}$. While $\mathcal{S}$ avoids the difficulty of discretizing an unbounded domain, it requires, unlike $\hat{\mathcal{S}}$, rough estimates of the spectral region of $\mathbf{L}$. Even though the rational Ritz values need to be recomputed in each step, which makes spectral poles computationally more demanding than their Zolotarëv competitors, the parameters $\xi_0, \ldots, \xi_k$ remain unchanged and are thus convenient for adaptively building the rational Krylov space. Moreover, in contrast to the poles based on the third Zolotarëv problem, the presence of $(\mu_j^{(k)})_{j=0}^k$ allows for better adjustments towards the true discrete spectrum of $\mathbf{L}$. Even though no analytical results are available, there is empirical evidence that such spectral methods outperform both Zolotarëv and EDS poles whenever the operator

exhibits a strong nonuniform spectral density.

10.1.4 Weak Greedy Poles

The pole sets presented up to this point exhibit different degrees of adaption to the data of the problem. Zolotarëv and EDS poles are the most general ones in the sense that they only require some rough bounds on the extremal eigenvalues of $\mathbf{L}$. Spectral poles are tailored to the particular matrix and differ even if two matrices share the same extremal eigenvalues. For poles chosen according to so-called *weak greedy algorithms*, we go one step further and adjust the poles not only to the matrix but also to the vector $\mathbf{b}$. The selection of these parameters is very popular in the reduced basis literature to alleviate the computational costs when evaluating solutions to parametric PDEs but has not attracted as much attention in RKMs yet. We refer to [DPW13, CD15] for a general survey over weak greedy algorithms and [ACN19, BGZ20, DS21, DAC+21, DH21, DH21] for their application to fractional diffusion problems. Using $f^{\boldsymbol{\tau}} \in \mathcal{CS}$ as a starting point, one can bound the rational Krylov approximation error by the triangle inequality

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq \int_0^\infty \mu_C^{\boldsymbol{\tau}}(\zeta)\|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\|\, d\zeta, \tag{10.22}$$

where $\mathbf{V}$ is an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$ and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$. Unlike in the previous two sections, we do not apply spectral arguments to bound (10.22) by a scalar rational approximation problem. Instead, the goal is to directly minimize this upper bound by choosing each consecutive pole according to

$$\xi_{k+1} := -\operatorname*{arg\,max}_{\zeta\in\mathbb{R}_0^+} E_{k+1}(\zeta), \quad E_{k+1}(\zeta) := \|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\|. \tag{10.23}$$

Clearly, (10.23) cannot be realized in practice. Even if we replace $\mathbb{R}_0^+$ with a discrete training set $\mathcal{T}_{\text{train}}^{\text{n}_\text{c}^\pm} \subset \mathbb{R}_0^+$ of finite cardinality, the scheme requires the evaluation of $(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b}$ for multiple values of ζ which is precisely what weak greedy algorithms seek to avoid. To counteract this, the idea is to replaces the true approximation error with a computable quantity which allows the selection to be done in a practically feasible manner while retaining the same performance as (10.23). The following lemma is dedicated to the identification of such an error indicator; see also [DAC+21, Lemma 5.1].

Lemma 10.19. *Let $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\zeta \in \mathbb{R}_0^+$. Then there holds*

$$\frac{1}{\lambda_{\max}+\zeta}\|\mathbf{r}_{k+1}(\zeta)\| \leq E_{k+1}(\zeta) \leq \frac{1}{\lambda_{\min}+\zeta}\|\mathbf{r}_{k+1}(\zeta)\|, \tag{10.24}$$

where $\mathbf{r}_{k+1}(\zeta) := \mathbf{b} - (\mathbf{L}+\zeta\mathbf{I})\mathbf{V}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}$ is the residual.

Proof. Using $E_{k+1}(\zeta) = \|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{r}_{k+1}(\zeta)\|$, we deduce

$$E_{k+1}(\zeta) \leq \|(\mathbf{L}+\zeta\mathbf{I})^{-1}\|\|\mathbf{r}_{k+1}(\zeta)\| = \frac{1}{\lambda_{\min}+\zeta}\|\mathbf{r}_{k+1}(\zeta)\|,$$

where the norms are understood as the vector and the associated matrix norm, respectively. This proves the second inequality in (10.24). On the other hand, there holds $\mathbf{r}_{k+1}(\zeta) = (\mathbf{L}+\zeta\mathbf{I})((\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1}+\zeta\mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b})$. Hence,

$$\|\mathbf{r}_{k+1}(\zeta)\| \leq \|\mathbf{L}+\zeta\mathbf{I}\|E_{k+1}(\zeta) \leq (\lambda_{\max}+\zeta)E_{k+1}(\zeta),$$

which implies the first inequality in (10.24). □

The quantity $\|\mathbf{r}_{k+1}(\zeta)\|$ is called *residual based error estimator*. On bounded domains, (10.24) shows that $\|\mathbf{r}_{k+1}(\zeta)\|$ is equivalent to the true approximation error $E_{k+1}(\zeta)$. Hence, if we restrict ζ to some bounded interval $[a,b] \subset \mathbb{R}_0^+$ and select the samples using $\|\mathbf{r}_{k+1}(\zeta)\|$ as surrogate for $E_{k+1}(\zeta)$, we obtain

$$\begin{aligned} E_{k+1}(\xi_{k+1}) &\geq \frac{1}{\lambda_{\max}+\xi_{k+1}}\|\mathbf{r}_{k+1}(\xi_{k+1})\| \\ &\geq \frac{1}{\lambda_{\max}+b}\max_{\zeta\in[a,b]}\|\mathbf{r}_{k+1}(\zeta)\| \geq \frac{\lambda_{\min}+a}{\lambda_{\max}+b}\max_{\zeta\in[a,b]}E_{k+1}(\zeta). \end{aligned}$$

In view of these results, a competitive selection of poles for approximating the matrix resolvent on $[a,b] \subset \mathbb{R}_0^+$ might be obtained inductively by

$$\xi_{k+1} := -\arg\max_{\zeta\in[a,b]}\|\mathbf{r}_{k+1}(\zeta)\|. \tag{10.25}$$

The following theoretical tool is instrumental to quantify the quality of these poles.

Definition 10.20. *For all $k \in \mathbb{N}$ we define the Kolmogorov k-width as*

$$\mathscr{K}_k := \inf_{\dim(V_k)\leq k}\sup_{\zeta\in\mathbb{R}_0^+}\inf_{\mathbf{v}_k\in V_k}\|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b}-\mathbf{v}_k\|$$

and set $\mathscr{K}_0 := \sup_{\zeta\in\mathbb{R}_0^+}\|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b}\|$ by convention.

The quantity $\mathscr{K}_k$ represents the best achievable decay for approximating the matrix resolvent. In practice, however, the optimal space, for which the infimum is attained, is computationally out of reach. Therefore, the above quantity should be viewed as a benchmark for more practical choices of the approximation space. The success of weak greedy algorithms is due to the following powerful result.

Theorem 10.21. *Let $0 \leq a < b < \infty$, $\Xi = \{\xi_0,\dots,\xi_k\}$ with $\xi_0 = \infty$ and $\xi_1,\dots,\xi_k$ chosen according to (10.25). Then there holds for any C_0, $c_0 \in \mathbb{R}^+$*

$$\forall k\in\mathbb{N}_0 : \mathscr{K}_k \leq C_0 e^{-c_0 k} \quad\Longrightarrow\quad \sup_{\zeta\in[a,b]}E_{k+1}(\zeta) \leq C_0\max\left\{\sqrt{2}\frac{\lambda_{\max}+b}{\lambda_{\min}+a}, e^{\frac{c_0}{6}}\right\}e^{-\frac{c_0 k}{6}}.$$

Proof. See [CD15, Corollary 8.4(iii)] and [DPW13, Corollary 3.3(iii)]. □

According to Theorem 10.21, exponential convergence of the Kolmogorov k-width guarantees exponential convergence of the weak greedy algorithm. Thanks to Corollary 8.25, we know that the former can be bounded by

$$\mathscr{K}_k \leq \sup_{\zeta\in[a,b]} \inf_{\mathbf{v}_k\in\mathcal{Q}_k^{\tilde{\mathcal{Z}}}(\mathbf{L},\mathbf{b})} \|(\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{b}-\mathbf{v}_k\| \leq \frac{2}{\lambda_{\min}+a} Z_k(\Sigma,[-b,-a])\|\mathbf{b}\| \tag{10.26}$$

for all $k \in \mathbb{N}$, where $\tilde{\mathcal{Z}}$ is the set containing the poles of Zolotarëv's extremal rational function on the condenser $(\Sigma,[-b,-a])$ in the sense of Theorem 9.30. This allows us to prove the following result; c.f. [BGZ20, Lemma 3.1].

Corollary 10.22. *Let $0 \leq a < b < \infty$, $\Xi = \{\xi_0,\dots,\xi_k\}$ with $\xi_0 = \infty$ and $\xi_1,\dots,\xi_k$ chosen according to (10.25), and $\delta = \delta_{[-b,-a;\lambda_{\min},\lambda_{\max}]}$ defined by (9.29). Then there holds*

$$\sup_{\zeta\in[a,b]} E_{k+1}(\zeta) \leq C\rho_{[\delta,1]}^{-\frac{k}{6}}\|\mathbf{b}\|, \qquad C = \frac{8}{\lambda_{\min}+a}\max\left\{\sqrt{2}\frac{\lambda_{\max}+b}{\lambda_{\min}+a}, \rho_{[\delta,1]}^{\frac{1}{6}}\right\}.$$

Proof. It follows from Theorem 9.30 that

$$Z_k(\Sigma,[-b-a]) \leq 4\rho_{[\delta,1]}^{-k},$$

Together with (10.26), this yields

$$\mathscr{K}_k \leq \frac{8}{\lambda_{\min}+a}\rho_{[\delta,1]}^{-k}\|\mathbf{b}\|$$

for all $k \in \mathbb{N}$. The inequality remains valid if $k = 0$. Hence, the conjecture follows by Theorem 10.21 with

$$c_0 = \frac{\pi^2}{\ln(4\delta^{-1})}, \qquad C_0 = \frac{8}{\lambda_{\min}+a}. \qquad \square$$

The major disadvantage of Corollary 10.22 is the fact that (10.22) is limited to bounded intervals $[a,b] \subset \mathbb{R}^+$. If by chance, however, the Cauchy-Stieltjes density μ_C^τ of f^τ has compact support, one can apply the weak greedy algorithm to any bounded superset $[a,b] \supset \operatorname{supp}\mu_C$ to arrive at the following result.

Corollary 10.23. *Let $f^\tau \in \mathcal{CS}$ with Cauchy-Stieltjes density μ_C^τ satisfying $\operatorname{supp}\mu_C^\tau \subset [a,b]$ for $0 \leq a < b < \infty$, $\Xi = \{\xi_0,\dots,\xi_k\}$ with $\xi_0 = \infty$ and $\xi_1,\dots,\xi_k$ chosen according to (10.25), δ and C as in Corollary 10.22, $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}f^\tau(\mathbf{L}_{k+1})\mathbf{V}^\dagger\mathbf{b}$. Then*

$$\|f^\tau(\mathbf{L})\mathbf{b}-\mathbf{u}_{k+1}\| \leq \tilde{C}\rho_{[\delta,1]}^{-\frac{k}{6}}\|\mathbf{b}\|, \qquad \tilde{C} := C\int_a^b \mu_C^\tau(\zeta)\,d\zeta.$$

Proof. If μ_C^τ has compact support, it follows from (10.22) that

$$\|f^\tau(\mathbf{L})\mathbf{b}-\mathbf{u}_{k+1}\| \leq \int_a^b \mu_C^\tau(\zeta)\,d\zeta \sup_{\zeta\in[a,b]} E_{k+1}(\zeta).$$

The claim now follows from Corollary 10.22. $\square$

Although Cauchy-Stieltjes functions exist whose density functions have compact support, see e.g., (8.5), Corollary 10.23 is only of limited use since most of the functions that we are interested in do not possess such a property. One possibility to mitigate this problem is to sample the poles over a sufficiently large parameter domain so that the truncation error falls below a user-defined threshold. Exemplary, we state the following result for the case where $f^\tau(\lambda) = \lambda^{-s}$; cf. [BLP19b].

Theorem 10.24. *Let* $\mathrm{n}_\mathrm{c}^-, \mathrm{n}_\mathrm{c}^+ \in \mathbb{R}^+$, $a = e^{-\mathrm{n}_\mathrm{c}^-}$, $b = e^{\mathrm{n}_\mathrm{c}^+}$, $\Xi = \{\xi_0, \dots, \xi_k\}$ *with* $\xi_0 = \infty$ *and* $\xi_1, \dots, \xi_k$ *chosen according to* (10.25), C *as in Corollary 10.22,* $s \in (0,1)$, $\mathbf{V}$ *an orthonormal basis of* $\mathcal{Q}_{k+1}^\Xi(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, *and* $\mathbf{u}_{k+1} = \mathbf{V}\mathbf{L}_{k+1}^{-s}\mathbf{V}^\dagger\mathbf{b}$. *Then*

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \leq \left(\frac{2e^{-(1-s)\mathrm{n}_\mathrm{c}^-}}{\lambda_{\min}(1-s)} + \frac{2e^{-s\mathrm{n}_\mathrm{c}^+}}{s} + C\frac{e^{(1-s)\mathrm{n}_\mathrm{c}^+} - e^{-(1-s)\mathrm{n}_\mathrm{c}^-}}{1-s} \rho_{[\lambda_{\min}, 4\lambda_{\max}]}^{-\frac{k}{6}} \right) \|\mathbf{b}\|. \tag{10.27}$$

Proof. Starting from Balakrishnan's formula (8.10), we apply the substitution $\zeta \mapsto \ln(\zeta)$ to deduce from Theorem 2.37

$$\begin{aligned}
\mathbf{L}^{-s}\mathbf{b} &= \frac{\sin(\pi s)}{\pi} \int_{-\infty}^{\infty} e^{(1-s)\zeta} (\mathbf{L} + e^\zeta \mathbf{I})^{-1} \mathbf{b} \, d\zeta, \\
\mathbf{u}_{k+1} &= \frac{\sin(\pi s)}{\pi} \int_{-\infty}^{\infty} e^{(1-s)\zeta} \mathbf{V} (\mathbf{L}_{k+1} + e^\zeta \mathbf{I}_{k+1})^{-1} \mathbf{V}^\dagger \mathbf{b} \, d\zeta.
\end{aligned}$$

We split each of the two integrals in its three contributions over $(-\infty, -\mathrm{n}_\mathrm{c}^-]$, $[-\mathrm{n}_\mathrm{c}^-, \mathrm{n}_\mathrm{c}^+]$, and $[\mathrm{n}_\mathrm{c}^+, \infty)$, apply the triangle inequality, and deduce, by definition of $E_{k+1}(\zeta)$,

$$\begin{aligned}
\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \leq \int_{-\infty}^{-\mathrm{n}_\mathrm{c}^-} e^{(1-s)\zeta} E_{k+1}(e^\zeta) \, d\zeta + \int_{-\mathrm{n}_\mathrm{c}^-}^{\mathrm{n}_\mathrm{c}^+} e^{(1-s)\zeta} E_{k+1}(e^\zeta) \, d\zeta \\
+ \int_{\mathrm{n}_\mathrm{c}^+}^{\infty} e^{(1-s)\zeta} E_{k+1}(e^\zeta) \, d\zeta.
\end{aligned}$$

Invoking the orthonormal property of $\mathbf{V}$ and the fact that the rational Ritz values are contained in Σ, we find

$$\begin{aligned}
\int_{-\infty}^{-\mathrm{n}_\mathrm{c}^-} e^{(1-s)\zeta} E_{k+1}(e^\zeta) \, d\zeta &\leq \int_{-\infty}^{-\mathrm{n}_\mathrm{c}^-} e^{(1-s)\zeta} \big(\|(\mathbf{L} + e^\zeta \mathbf{I})^{-1}\mathbf{b}\| + \|(\mathbf{L}_{k+1} + e^\zeta \mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger\mathbf{b}\|_2 \big) d\zeta \\
&\leq \|\mathbf{b}\| \int_{-\infty}^{-\mathrm{n}_\mathrm{c}^-} \frac{e^{(1-s)\zeta}}{\lambda_{\min}} \, d\zeta + \|\mathbf{V}^\dagger\mathbf{b}\|_2 \int_{-\infty}^{-\mathrm{n}_\mathrm{c}^-} \frac{e^{(1-s)}\zeta}{\lambda_{\min}} \, d\zeta.
\end{aligned}$$

Once more, we make use of the fact that $\mathbf{V}$ is an orthonormal basis to find $\|\mathbf{V}^\dagger\mathbf{b}\|_2 = \|\mathbf{P}\mathbf{b}\|$, where $\mathbf{P} := \mathbf{V}\mathbf{V}^\dagger$ is the orthogonal projector onto $\mathcal{Q}_{k+1}^\Xi(\mathbf{L}, \mathbf{b})$. Hence $\|\mathbf{V}^\dagger\mathbf{b}\|_2 = \|\mathbf{P}\mathbf{b}\| \leq \|\mathbf{b}\|$. Direct evaluations of the integrals reveal

$$\int_{-\infty}^{-\mathrm{n}_\mathrm{c}^-} e^{(1-s)\zeta} E_{k+1}(e^\zeta) \, d\zeta \leq 2\frac{e^{-(1-s)\mathrm{n}_\mathrm{c}^-}}{(1-s)\lambda_{\min}} \|\mathbf{b}\|.$$

Similar computations show

$$\begin{aligned}\int_{\mathrm{n_c^+}}^{\infty} e^{(1-s)\zeta} E_{k+1}(e^{\zeta})\,d\zeta &\leq \int_{\mathrm{n_c^+}}^{\infty} e^{(1-s)\zeta}\left(\|(\mathbf{L}+e^{\zeta}\mathbf{I})^{-1}\mathbf{b}\| + \|(\mathbf{L}_{k+1}+e^{\zeta}\mathbf{I}_{k+1})^{-1}\mathbf{V}^{\dagger}\mathbf{b}\|_2\right)d\zeta \\ &\leq \|\mathbf{b}\|\int_{\mathrm{n_c^+}}^{\infty}\frac{e^{(1-s)\zeta}}{e^{\zeta}}\,d\zeta + \|\mathbf{b}\|\int_{\mathrm{n_c^+}}^{\infty}\frac{e^{(1-s)}\zeta}{e^{\zeta}}\,d\zeta \\ &= 2\|\mathbf{b}\|\int_{\mathrm{n_c^+}}^{\infty} e^{-\zeta s}\,d\zeta = 2\frac{e^{-s\mathrm{n_c^+}}}{s}\|\mathbf{b}\|.\end{aligned}$$

To bound the third integral, we apply Corollary 10.22 and find

$$\begin{aligned}\int_{-\mathrm{n_c^-}}^{\mathrm{n_c^+}} e^{(1-s)\zeta} E_{k+1}(\zeta)\,d\zeta &\leq \int_{-\mathrm{n_c^-}}^{\mathrm{n_c^+}} e^{(1-s)\zeta}\,d\zeta \sup_{\zeta\in[e^{-\mathrm{n_c^-}},e^{\mathrm{n_c^-}}]} E_{k+1}(\zeta) \\ &= \frac{e^{(1-s)\mathrm{n_c^+}} - e^{-(1-s)\mathrm{n_c^-}}}{1-s} \sup_{\zeta\in[e^{-\mathrm{n_c^-}},e^{\mathrm{n_c^+}}]} E_{k+1}(\zeta).\end{aligned}$$

Combining all three integral estimates together with Corollary 10.23 we deduce

$$\|\mathbf{L}^{-s}\mathbf{b}-\mathbf{u}_{k+1}\| \leq \left(\frac{2e^{-(1-s)\mathrm{n_c^-}}}{\lambda_{\min}(1-s)} + \frac{2e^{-s\mathrm{n_c^+}}}{s} + C\frac{e^{(1-s)\mathrm{n_c^+}} - e^{-(1-s)\mathrm{n_c^-}}}{1-s}\rho_{[\delta,1]}^{-\frac{k}{6}}\right)\|\mathbf{b}\|.$$

The claim now follows from the observation that $\delta \geq \frac{\lambda_{\min}}{4\lambda_{\max}}$. Therefore, $\rho_{[\delta,1]}^{-\frac{k}{6}} \leq \rho_{[\frac{\lambda_{\min}}{4\lambda_{\max}},1]}^{-\frac{k}{6}} = \rho_{[\lambda_{\min},4\lambda_{\max}]}^{-\frac{k}{6}}$. □

Theorem 10.24 can be seen as improvement of the results provided in [BGZ20], where the slower convergence rates obtained by $\mathcal{Z}$ were used to show exponential convergence of the Kolmogorov k-width. The first two contributions on the right-hand side of (10.27) are caused by sampling the parameters only over the bounded domain $[e^{-\mathrm{n_c^-}}, e^{\mathrm{n_c^+}}]$ instead of $\mathbb{R}_0^+$. What these terms are concerned, it is desirable to choose $\mathrm{n_c^-}$ and $\mathrm{n_c^+}$ as large as possible. On the other hand, large values of these parameters enlarge the sample domain of the weak greedy algorithm, increasing the constant in the last term of (10.27).

Although analytically not fully understood, it is observed numerically that also for $f^{\tau}(\lambda) \in \{\lambda^s, E_{\alpha,\beta}(-t^{\alpha}\lambda^s)\}$ the poles obtained by (10.25) provide an excellent choice for building rational Krylov approximations if $[a,b]$ is sufficiently large. However, our previous results show that the approximability of $f^{\tau}(\mathbf{L})\mathbf{b}$ for any $f^{\tau} \in \mathcal{CS}\cup\mathcal{CB}\cup\mathcal{LS}$ is closely related to the approximability of the (scalar-valued) resolvent function on the spectral interval of $\mathbf{L}$. In particular, the pole sets $\mathcal{Z}$, $\mathcal{E}$, and $\mathcal{S}$ are all contained in $-\Sigma$. A reasonable choice for $[a,b]$, which avoids the selection of the cut-off parameters $\mathrm{n_c^-}, \mathrm{n_c^+}$ is thus given by $[a,b] = [\lambda_{\min}, \lambda_{\max}]$. Before we manifest these observations in the central definition of this section, we mention that the analysis provided above remains valid if we replace $[a,b]$ in (10.25) by a discrete training set, chosen either sufficiently fine to retain the accuracy of the algorithm [CD15] or based on random selections of moderate size [CDDN20].

Definition 10.25. *Let $\mathcal{T}_{\text{train}}$ be a training set of Σ. The pole set $\mathcal{G}_\infty = \{\xi_0, \dots, \xi_k\}$ is defined inductively by $\xi_0 := \infty$ and*

$$\xi_{j+1} := -\operatorname*{arg\,max}_{\zeta \in \mathcal{T}_{\text{train}}} \|\mathbf{r}_{k+1}(\zeta)\|, \tag{10.28}$$

where $\mathbf{r}_{k+1}(\zeta)$ is the residual from Lemma 10.19. We call $\mathcal{G} := \{\xi_1, \dots, \xi_k\}$ the weak greedy poles on $-\Sigma$. Accordingly, the pole set $\hat{\mathcal{G}}_\infty = \{\hat{\xi}_0, \dots, \hat{\xi}_k\}$ is defined by $\hat{\xi}_0 := \infty$ and

$$\hat{\xi}_{j+1} := -\operatorname*{arg\,max}_{\zeta \in \mathcal{T}_{\text{train}}^{\mathrm{n_c}^\pm}} \|\mathbf{r}_{k+1}(\zeta)\|, \tag{10.29}$$

where $\mathcal{T}_{\text{train}}^{\mathrm{n_c}^\pm} \subset \mathbb{R}_0^-$ is a training set of $[e^{-\mathrm{n_c}^-}, e^{\mathrm{n_c}^+}]$. We call $\hat{\mathcal{G}} := \{\hat{\xi}_1, \dots, \hat{\xi}_k\}$ the weak greedy poles on $\mathbb{R}_0^-$.

Remark 10.26. *It goes without saying that numerous other error estimators as well as a priori parameter selections exist to approximate parametric PDEs of reaction-diffusion type; see e.g., [RHP08, QRM11, MPT02] and references therein.*

By construction, the weak greedy poles are nested. More importantly, they are independent of the parameter and thus allow for an efficient querying of the solution map $\boldsymbol{\tau} \mapsto f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ in the course of an online-offline decomposition. The bounds provided in Theorem 10.27, however, suggest that the surrogate degenerates when s approaches an integer. Hence, we cannot prove uniform convergence for $s \in [0, 1]$ with the tools presented above. Just like $\mathcal{S}$ and $\hat{\mathcal{S}}$, weak greedy poles require a user-provided training set to carry out the maximization. Unlike their spectral counterparts, however, they are computed using spatial error estimators which makes their implementation more demanding, both from a computational and theoretical point of view. The key ingredient for an efficient realization of (10.28) is the implementation of an online-offline routine. After an initial computational investment, this allows one to query $\zeta \mapsto \|\mathbf{r}_{k+1}(\zeta)\|$ with complexity only depending on k. To make matters precise, let $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ and $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$. If we define $\mathbf{u}_{k+1}^{\mathrm{C}}(\zeta) := (\mathbf{L}_{k+1} + \zeta \mathbf{I}_{k+1})^{-1} \mathbf{V}^\dagger \mathbf{b}$ to be the coordinate vector of $\mathbf{V}(\mathbf{L}_{k+1} + \zeta \mathbf{I}_{k+1})^{-1} \mathbf{V}^\dagger \mathbf{b}$, we observe

$$\begin{aligned}
\|\mathbf{r}_{k+1}(\zeta)\|^2 &= \|\mathbf{b} - (\mathbf{L} + \zeta \mathbf{I}) \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)\|^2 \\
&= \|\mathbf{b}\|^2 - 2(\mathbf{b}, (\mathbf{L} + \zeta \mathbf{I}) \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)) + \|(\mathbf{L} + \zeta \mathbf{I}) \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)\|^2 \\
&= \|\mathbf{b}\|^2 - 2(\mathbf{b}, \mathbf{L} \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)) - 2\zeta(\mathbf{b}, \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)) + \|\mathbf{L} \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)\|^2 \\
&\qquad + 2\zeta(\mathbf{L} \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta), \mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)) + \zeta^2 \|\mathbf{V} \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)\|^2.
\end{aligned}$$

Recalling $\mathbf{L} = \mathbf{M}^{-1} \mathbf{A}$, we define the quantities

$$\begin{aligned}
c_{\mathbf{b}} &:= \|\mathbf{b}\| \in \mathbb{R}, & \mathbf{v}_1 &:= \mathbf{V}^T \mathbf{A} \mathbf{b} \in \mathbb{R}^{k+1}, \\
\mathbf{v}_2 &:= \mathbf{V}^T \mathbf{M} \mathbf{b} \in \mathbb{R}^{k+1}, & \mathbf{L}_{2,k+1} &:= \mathbf{V}^T \mathbf{A} \mathbf{M}^{-1} \mathbf{A} \mathbf{V} \in \mathbb{R}^{(k+1) \times (k+1)},
\end{aligned} \tag{10.30}$$

so that, invoking the second property in Lemma 7.10,

$$\begin{aligned}
\|\mathbf{r}_{k+1}(\zeta)\|^2 = c_{\mathbf{b}}^2 - 2(\mathbf{v}_1, \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta))_2 - 2\zeta(\mathbf{v}_2, \mathbf{u}_{k+1}^{\mathrm{C}}(\zeta))_2 + \|\mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)\|_{\mathbf{L}_{2,k+1}}^2 \\
+ 2\zeta \|\mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)\|_{\mathbf{L}_{k+1}}^2 + \zeta^2 \|\mathbf{u}_{k+1}^{\mathrm{C}}(\zeta)\|_2^2.
\end{aligned} \tag{10.31}$$

An efficient realization of (10.28) and (10.29) now

1. computes (10.30) once and for all in the offline phase,
2. queries the right-hand side of (10.31) in the online phase for all ζ in the training set.

