DEPARTMENT OF MATHEMATICS

CHAIR OF MATHEMATICAL FINANCE (M13)

TECHNICAL UNIVERSITY OF MUNICH

PIDE methods and concepts for parametric option pricing

Maximilian Gaß

Vollständiger Abdruck der von der Fakultät für Mathematik der
Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Massimo Fornasier

Prüfer der Dissertation:
1. Prof. Dr. Kathrin Glau
2. Prof. Dr. Christoph Reisinger, University of Oxford, UK (schriftliche Beurteilung)
   Prof. Dr. Barbara Wohlmuth (mündliche Prüfung)
3. Prof. Dr. Josef Teichmann, ETH Zürich, Schweiz (schriftliche Beurteilung)

Die Dissertation wurde am 29.06.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 03.11.2016 angenommen.

to Eva

Abstract

This thesis is concerned with methods for option pricing that we investigate both theoretically and numerically. The first main part interprets option prices as solutions to partial integro differential equations (PIDEs). Focusing on exponential Lévy models, we implement a numerical tool for solving PIDEs using a Galerkin finite element approach that is flexible in the driving asset process. Many numerical examples provide evidence for the numerical feasibility of the method. Furthermore we establish a stability and convergence analysis for PIDEs with time-inhomogeneous operators of Gårding type. The second part of the thesis applies Chebyshev polynomial interpolation to option pricing by interpreting option prices as functions of option and model parameters. A numerical implementation of the pricing interpolation technique illustrates the method and emphasizes the gain in efficiency. The third part combines the empirical interpolation algorithm of Barrault et al. (2004) with Fourier based option pricing by interpolating associated Fourier integrands. Theoretical findings are numerically validated. Further numerical studies highlight the appealing features of the method, especially in higher dimensional parameter spaces. Additionally, the recursive nature of the interpolation operator is resolved which renders the method numerically accessible for the interpolation of multivariate Fourier integrands, as well.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit Methoden zur Optionspreisbewertung in theoretischer und numerischer Hinsicht. Der erste Teil der Arbeit betrachtet Optionspreise als Lösungen von partiellen Integro-Differentialgleichungen (PIDEs). Mit besonderer Berücksichtigung von exponentiellen Lévy-Modellen wird ein numerisches Tool zur Lösung solcher PIDEs implementiert, das sich durch eine große Flexibilität bezüglich des treibenden Lévy-Prozesses auszeichnet. Viele numerische Beispiele unterstreichen die numerische Umsetzbarkeit der Herangehensweise. Zudem wird eine Stabilitäts- und Konvergenzanalyse für PIDEs mit zeitinhomogenem Operator, der eine Gårding-Ungleichung erfüllt, hergeleitet. Der zweite Teil der Arbeit verwendet die Chebyshev'sche Interpolationsmethode zur Optionspreisbewertung. Optionspreise werden dazu als Funktionen von Options- und Modellparametern behandelt. Eine numerische Implementierung der Methode unterstreicht den resultierenden Effizienzgewinn. Der dritte Teil kombiniert schließlich die Empirische Interpolation von Barrault et al. (2004) mit Fourier-Techniken zur Optionspreisbestimmung durch die Interpolation der zugehörigen Fourier-Integranden. Theoretische Ergebnisse der Untersuchung werden numerisch validiert. Weitere numerische Studien heben die attraktiven Eigenschaften der Methode hervor, insbesondere im Hinblick auf Parameterräume höherer Dimension. Zudem wird der rekursive Aufbau des Interpolationsoperators aufgelöst und die Interpolation so auch der Anwendung auf multivariate Fourier-Integranden numerisch zugänglich gemacht.

Acknowledgements

First of all, I sincerely thank my supervisor Prof. Dr. Kathrin Glau. Her unconditional passion for research, her scientific curiosity and her strong commitment to our joint projects have driven this thesis forward. I am deeply thankful for her continuous support when questions or problems came up. Additionally, the friendly atmosphere that she created during commonly spent hours reflecting on mathematical problems has made this research trip through academia not only enriching and rewarding but also simply fun.

I furthermore thank Prof. Dr. Matthias Scherer. A small remark of his after the defense of my Bachelor's thesis had encouraged me to sincerely consider the idea of pursuing a Ph.D. in the first place. My sincere gratitude also goes to Prof. Dr. Rudi Zagst, the chairholder of M13. I deeply appreciate his support and goodwill that allowed me to adapt my university contract several times in order to enable the pursuit of other projects outside of university.

In this context, I thank the managing board of the KPMG Center of Excellence in Risk Management for financing the last year of my Ph.D. studies and providing the possibility of a three-month on-the-job insight into consulting. Special thanks in this regard go to Franz Lorenz for his support and supervision during my stay at the company.

The last 3.5 years have been a lot of work but nonetheless a lot of fun especially due to my great colleagues at the chair. I therefore thank all fellow Ph.D. students and coworkers at the chair of financial mathematics for uncountable table soccer matches, emotional "Weißwurst lunch duty" coin tosses, serious discussions about trivia and absurdities of life and for all the friendships that began or intensified here and that will last for sure. In this regard I especially thank Thorsten Schulz for $9-1$ shared years of friendship and studies of mathematics that started during TUM pre–courses when we apparently did not even know yet, what the natural numbers were. I am also greatly indebted to Maximilian Mair for many casual discussions in front of the whiteboard about inexplicable numerical effects that have repeatedly opened seeming dead ends.

Furthermore I thank my parents and brothers for their steady support during my studies, encouragement in times of trouble and distraction during home visits.

Most of all I thank Eva, who during these intense years has been the answer to all those questions that really matter in life.

Maximilian Gaß
March 29, 2016

Contents

1 Introduction

Option pricing is a key task in mathematical finance. The statement itself seems clear and unambiguous at first, yet it offers a variety of interpretations with equally manifold consequences to mathematical finance.

Speculation and risk appetite interpret options as means to benefit from market behavior. Anticipated developments of the economy like ups and downs of exchange rates or cyclically recurring events with economic impact like central bank chair meetings provide an opportunity for financial profit from occasions that might otherwise be insignificant to individual interest. In this interpretation, options suddenly give financial value to originally unrelated events and option pricing becomes a sophisticated gambling instrument.
A different interpretation emphasizes the contribution of options in enabling other trading activities. Market participants engaging in mutual trading activities cherish the ability of options to seal sources of risk that threaten their primary commercial transactions. Here, option pricing enables trade and supports a running economy.
Capturing the market in terms of model assumptions and an associated parametrization fosters a third interpretation. Equipped with option pricing tools, a parametrized market model not only yields prices of financial instruments but also allows a description of the current state of the real world economy that it portrays. Risks that prevail in the markets are thus mirrored by the parameter values of the simulating model. In this perspective, option pricing methods not only map parameter values to option prices but implicitly provide a link between observed option prices and the current state of the economy. Option pricing routines then drive the calibration of market models and carry out the first step for risk measurement and risk assessment purposes.

Each interpretation provokes its own reaction by financial mathematics. Speculation identifies market behavior that it intends to benefit from and stimulates the development of mathematical valuation methods for respective sophisticated financial instruments. Hedging purposes require the capacity to provide options that exhaustively capture all relevant sources of risk and obtain prices for them. Finally, risk management purposes demand reliable quantification of risk, a requirement which translates into option pricing methods that yield precise results and and maintain trustworthy numerical routines.

The actual interpretation thus matters indeed and guides research in different directions. In this thesis we follow the third interpretation. We adopt the view that a market and the structure of its movements can be described by model assumptions and associated parameters, a view that is emphasized by the expression *parametric option pricing* or POP in short. The literature on parametric option pricing has largely followed the seminal work of Carr and Madan (1999) and Raible (2000). It has thus almost exclusively been devoted to the development of algorithms based on fast Fourier transforms, see Lee (2004), Lord, Fang, Bervoets and Oosterlee (2008), Feng and Linetsky (2008), Kudryavtsev and Levendorskiĭ (2009), Boyarchenko and Levendorskiĭ (2014). Furthermore, we refer to Sachs and Schu (2010), Cont, Lantos and Pironneau (2011) and Haasdonk, Salomon and Wohlmuth (2012) that apply the so-called reduced basis method to parametric option pricing in finance. Prices of financial products are thus functions which link parameters describing both the current condition of the market and the characteristics of the product to the prices of the instrument. As sketched above, this link applies in both directions. On the basis of a parametrized model, the pricing method of choice yields option prices which match the observed market valuation whenever the model parametrization matches the current state of the market. In return, observed option prices in the market serve as reference points for calibrating the parametric model to market reality. A model aligned to observed market reality then facilitates risk assessment. Reliable risk quantification, however, requires reliable pricing tools.

Mathematical finance faces several challenges of theoretical and numerical nature in establishing that reliable link between market reality and its model equivalent. First, the theoretical frameworks need to comprise the capabilities for thorough error control. Proper risk assessment relies on theoretical error bounds and convergence results to justify its claims. The requirements to option pricing approaches thus go beyond the deployment of pure concepts but rather additionally expect estimates on the errors inevitably occurring when those concepts are applied practically. Second, the approaches that prevail in theory must maintain numerical feasibility. Risk measurement techniques operate on actual data retrieved from the market and are implemented numerically. Today's numerical limitations thus restrict the set of solution approaches to the option pricing problem even though it might be unlimited in theory.

Theoretical concepts and numerical implementations in mathematical finance have come under additional distress in recent years. With the crisis of 2007–2009 hitting the global economy, neglected sources of risk in the markets had become visible. As a consequence, models have grown considerably in complexity in order to better reflect the observed market reality. Considering a few examples we mention stochastic volatility and Lévy models as well as models based on further classes of stochastic processes. See for instance Heston (1993), Eberlein, Keller and Prause (1998), Duffie, Filipović and Schachermayer (2003), Cuchiero, Keller-Ressel and Teichmann (2015) for asset models and see Eberlein and Özkan (2005), Keller-Ressel, Papapantoleon and Teichmann (2013), Filipović, Larsson and Trolle (2014) for fixed income models. Given these developments, the model of Black and Scholes (1973) and Merton (1973) that had originally initiated mathematical finance today comes across like an anecdotal special case in that expanded model universe.

Increases in model complexity naturally resonate in the respective numerical implementations. While the Black&Scholes model allowed for (semi-)explicit formulas for European plain vanilla options, a whole new generation of pricing tools has been developed to numerically process the advancements on the theoretical side. These pricing tools fall into three distinct main families. A first family contains Monte-Carlo techniques. Here, market movements are simulated path-wise and option prices are derived by taking averages over the simulated option payoffs for each path. The idea of this approach is very appealing given the wide applicability of the method concerning both models and options. At the same time, the method suffers from comparably low accuracy and slow runtimes. A second family consists in the collection of Fourier techniques. Option pricing based on the Fourier transform has been intensively studied and applied in recent years. The approach that had been pioneered by Stein and Stein (1991) and Heston (1993) for Brownian models unveiled a great flexibility in terms of capturing a large class of models and option types. Fourier pricing of European options in Lévy and the large class of affine jump models has first been developed by Carr and Madan (1999), Raible (2000) and Duffie et al. (2000). There is a large and further growing literature on Fourier methods to price path dependent options and we refer to Boyarchenko and Levendorskiĭ (2002b), Feng and Linetsky (2008), Kudryavtsev and Levendorskiĭ (2009), Zhylyevskyy (2010), Fang and Oosterlee (2011), Levendorskiĭ and Xie (2012), Feng and Lin (2013) and Zeng and Kwok (2014) in this regard. Additionally consider Eberlein, Glau and Papapantoleon (2010) for a general framework and analysis. For plain vanilla options, Fourier integration combines the advantages of theoretically and numerically proven efficiency with implementational ease. Yet the restriction to plain vanilla options excludes many products of American type that are in general more liquidly traded in the market and would thus be the preferred choice for example for the purpose of model calibration. Finally, a third family comprehends the partial integro differential equations (PIDE) approach. Here, option prices are interpreted as solutions to partial differential equations additionally containing an integral term, see Hilber et al. (2013), Hilber et al. (2009), Dang et al. (2016), Eberlein and Glau (2014) and others for an overview over PIDE theory as such. Numerical solutions to PIDEs based on finite difference schemes are proposed for example in Cont and Voltchkova (2005), Fakharany et al. (2016), Coclite et al. (2016), Chen and Wang (2015) and Company et al. (2013). For solution schemes relying on the finite element method we refer to Matache et al. (2004), Matache et al. (2005b), Matache et al. (2005a) and Winter (2009). Lin and Yang (2012) and Florescu et al. (2014) describe numerical solutions to PIDEs based on other schemes. While the PIDE method provides a great flexibility in terms of both models and options, the implementation of numerical PIDE solvers is rather sophisticated, indeed.

In summary we observe, that each of the three methods conveys a certain appeal which in return comes at a certain cost. Fast runtimes are paid by limited flexibility while an extensive scope of applicability corresponds to numerical expenses. In this thesis we try to resolve that seeming contradiction. We aim at

- exploiting the flexibility that option pricing techniques offer
- ensuring numerical feasibility of pricing methods especially in terms of runtimes
- developing error control measures wherever possible

We will pursue these goals in a two-step approach. In a first step, we focus on the flexibility that a special class of partial differential equations offers and study its potential for option pricing in detail. That class is the family of PIDEs, where the differential operator is allowed to contain an additional integral term that accounts for the modeling of jumps in the trajectories of market asset. Jumps are the characteristic feature of Lévy model theory which can indeed be cast in PIDE terms and which will provide examples that make the abstract model framework concrete. As we have indicated earlier, however, the flexibility that PIDE theory offers to option pricing carries a burden in numerical terms in turn. Numerical runtimes of PIDE solvers often fall short of the high expectations and practical needs of the industry. Therefore, in a second step, we focus on the expectation of fast numerical runtimes and the desire for efficient numerical schemes expressed by the industry. A first approach to improving numerical runtimes easily connects to arbitrary pricing methods thus including PIDE solvers, as well. A second approach that we investigate for fast and efficient option pricing will be taylored to Fourier pricing in particular. In both steps we balance thorough theoretical investigations with extensive numerical case studies. Neither theory nor implementation shall seem neglected throughout this thesis.

Before we are able to present our main results, Chapter 2 briefly surveys basic elements of the theories that this thesis relies on. Furthermore, it presents a variety of asset models that will serve as examples throughout the numerical studies done in this manuscript.

In Chapter 3 we consider prices u as solutions to partial integro differential equations

$$\begin{aligned} \partial_t u + \mathcal{A}u &= f, \\ u(0) &= g, \end{aligned}$$

with a model specific operator $\mathcal{A}$ and an initial condition g that depends on the payoff profile of the option. We address the issue of finding solutions to PIDEs both theoretically and numerically. Introducing the Galerkin method serves both ends. Interpreted as a theoretical concept it provides a solution framework that is compatible with the functional analysis behind PIDE theory. Interpreted as an algorithmic guideline it describes a numerical implementation for a PIDE solver. In the chapter we illustrate this duality. After a theoretical description of the method we take the Merton model as an example and implement a pricing tool based on the finite elements method (FEM). In a third step, we exploit Fourier techniques to resolve the model dependence of that FEM solver rendering it accessible to a variety of asset models simultaneously. Many numerical studies enrich the topics of the chapter. It closes with a major proof on stability and convergence for approximate solutions of time dependent PIDEs. The contents of this chapter appear in Gaß and Glau (2016) and parts of the implementation support the studies in Burkovska et al. (2016).

In the subsequent Chapter 4, we shift our focus to improving numerical runtimes of option pricing methods in general. To this end we introduce the Chebyshev polynomial interpolation method for option pricing, a technique using the well understood Chebyshev polynomials, see Platte and Trefethen (2008) and Trefethen (2013). The method interprets an option price as a function of model and option parameters. It demands option prices at prespecified nodes in the parameter space $\mathcal{P}$ and interpolates prices for arbitrary parameters $p \in \mathcal{P}$ inbetween,

$$Price^p \approx I_{\overline{N}}(Price^{(\cdot)})(p) = \sum_{j_1=0}^{N_1} \dots \sum_{j_D=0}^{N_D} c_{(j_1,\dots,j_D)} T_{(j_1,\dots,j_D)}(p), \qquad p \in \mathcal{P},$$

wherein $c_{(j_1,\dots,j_D)}$ are parameter independent, precomputed coefficients and $T_{(j_1,\dots,j_D)}$ are model independent Chebyshev polynomials. The Chebyshev method thus builds on arbitrary option pricing tools but reduces their application to providing prices at the prespecified nodes that the interpolation is built on. Pricing then consists in assembling a weighted sum with known coefficients and polynomials that are easy to evaluate thus improving pricing runtimes tremendously. Under certain smoothness conditions on the underlying price we state an exponential convergence result for the algorithm. The contents of the chapter are also presented in Gaß et al. (2016).

Chapter 5 pursues a similar objective. Tayloring the capacity of the empirical magic point interpolation method by Barrault et al. (2004) and the results of Maday et al. (2009) to Fourier pricing, we achieve a significant gain in efficiency and numerical runtimes in option pricing. The resulting magic point integration method interpolates Fourier integrands by achieving their separation into parts that depend on the parameter $p \in \mathcal{P}$ and parts that depend on the integration variable alone,

$$Price^p \approx \mathcal{I}_M(h)(p) := \sum_{m=1}^{M} h_p(z_m^*) \int_\Omega \theta_m^M(z)\,\mathrm{d}z, \qquad p \in \mathcal{P}.$$

The sum in the interpolator $\mathcal{I}_M$ thus consists of parameter independent integrals that are computed beforehand and parameter dependent coefficients that are cheap to evaluate. Pricing has again turned into the evaluation of a sum. By exploiting the structure of the model specific Fourier integrands, the algorithm detects those local nodes in the parameter space $\mathcal{P}$ that explain the structure of all parametrized Fourier integrands at a given precision, globally. Enjoying this flexibility renders the algorithm less affected by the curse of dimensionality that other methods suffer from. We state theoretical conditions for exponential convergence of the algorithm. Numerous case studies and pricing examples validate and illustrate our theoretical claims empirically. In the context of pricing, the method is presented in Gaß et al. (2015), as well. The general applicability for parametric integration is furthermore demonstrated in Gaß and Glau (2015).

In the appendix we gather supplementary material for the main chapters sketched above. An integration technique for oscillating integrands that we encounter in Chapter 3 is presented in Appendix A. Properties of the empirical interpolation method being the key ingredient for the pricing algorithm of Chapter 5 are stated in Appendix B. Finally, a proof of Gronwall's lemma in a version crucial to our convergence result at the end of Chapter 3 is provided in Appendix C.

Research aims at pushing boundaries of knowledge further into the unknown. Yet any research must acknowledge its own limitations. The discipline it is located in, the topics within this discipline that it devotes itself to and the process in itself eventually determine that special spot that individual research occupies. As naturally as that spot emerges and as inevitable as the process leading to it seems, research should be prepared to answer the question of which purpose it serves. Research questions arise from various observations and occasions and hence the answers to that question might be as diverse as individual research is.

In this thesis we investigate aspects of parametric option pricing. We pursue thorough theoretical investigations, propose numerical implementations that meet practical needs and embed our results into thorough error control regimes. In this regard the diffuse noise from a collapsing global economy in 2007 that echoes until today was the question we encountered and we offer our results as parts of an answer.

We briefly summarize the main contributions of this thesis.

Chapter 3 First, we introduce a method for solving pricing PIDEs using a finite element approach that is highly flexible in the model choice and numerically feasible. We implement the method using mollified hat functions and splines as basis functions and empirically confirm theoretically prescribed convergence rates. In the second part of the chapter we generalize stability and convergence results for approximate solutions to PIDEs of von Petersdorff and Schwab (2003) to time-dependent bilinear forms of Gårding type.

These contributions are separately presented in Gaß and Glau (2016) and support the studies in Burkovska, Gaß, Glau, Mahlstedt, Mair, Schoutens and Wohlmuth (2016). Parts of this chapter also appear in Gaß and Glau (2014).

Chapter 4 We apply the Chebyshev interpolation method of Trefethen (2013) to option pricing. Interpreting the characteristic function of a Lévy model as a function of the model parameters, we derive areas in the parameter space that these functions are analytic on thus providing examples that fulfill theoretical requirements for exponential convergence of the method. We perform thorough numerical studies that validate the theoretical claims of exponential convergence and emphasize the gain in efficiency.

These contributions are separately presented in Gaß, Glau, Mahlstedt and Mair (2016).

Chapter 5 We apply the empirical interpolation method of Barrault et al. (2004) to option pricing. For a variety of Lévy models we derive conditions on the parameter space that guarantee the existence of a strip of analyticity of the associated characteristic function. We present a variety of numerical studies that validate theoretical claims of exponential convergence of the method and emphasize its suitability for the approximation of option prices in several free parameters in the one-asset case. In the second part of the chapter we resolve the recursive nature of the interpolation operator and thus provide the possibility to apply the method numerically feasibly for pricing options on several assets, as well.

These contributions are separately presented in Gaß, Glau and Mair (2015) and Gaß and Glau (2015).

2 Preliminaries

In this chapter we gather some elementary concepts and results that the main parts of this thesis rely on. The following sections of this chapter are by no means exhaustive regarding the topics they present. Yet, they aim at providing a theoretic overview containing the most important cornerstones necessary for a full understandings of the main concepts that the following chapters investigate.

2.1 Fourier theory

The first section in this preliminary chapter is devoted to Fourier theory. The Fourier transform has been extensively studied, see Bracewell (1999) for an introduction. Today, the transform lies at the heart of many applications in statistics and beyond. Appendix 1 of Kammler (2007) provides an idea of the rich scope of Fourier analysis.

The following definitions set up the Fourier transform framework that we shall use in this thesis. Since there are different various of Fourier transforms we emphasize that all Fourier related content of this work traces back to the concept of the Fourier transform as outlined by the following Definition 2.1.

Definition 2.1 (Fourier transform)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be an integrable real valued function, $f \in L^1(\mathbb{R}^d)$. We define denote by $\widehat{f}$ or $\mathcal{F}(f)$ the Fourier transform *of f, defined by*

$$\widehat{f}(\xi) = \mathcal{F}(f)(\xi) = \int_{\mathbb{R}^d} e^{i\langle \xi, x\rangle} f(x)\, \mathrm{d}x, \qquad \forall \xi \in \mathbb{R}^d. \tag{2.1}$$

In (2.1), the bilinear form $\langle \cdot, \cdot \rangle$ denotes the Euclidian scalar product.

Under certain conditions, an integrable function f can be reconstructed by inverting the associated Fourier transform. The following lemma provides an inversion theorem for smooth functions in one dimension, $d = 1$, that we cite from Stein and Shakarchi (2003).

Lemma 2.2 (Fourier inversion)
Let $\widehat{f} : \mathbb{R} \to \mathbb{R}$ be the Fourier transform of a function $f \in \mathcal{S}(\mathbb{R})$, where

$$\mathcal{S}(\mathbb{R}) = \big\{ f \in C^\infty(\mathbb{R}) \,\big|\, \sup_{x \in \mathbb{R}} |x|^k \, |f^{(l)}(x)| < \infty, \quad \textit{for every } k, l \geq 0 \big\},$$

the so called Schwartz space. Then the relation

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\langle \xi, x\rangle} \widehat{f}(\xi) \, \mathrm{d}\xi, \qquad \forall x \in \mathbb{R} \tag{2.2}$$

holds.

Proof
We refer to the proof of Theorem 1.9 in Stein and Shakarchi (2003). □

When the function f in expression (2.1) is taken to be a probability density function, the respective integral can be cast as an expected value. In this sense, Fourier analysis is easily linked to probability theory. Thus, unsurprisingly, Fourier transforms for many probability density functions have been derived and analyzed. The following lemma gives the Fourier transform of the normal distribution as an example.

Lemma 2.3 (Fourier transform of the Normal density)
Let $f^{\mu,\sigma}$ be the density of the univariate Normal distribution $\mathcal{N}(\mu, \sigma)$ with expected value $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$,

$$f^{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x. \tag{2.3}$$

The Fourier transform $\widehat{f^{\mu,\sigma}} = \mathcal{F}(f^{\mu,\sigma})$ of $f^{\mu,\sigma}$ exists and is given by

$$\widehat{f^{\mu,\sigma}}(\xi) = e^{i\mu\xi} e^{-\frac{1}{2}\sigma^2\xi^2} \tag{2.4}$$

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

Proof
See Theorem 15.12 in Klenke (2008). □

The Fourier transform possesses many convenient properties that we exploit heavily throughout this theses. The following lemma collects some of these properties.

Lemma 2.4 (Properties of the Fourier transform)
Let $y \in \mathbb{R}^d$ and $a \in \mathbb{R}\backslash\{0\}$ be given and let $f, g \in L^1(\mathbb{R}^d)$. Define $f_y = f(\cdot - y)$ and $f^a = f(a\cdot)$. Then, the following equalities hold.

i) The Fourier transform of f shifted by y computes to

$$\widehat{f_y}(\xi) = e^{i\langle \xi, y\rangle} \widehat{f}(\xi), \qquad \forall \xi \in \mathbb{R}^d.$$

ii) The Fourier transform of f with its argument scaled by a computes to

$$\widehat{f^a}(\xi) = \frac{1}{|a|} \widehat{f}(\xi/a), \qquad \forall \xi \in \mathbb{R}^d.$$

iii) *The Fourier transform of a convolution is given by the product of the two individual Fourier transforms,*

$$\widehat{(f * g)}(\xi) = \widehat{f}(\xi)\widehat{g}(\xi), \qquad \forall \xi \in \mathbb{R}^d.$$

Proof

i)–ii) Elementary calculations.

iii) With $f, g \in L^1(\mathbb{R}^d)$, also $f * g \in L^1(\mathbb{R}^d)$. The Fourier transform of the convolution thus exists. Inserting the definition of both the Fourier transform and the convolution we derive for $\xi \in \mathbb{R}^d$

$$\begin{aligned}\widehat{(f * g)}(\xi) &= \int_{\mathbb{R}^d} e^{i\langle \xi, x\rangle} \,(f * g)\,(x) \,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} e^{i\langle \xi, x\rangle} \left[\int_{\mathbb{R}^d} f(x-y)g(y)\,\mathrm{d}y\right] \mathrm{d}x.\end{aligned}$$

By applying Fubini's theorem twice and with the substitution $z = x - y$ we have

$$\begin{aligned}\int_{\mathbb{R}^d} e^{i\langle \xi, x\rangle} \left[\int_{\mathbb{R}^d} f(x-y)g(y)\,\mathrm{d}y\right] \mathrm{d}x &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle \xi, x\rangle} f(x-y)g(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle \xi, z+y\rangle} f(z)g(y)\,\mathrm{d}z\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} e^{i\langle \xi, z\rangle} f(z)\,\mathrm{d}z \int_{\mathbb{R}^d} e^{i\langle \xi, y\rangle} g(y)\,\mathrm{d}y \\ &= \widehat{f}(\xi)\widehat{g}(\xi),\end{aligned}$$

which proves the claim. □

Remark 2.5 (Dampening)

When a function $f : \mathbb{R}^d \to \mathbb{R}$ is not integrable, $f \notin L^1(\mathbb{R}^d)$, its Fourier transform doesn't exist. Yet, if there exists $\eta \in \mathbb{R}^d$ such that

$$f_\eta(x) = e^{\langle \eta, x\rangle} f(x), \qquad \forall x \in \mathbb{R}^d, \tag{2.5}$$

is in $L^1(\mathbb{R}^d)$, we can derive the Fourier transform of f_η and thus introduce the concept of a generalized Fourier transform.

Definition 2.6 (Generalized Fourier transform)

Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\eta \in \mathbb{R}^d$ such that $f_\eta = e^{\langle \eta, \cdot\rangle} f \in L^1(\mathbb{R}^d)$. We call

$$\widehat{f_\eta}(\xi) = \widehat{e^{\langle \eta, \cdot\rangle} f}(\xi), \qquad \forall \xi \in \mathbb{R}^d \tag{2.6}$$

the generalized Fourier transform *of f. We sometimes write*

$$\widehat{f_\eta} = \widehat{f}(\cdot - i\eta). \tag{2.7}$$

We call $\eta \in \mathbb{R}^d$ such that $f_\eta \in L^1(\mathbb{R}^d)$ a dampening constant *and the term $e^{\langle \eta, \cdot\rangle}$ a* dampening factor *of f.*

The following theorem introducing Parseval's identity will be a crucial cornerstone of this thesis. It allows computing the integral of a product of functions by integrating the product of the two respective Fourier transforms, instead. The value of this identity for practical applications becomes evident, when numerical integration of functions is concerned which are difficult to evaluate but posses a Fourier transform in closed form at the same time.

Theorem 2.7 (Parseval's identity)
Let $f, g \in L^2(\mathbb{R}^d)$. Then we have the identity

$$\langle f, g \rangle_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} f(x) g(x) \,\mathrm{d}x = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(\xi) \overline{\widehat{g}(\xi)} \,\mathrm{d}\xi$$

which is called Parseval's identity.

Proof
See Equation (10) on page 187 in Rudin (1987). □

Parseval's identity of Theorem 2.7 draws our attention to integrability properties of Fourier transformed functions. While a function f might be difficult to evaluate, its Fourier transform $\widehat{f}$ might be easy to evaluate, but difficult to integrate. The next remark expands on this issue.

Remark 2.8 (On the relation between smoothness of f and decay of $\widehat{f}$)
There is an interesting relation between the smoothness of a function and the rate of decay of its Fourier transform. Let $f \in C^n(\mathbb{R})$ and $f^{(n)} = \frac{\partial^n}{\partial x^n} f \in L^1(\mathbb{R})$. Then, the Fourier transform of $f^{(n)}$ exists. By repeated integration by parts it can be expressed in terms of $\widehat{f}$ by

$$\begin{aligned}
\widehat{f^{(n)}}(\xi) &= \int_{\mathbb{R}} e^{i\xi x} \frac{\partial^n}{\partial x^n} f(x) \,\mathrm{d}x \\
&= (-i\xi) \int_{\mathbb{R}} e^{i\xi x} f^{(n-1)}(x) \,\mathrm{d}x \\
&= (-i\xi)^n \int_{\mathbb{R}} e^{i\xi x} f(x) \,\mathrm{d}x \\
&= (-i\xi)^n \widehat{f}(\xi)
\end{aligned} \tag{2.8}$$

for all $\xi \in \mathbb{R}$. The Fourier transform of a function in $L^1(\mathbb{R})$ is also in $L^1(\mathbb{R})$. Consequently, $\widehat{f^{(n)}} = (-i\,\cdot)^n \,\widehat{f} \in L^1(\mathbb{R})$. We conclude that $\widehat{f}$ decays faster to zero than $|\xi|^n$ diverges to infinity for $|\xi| \to \pm\infty$. In the same manner, decay properties of the Fourier transform of a function translate into smoothness properties of the function itself.

Relation (2.8) of Remark 2.8 will have a material impact with regards to numerical implications in Chapter 3.

2.2 Lévy process theory

We have already briefly touched upon the relation between Fourier analysis and probability theory in the remarks preceding Lemma 2.3 above. In this section we introduce a class of distributions, or rather a class of stochastic processes, that can even be characterized in Fourier terms, that is the class of Lévy processes. The majority of asset models that we consider in this thesis falls into this class. Models contained therein share the property that the log-asset process is modeled by a Lévy process. We therefore introduce the fundamental definitions and results of Lévy process theory in the following. We begin by citing Sato (2007) for the definition of a probability space and a Lévy process.

Definition 2.9 (Lévy process)
We call a d variate stochastic process $(L_t)_{t\geq 0}$ on a probability space $(\Omega, \mathcal{F}, P)$ a Lévy process *if the following conditions are satisfied.*

i) For any choice of $n \geq 1$ and $0 \leq t_0 < t_1 < \cdots < t_n$, random variables L_{t_0}, $L_{t_1} - L_{t_0}$, $L_{t_2} - L_{t_1}, \ldots, L_{t_n} - L_{t_{n-1}}$ are independent (independent increments property)

ii) $L_0 = 0$ a.s.

iii) The distribution of $L_{s+t} - L_s$ does not depend on s (temporal homogeneity or stationary increments property)

iv) It is stochastically continuous

v) There is $\Omega_0 \in \mathcal{F}$ with $P(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$, $L_t(\omega)$ is right-continuous in $t \geq 0$ and has left limits in $t > 0$.

With $(L_t)_{t\geq 0}$ being a Lévy process, the random variable L_t for $t \geq 0$ belongs to the large class of infinitely divisible distributions. Such distributions and thus also Lévy processes can be beautifully characterized via their Fourier transform.

Lemma 2.10 (Fourier transform of a Lévy process)
Let $(L_t)_{t\geq 0}$ be a Lévy process on $\mathbb{R}^d$. Let $t \geq 0$ arbitrary but fix. The characteristic function *$\widehat{L_t}$ of the random variable L_t is defined as*

$$\widehat{L_t}(\xi) = \mathbb{E}[e^{i\langle \xi, L_t\rangle}], \qquad \forall \xi \in \mathbb{R}^d, \tag{2.9}$$

and there exists a cumulant generating function *θ such that the characteristic function of L_t can be represented by*

$$\widehat{L_t}(\xi) = e^{t\theta(i\xi)}, \qquad \forall \xi \in \mathbb{R}^d, \tag{2.10}$$

with θ given by

$$\theta(i\xi) = i\langle \xi, b\rangle - \frac{1}{2}\langle \xi, \sigma\xi\rangle + \int_{\mathbb{R}^d} e^{i\langle \xi, y\rangle} - 1 - i\langle \xi, h(y)\rangle F(\mathrm{d}y), \qquad \forall \xi \in \mathbb{R}^d, \tag{2.11}$$

with $\sigma \in \mathbb{R}^{d\times d}$ a symmetric, positive semi-definite matrix, a drift term $b \in \mathbb{R}^d$ and a Borel Lévy measure *F satisfying*

$$F(\{0\}) = 0, \qquad \int_{\mathbb{R}^d} \min\{1, |y|^2\} F(\mathrm{d}y) < \infty, \tag{2.12}$$

and for some cut-off function *$h : \mathbb{R}^d \to \mathbb{R}$ that is a bounded measurable function with compact support and*

$$h(x) = x \tag{2.13}$$

in an environment of the origin.

Proof
Confer the proof of Theorem 8.1 in Sato (2007). □

Due to its significance to Lévy theory, the triplet (b, σ, F) characterizing a Lévy process through its cumulant generating function in (2.11) is given a name by the following definition.

Definition 2.11 (Characteristic triplet)
Let $(L_t)_{t\geq 0}$ be a Lévy process. We call the triplet (b, σ, F) of Lemma 2.10 the characteristic triplet *of the Lévy process $(L_t)_{t\geq 0}$.*

Note that the characteristic triplet of a Lévy process depends on the cut-off function h in (2.11). Given an additional property that not all Lévy processes share, the cut-off function can be replaced and the cumulant generating function can be rewritten in the sense of the following remark.

Remark 2.12 (Disregarding the cut-off function)
Let $(L_t)_{t\geq 0}$ be a Lévy process with characteristic triplet (b, σ, F). Identity (2.11) *of Lemma 2.10 states the general form of the cumulant generating function of a Lévy process. If the Lévy measure F additionally satisfies*

$$\int_{|x|\leq 1} |x| F(\mathrm{d}x) < \infty \tag{2.14}$$

we may use the zero function as cut-off function, $h \equiv 0$, leaving us with

$$\theta(i\xi) = i\langle \xi, \widetilde{b}\rangle - \frac{1}{2}\langle \xi, \sigma\xi\rangle + \int_{\mathbb{R}^d} (e^{i\langle \xi, y\rangle} - 1) F(\mathrm{d}y), \qquad \forall \xi \in \mathbb{R}^d, \tag{2.15}$$

with an appropriately adjusted $\widetilde{b} \in \mathbb{R}^d$ given by

$$\widetilde{b} = b - \int_{\mathbb{R}^d} h(y) F(\mathrm{d}y) \tag{2.16}$$

and thus an equivalent characteristic triplet $(\widetilde{b}, \sigma, F)$ with the zero function as cut-off function, compare Remark 8.4 in Sato (2007).

We will need to extend the domain of the characteristic function of a Lévy process to parts of the complex plane. This extension is well-defined under the assumptions of the following theorem taken from Sato (2007).

Theorem 2.13 (Exponential Moments)
Let $(L_t)_{t\geq 0}$ be a Lévy process on $\mathbb{R}^d$ with characteristic triplet (b, σ, F). Let

$$C = \left\{ c \in \mathbb{R}^d \mid \int_{|x|>1} e^{\langle c,x \rangle} F(\mathrm{d}x) < \infty \right\}. \tag{2.17}$$

i) The set C is convex and contains the origin.

ii) $c \in C$ if and only if $\mathbb{E}[e^{\langle c, L_t \rangle}] < \infty$ for some $t > 0$ or, equivalently, for every $t > 0$.

iii) If $w \in \mathbb{C}^d$ is such that $\Re(w) \in C$, then

$$\Psi(w) = \langle b, w \rangle + \frac{1}{2}\langle w, \sigma w \rangle + \int_{\mathbb{R}^d} (e^{\langle w,y \rangle} - 1 - \langle w, h(y) \rangle) F(\mathrm{d}y) \tag{2.18}$$

is definable, $\mathbb{E}[|e^{\langle w, L_t \rangle}|] < \infty$, and

$$\mathbb{E}[|e^{\langle w, L_t \rangle}|] = e^{t\Psi(w)}. \tag{2.19}$$

Proof
For a proof confer the proof of Theorem 25.17 in Sato (2007). □

We are now equipped with the quantities needed to introduce the notion of the symbol of a Lévy process. It will become clear later in the thesis that this concept builds a bridge from Fourier representations of Lévy processes to the theory of partial (integro-) differential equations.

Definition 2.14 (Symbol of a Lévy process)
Let $(L_t)_{t\geq 0}$ be a Lévy process on $\mathbb{R}^d$ with characteristic triplet (b, σ, F). The symbol *$A : \mathbb{R}^d \to \mathbb{C}$ of the Lévy process $(L_t)_{t\geq 0}$ is defined by*

$$A(\xi) = i\langle \xi, b \rangle + \frac{1}{2}\langle \xi, \sigma\xi \rangle - \int_{\mathbb{R}^n} \left(\exp(-i\langle \xi, y \rangle) - 1 + i\langle \xi, h(y) \rangle \right) F(dy) \tag{2.20}$$

for all $\xi \in \mathbb{R}^d$.

The symbol A of a Lévy process is a crucial quantity in this thesis. As pointed out in Glau (2015) one may show that there exists a constant $C > 0$ such that

$$|A(\xi)| \leq C(1 + \|\xi\|)^2, \qquad \forall \xi \in \mathbb{R}^d. \tag{2.21}$$

The notion of a symbol, however, is not exclusively reserved for Lévy processes. Indeed, other measurable functions satisfying inequalities in the fashion of (2.21) are called symbols as well and establish a link between the roots of these quantities in Fourier theory to the topic of partial (integro-)differential equations. To properly generalize the concept of symbols, we first need to cite the definitions of the Schwartz space $S(\mathbb{R}^d)$ from Eskin (1981), that we have already encountered in the special case of $d = 1$ in Lemma 2.2 above.

Definition 2.15 (The Schwartz space $S(\mathbb{R}^d)$)
The space $S = S(\mathbb{R}^d)$ is defined as the totality of all infinitely differentiable functions φ in the d-dimensional space $\mathbb{R}^d$ such that $\varphi(x)$ and all derivatives $\frac{\partial^p}{\partial x^p}\varphi(x)$ with multi-index $p = (p_1, \dots, p_d)$ of nonnegative integers decrease more rapidly than any negative power of $\|x\|$ as $\|x\| = \sqrt{x_1^2 + \cdots + x_d^2} \to \infty$.

Eberlein and Glau (2011) extend the notion of a Schwartz space to the weighted Schwartz space.

Definition 2.16 (The exponentially weighted Schwartz space $S_\eta(\mathbb{R}^d)$)
For $\eta \in \mathbb{R}^d$ let

$$S_\eta(\mathbb{R}^d) = \{u \in C^\infty(\mathbb{R}^d, \mathbb{C}) \,|\, \|u\|_{m,\eta} < \infty, \ \forall m \in \mathbb{N}_0\} \tag{2.22}$$

with

$$\|\varphi\|_{m,\eta} = \left\| e^{\langle \eta, \cdot \rangle} \varphi \right\|_m, \tag{2.23}$$

wherein for every $m \in \mathbb{N}_0$ the norms $\|\cdot\|_m$ are defined by

$$\|\varphi\|_m = \sup_{|p| \leq m} \sup_{x \in \mathbb{R}^d} (1 + |x|^2)^m |D^p \varphi(x)|. \tag{2.24}$$

We denote the dual space of $S_\eta(\mathbb{R}^d)$ by $S^*_\eta(\mathbb{R}^d)$.

Following Eskin (1981) and Glau (2015), we define the general notion of a symbol $A : \mathbb{R}^d \to \mathbb{C}$ and connect it with the concept of pseudo-differential operators.

Definition 2.17 (The class S^0_α and related pseudodifferential operators)
Let $(A_{t \in [0,T]})$ be a family of measurable functions $A : [0, T] \times \mathbb{R}^d \to \mathbb{C}$ satisfying with $\alpha \in (0, 2]$ and $0 \leq \beta < \alpha$

$$\begin{aligned} |A_t(\xi)| &\leq C_1(1 + \|\xi\|^2)^{\alpha/2}, && \forall t \in [0, T], \xi \in \mathbb{R}^d, \\ \Re(A_t(\xi)) &\geq C_2\|\xi\|^\alpha - C_3(1 + \|\xi\|^2)^{\beta/2}, && \forall t \in [0, T], \xi \in \mathbb{R}^d, \end{aligned} \tag{2.25}$$

for some $C_1, C_2 \in \mathbb{R}^+$ and $C_3 \geq 0$ independent of $t \in [0, T]$. Each function A_t is called a symbol. *We denote the set of functions satisfying* (2.25) *by S^0_α. With $t \in [0, T]$, the operator $\mathcal{A}_t$ defined on $S(\mathbb{R}^d)$ by*

$$\mathcal{A}_t u = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} A_t(\xi) \widehat{u}(\xi) e^{-i\langle \cdot, \xi \rangle} \, \mathrm{d}\xi, \qquad \forall u \in S(\mathbb{R}^d), \tag{2.26}$$

is called pseudodifferential operator with symbol *A_t.*

Definition 2.18 (Sobolev index α)
Let A be a symbol. Following Glau (2015) we call the parameter $\alpha \in (0,2]$ of (2.25) the Sobolev index *of symbol A or the* order *of the associated operator $\mathcal{A}$, respectively.*

Remark 2.19 (On the symbol and the Fourier transform of a Lévy process)
Let $(L_t)_{t\geq 0}$ be a Lévy process with characteristic triplet (b,σ,F). Considering Lemma 2.10 and Definition 2.14, we note that the associated symbol A satisfies the relation

$$\begin{aligned} A(\xi) &= i\langle\xi,b\rangle + \frac{1}{2}\langle\xi,\sigma\xi\rangle - \int_{\mathbb{R}^n} \left(\exp(-i\langle\xi,y\rangle) - 1 + i\langle\xi,h(y)\rangle\right) F(dy) \qquad (2.27)\\ &= -\,\theta(-i\xi)\\ &= -\,\theta(i(-\xi)). \end{aligned}$$

Thus, we realize an interesting connection between the Fourier transform of a Lévy process, its cumulant generating function and the symbol in the sense that

$$\widehat{L_t}(\xi) = \exp(t\theta(i\xi)) = \exp(-tA(-\xi)),$$

for all $\xi \in \mathbb{R}^d$.

2.3 Some Lévy asset price models

We present a selection of asset models of Lévy type that will accompany us throughout the whole thesis. Some of these model introductions are taken from Gaß et al. (2015). In what follows we denote by $\widetilde{\mathcal{Q}}$ the parameter space that the model as such is defined on. In later chapters, we will consider these models on possibly restricted parameter spaces $\mathcal{Q} \subseteq \widetilde{\mathcal{Q}}$ only, which is the reason for this minor notational inconvenience. Throughout all model introductions, the constant $r \geq 0$ denotes the risk-free interest rate. Each model is driven by an appropriately chosen Lévy process $(L_t^q)_{t\geq 0}$, $q \in \widetilde{\mathcal{Q}}$. The asset price process is then given by

$$S_t = S_0 e^{L_t^q}, \qquad S_0 > 0, \quad \forall t \in \mathbb{R}^+, \tag{2.28}$$

where (2.28) is understood componentwise when a d-variate model is concerned. For each model we state the characteristic function of L_T^q, $T \in \mathcal{T}$, for some chosen time horizon $\mathcal{T}$ and $q \in \widetilde{\mathcal{Q}}$ that is

$$\varphi_{T,q}(z) = \widehat{L_T^q}(z) = \mathbb{E}\big[e^{\langle iz, L_T^q\rangle}\big], \qquad z \in \mathbb{R}^d. \tag{2.29}$$

In finance, the characteristic function (2.29) of a Lévy process is a useful quantity in pricing, as we will see in the next section. To this end, however, the drift b of the process must be adjusted for the discounted asset process $(S_0 e^{-rt+L_t^q})_{t\geq 0}$ to become a martingale. This is ensured by the so called drift condition. Let $r \geq 0$ denote the risk-less interest rate and (b, σ, F) the triplet of $(L_t^q)_{t\geq 0}$ in (2.29), then the requirement for the discounted asset process of (2.28) to be a martingale is equivalent to

$$\int_{|y|>1} e^y F(\mathrm{d}y) < \infty \tag{2.30}$$

with the drift b being set to

$$b = r - \frac{\sigma^2}{2} - \int_{\mathbb{R}} (e^y - 1 - h(y))F(\mathrm{d}y), \tag{2.31}$$

compare for example Achdou and Pironneau (2005). We present some typical examples for such exponential Lévy models below.

2.3.1 Multivariate Black&Scholes model

The famous model of Black and Scholes (1973) marks the big bang of mathematical finance and earned its two inventors the Nobel price. The model allows the modeling of asset-price movement, albeit on an elementary level from today's point of view. A volatility parameter of the log-asset price process – and additional covariance parameters in the multivariate case – suffice to set up the mathematical model. More precisely, the d-variate Black-Scholes model is driven by a d-variate Brownian motion. The parameter space of the model solely consists of values determining the underlying covariance matrix $\sigma \in \mathbb{R}^{d\times d}$, which is symmetric and positive definite. For a concise representation of the parameter space, we define $\widetilde{\mathcal{Q}}$ as

$$\widetilde{\mathcal{Q}} = \{q \in \mathbb{R}^{d(d+1)/2} \mid \det(\sigma(q)) > 0\} \subset \mathbb{R}^{d(d+1)/2} \tag{2.32}$$

with the function $\sigma : \mathbb{R}^{d(d+1)/2} \to \mathbb{R}^{d\times d}$ defined by

$$\sigma(q)_{ij} = q_{(\max\{i,j\}-1)\max\{i,j\}/2+\min\{i,j\}}, \qquad i, j \in \{1, \dots, d\}. \tag{2.33}$$

By construction, $\sigma(q)$ is symmetric. The characteristic function of the process L_T^q, $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, driving log-returns in the model is then given by

$$\varphi_{T,q}(z) = \exp\Big(T\big(i\langle b, z\rangle - \frac{1}{2}\langle z, \sigma z\rangle\big)\Big), \tag{2.34}$$

for all $z \in \mathbb{R}^d$ with drift $b = b(q) \in \mathbb{R}^d$ adhering to the no-arbitrage condition (2.31)

$$b_i = r - \frac{1}{2}\sigma_{ii}, \qquad i \in \{1, \dots, d\}. \tag{2.35}$$

Note that for each $q \in \widetilde{\mathcal{Q}}$ given by (2.32), the characteristic function of the d-variate Black&Scholes model is analytic in z on the whole of $\mathbb{C}^d$. Figure 2.1 displays some asset price trajectories $(S_t)_{t\in[0,1]}$ in the univariate Black&Scholes model for various values of $\sigma \in \widetilde{\mathcal{Q}}$.

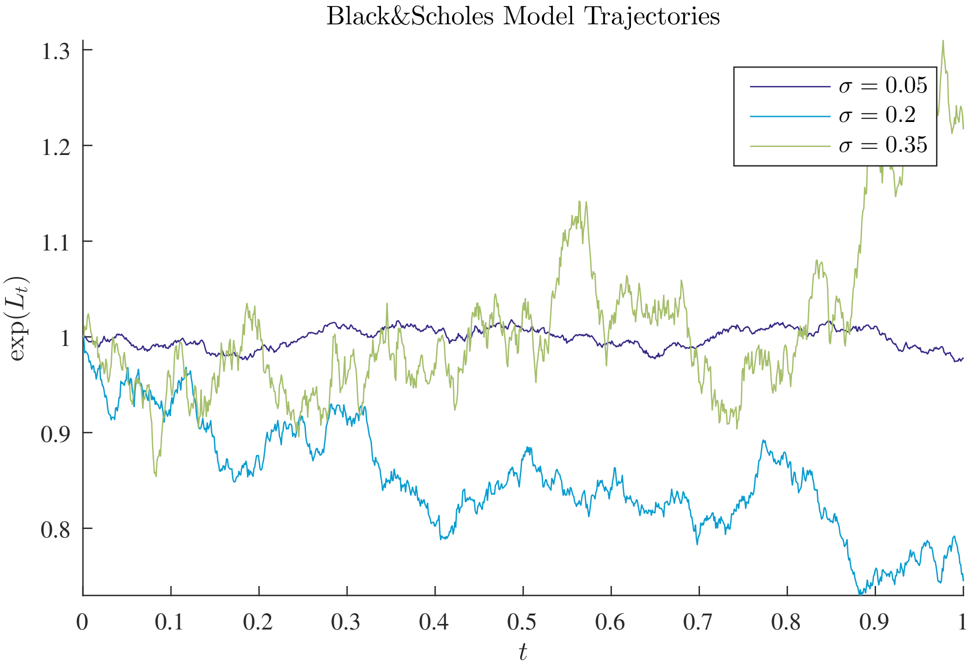

Figure 2.1 Three asset price trajectories in the univariate Black&Scholes model for different parameter sets with $S_0 = 1$ and $r = 0.03$.

2.3.2 Univariate Merton jump diffusion model

In the univariate case, the Merton Jump Diffusion model by Merton (1976) naturally extends the Black&Scholes model to a jump diffusion setting. The logarithm of the asset price process is composed of a Brownian part with variance $\sigma^2 > 0$ and a compound Poisson jump part consisting of normally $\mathcal{N}(\alpha, \beta^2)$ distributed jumps arriving at a rate $\lambda > 0$. The model parameter space is thus given by

$$\widetilde{\mathcal{Q}} = \{(\sigma, \alpha, \beta, \lambda) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}_0^+ \times \mathbb{R}^+\} \subset \mathbb{R}^4 \tag{2.36}$$

and the characteristic function of L_T^q with $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$ computes to

$$\varphi_{T,q}(z) = \exp\left(T\left(ibz - \frac{\sigma^2}{2}z^2 + \lambda\left(e^{iz\alpha - \frac{\beta^2}{2}z^2} - 1\right)\right)\right), \tag{2.37}$$

for all $z \in \mathbb{R}$, with no-arbitrage condition (2.31) demanding

$$b = r - \frac{\sigma^2}{2} - \lambda\left(e^{\alpha + \frac{\beta^2}{2}} - 1\right). \tag{2.38}$$

As in the univariate Black&Scholes model, for each $q \in \mathcal{Q}$ and $T > 0$, the characteristic function $\varphi_{T,q}$ of the Merton model is holomorphic. In Figure 2.2, we simulate three trajectories of the Merton jump diffusion model. Both the structural proximity to the Black&Scholes model and the distinguishing jump feature are clearly visible.

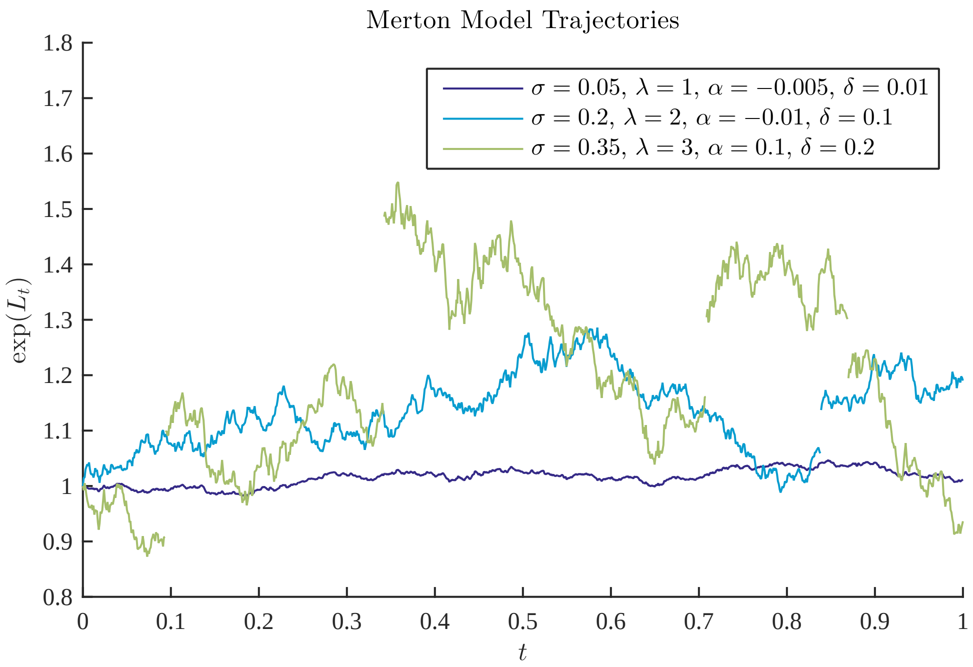

Figure 2.2 Three asset price trajectories in the Merton model for different parameter sets with $S_0 = 1$ and $r = 0.03$.

2.3.3 Univariate CGMY model

Another well known Lévy model that we consider is the univariate CGMY model by Carr et al. (2002). This class is also known as Koponen and KoBoL in the literature, see also Boyarchenko and Levendorskiĭ (2002a) and as tempered stable processes. With the model parameter space given by

$$\widetilde{\mathcal{Q}} = \{(C, G, M, Y) \in \mathbb{R}^+ \times \mathbb{R}_0^+ \times \mathbb{R}_0^+ \times (1, 2) \,|\, (M-1)^Y \in \mathbb{R}\} \subset \mathbb{R}^4, \tag{2.39}$$

the associated characteristic function of L_T^q with $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$ computes to

$$\begin{aligned}\varphi_{T,q}(z) = \exp\big(T\big(ibz + C\Gamma(-Y)\\ \big[(M-iz)^Y - M^Y + (G+iz)^Y - G^Y\big]\big)\big),\end{aligned} \tag{2.40}$$

for all $z \in \mathbb{R}$, where $\Gamma(\cdot)$ denotes the Gamma function. For no-arbitrage pricing we set the drift $b \in \mathbb{R}$ to

$$b = r - C\Gamma(-Y)\left[(M-1)^Y - M^Y + (G+1)^Y - G^Y\right]. \tag{2.41}$$

2.3.4 Univariate Normal Inverse Gaussian model

Another Lévy model we present is the univariate Normal Inverse Gaussian (NIG) model. The parameterization consists of $\delta, \alpha > 0$, $\beta \in \mathbb{R}$, with $\alpha^2 > \beta^2$. The model parameter

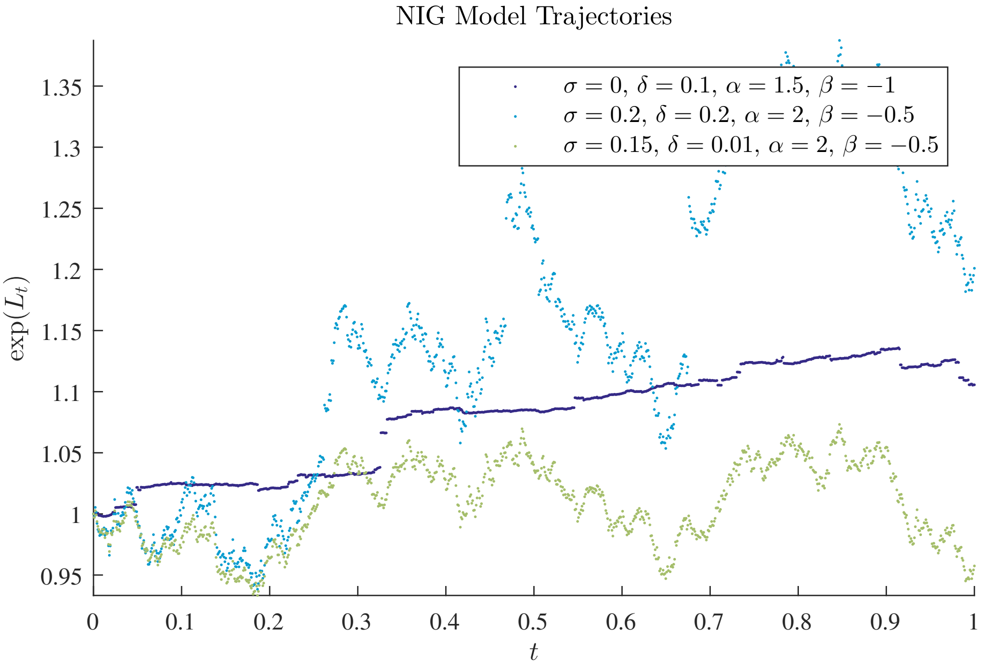

Figure 2.3 Three asset price trajectories in the NIG model for different parameter sets with $S_0 = 1$ and $r = 0.03$.

set $\widetilde{\mathcal{Q}}$ is thus given by

$$\widetilde{\mathcal{Q}} = \left\{(\delta, \alpha, \beta) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R} \,|\, \alpha^2 > \beta^2, \alpha^2 \geq (\beta + 1)^2\right\} \subset \mathbb{R}^3. \tag{2.42}$$

The characteristic function of L_T^q for this model is given by

$$\varphi_{T,q}(z) = \exp\left(T\left(ibz + \delta\left(\sqrt{\alpha^2 - \beta^2} - \sqrt{\alpha^2 - (\beta + iz)^2}\right)\right)\right) \tag{2.43}$$

for $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, wherein the no-arbitrage condition requires

$$b = r - \delta\left(\sqrt{\alpha^2 - \beta^2} - \sqrt{\alpha^2 - (\beta + 1)^2}\right). \tag{2.44}$$

The second condition in (2.42), $\alpha^2 \geq (\beta + 1)^2$, guarantees $b \in \mathbb{R}$. Figure 2.3 displays three sample paths of the NIG model. Graphically, the pure jump characteristic result in paths consisting of dots rather than connected lines.

2.3.5 Multivariate Normal Inverse Gaussian model

The NIG Lévy model exists in a multivariate version. Then, the parameterization consists of $\delta, \alpha > 0$, $\beta \in \mathbb{R}^d$, $\Lambda \in \mathbb{R}^{d \times d}$ symmetric with $\det(\Lambda) = 1$ and $\alpha^2 > \langle \beta, \Lambda\beta \rangle$. The model parameter set $\widetilde{\mathcal{Q}}$ is thus given by

$$\begin{aligned}\widetilde{\mathcal{Q}} = \big\{&(\delta, \alpha, \beta, \lambda) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^{d(d+1)/2} \\ &\mid \alpha^2 > \langle \beta, \Lambda(\lambda)\beta \rangle,\ \det(\Lambda(\lambda)) = 1, \\ &\quad \alpha^2 \geq \langle (\beta + e_i), \Lambda(\lambda)(\beta + e_i) \rangle,\ \forall i \in \{1, \ldots, d\}\big\} \subset \mathbb{R}^{2+d+d^2},\end{aligned} \tag{2.45}$$

where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ for all $i \in \{1, \ldots, d\}$ and wherein we define the function $\Lambda : \mathbb{R}^{d(d+1)/2} \to \mathbb{R}^{d \times d}$ by

$$\Lambda(\lambda)_{ij} = \lambda_{(\max\{i,j\}-1)\max\{i,j\}/2+\min\{i,j\}}, \qquad i, j \in \{1, \ldots, d\}. \tag{2.46}$$

The characteristic function in the d variate NIG model is given by

$$\varphi_{T,q}(z) = \exp\left(T\left(i\langle b, z\rangle + \delta\left(\sqrt{\alpha^2 - \langle \beta, \Lambda\beta\rangle} - \sqrt{\alpha^2 - \langle \beta + iz, \Lambda(\beta + iz)\rangle}\right)\right)\right) \tag{2.47}$$

with $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$. In a multivariate model, the no-arbitrage condition (2.31) must hold componentwise and thus requires

$$b_i = r - \delta\left(\sqrt{\alpha^2 - \langle \beta, \Lambda\beta\rangle} - \sqrt{\alpha^2 - \langle (\beta + e_i), \Lambda(\beta + e_i)\rangle}\right), \tag{2.48}$$

for all $i \in \{1, \ldots, d\}$. Equivalently to its univariate version, the third condition in (2.45), $\alpha^2 \geq \langle (\beta + e_i), \Lambda(\beta + e_i)\rangle$ for all $i \in \{1, \ldots, d\}$, guarantees $b \in \mathbb{R}^d$. Note that for $d = 1$, we have $\Lambda \equiv 1$ and the expression for the d variate characteristic function for the NIG model collapses to its unvariate counterpart. For notational convenience when dealing with the univariate model in numerical experiments, later, however, we decided to split the introduction of the model in the two cases $d = 1$ and $d > 1$.

2.4 Parametric option pricing with Fourier transform

Combining Fourier theory of Section 2.1 with Lévy theory of Section 2.2 in general and invoking the Lévy models we presented in the preceding Section 2.3 in particular now allows us to introduce the main concepts and prerequisites for option pricing based on the Fourier transform. The approach of pricing options using Fourier concepts has been initiated by Stein and Stein (1991) and Heston (1993) and has gained tremendous success in both academia and industry alike. A special emphasis on Lévy models and related models in Fourier pricing has been taken by Carr and Madan (1999), Raible (2000) and Duffie et al. (2000) to which we refer for an in-depth analysis of the matter.

We have given the following introduction into option pricing with Fourier transform methods in Gaß et al. (2015) already where we compute option prices of the form

$$\mathit{Price}^{K,T,q} := \mathbb{E}\big[f_K(L_T^q)\big] \tag{2.49}$$

with parametrized payoff function $f_K : \mathbb{R}^d \to \mathbb{R}$ and a parametric $\mathcal{F}_T$-measurable $\mathbb{R}^d$-valued random variable X_T^q for *payoff and model parameters* $K \in \mathcal{K} \subset \mathbb{R}^{D_1}$, $T \in \mathcal{T} \subset \mathbb{R}^{D_2}$, $q \in \mathcal{Q} \subset \mathbb{R}^{D_3}$ denoting $D = D_1 + D_2 + D_3$. Furthermore, let

$$p = (K, T, q) \in \mathcal{P} \quad \text{where } \mathcal{P} = \mathcal{K} \times \mathcal{T} \times \mathcal{Q}.$$

In order to pass to the pricing formula in terms of Fourier transforms, we impose the following *exponential moment condition* for $\eta \in \mathbb{R}^d$,

$$\mathbb{E}\big[e^{-\langle \eta, X_T^q\rangle}\big] < \infty \quad \text{for all } (T,q) \in \mathcal{T} \times \mathcal{Q}, \tag{Exp}$$

which allows us to define for every $(T,q) \in \mathcal{T} \times \mathcal{Q}$ the extension of the characteristic function of X_T^q to the complex domain $\mathbb{R}^d + i\eta$,

$$\varphi_{T,q}(z) := \mathbb{E}\big[e^{i\langle z, X_T^q\rangle}\big], \qquad \text{for all } z = \xi + i\eta,\ \xi \in \mathbb{R}^d. \tag{2.50}$$

We further introduce the following integrability condition

$$x \mapsto e^{\langle \eta, x\rangle} f_K(x),\ \xi \mapsto \varphi_{T,q}(\xi + i\eta) \in L^1(\mathbb{R}^d) \text{ for all } (K,T,q) \in \mathcal{P}. \tag{Int}$$

As indicated above, the Fourier representation of option prices traces back to the pioneering works of Carr and Madan (1999) and Raible (2000). The following version is an immediate consequence of Theorem 3.2 in Eberlein et al. (2010).

Proposition 2.20 (Fourier pricing)
Let $\eta \in \mathbb{R}^d$ such that (Exp) *and* (Int) *are satisfied. Then for every $(K,T,q) \in \mathcal{P}$,*

$$\mathit{Price}^{K,T,q} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d + i\eta} \widehat{f_K}(-z)\varphi_{T,q}(z)\,\mathrm{d}z. \tag{2.51}$$

Typically, that is for the most common option types, the generalized Fourier transform of f_K is of the form

$$\widehat{f_K}(z) = K^{iz+c} F(z) \tag{2.52}$$

for every $z \in \mathbb{R}^d + i\eta$ with some constant $c \in \mathbb{R}$ and a function $F : \mathbb{R}^d + i\eta \to \mathbb{C}$. Then the option prices (2.51) are indeed parametric Fourier integrals of the form

$$\mathit{Price}^{K,T,q} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d + i\eta} e^{-i\langle z, \log(K)\rangle} K^c F(z) \varphi_{T,q}(z)\,\mathrm{d}z. \tag{2.53}$$

As a first step in the numerical evaluation of (2.53) we employ an elementary symmetry and obtain

$$\int_{\mathbb{R}^d + i\eta} \widehat{f_K}(-z)\varphi_{T,q}(z)\,\mathrm{d}z = 2\int_{\mathbb{R}_+ \times \mathbb{R}^{d-1} + i\eta} \Re\Big(\widehat{f_K}(-z)\varphi_{T,q}(z)\Big)\,\mathrm{d}z, \tag{2.54}$$

which reduces the numerical effort by half.

Lemma 2.21 (Generalized Fourier transform of European vanilla options) *Let $g : \mathbb{R} \to \mathbb{R}_0^+$ be the payoff profile of a European option, that is*

$$g(x) = (e^x - K)^+, \qquad \forall x \in \mathbb{R}, \tag{2.55}$$

for the European call option and

$$g(x) = (K - e^x)^+, \qquad \forall x \in \mathbb{R}, \tag{2.56}$$

for the European put option, respectively. In both payoff profile functions, $K \in \mathbb{R}^+$ denotes the strike price. Then, the generalized Fourier transform computes to

$$\mathcal{F}(g_\eta)(\xi) = \widehat{g_\eta}(\xi) = \frac{K^{i\xi+\eta+1}}{(i\xi+\eta)(i\xi+\eta+1)} \tag{2.57}$$

wherein we choose

$$\begin{aligned} &\eta < \; -1, && \text{for the call option, and} \\ &\eta > 0, && \text{for the put option,} \end{aligned} \tag{2.58}$$

for the generalized Fourier transform to exist.

Proof
The lemma is proved by a straight-forward calculation. □

The structure of the Fourier transform of the payoff profiles of univariate plain vanilla European options extends to the multivariate case as well, as the following lemma demonstrates.