Note that the right-hand side of (10.31) only depends on the Krylov parameter k such that its evaluation for several thousand values of ζ is feasible. Although efficient, the reader should be warned that the implementation in its present form is prone to numerical round-off errors. In double precision arithmetic, this leads to stagnation of the residual based error indicator whenever $\|\mathbf{r}_{k+1}(\zeta)\|$ is in the range of 10^{-8}. More careful computations allow one to overcome these difficulties [Cas12, CEL14, BEOR14, YJN19].

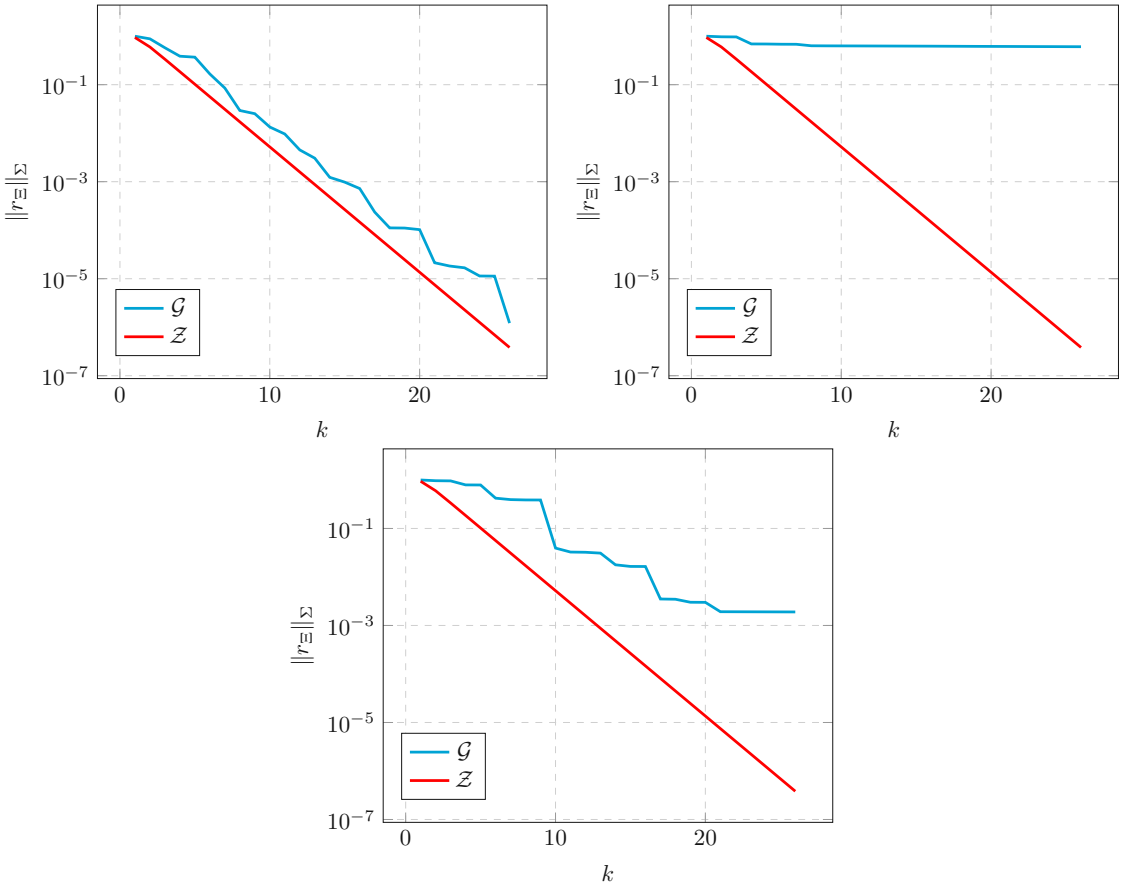

Figure 10.5: Maximum norm $\|r_\Xi\|_\Sigma$ on $\Sigma = [1, 1000]$ for $\Xi \in \{\mathcal{G}, \mathcal{Z}\}$ with $\mathbf{L}_1$ and $\mathbf{b}_1$ (top left), $\mathbf{L}_2$ and $\mathbf{b}_1$ (top right), and $\mathbf{L}_1$ and $\mathbf{b}_2$ (bottom).

As shown in the previous section, the pole set $\mathcal{S}$ allows for an interpretation as approximate solution to the modified Zolotarëv deviation problem (10.21). This motivates us to consider the following matrix-valued variant to the third Zolotarëv problem: Find $\Psi \subset \mathbb{B} \subset \mathbb{R}_0^-$ with $|\Psi| = k$, such that

$$\frac{\|r_{\Lambda,\Psi}(\mathbf{L})\mathbf{b}\|}{\inf\{|r_{\Lambda,\Psi}(\lambda)| : \lambda \in \mathbb{B}\}} = \min_{\substack{\Xi \subset \mathbb{B} \\ |\Xi| = k}} \frac{\|r_{\Lambda,\Xi}(\mathbf{L})\mathbf{b}\|}{\inf\{|r_{\Lambda,\Xi}(\lambda)| : \lambda \in \mathbb{B}\}}, \tag{10.32}$$

where Λ contains the rational Ritz values of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$. Unlike (10.21), (10.32) not only incorporates spectral information about the matrix $\mathbf{L}$ but also the one of the vector $\mathbf{b}$. Provided that $\mathbf{b}$ is uniformly excited by all eigenfunctions, it is thus reasonable to believe that solutions to (10.32) behave qualitatively similar to the ones obtained by (10.21) whenever $\mathbb{B} = -\Sigma$. Likewise, if the spectral density of $\mathbf{L}$ is uniform, (10.32) should yield solutions that are competitive with the ones obtained by the classical third Zolotarëv problem.

To establish a connection to weak greedy poles, we assume that the rational Ritz values are pairwise distinct. Then it follows from Theorem 7.14 that $\mathbf{V}(\mathbf{L}_{k+1} + \zeta \mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger \mathbf{b}$ coincides with the rational interpolant $r_{\Lambda,\Xi}^{\zeta}(\mathbf{L})\mathbf{b}$ defined in Definition 8.23 and, due to (8.27),

$$\begin{aligned}\sup_{\zeta \in -\mathbb{B}} E_{k+1}(\zeta) &= \sup_{\zeta \in -\mathbb{B}} \|(\mathbf{L} + \zeta \mathbf{I})^{-1}\mathbf{b} - \mathbf{V}(\mathbf{L}_{k+1} + \zeta \mathbf{I}_{k+1})^{-1}\mathbf{V}^\dagger \mathbf{b}\| \\ &= \sup_{\zeta \in -\mathbb{B}} \|(\mathbf{L} + \zeta \mathbf{I})^{-1} \frac{r_{\Lambda,\Xi}(\mathbf{L})\mathbf{b}}{r_{\Lambda,\Xi}(-\zeta)}\|.\end{aligned}$$

Hence,

$$\frac{1}{\lambda_{\max} + |\inf \mathbb{B}|} \frac{\|r_{\Lambda,\Xi}(\mathbf{L})\mathbf{b}\|}{\inf\{|r_{\Lambda,\Xi}(\zeta)| : \zeta \in \mathbb{B}\}} \leq \sup_{\zeta \in -\mathbb{B}} E_{k+1}(\zeta) \leq \frac{1}{\lambda_{\min}} \frac{\|r_{\Lambda,\Xi}(\mathbf{L})\mathbf{b}\|}{\inf\{|r_{\Lambda,\Xi}(\zeta)| : \zeta \in \mathbb{B}\}},$$

which proves that

$$\frac{\|r_{\Lambda,\Xi}(\mathbf{L})\mathbf{b}\|}{\inf\{|r_{\Lambda,\Xi}(\zeta)| : \zeta \in \mathbb{B}\}}$$

is equivalent to the true approximation error $E_{k+1}(\zeta)$ if $\mathbb{B}$ is bounded. As such, the weak greedy poles can be seen as approximation to the matrix-valued third Zolotarëv problem (10.32). To illuminate this relation in more detail, we report the maximal deviation of $r_{\mathcal{G}}$ on Σ in Figure 10.5 for the diagonal matrices $\mathbf{L}_1$, $\mathbf{L}_2 \in \mathbb{R}^{N \times N}$, $N = 1000$, used in Figure 10.3 and different vectors $\mathbf{b}_1$, $\mathbf{b}_2 \in \mathbb{R}^N$. The vector $\mathbf{b}_1$ is the constant 1-vector, which is uniformly excited by all eigenfunctions of the respective matrix. The second vector $\mathbf{b}_2 = (b_{2,1}, \ldots, b_{2,N})^T \in \mathbb{R}^{N \times 1}$ is a linear combination of the first 100 eigenvectors only and is defined by

$$b_{2,j} := \begin{cases} 1, & \text{if } j \leq 100, \\ 0, & \text{if } j > 100. \end{cases}$$

It is observed in Figure 10.5 that the weak greedy poles on $-\Sigma$ provide an excellent approximation to Zolotarëv's minimal deviation problem if the eigenvalues of the matrix are roughly equispaced and the vector is uniformly excited by all eigenfunctions. If either $\mathbf{L}$ or $\mathbf{b}$ do not exhibit such a uniform pattern, the quantity $\|r_{\mathcal{G}}\|_\Sigma$ stagnates for increasing k.

The pole set $\mathcal{G}$ is illustrated in Figure 10.6 for different matrices, vectors, and orders. As expected, $\mathcal{G}$ behaves qualitatively similar to $\mathcal{Z}$ for the uniform configuration $\mathbf{L}_1$ and $\mathbf{b}_1$. The weak greedy poles accumulate at the right end point of $-\Sigma$ if the biased matrix $\mathbf{L}_2$ is involved. This effect is more dramatic than for the spectral poles. Unlike the latter, however, $\mathcal{G}$ also depends on the particular vector. Therefore, the poles tend to cluster in those regions of the spectral interval whose corresponding eigenspace contributes to the excitation of $\mathbf{b}$.

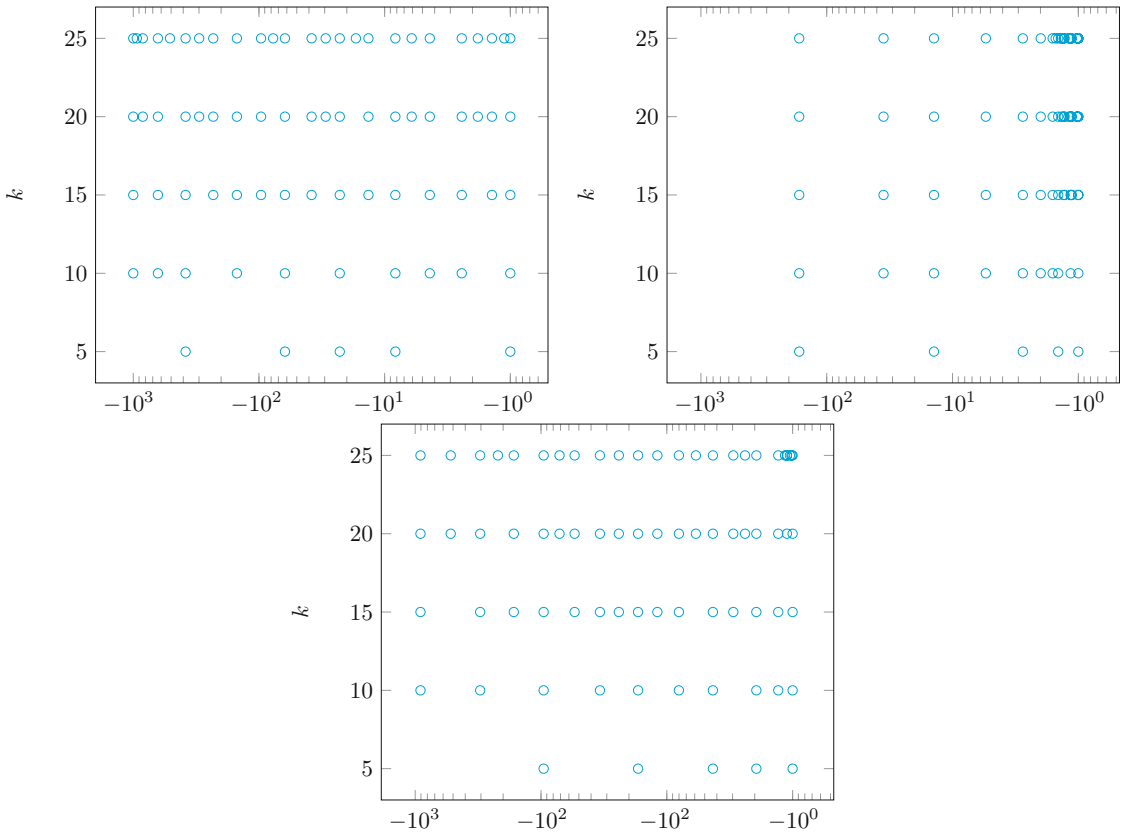

Figure 10.6: Weak greedy poles $\mathcal{G}$ on $-\Sigma = [-1000, -1]$ for $\mathbf{L}_1$ and $\mathbf{b}_1$ (top left), $\mathbf{L}_2$ and $\mathbf{b}_1$ (top right), and $\mathbf{L}_1$ and $\mathbf{b}_2$ (bottom).

10.1.5 Poles based on Rational Approximation

Up to this point, all presented pole configurations justify their selection based on the (simultaneous) approximability of the matrix kernels $g(\mathbf{L}, \zeta) \in \{(\mathbf{L} - \zeta\mathbf{I})^{-1}, \mathbf{L}(\mathbf{L} - \zeta\mathbf{I})^{-1}, e^{-\zeta\mathbf{L}}\}$. While this philosophy is the key ingredient for choosing parameter-independent pole sets, one withholds at the same time valuable information which might allow for a better adjustment towards the particular problem. If an approximation of $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ is desired only for one single value of $\boldsymbol{\tau}$, it might be worthwhile to incorporate the particular parameter in the choice of the poles. A conceptually straightforward approach to achieve this is provided by Theorem 7.16 which states

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2\|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_\Xi} \|f^{\boldsymbol{\tau}} - r_k\|_\Sigma. \tag{10.33}$$

To simplify matters, we set $\xi_0 = \infty$ henceforth. Then, according to (10.33), the poles of any rational function $r_k^{\boldsymbol{\tau}} \in \mathcal{R}_{k,k}$ that approximates $f^{\boldsymbol{\tau}}$ uniformly on Σ should constitute a good selection of parameters for building the rational Krylov space. Using [AN17, AN18, AN19] as a starting point, the authors of [ABDN19] make use of a Padé approximation of λ^{-s} and $(1 + \zeta\lambda^s)^{-1}$, $(s, \zeta) \in (0, 1) \times \mathbb{R}^+$, in Σ to approximate $\mathbf{L}^{-s}\mathbf{b}$ and $(\mathbf{I} + \zeta\mathbf{L}^s)^{-1}\mathbf{b}$ efficiently.

Numerical experiments in [ABDN19, DH21] confirm that its poles provide a competitive choice for approximating the matrix-vector product for fixed parameters s and ζ.

What the upper bound (10.33) is concerned, it is desirable to choose the pole set Ξ according to the ones of a rational function $r_k^{\mathcal{B}_\tau} \in \mathcal{R}_{k,k}$ that satisfies

$$\|f^\tau - r_k^{\mathcal{B}_\tau}\|_\Sigma = \min_{r_k \in \mathcal{R}_{k,k}} \|f^\tau - r_k\|_\Sigma.$$

Any such function is said to be *the best uniform rational approximation (BURA) of* f^τ *on* Σ. Existence and uniqueness of this problem is a classical result [Ach92].

Theorem 10.27. *If* $f^\tau \in C(\Sigma)$*, then there exists a unique best uniform rational approximation of* f^τ *on* Σ*.*

The *defect* $d \in \mathbb{N}_0$ of the BURA $r_k^{\mathcal{B}_\tau} \in \mathcal{R}_{k,k}$ is defined by the integer

$$d := k - \max\{\deg(p_k^\tau), \deg(q_\Xi^\tau)\},$$

where $p_k^\tau \in \mathcal{P}_k$ and $q_\Xi^\tau \in \mathcal{P}_k$ are polynomials of minimal degree such that $r_k^{\mathcal{B}_\tau} = p_k^\tau / q_\Xi^\tau$. The defect is a useful tool in the identification of BURAs. Here, and in all what follows, we say that a function g *equioscillates* between k extreme points in Σ if there exist k distinct points $\lambda_{\min} \leq \lambda_1^* < \cdots < \lambda_k^* \leq \lambda_{\max}$ such that

$$f(\lambda_j^*) = (-1)^{j+i_{\text{ini}}} \|g\|_\Sigma, \qquad j = 1, \ldots, k,$$

for some $i_{\text{ini}} \in \{0, 1\}$.

Lemma 10.28. *Let* $f^\tau \in C(\Sigma)$*,* $r_k^{\mathcal{B}_\tau} \in \mathcal{R}_{k,k}$ *the BURA of* f^τ *on* Σ*, and* d *its defect. Then there holds for any* $r_k \in \mathcal{R}_{k,k}$ *that* $r_k = r_k^{\mathcal{B}_\tau}$ *if and only if the error* $f^\tau - r_k$ *equioscillates between at least* $2k + 2 - d$ *points in* Σ*.*

Proof. See [Ach92, Tre19]. □

The computation of BURAs is a highly nontrivial task. Only for a few particular configurations of f^τ the latter is known analytically whence efficient numerical approximations are essential. The main difficulty of such algorithms lies in the implementation of *stable* procedures that yield accurate approximations for large values of k. The so-called *BRASIL algorithm* (best rational approximation by successive interval length adjustment) has been presented in [Hof21] and satisfies remarkable stability properties which allow one to compute BURAs of high degree to many functions of fractional diffusion type. In light of Lemma 10.28, its main idea is based on the observation that the BURA error $f^\tau - r_k^{\mathcal{B}_\tau}$ interpolates the function f^τ at a certain number of interpolation nodes $(x_j)_{j=0}^l \subset \Sigma$, $l \in \mathbb{N}$, and equioscillates between a sequence of extremal points $\lambda_j^* \in (x_{j-1}, x_j)$. The BRASIL algorithm now iteratively rescales the lengths of these intervals in order to equilibrate the local errors. The proposed method is available in the `baryrat`[1] Python package and can be used as a black-box pole generator for building the rational Krylov space. In light of these results, we provide the main definition of this section.

[1] https://github.com/c-f-h/baryrat

Definition 10.29. *In dependency of the function $f^{\boldsymbol{\tau}} \in C(\Sigma)$, we define the BURA poles $\mathcal{B}_{\boldsymbol{\tau}} := \{\xi_{\boldsymbol{\tau},1}^{(k)}, \dots, \xi_{\boldsymbol{\tau},k}^{(k)}\}$ as the poles of the BURA $r_k^{\mathcal{B}_{\boldsymbol{\tau}}} \in \mathcal{R}_{k,k}$ of $f^{\boldsymbol{\tau}}$ on Σ. We set $\mathcal{B}_{\boldsymbol{\tau}}^{\infty} := \mathcal{B}_{\boldsymbol{\tau}} \cup \{\infty\}$.*

In general, the BURA poles are neither nested nor independent of the parameter $\boldsymbol{\tau}$. Unlike $\mathcal{Z}$, $\mathcal{E}$, $\mathcal{S}$, and $\mathcal{G}$, they are not contained in $-\Sigma$. More importantly, they might not even be real as the function $f^{\boldsymbol{\tau}}(\lambda) = f(\lambda) = 1/(1+\lambda^2) \in \mathcal{LS}$ shows. Despite this inconvenience, we observe experimentally that $\mathcal{B}_{\boldsymbol{\tau}} \subset \mathbb{R}$ whenever $f^{\boldsymbol{\tau}}(\lambda) \in \{\lambda^{-s}, \lambda^s, E_{\alpha,\beta}(-t^\alpha \lambda^s)\}$.

Due to (10.33), the performance of the pole set $\mathcal{B}_{\boldsymbol{\tau}}$ is inherently related to the speed of convergence of the BURA as k approaches infinity. This subject has been studied intensively throughout the second half of the last century [Gon67, Gon78, Par88, ST92, Pro94]. To make matters precise, let $r_k^{\boldsymbol{\tau}} \in \mathcal{R}_{k,k}$ be a rational function with poles in $\Xi = \{\xi_1^{(k)}, \dots, \xi_k^{(k)}\}$ that interpolates $f^{\boldsymbol{\tau}} \in \mathcal{CS}$ in the nodes $\Lambda = \{\sigma_0^{(k)}, \dots, \sigma_k^{(k)}\}$. Then by definition of $r_k^{\mathcal{B}_{\boldsymbol{\tau}}}$ and the Hermite-Walsh formula for Cauchy-Stieltjes functions (10.15) it follows that

$$\|f^{\boldsymbol{\tau}} - r_k^{\mathcal{B}_{\boldsymbol{\tau}}}\|_\Sigma \le \|f^{\boldsymbol{\tau}} - r_k^{\boldsymbol{\tau}}\|_\Sigma \le f^{\boldsymbol{\tau}}(\lambda_{\min}) \frac{\|r_{\Lambda,\Xi}\|_\Sigma}{\inf\{|r_{\Lambda,\Xi}(\lambda)| : \lambda \in \mathbb{R}_0^-\}},$$

where we assume $\xi_0 = \infty$. Distributing Λ and Ξ asymptotically according to the equilibrium measure of the condenser $(\Sigma, \mathbb{R}_0^-)$, Theorem 9.20 yields

$$\lim_{k\to\infty} \|f^{\boldsymbol{\tau}} - r_k^{\mathcal{B}_{\boldsymbol{\tau}}}\|_\Sigma^{\frac{1}{k}} \le \lim_{k\to\infty} \left(\frac{\sup\{r_{\Lambda,\Xi}(\lambda) : \lambda \in \Sigma\}}{\inf\{r_{\Lambda,\Xi}(\lambda) : \lambda \in \mathbb{R}_0^-\}}\right)^{\frac{1}{k}} = e^{-\frac{1}{\operatorname{cap}(\Sigma,\mathbb{R}_0^-)}} \tag{10.34}$$

whenever $f^{\boldsymbol{\tau}} \in \mathcal{CS}$. Invoking (10.20) accordingly, it follows in complete analogy that (10.34) remains valid if $f^{\boldsymbol{\tau}} \in \mathcal{CB}$. Finally, if $f^{\boldsymbol{\tau}} \in \mathcal{LS}$ extends continuously to the imaginary axis, we infer from (10.14) with $\mathcal{C} = i\mathbb{R}$

$$\|f^{\boldsymbol{\tau}} - r_k^{\mathcal{B}_{\boldsymbol{\tau}}}\|_\Sigma \le \|f^{\boldsymbol{\tau}} - r_k^{\boldsymbol{\tau}}\|_\Sigma \le c_{f^{\boldsymbol{\tau}}} \frac{\|r_{\Lambda,\Xi}\|_\Sigma}{\inf\{|r_{\Lambda,\Xi}(z)| : z \in i\mathbb{R}\}},$$

where $c_{f^{\boldsymbol{\tau}}}$ is defined by (8.35). Hence, if $c_{f^{\boldsymbol{\tau}}} < \infty$ and Λ and Ξ are distributed according to the equilibrium measure of the condenser $(\Sigma, i\mathbb{R})$, it follows from Theorem 9.20

$$\lim_{k\to\infty} \|f^{\boldsymbol{\tau}} - r_k^{\mathcal{B}_{\boldsymbol{\tau}}}\|_\Sigma^{\frac{1}{k}} = e^{-\frac{1}{\operatorname{cap}(\Sigma,i\mathbb{R})}} \le e^{-\frac{1}{\sqrt{\operatorname{cap}(\Sigma,-\Sigma)}}}, \tag{10.35}$$

where the last inequality holds due to Proposition 9.34. We are now in position to show that the rational Krylov surrogate based on BURA poles asymptotically performs at least as good as the one based on Zolotarëv's poles.

Theorem 10.30. *Let $\mathbf{V}$ be an orthonormal basis of $\mathcal{Q}_{k+1}^{\mathcal{B}_{\boldsymbol{\tau}}^{\infty}}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V} f^{\boldsymbol{\tau}}(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}$.*

1. *If $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(t^\alpha \lambda^s)$ and $(\alpha, \beta, t, s) \in \Theta_{\mathrm{L}}$, then*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \rho_{[\lambda_{\min},\lambda_{\max}]}^{-\frac{k}{2}} \|\mathbf{b}\|. \tag{10.36}$$

2. *If* $f^\tau(\lambda) \in \{\lambda^{-s}, \lambda^s\}$ *with* $s \in (0,1)$*, then*

$$\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \rho^{-k}_{[\lambda_{\min}, 4\lambda_{\max}]}\|\mathbf{b}\|. \tag{10.37}$$

Moreover, (10.37) *remains valid for* $f^\tau(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s)$ *if* $(\alpha,\beta,t,s) \in \Theta_{\mathrm{C}}$.

Proof. At first, we apply Theorem 7.16 to see that

$$\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \|\mathbf{b}\| \min_{r_k \in \mathcal{P}_k/q_{\mathcal{B}_\tau}} \|f^\tau - r_k\|_\Sigma = \|\mathbf{b}\| \|f^\tau - r_k^{\mathcal{B}_\tau}\|_\Sigma. \tag{10.38}$$

If $f^\tau(\lambda) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$ and $(\alpha,\beta,t,s) \in \Theta_{\mathrm{L}}$, then $f^\tau \in \mathcal{LS}$ by Theorem 8.20. Moreover, $c_{f^\tau} < \infty$ thanks to Lemma 2.30. Thus, we may apply (10.35) to find

$$\|f^\tau - r_k^{\mathcal{B}_\tau}\|_\Sigma \preceq e^{-\frac{k}{\sqrt{\operatorname{cap}(\Sigma,-\Sigma)}}}.$$

Invoking Theorem 9.20 and (9.23), we deduce

$$\|f^\tau - r_k^{\mathcal{B}_\tau}\|_\Sigma \preceq \sqrt{Z_k(\Sigma,-\Sigma)} \leq \rho^{-\frac{k}{2}}_{[\lambda_{\min},\lambda_{\max}]},$$

whence (10.36) follows from (10.38) if $f^\tau(\lambda) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$ and $(\alpha,\beta,t,s) \in \Theta_{\mathrm{L}}$. Thanks to (10.7), it suffices to prove the second part of the conjecture. Let therefore $f^\tau(\lambda) \in \{\lambda^{-s}, \lambda^s\}$ with $s \in (0,1)$ or $f^\tau(\lambda) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$ with $(\alpha,\beta,t,s) \in \Theta_{\mathrm{C}}$, then $f^\tau \in \mathcal{CS} \cup \mathcal{CB}$. Therefore, (10.34) can be consulted to deduce by Theorem 9.20

$$\|f^\tau - r_k^{\mathcal{B}_\tau}\|_\Sigma \preceq e^{-\frac{k}{\operatorname{cap}(\Sigma,\mathbb{R}_0^-)}} \preceq Z_k(\Sigma,\mathbb{R}_0^-) \preceq \rho^{-k}_{[\lambda_{\min},4\lambda_{\max}]},$$

where the last inequality follows from Theorem 9.31. Once more, we invoke (10.38) to see that the claim is valid. □

In practice, the exponential convergence rates provided by Theorem 10.30 turn out to be rather pessimistic and one can hope for faster convergence rates in the form of

$$\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \preceq e^{-Ck}, \tag{10.39}$$

for some constant C larger than the ones encoded in Theorem 10.30. As shown in [Par88, Pro94, Pro05], see also [Rak16, Theorem 1] and [Güt10, Remark 7.8], there holds

$$\lim_{k\to\infty} \|f^\tau - r_k^{\mathcal{B}_\tau}\|_\Sigma^{\frac{1}{k}} \geq e^{-\frac{2}{\operatorname{cap}(\Sigma,\mathcal{C})}}, \tag{10.40}$$

where $\mathcal{C} \subset \mathbb{C}$ is the integration contour that encloses the largest possible domain in which f^τ is still analytic. Due to Propositions 8.3, 8.10, and Lemma 8.16, we have, in the limit case, $\mathcal{C} = \mathbb{R}_0^-$ whenever $f^\tau \in \mathcal{CS} \cup \mathcal{CB}$ and $\mathcal{C} = i\mathbb{R}$ if $f^\tau \in \mathcal{LS}$. Hence, any possible constant C satisfying (10.39) is at most two times larger than the respective constant obtained by Theorem 10.30. It remains to be clarified whether equality can be attained in (10.40). As

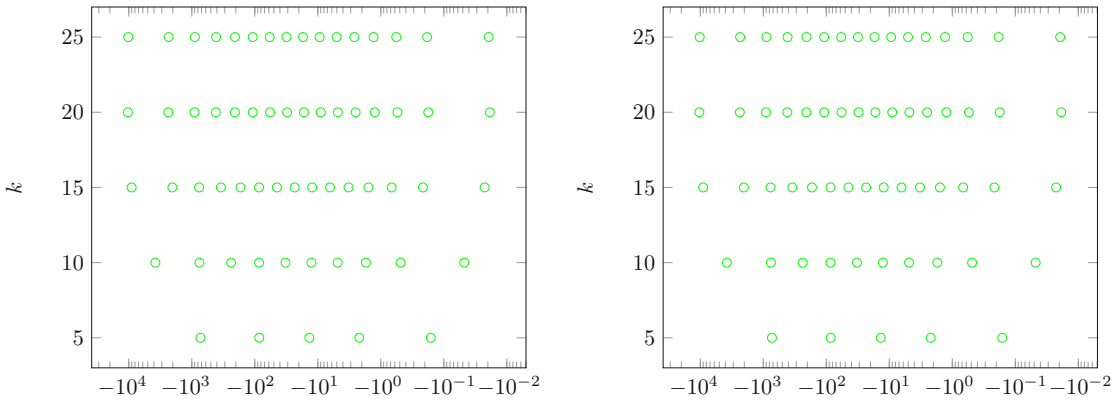

Figure 10.7: BURA poles $\mathcal{B}_{\boldsymbol{\tau}}$ of different orders k for $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-\frac{1}{2}}$ (left) and $f^{\boldsymbol{\tau}}(\lambda) = e^{-\sqrt{\lambda}}$ (right).

shown in [ST92, Theorem 6.2.2], the latter applies to the class of so-called *Markov functions*, that is, functions of the form

$$f^{\boldsymbol{\tau}}(\lambda) = \int \frac{d\nu^{\boldsymbol{\tau}}(\zeta)}{\lambda - \zeta},$$

where $\nu^{\boldsymbol{\tau}}$ is a complex measure with $\operatorname{supp} \nu^{\boldsymbol{\tau}} \subset [a, b]$, $-\infty \leq a < b < \infty$. Since $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ is a Markov function for all $s \in (0, 1)$ with $a = -\infty$, $b = 0$, and

$$d\nu^{s}(\zeta) = \frac{\sin(\pi s)}{\pi} |\zeta|^{-s} \, d\lambda,$$

where $d\lambda$ denotes the Lebesgue measure, we can improve the respective result of Theorem 10.30 in the following "optimal" manner.

Theorem 10.31. *Let $s \in (0, 1)$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+1}^{\mathcal{B}_{\boldsymbol{\tau}}^{\infty}}(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^{\dagger}\mathbf{L}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}\mathbf{L}_{k+1}^{-s}\mathbf{V}^{\dagger}\mathbf{b}$. Then there holds for all $s \in (0, 1)$*

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}\| \preceq \rho_{[\lambda_{\min}, 4\lambda_{\max}]}^{-2k} \|\mathbf{b}\|.$$

Proof. This is a consequence of Theorem 7.16 and [ST92, Theorem 6.2.2]. □

Theorem 10.31 shows that RKMs can benefit from the choice $\Xi = \mathcal{B}_{\boldsymbol{\tau}}$ when $\mathbf{u}_{k+1} \approx \mathbf{L}^{-s}\mathbf{b}$ is required for one single value of the parameter $s \in (0, 1)$. Although we are not aware of any comparable results for complete Bernstein and Laplace-Stieltjes functions, we observe numerically that BURA poles are among the most competitive pole sets if the solution map is approximated for one single value of the parameter only.

To complete the study of the BURA poles, we illustrate $\mathcal{B}_{\boldsymbol{\tau}} = \{\xi_{\boldsymbol{\tau},1}^{(k)}, \ldots, \xi_{\boldsymbol{\tau},k}^{(k)}\}$ for the functions $f^{\boldsymbol{\tau}}(\lambda) \in \{\lambda^{-\frac{1}{2}}, e^{-\sqrt{\lambda}}\}$ and different values of k in Figure 10.7. For both configurations of $f^{\boldsymbol{\tau}}$, the pole pattern looks rather similar. Although $\mathcal{B}_{\boldsymbol{\tau}} \not\subset -\Sigma$ already for small values of k, we observe that the poles tend to accumulate in the negative spectral interval.

10.2 Stopping Criteria

Nested pole sequences allow for an adaptive enrichment of the rational Krylov space until the sought accuracy is obtained. To take full advantage of their hierarchical structure, reliable error estimators are indispensable to assess the quality of the surrogate. The central objective of this section lies in the presentation of such estimators. In the first part, we present two well-established a posteriori error estimators that are suitable for the treatment of fractional diffusion problems. Although approved in many practical scenarios, these tools might fail to replicate the true approximation error since a rigorous analysis of their reliability is lacking. To mitigate this problem, we present, in the second part of this section, the implementation of a computable upper bound which allows one to assess the quality of approximations for a large class of poles where no analytical results are available.

10.2.1 Error Indicators

Throughout this section, let $f^\tau \in \mathcal{CS} \cup \mathcal{CB} \cup \mathcal{LS}$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_k(\mathbf{L}, \mathbf{b})$, $\mathbf{L}_k = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, and $\mathbf{u}_k = \mathbf{V} f^\tau(\mathbf{L}_k) \mathbf{V}^\dagger \mathbf{b}$, where we refrain from our usual "$k+1$-indexing" for more clarity in exposition. Our ambition lies in the description of a quantity $\eta_k \in \mathbb{R}_0^+$ with the property

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_k\| \approx \eta_k. \tag{10.41}$$

For this purpose, we review two well-known techniques which can be found in existing literature [Güt13, Güt10, DK94, KS10, DS11].

Difference of Iterates

A naive but practicable error indicator uses the triangle inequality as a starting point to bound the rational Krylov approximation error by

$$\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_k\| \leq \|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+d}\| - \|\mathbf{u}_{k+d} - \mathbf{u}_k\|$$

for some $d \in \mathbb{N}$. Under the assumption that the RKM converges sufficiently fast, one can expect the contribution $\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+d}\|$ to be small compared to $\|\mathbf{u}_{k+d} - \mathbf{u}_k\|$. After fixing the parameter d, a conceptually straightforward approach to obtain an error indicator satisfying (10.41) is

$$\eta_k := \|\mathbf{u}_{k+d} - \mathbf{u}_k\|.$$

The integer d is called *delay integer*. Typical choices of this parameter are e.g., $d = 2$ or $d = 3$. The evaluation of η_k can be carried out efficiently in the coordinate space with complexity only depending on k and d. The indicator may fail to replicate the error adequately when the approximation stagnates for $n \geq d$ iterations. In this case the value of η_k is too optimistic and does not provide a reliable estimate for $\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_k\|$.

Estimates based on Geometric Convergence

For any pole set presented in the previous section, there is proof or at least evidence that the corresponding rational Krylov approximation error decays at exponential rates. Hence, is is reasonable to assume that the rational Krylov iterates satisfy the identities

$$\|\mathbf{u}_{k+d} - \mathbf{u}_k\| \approx c\rho^{-k}, \qquad \|\mathbf{u}_{k+2d} - \mathbf{u}_{k+d}\| \approx c\rho^{-(k+d)}, \tag{10.42}$$

for some $c, \rho \in \mathbb{R}^+$ and the delay integer $d \in \mathbb{N}$. Following [KS10, Güt13] we define

$$\chi_k := \ln \|\mathbf{u}_{k+d} - \mathbf{u}_k\|,$$

which can be computed explicitly once the iterates $\mathbf{u}_k$ and $\mathbf{u}_{k+d}$ are available. This in turn allows us to estimate the value of ρ and c by

$$e^{-\frac{\chi_{k+d}-\chi_k}{d}} = \frac{\|\mathbf{u}_{k+d} - \mathbf{u}_k\|^{\frac{1}{d}}}{\|\mathbf{u}_{k+2d} - \mathbf{u}_{k+d}\|^{\frac{1}{d}}} = \frac{c^{\frac{1}{d}}\rho^{\frac{k}{d}}}{c^{\frac{1}{d}}\rho^{\frac{k}{d}-1}} \approx \rho,$$

and similarly

$$e^{\frac{(k+d)\chi_k - k\chi_{k+d}}{d}} \approx c,$$

which follows from (10.42) by direct substitution. We apply the triangle inequality to deduce

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_k\| \leq \|\mathbf{u}_k - \mathbf{u}_{k+d}\| + \|\mathbf{u}_{k+d} - \mathbf{u}_{k+2d}\| + \cdots + \|\mathbf{u}_{k+jd} - f(\mathbf{L})\mathbf{b}\|$$

for any $j \in \mathbb{N}$. Sending j to infinity yields

$$\begin{aligned}\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_k\| &\leq \|\mathbf{u}_k - \mathbf{u}_{k+d}\| + \sum_{j=1}^{\infty} \|\mathbf{u}_{k+jd} - \mathbf{u}_{k+(j+1)d}\| \\ &\approx c\rho^{-k} + c\rho^{-k}\sum_{j=1}^{\infty} \rho^{-jd}.\end{aligned}$$

Assuming $\rho > 1$, we recognize the latter as geometric series which evaluates to

$$\|f(\mathbf{L})\mathbf{b} - \mathbf{u}_k\| \approx \frac{c\rho^{-k}}{1-\rho^{-d}} =: \eta_k,$$

cf. [DK94, KS10, Güt13]. In practice, it might happen that $\chi_k \leq \chi_{k+d}$ in which case $\rho \leq 1$ such that η_k fails to mirror the error of the RKM. This typically indicates a stagnation of the error and one should iterate further to recover a reliable error estimator.

10.2.2 A Certified Error Estimate

Although approved in many practical scenarios, it might happen that the error estimators presented in Section 10.2.1 advocate to stop the enrichment of the search space *before* the required accuracy is achieved. To address this inconvenience, we present the implementation

of an error estimator that is *guaranteed* to bound the rational Krylov error. As a starting point we use the estimates provided by (10.2), (10.3), and (8.34) to bound

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2c_k^{\boldsymbol{\tau}}\|r_\Xi\|_\Sigma\|\mathbf{b}\|, \tag{10.43}$$

where we assume $\xi_0 = \infty$ and $\omega = \theta = 0$ in (8.4) and (8.12) for simplicity. The constant $c_k^{\boldsymbol{\tau}}$ is defined by

$$c_k^{\boldsymbol{\tau}} := \begin{cases} f^{\boldsymbol{\tau}}(\lambda_{\min}), & \text{if } f^{\boldsymbol{\tau}} \in \mathcal{CS}, \\ f^{\boldsymbol{\tau}}(\lambda_{\max}), & \text{if } f^{\boldsymbol{\tau}} \in \mathcal{CB}, \\ 4\gamma_k f^{\boldsymbol{\tau}}(0^+), & \text{if } f^{\boldsymbol{\tau}} \in \mathcal{LS}. \end{cases}$$

Provided that $\|r_\Xi\|_\Sigma$ is available, the right-hand side of (10.43) gives a computable upper bound of the rational Krylov error which allows us to assess the quality of the approximation even if no analytical results are known. It is clear that this bound can be crude, e.g., if the poles are sampled over $\mathbb{R}_0^-$, which applies to $\mathcal{B}_{\boldsymbol{\tau}}$ and the respective "hat" pole sets presented in Section 10.1. If we restrict ourselves to the poles contained in $-\Sigma$, however, our analytical and numerical discussions support the conjecture that $\Xi \in \{\mathcal{E}, \mathcal{S}, \mathcal{G}\}$ can be seen as reasonable approximation to Zolotarëv's minimal deviation problem on Σ if the spectral properties of the data is uniform.

To get ones hands on the error certificate (10.43), we need to determine $\|r_\Xi\|_\Sigma$ in a reliable way. A conceptually straightforward approach to compute the maximal deviation of r_Ξ is to evaluate its absolute value over a discrete training set $\mathcal{T}_{\text{train}} \subset \Sigma$ and choose its maximizer as approximation for $\|r_\Xi\|_\Sigma$. Somewhat cumbersomely, the training set must be provided by the user and needs to be chosen sufficiently fine to achieve a good approximation of the true global extremum. To counteract this, we present the following lemma which is instrumental in our computation of $\|r_\Xi\|_\Sigma$.

Lemma 10.32. *Let $\Xi = \{\xi_1, \dots, \xi_k\} \subset -\Sigma$ be a set of pairwise distinct poles with $\xi_k < \dots < \xi_1$. Then $r'_\Xi(\lambda)$ has exactly $k-1$ zeros $\lambda_1^*, \dots, \lambda_{k-1}^*$ in $\mathbb{R}^+$ that are local extrema of $r_\Xi(\lambda)$. There holds $-\xi_j < \lambda_j^* < -\xi_{j+1}$ for all $j = 1, \dots, k-1$ and*

$$r'_\Xi(\lambda) = -2\sum_{j=1}^{k} \frac{\xi_j}{(\lambda-\xi_j)^2} r_{\Xi_j}(\lambda), \quad r''_\Xi(\lambda) = -2\sum_{j=1}^{k} \frac{\xi_j}{(\lambda-\xi_j)^2}\left(r'_{\Xi_j}(\lambda) - \frac{2}{(\lambda-\xi_j)} r_{\Xi_j}(\lambda)\right), \tag{10.44}$$

where $\Xi_j = \{\xi_1, \dots, \xi_{j-1}, \xi_{j+1}, \dots, \xi_k\}$.

Proof. The function r_Ξ is smooth on the positive real axis and its roots are given by $-\xi_1, \dots, -\xi_k$. As such, Rolle's theorem guarantees the existence of at least $k-1$ local extrema $\lambda_1^*, \dots, \lambda_{k-1}^*$ such that $\lambda_j^* \in (-\xi_j, -\xi_{j+1})$ and $r'_\Xi(\lambda_j^*) = 0$ for all $j = 1, \dots, k-1$. To see that there cannot be more roots of r'_Ξ in $\mathbb{R}^+$, we note that r_Ξ has no zeros in $\mathbb{R}^-$ and thus at least one local extremum in each interval (ξ_j, ξ_{j-1}). Since $r_\Xi \in \mathcal{R}_{k,k}$ and $r_k \not\equiv 0$, it has $2k-2$ extremal points in total. Therefore, the first part of the proof is valid. The

identities in (10.44) follow from the generalized product rule

$$\frac{d}{d\lambda}\left(\prod_{j=1}^{k} g_j(\lambda)\right) = \sum_{j=1}^{k} \frac{dg_j}{d\lambda}(\lambda) \prod_{\substack{i=1 \\ i\neq j}}^{k} g_i(\lambda),$$

which holds for any collection of scalar and differentiable functions $(g_j)_{j=1}^k$. □

Thanks to Lemma 10.32, it suffices to compare the absolute values of $r_\Xi(\lambda_{\min})$ and $r_\Xi(\lambda_{\max})$ with those obtained by the $k-1$ local extrema $r_\Xi(\lambda_j^*)$ to determine the maximal deviation of $r_\Xi(\lambda)$ in $\Sigma = [\lambda_{\min}, \lambda_{\max}]$. We propose to compute $(\lambda_j^*)_{j=1}^{k-1}$ by Newton's method on each subinterval utilizing the tools provided by Lemma 10.32. Feasible initial values can be obtained by evaluating $r_\Xi(\lambda)$ over a discrete training set $\mathcal{T}_{\text{train}}^j \subset (-\xi_j, -\xi_{j+1})$, $j = 1, \dots, k-1$, of small cardinality. Provided that the initial value is sufficiently close to the true zero of r_Ξ', Newton's algorithm is guaranteed to converge to the desired solution as the following lemma shows.

Lemma 10.33. *Let $\Xi = \{\xi_1, \dots, \xi_k\} \subset -\Sigma$ be a set of poles with $\xi_k < \dots < \xi_1$ and λ_j^* the unique root of r_Ξ' in $(-\xi_j, -\xi_{j+1})$ for some $j \in \{1, \dots, k-1\}$. Then there exists some $\varepsilon > 0$ with $B_\varepsilon(\lambda_j^*) \subset (-\xi_j, -\xi_{j+1})$ such that for all initial values $\lambda_{j,0}^* \in B_\varepsilon(\lambda_j^*)$ Newton's method converges to λ_j^*.*

Proof. This is a direct consequence of Kantorovich's theorem [FV20]. □

We cannot quantify ε in Lemma 10.33 to choose $\mathcal{T}_{\text{train}}^j \subset (-\xi_j, -\xi_{j+1})$ sufficiently fine to guarantee $\lambda_{j,0}^* \in B_\varepsilon(\lambda_j^*)$ and thus convergence of Newton's method in $(-\xi_j, -\xi_{j+1})$. If it does converge, however, we can be certain that its limit yields the desired local extrema of r_Ξ. This observation suggests to start with a coarse training set and, if the iteration does not converge after a few steps or leaves the interval $(-\xi_j, -\xi_{j+1})$, restart Newton with an initial value extracted from a refined training set. Even though this procedure guarantees convergence only after some finite amount of refinements, we observe that the iteration is fairly robust in the initial value and usually converges in a few steps if we choose $|\mathcal{T}_{\text{train}}^j| = 20$ using equispaced points. We summarize this approach in Algorithm 2.

Remark 10.34. *The pole set $\mathcal{B}_\tau$ is not necessarily contained in $-\Sigma$ such that Algorithm 2 cannot be consulted to obtain a meaningful error indicator. Nevertheless, we can apply Theorem 7.16 to assess the quality of the rational Krylov surrogate obtained by the BURA $r_k^{\mathcal{B}_\tau}$ of f^τ on Σ by*

$$\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| \leq 2\Delta_{\mathcal{B}_\tau}\|\mathbf{b}\|, \tag{10.45}$$

where $\Delta_{\mathcal{B}_\tau} := \|f^\tau - r_k^{\mathcal{B}_\tau}\|_\Sigma$. Although analytically not available, $\Delta_{\mathcal{B}_\tau}$ can be recovered numerically as a by-product while computing $\mathcal{B}_\tau$ and is thus directly at hand.

Algorithm 2 Error Certification

Input: $0 < \lambda_L < \lambda_U$ with $\Sigma \subset [\lambda_L, \lambda_U]$, pole set $\Xi = \{\xi_1, \ldots, \xi_k\} \subset -[\lambda_L, \lambda_U]$ with $\xi_k < \cdots < \xi_1$ and $k \geq 2$, initial trainset size $n \in \mathbb{N}$, and maximum iteration number $i_{\max} \in \mathbb{N}$ for Newton's method

function ComputeInnerMax(Ξ, $n = 20$, $i_{\max} = 50$) // default parameter
  $\bar{\lambda} = 0$ // global maximum
  for $j = 1, \ldots, k-1$ **do**
    $\lambda^* = 0$
    while $\lambda^* \notin (-\xi_j, -\xi_{j+1})$ **do**
      $\mathcal{T}_{\text{train}} = [-\xi_j + \frac{1}{2n}, -\xi_j + \frac{1}{2n} + \frac{\xi_j - \xi_{j+1}}{n}, \ldots, -\xi_j + \frac{1}{2n} + (n-1)\frac{\xi_j - \xi_{j+1}}{n}]$
      $\lambda_0 = \arg\max_{\lambda \in \mathcal{T}_{\text{train}}} |r_\Xi(\lambda)|$
      $\lambda^* = \text{Newton}(\lambda_0, \Xi, i_{\max})$ // Computes root of r'_Ξ using (10.44)
      $n = 2n$
    end while
    if $|r_\Xi(\lambda^*)| > |r_\Xi(\bar{\lambda})|$ **then**
      $\bar{\lambda} = \lambda^*$
    end if
  end for
  return $\bar{\lambda}$
end function

function ComputeCertificate(λ_L, λ_U, Ξ)
  $\bar{\lambda} = \text{ComputeInnerMax}(\Xi)$
  $\Delta_\Xi = \max\{|r_\Xi(\lambda_L)|, |r_\Xi(\bar{\lambda})|, |r_\Xi(\lambda_U)|\}$
  return Δ_Ξ
end function

Output: Error certificate $\Delta_\Xi = \|r_\Xi\|_\Sigma$ with $\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{V} f^\tau(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}\| \leq 2c_k^\tau \Delta_\Xi \|\mathbf{b}\|$, where c_k^τ is as in (10.43).

10.3 Novel Pole Selection Algorithms

In real-world scenarios, one is interested in identifying the smallest parameter $k \in \mathbb{N}$ such that the approximation error is smaller than a user-defined threshold. One possibility to achieve this is to adaptively construct a pole set Ξ, compute the error certificate using Algorithm 2, and stop the procedure once the upper bound is smaller than the desired tolerance. We propose, using a similar concept as the one employed in [Bag69, DLZ10, DS11, GK13], a novel pole distribution $\mathcal{A} = \{\xi_1, \ldots, \xi_k\}$ which combines the first two stages and builds the rational Krylov space on the basis of the error certificate. The scheme is specified in Algorithm 3 and shall serve us as definition for the pole set $\mathcal{A}$, whose elements we call *automatic poles on* $-\Sigma$.

In accordance with our previous notation, we set $\mathcal{A}_\infty := \{\infty, \xi_1, \ldots, \xi_k\}$. The automatic poles on $-\Sigma$ are nested, independent of $\mathbf{b}$ and the parameter $\boldsymbol{\tau}$, and only require the knowledge of some rough extremal bounds for the spectral interval. They aim to approximate Zolotarëv's minimal deviation problem in a greedy manner. Due to the symmetric relation

Algorithm 3 Automatic Pole Selection Algorithm - $\mathcal{A}$

Input: tolerance $\varepsilon \in \mathbb{R}^+$, $0 < \lambda_L < \lambda_U$ with $\Sigma \subset [\lambda_L, \lambda_U]$, $\mathbf{b} \in \mathbb{R}^N$

$\xi_1 = -\lambda_L$
$\xi_2 = -\lambda_U$
$\mathcal{A} = \{\xi_1, \xi_2\}$
$k = 2$
do
  $\bar{\lambda} = \textsc{ComputeInnerMax}(\mathcal{A})$
  $\Delta_{\mathcal{A}} = |r_{\mathcal{A}}(\bar{\lambda})|$ // $|r_{\mathcal{A}}(\lambda)| = 0$ for $\lambda \in \{\lambda_L, \lambda_U\}$
  $\xi_{k+1} = -\bar{\lambda}$
  $\mathcal{A} = \mathcal{A} \cup \{\xi_{k+1}\}$
  Relabel the poles so that $\xi_{k+1} < \cdots < \xi_1$
  $k = k + 1$
while $2c_k^\tau \Delta_{\mathcal{A}} \|\mathbf{b}\| > \varepsilon$ // c_k^τ defined as in (10.43)

Output: Pole set $\mathcal{A}$ such that $\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| < \varepsilon$.

between its roots and poles, the global maximum of $|r_{\mathcal{A}}|$ can be detected automatically without the risk of missing a critical value. While $\mathcal{A}$ does not adapt to the actual spectrum of $\mathbf{L}$ like the spectral poles $\mathcal{S}$ on $-\Sigma$, the latter only approximates the modified Zolotarëv problem (10.21) which does not possess such a symmetry and thus requires to extract the maximizer over a discrete training set. Moreover, the computation of $\mathcal{A}$ allows us to directly access the maximal deviation of $r_{\mathcal{A}}$ over Σ to obtain the certificate provided by Algorithm 2 as a by-product. We cannot give a proof that our greedy algorithm generates an asymptotically optimal solution to Zolotarëv's minimal deviation problem. Nevertheless, our empirical findings reported in Figure 10.8 indicate that the algorithm possesses such a property.

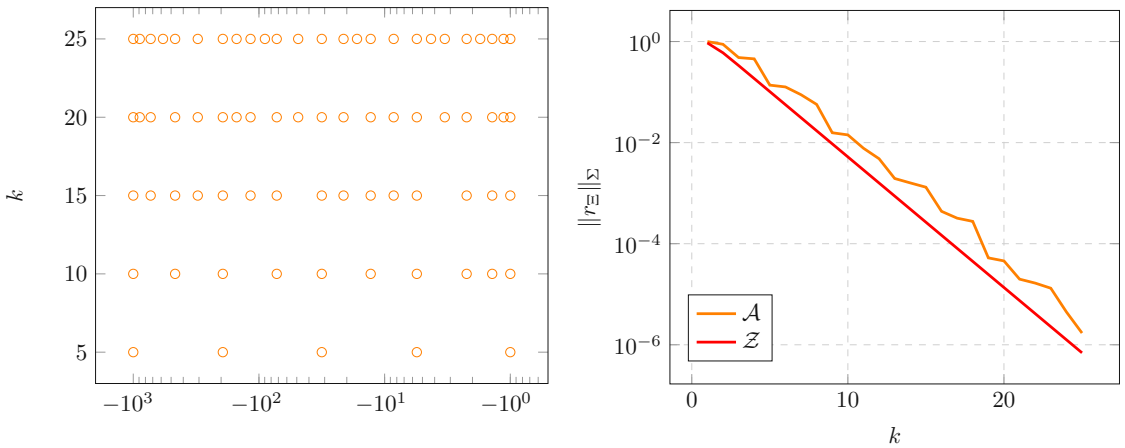

Figure 10.8: Automatic poles $\mathcal{A}$ on $-\Sigma = [-1000, -1]$ for different values of k (left) and the maximum norm $\|r_\Xi\|_\Sigma$ for $\Xi \in \{\mathcal{A}, \mathcal{Z}\}$ (right).