Lemma 2.22 (Generalized Fourier transform of European call on *d* assets) *The payoff profile of a call option on the minimum of d assets with strike $K \in \mathbb{R}^+$ is defined as*

$$f_K(x) = \left(e^{x_1} \wedge e^{x_2} \wedge \cdots \wedge e^{x_d} - K\right)^+, \tag{2.59}$$

for $x = (x_1, \ldots x_d)' \in \mathbb{R}^d$. With weight value $\eta \in \mathbb{R}^d$, $\eta_j < -1$, for all $j \in \{1, \ldots d\}$, the generalized Fourier transform of the multivariate f_K is

$$\widehat{f_K}(z + i\eta) = (-1)^d \frac{-K^{1+\sum_{j=1}^d (iz_j+\eta_j)}}{\prod_{j=1}^d (iz_j + \eta_j)\left(1 + \sum_{j=1}^d (iz_j + \eta_j)\right)}. \tag{2.60}$$

Proof
The result is taken from Example 5.7 in Eberlein et al. (2010). □

2.5 Sobolev spaces

Fourier theory has presented itself as an established theoretical framework for option pricing. By Proposition 2.20, option prices based on the stochastic nature of stock movements are expressed in terms of expected values and transformed to Fourier integrals. Recalling the seminal paper of Black and Scholes (1973), however, we understand that the pricing problem has initially been embedded in the theory of partial differential equations, a field that seems totally unrelated at first sight.

Yet, these two theories are just two different perspectives on the same problem. The first main chapter of this thesis will consider option pricing through the lens that it has been originally discovered with, that is the theory of partial differential equations. As we shall see in the following chapter, for solutions to partial differential equations in finance to exist, the notion of differentiability needs to be weakened. For a univariate, real-valued function f, the *classic* or *strong* derivative at $x \in \operatorname{supp}(f) \subseteq \mathbb{R}$ is given by the limit

$$f'(x) = \frac{\partial}{\partial x} f(x) = \lim_{\substack{h \to 0 \\ x+h \,\in\, \operatorname{supp}(f)}} \frac{f(x+h) - f(x)}{h}, \tag{2.61}$$

so it exists. By this definition, however, the function $\varphi_0 : \mathbb{R} \to \mathbb{R}$, defined by

$$\varphi_0(x) = (1 - |x|) \cdot \mathbb{1}_{|x| \leq 1}$$

is not differentiable at $x \in \{-1, 0, 1\}$ because the limit does not exist for these values. The choice of φ_0 as an example for a function not differentiable everywhere might appear random right now. Yet, precisely functions of this kind will play a key role in the theory of solving partial differentiable equations in the next chapter, both theoretically and numerically. The concept of differentiability must thus be widened until it contains functions like φ_0, as well.

We thus introduce the new concept of so called *weakly differentiable* functions which in a second step will constitute function spaces that solutions to partial differential equations in finance live in. We follow Seydel (2012) in defining the concept that generalizes the classic understanding of a derivative.

Definition 2.23 (Weak derivative)
Let $\Omega \subset \mathbb{R}^n$ and let

$$C_0^\infty(\Omega) = \{v \in C^\infty(\Omega) \mid \operatorname{supp}(v) \textit{ is a compact subset of } \Omega\}.$$

For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $\alpha_i \in \mathbb{N}_0$ for all $i \in \{1, \ldots, n\}$ define

$$|\alpha| = \sum_{i=1}^{n} \alpha_i. \tag{2.62}$$

With α a multi-index we call

$$(D^\alpha v)(x_1, \ldots, x_n) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} v(x_1, \ldots, x_n) \tag{2.63}$$

the partial derivative of v of order $|\alpha|$. *Let* $u : \Omega \to \mathbb{R}$ *be a real-valued function. If there exists* $w \in L^2(\Omega)$ *with*

$$\int_\Omega u\, D^\alpha v \,\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega w\, v \,\mathrm{d}x, \qquad \textit{for all } v \in C_0^\infty(\Omega), \tag{2.64}$$

we define $D^\alpha u = w$ *the* weak derivative of u *with multi-index* α. *Sometimes we also call* $D^\alpha u$ *the derivative of* u in distributional sense.

From Definition 2.23 we understand that weak differentiability is not a pointwise property like strong differentiability is but rather a global property that acts on integration against test functions. Having Definition 2.23 at hand, we can now build up new function spaces and introduce the notion of Sobolev spaces.

Definition 2.24 (Sobolev spaces $\boldsymbol{H^k}$)
Let $\Omega \subset \mathbb{R}^n$ *and* $k \in \mathbb{N}_0$. *We define the* Sobolev space

$$H^k(\Omega) = \left\{ v \in L^2(\Omega) \,|\, D^\alpha v \in L^2(\Omega) \textit{ for } |\alpha| \le k \right\}, \tag{2.65}$$

with $D^\alpha\cdot$ *being the weak derivative of Definition 2.23. For* $a < b \in \Omega$ *we define the subspace* $H_0^k(a,b) \subset H^k(\Omega)$ *by*

$$H_0^k(a,b) = \Big\{ v \in H^k(\Omega) \,|\, v(a) = v(b) = 0 \Big\}. \tag{2.66}$$

For the upcoming definition of fractional Sobolev spaces H^s, $s \in \mathbb{R}^+$, we follow Glau (2010).

Definition 2.25 (Fractional Sobolev spaces $\boldsymbol{H^s}(\mathbb{R}^{\boldsymbol{d}})$)
Let $s \in \mathbb{R}^+$. *We define*

$$H^s(\mathbb{R}^d) = \Big\{ v \in S'(\mathbb{R}^d) \,\Big|\, \mathcal{F}(v) \in L^1_{loc}(\mathbb{R}^d), \textit{ such that } \|v\|_{H^s(\mathbb{R}^d)} < \infty \Big\}, \tag{2.67}$$

wherein $\mathcal{F}(v)$ *denotes the Fourier transform of* v, *see Definition 2.1 and the norm* $\|\cdot\|_{H^s}$ *is given by*

$$\|v\|_{H^s(\mathbb{R}^d)} = \sqrt{\int_{\mathbb{R}^d} |\mathcal{F}(v)(\xi)|^2 \,(1+|\xi|)^{2s} \,\mathrm{d}\xi}, \qquad \forall v \in S'(\mathbb{R}^d). \tag{2.68}$$

We call the space $H^s(\mathbb{R}^d)$ *a* fractional Sobolev space of order s.

Definition 2.26 (Fractional Sobolev spaces $\widetilde{\boldsymbol{H}}^{\boldsymbol{s}}(\boldsymbol{a},\boldsymbol{b})$)
For $s \in \mathbb{R}^+$ *and* $a < b \in \mathbb{R}$ *we define by*

$$\widetilde{H}^s(a,b) = \left\{ v \in H^s(\mathbb{R}) \,\middle|\, v|_{\mathbb{R}\setminus[a,b]} = 0 \right\} \tag{2.69}$$

a subspace $\widetilde{H}^s(\mathbb{R}) \subset H^s(\mathbb{R})$ *of the fractional Sobolev space of Definition 2.25.*

Definition 2.27 (Sobolev space $H^1(\Omega)$)
The space $H^1(\Omega)$ denotes the space of functions $u \in L^2(\Omega)$ that possess a weak derivative (of first order) in $L^2(\Omega)$. The scalar product of $H^1(\Omega)$ is defined by

$$(u,v)_{H^1(\Omega)} := (\partial u, \partial v)_{L^2(\Omega)} + (u,v)_{L^2(\Omega)} = \int_\Omega \partial u(x)\partial v(x)\,\mathrm{d}x + \int_\Omega u(x)v(x)\,\mathrm{d}x. \quad (2.70)$$

Consequently, the norm of the space, $\|\cdot\|_{H^1(\Omega)}$ is given by

$$\|u\|_{H^1(\Omega)} = \sqrt{(u,u)_{H^1(\Omega)}}, \quad (2.71)$$

for all $u \in H^1(\Omega)$.

Even though Sobolev spaces contain functions that are not even differentiable in the strong sense, they maintain a close relationship to infinitely smooth functions in the strong sense, as the following theorem emphasizes.

Theorem 2.28 ($C^\infty(\Omega) \cap H^1(\Omega)$ dense in $H^1(\Omega)$)
The intersection of $C^\infty(\Omega)$ with $H^1(\Omega)$, $C^\infty(\Omega) \cap H^1(\Omega)$ is dense in $H^1(\Omega)$.

Proof
The claim follows from Theorem 3.5 in Wloka (2002) where a proof is provided. □

Definition 2.29 ($H_0^1(\Omega)$)
The completion of the space $C_0^\infty(\Omega)$ in the norm $\|\cdot\|_{H^1(\Omega)}$, is denoted by $H_0^1(\Omega)$,

$$H_0^1(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^1(\Omega)}}. \quad (2.72)$$

In the Fourier section above, we have already encountered the idea of exponentially weighting non-integrable functions with an appropriately chosen value $\eta \in \mathbb{R}^d$ to achieve integrability of the transformed result. There, the weighting approach aimed at making Fourier pricing accessible to plain vanilla European call and put options, the payoff functions of which lack integrability and can thus not be Fourier transformed. The following definition extends the weighting approach to Sobolev spaces that we have just introduced. In the context of pricing plain vanilla European call and put options, weighted Sobolev spaces will be as important to PDE theory as the generalized Fourier transform has been to Fourier pricing. We give the respective definition following Eberlein and Glau (2011).

Definition 2.30 (Weighted Sobolev-Slobodeckii space $H_\eta^s(\mathbb{R}^d)$)
Let $s \in \mathbb{R}$ and $\eta \in \mathbb{R}^d$. The weighted Sobolev-Slobodeckii space *$H_\eta^s(\mathbb{R}^d)$ is defined by*

$$H_\eta^s(\mathbb{R}^d) = \left\{ u \in S_\eta^*(\mathbb{R}^d) \mid \left\| \mathcal{F}(e^{\langle \eta, \cdot \rangle} u) \right\|_{\widehat{H}^s} < \infty \right\} \quad (2.73)$$

with the scalar product

$$\langle u, v \rangle_{H_\eta^s} = \langle \mathcal{F}(e^{\langle \eta, \cdot \rangle} u), \mathcal{F}(e^{\langle \eta, \cdot \rangle} v) \rangle_{\widehat{H}^s} \quad (2.74)$$

with

$$\langle \varphi, \psi \rangle_{\widehat{H}^s} = \int_{\mathbb{R}^d} \varphi(\xi)\overline{\psi(\xi)}(1 + |\xi|)^{2s}\,\mathrm{d}\xi. \quad (2.75)$$

2.6 Other concepts

This final section of the preliminary chapter summarizes some other concepts that we will encounter within the thesis. We begin by stating some definitions concerning Banach, Hilbert and related spaces. The section closes with a repetition of other elementary definitions. Stating them now in the preliminary section will later allow us to present our main results without unnecessary distractions.

We state the definition of a Banach and a Hilbert space that we took from Grossmann et al. (2007).

Definition 2.31 (Banach space)
Let U be a linear space endowed with a norm $\|\cdot\|_U : U \to \mathbb{R}$ that is a mapping with the following properties

i) $\|u\|_U \geq 0, \qquad$ *for all* $u \in U, \qquad \|u\|_U = 0 \Leftrightarrow u = 0,$

ii) $\|\lambda u\|_U = |\lambda| \|u\|_U, \qquad$ *for all* $u \in U,\ \lambda \in \mathbb{R},$

iii) $\|u + v\|_U \leq \|u\|_U + \|v\|_U, \qquad$ *for all* $u, v \in U.$

The space U endowed with the norm $\|\cdot\|_U$ is called a normed space. *A normed space is called* complete *if every Cauchy sequence $(u_k)_{k\geq 1} \subset U$ converges in U. Complete normed spaces are called* Banach spaces.

Definition 2.32 (Hilbert space)
Let $\mathcal{H}$ be a Banach space. If the norm $\|\cdot\|_{\mathcal{H}}$ in the space is induced by the scalar product $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$,

$$\|u\|_{\mathcal{H}} = \sqrt{\langle u, u\rangle_{\mathcal{H}}}, \qquad \forall u \in \mathcal{H}, \tag{2.76}$$

we call the space $\mathcal{H}$ a Hilbert space.

Definition and Theorem 2.33 (Separability of Hilbert spaces)
Let $\mathcal{H}$ be a Hilbert space. If $\mathcal{H}$ is finite dimensional, then it is separable, that is it contains a countable dense subset. If $\mathcal{H}$ is infinite dimensional, it is separable if and only if it has an orthonormal basis.

Proof
Consider the proof of Theorem 3.52 in Rynne and Youngson (2000). □

The next few definitions and results build on the Hilbert space theory and prepare it for the notion of solution spaces to partial differential equations in finance.

Definition 2.34 (The space $L^2(0,T;\mathcal{H})$)
For each Hilbert space $\mathcal{H}$ we define the function space $L^2(0,T;\mathcal{H})$ by

$$L^2(0,T;\mathcal{H}) = \big\{u : [0,T] \to \mathcal{H} \,\big|\, \int_0^T \|u(t)\|^2_{\mathcal{H}}\,\mathrm{d}t < \infty\big\}. \tag{2.77}$$

We take the definition of a Riesz basis from Christensen (2013).

Definition 2.35 (Riesz basis)
Let $\mathcal{H}$ be a Hilbert space. A Riesz basis *for $\mathcal{H}$ is a family of the form $(U\, e_k)_{k\geq 1}$, where $(e_k)_{k\geq 1}$ is an orthonormal basis for $\mathcal{H}$ and $U : \mathcal{H} \to \mathcal{H}$ is a bounded bijective operator.*

We follow page 15 from Arendt et al. (2011) and give the following definition.

Definition 2.36 (The space $C^n([0,T];\mathcal{H})$)
Let $\mathcal{H}$ be a Banach space and $T > 0$. We denote by $C([0,T];\mathcal{H})$ the vector space of all continuous functions $f : [0,T] \to \mathcal{H}$. With $n \in \mathbb{N}$ we denote by $C^n([0,T];\mathcal{H})$ the vector space of all n times differentiable functions with continuous n-th derivative, that is the space of all functions f such that for all $k \in \{0, \dots, n-1\}$ the limits

$$f^{(k+1)}(t) = \lim_{\substack{\Delta t \to 0 \\ t+\Delta t \in [0,T]}} \frac{f^{(k)}(t+\Delta t) - f^{(k)}(t)}{\Delta t}$$

exist for all $t \in [0,T]$ with $f^{(0)}, \dots, f^{(n)}$ being continuous and the convention $f^{(0)} \equiv f$.

We cite the following Definition 2.37 from page 15 of Arendt et al. (2011).

Definition 2.37 (Absolute continuity of a function)
Let $a < b \in \mathbb{R}$ and let X be a Banach space. Let $f : [a,b] \to X$. We say that f is absolutely continuous *on $[a,b]$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\sum_{i\in I} \|f(b_i) - f(a_i)\|_X < \varepsilon \tag{2.78}$$

for every finite set $\{(a_i, b_i)\}_{i\in I}$, $I \subset \mathbb{N}$, $|I| < \infty$, of disjoint intervals in $[a,b]$ with $\sum_{i\in I}(b_i - a_i) < \delta$.

Consider also Chapter VII in Elstrodt (2011) on the notion of absolute continuity. It is well known that absolute continuity is a weaker concept than continuous differentiability as far as functions on compacts are concerned. In other words, continuous differentiability of a function defined on a compact interval implies absolute continuity as the following lemma demonstrates. We give a short proof for the reader's convenience.

Lemma 2.38 (Absolute continuity of continuously differentiable functions)
Let $f : [a,b] \to X$ with $|a|, |b| < \infty$ and X a normed vector space and assume f to be continuously differentiable. Then f is absolutely continuous on $[a,b]$.

Proof
Let $\varepsilon > 0$. With f being continuously differentiable, f' is continuous and as a function defined on a compact set it is thus bounded. Let

$$M = \max_{x\in[a,b]} \left\|f'(x)\right\|_X. \tag{2.79}$$

Choose $\delta < \varepsilon/M$. Now choose an arbitrary finite set $\{(a_i, b_i)\}_{i\in I}$ of disjoint intervals in $[a, b]$ with $\sum_{i\in I}(b_i - a_i) < \delta$. Without loss of generality we may assume $a_i < b_i$ for all $i \in I$. Then

$$\sum_{i\in I} \|f(b_i) - f(a_i)\|_X = \sum_{i\in I} \left\| \frac{f(b_i) - f(a_i)}{b_i - a_i} \right\|_X (b_i - a_i) \leq M \sum_{i\in I} (b_i - a_i) < \varepsilon \tag{2.80}$$

which proves that f is absolutely continuous on $[a, b]$. □

We introduce the notion of the Bochner integral strictly following Definitions and Theorem 24.6 in Wloka (2002).

Definitions and Theorem 2.39 (Bochner integral)
Let $\mathcal{H}$ be a separable Hilbert space.

i) Let E denote the set of finitely valued functions $x : S \to \mathcal{H}$. E is a linear set and $E \subset L^1(S, \mathcal{H})$. If $x \in E$ we define

$$\int_S x(s)\,\mathrm{d}m(s) = \sum_{i=1}^{n} x_i\, m(B_i), \tag{2.81}$$

where $im(x) = \{x_1, \ldots, x_n, 0\}$ and $B_i = x^{-1}(x_i)$ for $i \in \{1, \ldots, n\}$. The integral is linear and

$$\left\| \int_S x(s)\,\mathrm{d}m(s) \right\| \leq \int_S \|x(s)\|\,\mathrm{d}m(s). \tag{2.82}$$

ii) We write $B^1(S, \mathcal{H}) = \overline{E}^{L^1(S;\mathcal{H})}$ and call $B^1(S, \mathcal{H})$ the set of Bochner integrable *functions. If $x \in B^1(S, \mathcal{H})$ there exists a sequence $(x_n)_{n\geq 1}$, $x_n \in E$ for all $n \geq 1$, with $x_n \to x$ in $L^1(S; \mathcal{H})$ as $n \to \infty$. We put*

$$\int_S x(s)\,\mathrm{d}m(s) = \lim_{n\to\infty} \int_S x_n(s)\,\mathrm{d}m(s). \tag{2.83}$$

In the theory of real-valued functions that are differentiable, Taylor's theorem links the evaluation of a differentiable function to a weighted sum of its derivatives and a remainder term that can be expressed in a (Riemann) integral form. Using the Bochner integral of Definitions and Theorem 2.39, the theorem extends to functions mapping real values to Hilbert spaces. The Taylor theorem for these Hilbert space valued functions will be central to the error and convergence analysis of approximate solutions to partial differential equations in finance, later.

Theorem 2.40 (Taylor's theorem)
With $n \in \mathbb{N}$, $T > 0$ and $\mathcal{H}$ a separable Hilbert space, assume $f \in C^n([0, T]; \mathcal{H})$. Let $t_0 \in [0, T]$ and $\Delta t > 0$ such that $t_0 + \Delta t \leq T$. Then

$$f(t_0 + \Delta t) = \sum_{k=0}^{n-1} \frac{1}{k!} f^{(k)}(t_0)\Delta t^k + \int_{t_0}^{t_0+\Delta t} \frac{(t_0 + \Delta t - s)^{n-1}}{(n-1)!} f^{(n)}(s)\,\mathrm{d}s \tag{2.84}$$

holds.

Proof
Assume first that $n = 1$. With f being continuously differentiable on a compact interval, Lemma 2.38 yields that f is absolutely continuous. We may thus apply Proposition 1.2.3 in (Arendt et al., 2011) which gives

$$f(t_0 + \Delta t) - f(t_0) = \int_{t_0}^{t_0+\Delta t} f^{(1)}(s)\,\mathrm{d}s \tag{2.85}$$

and thus confirms formula (2.84) for $n = 1$. For general $n \in \mathbb{N}$, taking (2.85) as induction assumption, the claim now follows from induction using integration by parts. □

We will derive approximate solutions to partial differential equations numerically by a so-called finite element approach. The method consists of an iterative scheme that is driven by two key matrices. When we investigate the method more closely, the two core matrices will usually have be of a so-called Toeplitz structure in the sense of the following definition.

Definition 2.41 (Toeplitz matrix)
Let $M \in \mathbb{R}^{N\times N}$ be a real valued matrix. We call M a Toeplitz matrix *if there exists a set $\{v_{-(N-1)}, \dots, v_{-1}, v_0, v_1, \dots v_{N-1}\} \subset \mathbb{R}$ such that*

$$M = \begin{pmatrix} v_0 & v_1 & v_2 & \cdots & v_{N-1} \\ v_{-1} & v_0 & v_1 & \ddots & \vdots \\ v_{-2} & \ddots & \ddots & \ddots & v_2 \\ \vdots & \ddots & v_{-1} & v_0 & v_1 \\ v_{-(N-1)} & \cdots & v_{-2} & v_{-1} & v_0 \end{pmatrix}.$$

We sometimes also say M has a Toeplitz structure.

We state Hölder's well known inequality which will contribute significantly in Chapter 3 during the derivation of stability and convergence results of approximate solutions to partial (integro) differential equations.

Theorem 2.42 (Hölder's inequality)
Let $f, g \in L^1(\mathbb{R})$ be real valued integrable functions. Let $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then the inequality

$$\int_{\mathbb{R}} |f(x)g(x)|\,\mathrm{d}x \le \left(\int_{\mathbb{R}} |f(x)|^p\,\mathrm{d}x\right)^{1/p} \left(\int_{\mathbb{R}} |g(x)|^q\,\mathrm{d}x\right)^{1/q}.$$

holds.

Finally, recall the definition of a Bernstein ellipse as introduced by Bernstein (1912). It describes an ellipse in the complex plane with foci at ±1, as the following definition

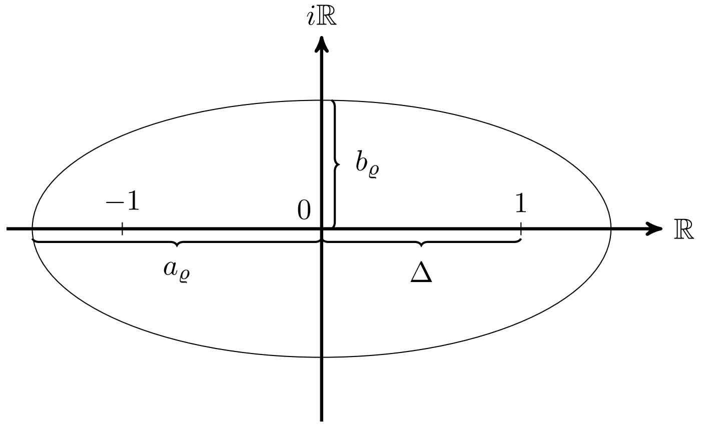

Figure 2.4 A Bernstein Ellipse $B([-1,1],\varrho)$ with foci -1, 1 and ellipse parameter $\varrho > 1$. The semimajor a_ϱ is part of the real line, the semiminor b_ϱ is part of the complex line. Here, Δ denotes the distance from either of the two foci to the center of the ellipse. For the ellipse parameter ϱ, the identity $\varrho = a_\varrho + b_\varrho$ holds.

states. In Chapter 4, the ellipse will characterize areas of analyticity of functions that we approximate with an interpolation approach. Therein, we reserve the flexibility of reshaping the classic Bernstein ellipse to a more general one in order to capture more individual areas of analyticity that the functions we approximate possess.

Definition 2.43 ((Generalized) Bernstein ellipse)
We define the Bernstein ellipse $B([-1,1],\varrho) \subset \mathbb{C}$ with parameter $\varrho > 1$ as the open region in the complex plane bounded by the ellipse with foci ± 1 and semiminor and semimajor axis lengths summing up to ϱ. We set the origin as the center and set the semimajor axis to lie on the real axis. Based on the concept of the Bernstein ellipse we define for $\underline{b} < \overline{b} \in \mathbb{R}$ the generalized Bernstein ellipse by

$$B([\underline{b},\overline{b}],\varrho) := \tau_{[\underline{b},\overline{b}]} \circ B([-1,1],\varrho), \tag{2.86}$$

where the transform $\tau_{[\underline{b},\overline{b}]} : \mathbb{C} \to \mathbb{C}$ is defined for every $z \in \mathbb{C}$ as

$$\tau_{[\underline{b},\overline{b}]}(z) = \overline{b} + \frac{\underline{b} - \overline{b}}{2}\big(1 - \Re(z)\big) + i\,\frac{\overline{b} - \underline{b}}{2}\Im(z). \tag{2.87}$$

Additionally, for an arbitrary set $Z \subset \mathbb{R}$, we define the generalized Bernstein ellipse by

$$B(Z,\varrho) := B([\inf Z, \sup Z],\varrho). \tag{2.88}$$

We call $\varrho > 1$ the ellipse parameter *of the (generalized) Bernstein ellipse.*

A Bernstein ellipse is depicted in Figure 2.4. The figure also depicts the relation between the ellipse semimajor a_ϱ and the semiminor b_ϱ in comparison to the location of the ellipse foci. The sum of the two ellipse axis lengths determines the ellipse parameter ϱ. The following remark states the relations between these quantities for later reference.

Remark 2.44 (Ellipse semiminor and semimajor)
Let $B([-1,1],\varrho)$ with $\varrho > 1$ be a Bernstein ellipse with semimajor a_ϱ and semiminor b_ϱ satisfying $a_\varrho + b_\varrho = \varrho$. Let Δ be the difference from either of the two foci of an ellipse to the center of the ellipse, then

$$\Delta = \sqrt{a_\varrho^2 - b_\varrho^2} \tag{2.89}$$

holds. In a Bernstein ellipse, $\Delta = 1$. From this, the well known relations

$$a_\varrho = \frac{\varrho + \frac{1}{\varrho}}{2}, \qquad b_\varrho = \frac{\varrho - \frac{1}{\varrho}}{2} \tag{2.90}$$

immediately follow.

3 PIDEs and option pricing

This chapter addresses some aspects of the theory of partial integro-differential equations in the context of finance and beyond. In abstract terms, we are considering problems of form

$$\begin{aligned} \partial_t u + \mathcal{A}u &= f, \\ u(0) &= g, \end{aligned}$$

for so-called partial integro-differential operators $\mathcal{A}$. Confronted with such a problem many questions naturally arise. Is there a function u that solves the problem? Is it unique? Is this u numerically accessible, or can we only dispose of it in theory? Are there methods to approximate u and how accurately are they? Do they converge?

Some of these questions address purely theoretical aspects of the problem. Others concern rather numerical issues and are answered in algorithmic terms. Still others cannot be assigned to either of these two categories but lie in the intersection where PIDE theory and numerical concepts blend.

This ambiguity draws through the whole chapter. On the one hand, it challenges the reader by confronting him with separate fields neither of which can be omitted in the derivation of numerical solutions that rest on solid theoretical grounds. On the other hand, it provides two perspectives onto the same problem that complement each other and foster extensive understanding.

Similarly, the contents of the following sections do not fall into strictly separated categories. Some have a strong theoretical focus, some emphasize numerical implications and some address the intersection of both realms.

We therefore highlight the four main sections in this chapter and briefly comment on their main emphasis. In Section 3.1 we present the theoretical framework of PIDE theory. The sections answers the question of existence and uniqueness of solutions u to problems as above and introduces the function spaces that a solution u lies in. The consecutive Section 3.2 illustrates the bridge from the theoretical problem to a numerical solution approach. It provides the theoretical foundation that approximate numerical solution schemes rely on. Section 3.3 is devoted to the development of a numerical solver for the PIDE of the well known asset model by Merton (1976). It implements the theoretical steps taken in the sections before and makes the theory explicitly comprehensible. Then, Section 3.4 abstracts from the Merton model and presents a very general framework for a FEM solver that easily adapts to many different models. After that, Section 3.5 compares all implementations empirically by presenting empirical order of convergence studies for several FEM implementations and a variety of models. Finally, Section 3.6 reconciles the numerical approximation with the solution provided by theory by deriving stability and convergence results in very general terms.

3.1 Existence and uniqueness of (weak) solutions to PDEs

Let us state the main interest of this chapter more concisely. We are interested in finding solutions $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ to problems of the form

$$\begin{aligned} \partial_t u + \mathcal{A}_t u &= f, \qquad \text{for almost all } t \in (0,T) \\ u(0) &= g, \end{aligned} \tag{3.1}$$

with $\mathcal{A} = (\mathcal{A}_t)_{t\in[0,T]}$ a time-inhomogeneous *Kolmogorov operator*, a *source term* or *right hand side* $f : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ and an *initial condition* $g : \mathbb{R}^d \to \mathbb{R}$.

Existence and uniqueness of such solutions u and the properties of the spaces that they live in depend heavily on the properties of the operator $\mathcal{A}$ as well as of properties like smoothness of the two functions f and g.

A well known example for a PDE in the form of (3.1) is the so-called heat equation,

$$\begin{aligned} \partial_t u - c^2 \frac{\partial^2}{\partial x^2} u &= 0, \qquad \text{for almost all } t \in (0,T) \\ u(0) &= g, \end{aligned} \tag{3.2}$$

with $c \in \mathbb{R}+$, $g \in C_0^\infty(\mathbb{R})$. By a Fourier approach one derives the solution $u \in C^{1,2}(\mathbb{R}^+, \mathbb{R})$ given by

$$u(t,x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\xi x} e^{-tc^2\xi^2} \widehat{g}(\xi)\, \mathrm{d}\xi, \qquad \forall (t,x) \in \mathbb{R}^+ \times \mathbb{R} \tag{3.3}$$

with $\widehat{g}$ the Fourier transform of g, see Cannon and Browder (2008).

The function u defined by (3.3) solves the heat equation of problem (3.2) pointwise. It is also called a *strong* solution, since it interprets the differential operator in the PDE in the strong sense of (2.61) as stated in the preliminary chapter above. Differentiability of this kind is indeed a strong property. In finance, we can not expect such strongly differentiable solutions to PDEs to exist, let alone smooth ones like u above. Think for example of the nondifferentiable payoff profiles of call and put options that lead to initial conditions $g \notin C_0^\infty(\mathbb{R})$ which affects the regularity of u accordingly. Consider in this context Eberlein and Glau (2014) for an approach deriving solutions in the form of (3.3) to PDEs in finance.

Consequently, the notion specifying the solution to a PDE must adapt to this issue of regularity.

One way of adjusting the concept of a solution to a PDE is pursued by the theory of *viscosity solutions*. Let us briefly touch upon this first possibility in the following before moving on. In Bardi et al. (1997), an analysis of viscosity solutions to PDEs of second order is provided. The authors analyze scalar-valued functions $u : \Omega \to \mathbb{R}$ that solve partial differential equations in the general form of

$$F(x, u, Du, D^2u) = 0 \tag{3.4}$$

on the open set Ω in the sense of the following definitions taken from Bardi et al. (1997).

Definition 3.1 (Upper and lower semicontinuous envelope)
Let $u : \Omega \to \mathbb{R}$. The notions of the upper semicontinuous envelope u^* *and the* lower semicontinuous envelope u_* *of u are given by*

$$\begin{cases} u^*(x) = & \limsup\limits_{r\downarrow 0}\{u(y) \,:\, y \in \Omega,\ |y-x| \le r\} \\ u_*(x) = & \liminf\limits_{r\downarrow 0}\{u(y) \,:\, y \in \Omega,\ |y-x| \le r\} \end{cases}$$

and u is upper semicontinuous *if $u = u^*$ and u is* lower semicontinuous *if $u = u_*$.*

Definition 3.2 (Viscosity solution)
Let $\mathcal{S}(N)$ be the set of real symmetric $N \times N$ matrices and F of (3.4) *be a function $F : \Omega \times \mathbb{R} \times \mathbb{R}^N \times \mathcal{S}(N) \to \mathbb{R}$ with $F(x, r, p, X) \le F(x, r, p, Y)$ for $Y \le X$ with the ordering $X \le Y$, if $\langle X\xi, \xi\rangle \le \langle Y\xi, \xi\rangle$ for all $\xi \in \mathbb{R}^N$, and let further F be nondecreasing in the second argument. Then u is a* viscosity subsolution (supersolution) *to PDE* (3.4) *in Ω if it is upper (lower) semicontinuous and for every $\varphi \in C^2(\Omega)$ and local maximum (minimum) point $\widehat{x} \in \Omega$ of $u - \varphi$ we have*

$$\begin{aligned} F(\widehat{x}, u(\widehat{x}), D\varphi(\widehat{x}), D^2\varphi(\widehat{x})) \le & 0 \\ (F(\widehat{x}, u(\widehat{x}), D\varphi(\widehat{x}), D^2\varphi(\widehat{x})) \ge & 0). \end{aligned}$$

And finally, u is called a viscosity solution *to* (3.4) *if it is a viscosity subsolution and a viscosity supersolution.*

We have encountered two different concepts in interpreting the notion of a solution to PDEs or PIDEs, respectively, the first one being the strong solution, the second one being the viscosity solution and further concepts exist as well. The fact that we actually have a choice in selecting a solution scheme to solve PIDEs fuels the suspicion that the eventual decision critically depends on the features that we expect from that solution scheme and the subsequent solution itself. Let us highlight the main goals that we pursue in deriving solutions to PIDEs. These are

i) possibility for thorough error control

ii) algorithmic accessibility

iii) numerical feasibility

A scheme that provides all these features is the Galerkin method which is based on the notion of weak differentiability. It powerfully combines a theoretical concept with an algorithmic translation that opens the method to numerically feasible implementations. At the same time it offers error control methods that manage to monitor inaccuracies inevitably arising from those numerical schemes. In order to be able to apply the Galerkin method we need to weaken the idea of differentiability of a function by replacing the strong derivative by a more general concept. The new notion of a derivative does no longer take effect in a pointwise fashion. Instead, *weak differentiability* acts on integration against test functions. Preparing our introduction of the associated idea of a *weak solution*, we cite Definition 17.1 of a Gelfand triplet from Wloka (2002).

Definition 3.3 (Gelfand triplet)
Let V be an (anti)reflexive Banach space and H a Hilbert space. Suppose $V \underset{i}{\hookrightarrow} H$ and that the embedding i is continuous, injective and that $\operatorname{im} i$ is dense in H. Let $i' : H \to V^*$ be continuous and injective and $\operatorname{im} i'$ dense in V^*. Altogether we have

$$V \underset{i}{\hookrightarrow} H \underset{i'}{\hookrightarrow} V^*, \tag{3.5}$$

where both embeddings i, i' are continuous, injective and have dense images in H and V^*. A scheme of this kind is called a Gelfand triplet. *For notational convenience we omit the symbols i and i' from here on.*

Based on Definition 2.34 and Definition 3.3 we define the *solution space* $W^1(0,T;V,H)$ for special choices of separable Hilbert spaces V and H.

Definition 3.4 (The solution space $\boldsymbol{W^1(0,T;V,H)}$)
Assume separable Hilbert spaces V and H which together with V^*, the dual space of V, form a Gelfand triplet,

$$V \hookrightarrow H \cong H^* \hookrightarrow V^*. \tag{3.6}$$

We define the solution space $W^1(0,T;V,H)$ *by*

$$W^1(0,T;V,H) = \{u \in L^2(0,T;V) \,|\, \partial_t u \in L^2(0,T;V^*)\}, \tag{3.7}$$

wherein the time derivative $\partial_t u$ is meant in the distributional or weak sense of Definition 2.23.

Before we can state the notion of a weak solution, we introduce a notion of associating an operator $\mathcal{A}$ with a bilinear form.

Definition 3.5 (Bilinear forms with associated operators)
Let $(a_t)_{t\in[0,T]}$ be a family of bilinear forms $a : [0,T] \times V \times V \to \mathbb{R}$ that are measurable in t. We say that this family of bilinear forms, is associated with linear operators $\mathcal{A}_t : V \to V^*$ *if for almost all $t \in [0,T]$*

$$\langle \mathcal{A}_t u, v \rangle_{V^* \times V} = a_t(u,v) \tag{3.8}$$

holds $\forall u, v \in V$.

With these tools we are now able to introduce the notion of a weak solution to problem (3.1). Analogously to Definition 1 in Glau (2016) we give the following definition.

Definition 3.6 (Weak solution)
Let V, H be separable Hilbert spaces which together with V^*, the dual of V, form a Gelfand triplet,

$$V \hookrightarrow H \cong H^* \hookrightarrow V^*.$$

Let $f \in L^2(0,T;V^*)$ and $g \in H$. Then we call $u \in W^1(0,T;V,H)$ a weak solution *to problem* (3.1)*, if for almost every $t \in (0,T)$*

$$(\partial_t u(t), v)_H + a_t(u(t), v) = \langle f(t), v\rangle_{V^* \times V} \tag{3.9}$$

holds for all $v \in V$, where for each $t \in [0,T]$ a_t is the bilinear form associated with operator $\mathcal{A}_t$ and if additionally

$$\lim_{t \downarrow 0} \|g - u(t)\|_H = 0 \tag{3.10}$$

for t converging to zero from above holds as well. Then for every $v \in V$ and $\chi \in C_0^\infty([0,T])$ we have

$$-\int_0^T (u(t), v)_H \dot{\chi}(t)\,\mathrm{d}t + \int_0^T a_t(u,v)\chi(t)\,\mathrm{d}t = \int_0^T \langle f(t), v\rangle_{V^\star \times V}\chi(t)\,\mathrm{d}t, \tag{3.11}$$

which we state here for later reference.

Under certain conditions, unique weak solutions $u \in W^1(0,T;V,H)$ to partial differential equations exist. We cite the classic result from Wloka (2002).

Theorem 3.7 (Existence and uniqueness of weak solutions)
Let $0 < T < \infty$. Let $V \hookrightarrow H \hookrightarrow V^*$ be a Gelfand triplet with separable Hilbert spaces V and H over $\mathbb{R}$. Let $a : [0,T] \times V \times V \to \mathbb{R}$, $(t, \varphi, \psi) \mapsto a_t(\varphi, \psi)$ be a bilinear form that satisfies the following three conditions.

i) The mapping $(t, \varphi, \psi) \mapsto a_t(\varphi, \psi)$ is a measurable mapping on $[0,T]$ for fixed $\varphi, \psi \in V$.

ii) There exists a constant $\alpha > 0$ independent of t, such that

$$|a_t(\varphi, \psi)| \leq \alpha \|\varphi\|_V \|\psi\|_V, \qquad \forall t \in [0,T] \text{ and } \forall \varphi, \psi \in V. \tag{3.12}$$

iii) There exist constants $\beta > 0$ and $\lambda \geq 0$ independent of t such that

$$a_t(\varphi, \psi) \geq \beta \|\varphi\|_V^2 - \lambda \|\psi\|_H^2, \qquad \forall t \in [0,T] \text{ and } \forall \varphi, \psi \in V. \tag{3.13}$$

Further, let $(\mathcal{A}_t)_{t\in[0,T]}$ be defined via the relation (3.8)*. Then there exists a unique weak solution $u \in W^1(0,T;V,H)$ to the linear parabolic problem* (3.1)*.*
Additionally, the operator $\mathcal{L}$ relating the pair $(f,g) \in L^2(0,T;V^*) \times H$ to that unique weak solution $u \in W^1(0,T;V,H)$ of the linear parabolic problem (3.1) *is a linear and continuous mapping,*

$$\mathcal{L} : L^2(0,T;V^*) \times H \to W^1(0,T;V,H).$$

Proof
For a proof of the theorem we refer the reader to the proof of Theorem 26.1 in Wloka (2002). □

Remark 3.8 (On the existence and uniqueness result)
Actually, the existence and uniqueness result of Theorem 3.7 also holds under more general assumptions. In condition iii) for example, the bilinear form may map to $\mathbb{C}$ instead of $\mathbb{R}$. Then, the left side of the inequality is replaced by $\Re(a_t(\varphi,\psi))$. We decided to focus on the real-valued case, however, since it lays out the scope for option pricing purposes. In the error and convergence analysis section later, Conditions ii) and iii) will play a most prominent role.

Remark 3.9 (Existence and uniqueness result for Lévy models)
The claim of Theorem 3.7 comprises partial differential equations from many model classes. In Eberlein and Glau (2011), the authors translate the result to the class of Lévy models. To that extent they transform the assumptions of the theorem into requirements onto the characteristic triplet (b,σ,F) of the underlying process and even allow for time-dependence of that triplet. Theorem 5.3 in Eberlein and Glau (2011) then yields the claim of existence and uniqueness of weak solutions to problems of form (3.1) *in the Lévy model case.*

Theorem 3.10 (Feynman-Kac)
Let $(L_t)_{t\geq 0}$ be a (time-homogeneous) Lévy process. Consider the PIDE (3.1) *where $\mathcal{A}_t \equiv \mathcal{A}$ is assumed to be the operator associated with the symbol of $(L_t)_{t\geq 0}$ and $f \equiv 0$. Assume further the assumptions (A1)–(A3) of Eberlein and Glau (2011) to hold. Then* (3.1) *possesses a unique weak solution*

$$u \in W^1(0,T;H^{\alpha/2}_\eta(\mathbb{R}^d),L^2_\eta(\mathbb{R}^d)) \tag{3.14}$$

where $\alpha > 0$ is the Sobolev index of the symbol of $(L_t)_{t\geq 0}$ and $\eta \in \mathbb{R}^d$ is chosen according to Theorem 6.1 in Eberlein and Glau (2011). If additionally $g_\eta \in L^1(\mathbb{R}^d)$, then the relation

$$u(T-t,x) = \mathbb{E}\left[g(L_{T-t}+x)\right] \tag{3.15}$$

holds for all $t \in [0,T]$, $x \in \mathbb{R}^d$.

Proof
The result is proved in Eberlein and Glau (2011) and follows from Theorem 6.1 therein. Their claim applies beyond the scope of time-homogeneous Lévy processes and includes so-called time-inhomogeneous PIIAC processes, as well. □

The analysis of Feynmac-Kac theorems in the fashion of Theorem 3.10, which link stochastic quantities via their expected value to the solution of PIDEs, is a topic of its own. In the context of finance, where u is the price of an option with payoff profile g in an asset model driven by a stochastic process $(L_t)_{t\geq 0}$, this link opens a second access to the classic pricing problem. Either one solves the associated PIDE or one computes the expected value. Depending on the given model and option, the one or the other way might be better suited to determine the option price. For a thorough investigation of the Feynman-Kac formula we refer the reader to the recent publication of Glau (2016), where the result is derived for Lévy processes with discontinuous killing rate.

3.2 The Galerkin method

By now, we have introduced the core definitions and theorems of the classic theory of partial differential equations in an abstract framework. We now know that solutions to PDEs exist under certain conditions and we have introduced the spaces that they live in. For practical use of these solutions, however, for example for pricing or calibration purposes, we also need numerical representations of these solutions. In general, the solution spaces we have considered so far are infinite dimensional. Clearly, a numerical solution can not provide such richness. Instead, its numerical means are limited to finite dimensionality. We thus have to transform the original, infinite dimensional problem to a finite dimensional, approximative setting. We consider the pricing PDE

$$\begin{aligned} \partial_t u(t,x) + (\mathcal{A}u)(t,x) + ru(t,x) &= 0, \qquad && \forall (t,x) \in (0,T)\times\mathbb{R} \\ u(0,x) &= g(x), && \forall x \in \mathbb{R}. \end{aligned} \tag{3.16}$$

The time-homogeneous operator $\mathcal{A}$ carries the model information. We state the operators $\mathcal{A}$ in (3.16) for some well known time-homogeneous univariate asset models from the Lévy class. Since a Lévy model is identified by its characteristic triplet (b,σ,F), so is the operator $\mathcal{A}$ of the associated PIDE, which is in general given by

$$\begin{aligned} (\mathcal{A}f)(x) = &-b\partial_x f(x) - \frac{1}{2}\sigma^2\partial_{xx} f(x) \\ &- \int_{\mathbb{R}} \left(f(x+z) - f(x) - \partial_x f(x) h(z) \right) F(dz), \end{aligned} \tag{3.17}$$

for all $f \in C_0^\infty(\mathbb{R})$ and $x \in \mathbb{R}$, see for example Eberlein and Glau (2011). Here, we are only interested in the operator representation of each model. For a more detailed overview we refer the reader to Papapantoleon (2008). In the general Lévy model framework, the operator $\mathcal{A}$ as stated in (3.17) contains an integral term. The respective PDE is more precisely a partial *integro* differential equation, PIDE. The following examples offer an overview over the operators of some well known Lévy models.

Example 3.11 (Black&Scholes (BS) model)
In the Black&Scholes model of Black and Scholes (1973), the log-asset price process is modeled without jumps. The Brownian part drives the model exclusively. Therefore, we

have $F \equiv 0$. The operator of the Black&Scholes PDE thus reduces to setting

$$\sigma > 0, \qquad F \equiv 0 \tag{3.18}$$

in (3.17)*. The drift term b is set to*

$$b = r - \frac{1}{2}\sigma^2 \tag{3.19}$$

for martingale pricing.

Example 3.12 (Merton model)
The model of Merton (1976) enriches the Brownian part from the Black&Scholes model by a jump part. The log-asset prices process thus consists of a Brownian motion together with a compound Poisson process with independent normally $\mathcal{N}(\alpha, \beta^2)$ distributed jumps arriving at a rate $\lambda > 0$. From this, the characteristic triplet (b, σ, F) is derived as

$$\sigma > 0, \qquad F(\mathrm{d}z) = \frac{\lambda}{\sqrt{2\pi\beta^2}} \exp\left(-\frac{(z-\alpha)^2}{2\beta^2}\right) \mathrm{d}z, \tag{3.20}$$

with drift set to

$$b = r - \frac{1}{2}\sigma^2 - \lambda\left(e^{\alpha + \frac{\beta^2}{2}} - 1\right), \tag{3.21}$$

as required by the no-arbitrage condition.

Example 3.13 (CGMY model)
The CGMY model by Carr et al. (2002) is a so-called pure jump model. In contrast to the Merton model, jumps do not arrive discretely in time. Instead, in each finite time interval, infinitely many jumps occur. The model inherits its name from the parameterization

$$C > 0, \qquad G \geq 0, \qquad M \geq 0, \qquad Y \in (1, 2). \tag{3.22}$$

The characteristic triplet determining the operator $\mathcal{A}$ is given by

$$\sigma > 0, \qquad F(\mathrm{d}z) = C\frac{\exp(-Mz)}{z^{1+Y}} \mathbb{1}_{z>0}\, \mathrm{d}z + C\frac{\exp(Gz)}{|z|^{1+Y}} \mathbb{1}_{z<0}\, \mathrm{d}z, \tag{3.23}$$

with drift term b

$$b = r - \frac{1}{2}\sigma^2 - C\Gamma(-Y)\left[(M-1)^Y - M^Y + (G+1)^Y - G^Y\right] \tag{3.24}$$

by the no-arbitrage condition.

Example 3.14 (Univariate Normal Inverse Gaussian (NIG) model)
Finally, we present the NIG model by Barndorff-Nielsen (1997). With

$$\delta > 0, \qquad \alpha > 0, \qquad \beta \in \mathbb{R} \tag{3.25}$$

and the parameter condition $\alpha^2 > \beta^2$, the characteristic triplet is given by

$$\sigma > 0, \qquad F(\mathrm{d}z) = \exp(\beta z)\frac{\delta\alpha}{\pi|z|}K_1(\alpha|z|)\,\mathrm{d}z, \tag{3.26}$$

wherein K_1 denotes the Bessel function which for $z \in \mathbb{R}^+0$ allows the representation

$$K_1(z) = \int_0^\infty e^{-z\cosh(t)}\cosh(t)\,\mathrm{d}t, \tag{3.27}$$

see Chapter VI in Watson (1995). The drift term b is set to

$$b = r - \frac{1}{2}\sigma^2 - \delta\left(\sqrt{\alpha^2-\beta^2} - \sqrt{\alpha^2-(\beta+1)^2}\right) \tag{3.28}$$

to satisfy the no-arbitrage condition.

This reduction is achieved by the so-called Galerkin method that we introduce now. It consists of several steps that we discuss one by one. The identification of these steps that we present below is in major parts taken from Section VI.1 in Glau (2010) and inspired by Zeidler (1990). They lead from the general PIDE (3.16) to a numerically tractable approximative scheme that we consider in the next section with the Merton model as a specific example. The transition steps are the following.

i) **Modification to a problem with fast decaying solution.**
We will not solve problem (3.16) directly. One of the main obstacles that prevents an immediate numerical solution is the unbounded spacial domain of problem (3.16). This unbounded domain needs to be reduced to a bounded on. As a preparation for this localization, we modify problem (3.16) to a new problem which we know to possess a solution that rapidly decays to zero as $x \to \pm\infty$. This adjustment prepares step ii), where the motivation of this modification will be clarified. In order for the modification to result in a new the solution to which quickly decays to zero, we subtract a function ψ that we know to approximately mimic the behavior of u for large absolute values of $x \in \mathbb{R}$. The modification of this step i) thus consists in subtracting ψ from u and considering the resulting problem for $\phi = u - \psi$ given by

$$\begin{aligned} \partial_t\phi(t,x) + (\mathcal{A}\phi)(t,x) + r\phi(t,x) &= f(t,x), \qquad &&\forall(t,x) \in (0,T)\times\mathbb{R} \\ \phi(0,x) &= g_\Psi(x), &&\forall x \in \mathbb{R}, \end{aligned} \tag{3.29}$$

where $g_\Psi(x) = g(x) - \psi(0,x)$ for all $x \in \mathbb{R}$ and the right hand side f is given by

$$f(t,x) := -\left(\partial_t\psi(t,x) + (\mathcal{A}\psi)(t,x) + r\psi(t,x)\right).$$

The solution u to the original problem (3.16) can easily be restored by $u = \phi + \psi$. We establish the properties that ψ needs to provide, later, where we will present some examples, as well.

ii) **Localization to a boundary value problem.**
At first glance, the modification of the original problem (3.16) to the modified problem (3.29) complicated the derivation of numerical solution. Yet, now that we know ϕ to decay to zero for $|x| \to \infty$, we may cut the domain $\mathbb{R}$ to a finite interval (a, b) and assume the solution to the cut domain problem to be equal to zero outside of that interval. We denote the solution to the cut domain problem by $\overline{\phi}$. Instead of (3.29) we thus now and consider

$$\begin{aligned} \partial_t\overline{\phi}(t,x) + \left(\mathcal{A}\overline{\phi}\right)(t,x) + r\overline{\phi}(t,x) &= f(t,x), && \forall (t,x) \in (0,T) \times (a,b) \\ \overline{\phi}(t,a) = \overline{\phi}(t,b) &= 0, && \forall t \in (0,T), \\ \overline{\phi}(0,x) &= g_\phi(x), && \forall x \in (a,b), \end{aligned} \tag{3.30}$$

wherein $g_\phi = g_\Psi$ and where the right hand side remains unchanged.

iii) **Weak formulation of the resulting problem.**
Solution $\overline{\phi}$ to problem (3.30) still lives in the same function space as solution u to the original problem 3.16. We thus now cast problem (3.30) in an appropriate functions space setting which reflects our restriction of the infinite domain $\mathbb{R}$ to the finite domain of interest (a, b). Choosing an appropriate Gelfand triplet guarantees a weak solution $v \in W^1(0,T; V, H)$ to the localized problem (3.30)

$$\begin{aligned} \partial_t v + \mathcal{A}v + rv &= f, \\ v(0) &= g_\psi, \end{aligned} \tag{3.31}$$

where V and H build on the finite domain (a, b) and are assumed to be separable Hilbert spaces. The actual choices of V and H depend on the properties of the operator $\mathcal{A}$ and thus on the regularity that is required for a weak solution v to exist.

iv) **Variational formulation.**
We make the meaning of the weak formulation of problem (3.31) explicit. This step serves again as a preparatory step for the discretizations soon to follow. A function $v \in W^1(0,T; V, H)$ solves the weak problem (3.31) of step iii), if v satisfies the initial condition as a limit in H and if

$$\begin{aligned} -\int_0^T \langle v(t,\cdot), \varphi\rangle_H \partial_t\nu(t)\,\mathrm{d}t + \int_0^T a(v(t,\cdot),\varphi)\nu(t)\,\mathrm{d}t + r\int_0^T \langle v(t,\cdot),\varphi\rangle_H \nu(t)\,\mathrm{d}t \\ = -\Bigg(-\int_0^T \langle \psi(t,\cdot),\varphi\rangle_H \partial_t\nu(t)\,\mathrm{d}t \\ + \int_0^T a(\psi(t,\cdot),\varphi)\nu(t)\,\mathrm{d}t + r\int_0^T \langle\psi(t,\cdot),\varphi\rangle_H \nu(t)\,\mathrm{d}t \Bigg) \end{aligned} \tag{3.32}$$

for all $\nu \in C_0^\infty(0,T)$ that serve as test functions with respect to the time domain and for all $\varphi \in V$ that serve as test functions with respect to the spacial domain. In (3.32), the weak derivatives with respect to time have been transfered to the test function ν, the expression for the right hand side f has been resolved and $a(\cdot,\cdot)$ denotes the bilinear form associated with the operator $\mathcal{A}$ in the sense of Definition 3.5.

v) **Space discretization.**
In general, the solution $v \in W^1(0,T;V,H)$ to problem 3.31 lives in a Hilbert space of infinite dimension. Clearly, we will not be able to capture its infinite dimensionality numerically. Instead, we choose a sequence of finite dimensional Hilbert spaces V_n, $n \in \mathbb{N}$, with $V_n \subset V$ for all $n \in \mathbb{N}$ and reformulate problem (3.31) on these subspaces. A finite set of $n \in \mathbb{N}$ basis functions suffices to span each subspace V_n which thus renders numerical solutions schemes applicable. By assumption in step iii), the space V is separable.We choose a countable Riesz basis $\{\varphi_1, \varphi_2, \varphi_3, \dots\}$ of V. Since by virtue of the Gelfand triplet V is dense in H, there exists a sequence $(h_n)_{n\in\mathbb{N}}$ with

$$h_n \to g_\psi|_{(a,b)}$$

in H and $h_n \in V_n = \operatorname{span}\{\varphi_1^{(n)}, \varphi_2^{(n)}, \dots, \varphi_n^{(n)}\}$ for each $n \in \mathbb{N}$ where $\varphi_i^{(j)} \in \{\varphi_1, \varphi_2, \varphi_3, \dots\}$ for all $i \leq j \in \mathbb{N}$. The approximation $v_n \in W^1(0,T;V_n, H \cap V_n)$ of v and h_n are thus given by

$$v_n(t) := \sum_{k=1}^{n} V_k^{(n)}(t)\varphi_k^{(n)}, \qquad h_n = \sum_{k=1}^{n} \alpha_k^{(n)}\varphi_k^{(n)}. \tag{3.33}$$

By its definition in (3.33), for each $n \in \mathbb{N}$, v_n is given as a linear combination of basis functions $\varphi_k^{(n)}$, $k \in \{1,\dots,n\}$, of V_n. These basis functions are weighted by time dependent weights. Consequently we have for each $t \in (0,T)$ that $v_n(t) \in V_n$. Considering the consequences of this reduction in dimensionality we now face instead of finding v in (3.32) the new problem of finding $v_n \in W^1(0,T;V_n, H \cap V_n)$ such that

$$\begin{aligned}
-\int_0^T \langle v_n(t,\cdot),\varphi\rangle_H \partial_t\nu(t)\,\mathrm{d}t &+ \int_0^T a(v_n(t,\cdot),\varphi)\nu(t)\,\mathrm{d}t + r\int_0^T \langle v_n(t,\cdot),\varphi\rangle_H \nu(t)\,\mathrm{d}t \\
&= -\Big(-\int_0^T \langle \psi(t,\cdot),\varphi\rangle_H \partial_t\nu(t)\,\mathrm{d}t \\
&\qquad + \int_0^T a(\psi(t,\cdot),\varphi)\nu(t)\,\mathrm{d}t + r\int_0^T \langle \psi(t,\cdot),\varphi\rangle_H \nu(t)\,\mathrm{d}t \Big)
\end{aligned} \tag{3.34}$$

for all $\nu \in C_0^\infty(0,T)$ and for all test functions $\varphi \in V_n$. By assumption, the bilinear form $a(\cdot,\cdot)$ satisfies conditions ii) and iii) with respect to the space V. As a consequence, so does the bilinearform $a|_{V_n \times V_n}$ with respect to V_n. The classic Theorem 3.7 thus guarantees the existence and uniqueness of a weak solution v_n to the variational problem (3.34) for each $n \in \mathbb{N}$. Additionally, the sequence $(v_n)_{n\geq 1}$ converges to the solution v of the infinite dimensional original problem (3.32) in the sense that

$$v_n \to v \text{ in } L^2(0,T;V), \qquad \max_{0\le t\le T} \|v_n(t) - v(t)\|_H \to 0, \tag{3.35}$$

see Theorem 23.A and Remark 23.25 in Zeidler (1990).

vi) **Matrix formulation.**
For $t \in (0,T)$ we can represent each $v_n(t,\cdot)$ using the basis functions of V_n,

$$v_n(t,\cdot) = \sum_{k=1}^{n} V_k(t)\varphi_k^{(n)}, \tag{3.36}$$

with t dependent coefficients $V_k(t)$, $k \in \{1,\dots,n\}$. A matrix representation of (3.34) arises. Let $n \in \mathbb{N}$ arbitrary but fix. All operators in (3.34) are linear. Therefore, using only the basis function $\varphi_j^{(n)}$, $j \in \{1,\dots,n\}$, of V_n as test functions does not result in a loss of generality. It allows, however, transforming (3.34) into a matrix form. We get

$$\begin{aligned} \sum_{k=1}^{n} \partial_t V_k(t)\langle\varphi_k^{(n)},\varphi_j^{(n)}\rangle_H + \sum_{k=1}^{n} V_k(t) a(\varphi_k^{(n)},\varphi_j^{(n)}) & \\ +r\sum_{k=1}^{n} V_k(t)\langle\varphi_k^{(n)},\varphi_j^{(n)}\rangle_H = F_j(t),& \\ V_k(0) = \alpha_k, \qquad k \in \{1,\dots,n\},& \end{aligned} \tag{3.37}$$

with appropriately chosen α_k, $k \in \{1,\dots,n\}$, to approximate the initial condition and wherein for $j \in \{1,\dots,n\}$

$$F_j(t) = -\left(\langle\partial_t\psi(t,\cdot),\varphi_j^{(n)}\rangle + a(\psi(t,\cdot),\varphi_j^{(n)}) + r\langle\psi(t,\cdot),\varphi_j^{(n)}\rangle\right). \tag{3.38}$$

We rewrite (3.37) in matrix notation by

$$\begin{aligned} M\,\dot V(t) + A\,V(t) &= F(t), \qquad \text{for almost all } t \in [0,T], \\ V(0) &= \alpha, \end{aligned} \tag{3.39}$$

wherein $F(t) = (F_1(t),\dots,F_n(t))'$ and equivalently $\alpha = (\alpha_1,\dots,\alpha_n)'$ and the central matrices $M \in \mathbb{R}^{n\times n}$ and $A \in \mathbb{R}^{n\times n}$ are given by

$$M_{jk} = \langle\varphi_k^{(n)},\varphi_j^{(n)}\rangle_H, \qquad \forall 1 \le j,k \le n, \tag{3.40}$$

$$A_{jk} = a(\varphi_k^{(n)},\varphi_j^{(n)}) + r\langle\varphi_k^{(n)},\varphi_j^{(n)}\rangle_H, \qquad \forall 1 \le j,k \le n. \tag{3.41}$$

We call M the *mass matrix* and A the *stiffness matrix.* To solve problem (3.39) we thus now need to determine the time dependent vector

$$V(t) = (V_1(t),\dots,V_n(t))' \tag{3.42}$$

that satisfies the ODE therein.

vii) **Time discretization.**
We have reduced the dimensionality in space. Equally, we now discretize (3.39) with respect to time to receive a so-called fully discretized problem. To this end we choose $M \in \mathbb{N}$ and set up a time grid

$$0 = t_0 < t_1 < \cdots < t_M = T. \tag{3.43}$$

We introduce the notation $V^k := V(t^k)$, $k \in \{0, \dots, M\}$, and $\Delta t^k = t^{k+1} - t^k$, $k \in \{0, \dots, M-1\}$. We choose a $\theta \in [0, 1]$, approximate the time derivative by a finite difference approach and get from (3.39) the fully discrete scheme

$$\begin{aligned} M \, \frac{V^{k+1} - V^k}{\Delta t_k} + A \, V^{k+\theta} &= F^{k+\theta}, \qquad k \in \{0, \dots, M-1\} \\ V^0 &= \alpha, \end{aligned} \tag{3.44}$$

with

$$\begin{aligned} V^{k+\theta} &= \theta V^{k+1} + (1-\theta) V^k, \qquad k \in \{0, \dots, M-1\}, \\ F^{k+\theta} &= \theta F^{k+1} + (1-\theta) F^k, \qquad k \in \{0, \dots, M-1\}. \end{aligned}$$

Different values of $\theta \in [0, 1]$ result in variations in stability of the numerical procedures as we shall see later. Typically, we set $\theta = 1/2$, yielding the so-called Crank-Nicolson scheme.

The matrix-vector formulation in the fully discretized scheme links the solution V^k at time grid point t^k to the solution V^{k+1} at time grid point t^{k+1} The initial condition provides the values for $V^0 \in \mathbb{R}^n$. Thus, rewriting (3.44) and sorting by exponent we get the relation

$$(M + \Delta t^k \theta A) V^{k+1} = \Big(M - \Delta t^k (1-\theta) A \Big) \, V^k + F^{k+\theta},$$

for $k \in \{0, \dots, M-1\}$, which is equivalent to

$$V^{k+1} = (M + \Delta t^k \theta A)^{-1} \left(\Big(M - \Delta t^k (1-\theta) A \Big) \, V^k + F^{k+\theta} \right), \tag{3.45}$$

for $k \in \{0, \dots, M-1\}$. By iteratively applying (3.45), the solution to (3.44) on the whole space-time grid is derived.

Remark 3.15 (On the fully discrete solution)
Steps i)–vii) impose several layers of approximation on the original problem (3.16)*. Loosely speaking they first introduce a discretization in space, and a discretization in time, thereafter. When the PIDE is discretized in space, at the end of step v) we cite the convergence result* (3.35) *of Zeidler (1990) for the semi-discrete approximate solution that is still continuous in time. Convergence results for the fully discrete approximate solution V^k, $k \in \{0, \dots, N\}$, in* (3.45) *are provided in Section 3.6.*

Steps i) to vii) provide us with the theoretical background to set up a numerical Galerkin solver to solve pricing PIDEs of type (3.16). For an actual implementation of the method we need to decide on basis functions $w_k^{(n)}$, $k \in \{1, \dots, n\}$, spanning the solution spaces, a European payoff profile g and a pricing model represented by the PIDE operator $\mathcal{A}$. The core challenge then lies in calculating the key numerical ingredients, those being the mass matrix $M \in \mathbb{R}^{n \times n}$ as defined in (3.40), the stiffness matrix $A \in \mathbb{R}^{n \times n}$ as defined in (3.41), and the right hand side $F \in \mathbb{R}^n$ of (3.38). We consider the numerical difficulties arising from these quantities in the next section, taking the pricing problem of a European plain vanilla option in the Merton model as an example.

3.3 A FEM solver for the Merton model using hat functions

In this section, we build an actual Galerkin solver for pricing plain vanilla options in an elementary yet well known Lévy jump diffusion model. The computational steps that follow reflect the theoretical steps of the abstract framework of Section 3.2. We consider the Merton model as an example.

3.3.1 The model

We briefly stated the Merton jump-diffusion asset model of Merton (1976) in Fourier terms in Section 2.3.2. Throughout the rest of this chapter, it will serve as the example that the numerical PIDE solver being developed in this chapter will be based on. Let us therefore highlight its features in more detail. Consider a stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{Q})$. In the Merton model, the price process $(S_t)_{t \geq 0}$ of the underlying asset is modeled by

$$S_t = S_0 e^{L_t}, \tag{3.46}$$

with $S_0 = e^{x_0} > 0$ being today's value of the underlying, and wherein $(L_t)_{t \geq 0}$ is a Lévy jump diffusion process composed of a drift $b \in \mathbb{R}$, a Brownian part $\sigma > 0$ and a compound Poisson distributed jump part with jump intensity $\lambda > 0$ and Normally $\mathcal{N}(\alpha, \beta^2)$ distributed jump sizes,

$$L_t = bt + \sigma W_t + \sum_{i=1}^{N_t} X_i, \tag{3.47}$$

wherein $(W_t)_{t \geq 0}$ is a standard Brownian motion and $X_i \sim \mathcal{N}(\alpha, \beta^2)$, for all $i \in \mathbb{N}$. The Brownian motion $(W_t)_{t \geq 0}$, the Poisson process $(N_t)_{t \geq 0}$ and the normally distributed random variables are independent from another. From (3.47) we read off the the triplet (b, σ, F) that characterizes the model. The Lévy measure is given by

$$F(\mathrm{d}x) = \lambda \frac{1}{\sqrt{2\pi\beta^2}} \exp\left(-\frac{(x-\alpha)^2}{2\beta^2}\right) \mathrm{d}x. \tag{3.48}$$

In order to determine the drift value we consider the no-arbitrage condition. The process $(S_t)_{t\geq 0}$ defined in (3.46) discounted by the constant risk-free interest rate $r \geq 0$ must be a martingale under $\mathbb{Q}$, such that

$$e^{rt} = \mathbb{E}_{\mathbb{Q}}[e^{L_t}] = e^{t\theta(i(-i))}, \qquad \forall t \geq 0, \tag{3.49}$$

holds. By the definition of the cumulant generating function θ in (2.11) of Lemma 2.10, the identity (3.49) holds if

$$r = \theta(i(-i)) = b + \frac{1}{2}\sigma^2 + \int_{\mathbb{R}} (e^y - 1)F(\mathrm{d}y), \tag{3.50}$$

where we set the cut-off function to zero, $h \equiv 0$, by Remark 2.12. Note that we may choose the ξ argument of the cumulant generating function to be complex by Theorem 2.13. In accordance with the no-arbitrage condition stated in generality by identity (2.31), the drift b is thus set to

$$\begin{aligned} b &= r - \frac{\sigma^2}{2} - \int_{\mathbb{R}} (e^y - 1)F(\mathrm{d}y) \\ &= r - \frac{\sigma^2}{2} - \lambda\left(\exp\left(\frac{(\alpha+\beta^2)^2 - \alpha^2}{2\delta^2}\right) - 1\right) \\ &= r - \frac{\sigma^2}{2} - \lambda\left(\exp\left(\alpha + \frac{\beta^2}{2}\right) - 1\right), \end{aligned} \tag{3.51}$$

which completes the triplet (b, σ, F). Figure 2.2 displays a typical asset price trajectory $(S_0 \exp(L_t))_{t\geq 0}$ in the Merton model for $t \in [0, 1]$.