In a sense, Algorithm 3 is not fully automatic since it necessitates the availability of some

rough spectral bounds of the matrix $\mathbf{L}$. A heuristic approach to overcome this restriction is based on the observation that the extremal eigenvalues of $\mathbf{L}_{k+1}$ typically provide good approximations to the extremal eigenvalues of $\mathbf{L}$. In light of the fact that the rational Ritz values are contained in Σ, an automated variant of Algorithm 3 is obtained by iteratively adapting the underlying spectral interval based on the extremal eigenvalues of $\mathbf{L}_{k+1}$. This allows us to generate information about the spectral region without any user-provided data. Recognizing this fact, we present the fully automatic pole selection strategy for incrementally building the set $\mathcal{F} = \{\xi_1, \dots, \xi_k\}$ in Algorithm 4. We call $\mathcal{F}$ *fully automatic poles on* $-\Sigma$ and set $\mathcal{F}_\infty := \{\infty, \xi_1, \dots, \xi_k\}$.

Algorithm 4 Fully Automatic Pole Selection Algorithm - $\mathcal{F}$

Input: tolerance $\varepsilon \in \mathbb{R}^+$, $\mathbf{L} \in \mathbb{R}^{N\times N}$, $\mathbf{b} \in \mathbb{R}^N$

$\mathbf{V} = \textsc{RationalArnoldi}(\mathbf{L}, \mathbf{b}, \Xi = \{\infty, \infty\})$
$\mathbf{L}_2 = \mathbf{V}^\dagger \mathbf{L}\mathbf{V}$
Compute eigenvalues $\mu_0^{(1)}$, $\mu_1^{(1)}$ of $\mathbf{L}_2$ and set $\mu_{\min} = \mu_0^{(1)}$, $\mu_{\max} = \mu_1^{(1)}$
$\xi_1 = -\mu_{\min}$, $\xi_2 = -\mu_{\max}$
$\mathcal{F} = \{\xi_1, \xi_2\}$
$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2] = \textsc{RationalArnoldi}(\mathbf{L}, \mathbf{b}, \Xi = \mathcal{F})$
$k = 2$
do
    $\bar{\lambda} = \textsc{ComputeInnerMax}(\mathcal{F})$
    $\xi_{k+1} = -\arg\max\{|r_\mathcal{F}(\mu_{\min})|, |r_\mathcal{F}(\bar{\lambda})|, |r_\mathcal{F}(\mu_{\max})|\}$
    $\Delta_\mathcal{F} = |r_\mathcal{F}(-\xi_{k+1})|$
    $\mathcal{F} = \mathcal{F} \cup \{\xi_{k+1}\}$
    $\mathbf{w} = (\mathbf{I} - \xi_{k+1}^{-1}\mathbf{L})^{-1}\mathbf{L}\mathbf{v}_k$
    Orthonormalize $\mathbf{w}$ against $\mathbf{V}$ to obtain new basis vector $\mathbf{v}_{k+1}$
    Set $\mathbf{V} = [\mathbf{V}, \mathbf{v}_{k+1}]$
    $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L}\mathbf{V}$
    Compute extremal eigenvalues $\mu_0^{(k)}$ and $\mu_k^{(k)}$ of $\mathbf{L}_{k+1}$ and set $\mu_{\min} = \mu_0^{(k)}$, $\mu_{\max} = \mu_k^{(k)}$
    Relabel the poles so that $\xi_{k+1} < \cdots < \xi_1$
    $k = k + 1$
while $2c_k^\tau \|\mathbf{b}\| \Delta_\mathcal{F} > \varepsilon$ // c_k^τ defined as in (10.43)

Output: Pole set $\mathcal{F}$ such that $\|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{u}_{k+1}\| < \varepsilon$.

Just like $\mathcal{A}$, the fully automatic poles on $-\Sigma$ are independent of $\mathbf{b}$ and, more importantly, the parameter $\boldsymbol{\tau}$. They can be viewed as greedy approximation of Zolotarëv's minimal deviation problem where the underlying spectral interval Σ of $\mathbf{L}$ is adaptively replaced by the spectral interval of the compression $\mathbf{L}_{k+1}$. As a consequence, the pole set $\mathcal{F}$ requires the computation of the smallest and largest rational Ritz value of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^\mathcal{F}(\mathbf{L}, \mathbf{b})$ in each iteration, which makes its computation more demanding than the one of $\mathcal{A}$. The former, however, can be computed even if no information about the spectral region is available.

We illustrate the pole set $\mathcal{F}$ for different values of k in Figure 10.9 for $\lambda_L = 1$ and $\lambda_U = 1000$ using the same matrices as in Figure 10.4. For $k = 5$ and the matrix $\mathbf{L}_1$, the extremal rational Ritz values are $\mu_0^{(5)} = 2.3$ and $\mu_5^{(5)} = 855.8$, which explains why $\mathcal{F}$ is

geometrically distributed across $[-\mu_5^{(5)}, -\mu_0^{(5)}]$ instead of $-\Sigma$. For increasing orders, the approximation of Σ through the smallest and largest rational Ritz value improves such that $\mathcal{F}$ does not significantly differ from the automatic pole set $\mathcal{A}$. For the matrix $\mathbf{L}_2$, we observe $[\mu_0^{(5)}, \mu_5^{(5)}] \approx [1.03, 1000]$ so that the true spectral interval of $\mathbf{L}_2$ is well approximated already for small values of k. Therefore, it is not surprising that $\|r_{\mathcal{F}}\|_\Sigma$ yields similar results compared to the ones obtained by $\mathcal{A}$, as shown in Figure 10.10.

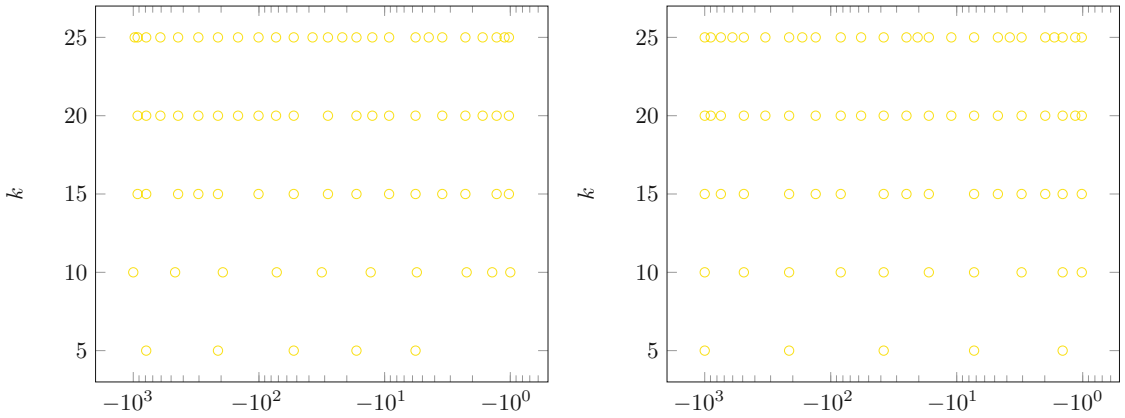

Figure 10.9: Fully automatic poles $\mathcal{F}$ on $-\Sigma = [-1000, -1]$ for $\mathbf{L}_1$ (left) and $\mathbf{L}_2$ (right).

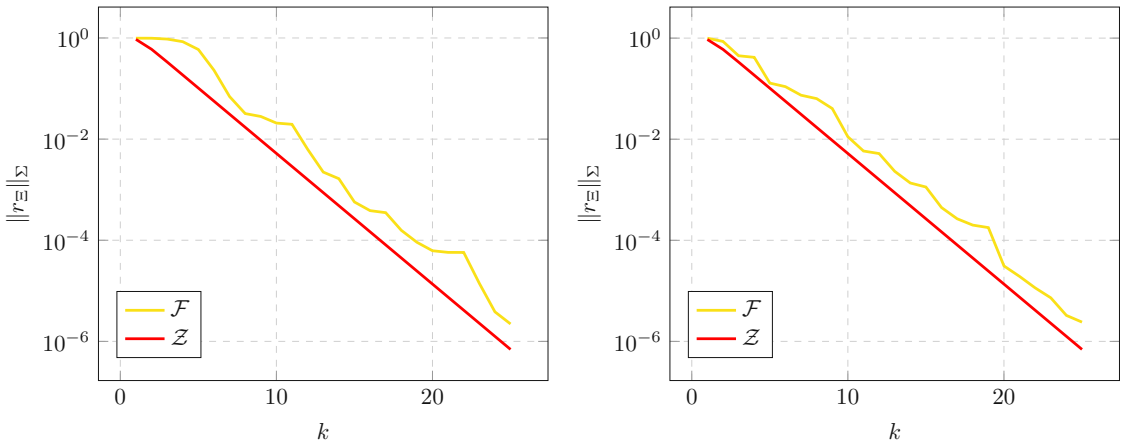

Figure 10.10: Maximum norm $\|r_\Xi\|_\Sigma$ on $\Sigma = [1, 1000]$ for $\Xi \in \{\mathcal{F}, \mathcal{Z}\}$ with $\mathbf{L}_1$ (left) and $\mathbf{L}_2$ (right).

10.4 Numerical Examples

In the final section of this chapter we underpin the effectiveness of the presented poles and compare their performance in the course of a few prototypical fractional diffusion problems. Throughout, we set $\mathcal{L} = -\Delta$ and denote with $V_h = \mathcal{P}_1^0(\mathcal{T}_h)$ the Lagrangian finite element

space of order one with vanishing trace constructed through a quasi-uniform triangular mesh on $\Omega = (0,1)^2$ of mesh size $h = 0.08$. The FEM space gives rise to the matrix approximation $\mathbf{L} = \mathbf{M}^{-1}\mathbf{A}$ of the Laplacian, where $\mathbf{M}$ and $\mathbf{A}$ label the mass and stiffness matrix, respectively. We choose $\mathbf{b}$ to be the coefficient vector of the L_2-orthogonal projection of the constant 1 function onto V_h in each of our experiments.

All numerical examples are implemented within the finite element library Netgen/NGSolve[2] [Sch97, Sch14]. The evaluation of the Mittag-Leffler function is performed using the `jscatter` software package[3]. Further details on the implementation of the poles are listed below.

- We define $\lambda_L := 19$ and $\lambda_U := 489580$ to be our upper and lower bounds of the extremal eigenvalues of $\mathbf{L}$, obtained by a numerical approximation.
- We use the special function library from `Scipy`[4] to evaluate the elliptic integrals and the Jacobi elliptic functions in the computation of $\Xi \in \{\mathcal{Z}, \hat{\mathcal{Z}}, \mathcal{E}, \hat{\mathcal{E}}\}$.
- The spectral poles $\mathcal{S}$ are computed using a discrete training set $\mathcal{T}_{\text{train}} \subset [-\lambda_U, -\lambda_L]$ consisting of 10^6 geometrically distributed sampling points. For the pole set $\hat{\mathcal{S}}$, we set $\mathrm{n}_\mathrm{c}^- = \mathrm{n}_\mathrm{c}^+ = 20$ and use 10^6 equispaced sampling points in $[-\mathrm{n}_\mathrm{c}^-, \mathrm{n}_\mathrm{c}^+]$ to obtain the training set $\mathcal{T}_{\text{train}}^{\mathrm{n}_\mathrm{c}^\pm}$ after the transformation $\lambda \mapsto -e^\lambda$.
- The weak greedy poles $\mathcal{G}$ are computed using a discrete training set $\mathcal{T}_{\text{train}} \subset [\lambda_L, \lambda_U]$ consisting of 10^4 geometrically distributed sampling points. For the pole set $\hat{\mathcal{G}}$, we choose $\mathrm{n}_\mathrm{c}^- = \mathrm{n}_\mathrm{c}^+ = 20$ and use 10^4 equispaced sampling points in $[-\mathrm{n}_\mathrm{c}^-, \mathrm{n}_\mathrm{c}^+]$ to obtain the training set $\mathcal{T}_{\text{train}}^{\mathrm{n}_\mathrm{c}^\pm}$ after the transformation $\lambda \mapsto e^\lambda$.
- The BURA poles are computed using the BRASIL algorithm [Hof21] contained in the `baryrat`[5] Python package.

10.4.1 Parameter Study

The goal of this section is to illuminate the impact of the parameters on the rational Krylov approximation. In particular, we are interested in the limit case as the fractional parameters approach an integer or the time t approaches zero. Throughout, we limit ourselves to the study of parameter independent pole sets. To make matters precise, we introduce the discrete L_2-error

$$\mathrm{E}(k, \Xi) := \|f^\tau(\mathbf{L})\mathbf{b} - \mathbf{V} f^\tau(\mathbf{L}_{k+1})\mathbf{V}^\dagger \mathbf{b}\|, \tag{10.46}$$

where $\mathbf{V}$ is an orthonormal basis of $\mathcal{Q}_{k+1}^\Xi(\mathbf{L}, \mathbf{b})$. Starting with the stationary problem, we depict the error $\mathrm{E}(6, \Xi)$ for $f^\tau(\lambda) = \lambda^s$, $s \in [-1, 1]$, and different configurations of Ξ in Figure 10.11. In light of Theorem 10.2 and 10.6, the error is expected to behave like

$$\lambda^*(s) := \begin{cases} \lambda_L^s, & \text{if } s \in [-1, 0], \\ \lambda_U^s, & \text{if } s \in (0, 1], \end{cases}$$

[2] https://ngsolve.org/
[3] https://pypi.org/project/jscatter/
[4] https://docs.scipy.org/doc/scipy/reference/special.html
[5] https://github.com/c-f-h/baryrat

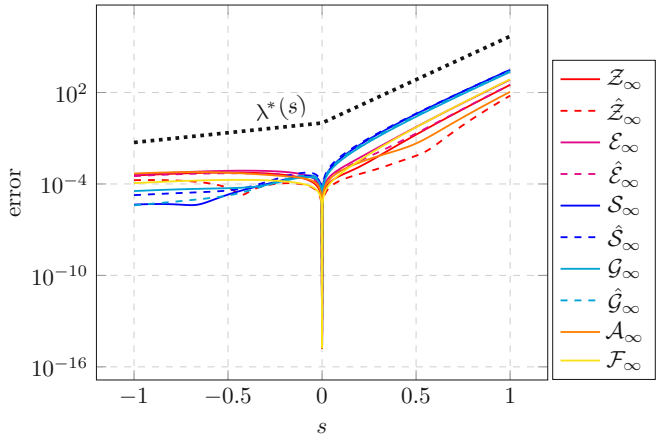

Figure 10.11: Error $\mathrm{E}(6, \Xi)$ for $f^{\tau}(\lambda) = \lambda^s$, $s \in [-1, 1]$, and various pole sets Ξ.

whenever $\Xi \in \{\mathcal{Z}, \hat{\mathcal{Z}}\}$. The example shows that this is indeed the case, even more, the same seemingly applies also to all other pole sets. In particular, we observe for $\Xi = \hat{\mathcal{G}}_\infty$ that the upper bound obtained by Theorem 10.24 is to pessimistic as s approaches an integer. At $s = 0$ the error reaches machine precision for all pole configurations since $\mathbf{L}^0\mathbf{b} = \mathbf{b} \in \mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ in which case the RKM is exact. Furthermore, it seems that the RKM error of $\hat{\mathcal{Z}}$ remains bounded as $s \to 1^-$ even without the additional pole at infinity, advocated by the second part of Theorem 10.4.

In the time-dependent regime one is interested in (10.46) for $f^{\tau}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s)$. We fix $\beta = 1$ and report the evolution of this quantity in Figure 10.12 using $(\alpha, s) \in \{(0.5, 0.5), (0.6, 0.75)\}$. In the case of $(\alpha, s) = (0.5, 0.5)$ we are, according to Proposition 8.22, in the Cauchy-Stieltjes regime. Theorem 10.2 and 10.6 predict that the error decays like $\mathcal{O}(E_{\alpha,1}(-\sqrt{t\lambda_L}))$ when $t \to \infty$ and $\Xi \in \{\mathcal{Z}, \hat{\mathcal{Z}}\}$, which is precisely what we observe in Figure 10.12. Since $E_{\alpha,1}(-t^\alpha \lambda^s) \in \mathcal{LS} \setminus \mathcal{CS}$ whenever $s + \frac{\alpha}{2} \geq 1$, we cannot confirm analytically that the error satisfies such a property if $(\alpha, s) = (0.6, 0.75)$ or $\Xi \in \{\mathcal{E}, \hat{\mathcal{E}}, \mathcal{S}, \hat{\mathcal{S}}, \mathcal{G}, \hat{\mathcal{G}}, \mathcal{A}, \mathcal{F}\}$. However, our numerical experiments suggest that (10.46) can be bounded using $f^{\tau}(\lambda_L)$ regardless of the parameters and the poles.

Assuming that β is bounded away from zero, our analysis assures that the rational Krylov surrogate of $E_{\alpha,\beta}(-t^\alpha \mathbf{L}^s)\mathbf{b}$ converges uniformly in the parameters when Zolotarëv's poles are used. To confirm that $\mathrm{E}(6, \Xi)$ does not degenerate for *any* pole selection provided in this chapter, we plot the limiting behaviour of the error for $t \to 0$ in Figure 10.13. Not only does the error remain uniformly bounded, it even converges to zero like $\mathcal{O}(t^\alpha)$. This is due to $E_{\alpha,\beta}(-t^\alpha \lambda^s) \equiv 1$ for $t = 0$, in which case the rational Krylov approximation is exact.

To understand the sensitivity of the error with respect to the fractional parameters, we fix $\beta = 1$ and $t = 1.5$ to illustrate the spatial error as function of α and s in Figure 10.14. The quantity $\mathrm{E}(6, \mathcal{Z}_\infty)$ is evaluated over a discrete parameter grid contained in $[0, 1]^2$, where the extended definition $f^{\tau}(\lambda) = \mathrm{e}_{\alpha,\beta}(-t^\alpha, \lambda^s)$, defined by (8.21), is used. Whenever the Euclidean norm of $(\alpha, s) \in \mathbb{R}^2$ is close to $\sqrt{2}$ or $s \ll 1$, we see that the error is small compared to other configurations of the fractional parameters. The former, in a sense, underpins our observations from Figure 10.12 that the error is proportional to $f^{\tau}(\lambda_L)$

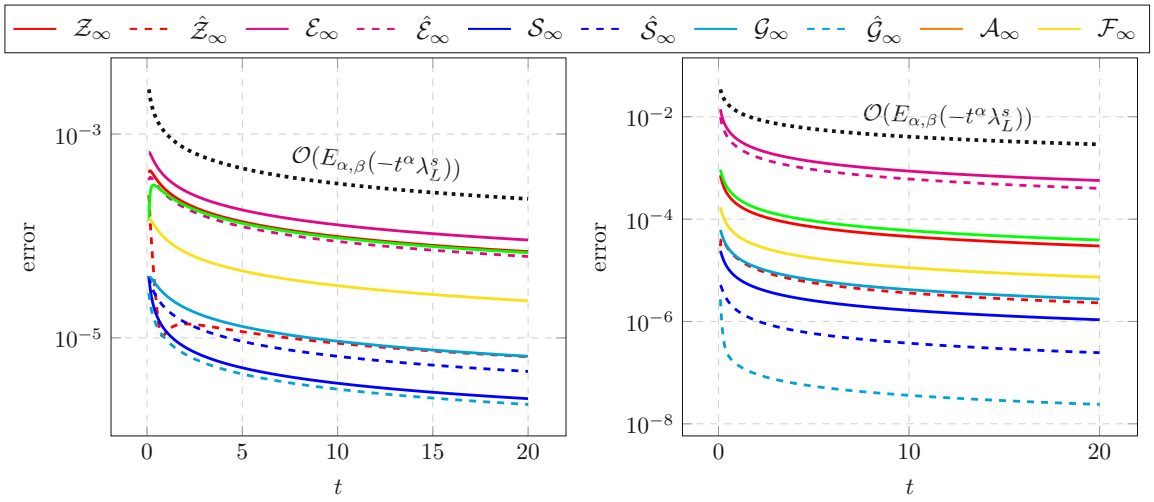

Figure 10.12: Error $\mathrm{E}(6,\Xi)$ for $f^{\mathcal{T}}(\lambda) = E_{\alpha,1}(-t^\alpha\lambda^s)$ with $(\alpha,s) = (0.5,0.5)$ (left) and $(\alpha,s) = (0.6,0.75)$ (right) with $t \in [0.1,20]$ and various pole sets Ξ.

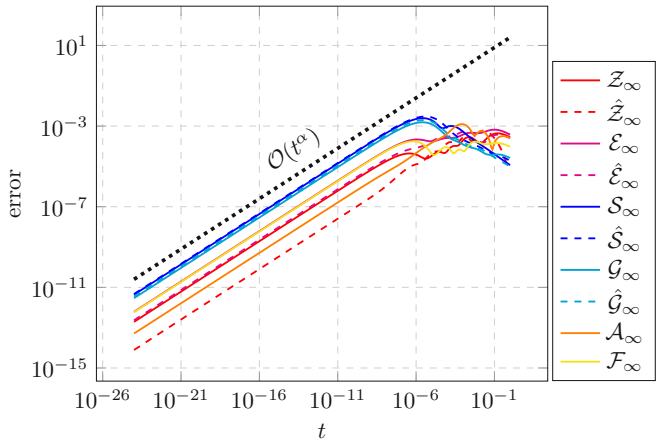

Figure 10.13: Error $\mathrm{E}(6,\Xi)$ for $f^{\mathcal{T}}(\lambda) = E_{\alpha,1}(-t^\alpha\lambda^s)$ with $(\alpha,s) = (0.6,0.75)$, various pole sets Ξ, and decreasing values of t.

irrespectively of $(\alpha,s) \in [0,1]^2$ and thus decrease whenever α or s approach one. Contrary to the proportionality to $f^{\mathcal{T}}(\lambda_L)$, the quality of the surrogate improves also for small values of the spatial fractional parameter. This can be seen as a local approximation effect since $\mathrm{e}_{\alpha,1}(-t^\alpha,\lambda^s) \equiv$ const. if $s = 0$, in which case the rational Krylov approximation is exact. In this regime, the error appears to be less prone to increasing values of α. For the other pole sets, we observe that the respective parameter plots look qualitatively similar.

10.4.2 Convergence Study

We now focus on a numerical confirmation of the convergence rates predicted by our analysis. Starting with the stationary case, we monitor the error involving $f^{\mathcal{T}}(\lambda) = \lambda^s$ for $s = -\frac{1}{2}$

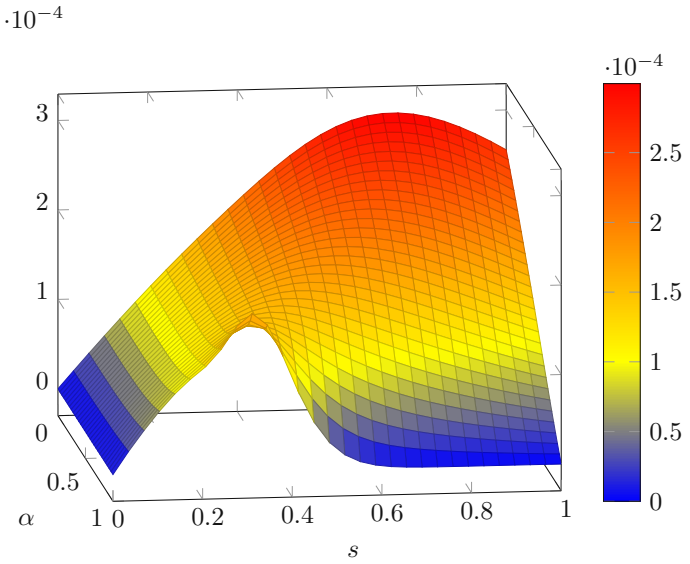

Figure 10.14: Error $\mathrm{E}(6, \mathcal{Z}_\infty)$ for $f^{\boldsymbol{\tau}}(\lambda) = \mathrm{e}_{\alpha,1}(-1.5^\alpha, \lambda^s)$ and $(\alpha, s) \in [0,1]^2$.

and $s = \frac{1}{2}$ in Figure 10.15.

- In accordance with Theorem 10.31, the BURA poles yield a decay rate of order $\mathcal{O}(\rho_{[\lambda_L, 4\lambda_U]}^{-2k})$ if $s = -\frac{1}{2}$. The same seemingly holds true if $s = \frac{1}{2}$, which indicates that our analysis, encoded in Theorem 10.30, might be too pessimistic in the regime of positive exponents s.
- While RKMs build upon $\hat{\mathcal{Z}}_\infty$ are expected to converge like $\rho_{[\lambda_L, 4\lambda_U]}^{-k}$, we observe experimentally that the error behaves competitive with the one obtained by $\mathcal{B}_{\boldsymbol{\tau}}^\infty$ and frequently reaches machine precision before the predicted convergence rates become visible. Unlike $\mathcal{B}_{\boldsymbol{\tau}}^\infty$, however, the very same search space can be employed to approximate $\mathbf{L}^s\mathbf{b}$ for any $s \in [-1, 1]$ if Zolotarëv's poles are used, while $\mathcal{Q}_{k+1}^{\mathcal{B}_{\boldsymbol{\tau}}}(\mathbf{L}, \mathbf{b})$ needs to be computed for $s = -\frac{1}{2}$ and $s = \frac{1}{2}$ individually.
- As predicted by Theorem 10.6, the surrogate extracted from $\mathcal{Q}_{k+1}^{\mathcal{Z}_\infty}(\mathbf{L}, \mathbf{b})$ converges like $\mathcal{O}(\rho_{[\lambda_L, \lambda_U]}^{-k/2})$ if $s = \frac{1}{2}$. In the Cauchy-Stieltjes regime, i.e., $s = -\frac{1}{2}$, the surrogate first outperforms the predictions but finally, after leaving the preasymptotic range, converges with the expected rate. A similar behaviour is observed for $\Xi \in \{\mathcal{E}, \mathcal{A}, \mathcal{F}\}$, which is reasonable since any of these poles imitates the optimality condition of $\mathcal{Z}$.
- Similarly to what is said above, the error $\mathrm{E}(k, \hat{\mathcal{E}}_\infty)$ for $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{\frac{s}{2}}$ decays with the rate predicted by the second claim in Theorem 10.14 but exceeds our expectations for $s = -\frac{1}{2}$.
- Even though Theorem 10.24 suggests that $\hat{\mathcal{G}}_\infty$ yields considerably slower convergence rates than $\hat{\mathcal{Z}}_\infty$, our experiment shows that the discrepancy between these two pole sets is not that dramatic. Both weak greedy configurations yield attractive *nested* alternatives to $\Xi \in \{\hat{\mathcal{Z}}_\infty, \mathcal{Z}_\infty\}$ and perform qualitatively similar to the EDS poles.

– Only for the spectral poles, we observe that the pole set on $-\Sigma$ outperforms its respective "hat" counterpart. For the other configurations, it might be worthwhile to invest in poles that are contained in the entire negative real axis. In this regime, however, our developed error certificate, presented in Algorithm 2, cannot be expected to provide a reliable prediction about the true approximation error.