3.3.2 Pricing P(I)DE

The Merton model introduces the forward pricing PIDE

$$\begin{aligned} \partial_t u + \mathcal{A}u + ru &= 0 \quad \text{in } (0,T) \times \mathbb{R} \\ u(0) &= g \quad \text{in } \mathbb{R}, \end{aligned} \tag{3.52}$$

where by (3.17) the operator takes the form

$$(\mathcal{A}f)(x) = -b\partial_x f(x) - \frac{1}{2}\sigma^2 \partial_{xx} f(x) - \int_{\mathbb{R}} (f(x+y) - f(x))F(\mathrm{d}y) \tag{3.53}$$

for all $f \in C_0^\infty(\mathbb{R})$, with (b, σ, F) the characteristic triplet from above. In Section 3.1 we underlined, that the existence and uniqueness of (weak) solutions to PIDEs of form (3.52) depend on the choice of the solution space $W^1(0, T; V, H)$ yielded by the Hilbert spaces V and H that generate a Gelfand triplet together with V^*, the dual of V. In Section 3.2 we have taken several theoretical steps that demonstrated how to simplify a PIDE of form (3.52) to an approximate problem that is numerically tractable. Now we want to

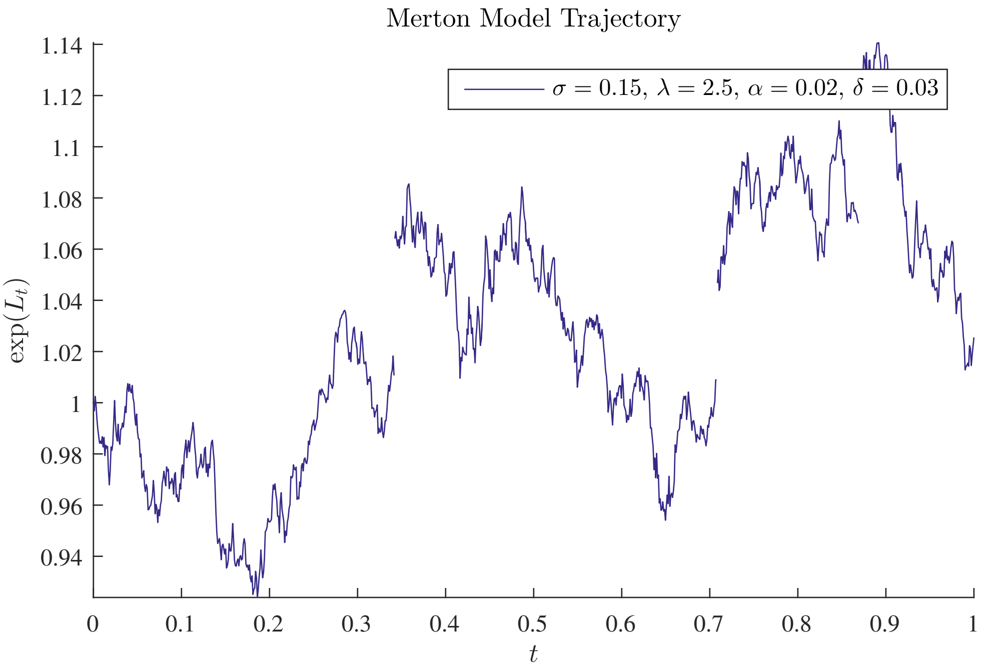

Figure 3.1 A single asset price trajectory in the Merton model. The Brownian component is parameterized by $\sigma = 0.15$. Jumps arrive at a rate of $\lambda = 2.5$ with expected value $\alpha = 0.02$ and standard deviation $\delta = 0.03$. The asset price process starts at $S_0 = 1$. The constant riskless interest rate is set to $r = 0.03$.

focus on the actual numerical implementation of the rather theoretical perspective of the previous two sections. We thus do not explicitly state the various solution spaces as we put those abstract simplification steps into concrete terms. As usual, however, we chose

$$H = L^2(\mathbb{R}), \qquad \text{or respectively} \qquad H = L^2(a, b). \tag{3.54}$$

The pricing of classic European options requires the notion of weighted Sobolev spaces to determine V and V^*, consider Definition 2.30. Weighted Sobolev spaces have also played a role in Theorem 3.10 where they were needed to link the solution to a PIDE of type (3.52) to an expected value via a Feynman-Kac approach. We thus emphasize, that weighted Sobolev spaces are crucial for the theoretical framework required for a unique weak solution to (3.52) to exist. Nevertheless, from here on we focus on implementational issues and thus try to avoid direct contact with the functional analysis in the background wherever possible. We recommend Eberlein and Glau (2011) for the proper treatment of the underlying spaces.

The numerical objects that we need in order to numerically approximate the weak solution to the Merton pricing PIDE (3.52) almost all depend on the bilinear form associated with the operator. The operator $\mathcal{A}$ of (3.53) yields a time-homogeneous bilinear form

$$\begin{aligned} a(\varphi, \psi) = & - b \int_{\mathbb{R}} (\partial_x \varphi(x))\, \psi(x)\, \mathrm{d}x \\ & - \frac{1}{2}\sigma^2 \int_{\mathbb{R}} (\partial_{xx} \varphi(x))\, \psi(x)\, \mathrm{d}x \\ & - \int_{\mathbb{R}} \left(\int_{\mathbb{R}} (\varphi(x+y) - \varphi(x)) F(\mathrm{d}y) \right) \psi(x)\, \mathrm{d}x \\ & + r \int_{\mathbb{R}} \varphi(x)\psi(x)\, \mathrm{d}x, \end{aligned} \tag{3.55}$$

defined for all $\varphi, \psi \in C_0^\infty(\mathbb{R})$. The bilinear form $a(\cdot, \cdot)$ of (3.55) is continuous as a mapping from $H_0^1(\mathbb{R}) \times H_0^1(\mathbb{R}) \to \mathbb{R}$. As such, it has a unique extension to an associated bilinear form $a : H_0^1(\mathbb{R}) \times H_0^1(\mathbb{R}) \to \mathbb{R}$ given by

$$\begin{aligned} a(\varphi, \psi) = & - b \int_{\mathbb{R}} (\partial_x \varphi(x))\, \psi(x)\, \mathrm{d}x \\ & + \frac{1}{2}\sigma^2 \int_{\mathbb{R}} (\partial_x \varphi(x))\, (\partial_x \psi(x))\, \mathrm{d}x \\ & - \int_{\mathbb{R}} \left(\int_{\mathbb{R}} (\varphi(x+y) - \varphi(x)) F(\mathrm{d}y) \right) \psi(x)\, \mathrm{d}x \\ & + r \int_{\mathbb{R}} \varphi(x)\psi(x)\, \mathrm{d}x, \end{aligned} \tag{3.56}$$

for all $\varphi, \psi \in H_0^1(\mathbb{R})$, where the transition from (3.55) to (3.56) is achieved by applying integration by parts in the first summand. We proceed with the bilinear form $a(\cdot, \cdot)$ of (3.56).

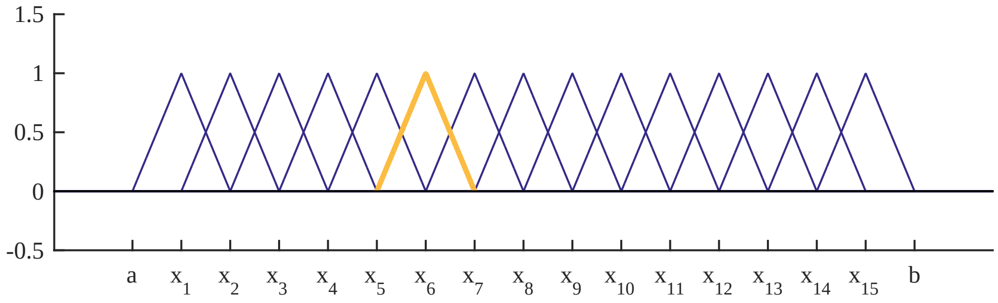

Figure 3.2 A plot of $N = 15$ hat functions φ_i, $i \in \{1, \dots, N\}$, spanning the bounded domain (a, b) as given by Definition 3.16 on an equidistant grid. For them to better distinguish, φ_6 is highlighted.

3.3.3 Basis functions: The hat functions

In accordance with step ii) of the abstract scheme in Section 3.2, we limit the unbounded spacial domain $\mathbb{R}$ of the Merton pricing PIDE problem to a bounded domain $(a, b) \subset \mathbb{R}$. On this bounded domain we establish a finite set of basis functions that span the finite dimensional space with respect to the spacial variable x. Key ingredients of a numerical PIDE solver depend heavily on the choice of basis functions. In this implementation we choose the well known hat functions as basis functions.

Definition 3.16 (FEM hat functions)
Let $N \in \mathbb{N}$ and $a < b \in \mathbb{R}$. Assume an equidistant grid $\Omega = \{x_0, x_1, \dots, x_N, x_{N+1}\}$ on (a, b) with mesh fineness $h > 0$,

$$a = x_0 < x_1 < \cdots < x_N < x_{N+1} = b, \tag{3.57}$$

with $x_i = a + ih$ for all $i \in \{1, \dots, N+1\}$, then the N hat functions φ_i, $i \in \{1, \dots, N\}$, are given by

$$\varphi_i(x) = \left(1 - \frac{|x - x_i|}{h}\right) \mathbb{1}_{|x - x_i| < h}, \qquad i \in \{1, \dots, N\}, \tag{3.58}$$

with derivative in the distributional or weak sense of Definition 2.23 given by

$$\frac{\partial}{\partial x}\varphi_i(x) = \begin{cases} h^{-1}, & x \in (x_i - h,\ x_i], \\ -h^{-1}, & x \in (x_i,\ x_i + h), \end{cases} \tag{3.59}$$

for all $i \in \{1, \dots, N\}$.

Clearly, the hat functions of Definition 3.16 are piecewise linear as Figure 3.2 illustrates. Later, we will also need the Fourier transform of the hat functions on an equidistant grid. This is provided by the following lemma.

Lemma 3.17 (Fourier transform of hat functions)
Assume $N \in \mathbb{N}$ and let φ_i, $i \in \{1, \dots, N\}$, be the hat functions on an equidistant grid $\{x_1, \dots, x_N\}$ with grid fineness $h > 0$ as introduced in Definition 3.16. Denote by φ_0 the hat function associated with the origin,

$$\varphi_0(x) = \left(1 - \frac{|x|}{h}\right) \mathbb{1}_{|x|<h}, \tag{3.60}$$

with appropriately scaled support, $\operatorname{supp} \varphi_0 \subset [-h, h]$. *Then, the characteristic function of hat function φ_j, $j \in \{1, \dots, N\}$, is given by*

$$\widehat{\varphi_j}(\xi) = e^{i\xi x_j} \widehat{\varphi_0}(\xi), \tag{3.61}$$

for all $\xi \in \mathbb{R}$, where

$$\widehat{\varphi_0}(\xi) = \frac{2}{\xi^2 h}(1 - \cos(\xi h)), \tag{3.62}$$

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

Proof
The derivation of the characteristic function of φ_0 is a straightforward calculation,

$$\begin{aligned}
\widehat{\varphi_0}(\xi) =& \int_{\mathbb{R}} e^{i\xi x} \varphi_0(x) \, \mathrm{d}x \\
=& \frac{1}{h} \int_{-h}^{0} (h + x) e^{i\xi x} \, \mathrm{d}x + \frac{1}{h} \int_{0}^{h} (h - x) e^{i\xi x} \, \mathrm{d}x \\
=& \frac{1}{h} \left(\int_0^h h e^{-i\xi x} + h e^{i\xi x} \, \mathrm{d}x + \int_0^h -x e^{-i\xi x} - x e^{i\xi x} \, \mathrm{d}x \right) \\
=& \frac{1}{h} \left(2h \int_0^h \cos(\xi x) \, \mathrm{d}x - 2 \int_0^h x \cos(\xi x) \, \mathrm{d}x \right) \\
=& \frac{2}{h} \left(h \left[\frac{1}{\xi} \sin(\xi x) \right]_0^h - \frac{1}{\xi^2} \left[\xi x \sin(\xi x) + \cos(\xi x) \right]_0^h \right) \\
=& \frac{2}{h} \left(\frac{h}{\xi} (\sin(\xi h) - 0) - \frac{1}{\xi^2} (\xi h \sin(\xi h) + \cos(\xi h) - (0 + 1)) \right) \\
=& \frac{2}{\xi} \sin(\xi h) - \frac{2}{\xi^2 h} (\xi h \sin(\xi h) + \cos(\xi h) - 1) \\
=& \frac{2}{\xi^2 h} (1 - \cos(\xi h)).
\end{aligned}$$

From this we deduce with $\varphi_j = \varphi_0(\cdot - x_j)$ and by property i) of Lemma 2.4, that the characteristic function of φ_j is given by

$$\widehat{\varphi_j}(\xi) = e^{i\xi x_j} \widehat{\varphi_0}(\xi),$$

for all $\xi \in \mathbb{R}$ and for all $j \in \{1, \dots, N\}$, which proves the lemma. □

3.3.4 Mass and stiffness matrix - an explicit derivation

As we have seen in Section 3.2, the key ingredients of a numerical solver are the mass matrix M and the stiffness matrix A. They drive the so-called time stepping scheme of (3.44) or (3.45), respectively, that iteratively derives the fully discrete solution on the space-time grid. Both matrices depend on the choice of basis functions φ_i, $i \in \{1, \ldots, N\}$, spanning the finite dimensional solution spaces built on V_N.

Lemma 3.18 (Mass matrix for hat functions)
Let $N \in \mathbb{N}$, and assume N hat functions φ_i, $i \in \{1, \ldots, N\}$, spanning a bounded domain given by an equidistantly spaced grid with mesh fineness $h > 0$. Then the mass matrix $M \in \mathbb{R}^{N \times N}$ given by

$$M_{ij} = \int_{\mathbb{R}} \varphi_j(x)\varphi_i(x) \, \mathrm{d}x, \qquad i, j \in \{1, \ldots, N\}, \tag{3.63}$$

computes to

$$M = \frac{h}{6} \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 \\ 1 & 4 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 4 & 1 \\ 0 & \cdots & 0 & 1 & 4 \end{pmatrix}, \tag{3.64}$$

with $M \in \mathbb{R}^{N \times N}$.

Proof
The entries of the mass matrix M are derived by elementary calculations. □

Mass matrix entries M_{ij} for $i, h \in \{1, \ldots, N\}$, as defined by (3.63) are only nonzero when the domains of the associated basis functions φ_i and φ_j overlap. Therefore, the mass matrix M of (3.64) is a sparse matrix when the underlying grid is populated by finitely supported hat functions and the degree of sparsity grows in N, as $M_{i,j} \neq 0$ if and only of $|i - j| \leq 1$.

The derivation of the stiffness matrix is a lot more involved. We recall the definition of the stiffness matrix in Equation (3.41) as

$$A_{ij} = a(\varphi_j, \varphi_i) + r\langle \varphi_j, \varphi_i \rangle_H, \qquad i, j \in \{1, \ldots, N\}.$$

We split up the stiffness matrix $A \in \mathbb{R}^{N \times N}$ into several parts that we compute individually,

$$A = A^{(1)} + A^{(2)} + A^{(3)} + A^{(4)} \tag{3.65}$$

where

$$A_{ij}^{(1)} = -b\int_{\mathbb{R}} \left(\frac{\partial}{\partial x}\varphi_j(x)\right)\varphi_i(x)\,\mathrm{d}x \tag{3.66}$$

$$A_{ij}^{(2)} = \frac{1}{2}\sigma^2\int_{\mathbb{R}} \frac{\partial}{\partial x}\varphi_j(x)\frac{\partial}{\partial x}\varphi_i(x)\,\mathrm{d}x \tag{3.67}$$

$$A_{ij}^{(3)} = -\int_{\mathbb{R}}\int_{\mathbb{R}} (\varphi_j(x+y)-\varphi_j(x))F(\mathrm{d}y)\varphi_i(x)\,\mathrm{d}x \tag{3.68}$$

$$A_{ij}^{(4)} = rM_{ij} \tag{3.69}$$

for $i,j \in \{1,\dots,N\}$, where we implicitly use $H = L^2(a,b)$ as set in (3.54). The stiffness matrix carries the information describing the behavior of the underlying asset price process as represented by the characteristic triplet (b,σ,F). Especially the existence of a Lévy measure F carrying jump information, $F(\mathrm{d}y) \neq 0$, in general complicates the derivation of (semi-)explicit formulas of the stiffness matrix considerably. Section 3.2, where we stated the Lévy measures F of some well known models, underlines the challenge of numerical integration with respect to Lévy measures.

Yet, for the Merton model we will derive (semi-)explicit formulas for the stiffness matrix entries, including the jump part $A_{ij}^{(3)}$ in (3.68). We will analytically solve the integrals in $A_{ij}^{(k)}$, $k \in \{1,2,3,4\}$, until explicit formulas are derived or until the expressions depend on integrals with respect to the Lévy measure F that the following Lemma can solve.

Lemma 3.19 (Important integrals with respect to $F(\mathrm{d}y)$ in the Merton model) *Let $F(\mathrm{d}y)$ be the Lévy measure of the Merton model,*

$$F(\mathrm{d}y) = \lambda\frac{1}{\sqrt{2\pi}\beta}\exp\left(-\frac{(y-\alpha)^2}{2\beta^2}\right)\mathrm{d}y \tag{3.70}$$

with $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^+$ and $\lambda \in \mathbb{R}^+$. Then we have the following identities,

$$\int\limits_{-\infty}^{x} F(\mathrm{d}y) = \frac{1}{2}\lambda\,\mathrm{erf}\left(\frac{x-\alpha}{\sqrt{2}\beta}\right). \tag{3.71}$$

Let $c \in \mathbb{R}$, then

$$\int\limits_{-\infty}^{x} (y-c)F(\mathrm{d}y) = \frac{\lambda}{\sqrt{2\pi}\beta}\left(\beta^2\left(-e^{-\frac{(x-\alpha)^2}{2\beta^2}}\right) - \sqrt{\frac{\pi}{2}}\beta(c-\alpha)\,\mathrm{erf}\left(\frac{x-\alpha}{\sqrt{2}\beta}\right)\right), \tag{3.72}$$

and

$$\int\limits_{-\infty}^{x} (y-c)^2F(\mathrm{d}y) = \frac{\lambda}{2\sqrt{2\pi}}\Bigg(\sqrt{2\pi}\left((\alpha-c)^2+\beta^2\right)\mathrm{erf}\left(\frac{x-\alpha}{\sqrt{2}\beta}\right) - 2\beta(\alpha-2c+x)e^{-\frac{(\alpha-y)^2}{2\beta^2}}\Bigg), \tag{3.73}$$

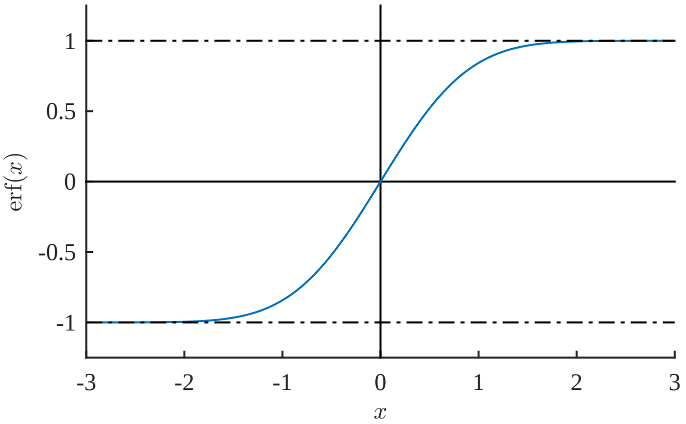

Figure 3.3 The erf function as given by (3.76) in Lemma 3.19, evaluated over $x \in [-3, 3]$.

further

$$\int\limits_{-\infty}^{x} (y-c)^3 F(\mathrm{d}y) = -\frac{\lambda}{2\sqrt{2\pi}} \left(\sqrt{2\pi}(c-\alpha)((\alpha-c)^2 + 3\beta^2) \operatorname{erf}\left(\frac{x-\alpha}{\sqrt{2}\beta}\right) \right.$$
$$\left. + 2\beta e^{-\frac{(\alpha-x)^2}{2\beta^2}} (\alpha^2 + 3c^2 - 3c(\alpha+x) + 2\beta^2 + x^2 + \alpha x) \right), \quad (3.74)$$

and finally

$$\int\limits_{-\infty}^{x} e^y F(\mathrm{d}y) = -\frac{\lambda}{2} e^{\alpha + \frac{\beta^2}{2}} \operatorname{erf}\left(\frac{\alpha + \beta^2 - x}{\sqrt{2}\beta}\right). \quad (3.75)$$

In all identities (3.71)–(3.75), erf *denotes the so called error function,*

$$\operatorname{erf} : \mathbb{R} \to (-1, 1),$$

defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \mathrm{d}t, \qquad \forall x \in \mathbb{R}. \quad (3.76)$$

Proof
All integrals have been solved using `http://www.wolframalpha.com` and performing elementary transformations on the results. □

A plot of the erf function is depicted in Figure 3.3.

Remark 3.20 (erf and normal distribution)
The relation between the erf *function and the cumulative distribution function of the standard normal distribution is obvious,*

$$\begin{aligned}\operatorname{erf}(x) &= \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{d}t = \frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{2}x} e^{-\frac{t^2}{2}}\,\mathrm{d}t\\ &= 2\left[\Phi_{0,1}\left(\sqrt{2}x\right) - \Phi_{0,1}(0)\right]\\ &= 2\Phi_{0,1}\left(\sqrt{2}x\right) - 1,\end{aligned}$$

where Φ_{μ,σ^2} denotes the cumulative distribution function of the normal distribution with expected value $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}^+$.

We begin the derivation of the individual parts $A^{(k)}$, $k \in \{1,2,3,4\}$, of the stiffness matrix A of (3.65).

($A_{ij}^{(1)}$) Elementary calculations result in

$$A_{ij}^{(1)} = \frac{1}{2}\,b\begin{cases} -1, & j-i=1,\\ 1, & j-i=-1,\end{cases}$$

yielding the matrix

$$A^{(1)} = \frac{1}{2}\,b\begin{pmatrix} 0 & -1 & 0 & \cdots & 0\\ 1 & 0 & -1 & \ddots & \vdots\\ 0 & \ddots & \ddots & \ddots & 0\\ \vdots & \ddots & 1 & 0 & -1\\ 0 & \cdots & 0 & 1 & 0\end{pmatrix},$$

with b as in (3.51).

($A_{ij}^{(2)}$) Elementary calculations result in

$$A_{ij}^{(2)} = \frac{1}{2h}\sigma^2\begin{cases} 2, & |i-j|=0,\\ -1, & |i-j|=1,\end{cases}$$

yielding the matrix

$$A^{(2)} = \frac{1}{2h}\sigma^2\begin{pmatrix} 2 & -1 & 0 & \cdots & 0\\ -1 & 2 & -1 & \ddots & \vdots\\ 0 & \ddots & \ddots & \ddots & 0\\ \vdots & \ddots & -1 & 2 & -1\\ 0 & \cdots & 0 & -1 & 2\end{pmatrix}. \tag{3.77}$$

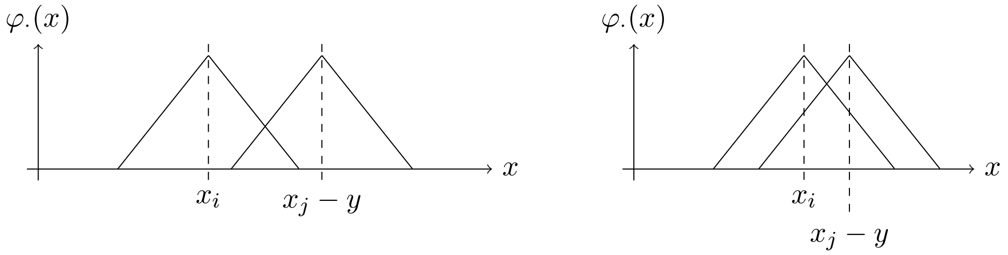

Figure 3.4 Two types of overlap have to be distinguished during the derivation of $A_{ij}^{(3)}$ in (3.78). On the left, y is such that $h < |x_i - (x_j - y)| < 2h$, while on the right, y is such that $|x_i - (x_j - y)| < h$. Whenever $|x_i - (x_j - y)| > 2h$, there is no overlap and the respective integral (3.80) is equal to zero.

($A_{ij}^{(3)}$) We compute $A^{(3)}$ of (3.68).

$$\begin{aligned} A_{ij}^{(3)} &= -\int_{\mathbb{R}}\int_{\mathbb{R}} (\varphi_j(x+y) - \varphi_j(x))\, F(\mathrm{d}y)\varphi_i(x)\,\mathrm{d}x \\ &= -\int_{\mathbb{R}}\int_{\mathbb{R}} (\varphi_j(x+y) - \varphi_j(x))\, \varphi_i(x)\,\mathrm{d}x\, F(\mathrm{d}y) \\ &= -\int_{\mathbb{R}}\left[\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x\right] F(\mathrm{d}y) + \int_{\mathbb{R}}\left[\int_{\mathbb{R}} \varphi_j(x)\varphi_i(x)\,\mathrm{d}x\right] F(\mathrm{d}y) \\ &= -\int_{\mathbb{R}}\left[\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x\right] F(\mathrm{d}y) + \lambda M_{ij}. \end{aligned} \tag{3.78}$$

Let now $y \in \mathbb{R}$ and $i, j \in \{1, \ldots, N\}$ fix such that

$$y \le x_j - x_i \quad \Leftrightarrow \quad x_i \le x_j - y \tag{3.79}$$

and consider $\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x$. Then, by the definition of the hat functions φ_i, $i \in \{1, \ldots, N\}$, in Definition 3.16 we have

$$\begin{aligned} &\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}} \left(1 - \frac{|x - (x_j - y)|}{h}\right) \mathbb{1}_{|x-(x_j-y)|<h} \left(1 - \frac{|x - x_i|}{h}\right) \mathbb{1}_{|x-x_i|<h}\,\mathrm{d}x. \end{aligned} \tag{3.80}$$

The integral (3.80) is nonzero only if the two functions in the integrand overlap, which is the case if $|x_i - (x_j - y)| < 2h$. Then, two different kinds of overlapping have to be distinguished, see Figure 3.4.

Define $d = x_j - x_i$. Then, by some tedious but elementary calculations, (3.80) computes to

$$\begin{aligned}\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x = \mathbb{1}_{h<|y-d|\le 2h} &\left[\frac{(2h-|y-d|)^3}{6h^2}\right] \\ + \mathbb{1}_{|y-d|\le h} &\left[\frac{1}{2h^2}|y-d|^3 - \frac{1}{h}|y-d|^2 + \frac{2}{3}h\right].\end{aligned} \tag{3.81}$$

Concerning the indicator functions in (3.81) we find that

$$\begin{aligned}\mathbb{1}_{h<|y-d|\le 2h} &= 1 \Leftrightarrow y \in (d+h,\ d+2h] \cup [d-2h,\ d-h), \\ \mathbb{1}_{|y-d|\le h} &= 1 \Leftrightarrow y \in [d-h,\ d+h].\end{aligned} \tag{3.82}$$

We use the intervals of (3.82), where the indicator functions in (3.81) are nonzero, to integrate $\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x$ with respect to the Lévy measure ν of the Merton model. Until now, all derivations have been conducted independently of the model represented by the Lévy measure. At this point, the derivations depend on the model. We derive

$$\begin{aligned}\int_{\mathbb{R}}\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x\,F(\mathrm{d}y) = &\int_{\mathbb{R}} \mathbb{1}_{h<|y-d|\le 2h}\left[\frac{(2h-|y-d|)^3}{6h^2}\right]F(\mathrm{d}y) \\ + &\int_{\mathbb{R}} \mathbb{1}_{|y-d|\le h}\left[\frac{1}{2h^2}|y-d|^3 - \frac{1}{h}|y-d|^2 + \frac{2}{3}h\right]F(\mathrm{d}y).\end{aligned} \tag{3.83}$$

We integrate both summands in (3.83) separately. For the first we find

$$\begin{aligned}&\int_{\mathbb{R}} \mathbb{1}_{h<|y-d|\le 2h}\left[\frac{(2h-|y-d|)^3}{6h^2}\right]F(\mathrm{d}y) \\ &= \frac{1}{6h^2}\left(\int_{d-2h}^{d-h}(2h+y-d)^3F(\mathrm{d}y) + \int_{d+h}^{d+2h}(2h-(y-d))^3F(\mathrm{d}y)\right) \\ &= \frac{1}{6h^2}\left(\int_{d-2h}^{d-h}(y-(d-2h))^3F(\mathrm{d}y) - \int_{d+h}^{d+2h}(y-(d+2h))^3F(\mathrm{d}y)\right)\end{aligned} \tag{3.84}$$

For the integration of the second summand in (3.81) we have

$$\begin{aligned}&\int_{\mathbb{R}} \mathbb{1}_{|y-d|\le h}\left[\frac{1}{2h^2}|y-d|^3 - \frac{1}{h}|y-d|^2 + \frac{2}{3}h\right]F(\mathrm{d}y) \\ &= \frac{1}{2h^2}\int_{d-h}^{d+h}|y-d|^3F(\mathrm{d}y) - \frac{1}{h}\int_{d-h}^{d+h}|y-d|^2F(\mathrm{d}y) + \frac{2}{3}h\int_{d-h}^{d+h}F(\mathrm{d}y) \\ &= \frac{1}{2h^2}\left(\int_{d}^{d+h}(y-d)^3F(\mathrm{d}y) - \int_{d-h}^{d}(y-d)^3F(\mathrm{d}y)\right) \\ &\qquad - \frac{1}{h}\int_{d-h}^{d+h}(y-d)^2F(\mathrm{d}y) + \frac{2}{3}h\int_{d-h}^{d+h}F(\mathrm{d}y)\end{aligned} \tag{3.85}$$

All the individual integral values in (3.84) and (3.85) are now provided by Lemma 3.19.

This finishes the derivation of $\int_{\mathbb{R}}\int_{\mathbb{R}} \varphi_j(x+y)\varphi_i(x)\,\mathrm{d}x\,F(\mathrm{d}y)$ in (3.78) and thereby also the computation of the third part of the stiffness matrix A as given by (3.68).

($A^{(4)}_{ij}$) For the final part of the stiffness matrix, $A^{(4)}$ there is nothing left to do. By definition,

$$A^{(4)}_{ij} = rM_{ij},$$

so the forth part of the stiffness matrix A is given by the definition of the mass matrix in (3.63).

3.3.5 The right hand side F - a Fourier approach

As the goal of this implementation is to derive prices of plain vanilla European call and put options, the solution to pricing PIDE (3.52) will not possess zero boundaries. Linear combinations of classic hat functions, however, can only represent functions with zero boundaries (where we will not pursue the concept of *special* hat functions that are basically half hats associated with node x_0 or x_{N+1}, respectively, that circumvent this restriction. We direct the reader to Chapter 5.2 in Seydel (2012) for these special basis functions, instead). Consequently, the original Merton pricing PIDE needs to be transformed to a new problem which we assume to be equal to zero at the boundaries of the bounded domain (a, b). We have seen the theoretical concept of the enforcement of Dirichlet zero-boundaries in Section 3.2 and in Steps i) and ii) therein. For the numerical implementation, we need to decide on a specific function ψ to subtract from the solution to the original problem. The choice of this function depends first and foremost on the payoff profile of the option that we derive prices for.

For plain vanilla European call and put options, there are standard boundary conditions in the literature see Example 15.5 in Hull (2015). These are inherited from the price value V^C of a call option and the price value V^P of a put option that behave for $|x| \to \infty$ and $|x| \to 0$ as

$$\begin{aligned} &V^C(x,t) \to 0, && x \to -\infty,\ t \in [0,T] \\ &V^C(x,t) \to e^x - Ke^{-rt}, && x \to +\infty,\ t \in [0,T] \end{aligned} \tag{3.86}$$

for call options and

$$\begin{aligned} &V^P(x,t) \to Ke^{-rt} - e^x, && x \to -\infty,\ t \in [0,T] \\ &V^P(x,t) \to 0, && x \to +\infty,\ t \in [0,T] \end{aligned} \tag{3.87}$$

for put options. In Figure 3.5 we assess the accuracy of these boundary conditions in the Black&Scholes model. For the localization of the pricing PIDE (3.52) to a bounded space-time region $(0,T) \times (a,b)$ with $a \ll \log(K) \ll b \in \mathbb{R}$ and $T > 0$, a function ψ,

$$\psi : [0,T] \times [a,b] \to \infty, \tag{3.88}$$

to subtract would need to fulfill

$$\begin{aligned} &\psi^C(t,a) = 0, && \forall t \in [0,T], \\ &\psi^C(t,b) = e^b - Ke^{-rt}, && \forall t \in [0,T], \end{aligned} \tag{3.89}$$

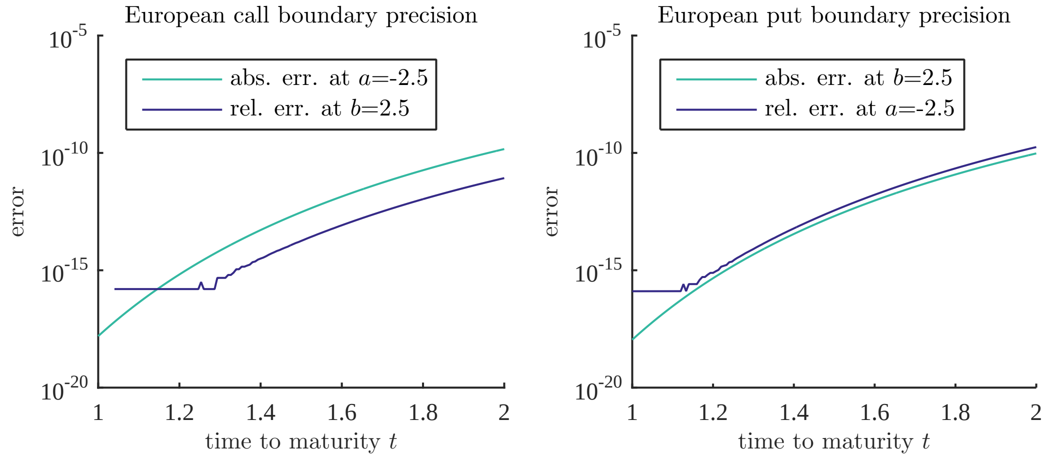

Figure 3.5 Precision study of the classic boundary conditions for European call (left) and put options (right). We compare ψ defined according to (3.89) or (3.90), respectively, to prices of the Black&Scholes model generated by Matlab's `blsprice` routine. We set $r = 0.05$, $K = 1$, $\sigma = 0.3$ and evaluate European call and put prices for $S_0^{\max} = e^b$ with $b = 2.5$ and $S_0^{\min} = e^a$ with $a = -2.5$ for time to maturity values of $t \in [1, 2]$. With values for r and σ being rather large and $|b|$ and $|a|$ being rather small, both model as well as grid parameters have been chosen rather conservatively. Results in more realistic settings are even better than the depicted ones.

for call options and

$$\begin{aligned} \psi^P(t,a) &= Ke^{-rt} - e^a, && \forall t \in [0,T], \\ \psi^P(t,b) &= 0, && \forall t \in [0,T], \end{aligned} \tag{3.90}$$

for put options. Naive choices for both European options are

$$\begin{aligned} \widetilde{\psi}^C(t,x) &= \left(e^x - Ke^{-rt}\right)^+, \\ \widetilde{\psi}^P(t,x) &= \left(Ke^{-rt} - e^x\right)^+. \end{aligned} \tag{3.91}$$

Both candidates in (3.91) fulfill the boundary conditions (3.89) and (3.90), respectively.