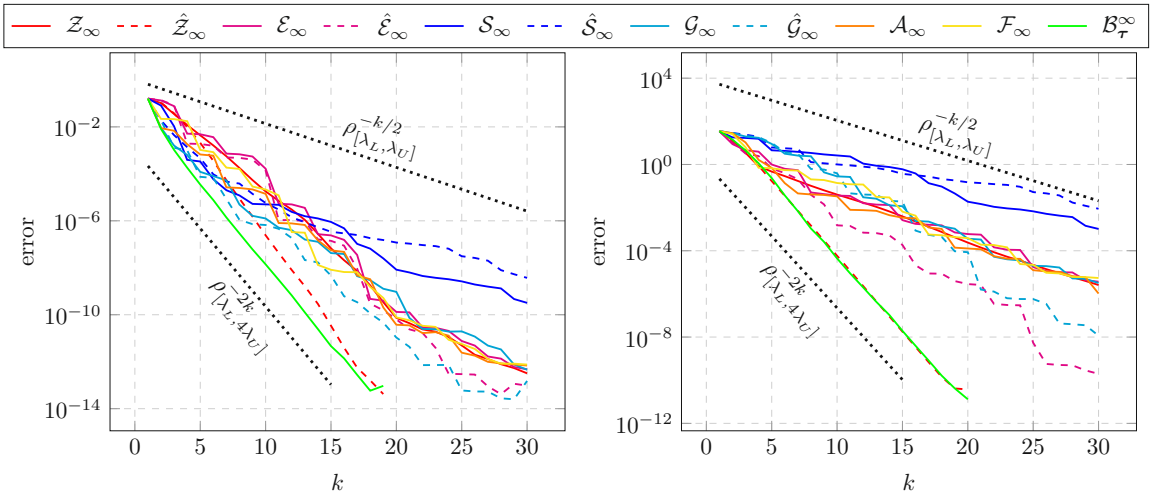

Figure 10.15: Error $\mathrm{E}(k,\Xi)$ for different pole sets, $f^{\boldsymbol{\tau}}(\lambda) = \lambda^s$, $s = -\frac{1}{2}$ (left), and $s = \frac{1}{2}$ (right).

We are also interested in the numerical approximation of time-dependent problems. For this purpose, we fix $\beta = 1$ and $t = 1.5$ to monitor (10.46) with $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s)$ as a function in k for $(\alpha, s) \in \{(0.5, 0.5), (0.9, 0.75)\}$ in Figure 10.16.

– Similarly to the stationary case, we observe that the rational Krylov errors of $\mathcal{Z}$, $\mathcal{E}$, $\mathcal{A}$, and $\mathcal{F}$ are almost coincident. In a sense, this is reasonable since all of these poles aim directly for a minimization of $\|r_\Xi\|_\Sigma$. The error initially decreases faster than predicted by our theory and frequently reaches machine precision before the expected convergence rates become visible.

– If $(\alpha, s) = (0.9, 0.75)$, there holds $\frac{\alpha}{2} + s \geq 1$ whence $f^{\boldsymbol{\tau}} \in \mathcal{LS} \setminus \mathcal{CS}$. Even though Theorem 10.2 and the second claim in Theorem 10.14 and 10.30 are no longer applicable, our experiment provides evidence that the rate of convergence proven therein remains valid even if $\frac{\alpha}{2} + s \geq 1$.

– Unlike in Figure 10.15, the BURA poles outperform each of its competitors by a significant margin. If $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ needs to be computed for one single value of $\boldsymbol{\tau}$ only, $\mathcal{B}^\infty_{\boldsymbol{\tau}}$ provides the best pole set for building the rational Krylov space.

Remark 10.35. *In view of Remark 5.24, it might be of interest to approximate $E_{\alpha,\beta}(-t^\alpha \mathbf{L}^s)\mathbf{b}$ for $\alpha > 1$. In this regime, however, we have $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha \lambda^s) \notin \mathcal{LS}$. Even though our analysis is not applicable in this case, we observe numerically that the poles listed above yield exponential convergence comparable to the ones obtained for $\alpha \in (0, 1]$.*

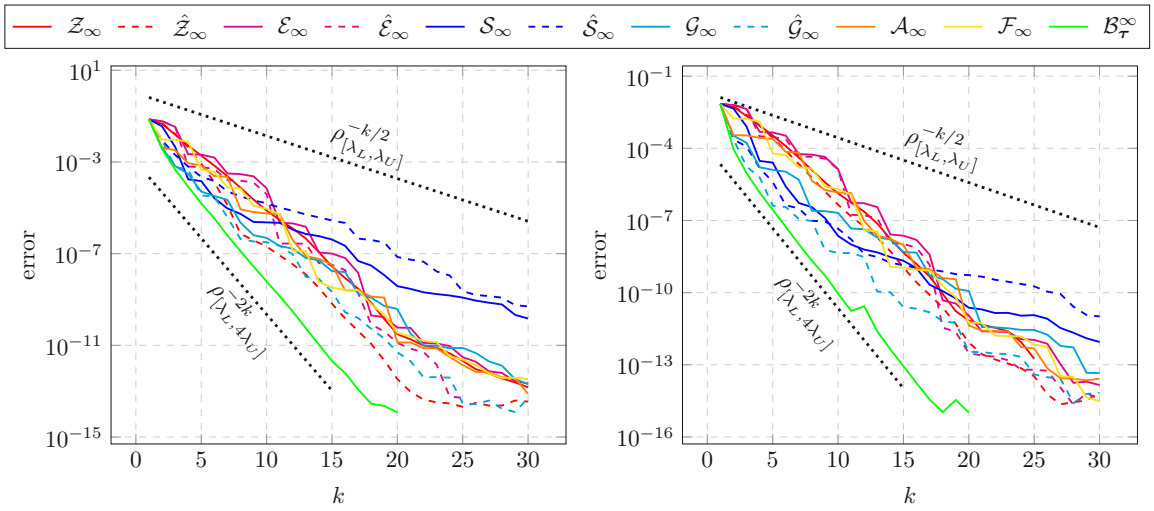

Figure 10.16: Error $\mathrm{E}(k,\Xi)$ for different pole sets, $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$, $\beta = 1$, $t = 1.5$, $(\alpha, s) = (0.5, 0.5)$ (left), and $(\alpha, s) = (0.9, 0.75)$ (right).

We conclude this section with a systematic comparison of the presented pole configurations in Table 10.1, incorporating (from top to bottom)

1. their ability to efficiently query the solution map $\boldsymbol{\tau} \mapsto f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ for multiple instances of $\boldsymbol{\tau}$ using the same search space,
2. nestedness of the poles as k increases, and hence their ability to incrementally improve the accuracy of the surrogate,
3. the required user-provided data,
4. their ability to adapt to the spectral density of $\mathbf{L}$,
5. their ability to adapt to the vector $\mathbf{b}$,
6. the availability of theoretical convergence results,
7. the availability of a (reasonable) error certificate.

With the exception of the BURA poles, all pole sets are independent of the parameter and thus suitable for multi-query problems in fractional diffusion. Among them, only $\mathcal{Z}$ and $\hat{\mathcal{Z}}$ are not nested. In terms of available analytical results, the poles can be classified in four groups. The first group contains $\mathcal{Z}$ and is the only one that allows for explicit error bounds for arbitrary $f^{\boldsymbol{\tau}}$ of Stieltjes and complete Bernstein type. The second group comprises $\hat{\mathcal{Z}}$ and $\hat{\mathcal{G}}$ which provide a rigorous analysis only for some functions of fractional diffusion type, namely $f^{\boldsymbol{\tau}} \in \mathcal{CS} \cup \mathcal{CB}$ and $f^{\boldsymbol{\tau}} \in \mathcal{CS}$, respectively. The BURA and EDS poles can be quantified in terms of their asymptotic convergence rates but explicit bounds for finite k are not available. The last group entails the remaining configuration where no analytical results are known to the author.

In contrast to their respective "hat"-counterpart, the pole sets $\mathcal{E}, \mathcal{S}$, and $\mathcal{G}$ are amenable to our developed error certificate. Their computation requires the availability of explicit bounds for $\lambda_{\min}$ and $\lambda_{\max}$. On the other hand, $\hat{\mathcal{S}}$ and $\hat{\mathcal{G}}$ require the choice of the cut-off parameters $\mathrm{n_c^-}, \mathrm{n_c^+}$ as the underlying parameter domain is unbounded. Clearly, Algorithm 2 cannot be consulted to assess the quality of $\mathcal{B}_{\boldsymbol{\tau}}^{\infty}$. Nevertheless one can resort to (10.45) to obtain a meaningful error indicator whenever the BURA poles are employed. Theoretically, our error certificate is applicable to $\mathcal{F}$. However, since $\mathcal{F}$ seeks to avoid the explicit computation of $\lambda_{\min}$ and $\lambda_{\max}$, computing $\|r_{\mathcal{F}}\|_{\Sigma}$ to control the rational Krylov error is only of limited use.

Pole set Ξ	$\mathcal{Z}$	$\hat{\mathcal{Z}}$	$\mathcal{E}$	$\hat{\mathcal{E}}$	$\mathcal{S}$	$\hat{\mathcal{S}}$	$\mathcal{G}$	$\hat{\mathcal{G}}$	$\mathcal{B}_{\boldsymbol{\tau}}$	$\mathcal{A}$	$\mathcal{F}$
multi-query	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
nested	×	×	✓	✓	✓	✓	✓	✓	×	✓	✓
user-provided	S	S	S	S	$\mathrm{S}, \mathcal{T}_{\mathrm{train}}$	$\mathcal{T}_{\mathrm{train}}^{\mathrm{n_c^{\pm}}}$	$\mathrm{S}, \mathcal{T}_{\mathrm{train}}$	$\mathcal{T}_{\mathrm{train}}^{\mathrm{n_c^{\pm}}}$	S	S	-
spectral adapt.	×	×	×	×	✓	✓	✓	✓	×	×	×
vector adapt.	×	×	×	×	×	×	✓	✓	×	×	×
analysis	✓	$\mathcal{CS}, \mathcal{CB}$	∼	∼	×	×	×	$\mathcal{CS}$	∼	×	×
certificate	✓	×	✓	×	✓	×	✓	×	✓	✓	∼

Table 10.1: Properties of the pole sets, where $\mathrm{S} = \{\lambda_L, \lambda_U\}$ contains bounds for the spectral region of $\mathbf{L}$, $\mathrm{n_c^-}, \mathrm{n_c^+} \in \mathbb{R}^+$ are cut-off parameters, and $\mathcal{T}_{\mathrm{train}}$, $\mathcal{T}_{\mathrm{train}}^{\mathrm{n_c^{\pm}}}$ training sets of the respective parameter domain.

11 Selected MOR Methods Based on Rational Approximation

In the previous section it is shown that rational Krylov methods provide an attractive tool to approximate solutions to fractional diffusion problems. The purpose of this chapter is to draw parallels to existing model order reduction schemes that are based on or related to rational approximation methods. In line with [DH21], we present the class of Reduced Basis Methods (RBM) for fractional diffusion problems and prove that they can be interpreted as variants of certain rational Krylov methods. These theoretical insights allow us to harness our analysis for RKMs to develop new convergence proofs for several of the studied schemes. They suggest how to design novel and improve available methods and allow for a direct comparison of the algorithms.

11.1 Rational Approximation Methods

In Chapter 6 it is shown that the evaluation of the accurate but expensive discrete eigenfunction method boils down to the evaluation of a matrix-vector product of the form $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$, where $\mathbf{L} \in \mathbb{R}^{N\times N}$ is the discrete approximation of the differential operator, $\mathbf{b} \in \mathbb{R}^N$ a vector, and $f^{\boldsymbol{\tau}}$ a parametric function. Instead of computing the matrix function exactly, one class of methods replaces $f^{\boldsymbol{\tau}}$ by a suitable rational function $r_k^{\boldsymbol{\tau}} \in \mathcal{R}_{k,k}$ such that

$$\mathbf{u}_{r_k^{\boldsymbol{\tau}}} := r_k^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} \approx f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}. \tag{11.1}$$

Any such method is called *rational approximation method* [Hof20, DH21] and has been applied in e.g., [HLM+18, HLM+20, AN20, HKL+21b, HKL+21a, DH21, Vab21c]. The computational benefit of $\mathbf{u}_{r_k^{\boldsymbol{\tau}}}$ compared to $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ is due to the partial fraction decomposition of $r_k^{\boldsymbol{\tau}}$,

$$r_k^{\boldsymbol{\tau}}(\lambda) = c_0^{\boldsymbol{\tau}} + \sum_{j=1}^{k} \frac{c_j^{\boldsymbol{\tau}}}{\lambda - \xi_j^{\boldsymbol{\tau}}}, \tag{11.2}$$

where $(c_j^{\boldsymbol{\tau}})_{j=0}^k$ are the *residues* and $(\xi_j^{\boldsymbol{\tau}})_{j=1}^k$ the *poles*, which we assume to be pairwise distinct for simplicity. Thanks to (11.2), the computation of $\mathbf{u}_{r_k^{\boldsymbol{\tau}}}$ does not require the availability of *all* eigenvectors of $\mathbf{L}$ but instead can be obtained by

$$\mathbf{u}_{r_k^{\boldsymbol{\tau}}} = c_0^{\boldsymbol{\tau}}\mathbf{b} + \sum_{j=1}^{k} c_j^{\boldsymbol{\tau}}(\mathbf{L} - \xi_j^{\boldsymbol{\tau}}\mathbf{I})^{-1}\mathbf{b},$$

which only involves k solves to shifted linear systems of equations. As noted in [Hof20], the discrepancy between $\mathbf{u}_{r_k^{\boldsymbol{\tau}}}$ and the exact matrix-vector product can be bounded directly in terms of the approximation quality of the scalar function $r_k^{\boldsymbol{\tau}} \approx f^{\boldsymbol{\tau}}$.

Theorem 11.1. *For all $r_k^{\boldsymbol{\tau}} \in \mathcal{R}_{k,k}$ there holds*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{r_k^{\boldsymbol{\tau}}}\| \leq \|f^{\boldsymbol{\tau}} - r_k^{\boldsymbol{\tau}}\|_{\Sigma}\|\mathbf{b}\|. \tag{11.3}$$

Proof. Let $\mathbf{U} \in \mathbb{R}^{N\times N}$ denote the matrix of eigenvectors of $\mathbf{L}$ and $\mathbf{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_N)$ the diagonal matrix containing its eigenvalues. Invoking

$$f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{r_k^{\boldsymbol{\tau}}} = \mathbf{U}(f^{\boldsymbol{\tau}}(\mathbf{D}) - r_k^{\boldsymbol{\tau}}(\mathbf{D}))\mathbf{U}^{-1}\mathbf{b}$$

together with Lemma 6.7, we find

$$\begin{aligned}\|f^{\boldsymbol{\tau}}(\mathbf{b}) - \mathbf{u}_{r_k^{\boldsymbol{\tau}}}\| &= \|(f^{\boldsymbol{\tau}}(\mathbf{D}) - r_k^{\boldsymbol{\tau}}(\mathbf{D}))\mathbf{U}^{-1}\mathbf{b}\|_2 \\ &\leq \max_{j=1,\dots,N} |f^{\boldsymbol{\tau}}(\lambda_j) - r_k^{\boldsymbol{\tau}}(\lambda_j)|\|\mathbf{U}^{-1}\mathbf{b}\|_2 \leq \|f^{\boldsymbol{\tau}} - r_k^{\boldsymbol{\tau}}\|_{\Sigma}\|\mathbf{b}\|. \end{aligned}$$ □

Note the close relation of this result to the one stated in Theorem 7.16. Roughly spoken, the error obtained using the rational Krylov method with given poles $\xi_0 = \infty$ and $(\xi_j)_{j=1}^k \subset \mathbb{R} \setminus \Sigma$ is twice as large as the error obtained using the best possible rational approximation method having these same poles.

Remark 11.2. *Similar to Remark 7.17, (11.3) can be improved to a rational approximation problem on the discrete spectrum $\sigma(\mathbf{L})$ of $\mathbf{L}$*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{r_k^{\boldsymbol{\tau}}}\| \leq \|f^{\boldsymbol{\tau}} - r_k^{\boldsymbol{\tau}}\|_{\sigma(\mathbf{L})}\|\mathbf{b}\|,$$

which might be smaller than the one posed on the continuum Σ.

In view of Theorem 7.14, any RKM falls in the class of rational approximation methods, but whereas the denominator of the rational function is fixed (via selection of the pole set Ξ) a priori, the numerator is determined automatically via Rayleigh-Ritz extraction. In principle, this allows RKMs to approximate $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ for multiple instances of the parameter while a direct choice of the residues $(c_j^{\boldsymbol{\tau}})_{j=0}^k$ typically approximates $f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b}$ only for one single value of $\boldsymbol{\tau}$ reasonably well.

11.1.1 Direct Rational Approximation - The BURA Method

A variant of rational approximation methods are so-called *direct rational approximation methods*, where the numerator the denominator of $r_k^{\boldsymbol{\tau}} \in \mathcal{R}_{k,k}$ are chosen a priori. In view of Theorem 11.1, it is desirable to choose $r_k^{\boldsymbol{\tau}}$ as best uniform rational approximation of $f^{\boldsymbol{\tau}}$ in Σ [HLM+18, HLM+20, DH21, HKL+21b, HKL+21a], which can be computed numerically e.g., by means of the BRASIL algorithm developed in [Hof21]. As opposed to RKMs based on BURA poles, these schemes require the knowledge of the residues $(c_j^{\boldsymbol{\tau}})_{j=0}^k$ in (11.2) which makes them typically more prone to round-off errors. Due to the close relation between Theorem 7.16 and 11.1, it is not surprising that the analysis of RKMs with BURA poles immediately carries over to direct rational approximation schemes whenever $r_k^{\boldsymbol{\tau}}$ is chosen to be the BURA of $f^{\boldsymbol{\tau}}$ in Σ. For completeness, we state here the respective counterpart to Theorem 10.30 and 10.31.

Theorem 11.3. *Let $r_k^{\mathcal{B}_{\boldsymbol{\tau}}}$ denote the BURA of $f^{\boldsymbol{\tau}}$ in Σ.*

1. *If* $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ *and* $s \in (0,1)$*, then*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{r_k^{\mathcal{B}_{\boldsymbol{\tau}}}}\| \preceq \rho^{-2k}_{[\lambda_{\min},4\lambda_{\max}]}\|\mathbf{b}\|.$$

2. *If* $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{s}$ *and* $s \in (0,1)$*, then*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{r_k^{\mathcal{B}_{\boldsymbol{\tau}}}}\| \preceq \rho^{-k}_{[\lambda_{\min},4\lambda_{\max}]}\|\mathbf{b}\|.$$

3. *If* $f^{\boldsymbol{\tau}}(\lambda) = E_{\alpha,\beta}(-t^\alpha\lambda^s)$ *and* $(\alpha,\beta,t,s) \in \Theta_{\mathrm{L}}$*, then there holds*

$$\|f^{\boldsymbol{\tau}}(\mathbf{L})\mathbf{b} - \mathbf{u}_{r_k^{\mathcal{B}_{\boldsymbol{\tau}}}}\| \preceq \rho^{-\frac{k}{2}}_{[\lambda_{\min},\lambda_{\max}]}\|\mathbf{b}\|. \tag{11.4}$$

Moreover, if $(\alpha,\beta,t,s) \in \Theta_{\mathrm{C}}$*, then* (11.4) *remains valid if we replace* $\rho^{-\frac{k}{2}}_{[\lambda_{\min},\lambda_{\max}]}$ *by* $\rho^{-k}_{[\lambda_{\min},4\lambda_{\max}]}$.

11.1.2 Reduced Basis Methods

Several recently proposed numerical schemes exploit the fact that the nonlocal character of the fractional operator can be circumvented at the cost of parametric solutions to classical reaction-diffusion problems. The reduced basis method [RHP08, QRM11, HRS15, QMN15] is a prevalent choice for reducing the computational effort in the evaluation of these solutions for multiple instances of the parameter and has recently sparked a considerable amount of research activity in the fractional diffusion community [WGP17, ACN19, DS19, BGZ20, DS21, ACR21, DH21]. In this section, we show that several recently proposed schemes which are based on RBMs admit a representation in the rational Krylov framework. To make matters precise, we consider the discrete parametric reaction-diffusion equation

$$(\mathbf{L} + \zeta\mathbf{I})\mathbf{w}(\zeta) = \mathbf{b}$$

for a prescribed right-hand side $\mathbf{b} \in \mathbb{R}^n$ and a parameter $\zeta \in \overline{\mathbb{R}}_0^+ := \mathbb{R}_0^+ \cup \{\infty\}$ that encodes the variability of the problem. We set $\mathbf{w}(\infty) := \mathbf{b}$ by convention. The RBM seeks to approximate the manifold of solutions $(\mathbf{w}(\zeta))_{\zeta\in\mathbb{R}_0^+}$ in the low-dimensional *reduced basis space*

$$\mathcal{V}^Z_{k+1}(\mathbf{L},\mathbf{b}) := \operatorname{span}\{\mathbf{w}(\zeta_0),\dots,\mathbf{w}(\zeta_k)\}, \tag{11.5}$$

where $0 \le \zeta_0 < \cdots < \zeta_k$ are particular parameters which we refer to as *snapshots*[1] throughout this section and $Z := \{\zeta_0,\dots,\zeta_k\}$. There holds $\dim \mathcal{V}^Z_{k+1}(\mathbf{L},\mathbf{b}) = |Z| = k+1$ if $\mathbf{b}$ is excited by sufficiently many eigenfunctions of $\mathbf{L}$ [DS19]. Recalling (7.5), the reduced basis surrogate $\mathbf{w}_{k+1}(\zeta) \in \mathcal{V}^Z_{k+1}(\mathbf{L},\mathbf{b})$ of $\mathbf{w}(\zeta)$ is computed via Galerkin projection, i.e.,

$$\forall \mathbf{v} \in \mathcal{V}^Z_{k+1}(\mathbf{L},\mathbf{b}): \quad (\mathbf{b} - (\mathbf{L}+\zeta\mathbf{I})\mathbf{w}_{k+1}(\zeta), \mathbf{v}) = 0, \tag{11.6}$$

and is uniquely defined by this condition. As shown in [Güt10], it coincides with the Rayleigh-Ritz approximation of $\mathbf{w}(\zeta)$ extracted from the reduced basis space.

[1] Our terminology differs from standard RBM notation, where the term *snapshot* is typically employed to refer to the solution $\mathbf{w}(\zeta_j)$ instead of the parameter ζ_j itself.

Lemma 11.4. *Let* $\mathbf{W}$ *be an orthonormal basis of* $\mathcal{V}_{k+1}^Z(\mathbf{L},\mathbf{b})$, $\hat{\mathbf{L}}_{k+1} = \mathbf{W}^\dagger\mathbf{L}\mathbf{W}$, *and* $\mathbf{w}_{k+1}(\zeta)$ *the reduced basis approximation of* $\mathbf{w}(\zeta)$ *with snapshots in* Z. *Then there holds*

$$\mathbf{w}_{k+1}(\zeta) = \mathbf{W}(\hat{\mathbf{L}}_{k+1} + \zeta\mathbf{I}_{k+1})^{-1}\mathbf{W}^\dagger\mathbf{b}.$$

Proof. The proof follows the outline of [Güt10, Remark 3.5]. Since $\mathbf{W}\mathbf{W}^\dagger$ is the orthogonal projector onto $\mathcal{V}_{k+1}^Z(\mathbf{L},\mathbf{b})$, there holds for all $\mathbf{v} \in \mathcal{V}_{k+1}^Z(\mathbf{L},\mathbf{b})$

$$\begin{aligned}(\mathbf{b} - (\mathbf{L}+\zeta\mathbf{I})\mathbf{W}(\hat{\mathbf{L}}_{k+1} + \zeta\mathbf{I}_{k+1})^{-1}\mathbf{W}^\dagger\mathbf{b}, \mathbf{v}) \\ = (\mathbf{W}\mathbf{W}^\dagger\mathbf{b} - \mathbf{W}\mathbf{W}^\dagger(\mathbf{L}+\zeta\mathbf{I})\mathbf{W}(\hat{\mathbf{L}}_{k+1} + \zeta\mathbf{I}_{k+1})^{-1}\mathbf{W}^\dagger\mathbf{b}, \mathbf{v}) \\ = (\mathbf{W}\mathbf{W}^\dagger\mathbf{b} - \mathbf{W}(\hat{\mathbf{L}}_{k+1} + \zeta\mathbf{I}_{k+1})(\hat{\mathbf{L}}_{k+1} + \zeta\mathbf{I}_{k+1})^{-1}\mathbf{W}^\dagger\mathbf{b}, \mathbf{v}) = 0.\end{aligned}$$

Since the reduced basis surrogate is uniquely defined by (11.6), we conclude that the conjecture is valid. □

After an initial computational investment, the reduced space (11.5) allows us to evaluate the coordinate vector of $\mathbf{w}_{k+1}(\zeta)$ in the basis $\{\mathbf{w}(\zeta_0),\dots,\mathbf{w}(\zeta_k)\}$ for arbitrary ζ with complexity only depending on k. Due to the fourth claim in Lemma 7.5, we immediately obtain the following result.

Lemma 11.5. *Let* $Z = \{\zeta_0,\dots,\zeta_k\} \subset \overline{\mathbb{R}}_0^+$ *be pairwise distinct and* $\Xi = -Z$. *Then there holds*

$$\mathcal{V}_{k+1}^Z(\mathbf{L},\mathbf{b}) = \mathcal{Q}_{k+1}^\Xi(\mathbf{L},\mathbf{b}).$$

Note that Lemma 11.4 and 11.5 imply that the reduced basis approximation of $\mathbf{w}(\zeta)$ with snapshots in Z is nothing else but the rational Krylov approximation of $\mathbf{w}(\zeta)$ with poles in $-Z$. In the following, we show that similar results apply to two particular classes of reduced basis methods which have been applied to fractional diffusion problems, namely ones based on interpolation and on quadrature.

Interpolation-based Reduced Basis Methods

Two different model order reduction strategies have been recently proposed in [DS19, DS21], which couple interpolation theory with reduced basis technology. Exploiting one of the integral representations deduced in Section 4.1.2, the (forward) fractional operator with positive exponent $s \in (0,1)$ is written as weighted integral over parametrized reaction-diffusion problems

$$\mathbf{L}^s\mathbf{u} = \frac{\sin(\pi s)}{\pi}\int_0^\infty \zeta^{s-1}(\mathbf{u} - \zeta\hat{\mathbf{w}}(\zeta))\,d\zeta, \qquad \hat{\mathbf{w}}(\zeta) := (\mathbf{L}+\zeta\mathbf{I})^{-1}\mathbf{u}. \tag{11.7}$$

Based on a selection of snapshots $(\zeta_j)_{j=0}^k \subset \overline{\mathbb{R}}_0^+$, the integrand is approximated using a RBM, yielding

$$\mathbf{b}_{k+1} := \frac{\sin(\pi s)}{\pi}\int_0^\infty \zeta^{s-1}(\mathbf{u} - \zeta\hat{\mathbf{w}}_{k+1}(t))\,d\zeta,$$

where $\hat{\mathbf{w}}_{k+1}$ is defined as in (11.6) upon replacing $\mathbf{b}$ with $\mathbf{u}$. As shown in [DS19, Theorem 4.3], the surrogate evaluates to

$$\mathbf{b}_{k+1} = \mathbf{W}\hat{\mathbf{L}}_{k+1}^{s}\mathbf{W}^{\dagger}\mathbf{u}, \tag{11.8}$$

where $\mathbf{W}$ is an orthonormal basis of $\mathcal{V}_{k+1}^{Z}(\mathbf{L},\mathbf{u})$ and $\hat{\mathbf{L}}_{k+1} = \mathbf{W}^{\dagger}\mathbf{L}\mathbf{W}$. In [DS19] it was proven that the scheme approximates $\mathbf{L}^{s}\mathbf{u}$ at exponential convergence rates. Motivated by these results, the authors of [DS21] proposed a version of (11.8) for the backward operator. They confirmed experimentally that

$$\mathbf{u}_{k+1}^{\mathrm{RB}} := \mathbf{W} f^{\tau}(\hat{\mathbf{L}}_{k+1})\mathbf{W}^{\dagger}\mathbf{b} \tag{11.9}$$

converges exponentially to $\mathbf{L}^{-s}\mathbf{b}$ if $f^{\tau}(\lambda) = \lambda^{-s}$ and $\mathbf{W}$ is a basis of $\mathcal{V}_{k+1}^{Z}(\mathbf{L},\mathbf{b})$, but no rigorous proof was known so far. The following theorem has been published in [DH21] and allows one to close this gap in the literature.