However, we do not want to repeat tedious calculations of the kind we encountered in the derivation of semi-explicit expressions for entries of the stiffness matrix. Instead, we intend to apply a Fourier approach and compute the entries of the right hand side $F \in \mathbb{R}^N$ numerically. As we shall see below, for the application of this approach we need not only a closed expression of the function ψ which we subtract from the original problem, but additionally a closed expression of its Fourier transform $\widehat{\psi}$. For better numerical tractability, we require a fast decay of $|\widehat{\psi}(\xi)|$ for $|\xi| \to \infty$. The smoother ψ, the faster $|\widehat{\psi}|$ decays, compare Remark 2.8. Consequently, due to the kink at $x = \log(Ke^{-rt})$ for all $t \in [0,T]$, both $\widetilde{\psi}^C(\cdot,t)$ and $\widetilde{\psi}^P(\cdot,t)$ are only continuous, but not continuously differentiable and thus already lack elementary smoothness. We thus need different functions ψ to subtract that not only fulfill the appropriate boundary conditions (3.89) or (3.90) but that are also as smooth as possible. This additional requirement rules out naive candidates like (3.91). As we have seen in (3.38), the right hand side in vector notation is given by $F(t^k) = (F_1(t^k), \dots, F_N(t^k)) \in \mathbb{R}^N$ for each t^k on the time grid with $F_j(\cdot)$, $j \in \{1, \dots, N\}$, given by

$$F_j = -\int_{\mathbb{R}} \left(\partial_t \psi(t,x) + (\mathcal{A}\psi)(t,x) + r\psi(t,x)\right) \varphi_j(x)\,\mathrm{d}x \tag{3.92}$$

for all $j \in \{1, \dots, N\}$.

In contrast to the integrals in the stiffness matrix, we intend to avoid solving the integral in (3.92) analogously to derive the right hand side F. Instead, we follow Eberlein and Glau (2011) by invoking Parseval's identity of Theorem 2.7 in a way that we call the symbol method.

Lemma 3.21 (The symbol method)
Let A be the symbol of a Lévy process given by the characteristic triplet (b, σ, F). Denote by $\mathcal{A} : C_0^\infty(\mathbb{R}^d, \mathbb{C}) \to C^\infty(\mathbb{R}^d, \mathbb{C})$ the pseudodifferential operator associated with symbol A. Furthermore, denote by $a : C_0^\infty \times C_0^\infty \to \mathbb{C}$ the bilinear form associated with the operator $\mathcal{A}$. Let $\eta \in \mathbb{R}^d$. If

i) the exponential moment condition

$$\int_{|x|>1} e^{-\langle \eta', x\rangle} F(\mathrm{d}x) < \infty \tag{3.93}$$

holds for all $\eta' \in \operatorname{sgn}(\eta^1)[0, |\eta^1|] \times \cdots \times \operatorname{sgn}(\eta^d)[0, |\eta^d|]$ and

ii) there exists a constant $C_1 > 0$ with

$$|A(z)| \le C_1(1 + \|z\|)^\alpha \tag{3.94}$$

for all $z \in U_{-\eta}$ where

$$U_{-\eta} = U_{-\eta^1} \times \cdots \times U_{-\eta^d} \tag{3.95}$$

with $U_{-\eta^j} = \mathbb{R} - i\operatorname{sgn}(\eta^j)[0, |\eta^j|)$,

then $a(\cdot,\cdot)$ possesses a unique linear extension $a : H_\eta^{\alpha/2} \times H_\eta^{\alpha/2} \to \mathbb{C}$ which can be written as

$$a(\varphi, \psi) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} A(\xi - i\eta)\widehat{\varphi}(\xi - i\eta)\overline{\widehat{\psi}(\xi - i\eta)}\,\mathrm{d}\xi \tag{3.96}$$

for all $\varphi, \psi \in H_\eta^{\alpha/2}(\mathbb{R}^d)$.

Proof
The proof can be found in Eberlein and Glau (2011) using Theorem 4.1 therein and Parseval's identity 2.7. □

Lemma 3.21 enables us to avoid considering the operator $\mathcal{A}$ for evaluating the associated bilinear form and use the belonging symbol A, instead. Let us observe the effect of Lemma 3.21 on the derivation of the right hand side F_j, $j \in \{1, \ldots, N\}$ for our numerical FEM solver for the Merton model.

With this identity, we are able to derive the right hand side $(F_j)_{j\in\{1,\ldots,N\}}$ in terms of Fourier transforms. Consider a smooth function $\psi : [0,T] \times \mathbb{R} \to \mathbb{R}$ such that $\psi(t) \in H^{\alpha/2}_{\eta}(\mathbb{R})$ for all $t \in [0,T]$ for some $\eta \in \mathbb{R}$. With $t \in [0,T]$,

$$\begin{aligned} F_j(t) = &- \int_{\mathbb{R}} \left(\partial_t \psi(t,x) + (\mathcal{A}\psi)(t,x) + r\psi(t,x)\right) \varphi_j(x)\,\mathrm{d}x \\ = &- \left(\int_{\mathbb{R}} \partial_t \psi(t,x)\varphi_j(x)\,\mathrm{d}x + \int_{\mathbb{R}} (\mathcal{A}\psi)(t,x)\varphi_j(x)\,\mathrm{d}x + r\int_{\mathbb{R}} \psi(t,x)\varphi_j(x)\,\mathrm{d}x \right). \end{aligned} \tag{3.97}$$

We consider the three parts in (3.97) individually. In the last summand we use appropriate dampening to apply Parseval's identity of Theorem 2.7 and get

$$\begin{aligned} \int_{\mathbb{R}} \psi(t,x)\varphi_j(x)\,\mathrm{d}x &= \int_{\mathbb{R}} e^{\eta x}\psi(t,x)e^{-\eta x}\varphi_j(x)\,\mathrm{d}x \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \mathcal{F}(e^{\eta\cdot}\psi(\cdot,t))(\xi)\overline{\mathcal{F}(e^{-\eta\cdot}\varphi_j(\cdot,t))(\xi)}\,\mathrm{d}\xi \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{\psi_\eta}(\cdot,t)(\xi)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi. \end{aligned} \tag{3.98}$$

By the same means we get for the first summand in (3.97) that

$$\int_{\mathbb{R}} \partial_t\psi(t,x)\varphi_j(x)\,\mathrm{d}x = \frac{1}{2\pi}\int_{\mathbb{R}} \partial_t\widehat{\psi_\eta}(\cdot,t)(\xi)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi. \tag{3.99}$$

Finally, for the second summand we have by applying the symbol method of Lemma 3.21 to (3.92) that

$$\int_{\mathbb{R}} (\mathcal{A}\psi)(t,x)\varphi_j(x)\,\mathrm{d}x = \frac{1}{2\pi}\int_{\mathbb{R}} A(\xi - i\eta)\widehat{\psi(\cdot,t)}(\xi - i\eta)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi, \tag{3.100}$$

where A denotes the symbol of the Merton model.

Example 3.22 (Symbol in the Merton model)
In the Merton model where $\sigma > 0$, $\lambda > 0$, $\alpha \in \mathbb{R}$ and $\beta > 0$, the symbol computes to

$$A(\xi) = A^{merton}(\xi) = \frac{1}{2}\sigma^2\xi^2 + i\xi b - \lambda\left(e^{-i\alpha\xi - \frac{1}{2}\beta^2\xi^2} - 1\right) \tag{3.101}$$

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

We see from Example 3.22 that the symbol of the Merton model appears to be numerically accessible. Consequently, the symbol A for the Merton model is very suitable for numerical integration, as is $\widehat{\psi}$ given that the function ψ itself is smooth enough. In this case, solving the integral in (3.100) is numerically accessible – in stark contrast to the respective integral in (3.97).

The following remark summarizes the numerical requirements on ψ.

Remark 3.23 (Empirical criteria for ψ)
Consider a pricing PIDE (3.52) *for a European plain vanilla option with payoff profile g with weak solution $u \in W^1(0,T;H_\eta^{\alpha/2}(\mathbb{R}),L_\eta^2(\mathbb{R}))$ for some weight $\eta \in \mathbb{R}$ that shall be numerically approximated on a space time grid in $[a,b]\times[0,T]$. Assume $\psi \in H_\eta^{\alpha/2}(\mathbb{R})$ that (approximately) matches the boundary conditions on the boundaries of the space-time grid i.e. for the call option* (3.89) *and for the put option* (3.90). *Then ψ is numerically suitable for the purpose of localizing the pricing PIDE* (3.52) *if*

i) ψ is quickly evaluable on the region $[a,b]\times[0,T]$ and

ii) the integral

$$F_j = -\frac{1}{2\pi}\int_{\mathbb{R}}\Big(\widehat{\partial_t\psi(\cdot,t)}(\xi-i\eta,t)+A(\xi-i\eta)\widehat{\psi(\cdot,t)}(\xi-i\eta)+r\widehat{\psi}(\xi-i\eta,t)\Big)\overline{\widehat{\varphi_{j_{-\eta}}}(\xi)}\,\mathrm{d}\xi$$

can be numerically evaluated for all $j\in\{1,\dots,N\}$.

Criterium i) allows retransforming the solution of the localized problem into the solution of the original pricing PIDE, while criterium ii) grants the numerical derivation of the right hand side $F\in\mathbb{R}^N$.

In the following two subsections we will analyze two candidates for ψ that match the criteria of Remark 3.23.

A first suggestion for ψ consists in using Black&Scholes prices as functions in $x = \log(S_0)\in[a,b]$ and time to maturity $t\in[0,T]$ for localization of the pricing PIDE (3.52). We express the price of a European option with payoff profile f_K in the Black&Scholes model in terms of (generalized) Fourier transforms using Proposition 2.20 and define ψ accordingly, as the following lemma explains.

Lemma 3.24 (Subtracting Black&Scholes prices)
Let $\eta\in\mathbb{R}$ such that Conditions (Exp) and (Int) of Proposition 2.20 are satisfied. Choose a Black&Scholes volatility $\sigma^2>0$ and for European options set $r_\Psi = r$ with $r\geq 0$ the prevailing risk-free interest rate. Define ψ to be the associated Black&Scholes price,

$$\psi(t,x)=\psi^{bs,r_\psi}(t,x)=e^{-\eta x}e^{-r_\psi t}\frac{1}{2\pi}\int_{\mathbb{R}}e^{i\xi x}\widehat{f_K}(-(\xi+i\eta))\varphi_{t,\sigma}^{bs,r_\psi}(\xi+i\eta)\,\mathrm{d}\xi, \qquad (3.102)$$

wherein $f_K = g$, the initial condition and A the symbol of the associated operator $\mathcal{A}$ in (3.16). *Then, the right hand side $F:[0,T]\to\mathbb{R}^N$ takes the form*

$$F_j(t)=\frac{1}{2\pi}\int_{\mathbb{R}}\Big((r_\psi-r)+\left(A^{bs,r_\psi}-A\right)(\xi-i\eta)\Big)$$
$$\widehat{f_K}(\xi-i\eta)\exp\left(-t\left(r_\psi+A^{bs,r_\psi}(\xi-i\eta)\right)\right)\overline{\widehat{\varphi_j}(\xi+i\eta)}\,\mathrm{d}\xi. \qquad (3.103)$$

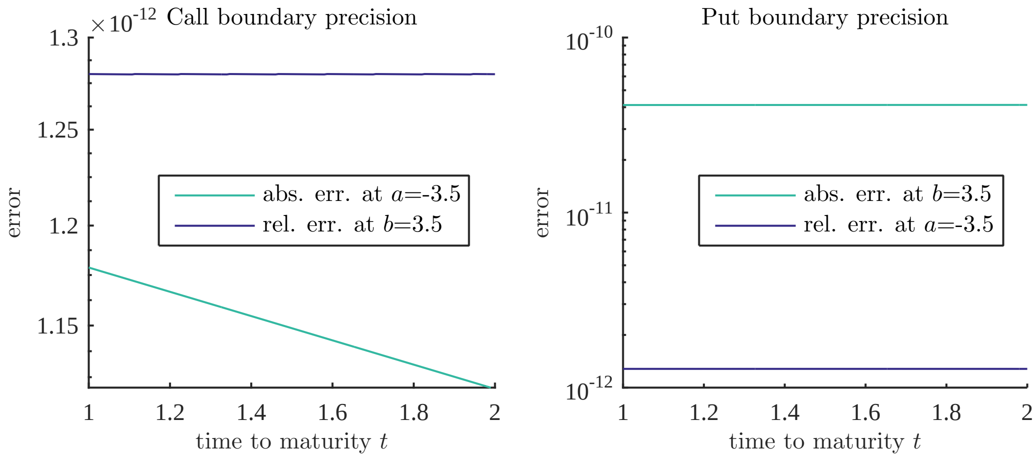

Figure 3.6 Precision study of the boundary conditions for European call (left) and put options (right) now given by (3.114). We compare $\psi \times \Phi_{0,\sigma_\Phi=0.5}$ or $\psi \times (1 - \Phi_{0,\sigma_\Phi=0.5})$, respectively, to prices of the Black&Scholes model. Equivalently to Figure 3.5 we set $r = 0.05$, $K = 1$, $\sigma = 0.3$ and evaluate European call and put prices for $S_0^{\max} = e^{x_{\max}}$ and $S_0^{\min} = e^{x_{\min}}$. In contrast to Figure 3.5, the absolute values of $x_{\min}$ and $x_{\max}$ have to be increased to $x_{\min} = -3.5$ and $x_{\max} = 3.5$ to achieve comparable accuracy.

for all $j \in \{1, \ldots, N\}$.

Proof

In order to derive the right hand side, we need to represent ψ in Fourier terms. Since for call and put options, $\psi \notin L^1(\mathbb{R})$, we compute the (generalized) Fourier transform of ψ or the Fourier transform of ψ_η, respectively. We get

$$\begin{aligned}\psi_\eta(t,x) &= e^{\eta x}\psi^{\mathrm{bs},r_\psi}(t,x)\\ &= e^{-r_\psi t}\frac{1}{2\pi}\int_{\mathbb{R}} e^{i\xi x}\widehat{f_K}(-(\xi+i\eta))\varphi_{t,\sigma}^{\mathrm{bs},r_\psi}(\xi+i\eta)\,\mathrm{d}\xi\\ &= e^{-r_\psi t}\frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi x}\widehat{f_K}(\xi-i\eta)\varphi_{t,\sigma}^{\mathrm{bs},r_\psi}(-(\xi-i\eta))\,\mathrm{d}\xi.\end{aligned} \tag{3.104}$$

The integral in (3.104) is a Fourier (inversion) integral. Hence,

$$\begin{aligned}\widehat{\psi_\eta}(\xi,t) &= e^{-r_\psi t}\widehat{f_K}(\xi-i\eta)\varphi_t^{\mathrm{bs},r_\psi}(-(\xi-i\eta))\\ &= e^{-r_\psi t}\widehat{f_K}(\xi-i\eta)\exp\left(-tA^{\mathrm{bs},r_\psi}(\xi-i\eta)\right)\\ &= \widehat{f_K}(\xi-i\eta)\exp\left(-t\left(r_\psi + A^{\mathrm{bs},r_\psi}(\xi-i\eta)\right)\right),\end{aligned} \tag{3.105}$$

where we used the relation between the characteristic function and the symbol of a process, confer Remark 2.19. Next, we prove that

$$\widehat{\frac{\partial}{\partial t}\psi_\eta}(t,\xi) = \frac{\partial}{\partial t}\widehat{\psi_\eta}(t,\xi) \tag{3.106}$$

for almost all $t \in [0, T]$. For notational convenience we assume $r = 0$ and $K = 1$ for the proof of (3.106). Let $0 < \varepsilon < T$ and define $I_\varepsilon = [\varepsilon, T]$. Thus, the claim (3.106) holds, if

$$\begin{aligned}\int_{\mathbb{R}} e^{i\xi x} \frac{\partial}{\partial t} \int_{\mathbb{R}} e^{-izx} \widehat{f}_1(z - i\eta) \varphi^{\mathrm{bs}}_{t,\sigma}(-(z - i\eta))\, \mathrm{d}z\, \mathrm{d}x& \\ = \int_{\mathbb{R}} e^{i\xi x} \int_{\mathbb{R}} e^{-izx} \widehat{f}_1(z - i\eta) \frac{\partial}{\partial t} \varphi^{\mathrm{bs}}_{t,\sigma}(-(z - i\eta))\, \mathrm{d}z\, \mathrm{d}x&,\end{aligned} \tag{3.107}$$

which holds if

$$\begin{aligned}\frac{\partial}{\partial t} \int_{\mathbb{R}} e^{-izx} \widehat{f}_1(z - i\eta) \varphi^{\mathrm{bs}}_{t,\sigma}(-(z - i\eta))\, \mathrm{d}z& \\ = \int_{\mathbb{R}} e^{-izx} \widehat{f}_1(z - i\eta) \frac{\partial}{\partial t} \varphi^{\mathrm{bs}}_{t,\sigma}(-(z - i\eta))\, \mathrm{d}z,& \quad \forall x \in \mathbb{R}.\end{aligned} \tag{3.108}$$

Fix $x \in \mathbb{R}$. The integrand on the left of (3.108) is integrable for all $t \in I_\varepsilon$ and it is differentiable for all $z \in \mathbb{R}$. Furthermore, the integrand on the right of (3.108) is bounded by a function h independent of $t \in I_\varepsilon$, since

$$\begin{aligned}&\left| e^{-izx} \widehat{f}_1(z - i\eta) \frac{\partial}{\partial t} \varphi^{\mathrm{bs}}_{t,\sigma}(-(z - i\eta)) \right| \\ &= \left| \widehat{f}_1(z - i\eta) \left(ib(-(z - i\eta)) - \frac{1}{2}\sigma^2 (z - i\eta)^2 \right) \varphi^{\mathrm{bs}}_{t,\sigma}(-(z - i\eta)) \right| \\ &= \left| \widehat{f}_1(z - i\eta) \left(ib(z - i\eta) + \frac{1}{2}\sigma^2 (z - i\eta)^2 \right) \right| \\ &\qquad \left| \exp\left(it[-bz + \sigma^2 \eta z] - tb\eta - t\frac{1}{2}\sigma^2 (z^2 - \eta^2) \right) \right| \\ &\leq \left| \widehat{f}_1(z - i\eta) \left(ib(z - i\eta) + \frac{1}{2}\sigma^2 (z - i\eta)^2 \right) \right| \\ &\qquad \max_{t \in [\varepsilon, T]} \exp\big(- t(b\eta - \frac{1}{2}\sigma^2\eta^2) \big) \exp\Big(- \varepsilon \frac{1}{2} \sigma^2 z^2 \Big) \\ &= h(z)\end{aligned} \tag{3.109}$$

wherein $b \in \mathbb{R}$ is the risk neutral drift chosen according to (2.35) from Section 2.3.1 in the preliminary chapter. The upper bound derived in (3.109) is integrable, $h \in L^1(\mathbb{R})$. We may therefore apply Lemma 16.2 from Bauer (1992) which validates identity (3.108) and thus proves identity (3.106) for all $t \in I_\varepsilon$. Since ε can be chosen arbitrarily small, identity (3.106) holds almost everywhere on $[0, T]$. We may thus exchange integration and differentiation and get

$$\begin{aligned}\widehat{\frac{\partial}{\partial t}\psi_\eta}(t, \xi) &= \frac{\partial}{\partial t} \widehat{\psi_\eta}(t, \xi) \\ &= \widehat{f_K}(\xi - i\eta) \left(- \left(r_\psi + A^{\mathrm{bs}, r_\psi}(\xi - i\eta) \right) \right) \exp\left(-t \left(r_\psi + A^{\mathrm{bs}, r_\psi}(\xi - i\eta) \right) \right) \\ &= - \left(r_\psi + A^{\mathrm{bs}, r_\psi}(\xi - i\eta) \right) \widehat{\psi_\eta}(t, \xi).\end{aligned} \tag{3.110}$$

Finally, since $\psi^{\mathrm{bs},r_\psi} \in H^{\alpha/2}_\eta(\mathbb{R})$, we have analogously to identity (3.100), that

$$\int_{\mathbb{R}} (\mathcal{A}\psi^{\mathrm{bs},r_\psi})(t,x)\varphi_j(x)\,\mathrm{d}x = \frac{1}{2\pi}\int_{\mathbb{R}} A(\xi - i\eta)\widehat{\psi^{\mathrm{bs},r_\psi}(t,\cdot)}(\xi - i\eta)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi. \tag{3.111}$$

So, collecting our results, for $\psi = \psi^{\mathrm{bs},r_\psi}$ we arrive at

$$\begin{aligned}
F_j(t) = &-\int_{\mathbb{R}} \left(\frac{\partial}{\partial t}\psi^{\mathrm{bs},r_\psi}(t,x) + (\mathcal{A}\psi^{\mathrm{bs},r_\psi})(t,x) + r\psi^{\mathrm{bs},r_\psi}(t,x)\right)\varphi_j(x)\,\mathrm{d}x\\
= &-\frac{1}{2\pi}\int_{\mathbb{R}} \Bigg(-\left(r_\psi + A^{\mathrm{bs},r_\psi}(\xi - i\eta)\right)\widehat{\psi_\eta^{\mathrm{bs},r_\psi}}(t,\xi)\\
&\qquad + A(\xi - i\eta)\widehat{\psi_\eta^{\mathrm{bs},r_\psi}}(t,\xi)\\
&\qquad + r\widehat{\psi_\eta^{\mathrm{bs},r_\psi}}(\xi,t)\Bigg)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi\\
= &\frac{1}{2\pi}\int_{\mathbb{R}} \Big((r_\psi - r) + \left(A^{\mathrm{bs},r_\psi}(\xi - i\eta) - A(\xi - i\eta)\right)\Big)\widehat{\psi_\eta^{\mathrm{bs},r_\psi}}(t,\xi)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi\\
= &\frac{1}{2\pi}\int_{\mathbb{R}} \Big((r_\psi - r) + \left(A^{\mathrm{bs},r_\psi} - A\right)(\xi - i\eta)\Big)\\
&\qquad \widehat{f_K}(\xi - i\eta)\exp\left(-t\left(r_\psi + A^{\mathrm{bs},r_\psi}(\xi - i\eta)\right)\right)\overline{\widehat{\varphi_j}(\xi + i\eta)}\,\mathrm{d}\xi,
\end{aligned} \tag{3.112}$$

which proves the claim. □

For the choice of η in Lemma 3.24, consider Proposition 2.20 or Lemma 2.21, respectively, for plain vanilla European options. The candidate $\psi = \psi^{\mathrm{bs},r_\Psi}$ matches the criteria of Remark 3.23. It is quickly evaluable, since functions for yielding Black&Scholes prices are implemented in many code libraries. Also, the integral in (3.103) is numerically accessible, since the integrand decays fast.

Remark 3.25 ($r_\Psi = 0$ for American options)
Choosing ψ to be Black&Scholes prices does not only suit the case of European plain vanilla options but American ones, as well. Only the value of r_Ψ needs to be adjusted. When no dividends are paid, the price of an American call options is equal to the price of a European call option. In this case, the Lemma applies identically. For put options, however, the boundary conditions change when an American put instead of a European put is considered. Then, the boundary conditions coincide with those of a European put when interest rates are assumed to be equal to zero,

$$\begin{aligned}
V^P_{Am}(x,t) &\to K - e^x, & x &\to -\infty,\\
V^P_{Am}(x,t) &\to 0, & x &\to +\infty,
\end{aligned} \tag{3.113}$$

confer also Chapter 11 in Hull (2015). Consequently, for American options, choose $r_\Psi = 0$ in (3.102).

A major disadvantage of choosing $\psi = \psi^{\mathrm{bs},r_\Psi}$, however, lies in the fact that neither $j \in \{1,\dots,N\}$ nor $t \in [0,T]$ can be separated from the integrand in (3.103). Consequently, $F_j(t^k)$, $j \in \{1,\dots,N\}$, $k \in \{1,\dots,M\}$, must be numerically evaluated on each grid node individually. This results in significant numerical cost. We therefore present a second candidate for ψ that avoids this issue.

Lemma 3.26 (Subtracting Quasi-Hockey stick multiplied by Normal)
Let $\sigma_\psi > 0$. In the European option case set $r_\psi = r$, with $r \geq 0$ the prevailing risk-free interest rate. Define ψ^C in the call option and ψ^P in the put option case by

$$\begin{aligned} \psi^C(t,x) &= \left(e^x - Ke^{-r_\psi t}\right)\Phi(x), && (t,x) \in [0,T]\times[a,b], \\ \psi^P(t,x) &= \left(Ke^{-r_\psi t} - e^x\right)(1-\Phi(x)), && (t,x) \in [0,T]\times[a,b], \end{aligned} \tag{3.114}$$

where Φ denotes the cumulative distribution function of the normal $\mathcal{N}(0,\sigma_\psi^2)$ distribution. Furthermore, in the call option case choose $\eta < -1$ and $\eta > 0$ in the put option case. Then, the right hand side $F : [0,T] \to \mathbb{R}^N$ is given by

$$\begin{aligned} F_j(t) = \frac{1}{2\pi}\Bigg(&\int_{\mathbb{R}} \left(A(\xi - i\eta) + r\right) \frac{\widehat{f^{\mathcal{N}}}(\xi - i(\eta+1))}{i\xi + (\eta+1)} \overline{\widehat{\varphi_j}(\xi + i\eta)}\,\mathrm{d}\xi \\ &- e^{-r_\psi t}K \int_{\mathbb{R}} \left(r - r_\psi + A(\xi - i\eta)\right) \frac{\widehat{f^{\mathcal{N}}}(\xi - i\eta)}{i\xi + \eta} \overline{\widehat{\varphi_j}(\xi + i\eta)}\,\mathrm{d}\xi \Bigg), \end{aligned} \tag{3.115}$$

for all $j \in \{1,\dots,N\}$ with $t \in [0,T]$, where A is the symbol of the associated operator $\mathcal{A}$ in PIDE (3.52) and with

$$\widehat{f^{\mathcal{N}}}(\xi) = \exp\left(-\frac{1}{2}\sigma_\psi^2\xi^2\right),$$

the Fourier transform of the normal $\mathcal{N}(0,\sigma_\psi^2)$ density derived in Lemma 2.3.

Proof
We consider the call option case, first. To derive the expression for F_j in (3.115) we need to compute the Fourier transform of (the appropriately weighted) ψ^C. We choose $\eta < -1$ arbitrary but fix and $t \in [0,T]$ arbitrary but fix and compute for $K = 1$,

$$\begin{aligned} \widehat{\psi_\eta^C(t,\cdot)}(\xi) &= \int_{\mathbb{R}} e^{i\xi x} e^{\eta x}\left(e^x - e^{-r_\psi t}\right)\Phi(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}} e^{i\xi x} e^{(\eta+1)x}\Phi(x)\,\mathrm{d}x - e^{-r_\psi t}\int_{\mathbb{R}} e^{i\xi x}e^{\eta x}\Phi(x)\,\mathrm{d}x. \end{aligned} \tag{3.116}$$

We take the first integral in (3.116) and get by applying integration by parts

$$\begin{aligned} \int_{\mathbb{R}} e^{i\xi x}e^{(\eta+1)x}\Phi(x)\,\mathrm{d}x &= \int_{\mathbb{R}} e^{i(\xi - i(\eta+1))x}\Phi(x)\,\mathrm{d}x \\ &= \left[\frac{e^{i(\xi-i(\eta+1))x}}{i\xi + (\eta+1)}\Phi(x)\right]_{-\infty}^{+\infty} - \int_{\mathbb{R}} \frac{e^{i(\xi-i(\eta+1))x}}{i\xi+(\eta+1)} f^{\mathcal{N}}(x)\,\mathrm{d}x, \end{aligned} \tag{3.117}$$

where

$$f^{\mathcal{N}}(x) = \frac{\partial}{\partial x}\Phi(x) \tag{3.118}$$

denotes the density of the normal $\mathcal{N}(0,\sigma_\psi^2)$. Since $\eta < -1$, the non-integral part in (3.117) tends to zero as $x \to +\infty$. Furthermore, by l'Hôpital's rule we have

$$\begin{aligned}
\lim_{x\to-\infty} e^{(\eta+1)x}\Phi(x) &= \lim_{x\to-\infty} \frac{\Phi(x)}{\exp\big(-(\eta+1)x\big)} \\
&= \lim_{x\to-\infty} \frac{f^{\mathcal{N}}(x)}{-(\eta+1)\exp\big(-(\eta+1)x\big)} \\
&= -\frac{1}{\eta+1}\frac{1}{\sqrt{2\pi\sigma_\psi^2}} \lim_{x\to-\infty} \exp\left(-\frac{1}{2}\sigma_\psi^2 x^2\right)\exp\big((\eta+1)x\big) \\
&= 0.
\end{aligned} \tag{3.119}$$

Hence, the non-integral part in (3.117) is equal to zero and we have

$$\int_{\mathbb{R}} e^{i\xi x}e^{(\eta+1)x}\Phi(x)\,\mathrm{d}x = -\frac{1}{i\xi+(\eta+1)}\int_{\mathbb{R}} e^{i(\xi-i(\eta+1))x}f^{\mathcal{N}}(x)\,\mathrm{d}x, \tag{3.120}$$

which can be expressed in terms of the Fourier transform of the normal distribution yielding

$$\int_{\mathbb{R}} e^{i\xi x}e^{(\eta+1)x}\Phi(x)\,\mathrm{d}x = -\frac{\widehat{f^{\mathcal{N}}}(\xi-i(\eta+1))}{i\xi+(\eta+1)}. \tag{3.121}$$

Equivalently, we obtain for the second integral in (3.116)

$$\int_{\mathbb{R}} e^{i\xi x}e^{\eta x}\Phi(x)\,\mathrm{d}x = -\frac{\widehat{f^{\mathcal{N}}}(\xi-i\eta)}{i\xi+\eta}. \tag{3.122}$$

Assembling these results we find

$$\widehat{\psi_\eta^C(t,\cdot)}(\xi) = -\frac{\widehat{f^{\mathcal{N}}}(\xi-i(\eta+1))}{i\xi+(\eta+1)} + e^{-r_\psi t}\frac{\widehat{f^{\mathcal{N}}}(\xi-i\eta)}{i\xi+\eta}. \tag{3.123}$$

As in the proof of Lemma 3.24, we exchange differentiation and integration and get

$$\widehat{\frac{\partial}{\partial t}\psi_\eta^C(t,\cdot)}(\xi) = \frac{\partial}{\partial t}\widehat{\psi_\eta^C(\cdot,t)}(\xi) = -r_\psi e^{-r_\psi t}\frac{\widehat{f^{\mathcal{N}}}(\xi-i\eta)}{i\xi+\eta}. \tag{3.124}$$

We thus have

$$\begin{aligned}
F_j(t) = &-\frac{1}{2\pi}\int_{\mathbb{R}}\left(\dot{\widehat{\psi_\eta^C}}(\xi,t) + A(\xi-i\eta)\widehat{\psi_\eta^C}(\xi,t) + r\widehat{\psi_\eta^C}(\xi,t)\right)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi \\
= &-\frac{1}{2\pi}\int_{\mathbb{R}}\Bigg(-r_\psi e^{-r_\psi t}\frac{\widehat{f^{\mathcal{N}}}(\xi-i\eta)}{i\xi+\eta} \\
&+\big(A(\xi-i\eta)+r\big)\left(-\frac{\widehat{f^{\mathcal{N}}}(\xi-i(\eta+1))}{i\xi+(\eta+1)} + e^{-r_\psi t}\frac{\widehat{f^{\mathcal{N}}}(\xi-i\eta)}{i\xi+\eta}\right)\Bigg)\overline{\widehat{\varphi_{j}}_{-\eta}(\xi)}\,\mathrm{d}\xi
\end{aligned}$$

from which we deduce by splitting the integral

$$
\begin{aligned}
F_j(t) = \frac{1}{2\pi}\Bigg(\int_{\mathbb{R}} \big(A(\xi - i\eta) + r\big) \frac{\widehat{f^{\mathcal{N}}}(\xi - i(\eta+1))}{i\xi + (\eta+1)} \overline{\widehat{\varphi_j}(\xi + i\eta)}\, \mathrm{d}\xi \\
- e^{-r_\psi t} \int_{\mathbb{R}} \big(r - r_\psi + A(\xi - i\eta)\big) \frac{\widehat{f^{\mathcal{N}}}(\xi - i\eta)}{i\xi + \eta} \overline{\widehat{\varphi_j}(\xi + i\eta)}\, \mathrm{d}\xi \Bigg)
\end{aligned}
\tag{3.125}
$$

with

$$
\widehat{f^{\mathcal{N}}}(\xi) = \exp\left(-\frac{1}{2}\sigma_\psi^2 \xi^2\right).
$$

For the put option case we choose as defined in (3.114),

$$
\begin{aligned}
\psi^P(x,t) &= \left(Ke^{-r_\psi t} - e^x\right)(1 - \Phi(x)) \\
&= \left(e^x - Ke^{-r_\psi t}\right)(\Phi(x) - 1).
\end{aligned}
\tag{3.126}
$$

Since

$$
\frac{\partial}{\partial x}(\Phi(x) - 1) = \frac{\partial}{\partial x}\Phi(x), \qquad \forall x \in \mathbb{R}, \tag{3.127}
$$

the computations for $\widehat{\psi_\eta^P}$ follow along the same lines as they do for the call and we get the relation

$$
\widehat{\psi_\eta^P(t,\cdot)}(\xi) = \widehat{\psi_\eta^C(t,\cdot)}(\xi), \qquad \forall (t,\xi) \in [0,T] \times \mathbb{R}, \tag{3.128}
$$

for η set to some $\eta > 0$, which proves the claim. □

Remark 3.27 (Computational features of ψ^C and ψ^P)
While ψ^C serves as localizing function for the call option case, ψ^P can be used in the put option case. Both candidates are based on their "naive" counterparts in (3.91) *but avoid the lack of differentiability with respect to x in $x = \log(Ke^{-rt})$ for $t \in [0,T]$. As a consequence, both ψ^C and ψ^P are very smooth functions and thus fulfill the requirements collected in Remark 3.23 when σ_ψ is chosen small enough. Additionally, the two integrals in* (3.115) *do not depend on the time variable $t \in [0,T]$ and thus need to be computed only once for each basis function φ_j. This results in a significant acceleration in computational time compared to the suggestion $\psi = \psi^{bs,\sigma_\psi}$ of Lemma 3.24.*

We implemented the FEM solver as sketched above in MATLAB and conducted a study of the empirical order of convergence. The results of this study can be found in Section 3.5 below.