Theorem 11.6. *Let $\zeta \in \mathbb{R}_0^+$, $Z = \{\zeta_0,\dots,\zeta_k\} \subset \overline{\mathbb{R}}_0^+$ pairwise distinct. Then the reduced basis approximation (11.9) with snapshots in Z coincides with the rational Krylov approximation $\mathbf{u}_{k+1} \in \mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$ of $f^{\tau}(\mathbf{L})\mathbf{b}$ with poles in $\Xi = -Z$.*

Proof. According to Lemma 7.9, the rational Krylov approximation is independent of the choice of the basis. In view of (11.9), it thus suffices to verify that the corresponding search spaces $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$ and $\mathcal{V}_{k+1}^{Z}(\mathbf{L},\mathbf{b})$ coincide. This is true due to Lemma 11.5. □

Since the reduced basis approximation (11.9) is essentially a RKM, all analytical results presented in Chapter 10 also apply to $\mathbf{u}_{k+1}^{\mathrm{RB}}$.

The second method presented in [DS21] is referred to as *dual reduced basis approximation*. It follows a similar idea but is based on $\mathbf{L}^{-1}$. Since $\mathbf{L}^{-s} = \mathbf{L}^{-1}\mathbf{L}^{1-s}$, it follows from (11.7) that the inverse operator can be expressed as

$$\mathbf{L}^{-s}\mathbf{b} = \frac{\sin(\pi s)}{\pi}\int_0^{\infty} \zeta^{-s}\mathbf{L}^{-1}(\mathbf{b} - \zeta\mathbf{w}(\zeta))\,d\zeta.$$

The latter is again approximated utilizing reduced basis technology with prescribed snapshots $Z = \{\zeta_0,\dots,\zeta_k\} \subset \overline{\mathbb{R}}_0^+$ by means of

$$\mathbf{u}_{k+1}^{\mathrm{DUAL}} := \frac{\sin(\pi s)}{\pi}\int_0^{\infty} \zeta^{-s}\mathbf{L}^{-1}(\mathbf{b} - \zeta\mathbf{w}_{k+1}(\zeta))\,d\zeta. \tag{11.10}$$

It was shown in [DS21, Theorem 3.4] that (11.10) can be computed via

$$\mathbf{u}_{k+1}^{\mathrm{DUAL}} = \mathbf{L}^{-1}\mathbf{W}\hat{\mathbf{L}}_{*,k+1}^{s-1}\mathbf{W}^{\dagger}\mathbf{b}, \qquad \hat{\mathbf{L}}_{*,k+1} := \mathbf{W}^{\dagger}\mathbf{L}^{-1}\mathbf{W},$$

whenever $\mathbf{W}$ is an orthonormal basis of $\mathcal{V}_{k+1}^{Z}(\mathbf{L},\mathbf{b})$. If $\zeta_j = \infty$ for one $j \in \{0,\dots,k\}$, this surrogate can be interpreted as a postprocessed rational Krylov approximation as follows.

Theorem 11.7. *Let $s \in (0,1)$, $f^{\tau}(\lambda) = \lambda^{s-2}$, $\mathbf{u}_{k+1}^{\mathrm{DUAL}}$ the dual reduced basis approximation (11.10) with snapshots $Z = \{\zeta_0,\dots,\zeta_k\} \subset \overline{\mathbb{R}}_0^+$, $\Xi = -1/Z$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}^{-1},\mathbf{L}^{-1}\mathbf{b})$, $\mathbf{L}_{*,k+1} = \mathbf{V}^{\dagger}\mathbf{L}^{-1}\mathbf{V}$, and $\mathbf{u}_{k+1} = \mathbf{V}f^{\tau}(\mathbf{L}_{*,k+1})\mathbf{V}^{\dagger}\mathbf{L}^{-1}\mathbf{b}$. Assume $\zeta = \infty$ for one $\zeta \in Z$ such that $-\frac{1}{\zeta} = -\frac{1}{\infty} = 0 \in \Xi$. Then there holds*

$$\mathbf{u}_{k+1}^{\mathrm{DUAL}} = \mathbf{L}^{-1}\mathbf{u}_{k+1}. \tag{11.11}$$

Proof. W.l.o.g. we assume that $\zeta_0 = \infty$ is the only infinite snapshot. Then there holds

$$\begin{aligned}\mathcal{V}_{k+1}^{Z}(\mathbf{L},\mathbf{b}) &= \operatorname{span}\{\mathbf{b}, (\mathbf{L}+\zeta_1\mathbf{I})^{-1}\mathbf{b}, \dots, (\mathbf{L}+\zeta_k\mathbf{I})^{-1}\mathbf{b}\}\\ &= \operatorname{span}\{\mathbf{b}, (\mathbf{I}+\zeta_1\mathbf{L}^{-1})^{-1}\mathbf{L}^{-1}\mathbf{b}, \dots, (\mathbf{I}+\zeta_k\mathbf{L}^{-1})^{-1}\mathbf{L}^{-1}\mathbf{b}\}\\ &= \operatorname{span}\{\mathbf{b}, (\zeta_1^{-1}\mathbf{I}+\mathbf{L}^{-1})^{-1}\mathbf{L}^{-1}\mathbf{b}, \dots, (\zeta_k^{-1}\mathbf{I}+\mathbf{L}^{-1})^{-1}\mathbf{L}^{-1}\mathbf{b}\}\\ &= \mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}^{-1},\mathbf{L}^{-1}\mathbf{b}).\end{aligned}$$

Since $0 \in \Xi$ there holds $(\mathbf{L}^{-1})^{-1}\mathbf{L}^{-1}\mathbf{b} = \mathbf{b} \in \mathcal{Q}_{k+1}(\mathbf{L}^{-1},\mathbf{L}^{-1}\mathbf{b})$. Hence $\mathbf{b} = \mathbf{V}\mathbf{V}^\dagger\mathbf{b}$ so that

$$\mathbf{u}_{k+1} = \mathbf{V}\mathbf{L}_{*,k+1}^{s-2}\mathbf{V}^\dagger\mathbf{L}^{-1}\mathbf{V}\mathbf{V}^\dagger\mathbf{b} = \mathbf{V}\mathbf{L}_{*,k+1}^{s-2}\mathbf{L}_{*,k+1}\mathbf{V}^\dagger\mathbf{b} = \mathbf{V}\mathbf{L}_{*,k+1}^{s-1}\mathbf{V}^\dagger\mathbf{b}.$$

On the other hand, we have

$$\mathbf{L}\mathbf{u}_{k+1}^{\text{DUAL}} = \mathbf{W}\hat{\mathbf{L}}_{*,k+1}^{s-1}\mathbf{W}^\dagger\mathbf{b}, \qquad \hat{\mathbf{L}}_{*,k+1} = \mathbf{W}^T\mathbf{L}^{-1}\mathbf{W},$$

for any orthonormal basis $\mathbf{W}$ of $\mathcal{V}_{k+1}^{Z}(\mathbf{L},\mathbf{b})$. Due to Lemma 11.5, $\mathbf{V}$ and $\mathbf{W}$ span the same space. Hence, by Lemma 7.9, we deduce $\mathbf{u}_{k+1} = \mathbf{L}\mathbf{u}_{k+1}^{\text{DUAL}}$ so that (11.11) is valid. □

In [DS21, Theorem 4.3], it was shown that

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}^{\text{DUAL}}\| \preceq \rho_{[\lambda_{\min},\lambda_{\max}]}^{-k/2} \begin{cases} \|\mathbf{b}\|, & s > \frac{1}{2},\\ \|\mathbf{L}^{\frac{1}{2}}\mathbf{b}\|, & s \le \frac{1}{2},\end{cases}$$

if $Z = -\mathcal{Z}_\infty$, in which case the quality of the approximation depends unfavorably on the condition number whenever $s \le \frac{1}{2}$. Even worse, the dual reduced basis approximation has the disadvantage that it is not online efficient: Even if $\mathbf{W}$ and $\mathbf{L}_{*,k+1}$ are available, the query $s \mapsto \mathbf{u}_{k+1}^{\text{DUAL}}$ requires the performance of a matrix-vector product with $\mathbf{L}^{-1}$ and thus depends on the problem size N. The latter can be avoided if one directly extracts the surrogate from $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}^{-1},\mathbf{L}^{-1}\mathbf{b})$ using the poles $\Xi = -1/\mathcal{Z} \cup \{0\}$ and $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{s-1}$, or equivalently, $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}^{-1},\mathbf{b})$ with $\Xi = -1/\mathcal{Z} \cup \{\infty\}$ and $f(\lambda) = \lambda^s$. Motivated by these results, we define the *modified dual approximation* as

$$\mathbf{u}_{k+1}^{\text{DUAL2}} := \mathbf{V}\mathbf{L}_{*,k+1}^{s}\mathbf{V}^\dagger\mathbf{b}, \qquad \mathbf{L}_{*,k+1} = \mathbf{V}^\dagger\mathbf{L}^{-1}\mathbf{V}, \tag{11.12}$$

where $\mathbf{V}$ denotes a basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}^{-1},\mathbf{b})$. Even though (11.12) can be efficiently queried once the basis and its compression is available, the computation of $\mathbf{L}_{*,k+1} = \mathbf{V}^\dagger\mathbf{L}^{-1}\mathbf{V}$ requires $k+1$ additional linear solves which makes $\mathbf{u}_{k+1}^{\text{DUAL2}}$ more expensive compared to conventional RKMs.

Quadrature-based Reduced Basis Methods

Another class of reduced basis schemes for elliptic fractional diffusion problems uses the quadrature scheme presented in Section 6.3 as a starting point [BGZ20, DAC+21]. Provided some weights $(\omega_j)_{j=1}^m$ and nodes $(\eta_j)_{j=1}^m$ defining a quadrature, we have, recalling (6.9),

$$\mathbf{u}_m^{\text{QUAD}} = \sum_{j=1}^{m} \omega_j \eta_j^{-s}\mathbf{w}(\eta_j) \approx \frac{\sin(\pi s)}{\pi}\int_0^\infty \zeta^{-s}\mathbf{w}(\zeta)\,d\zeta = \mathbf{L}^{-s}\mathbf{b}. \tag{11.13}$$

In every quadrature node a parametric reaction-diffusion problem must be approximated, which turns out to be the method's bottleneck. The overall costs of computing (11.13) is essentially m queries of finite element solves for $\mathbf{w}(\zeta)$. In practice, m is in the range of $\mathcal{O}(100)$, see [BGZ20, DAC+21], resulting in a substantial computational effort if the costs of computing $\mathbf{w}(\zeta)$ are high. To mitigate this problem, the idea is to add an additional layer of approximation in the form of a RBM. Given a collection of snapshots $Z = \{\zeta_0, \dots, \zeta_k\}$, the *quadrature-based reduced basis approximation* is defined by

$$\mathbf{u}_{m,k+1}^{\mathrm{QUAD}} := \sum_{j=1}^{m} \omega_j \eta_j^{-s} \mathbf{w}_{k+1}(\eta_j). \tag{11.14}$$

Unlike $\mathbf{u}_{k+1}^{\mathrm{RB}}$, one might not be able to extract $\mathbf{u}_{m,k+1}^{\mathrm{QUAD}}$ from the reduced basis space via Rayleigh-Ritz extraction using the exact matrix function $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$. This can be compensated upon replacing $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ with the rational approximation stemming from the quadrature.

Theorem 11.8. *Let $\mathbf{u}_{m,k+1}^{\mathrm{QUAD}}$ be the quadrature-based reduced basis approximation defined by (11.14) with snapshots $Z = \{\zeta_0, \dots, \zeta_k\} \subset \overline{\mathbb{R}}_0^+$, $\mathbf{V}$ an orthonormal basis of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L}, \mathbf{b})$ with poles in $\Xi = -Z$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger \mathbf{L} \mathbf{V}$, and*

$$r_m^{\mathrm{QUAD}}(\lambda) := \sum_{j=1}^{m} \omega_j \frac{\eta_j^{-s}}{\lambda + \eta_j}.$$

Then there holds

$$\mathbf{u}_{m,k+1}^{\mathrm{QUAD}} = \mathbf{V} r_m^{\mathrm{QUAD}}(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}. \tag{11.15}$$

Proof. Recalling Theorem 11.6, we make use of the fact that any rational Krylov approximation is independent of the choice of the basis to write

$$\mathbf{w}_{k+1}(\eta_j) = \mathbf{V}(\mathbf{L}_{k+1} + \eta_j \mathbf{I}_{k+1})^{-1} \mathbf{V}^\dagger \mathbf{b}$$

for all $j = 1, \dots, m$. Therefore,

$$\mathbf{u}_{m,k+1}^{\mathrm{QUAD}} = \mathbf{V} \sum_{j=1}^{m} \omega_j \eta_j^{-s} (\mathbf{L}_{k+1} + \eta_j \mathbf{I}_{k+1})^{-1} \mathbf{V}^\dagger \mathbf{b} = \mathbf{V} r_m^{\mathrm{QUAD}}(\mathbf{L}_{k+1}) \mathbf{V}^\dagger \mathbf{b}. \qquad \square$$

From a theoretical point of view, it is desirable to replace $r_m^{\mathrm{QUAD}}(\lambda)$ in (11.15) with the exact matrix function $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$, which has the benefit that the choice of a quadrature rule as well as the tuning of its parameters can be avoided. By construction, the rational Krylov surrogate so obtained coincides with $\mathbf{u}_{k+1}^{\mathrm{RB}}$ if $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ in (11.9). In particular, the integral in (11.13) can be computed exactly after replacing $\mathbf{w}(\zeta)$ with $\mathbf{w}_{k+1}(\zeta)$ and thus does not require the implementation of quadrature rules. Unlike $\mathbf{u}_{m,k+1}^{\mathrm{QUAD}}$, however, the computation of $\mathbf{u}_{k+1}^{\mathrm{RB}}$ requires the inversion and diagonalization of the compressed matrix $\mathbf{L}_{k+1}$ and is thus slightly more expensive than directly evaluating the quadrature sum in (11.14). If k is small, the additional computational effort is negligible.

Thanks to Theorem 11.8, the analysis of quadrature-based reduced basis approximation directly follows from standard tools of RKMs. To see this, we interpret $\mathbf{u}_{k+1}^{\mathrm{RB}}$ and $\mathbf{u}_{m,k+1}^{\mathrm{QUAD}}$ as rational Krylov approximation of $\mathbf{L}^{-s}\mathbf{b}$ extracted from $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$ in the sense of Theorem 11.6 and 11.8. By the triangle inequality it follows

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{m,k+1}^{\mathrm{QUAD}}\| \leq \|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}^{\mathrm{RB}}\| + \|\mathbf{u}_{k+1}^{\mathrm{RB}} - \mathbf{u}_{m,k+1}^{\mathrm{QUAD}}\|. \tag{11.16}$$

The first expression on the right-hand side of (11.16) can be seen as the error caused by the model order reduction scheme while the latter corresponds to the contribution of the quadrature. Provided an orthonormal basis $\mathbf{V}$ of $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$, $\mathbf{L}_{k+1} = \mathbf{V}^\dagger\mathbf{L}\mathbf{V}$, and $e(\lambda) := \lambda^{-s} - r_m^{\mathrm{QUAD}}(\lambda)$, the second term can be written as

$$\begin{aligned}\|\mathbf{u}_{k+1}^{\mathrm{RB}} - \mathbf{u}_{m,k+1}^{\mathrm{QUAD}}\| &= \|\mathbf{V}\mathbf{L}_{k+1}^{-s}\mathbf{V}^\dagger\mathbf{b} - \mathbf{V}r_m^{\mathrm{QUAD}}(\mathbf{L}_{k+1})\mathbf{V}^\dagger\mathbf{b}\| \\ &= \|e(\mathbf{L}_{k+1})\mathbf{V}^\dagger\mathbf{b}\|_2 \leq \max_{j=0,\dots,k}|e(\mu_j^{(k)})|\|\mathbf{V}^\dagger\mathbf{b}\|_2,\end{aligned}$$

where $(\mu_j^{(k)})_{j=0}^k$ are the rational Ritz values of $\mathbf{L}$ on $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$. Since $(\mu_j^{(k)})_{j=0}^k \subset \Sigma$ and $\mathbf{P} = \mathbf{V}\mathbf{V}^\dagger$ is the orthogonal projector on $\mathcal{Q}_{k+1}^{\Xi}(\mathbf{L},\mathbf{b})$, it follows from $\|\mathbf{V}^\dagger\mathbf{b}\|_2 = \|\mathbf{P}\mathbf{b}\| \leq \|\mathbf{b}\|$

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{m,k+1}^{\mathrm{QUAD}}\| \leq \|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}^{\mathrm{RB}}\| + \|\lambda^{-s} - r_m^{\mathrm{QUAD}}(\lambda)\|_\Sigma\|\mathbf{b}\|. \tag{11.17}$$

The first term in (11.17) is guaranteed to converge exponentially under a suitable choice of the snapshots (or rather poles) advocated in Chapter 10. The second expression depends on the particular *scalar* quadrature. Assuming that the latter converges exponentially, we immediately derive exponential convergence rates for the quadrature-based reduced basis approximation $\mathbf{u}_{m,k+1}^{\mathrm{QUAD}}$.

One particular example that fits in the general framework presented above has been studied in [BGZ20]. The authors choose the sinc approximation (6.10) as a starting point and apply a RBM on top, resulting in the sinc quadrature-based reduced basis approximation

$$\mathbf{u}_{\mathrm{q},k+1}^{\mathrm{SINC}} := \frac{\mathrm{q}\sin(\pi s)}{\pi}\sum_{j=-n_-}^{n_+} e^{(1-s)\eta_j}\mathbf{w}_{k+1}(e^{\eta_j}),$$

where $\mathrm{q} \in \mathbb{R}_0^+$ is the sinc parameter, $\eta_j = j\mathrm{q}$ for all $j = -n_-,\dots,n_+$, and $n_-, n_+ \in \mathbb{N}$. The snapshots are chosen as $Z = -\hat{\mathcal{G}} \cup \{\infty\}$ sampled over the parameter domain $[e^{-\mathrm{q}n_-}, e^{\mathrm{q}n_+}]$. Its quality can be assessed by bounding the two terms in (11.17). The first contribution can be quantified via Theorem 10.24 in terms of

$$\|\mathbf{L}^{-s}\mathbf{b} - \mathbf{u}_{k+1}^{\mathrm{RB}}\| \leq \left(\frac{2e^{(1-s)\mathrm{q}n_-}}{\lambda_{\min}(1-s)} + \frac{2e^{-s\mathrm{q}n_+}}{s} + C\frac{e^{(1-s)\mathrm{q}n_+} - e^{-(1-s)\mathrm{q}n_-}}{1-s}\rho_{[\lambda_{\min},4\lambda_{\max}]}^{-\frac{k}{6}}\right)\|\mathbf{b}\|. \tag{11.18}$$

The sinc quadrature is known to converge exponentially in q [BLP19b, Theorem 3.2],

$$\|\lambda^{-s} - r_{\mathrm{q}}^{\mathrm{SINC}}(\lambda)\|_\Sigma \preceq \left(\frac{e^{-\frac{\pi^2}{2\mathrm{q}}}}{\sinh(\frac{\pi^2}{2\mathrm{q}})} + e^{(1-s)\mathrm{q}n_+} + e^{-s\mathrm{q}n_-}\right)\|\mathbf{b}\|, \tag{11.19}$$

where $r_{\mathrm{q}}^{\textsc{Sinc}}$ is the rational function associated to $\mathbf{u}_{m,k+1}^{\textsc{Sinc}}$ in the sense of Theorem 11.8. Imposing the values of n_+ and n_- according to (6.12) guarantees the exponentials from above to be balanced. The estimates (11.18) and (11.19) combined with (11.17) yield exponential convergence of the scheme which has already been observed in [BGZ20, Theorem 2 & Lemma 3.2].

In its present form, the snapshots are sampled over $[e^{-\mathrm{q}n_-}, e^{\mathrm{q}n_+}]$. Therefore, the search space depends unfavorably on the fractional power s if n_- and n_+ are chosen according to (6.12). As discovered in [BGZ20], this problem can be mitigated if one chooses

$$n_- = \left\lceil \frac{\pi^2}{2s_{\min}\mathrm{q}^2} \right\rceil, \qquad n_+ = \left\lceil \frac{\pi^2}{2(1-s_{\max})\mathrm{q}^2} \right\rceil, \tag{11.20}$$

for some $0 < s_{\min} < s_{\max} < 1$ fixed. Along with this choice, the exponential convergence results from above is recovered with the benefit that $s \mapsto \mathbf{u}_{\mathrm{q},k+1}^{\textsc{Sinc}}$ can be efficiently queried for multiple values of $s \in [s_{\min}, s_{\max}]$.

11.2 Numerical Results

This section is devoted to a numerical comparison of the algorithms discussed above, incorporating efficiency, similarities, and performance with respect to several values of the parameter s. For this purpose, we choose Ω, the finite element space V_h, the spectral bounds, λ_L, and λ_U, and $\mathcal{L}$ as in Section 10.4 to study the convergence properties of

$$\hat{\mathrm{E}}(k,s) := \|\mathbf{L}^{-s}\mathbf{b} - \hat{\mathbf{u}}_{k+1}\|, \tag{11.21}$$

where $\hat{\mathbf{u}}_{k+1} \in \{\mathbf{u}_{k+1}^{\textsc{Zolo}}, \mathbf{u}_{k+1}^{\textsc{Dual}}, \mathbf{u}_{k+1}^{\textsc{Dual2}}, \mathbf{u}_{k+1}^{\textsc{Greedy}}, \mathbf{u}_{\mathrm{q},k+1}^{\textsc{Sinc}}, \mathbf{u}_{k+1}^{\textsc{Direct}}, \mathbf{u}_{k+1}^{\textsc{Bura}}\}$ denotes either of the following surrogates.

1. We choose $\mathbf{u}_{k+1}^{\textsc{Zolo}}$ as reduced basis approximation defined by (11.9) with $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$. The snapshots Z are chosen as $Z = -\hat{\mathcal{Z}} \cup \{\infty\}$. Due to Theorem 11.6, this choice is equivalent to a conventional rational Krylov approximation of $\mathbf{L}^{-s}\mathbf{b}$ with poles in $\hat{\mathcal{Z}}_\infty$.

2. By $\mathbf{u}_{k+1}^{\textsc{Dual}}$ we label the dual reduced basis approximation (11.10) with snapshots $Z = -1/\hat{\mathcal{Z}} \cup \{\infty\}$ as in [DS21].

3. The surrogate $\mathbf{u}_{k+1}^{\textsc{Dual2}}$ denotes the modified dual approximation (11.12) whose snapshots we choose according to $Z = -\hat{\mathcal{Z}} \cup \{\infty\}$.

4. For the sinc quadrature approximation $\mathbf{u}_{\mathrm{q},k+1}^{\textsc{Sinc}}$, we set $s_{\min} := 0.2$, $s_{\max} := 0.8$, $\mathrm{q} := 0.15$, and choose n_- and n_+ according to (11.20). The snapshots Z are chosen as $Z = -\hat{\mathcal{G}} \cup \{\infty\}$ sampled over a discrete training set $\mathcal{T}_{\mathrm{train}}^{n_\pm} = \mathcal{T}_{\mathrm{train}}^{\mathrm{q}} \subset [e^{-\mathrm{q}n_-}, e^{\mathrm{q}n_+}]$ as in [BGZ20]. The set $\mathcal{T}_{\mathrm{train}}^{\mathrm{q}}$ is constructed using 10^4 equispaced points over $[-\mathrm{q}n_-, \mathrm{q}n_+]$ which are transformed to the desired parameter domain under the transformation $\lambda \to e^{\lambda}$.

5. We choose $\mathbf{u}_{k+1}^{\text{Greedy}}$ as reduced basis approximation (11.9) with $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$. The snapshots Z are chosen to be the same ones as for $\mathbf{u}_{\text{q},k+1}^{\text{Sinc}}$. Due to Theorem 11.6, $\mathbf{u}_{k+1}^{\text{Greedy}}$ coincides with the conventional rational Krylov approximation of $\mathbf{L}^{-s}\mathbf{b}$ with poles in $\hat{\mathcal{G}}_\infty$.

6. For the direct rational approximation method $\mathbf{u}_{r_k^{\boldsymbol{\tau}}} =: \mathbf{u}_{k+1}^{\text{Direct}}$, we choose $r_k^{\boldsymbol{\tau}} \in \mathcal{R}_{k,k}$ as the best uniform rational approximation of $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ on Σ obtained by the BRASIL algorithm [Hof21].

7. Finally, we choose $\mathbf{u}_{k+1}^{\text{Bura}}$ as rational Krylov approximation of $\mathbf{L}^{-s}\mathbf{b}$ with poles in $\mathcal{B}_{\boldsymbol{\tau}}^{\infty}$.

The errors (11.21) between the exact matrix-vector product and its low-dimensional surrogates obtained by the seven methods listed above are reported in Figure 11.1 for $s \in \{0.2, 0.8\}$ and $\mathbf{b}$ denoting the coefficient vector of the L^2-orthogonal projection of the constant 1-function onto the FEM space.

– The two methods which rely on the BURA, that is, $\mathbf{u}_{k+1}^{\text{Bura}}$ and $\mathbf{u}_{k+1}^{\text{Direct}}$, provide the best approximation among all tested methods irrespectively of the fractional order. The observed rate of convergence matches our analytical findings stated in Theorem 10.31 and 11.3, respectively.

– In view of Theorem 7.16 and 11.1, it is not surprising that $\mathbf{u}_{k+1}^{\text{Bura}}$ and $\mathbf{u}_{k+1}^{\text{Direct}}$ perform qualitatively similar. In accordance with the theory, the direct BURA method slightly outperforms $\mathbf{u}_{k+1}^{\text{Bura}}$ for the given vector $\mathbf{b}$. This is reasonable since $\mathbf{u}_{k+1}^{\text{Bura}}$ only yields a quasi-optimal rational approximation to $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ on Σ, while the BURA provides the best uniform rational approximation of $f^{\boldsymbol{\tau}}$ by definition.

– After a few iterations, the dual reduced basis approximation $\mathbf{u}_{k+1}^{\text{Dual}}$ converges with the predicted rate of $\mathcal{O}(\rho_{[\lambda_L,\lambda_U]}^{-k/2})$. The modified dual reduced basis approximation $\mathbf{u}_{k+1}^{\text{Dual2}}$ comes at sufficiently less computational costs and outperforms its competitor for reasonably large values of k.

– The approximations $\mathbf{u}_{\text{q},k+1}^{\text{Sinc}}$ and $\mathbf{u}_{k+1}^{\text{Greedy}}$ coincide for all values of k and s. The additional quadrature discretization has no impact on the quality of $\mathbf{u}_{\text{q},k+1}^{\text{Sinc}}$ at all. This is due to the small value of the sinc spacing $\text{q} = 0.15$ that causes the second term in (11.16) to fall below machine precision, such that $\mathbf{u}_{\text{q},k+1}^{\text{Sinc}} \approx \mathbf{u}_{k+1}^{\text{Greedy}}$. Indeed, we observe numerically that for any $\lambda \in \Sigma$ there holds $\lambda^{-s} \approx r_{\text{q}}^{\text{Sinc}}(\lambda)$ up to machine precision, where $r_{\text{q}}^{\text{Sinc}}$ denotes the rational approximation of $f^{\boldsymbol{\tau}}(\lambda) = \lambda^{-s}$ stemming from the quadrature in the sense of Theorem 11.8.