We have also tested the implementation in a project analyzing a method commonly used by practitioners for model calibration purposes. Clearly, our implementation as outlined above is designed with European options in mind. As such, it is a valuable tool for calibrating the Merton model to European option prices in the market. Yet, practice argues that American options are traded more liquidly and thus would offer a more favorable source for reference prices in model calibration. Calibrating a model to American prices, however, depends on the ability to derive American model prices for a vast amount of model parameter constellations within a reasonable amount of time which is typically numerically unfeasible. Therefore, some practitioners take American option prices that they observe in the market, strip off the component that represents the price for the American feature and calibrate their models in a European fashion to quasi-European prices that result from that transformation. This method is known as *De-Americanization.* Its effect on pricing and calibration is studied extensively in Burkovska et al. (2016) to which our implementation contributed the results with regards to pricing call and put options in the Merton model.

3.4 A general FEM solver based on the symbol method

Section 3.3 has provided us with a FEM solver capable of deriving European call and put option prices in the Merton model. The key ingredients of the solver have been analytically derived. Let us emphasize our two main findings from that exercise. First, the analytic treatment of the Lévy measure presented a serious challenge during the computations. Especially the double integral term and the Lévy density required lengthy and tedious consideration. Second, the actual computations we performed are closely tied to the Merton model. Naively setting up a FEM solver for different models in the same way would put us in the position of having to adapt all of our Merton-specific calculations with respect to the Lévy measure of the new model. These two findings underline that our first approach above can hardly be generalized to other models without serious computational efforts for each new model individually.

Consequently, in this section we approach the calculation of FEM solver components differently. In Section 3.3.5, Parseval's identity of Theorem 2.7 has enabled us to compute the right hand side by numerical integration of the Fourier transforms of the involved quantities. We have seen that in the course of this transformation, dealing with the operator of the underlying model has vanished while the associated symbol appeared in the calculations, instead. In stark contrast to the operator, the symbol of a Lévy model is numerically accessible in many cases and we will present several examples in the following. This feature nourishes the hope of being able to renounce the treatment of the operator alltogether by shifting the focus to its Fourier counterpart, the symbol, instead. Investigating this shift in perspective, this section aims at establishing a numerical FEM solver framework that

i) provides flexibility in the choice of the asset model and thus

ii) avoids tedious individual consideration of different models but still

iii) maintains numerical feasibility.

As we will see, achieving these core aims comes at a certain cost. While considering the FEM solver components in Fourier space will be highly advantageous regarding some aspects, it will also pose new challenges regarding others. More precisely, while shifting our perspective to Fourier spaces solves the problem of having to consider the operator, at the same time it leaves us with new numerical challenges concerning the choice of basis functions. The contents of this section that focuses on the symbol method also appear in Gaß and Glau (2016).

Before we consider these new challenges, let us state the core lemma of this section.

Lemma 3.28 (Symbol method for bilinear forms)
Let $A \in S^0_\alpha$ be a univariate symbol as introduced in Definition 2.17 and let $\mathcal{A}$ be the associated operator in a PIDE of form (3.1). Further, let $a(\cdot,\cdot)$ be the associated bilinear form. If there exists a constant $c > 0$ such that

$$a(u,v) \le c\|u\|_{H^{\alpha/2}_0(\mathbb{R})}\|v\|_{H^{\alpha/2}_0(\mathbb{R})}, \qquad \forall u,v \in C^\infty_0(\mathbb{R}), \tag{3.129}$$

then the bilinear form possesses a unique linear extension

$$a : H^{\alpha/2}_0(\mathbb{R}) \times H^{\alpha/2}_0(\mathbb{R}) \to \mathbb{C}. \tag{3.130}$$

Assume further for $N \in \mathbb{N}$ a set of functions $\varphi_0, \varphi_1, \ldots, \varphi_N \in H^{\alpha/2}_0(\mathbb{R})$ and constants $x_1, \ldots, x_N \in \mathbb{R}$, such that for all $i \in \{1,\ldots,N\}$

$$\varphi_i(x) = \varphi_0(x - x_i), \qquad \forall x \in \mathbb{R},$$

holds. Then we have

$$a(\varphi_l, \varphi_k) = \frac{1}{2\pi}\int_{\mathbb{R}} A(\xi)e^{i\xi(x_l - x_k)}\left|\widehat{\varphi_0}(\xi)\right|^2 \mathrm{d}\xi. \tag{3.131}$$

for all $k, l \in \{1,\ldots,N\}$. If additionally

$$\Re(A(\xi)) = \Re(A(-\xi)) \quad \text{and} \quad \Im(A(\xi)) = -\Im(A(-\xi)), \tag{3.132}$$

then

$$a(\varphi_l, \varphi_k) = \frac{1}{\pi}\int_0^\infty \Re\left(A(\xi)e^{i\xi(x_l - x_k)}\right)\left|\widehat{\varphi_0}(\xi)\right|^2 \mathrm{d}\xi \tag{3.133}$$

for all $k, l \in \{1,\ldots,N\}$.

Proof
Due to property i) in Lemma 2.4

$$\widehat{\varphi_j}(\xi) = e^{i\xi x_j}\widehat{\varphi_0}(\xi). \tag{3.134}$$

Since $\varphi_i \in H^{\alpha/2}_0(\mathbb{R})$, for all $i \in \{1,\ldots,N\}$, the identity (3.131) follows from Theorem 4.1 and Remark 5.2 and the lines thereafter in Eberlein and Glau (2011), see also page 68 in Glau (2010). The second claim (3.133) is then elementary. □

Remark 3.29 (On the symbol method for bilinear forms)
Lemma 3.28 provides an appealing formula to derive the values of all entries in the stiffness matrix $(A_{ij})_{i,j\in\{1,\dots,N\}}$. It offers an alternative to explicitly considering the effect of the operator $\mathcal{A}$ on the basis functions that we presented in Section 3.3.4. Instead, it exploits the availability of the associated symbol A that often contains the model information in an explicit and numerically pleasing way, as the following examples show.

Corollary 3.30 (Symbol method for stiffness matrices)
Let $A \in S^0_\alpha$ be a univariate symbol associated with the operator $\mathcal{A}$ of a PIDE of form (3.52). Denote by $\varphi_i \in L^1(\mathbb{R})$, $i \in \{1,\dots,N\}$ the basis functions of a Galerkin solving scheme associated with an equidistantly spaced grid $\Omega = \{x_1,\dots,x_N\}$ possessing the property

$$\varphi_i(x) = \varphi_0(x - x_i), \qquad \forall x \in \mathbb{R}, \tag{3.135}$$

for some $\varphi_0 : \mathbb{R} \to \mathbb{R}$. Then, the stiffness matrix $A \in \mathbb{R}^{N\times N}$ of the scheme can be computed by

$$A_{kl} = \frac{1}{2\pi}\int_{\mathbb{R}} A(\xi)e^{i\xi(x_l - x_k)}\left|\widehat{\varphi_0}(\xi)\right|^2 \mathrm{d}\xi \tag{3.136}$$

for all $k, l \in \{1,\dots,N\}$.

Proof
The proof is an immediate consequence of Lemma 3.28. □

Earlier, we introduced operators $\mathcal{A}$ and the characteristic triplets (b, σ, F) of some well known asset models. In Example 3.22 we have already seen the symbol of the Merton model. The following examples present the symbols of the remaining models introduced before.

Example 3.31 (Symbol in the Black&Scholes (BS) model)
In the univariate Black&Scholes model, determined by the Brownian volatility $\sigma > 0$, the symbol is given by

$$A(\xi) = A^{bs}(\xi) = i\xi b + \frac{1}{2}\sigma^2\xi^2, \tag{3.137}$$

with drift set b to

$$b = r - \frac{1}{2}\sigma^2 \tag{3.138}$$

as seen in Example 3.11.

Example 3.32 (Symbol in the CGMY model)
In the CGMY model of Carr et al. (2002) with $\sigma > 0$, $C > 0$, $G \geq 0$, $M \geq 0$ and $Y \in (1, 2)$, the symbol computes to

$$\begin{aligned} A(\xi) = A^{cgmy}(\xi) = i\xi b + \frac{1}{2}\sigma^2\xi^2 \\ - C\Gamma(-Y)\left[(M + i\xi)^Y - M^Y + (G - i\xi)^Y - G^Y\right], \end{aligned} \tag{3.139}$$

for all $\xi \in \mathbb{R}$, with drift b set to

$$b = r - \frac{1}{2}\sigma^2 - C\Gamma(-Y)\left[(M-1)^Y - M^Y + (G+1)^Y - G^Y\right] \tag{3.140}$$

for martingale pricing.

Example 3.33 (Symbol in the NIG model)
With $\sigma > 0$, $\alpha > 0$, $\beta \in \mathbb{R}$ and $\delta > 0$ such that $\alpha^2 > \beta^2$, the symbol of the NIG model is given by

$$A(\xi) = A^{nig}(\xi) = \frac{1}{2}\sigma^2\xi^2 + i\xi b - \delta\left(\sqrt{\alpha^2-\beta^2} - \sqrt{\alpha^2-(\beta-i\xi)^2}\right) \tag{3.141}$$

for all $\xi \in \mathbb{R}$ with drift given by

$$b = r - \frac{1}{2}\sigma^2 - \delta\left(\sqrt{\alpha^2-\beta^2} - \sqrt{\alpha^2-(\beta+1)^2}\right) \tag{3.142}$$

Corollary 3.34 (Deriving the BS stiffness matrix using the symbol)
Denote by $r \geq 0$ the prevailing constant risk-free interest rate. Consider the pricing PDE of the univariate Black&Scholes model, that is (3.52) *with operator $\mathcal{A}$ given by Example 3.11. Consider the numerical implementation of a FEM solver assuming the hat functions φ_i, $i \in \{1,\ldots,N\}$, of Definition 3.16 for some $N > 0$ as basis functions on an equidistant grid with fineness $h > 0$. Then the respective stiffness matrix $A \in \mathbb{R}^{N\times N}$ is given by*

$$\begin{aligned} A_{ij} = \frac{2\sigma^2}{\pi h^2}\int_0^\infty \frac{1}{\xi^2}\cos(\xi h(j-i))(1-\cos(\xi h))^2\,\mathrm{d}\xi \\ - \frac{4b}{\pi h^2}\int_0^\infty \frac{1}{\xi^3}\sin(\xi h(j-i))(1-\cos(\xi h))^2\,\mathrm{d}\xi + rM_{ij} \end{aligned} \tag{3.143}$$

for all $i, j \in \{1,\ldots,N\}$, where $M \in \mathbb{R}^{N\times N}$ is the model-independent mass matrix given by Lemma 3.18

Proof
The stiffness matrix A is given by the bilinear form $a(\cdot,\cdot) : H_0^{\alpha/2}(\mathbb{R}) \times H_0^{\alpha/2}(\mathbb{R}) \to \mathbb{R}$ with $\alpha = 2$, associated with the operator $\mathcal{A}$ by

$$A_{ij} = a(\varphi_j, \varphi_i) + rM_{ij}.$$

Let φ_0 be the hat function centered over the origin with $\operatorname{supp}\varphi_0 \subset (-h,h)$, as defined in (3.60). Since the Black&Scholes symbol $A = A^{\mathrm{bs}}$ fulfills condition (3.132) of Lemma 3.28, we have

$$a(\varphi_j, \varphi_i) = \frac{1}{\pi}\int_0^\infty \Re\left(A^{\mathrm{bs}}(\xi)e^{i\xi(x_j-x_i)}\right)|\widehat{\varphi_0}(\xi)|^2\,\mathrm{d}\xi.$$

Inserting the formula of $\widehat{\varphi_0}$ from Lemma 3.17 and A^{bs} from (3.137) of Example 3.31 yields the claim. □

Remark 3.35 (Toeplitz structure of stiffness matrix)
The mass matrix M is a Toeplitz matrix given that the basis functions are defined on an equidistant grid and possess property (3.135). *We observe that the values of the integrals in Equation* (3.136) *in Corollary 3.30 only depend on the value of $j - i \in \{-(N-1), \ldots, -1, 0, 1, \ldots, N-1\}$. This means, that each individual diagonal of A is determined by only one single value in the sense of Definition 2.41. Consequently, the stiffness matrix is a Toeplitz matrix, as well. Thus, for its numerical derivation only $2N - 1$ instead of N^2 integrals have to be computed. This feature is lost, if the grid that the basis functions populate is not equidistantly spaced.*

Algorithm 1 A symbol method based FEM solver

```
Choose equidistant space grid $x_i$, $i \in \{1, \ldots, N\}$
Choose basis functions $\varphi_i$, $i \in \{1, \ldots, N\}$, with $\varphi_i(x) = \varphi_0(x - x_i)$ for some $\varphi_0$
Choose equidistant time grid $T_j$, $j \in \{0, \ldots, M\}$
procedure COMPUTE MASS MATRIX $M$
    Derive the mass matrix $M \in \mathbb{R}^{N \times N}$ by
        $M_{kl} = \int_{\mathbb{R}} \varphi_l(x)\varphi_k(x)\,\mathrm{d}x, \qquad \forall k, l \in \{1, \ldots, N\}$
procedure COMPUTE STIFFNESS MATRIX $A$
    Derive the stiffness matrix $A \in \mathbb{R}^{N \times N}$ by plugging the symbol $A$ of the chosen
    model into the following formula and computing
        $A_{kl} = \frac{1}{2\pi} \int_{\mathbb{R}} A(\xi)\, e^{i\xi(x_l - x_k)}\, |\widehat{\varphi_0}(\xi)|^2 \,\mathrm{d}\xi, \qquad \forall k, l \in \{1, \ldots, N\}$
    using numerical integration
procedure RUN THETA SCHEME
    Following the suggestions by Lemma 3.24 or Lemma 3.26 for plain vanilla European
    options choose a function $\psi$ to subtract from the original pricing problem
    to obtain a zero boundary problem and retrieve the respective basis function
    coefficient vectors $\overline{\psi}^k \in \mathbb{R}^N$, $k \in \{0, \ldots, M\}$
    Choose an appropriate basis function coefficient vector $V^1 \in \mathbb{R}^N$ matching the
    initial condition of the transformed problem
    Derive the right hand side vectors $F^k \in \mathbb{R}^N$, $k \in \{0, \ldots, M\}$, as defined in
    Lemma 3.24 or Lemma 3.26 matching the choice of $\psi$
    Choose $\theta \in [0, 1]$ and run the iterative scheme
    for $k = 0 : (M - 1)$
        $V^{k+1} = (M + \Delta t\,\theta\,A)^{-1} \big((M - \Delta t\,(1 - \theta)\,A)\,V^k + F^{k+\theta}\big)$
    end
procedure RECONSTRUCT SOLUTION TO ORIGINAL PROBLEM
    Add previously subtracted right hand side $\psi$ to the solution of the transformed
    problem by computing
        $\widetilde{V}^k = V^k + \overline{\psi}^k, \qquad k \in \{0, \ldots, M\}$
    to retrieve the basis function coefficient vectors $\widetilde{V}^k$, $k \in \{0, \ldots, M\}$, to the orig-
    inal pricing problem
```

Algorithm 1 summarizes the abstract structure of a general FEM solver based on the symbol method. By plugging the symbol associated to the model of choice into the computation of line 9 of the algorithm, the solver instantly adapts to that model. In other words, only one line needs to be specified to obtain a model specific solver for option pricing. As Examples 3.31, 3.32, 3.33 and others emphasize, the symbol exists in analytically (semi–)closed form for many models, indeed. Algorithm 1 thus provides a very appealing tool for FEM pricing in practice. Model specific computations that we had encountered earlier for the Merton model have become unnecessary.

3.4.1 Numerical aspects

By now we have seen two alternative ways to compute the stiffness matrix A. The derivation in Section 3.3.4 required the consideration of the Lévy measure F. Taking the Merton model as an example we understood that long and tedious calculations may come with this approach. Section 3.4 offered a different solution. By expressing its entries in terms of Fourier transforms, Corollary 3.30 displayed a formula for the stiffness matrix values that accesses model information not via the operator but via the related symbol, instead. Many examples have shown, that explicit formulas for the symbol exist for many interesting models.

From a numerical perspective, however, new challenges arise. Basis functions with bounded support alleviate numerical integration as they limit the area within the integration range that supporting nodes are distributed over. This is the case for classic hat functions φ_i since $\operatorname{supp}\varphi_i \subseteq [x_i - h, x_i + h]$. Transitioning into Fourier space, however, comes at the cost of numerical integration on an unbounded domain, since the support of $\widehat{\varphi_i}$ is not bounded in $\mathbb{R}$, $\operatorname{supp}\widehat{\varphi_i} = \mathbb{R}$, see Figure 3.7.

As an example, Figure 3.8 displays some stiffness matrix integrands for the Black&Scholes model in Fourier terms. More precisely, we show several integrands of $A \in \mathbb{R}^{N\times N}$ in the representation provided by (3.143) of Corollary 3.34. Each integrand is evaluated for a different value of $j-i$ over three different subintervals taken from the unbounded integration range. In the Fourier approach of calculating the stiffness matrix $A \in \mathbb{R}^{N\times N}$ via the respective symbol, the integrands of A_{ij} would have to be numerically integrated for all $j-i \in \{-(N-1), \dots, -1, 0, 1, \dots, N-1\}$. The larger $|j-i|$, however, the more severe the numerical challenges for evaluating the integrand, as Figure 3.8 demonstrates. All integrands illustrated therein decay rather slowly. Additionally, oscillations increase in $|j-i|$. In combination, these two observations seriously threaten a numerically reliable evaluation of the integral. With increasing values of $|j-i|$, the oscillations of the integrand accelerate and the number of necessary supporting points for accurate integration soars. In this toy example of the Black&Scholes model, pointing out the challenging integration of the stiffness matrix integrand for large values of $|j-i|$ might not be very convincing, since we know the stiffness matrix entries to be equal to zero for $|j-i| > 1$. For Lévy jump models, however, the stiffness matrix is in general fully populated and these oscillations have to be dealt with, indeed. In the following section

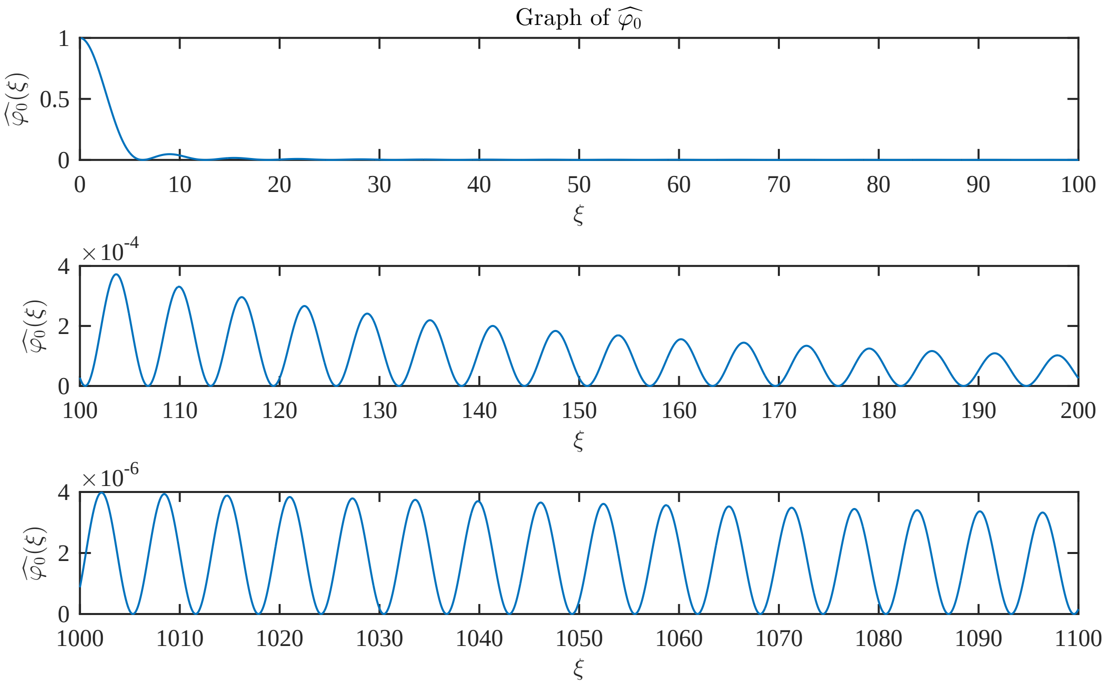

Figure 3.7 Graph of $\widehat{\varphi_0}$, the Fourier transform of the hat function φ_0 centered over the origin, evaluated over three subintervals of $\mathbb{R}^+$. The mesh is chosen with $h = 1$. The oscillations and the rather slow decay to zero complicate numerical integration with high accuracy requirements considerably when $\widehat{\varphi_0}$ is involved.

we investigate the influence of inaccurately calculated stiffness matrix entries onto the accuracy of option prices.

3.4.2 An accuracy study of the stiffness matrix

Using the classic hat functions as basis functions we thus have to accept that severe numerically challenges are attached to the computation of the $2N - 1$ entries of the stiffness matrix $A \in \mathbb{R}^{N\times N}$ via the Fourier approach of Corollary 3.30 due to heavily oscillating integrands. Investigating how material these challenges are, we conduct an empirical study of the propagation of integration errors in the stiffness matrix and their influence on the accuracy of the derived option prices. We have already performed a similar study of this kind in Gaß and Glau (2014) wherein the results are presented in more detail. We choose the Black&Scholes model parametrized by $\sigma = 0.2$ modeling price movements of a stock in a market with interest rate $r = 0.01$, where we price a put option with strike $K = 1$ and maturities $T \in [0, 3]$ for current values of the stock

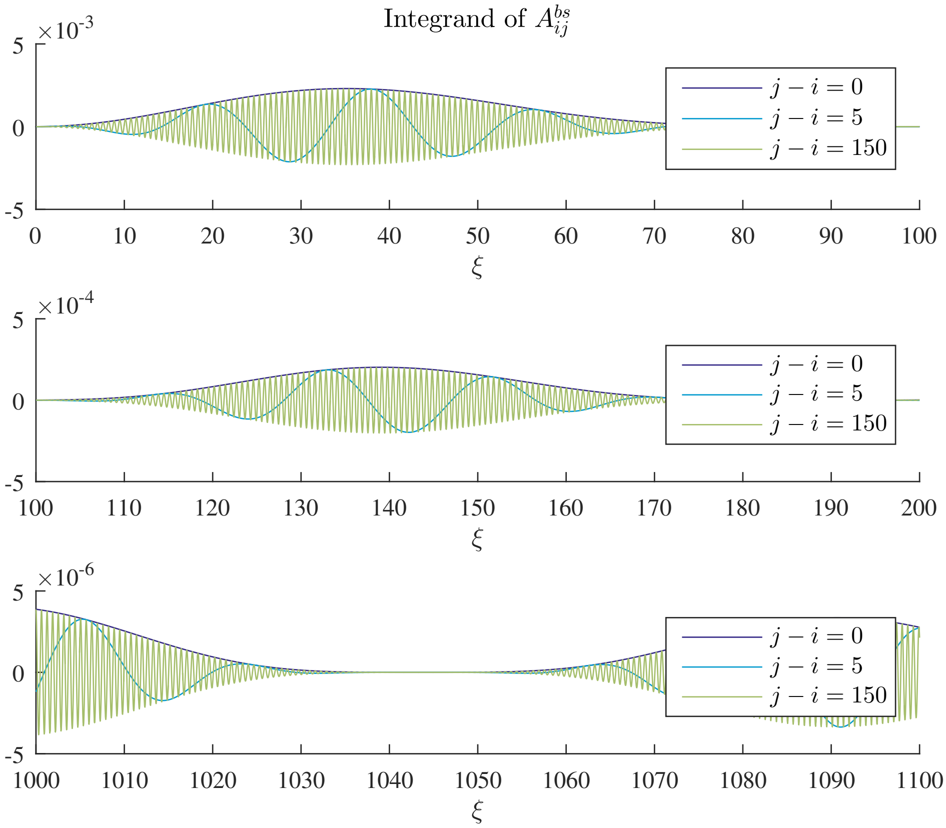

Figure 3.8 The first integrand of A_{ij} in (3.143) for several values of $j-i$. The grid of the hat functions spans the interval $[-5,5]$ with 150 equidistantly spaced inner nodes and grid fineness $h = 0.0662$. A Black&Scholes solution on this grid would thus be represented by the weighted sum of 150 hat functions. We observe that oscillations of the integrand increase in the value of $|j-i|$.

$S_0 \in [S_{\min}, S_{\max}]$ with $S_{\min} = 0.01$ and $S_{\max} = 10$. We set the number of involved FEM hat functions to $N = 150$, resulting in a mesh with 150 inner grid nodes and mesh fineness $h = 0.0464$. We know the mass matrix of the Black&Scholes model to be

$$M = \frac{h}{6} \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 \\ 1 & 4 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 4 & 1 \\ 0 & \cdots & 0 & 1 & 4 \end{pmatrix} \in \mathbb{R}^{N \times N},$$

and the stiffness matrix to be given by

$$A = A^{\mathrm{bs}} = A^{(1)} + A^{(2)} + rM \in \mathbb{R}^{N \times N}, \tag{3.144}$$

where

$$A^{(1)} = \frac{1}{2}\left(r - \frac{\sigma^2}{2}\right) \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 0 & -1 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}, \quad A^{(2)} = \frac{\sigma^2}{2} \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

With these matrices we set up a theta scheme, $\theta = 0.5$, and derive Black&Scholes put option prices. The resulting pricing surface is depicted in Figure 3.9. Since we can solve the integrals determining the entries of the stiffness matrix $A \in \mathbb{R}^{N \times N}$ explicitly in the case of the Black&Scholes model, we know their true value and can simulate how the resulting pricing surface is affected by inaccuracies that might occur when these integrals are solved numerically, instead. To this extent we take the correct stiffness matrix given by (3.144) and distort each of its entries randomly at different positions $D \in \mathbb{N}$ after the decimal point by adding $\varepsilon_i^D = 10^{-(D-1)}\varepsilon_i$ with random $\varepsilon_i \in (-1, 1)$ for $i \in \{-(N-1), \ldots, -1, 0, 1, \ldots, (N-1)\}$ onto the (side) diagonal i of Matrix A. Each individual (side) diagonal of the original stiffness matrix is thus affected evenly, keeping the Toeplitz structure of the matrix intact. Since the value of A_{ij} is only determined by the value of $j - i$, this distortion mimics the influence that integration inaccuracies would have.

So, for $D \in \mathbb{N}$ we define the distorted stiffness matrix by

$$A^D_{\mathrm{distort}} = A + \varepsilon^D \in \mathbb{R}^{N \times N}, \tag{3.145}$$

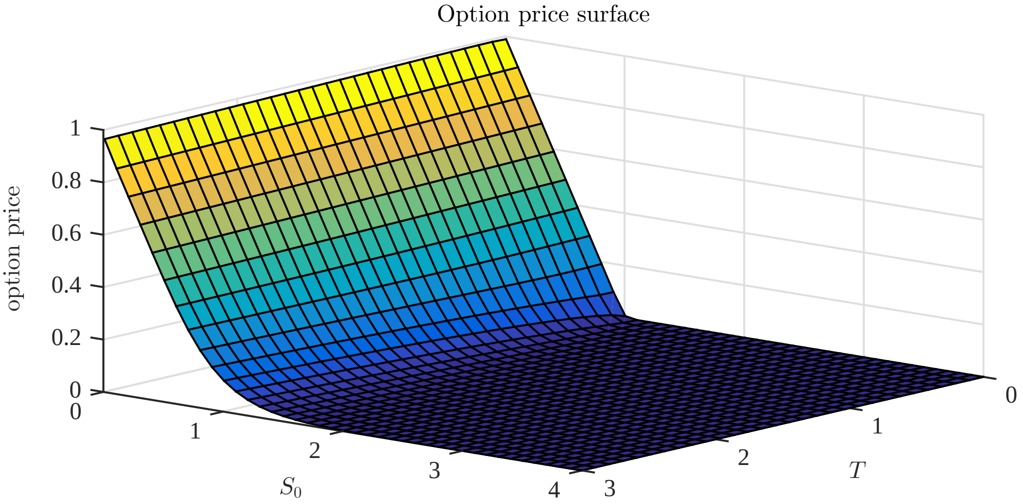

Figure 3.9 Pricing surface of a put option with strike $K = 1$ in the Black&Scholes model with parameter $\sigma = 0.2$ and interest rate $r = 0.01$. The space grid consists of $N = 150$ equidistant inner nodes with mesh fineness $h = 0.0464$. Only a part from the whole surface spanning from $S_{\text{min}} = 0.01$ to $S_{\text{max}} = 10$ that prices were computed for is shown. The considered maturities range from $T_{\text{min}} = 0$ to $T_{\text{max}} = 3$.

with

$$\varepsilon^D = 10^{-(D-1)} \begin{pmatrix} \varepsilon_0 & \varepsilon_1 & \varepsilon_2 & \cdots & \cdots & \cdots & \varepsilon_{N-1} \\ \varepsilon_{-1} & \varepsilon_0 & \varepsilon_1 & \ddots & \ddots & \ddots & \vdots \\ \varepsilon_{-2} & \varepsilon_{-1} & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \varepsilon_1 & \varepsilon_2 \\ \vdots & \ddots & \ddots & \ddots & \varepsilon_{-1} & \varepsilon_0 & \varepsilon_1 \\ \varepsilon_{-(N-1)} & \cdots & \cdots & \cdots & \varepsilon_{-2} & \varepsilon_{-1} & \varepsilon_0 \end{pmatrix} \in \mathbb{R}^{N\times N},$$

with uniformly distributed $\varepsilon_i \in (-1,1)$, $i \in \{-(N-1),\dots,-1,0,1,\dots,(N-1)\}$, that are drawn independently from each other. Using these distorted stiffness matrices A^D_{distort} for different values $D \in \mathbb{N}$, we derive again price surfaces of the put option in the Black&Scholes model and compare the difference between the prices coming from the distorted stiffness matrix $A^D_{\text{distort}} \in \mathbb{R}^{N\times N}$ to the prices from the intact stiffness matrix $A \in \mathbb{R}^{N\times N}$. The results are shown in Figure 3.10. We observe that the absolute price differences decrease almost linearly in D. An accuracy of $D = 3$ corresponds to integration results that are exact up to the third digit after the decimal point. Pricing resulting from stiffness matrices computed with such a low integration accuracy are unacceptable. The respective pricing errors observable in the top left corner of Figure 3.10 indicate

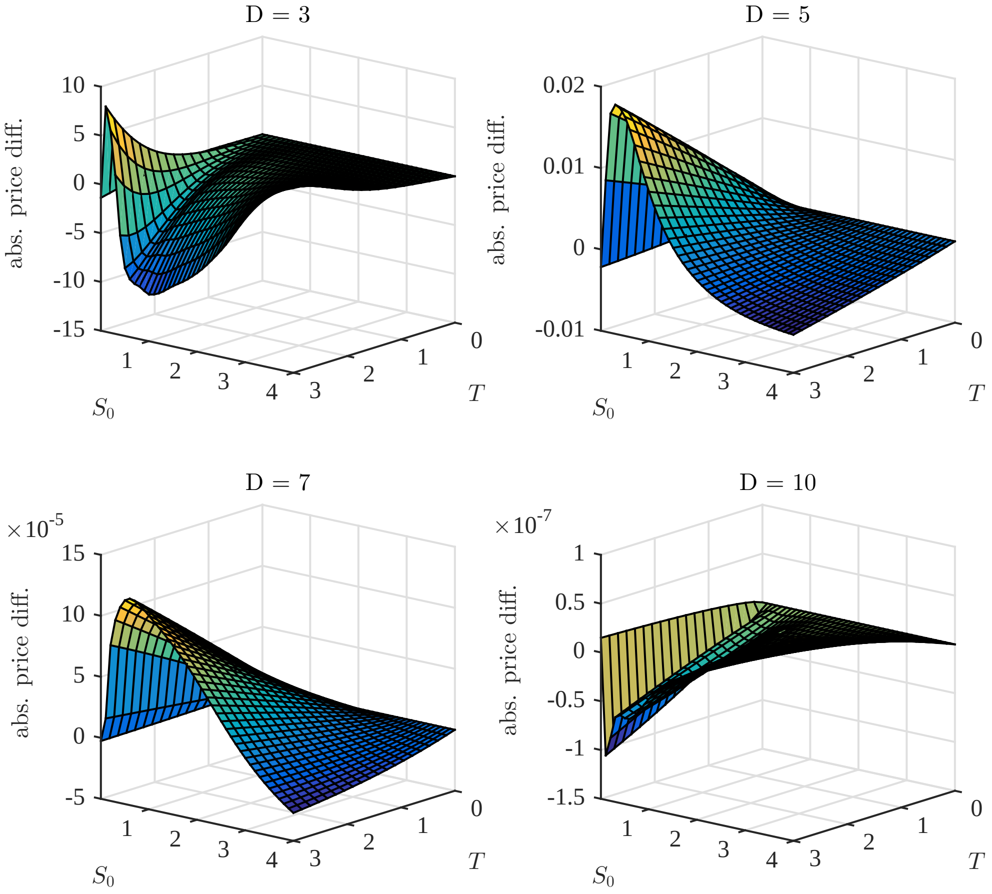

Figure 3.10 Absolute price differences resulting from a distortion of the stiffness matrix A. True and distorted prices describe the market value of a put option in the Black&Scholes model parametrized equivalently to the setting of Figure 3.9. We compare the price surfaces coming from a theta scheme using the stiffness matrix A given by (3.144) to the respective pricing surface when A is replaced by A^D_{distort}, the distorted version of A as defined in (3.145), for different values of $D \in \mathbb{N}$. The influence of the distortion decreases in D.

relative errors of several hundred percent points. With more precise integration results, the error decreases in D until highly appealing pricing results are achieved for $D = 7$ and beyond. The magnitude of the pricing error resulting from a distorted stiffness matrix emphasizes the necessity of being able to derive the stiffness matrix entries as accurately as possible. This poses a serious challenge to the numerical integration routines that have to handle strongly oscillating and slowly decaying integrands which we have seen in Figure 3.8. Yet, this problem of numerical integration of oscillating integrands has drawn attention by research for a long time. One example for an integration routine of approximating the integral

$$V_1 = \int_a^b f(x) \cos(c\, x)\, \mathrm{d}x, \qquad a \leq b \in \mathbb{R}, \quad c \in \mathbb{R} \tag{3.146}$$

is so called Filon's formula, see Abramowitz and Stegun (2014) for details. Unfortunately, Filon's approach focuses on the oscillation alone while lacking an emphasis on the integration of decaying functions. Consequently, $b < \infty$ is required which thus rules out an immediate application of the approach for our purposes, where coming back to our Black&Scholes model example expressions of the form

$$V_2 = \int_0^\infty \frac{g(x)}{x^k}\, \mathrm{d}x, \qquad 2 \leq k \in \mathbb{N} \tag{3.147}$$

for oscillating functions g in the sense that $\exists p > 0$ such that $g(x) = g(x+p)$ for all $x \in \mathbb{R}^+$ are considered. In Appendix A and Lemma A.2 therein, we present an integration algorithm for expressions of the form (3.147) tailor-made for the integration of that special class of decaying functions exhibiting the oscillatory behavior we observed above. Numerical experiments study the approximation power of the algorithm in detail.