– In line with our experiments presented in Section 10.4, the methods based on the BURA provide the most accurate approximations across all scenarios and are specifically tailored towards a particular choice of the fractional parameter. If, however, solutions to (4.24) for several values of s are required, $\mathbf{u}_{k+1}^{\text{Zolo}}$, $\mathbf{u}_{k+1}^{\text{Dual2}}$, $\mathbf{u}_{\text{q},k+1}^{\text{Sinc}}$, and $\mathbf{u}_{k+1}^{\text{Greedy}}$ outperform their competitors in terms of efficiency since they allow direct querying of the solution for arbitrary s after an initial offline computation phase.

The latter, however, is roughly twice as costly for the modified dual reduced basis approximation compared to $\mathbf{u}_{k+1}^{\text{Zolo}}$, $\mathbf{u}_{\text{q},k+1}^{\text{Sinc}}$, and $\mathbf{u}_{k+1}^{\text{Greedy}}$.

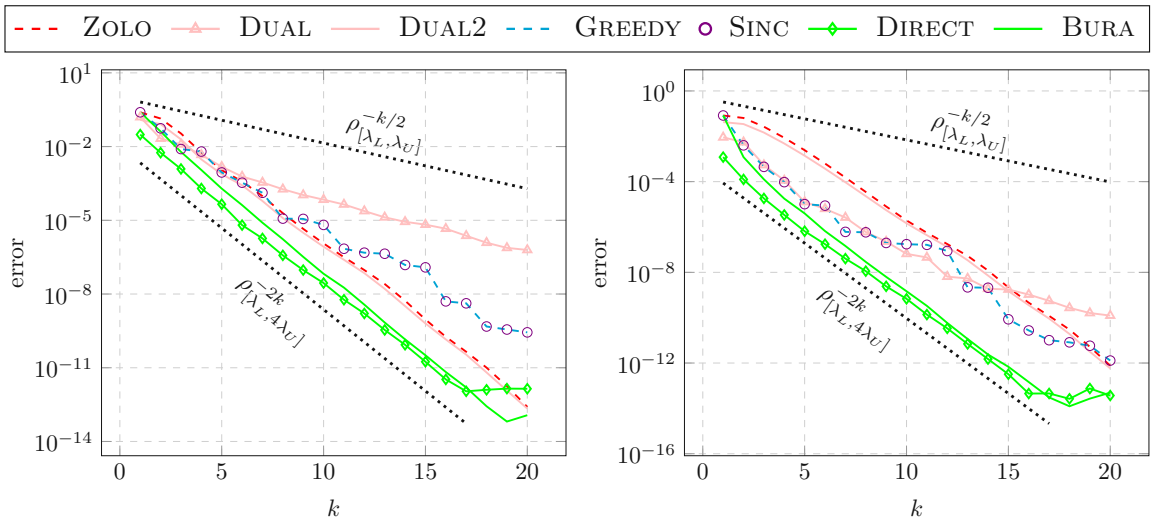

Figure 11.1: Error $\hat{\mathrm{E}}(k,s)$ for $s = 0.2$ (left) and $s = 0.8$ (right) when $\mathbf{b}$ is the coefficient vector of the L^2-orthogonal projection of the constant 1-function onto V_h.

The coefficient vector $\mathbf{b}$ of the L^2-orthogonal projection of the constant 1-function onto the FEM space is roughly uniformly excited by all eigenvectors of $\mathbf{L}$. In Figure 11.2, we consider the coefficient vector arising from $b(\mathbf{x}) = \sin(\pi x)\sin(\pi y) \in H_0^1(\Omega)$, $\mathbf{x} = (x,y) \in \Omega$, with fractional parameter $s = \frac{1}{2}$. We observe that all methods converge with the same rates as in the previous examples. Unlike in Figure 11.1, however, the direct BURA method appears to be the least accurate approximation among all tested methods. This is due to the smoothness of b, causing the excitations of $\mathbf{b}$ to decay quickly. Since the BURA is entirely independent of the vector $\mathbf{b}$, and thus not capable to adapt to its spectral properties, $\mathbf{u}_{k+1}^{\text{Direct}}$ requires substantially more linear solves to reach a prescribed accuracy compared to its rational Krylov competitors. The latter incorporate information about the vector in the construction of the search space and can thus bias the surrogate towards the particular choice of $\mathbf{b}$.

For the reader's convenience, we conclude this section with a systematic comparison of the discussed methods in Table 11.1, incorporating (from top to bottom)

1. their ability to efficiently query the solution map $s \mapsto \mathbf{L}^{-s}\mathbf{b}$ for multiple instances of s using the same search space,
2. nestedness of their poles as k increases, and hence their ability to incrementally improve the accuracy of the surrogate,
3. the required user-provided data.

For $\mathbf{u}_{k+1}^{\text{Zolo}}$, $\mathbf{u}_{k+1}^{\text{Dual}}$, and $\mathbf{u}_{k+1}^{\text{Dual2}}$, the respective search space is entirely independent of s and allows to query the solution map irrespectively of the fractional parameter. After

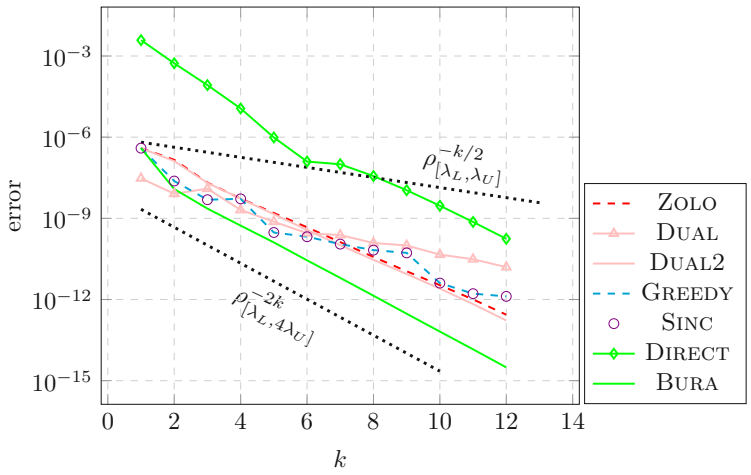

Figure 11.2: Error $\hat{\mathrm{E}}(k, 0.5)$ when $\mathbf{b}$ is the coefficient vector of the L^2-orthogonal projection of $b(\mathbf{x}) = \sin(\pi x)\sin(\pi y)$, $\mathbf{x} = (x, y) \in (0, 1)^2$, onto V_h.

restriction to a proper subset $[s_{\min}, s_{\max}]$ with $0 < s_{\min} < s_{\max} < 1$ fixed, the same also applies to $\mathbf{u}_{k+1}^{\text{Greedy}}$ and $\mathbf{u}_{\mathrm{q},k+1}^{\text{Sinc}}$. The remaining approximations presented in Table 11.1 rely on explicit rational approximations of $f^{\tau}(\lambda) = \lambda^{-s}$ and have to be computed for each $s \in [0, 1]$ individually. Neither the Zolotarëv points nor the BURA poles of order k are contained in the respective parameter set of order $k + 1$, which is computationally inconvenient when adaptive accuracy control is required. Conversely, in the computation of a basis spanning the search space of $\mathbf{u}_{k+1}^{\text{Greedy}}$ and $\mathbf{u}_{\mathrm{q},k+1}^{\text{Sinc}}$ only one new function is added at each stage to the k previously selected vectors which are left unchanged. As opposed to the other approximations, $\mathbf{u}_{k+1}^{\text{Greedy}}$ and $\mathbf{u}_{\mathrm{q},k+1}^{\text{Sinc}}$ do not require estimates on the spectral region of $\mathbf{L}$; however, they require a discrete training set $\mathcal{T}_{\text{train}}^{n_\pm} \subset [e^{-\mathrm{q}n_-}, e^{\mathrm{q}n_+}]$ whose range is encoded in the choice of the sinc parameter q.

	$\mathbf{u}_{k+1}^{\text{Zolo}}$	$\mathbf{u}_{k+1}^{\text{Dual}}$	$\mathbf{u}_{k+1}^{\text{Dual2}}$	$\mathbf{u}_{k+1}^{\text{Greedy}}$	$\mathbf{u}_{\mathrm{q},k+1}^{\text{Sinc}}$	$\mathbf{u}_{k+1}^{\text{Direct}}$	$\mathbf{u}_{k+1}^{\text{Bura}}$
multi-query	✓	×	✓	$[s_{\min}, s_{\max}]$	$[s_{\min}, s_{\max}]$	×	×
nested	×	×	×	✓	✓	×	×
user-provided	S	S	S	$\mathcal{T}_{\text{train}}^{\mathrm{q}}, \mathrm{q}$	$\mathcal{T}_{\text{train}}^{\mathrm{q}}, \mathrm{q}$	S	S

Table 11.1: Properties of the methods, where $\mathrm{S} = \{\lambda_L, \lambda_U\}$ contains bounds for the extremal eigenvalues of $\mathbf{L}$, $\mathcal{T}_{\text{train}}^{\mathrm{q}}$ a training set of $[e^{-n_-}, e^{n_+}]$, and q the sinc parameter.

The dominant computational effort for all methods is the solution of k parametric reaction-diffusion equations. When using an optimal solver such as a multigrid method, this requires $\mathcal{O}(kN)$ operations. Those methods which require diagonalization of a compressed matrix have to expend a further $\mathcal{O}(k^3)$ operations to do so, where typically $k \ll N$.

Bibliography

[AB17] H. Antil and S. Bartels. Spectral approximation of fractional PDEs in image processing and phase field modeling. *Computational Methods in Applied Mathematics*, 17(4):661–678, 2017.

[ABDN19] L. Aceto, D. Bertaccini, F. Durastante, and P. Novati. Rational Krylov methods for functions of matrices with applications to fractional partial differential equations. *Journal of Computational Physics*, 396:470–482, 2019.

[Ach92] N. I. Achieser. *Theory of Approximation*. Dover books on advanced mathematics. Dover Publications, 1992.

[ACN19] H. Antil, Y. Chen, and A. Narayan. Reduced basis methods for fractional Laplace equations via extension. *SIAM Journal on Scientific Computing*, 41(6):A3552–A3575, 2019.

[ACR21] H. Antil, A. N. Ceretani, and C. N. Rautenberg. The spatially variant fractional Laplacian. *arXiv preprint*, 2021. 2106.11471.

[ADS21] H. Antil, P. Dondl, and L. Striet. Approximation of integral fractional Laplacian and fractional PDEs via sinc-basis. *SIAM Journal on Scientific Computing*, 43(4):A2897–A2922, 2021.

[AF03] R. A. Adams and J. F. Fournier. In *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*, pages 301–305. Elsevier, 2003.

[AFM21] N. Angleitner, M. Faustmann, and J. M. Melenk. Approximating inverse FEM matrices on non-uniform meshes with $\mathcal{H}$-matrices. *Calcolo*, 58, 2021.

[AG18] M. Ainsworth and C. Glusa. Hybrid finite element-spectral method for the fractional Laplacian: Approximation theory and efficient solver. *SIAM Journal on Scientific Computing*, 40(4):A2383–A2405, 2018.

[Aga53] R. P. Agarwal. A propos d'une note de M. Pierre Humbert. 218, 1953.

[Akh90] N. I. Akhiezer. *Elements of the theory of elliptic functions*. American Mathematical Society, 1990.

[AM17] M. Ainsworth and Z. Mao. Analysis and approximation of a fractional Cahn-Hilliard equation. *SIAM Journal on Numerical Analysis*, 55(4):1689–1718, 2017.

[AN17] L. Aceto and P. Novati. Rational approximation to the fractional Laplacian operator in reaction-diffusion problems. *SIAM Journal on Scientific Computing*, 39(1):A214–A228, 2017.

[AN18] L. Aceto and P. Novati. Efficient implementation of rational approximations to fractional differential operators. *Journal of Scientific Computing*, 76:651–671, 2018.

[AN19] L. Aceto and P. Novati. Rational approximations to fractional powers of self-adjoint positive operators. *Numer. Math.*, 143(1):1–16, 2019.

[AN20] L. Aceto and P. Novati. Fast and accurate approximations to fractional powers of operators. *arXiv preprint*, 2020. 2004.09793.

[AN21] L. Aceto and P. Novati. Exponentially convergent trapezoidal rules to approximate fractional powers of operators. *arXiv preprint*, 2021. 2107.05860.

[Ana18] G. Anastassiou. *Intelligent Computations: Abstract Fractional Calculus, Inequalities, Approximations*. 2018.

[APR17] H. Antil, J. Pfefferer, and S. Rogovs. Fractional operators with inhomogeneous boundary conditions: Analysis, control, and discretization. *Communications in Mathematical Sciences*, 16, 2017.

[AR19] H. Antil and C. N. Rautenberg. Sobolev spaces with non-Muckenhoupt weights, fractional elliptic operators, and applications. *SIAM Journal on Mathematical Analysis*, 51(3):2479–2503, 2019.

[Aro55] N. Aronszajn. Boundary values of functions with finite Dirichlet integral, 1955.

[AS64] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. National Bureau of Standards Applied Mathematics Series, 1964.

[Bag69] T. Bagby. On interpolation by rational functions. *Duke Mathematical Journal*, 36:95–104, 1969.

[Bal60] A. V. Balakrishnan. Fractional powers of closed operators and the semigroups generated by them. *Pacific J. Math.*, 10(2):419–437, 1960.

[Bat06] P. W. Bates. On some nonlocal evolution equations arising in materials science. *Nonlinear dynamics and evolution equations*, pages 13–52, 2006.

[Baz01] E. Bazhlekova. *Fractional Evolution Equations in Banach Spaces*. PhD thesis, Eidhoven University of Technology, Netherlends, 2001.

[BBC03] K. Bogdan, K. Burdzy, and Z.-Q. Chen. Censored stable processes. *Probability Theory and Related Fields*, 127:89–152, 2003.

[BBNS18] A. Bonito, J. P. Borthagaray, R. H. Nochetto, and E. Otárola A. J. Salgado. Numerical methods for fractional diffusion. *Computing and Visualization in Science*, 2018.

[BCdH08] A. Brú, D. Casero, S. de Franciscis, and M. A. Herrero. Fractal analysis and tumour growth. *Mathematical and Computer Modelling*, 47(5):546–559, 2008.

[BCdPS13] C. Brändle, E. Colorado, A. de Pablo, and U. Sánchez. A concave-convex elliptic problem involving the fractional Laplacian. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 143(01):39–71, 2013.

[BDST12] D. Baleanu, K. Diethelm, E. Scalas, and J. Trujillo. *Fractional Calculus: Models and Numerical Methods*, volume 3. World Scientific Publishing Company, 2 edition, 2012.

[Bec11] B. Beckermann. An error analysis for rational Galerkin projection applied to the Sylvester equation. *SIAM Journal on Numerical Analysis*, 49(5/6):2430–2450, 2011.

[BEOR14] A. Buhr, C. Engwer, M. Ohlberger, and S. Rave. A numerically stable a posteriori error estimator for reduced basis approximations of elliptic equations. In X. Oliver E. Onate and A. Huerta, editors, *Proceedings of the 11th World Congress on Computational Mechanics*, pages 4094–4102. CIMNE, Barcelona, 2014.

[Ber29] S. Bernstein. Sur les fonctions absolument monotones. *Acta Math.*, 52:1–66, 1929.

[Ber08] C. Berg. *Positive Definite Functions: From Schoenberg to Space-Time Challenges*, chapter Stieltjes-Pick-Bernstein-Schoenberg and their connection to complete monotonicity, pages 15–45. 2008.

[BG12] 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.

[BG17] M. Berljafa and S. Güttel. Parallelization of the rational Arnoldi algorithm. *SIAM Journal on Scientific Computing*, 39(5):S197–S221, 2017.

[BGZ20] A. Bonito, D. Guignard, and A. R. Zhang. Reduced basis approximations of the solutions to spectral fractional diffusion problems. *Journal of Numerical Mathematics*, 28(3):147–160, 2020.

[BIG03] A. Ben-Israel and T. Greville. *Generalized inverses: Theory and applications*, volume 15. Springer, New York, 2003.

[BL76] J. Bergh and J. Lofstrom. *Interpolation spaces*, volume 223. Springer-Verlag, Berlin, 1 edition, 1976.

[BLP17a] A. Bonito, W. Lei, and J. E. Pasciak. The approximation of parabolic equations involving fractional powers of elliptic operators. *Journal of Computational and Applied Mathematics*, 315:32–48, 2017.

[BLP17b] A. Bonito, W. Lei, and J. E. Pasciak. Numerical approximation of space-time fractional parabolic equations. *Computational Methods in Applied Mathematics*, 17(4):679–705, 2017.

[BLP19a] A. Bonito, W. Lei, and J. E. Pasciak. Numerical approximation of the integral fractional Laplacian. *Numerische Mathematik*, 142, 2019.

[BLP19b] A. Bonito, W. Lei, and J. E. Pasciak. On sinc quadrature approximations of fractional powers of regularly accretive operators. *Journal of Numerical Mathematics*, 27(2):57–68, 2019.

[BM89] C. Bernardi and Y. Maday. Properties of some weighted Sobolev spaces and application to spectral approximations. *SIAM Journal on Numerical Analysis*, 26(4):769–829, 1989.

[BMN+18] L. Banjai, J. M. Melenk, R. H. Nochetto, E. Otárola, A. J. Salgado, and C. Schwab. Tensor FEM for spectral fractional diffusion. *Foundations of Computational Mathematics*, 2018.

[BMS20] L. Banjai, J. M. Melenk, and C. Schwab. Exponential convergence of hp FEM for spectral fractional diffusion in polygons. Technical report, SNF, Zurich, 2020.

[Boc33] S. Bochner. Integration von Funktionen, deren Werte die Elemente eines Vektorraumes sind. *Fundamenta Mathematicae*, 20(1):262–176, 1933.

[BOKG+14] A. Bueno-Orovio, D. Kay, V. Grau, B. Rodriguez, and K. Burrage. Fractional diffusion models of cardiac electrical propagation: Role of structural heterogeneity in dispersion of repolarization. *Journal of The Royal Society Interface*, 11(97):20140352, 2014.

[BP15] A. Bonito and J. E. Pasciak. Numerical approximation of fractional powers of elliptic operators. *Mathematics of Computation*, 84(295):2083–2110, 2015.

[BP16] A. Bonito and J. E. Pasciak. Numerical approximation of fractional powers of regularly accretive operators. *IMA Journal of Numerical Analysis*, 37(3):1245–1273, 2016.

[BR09] 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.

[Bra93] J. H. Bramble. *Multigrid Methods*. Chapman & Hall, New York, 1993.

[BRS16] G. Molica Bisci, V. D. Radulescu, and R. Servadei. *Variational Methods for Nonlocal Fractional Problems*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2016.

[BS87] M. S. Birman and M. Z. Solomjak. *Spectral theory of self-adjoint operators in Hilbert space*, volume 5. Springer Science & Business Media, 1 edition, 1987.

[BS88] C. Bennett and R. C. Sharpley. *Interpolation of operators*, volume 129. Academic press, 1988.

[BSV15] M. Bonforte, Y. Sire, and J. L. Vázquez. Existence, uniqueness and asymptotic behaviour for fractional porous medium equations on bounded domains. *Discrete & Continuous Dynamical Systems*, 35:5725, 2015.

[BT00] B. Le Bailly and J. P. Thiran. Optimal rational functions for the generalized Zolotarëv problem in the complex plane. *SIAM Journal on Numerical Analysis*, 38(5):1409–1424, 2000.

[BT17] B. Beckermann and A. Townsend. On the singular values of matrices with displacement structure. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1227–1248, 2017.

[BV16] C. Bucur and E. Valdinoci. *Nonlocal Diffusion and Applications*, chapter An Introduction to the Fractional Laplacian, pages 7–37. Springer International Publishing, Cham, 2016.

[BVY62] G. Birkhoff, R. S. Varga, and D. Young. Alternating direction implicit methods. volume 3 of *Advances in Computers*, pages 189–273. Elsevier, 1962.

[Cap67] M. Caputo. Linear models of dissipation whose Q is almost frequency independent—II. *Geophysical Journal International*, 13(5):529–539, 1967.

[Cap69] M. Caputo. *Elasticità e Dissipazione*. 1969.

[Cas12] F. Casenave. Accurate a posteriori error evaluation in the reduced basis method. *Comptes Rendus Mathematique*, 350(9):539–542, 2012.

[CD15] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numerica*, 24:1–159, 2015.

[CDDN20] A. Cohen, W. Dahmen, R. Devore, and J. Nichols. Reduced basis greedy selection using random training sets. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54(5):1509–1524, Sep 2020.

[CDDS11] A. Capella, J. Dávila, L. Dupaigne, and Y. Sire. Regularity of radial extremal solutions for some non-local semilinear equations. *Communications in Partial Differential Equations*, 36(8):1353–1384, 2011.

[CdTGG20] N. Cusimano, F. del Teso, and L. Gerardo-Giorda. Numerical approximations for fractional elliptic equations via the method of semigroups. *ESAIM Mathematical Modelling and Numerical Analysis*, 54:751–774, 2020.

[CEL14] F. Casenave, A. Ern, and T. Lelièvre. Accurate and online-efficient evaluation of the a posteriori error bound in the reduced basis method. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 48(1):207–229, 2014.

[CGGG21] N. Cusimano, L. Gerardo-Giorda, and A. Gizzi. A space-fractional bidomain framework for cardiac electrophysiology: 1D alternans dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(7):073123, 2021.

[Cha68] K. Chandrasekharan. *Introduction to Analytic Number Theory*. Springer Verlag, Berlin, 1968.

[Cia02] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Society for Industrial and Applied Mathematics, 2002.

[CKS10] Z.-Q. Chen, P. Kim, and R. Song. Two-sided heat kernel estimates for censored stable-like processes. *Probability theory and related fields*, 146(3-4):361, 2010.

[CM11] A. Chikrii and I. Matychyn. *Advances in Dynamic Games*, chapter Riemann–Liouville, Caputo, and Sequential Fractional Derivatives in Differential Games, pages 61–81. 2011.

[CS07] L. Caffarelli and L. Silvestre. An extension problem related to the fractional Laplacian. *Communications in Partial Differential Equations*, 32(8):1245–1260, 2007.

[CS16] L. A. Caffarelli and P. R. Stinga. Fractional elliptic equations, Caccioppoli estimates and regularity. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 33(3):767–807, 2016.

[CT10] X. Cabré and J. Tan. Positive solutions of nonlinear problems involving the square root of the Laplacian. *Advances in Mathematics*, 224(5):2052–2093, 2010.

[CWHM15] S. N. Chandler-Wilde, D. P. Hewett, and A. Moiola. Interpolation of Hilbert and Sobolev spaces: Quantitative estimates and counterexamples. *Mathematika*, 61(2):414–443, 2015.

[DAC+21] H. Dinh, H. Antil, Y. Chen, E. Cherkaev, and A. Narayan. Model reduction for fractional elliptic problems using Kato's formula. *Mathematical Control & Related Fields*, 2021.

[DH21] T. Danczul and C. Hofreither. On rational Krylov and reduced basis methods for fractional diffusion. *arXiv preprint*, 2021. 2102.13540.

[DHS21] T. Danczul, C. Hofreither, and J. Schöberl. A unified rational Krylov method for elliptic and parabolic fractional diffusion problems. *arXiv preprint*, 2021. 2103.13068.

[Die10] K. Diethelm. *The Analysis of Fractional Differential Equations.* Springer, Berlin, Heidelberg, 2010.

[DK94] V. Druskin and L. Knizhnerman. Spectral approach to solving three-dimensional Maxwell's equations in the time and frequency domains. *Radio Science*, 29:937–953, 1994.

[DK98] 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.

[DKZ09] V. Druskin, L. Knizhnerman, and M. Zaslavsky. Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts. *SIAM Journal on Scientific Computing*, 31(5):3760–3780, 2009.

[DL21] M. Daoud and E. H. Laamri. Fractional Laplacians : A short survey. *Discrete & Continuous Dynamical Systems - S*, 2021.

[DLZ10] V. Druskin, C. Lieberman, and M. Zaslavsky. On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. *SIAM Journal on Scientific Computing*, 32(5):2485–2496, 2010.

[DPW13] R. DeVore, G. Petrova, and P. Wojtaszczyk. Greedy algorithms for reduced bases in Banach spaces. *Constructive Approximation*, 37(3):455–466, 2013.

[DS88] N. Dunford and J. Schwartz. *Linear operators, part 1: general theory*, volume 10. John Wiley & Sons, 1988.

[DS11] V. Druskin and V. Simoncini. Adaptive rational Krylov subspaces for large-scale dynamical systems. *Systems & Control Letters*, 60(8):546–560, 2011.

[DS19] T. Danczul and J. Schöberl. A reduced basis method for fractional diffusion operators I. *arXiv preprint*, 2019. 1904.05599.

[DS21] T. Danczul and J. Schöberl. A reduced basis method for fractional diffusion operators II. *Journal of Numerical Mathematics*, 2021.

[DWZ17] S. Duo, H. Wang, and Y. Zhang. A comparative study on nonlocal diffusion operators related to the fractional Laplacian. *Discrete and Continuous Dynamical Systems - Series B*, 24, 2017.

[DZ21] B. Duan and Z. Zhang. A rational approximation scheme for computing Mittag-Leffler function with discrete elliptic operator as input. *Journal of Scientific Computing*, 87, 2021.

[EH06] J. Eshof and M. Hochbruck. Preconditioning Lanczos approximations to the matrix exponential. *SIAM J. Scientific Computing*, 27:1438–1457, 2006.

[EMOT54] A. Erdelyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi. *Tables of Integral Transforms: Vol.: 2.* McGraw-Hill Book Company, Incorporated, 1954.

[Eva10] L. C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 2010.

[EW91] N. S. Ellner and E. L. Wachspress. Alternating direction implicit iteration for systems with complex spectra. *SIAM Journal on Numerical Analysis*, 28(3):859–870, 1991.

[FKM20] M. Faustmann, M. Karkulik, and J. M. Melenk. Local convergence of the FEM for the integral fractional Laplacian. *arXiv preprint*, 2020. 2005.14109.

[FKR+21] M. Fritz, C. Kuttler, M. L. Rajendran, B. Wohlmuth, and L. Scarabosio. On a subdiffusive tumour growth model with fractional time derivative. *IMA Journal of Applied Mathematics*, 2021.

[FMP13] M. Faustmann, J. M. Melenk, and D. Praetorius. $\mathcal{H}$-matrix approximability of the inverses of FEM matrices. *Numerische Mathematik*, 131, 2013.

[FMP21] M. Faustmann, J. M. Melenk, and D. Praetorius. Quasi-optimal convergence rate for an adaptive method for the integral fractional Laplacian. *Mathematics of Computations*, 90(330):1557–1587, 2021.

[FRW21] M. Fritz, M. L. Rajendran, and B. Wohlmuth. Time-fractional Cahn-Hilliard equation: Well-posedness, degeneracy, and numerical solutions. *arXiv preprint*, 2021. 2104.03096.

[FS21] X. Feng and M. Sutton. A new theory of fractional differential calculus. *Analysis and Applications*, 19(04):715–750, 2021.

[FV20] J. A. Ezquerro Fernández and M. Á. Hernández Verón. *Mild Differentiability Conditions for Newton's Method in Banach Spaces*, chapter The Newton-Kantorovich Theorem, pages 1–22. Birkhäuser, Cham, 2020.

[FZ20] G. Failla and M. Zingales. Advanced materials modelling via fractional calculus: Challenges and perspectives. *Philosophical Transactions of the Royal Society A*, 378, 2020.

[Gag58] E. Gagliardo. Proprieta di alcune classi di funzioni in piu variabili. *Ricerche di Matematica*, 7(1):102–137, 1958.

[Gar15] R. Garrappa. Numerical evaluation of two and three parameter Mittag-Leffler functions. *SIAM J. Numer. Anal.*, 53:1350–1369, 2015.

[GH15] P. Gatto and J. S. Hesthaven. Numerical approximation of the fractional Laplacian via *hp*-finite elements, with an application to image denoising. *J. Sci. Comput.*, 65(1):249–270, 2015.

[GK13] S. Güttel and L. Knizhnerman. A black-box rational Arnoldi variant for Cauchy-Stieltjes matrix functions. *BIT Numerical Mathematics*, 53(3):595–616, 2013.

[GLY15] R. Gorenflo, Y. Luchko, and M. Yamamoto. Time-fractional diffusion equation in the fractional Sobolev spaces. *Fractional Calculus and Applied Analysis*, 18(3):799–820, 2015.

[GM05] Q.-Y. Guan and Z.-M. Ma. Boundary problems for fractional Laplacians. *Stochastics and Dynamics*, 5, 2005.

[GM06] Q.-Y. Guan and Z.-M. Ma. Reflected symmetric α-stable processes and regional fractional Laplacian. *Probability Theory and Related Fields*, 134:649–694, 2006.

[GO09] G. Gilboa and S. Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2009.

[Gon67] A. A. Gončar. Estimates of the growth of rational functions and some of their applications. *Mathematics of the USSR-Sbornik*, 1(3):445–456, 1967.

[Gon69] A. A. Gončar. Zolotarëv problems connected with rational functions. *Mathematics of the USSR-Sbornik*, 78 (120):640–654, 1969.

[Gon78] A. A Gončar. On the speed of rational approximation of some analytic functions. *Sbornik: Mathematics*, 34(2):131–145, 1978.

[GPTV15] S. Güttel, E. Polizzi, P. Tang, and G. Viaud. Zolotarëv quadrature rules and load balancing for the FEAST eigensolver. *SIAM Journal on Scientific Computing*, 37(4):A2100–A2122, 2015.

[GRN+17] P. Ginzburg, D. J. Roth, M. E. Nasir, P. Segovia, A. V. Krasavin, J. Levitt, L. M. Hirvonen, B. Wells, K. Suhling, D. Richards, et al. Spontaneous emission in non-local materials. *Light: Science & Applications*, 6(6):e16273–e16273, 2017.

[Gru16] G. Grubb. Regularity of spectral fractional Dirichlet and Neumann problems. *Mathematische Nachrichten*, 289(7):831–844, 2016.

[GS21] S. Güttel and M. Schweitzer. A comparison of limited-memory Krylov methods for Stieltjes functions of Hermitian matrices. *SIAM Journal on Matrix Analysis and Applications*, 42(1):83–107, 2021.

[Gua06] Q.-Y. Guan. Integration by parts formula for regional fractional Laplacian. *Communications in Mathematical Physics*, 266:289–329, 2006.

[Güt10] S. Güttel. *Rational Krylov Methods for Operator Functions*. PhD thesis, Technische Universität Bergakademie Freiberg, Germany, 2010. Dissertation available as MIMS Eprint 2017.39.

[Güt13] S. Güttel. Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection. *GAMM-Mitteilungen*, 36(1):8–31, 2013.

[Hac99] W. Hackbusch. A sparse matrix arithmetic based on H-matrices. part I: Introduction to H-matrices. *Computing*, 62(2):89–108, 1999.

[Hen93] P. Henrici. *Applied and Computational Complex Analysis, Volume 3 : Discrete Fourier Analysis, Cauchy Integrals, Construction of Conformal Maps, Univalent Functions*. John Wiley & Sons, Inc., USA, 1993.

[Hig08] N. J. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, 2008.

[Hil53] T. H. Hildebrandt. Integration in abstract spaces. *Bulletin of the American Mathematical Society*, 59:111–139, 1953.

[HJ91] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[HKL+21a] S. Harizanov, N. Kosturski, I. Lirkov, S. Margenov, and Y. Vutov. Reduced multiplicative (BURA-MR) and additive (BURA-AR) best uniform rational approximation methods and algorithms for fractional elliptic equations. *Fractal and Fractional*, 5(3), 2021.

[HKL+21b] S. Harizanov, N. Kosturski, I. Lirkov, S. Margenov, and Y. Vutov. Reduced sum implementation of the BURA method for spectral fractional diffusion problems. *arXiv preprint*, 2021. 2105.09048.

[HL97] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 34(5):1911–1925, 1997.

[HLM+18] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov, and Y. Vutov. Optimal solvers for linear systems with fractional powers of sparse SPD matrices. *Numerical Linear Algebra with Applications*, 25(5):e2167, 2018.

[HLM+20] S. Harizanov, R. Lazarov, S. Margenov, P. Marinov, and J. E. Pasciak. Analysis of numerical methods for spectral fractional elliptic equations based on the best uniform rational approximation. *Journal of Computational Physics*, 408:109285, 2020.

[HMP21] S. Harizanov, S. Margenov, and N. Popivanov. Spectral fractional Laplacian with inhomogeneous Dirichlet data: Questions, problems, solutions. In I. Georgiev, H. Kostadinov, and E. Lilkova, editors, *Advanced Computing in Industrial Mathematics*, pages 123–138, Cham, 2021. Springer International Publishing.

[Hof20] C. Hofreither. A unified view of some numerical methods for fractional diffusion. *Computers & Mathematics with Applications*, 80(2):332–350, 2020.

[Hof21] C. Hofreither. An algorithm for best rational approximation based on barycentric rational interpolation. *Numerical Algorithms*, 2021.

[HRS15] J. S. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. Springer, Switzerland, 1 edition, 2015.

[HSMR20] S. Hijazi, G. Stabile, A. Mola, and G. Rozza. Data-driven POD-Galerkin reduced order model for turbulent flows. *Journal of Computational Physics*, 416:109513, 2020.

[Hu98] C. Hu. Algorithm 785: A software package for computing Schwarz-Christoffel conformal transformation for doubly connected polygonal regions. 24(3):317–333, 1998.

[ILTA05] M. Ilic, F. Liu, I. Turner, and V. Anh. Numerical approximation of a fractional-in-space diffusion equation, I. *Fractional Calculus and Applied Analysis*, 8(3):323–341, 2005.

[ILTA06] M. Ilic, F. Liu, I. Turner, and V. Anh. Numerical approximation of a fractional-in-space diffusion equation (II) – with nonhomogeneous boundary conditions. *Fractional Calculus and Applied Analysis*, 9(4):333–349, 2006.

[IT95] M.-P. Istace and J.-P. Thiran. On the third and fourth Zolotarëv problems in the complex plane. *SIAM Journal on Numerical Analysis*, 32(1):249–259, 1995.

[JLZ15] B. Jin, R. Lazarov, and Z. Zhou. An analysis of the L1 scheme for the subdiffusion equation with nonsmooth data. *IMA Journal of Numerical Analysis*, 36(1):197–221, 2015.

[Kar18] M. Karkulik. Variational formulation of time-fractional parabolic equations. *Computers & Mathematics with Applications*, 75(11):3929–3938, 2018.

[Kat60] T. Kato. Note on fractional powers of linear operators. *Proceedings of the Japan Academy*, 36(3):94 – 96, 1960.

[KNBR21] E. N. Karatzas, M. Nonino, F. Ballarin, and G. Rozza. A reduced order cut finite element method for geometrically parameterized steady and unsteady Navier-Stokes problems. *arXiv preprint*, 2021. 2010.04953.

[Kry31] A. N. Krylov. On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined. *Izvestija AN SSSR (News of Academy of Sciences of the USSR), Otdel. mat. i estest. nauk*, 7(4):491–539, 1931.

[KS10] 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.

[KW21] U. Khristenko and B. Wohlmuth. Solving time-fractional differential equation via rational approximation. *arXiv preprint*, 2021. 2102.05139.

[Kwa17] M. Kwaśnicki. Ten equivalent definitions of the fractional Laplace operator. *Fractional Calculus and Applied Analysis*, 20(1):7–51, 2017.

[Lan72] N. S. Landkof. *Foundations of modern potential theory*. Springer-Verlag Berlin, New York, 1972.

[LB92] J. Lund and K. L. Bowers. *Sinc Methods for Quadrature and Differential Equations*. Society for Industrial and Applied Mathematics, 1992.

[Leb77] V. I. Lebedev. On a Zolotarëv problem in the method of alternating directions. *USSR Computational Mathematics and Mathematical Physics*, 17(2):58–76, 1977.

[Lei18] W. Lei. *Numerical Approximation of Partial Differential Equations Involving Fractional Differential Operators*. PhD thesis, Texas A & M University, USA, 2018.

[LM72] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications*. Springer-Verlag, New York, 1972.

[LM98] R. Lehoucq and K. Meerbergen. Using generalized Cayley transformations within an inexact rational Krylov sequence method. *SIAM Journal on Matrix Analysis and Applications*, 20(1):131–148, 1998.

[LP12] K. Li and J. Peng. Fractional resolvents and fractional evolution equations. *Applied Mathematics Letters*, 25(5):808–812, 2012.

[LPG+20] A. Lischke, G. Pang, M. Gulian, F. Song, C. Glusa, X. Zheng, Z. Mao, W. Cai, M. M. Meerschaert, M. Ainsworth, and G. E. Karniadakis. What is the fractional Laplacian? A comparative review with new results. *Journal of Computational Physics*, 404:109009, 2020.

[LQR12] T. Lassila, A. Quarteroni, and G. Rozza. A reduced basis model with parametric coupling for fluid-structure interaction problems. *SIAM Journal on Scientific Computing*, 34(2):A1187–A1213, 2012.

[LR10] T. Lassila and G. Rozza. Parametric free-form shape design with PDE models and reduced basis method. *Computer Methods in Applied Mechanics and Engineering*, 199(23):1583–1592, 2010.

[LS94] A. L. Levin and E. B. Saff. Optimal ray sequences of rational functions connected with the Zolotarëv problem. *Constructive Approximation*, 10, 1994.

[LS01] A. L. Levin and E. B. Saff. The distribution of zeros and poles of asymptotically extremal rational functions for Zolotarëv's problem. *Journal of Approximation Theory*, 110(1):88–108, 2001.

[LS06] E. Levin and E. Saff. Potential theoretic tools in polynomial and rational approximation. In *Harmonic Analysis and Rational Approximation*, volume 327, pages 71–94. Springer-Verlag, Berlin, 2006.

[LS21] P. Loreti and D. Sforza. Weak solutions for time-fractional evolution equations in Hilbert spaces. *Fractal and Fractional*, 5(4), 2021.

[Lub88] C. Lubich. Convolution quadrature and discretized operational calculus. I. *Numerische Mathematik*, 53(1), 1988.

[Lun09] A. Lunardi. *Interpolation Theory*. 2 edition, 2009.

[Mai10] F. Mainardi. *Fractional Calculus and Waves in Linear Viscoelasticity*. Imperial College Press, 2010.

[Mai20] F. Mainardi. Why the Mittag-Leffler function can be considered the queen function of the fractional calculus? *Entropy*, 22(12), 2020.

[McL00] W. C. McLean. *Strongly Elliptic Systems and Boundary Integral Equations*, volume 86. Cambridge University Press, 2000.

[Mer12] M. Merkle. Completely monotone functions: A digest. *Analytic Number Theory, Approximation Theory, and Special Functions: In Honor of Hari M. Srivastava*, 2012.

[Mik78] J. Mikusiński. The Bochner integral. In *The Bochner Integral*, volume 55, pages 15–22. Springer, 1978.

[Mil99] K. S. Miller. A note on the complete monotonicity of the generalized Mittag-Leffler function. *Real Anal. Exchange*, 23(2):753–756, 1999.

[ML03] G. Mittag-Leffler. Sur la nouvelle fonction $e_\alpha(x)$. 137, 1903.

[MN04] I. Moret and P. Novati. RD-rational approximations of the matrix exponential. *BIT Numerical Mathematics*, 44, 2004.

[MN11] I. Moret and P. Novati. On the convergence of Krylov subspace methods for matrix Mittag-Leffler functions. *SIAM Journal on Numerical Analysis*, 49(5):2144–2164, 2011.

[MN18] I. Moret and P. Novati. Krylov subspace methods for functions of fractional differential operators. *Mathematics of Computation*, 88(315):293–312, 2018.

[MPT02] Y. Maday, A. T. Patera, and G. Turinici. A priori convergence theory for reduced-basis approximations of single-parametric elliptic partial differential equations. *Journal of Scientific Computing*, 17:437–446, 2002.

[MR93] K. S. Miller and B. Ross. *An Introduction to the Fractional Calculus and Fractional Differential Equations*. Wiley, New York, 1993.

[MR20a] S. Massei and L. Robol. Rational Krylov for Stieltjes matrix functions: Convergence and pole selection. *BIT Numerical Mathematics*, 61(1):237–273, 2020.

[MR20b] J. M. Melenk and A. Rieder. hp-FEM for the fractional heat equation. *IMA Journal of Numerical Analysis*, 41(1):412–454, 2020.

[NBR21] M. Nonino, F. Ballarin, and G. Rozza. A monolithic and a partitioned, reduced basis method for fluid–structure interaction problems. *Fluids*, 6(6), 2021.

[NF16] Y. Nakatsukasa and R. W. Freund. Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of Zolotarëv's functions. *SIAM Review*, 58(3):461–493, 2016.

[NOS15] R. H. Nochetto, E. Otárola, and A. J. Salgado. A PDE approach to fractional diffusion in general domains: A priori error analysis. *Foundations of Computational Mathematics*, 15(3):733–791, 2015.

[NOS16] R. H. Nochetto, E. Otárola, and A. J. Salgado. A PDE approach to space-time fractional parabolic problems. *SIAM Journal on Numerical Analysis*, 54(2):848–873, 2016.

[NPV12] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *Bulletin des Sciences Mathematiques*, 136(5):521–573, 2012.

[NRMQ13] F. Negri, G. Rozza, A. Manzoni, and A. Quarteroni. Reduced basis method for parametrized elliptic optimal control problems. *SIAM Journal on Scientific Computing*, 35(5):A2316–A2340, 2013.

[OLBC10] F. Olver, D. Lozier, R. Boisvert, and C. Clark. *The NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY, 2010.

[OS74] K. B. Oldham and J. Spanier. The fractional calculus. volume 111 of *Mathematics in Science and Engineering*, pages 225–234. Elsevier, 1974.

[OS18] E. Otárola and A. J. Salgado. Regularity of solutions to space–time fractional wave equations: A PDE approach. *Fractional Calculus and Applied Analysis*, 21(5):1262–1293, 2018.

[Ose07] I. V. Oseledets. Lower bounds for separable approximations of the Hilbert kernel. *Sbornik: Mathematics*, 198(3):425–432, 2007.

[Par88] O. G. Parfenov. Estimates of the singular numbers of the Carleson imbedding operator. *Mathematics of the USSR-Sbornik*, 59(2):497, 1988.

[Pee63] J. Peetre. On the theory of interpolation spaces. 1963.

[PK09] I. Podlubny and M. Kacenak. Mittag-leffler function. *The MATLAB routine, http://www.mathworks.com/matlabcentral/fileexchange*, 2009.

[Pod99] I. Podlubny. *Fractional differential equations: An introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*. Mathematics in Science and Engineering. Academic Press, London, 1999.

[PP88] P. P. Petrushev and V. A. Popov. *Rational Approximation of Real Functions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1988.

[Pro94] V. A. Prokhorov. Rational Approximation of Analytic Functions. *Sbornik: Mathematics*, 78(1):139–164, 1994.

[Pro05] V. A. Prokhorov. On best rational approximation of analytic functions. *Journal of Approximation Theory*, 133(2):284–296, 2005.

[QMN15] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced Basis Methods for Partial Differential Equations*. Springer International Publishing, 2015.

[QR03] A. Quarteroni and G. Rozza. Optimal control and shape optimization of aorto-coronaric bypass anastomoses. *Mathematical Models and Methods in Applied Sciences*, 13(12):1801–1823, 2003.

[QR07] A. Quarteroni and G. Rozza. Numerical solution of parametrized Navier–Stokes equations by reduced basis methods. *Numerical Methods for Partial Differential Equations*, 23(4):923–948, 2007.

[QR14] A. Quarteroni and G. Rozza. *Reduced Order Methods for Modeling and Computational Reduction*. Springer, Switzerland, 1 edition, 2014.

[QRM11] A. Quarteroni, G. Rozza, and A. Manzoni. Certified reduced basis approximation for parametrized partial differential equations and applications. *Journal of Mathematics in Industry*, 1, 2011.

[Rak16] E. A. Rakhmanov. The Gonchar-Stahl ρ^2-theorem and associated directions in the theory of rational approximations of analytic functions. *Sbornik: Mathematics*, 207(9):1236–1266, 2016.

[Ran95] T. Ransford. *Potential Theory in the Complex Plane*. London Mathematical Society Student Texts. Cambridge University Press, 1995.

[RHP08] G. Rozza, D. B. P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15, 2008.

[Rie20] A. Rieder. Double exponential quadrature for fractional diffusion. *arXiv preprint*, 2020. 2012.05588.

[Ros97] J. Rostand. Computing logarithmic capacity with linear programming. *Experimental Mathematics*, 6(3):221 – 238, 1997.

[RTW20] D. Rubin, A. Townsend, and H. Wilber. Bounding Zolotarëv numbers using Faber rational functions. *arXiv preprint*, 2020. 1911.11882.

[Rud74] W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. Tata McGraw-Hill, 1974.

[Ruh84] A. Ruhe. Rational Krylov sequence methods for eigenvalue computation. *Linear Algebra and its Applications*, 58:391–405, 1984.

[Saa81] Y. Saad. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of Computation*, 37(155):105–126, 1981.

[Saf10] E. Saff. Logarithmic potential theory with applications to approximation theory. *Surveys in Approximation Theory*, 5, 2010.

[SBMR18] M. Strazzullo, F. Ballarin, R. Mosetti, and G. Rozza. Model reduction for parametrized optimal control problems in environmental marine sciences and engineering. *SIAM Journal on Scientific Computing*, 40(4):B1055–B1079, 2018.

[Sch96] W. R. Schneider. Completely monotone generalized Mittag-Leffler functions. *Expo. Math.*, 14, 1996.

[Sch97] J. Schöberl. Netgen an advancing front 2D/3D-mesh generator based on abstract rules. *Computing and Visualization in Science*, 1:41–52, 1997.

[Sch14] J. Schöberl. C++11 implementation of finite elements in NGSolve. 2014.

[SEV14] R. Servadei and Enrico E. Valdinoci. On the spectrum of two different fractional operators. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 144(4):831–855, 2014.

[SGF20] M. Shaat, E. Ghavanloo, and S. A. Fazelzadeh. Review on nonlocal continuum mechanics: Physics, material applicability, and mathematics. *Mechanics of Materials*, 150:103587, 2020.

[SJC20] P. Shengfang, Z. Junjie, and Z. Chunyu. Efficient aerodynamic shape optimization through reduced order CFD modeling. *Optimization and Engineering*, 21, 2020.

[SKM93] S. G. Samko, A. A. Kilbas, and O. I. Marichev. *Fractional integrals and derivatives: Theory and applications*. Gordon and Breach, Yverdon, 1993.

[Slo58] L. N. Slobodeckij. Generalized Sobolev spaces and their applications to boundary value problems of partial differential equations. *Leningrad. Gos. Ped. Inst. Ucep. Zap.*, 197:54–112, 1958.

[SR18] G. Stabile and G. Rozza. Finite volume POD-Galerkin stabilised reduced order methods for the parametrised incompressible Navier–Stokes equations. *Computers & Fluids*, 173:273–284, 2018.

[SSV12] R. L. Schilling, R. Song, and Z. Vondracek. *Bernstein Functions: Theory and Applications*. De Gruyter, 2012.

[ST92] H. Stahl and V. Totik. *General Orthogonal Polynomials*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1992.

[ST97] E. B. Saff and V. Totik. *Logarithmic Potentials with External Fields*. Springer, Berlin, Heidelberg, 1997.

[ST10] P. Raúl Stinga and J. L. Torrea. Extension problem and Harnack's inequality for some fractional operators. *Communications in Partial Differential Equations*, 35(11):2092–2122, 2010.

[Sta91] G. Starke. Optimal alternating direction implicit parameters for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 28(5):1431–1445, 1991.

[Sta92] G. Starke. Near-circularity for the rational Zolotarëv problem in the complex plane. *Journal of Approximation Theory*, 70(1):115–130, 1992.

[Sta93] G. Starke. Fejér-Walsh points for rational functions and their use in the ADI iterative method. *Journal of Computational and Applied Mathematics*, 46(1):129–141, 1993.

[Ste12] F. Stenger. *Numerical methods based on Sinc and analytic functions*, volume 20. Springer-Verlag New York, 2012.

[Sti10] P. R. Stinga. *Fractional powers of second order partial differential operators: Extension problem and regularity theory*. PhD thesis, Universidad Autonoma de Madrid, Spain, 2010.

[SV16] J. Sprekels and E. Valdinoci. A new type of identification problems: Optimizing the fractional order in a nonlocal evolution equation. *SIAM J. Control and Optimization*, 55:70–93, 2016.

[SY11] K. Sakamoto and M. Yamamoto. Initial value/boundary value problems for fractional diffusion-wave equations and applications to some inverse problems. *Journal of Mathematical Analysis and Applications*, 382(1):426–447, 2011.

[SZB+18] H. Sun, Y. Zhang, D. Baleanu, W. Chen, and Y. Chen. A new collection of real world applications of fractional calculus in science and engineering. *Communications in Nonlinear Science and Numerical Simulation*, 64:213–231, 2018.

[Tar07] L. Tartar. *An introduction to Sobolev spaces and interpolation spaces*, volume 3. Springer-Verlag, Berlin, 2007.

[Tho06] Vidar Thomée. *Galerkin Finite Element Methods for Parabolic Problems (Springer Series in Computational Mathematics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[Tod84] J. Todd. *Applications of Transformation Theory: A Legacy from Zolotarëv (1847–1878)*. Springer Netherlands, Dordrecht, 1984.

[Tre19] L. N. Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM-Society for Industrial and Applied Mathematics, 2019.

[Tri78] H. Triebel. *Interpolation theory, function spaces, differential operators*. North-Holland Pub., 1978.

[Vab15] P. N. Vabishchevich. Numerically solving an equation for fractional powers of elliptic operators. *Journal of Computational Physics*, 282:289–302, 2015.

[Vab21a] P. N. Vabishchevich. An approximate representation of a solution to fractional elliptical BVP via solution of parabolic IVP. *Journal of Computational and Applied Mathematics*, page 113460, 2021.

[Vab21b] P. N. Vabishchevich. Some methods for solving equations with an operator function and applications for problems with a fractional power of an operator. *arXiv preprint*, 2021. 2105.10432.

[Vab21c] P. N. Vabishchevich. Splitting schemes for non-stationary problems with a rational approximation for fractional powers of the operator. *Applied Numerical Mathematics*, 165:414–430, 2021.

[Val09] E. Valdinoci. From the long jump random walk to the fractional Laplacian. *Bol. Soc. Esp. Mat. Apl. SeMA*, 49, 2009.

[Wac88] E. L. Wachspress. The ADI minimax problem for complex spectra. *Applied Mathematics Letters*, 1(3):311–314, 1988.

[Wac13] E. Wachspress. *The ADI Model Problem*. Springer New York, 2013.

[Wal60] J. L. Walsh. *Interpolation and approximation by rational functions in the complex domain*. American Mathematical Society, Providence, RI, 3 edition, 1960.

[Wal65] J. L. Walsh. Hyperbolic capacity and interpolating rational functions. *Duke Mathematical Journal*, 32(3):369 – 379, 1965.

[Wei60] G. Weiss. The theory of matrices. *Science*, 131(3408):1216–1216, 1960.

[WGP17] D. R. Witman, M. Gunzburger, and J. Peterson. Reduced-order modeling for nonlocal diffusion problems. *International Journal for Numerical Methods in Fluids*, 83(3):307–327, 2017.

[Wid43] D. V. Widder. The Laplace transform. *The Mathematical Gazette*, 27(273):37–39, 1943.

[WR66] J. L. Walsh and H. G. Russell. Hyperbolic capacity and interpolating rational functions II. *Duke Mathematical Journal*, 33(2):275 – 279, 1966.

[WT07] J. A. C. Weideman and L. N. Trefethen. Parabolic and hyperbolic contours for computing the Bromwich integral. *Mathematics of Computation*, 76(259):1341–1356, 2007.

[Yag10] A. Yagi. *Abstract Parabolic Evolution Equations and their Applications*. Springer-Verlag Berlin Heidelberg, 2010.

[YJN19] C. Yanlai, J. Jiang, and A. Narayan. A robust error estimator and a residual-free error indicator for reduced basis methods. *Computers & Mathematics with Applications*, 77(7):1963–1979, 2019.

[Yos95] K. Yosida. *Functional Analysis*. Springer, Berlin, Heidelberg, 1995.

[YPK16] Y. Yu, P. Perdikaris, and G. E. Karniadakis. Fractional modeling of viscoelasticity in 3D cerebral arteries and aneurysms. *J. Comput. Phys.*, 323(C):219–242, 2016.

[YTLI11] Q. Yang, I. Turner, F. Liu, and M. Ilic. Novel numerical methods for solving the time-space fractional diffusion equation in two dimensions. *SIAM J. Scientific Computing*, 33:1159–1180, 2011.

[ZBF+20] Z. Zainib, F. Ballarin, S. Fremes, P. Triverio, L. Jiménez-Juan, and G. Rozza. Reduced order methods for parametric optimal flow control in coronary bypass grafts, toward patient-specific data assimilation. *International Journal for Numerical Methods in Biomedical Engineering*, 2020.

[Zol77] E. I. Zolotarëv. Collected works. *St.-Petersburg Academy of Sciences*, 1877.

Curriculum Vitae

Persönliche Daten

Name	**Tobias Danczul**
Geburtsdatum	01.05.1994
Geburtsort	Wien
Nationalität	Österreich
Email	`tobias.danczul@tuwien.ac.at`

Ausbildung

seit 11/2018	Projektassistent am Institut für Analysis und Scientific Computing, TU Wien, Österreich
10/2016–06/2018	Masterstudium Technische Mathematik, TU Wien, Österreich
10/2013–10/2016	Bachelorstudium Technische Mathematik, TU Wien, Österreich
06/2012	Matura, GRG3 Kundmanngasse, Wien, Österreich

Wissenschaftliche Publikationen

T. Danczul and J. Schöberl. A reduced basis method for fractional diffusion operators I. arXiv preprint, 2019. 1904.05599.

T. Danczul and J. Schöberl. A reduced basis method for fractional diffusion operators II. Journal of Numerical Mathematics, 2021.

T. Danczul and C. Hofreither. On rational Krylov and reduced basis methods for fractional diffusion. arXiv preprint, 2021. 2102.13540.

T. Danczul, C. Hofreither, and J. Schöberl. A unified rational Krylov method for elliptic and parabolic fractional diffusion problems. arXiv preprint, 2021. 2103.13068.

Wien, am 25. Oktober, 2021

Tobias Danczul