Yet, stiffness matrix integrals in general can not be cast in terms of expression (3.147). In some cases, a periodic behavior of the nominator is missing, in others the order of decay is not equal to an integer value. In these cases, again individual integration algorithms would be required which is exactly what the symbol method tries to avoid. Therefore, in the following section we take a different approach to arrive at stiffness matrix integrals that allow a feasible numerical evaluation.

3.4.3 New choices for the basis functions

Previously we had presented a Finite Element implementation for pricing European plain vanilla options in the Merton model using the well known classic hat functions as basis functions. As we have seen, the existence of a jump part with Lévy measure F in the operator $\mathcal{A}$ renders the derivation of the stiffness matrix numerically challenging. While the Merton model still allows quasi-explicit formulas for the stiffness matrix entries, this is in general no longer the case when more involved Lévy jump models are considered. Therefore we analyzed the possibility of accessing the jump information in the Fourier space, instead. Then, the model information is represented by the symbol instead of the operator, a quantity that is available in closed form in many cases. As a negative consequence of this shift into the Fourier space, however, we now have to integrate terms involving the Fourier transform of the considered FEM basis functions. In the case of classic hat functions, this translates into the necessity of integrating slowly decaying, heavily oscillating integrands. Classic hat functions therefore appear hardly compatible to the symbol method approach. Let us therefore investigate two alternative choices for FEM basis functions.

3.4.3.1 Mollified hat functions

Hat functions are piecewise linear functions. While being continuous they are not continuously differentiable everywhere and thus lack smoothness on an elementary level already. This lack of smoothness translates into a slow decay of their Fourier transform, compare Remark 2.8. A fast decay of the Fourier transform, however, is one of the crucial features that basis functions need to possess in order to become eligible in a symbol method based FEM implementation.

Due to its lack of smoothness, the classic hat function is thus ruled out as a FEM basis function candidate in such an implementation and needs to be replaced by an alternative. It is well known, however, that convolution with a smooth function has a smoothing effect on the function that the convolution is applied to. Our first basis function alternative will therefore be a classic hat function smoothed by convolution.

Definition 3.36 (Mollifier)
A smooth function $m \in C^\infty(\mathbb{R}^d)$, $m : \mathbb{R}^d \to \mathbb{R}$ is called mollifier, *if it fulfills*

i) $\int_{\mathbb{R}^d} m(x)\,\mathrm{d}x = 1$,

ii) $\lim\limits_{\varepsilon\to 0} m^\varepsilon(x) = \lim\limits_{\varepsilon\to 0} \frac{1}{\varepsilon^d} m\left(\frac{x}{\varepsilon}\right) = \delta(x)$, *where δ is the Dirac delta function and*

iii) m has compact support, $m \in C_0^\infty(\mathbb{R}^d)$.

Convoluting certain functions f with a mollifier m results in very smooth functions $f * m$ in the sense of the following lemma.

Lemma 3.37 (Mollifying a function)
Let $m \in C^\infty(\mathbb{R})$ be a univariate mollifier and $f \in C_0^0(\mathbb{R})$ a continuous function with compact support. Then the mollfied f, denoted by $f * m$, is infinitely smooth, $f * m \in C^\infty(\mathbb{R})$.

Proof
The claim is a direct consequence from Theorem E.25 in Schilling (2005). □

Example 3.38 (Standard mollifier)
A standard example of a mollifier $m : \mathbb{R} \to \mathbb{R}_0^+$ is given by

$$m(x) = \begin{cases} \frac{1}{C} \exp\left(-\frac{1}{1-|x|^2}\right), & |x| < 1, \\ 0, & \textit{otherwise}, \end{cases} \tag{3.148}$$

with the normalization constant defined by $C = \int_{\mathbb{R}} m(x)\,\mathrm{d}x$.

Let us investigate, how the standard mollifier of Example 3.38 operates on both smooth and non-smooth functions. We define

$$\begin{aligned} f_1 &: x \mapsto \mathbb{1}_{|x|<2}, \\ f_2 &: x \mapsto \frac{1}{6}(x+3)\,\mathbb{1}_{|x|\leq 3}, \\ f_3 &: x \mapsto \varphi_0^{h=1}(x), \\ f_4 &: x \mapsto m(x), \end{aligned} \tag{3.149}$$

so f_1 is a piecewise constant function, f_2 is a piecewise linear function, f_3 is the classic hat function centered over the origin as defined in (3.60) and f_4 is the mollifier of Example 3.38, itself. We apply the standard mollifier m defined in (3.148) of Example 3.38 to each of these functions by convolution. Figure 3.11 shows the graph of each f_i, $i \in \{1,2,3,4\}$, together with $f_i{*}m$, the convolution of that function with the standard mollifier. The smoothing effect is clear to see.

Mollifying functions has a smoothing effect on them. By Remark 2.8, smoothness of a function translates into decay rates of its Fourier transform. Lemma 3.30 presented a method to derive stiffness matrix entries in Fourier space. In the respective formula, the Fourier transform of the basis functions was needed. When hat functions are used as basis functions, however, we face numerical challenges since the Fourier transforms of hat functions oscillate heavily and decay rather slowly. Hat functions smoothed by mollifiers thus appear as interesting candidates to replace the classic hat functions as basis functions in a Finite Element implementation.

Before we can test the suitability of mollified hat functions as basis functions, however, we want to control the influence of the mollifier on functions it is applied to. Simply applying m to the hat function might distort it too strongly. After all, in Figure 3.11 the mollified hat function is hardly distinguishable from the mollified mollifier.

Remark 3.39 (The mollication parameter $\varepsilon > 0$)
Let $m : \mathbb{R}^d \to \mathbb{R}$ be a mollifier in the sense of Definition 3.36. Define for $\varepsilon > 0$

$$m^\varepsilon(x) = \frac{1}{\varepsilon^d} m\left(\frac{x}{\varepsilon}\right), \qquad \forall x \in \mathbb{R}^d.$$

We call ε the mollification parameter *of m. The function m^ε is still a mollifier. The parameter $\varepsilon > 0$ regulates the smoothing influence on the function that the mollifier is applied to. For decreasing values of ε the smoothing influence decreases, for increasing values of ε, the smoothing influence increases.*

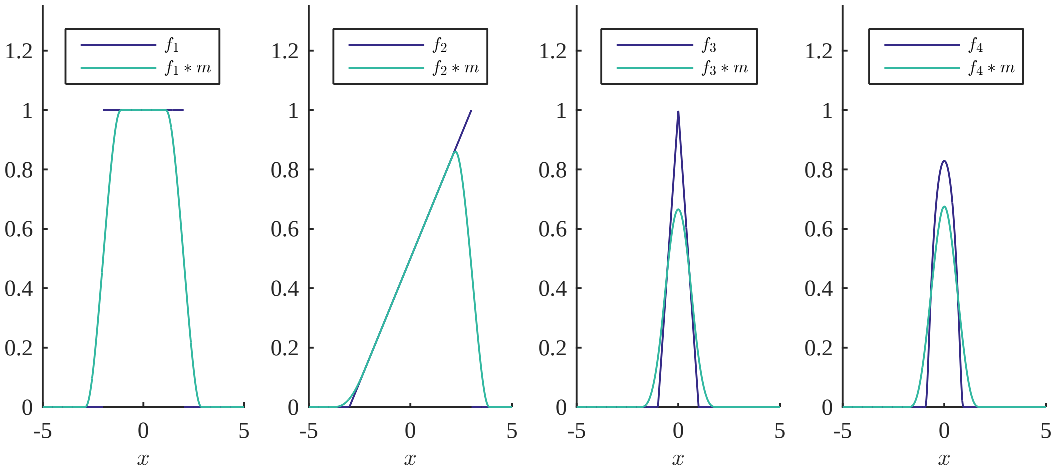

Figure 3.11 The effect of the classic mollifier defined in Example 3.38 on four exemplary functions f_i, $i \in \{1, 2, 3, 4\}$, defined in (3.149). The first two functions are not even continuous, the third one is not differentiable. After mollification, however, they all appear smoothed. Note two interesting observations. The mollifier leaves piecewise linear function parts unchanged when they are long enough (f_1, f_2). At the same time, it might further mollify functions that are already smooth (f_4).

Introducing the mollification parameter ε of Remark 3.39 we gain control over the mollification influence. In choosing $\varepsilon > 0$ smaller, the mollified function gravitates towards its untreated counterpart. Both are identical in the limit, as the following lemma shows.

Lemma 3.40 (Convergence of mollified functions)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous. Let m be a mollifier in the strict sense of Definition 3.36 with support in the unit ball, $\operatorname{supp} m \subseteq B_1^{\|\cdot\|}(0)$ *with respect to some norm. Then $f * m^\varepsilon \to f$ uniformly as $\varepsilon \to 0$ on any compact subset $K \subset \mathbb{R}^d$.*

Proof
The proof is taken from Loftin (2010), see also (Showalter, 2010, Chapter II, Lemma 1.2). By assumption, $K \subset \mathbb{R}^d$ is compact. Therefore, there exists $r > 0$ such that $K \subset \overline{B_r^{\|\cdot\|}(0)}$. The continuous function f is uniformly continuous on the compact set $\overline{B_{r+1}^{\|\cdot\|}(0)}$. Choose $\widetilde{\varepsilon} > 0$. There exists $\delta > 0$ such that for $z, w \in \overline{B_{r+1}^{\|\cdot\|}(0)}$ we have with $\|z - w\| < \delta$ also $|f(z) - f(w)| < \widetilde{\varepsilon}$. Now choose $\varepsilon \in (0, \min\{1, \delta\})$.

Let $x \in K \subset \overline{B_r^{\|\cdot\|}(0)}$. Then,

$$\begin{aligned} |(f * m^\varepsilon)(x) - f(x)| &= \left| \int_{\mathbb{R}^d} f(x-y) m^\varepsilon(y) \,\mathrm{d}y - f(x) \right| \\ &= \left| \int_{\mathbb{R}^d} f(x-y) m^\varepsilon(y) \,\mathrm{d}y - \int_{\mathbb{R}^d} f(x) m^\varepsilon(y) \,\mathrm{d}y \right| \\ &\leq \int_{\mathbb{R}^d} |f(x-y) - f(y)| \, m^\varepsilon(y) \,\mathrm{d}y. \end{aligned} \tag{3.150}$$

Since $\operatorname{supp} m = B_1(0)$ and $m^\varepsilon = \frac{1}{\varepsilon} m(\cdot/\varepsilon)$, $\operatorname{supp} m^\varepsilon = B_\varepsilon^{\|\cdot\|}(0)$. Thus, continuing in (3.150) we get

$$\begin{aligned} \int_{\mathbb{R}^d} |f(x-y) - f(y)| \, m^\varepsilon(y) \,\mathrm{d}y &= \int_{\overline{B_\varepsilon^{\|\cdot\|}(0)}} |f(x-y) - f(y)| \, m^\varepsilon(y) \,\mathrm{d}y \qquad (3.151) \\ &< \int_{\overline{B_\varepsilon^{\|\cdot\|}(0)}} \widetilde{\varepsilon} m^\varepsilon(y) \,\mathrm{d}y \\ &= \widetilde{\varepsilon}, \end{aligned}$$

which proves the claim. □

The mollification parameter ε and the claim of Lemma 3.40 are powerful tools in smoothing the nondifferentiable hat functions. Before the smoothed functions can be deployed, however, we need to derive their Fourier transform.

The Fourier transform of the convolution of two integrable functions is given by the product of the two individual Fourier transforms as Property iii) in Lemma 2.4 shows. In theory, this provides the link from using smoothed hat functions as basis functions to the numerical derivation of the stiffness matrix entries. The Fourier transform of the classic mollifier, however, is not known in closed form. Its numerical evaluation is thus challenging, especially when integration of the mollifier is concerned. Recently, Johnson (2015) has expanded on the issue of evaluating $\widehat{m}$ approximately, emphasizing the numerical difficulties involved.

Classic mollifiers or the standard mollifier of Example 3.38 at least thus don't suit our needs. We therefore mollify with a different class of functions that display very similar mollification effects. Following Proposition and Definition 2.14 in Alt (2011) we introduce the definition of a Dirac sequence.

Definition 3.41 (Dirac sequence)
We call a sequence $(\widetilde{m}_k)_{k\in\mathbb{N}}$, $\widetilde{m}_k \in L^1(\mathbb{R}^d)$ for all $k \in \mathbb{N}$, a Dirac sequence, *if*

i) $\widetilde{m}_k \geq 0$, $\forall k \in \mathbb{N}$,

ii) $\int_{\mathbb{R}^d} \widetilde{m}_k(x) \,\mathrm{d}x = 1$, *and* *iii) if for all $\varrho > 0$ we have the convergence*

$$\int_{\mathbb{R}^d \setminus B_\varrho(0)} \widetilde{m}_k(x)\,\mathrm{d}x \to 0,$$

for $k \to \infty$.

Again by Proposition and Definition 2.14 in Alt (2011) we have the following remark.

Remark 3.42 (Dirac ε)
Let $\widetilde{m} \in L^1(\mathbb{R}^d)$ with

$$\widetilde{m} \geq 0 \text{ and } \int_{\mathbb{R}^d} \widetilde{m}(x)\,\mathrm{d}x = 1. \tag{3.152}$$

Analogously to Remark 3.39 define

$$\widetilde{m}^\varepsilon = \frac{1}{\varepsilon^d}\widetilde{m}\left(\frac{\cdot}{\varepsilon}\right). \tag{3.153}$$

Then for each $\varrho > 0$ we have

$$\int_{\mathbb{R}^d} \widetilde{m}^\varepsilon(x)\,\mathrm{d}x = 1 \text{ and } \int_{\mathbb{R}^d \setminus B_\varrho(0)} \widetilde{m}^\varepsilon(x)\,\mathrm{d}x \to 0, \tag{3.154}$$

for $\varepsilon \to 0$. Consequently, for each null sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ the sequence $(\widetilde{m}^{\varepsilon_k})_{k\in\mathbb{N}}$ is a Dirac sequence in the sense of Definition 3.41.

Definition 3.41 generalizes the notion of a (positive) mollifier as defined in Definition 3.36. Each sequence of $(m^{\varepsilon_k})_{k\in\mathbb{N}}$, $m^{\varepsilon_k} = \varepsilon_k^{-d} m(\cdot/\varepsilon_k)$, with m a positive mollifier, $m : \mathbb{R} \to \mathbb{R}_0^+$, is a Dirac sequence.

Example 3.43 (A Dirac sequence based on the Normal distribution)
We present an example for a Dirac sequence. Define

$$\widetilde{m}_{Gaussian}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{3.155}$$

Define further

$$\widetilde{m}^\varepsilon_{Gaussian} = \frac{1}{\varepsilon}\widetilde{m}_{Gaussian}\left(\frac{\cdot}{\varepsilon}\right). \tag{3.156}$$

With $(\varepsilon_k)_{k\in\mathbb{N}}$ a null sequence we call $(\widetilde{m}^{\varepsilon_k}_{Gaussian})_{k\in\mathbb{N}}$ a Gaussian Dirac sequence*.*

A Gaussian Dirac sequence as given by Example 3.43 can be used for mollification of (non-smooth) functions, as well. For that matter, we take

$$\widetilde{m}^\varepsilon_{\text{Gaussian}} = \frac{1}{\varepsilon}\widetilde{m}_{\text{Gaussian}}\left(\frac{\cdot}{\varepsilon}\right) \tag{3.157}$$

of Example 3.43 and apply $\widetilde{m}^{\varepsilon}_{\text{Gaussian}}$ to the classic hat function by convolution for different values of $\varepsilon > 0$. As in the case of mollifiers, the value of ε governs the degree of the smoothing effect on the function that $\widetilde{m}^{\varepsilon}_{\text{Gaussian}}$ is applied to. Figure 3.12 shows the results of mollifying classic hat functions using the Dirac sequence of Example 3.43. Due to the smoothing effect of a Dirac sequence, we use the term *mollifier* in this context, as well, even though a Dirac sequence does not necessarily fulfill the requirement of compact support of Definition 3.36.

Corollary 3.44 (Fourier transform of Gaussian mollifier)
The characteristic function of the Gaussian mollifier is known in closed form,

$$\widehat{m^{\varepsilon}_{Gaussian}}(\xi) = \exp\left(-\frac{1}{2}\varepsilon^2\xi^2\right), \tag{3.158}$$

and exhibits exponential decay, which is the reason why this mollifier is especially interesting for our purposes.

Proof
Since $m^{\varepsilon}_{\text{Gaussian}}$ is identical to the density of a normally $\mathcal{N}(0, \varepsilon^2)$ distributed random variable, the claim is a direct consequence of Lemma 2.3. □

Analogously to Lemma 3.40 we also have a convergence result for functions f mollified by a Dirac sequence.

Lemma 3.45 (Convergence of mollification with a Dirac sequence)
Let $1 \leq p < \infty$. Let $f \in L^p(\mathbb{R}^d)$ and $(\widetilde{m}_k)_{k\in\mathbb{N}}$ be a Dirac sequence. Then

$$f * \widetilde{m}_k \to f \tag{3.159}$$

in $L^p(\mathbb{R}^d)$ for $k \to \infty$.

Proof
See the proof of Satz 2.15 in Alt (2011). □

We state an analogous result to Corollary 3.34 with mollified hat functions as basis functions.

Corollary 3.46 (Black&Scholes stiffness matrix with mollified hat functions)
Consider the pricing PDE of the univariate Black&Scholes model, that is a PDE of form (3.52) wherein the operator $\mathcal{A}$ is parametrized following Example 3.11 with $r \geq 0$ and $\sigma > 0$. Consider a numerical FEM solver and assume $N > 0$ mollified hat functions

$$\varphi_i^{\varepsilon} = \varphi_i * \widetilde{m}^{\varepsilon}_{Gaussian}, \qquad \forall i \in \{1, \ldots, N\} \tag{3.160}$$

on an equidistant grid with grid fineness $h > 0$ as basis functions, wherein $\widetilde{m}^{\varepsilon}_{Gaussian}$ denotes the Gaussian mollifier of Example 3.43 with mollification parameter $\varepsilon > 0$.

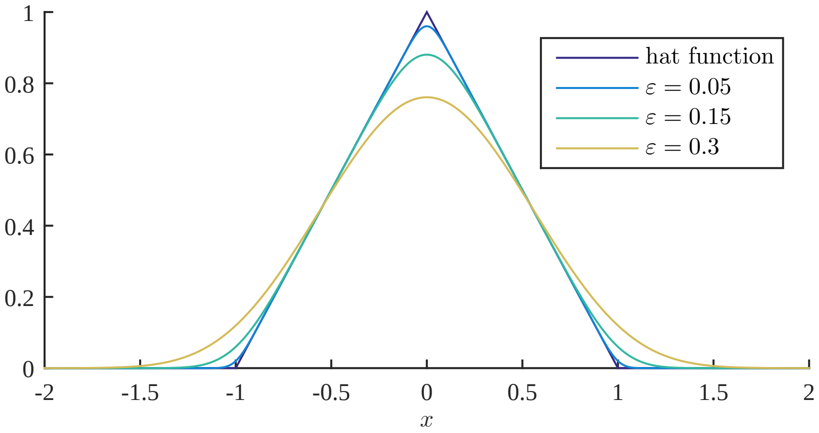

Figure 3.12 A comparison between the classic hat function φ_0 with $h = 1$ as defined in (3.60) and the mollified hat function $\varphi_0^\varepsilon = \varphi_0 * \widetilde{m}^\varepsilon_{\text{Gaussian}}$ for several values of $\varepsilon \in \{0.05, 0.15, 0.3\}$ using the Gaussian mollifier of Example 3.43.

Then the respective stiffness matrix $A \in \mathbb{R}^{N\times N}$ *is given by*

$$\begin{aligned} A_{ij} = \frac{2\sigma^2}{\pi h^2} \int_0^\infty \frac{1}{\xi^2} \cos(\xi h(j-i))(1-\cos(\xi h))^2 e^{-\varepsilon^2\xi^2} \,\mathrm{d}\xi \\ - \frac{4}{\pi h^2}(r - \frac{1}{2}\sigma^2) \int_0^\infty \frac{1}{\xi^3} \sin(\xi h(j-i))(1-\cos(\xi h))^2 e^{-\varepsilon^2\xi^2} \,\mathrm{d}\xi \\ + \frac{4r}{\pi h^2} \int_0^\infty \frac{1}{\xi^4} \cos(\xi h(j-i))(1-\cos(\xi h))^2 e^{-\varepsilon^2\xi^2} \,\mathrm{d}\xi, \end{aligned} \tag{3.161}$$

for all $i, j \in \{1, \dots, N\}$.

Proof
The result is proved analogously to Corollary 3.34, using

$$\widehat{\varphi_0^\varepsilon} = \widehat{\varphi_0}\, \widehat{m}^\varepsilon_{\text{Gaussian}}$$

by property iii) of Lemma 2.4. The Fourier transform $\widehat{m}^\varepsilon_{\text{Gaussian}}$ is given by Corollary 3.44. □

Figure 3.13 displays the integrand in (3.161). The integrand is evaluated on three subintervals of the semi-infinite integration region. The grid setting is identical to the one of Figure 3.8. Instead of classic hat functions their mollified counterparts have been employed as basis functions using the Gaussian mollifier of Example 3.43 as smoothing influence. Even with just a slight mollification influence, $\varepsilon = 0.05h$, the decay of the integrand accelerates. For moderate values of $\varepsilon = 0.3h$ the integrand decays to zero rapidly.

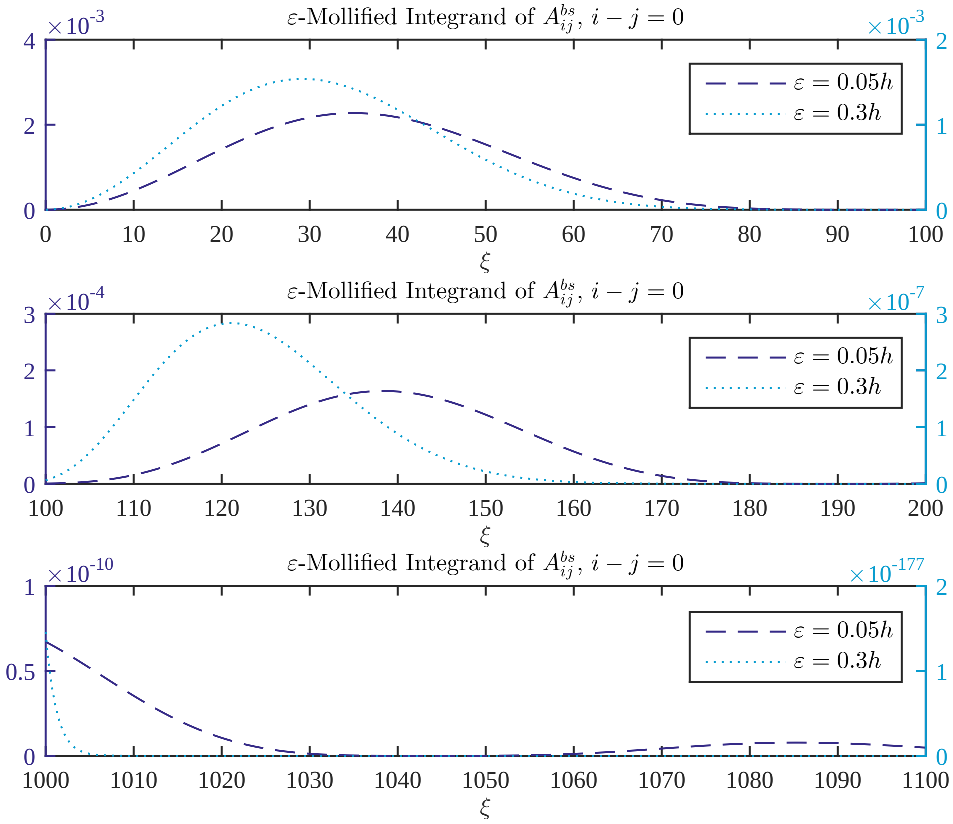

Figure 3.13 The integrand of A_{ij} in (3.161), the stiffness matrix of the Black&Scholes model with mollified hat functions as basis functions for the main diagonal entry, $j-i=0$.

We have implemented the symbol method using mollified hat functions as basis functions for several models and have conducted an empirical order of convergence study that we present at the end of the chapter. The results confirm that mollification is not only theoretically interesting but empirically solves the problem of lacking numerical integrability, as well. Mollifying the hat functions has thus proved to be numerically advantageous.

But let us consider the theoretical consequences, as well. The Fourier transform of a smoothed function decays faster than the Fourier transform of the original function itself. The integrals in the stiffness matrix thus become feasible. In our FEM implementation, the non-smooth hat functions span a finite dimensional subspace of the solution space of the underlying PDE. But smoothing a function changes it. Therefore, smoothing the hat basis functions changes the spaces they span.

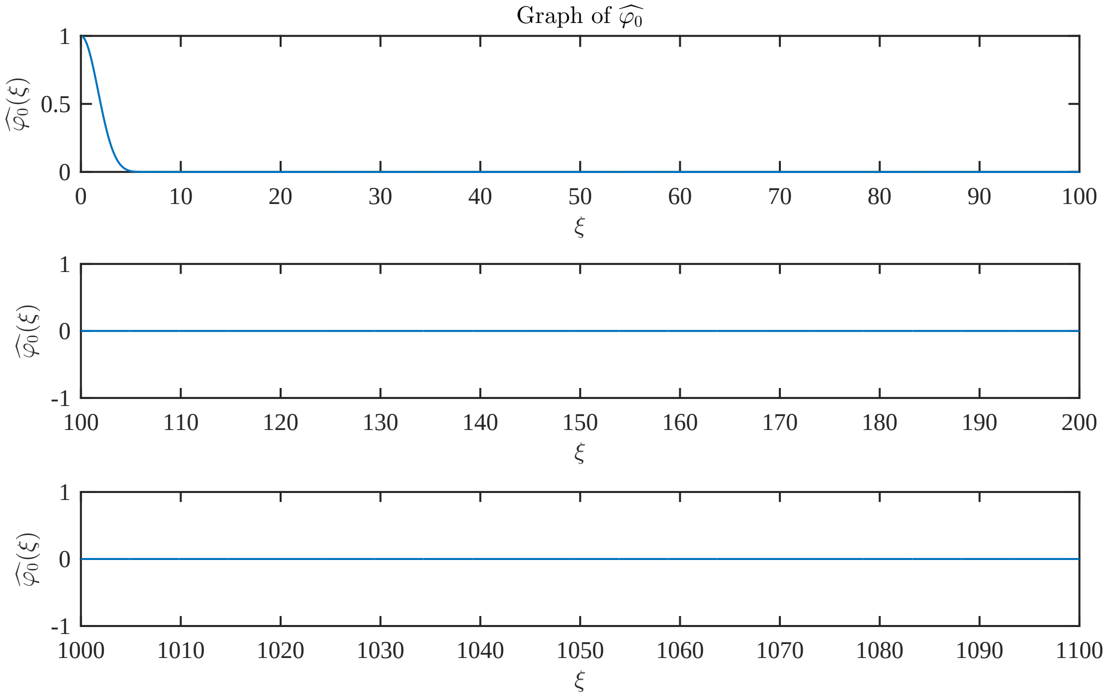

Figure 3.14 Graph of $\widehat{\varphi_0^\varepsilon}$, the Fourier transform of the mollified hat function φ_0^ε centered over the origin, evaluated over three subintervals of $\mathbb{R}^+$. The mesh is chosen with $h = 1$ and the mollification parameter is set to $\varepsilon = 0.3h$. The oscillations and the rather slow decay to zero that we observe in Figure 3.7 where the Fourier transform of the classic hat function is displayed, have vanished completely.

In other words, the discretization in space by mollified hat functions might not fall into the scope of step v) of Section 3.2. Principally, there are two ways to deal with this modification theoretically.

i) Investigate the function spaces that are spanned by mollified hat functions

ii) Treat mollified hat functions as classic hat functions and interpret the contribution of mollification to the algorithm's quantities as a numerical inaccuracy that is addressed by error control methods separately

The appeal of possibility i) lies in the straightforwardness with which the situation would be assessed. The mollification takes effect on the level of the basis functions and modifies them immediately. Investigating the basis properties of the resulting functions from a theoretical point of view would address mollification directly instead of avoiding that confrontation. At the same this, the approach could be cumbersome as the theoretical effect of mollification is rather severe, for example regarding the support of the mollified hat functions which is infinite in theory.

Possibility ii) avoids the issue by viewing the effect of mollification not as a theoretical adaptation but rather as a purely numerical influence, instead. The theoretically expected values of the algorithm's output, for example the stiffness matrix, would thus still be based on the classic hat functions. Independently from the accuracy of the applied numerical integration routine, however, the actual result of the derivations would deviate due to the effect of the mollifier. That difference in the respective quantity would be interpreted as a kind of commonly observed numerical noise that one tries to measure and control. In this regard, the mollification parameter ε becomes the trigger of the numerical disturbance the influence of which can be limited and reduced by shifting ε closer to zero. The actually chosen value of the parameter would then result from a compromise between feasible integrability and desired accuracy of the output. The challenge of this approach would consist in investigating whether this compromise can be reached in all cases of interest. In von Petersdorff and Schwab (2003), the authors provide a framework with which that kind of noise control could be achieved.

Both of these possibilities might stimulate further research to reconcile (mollified) hat functions with the challenges arising from the Fourier aspect of the symbol method. On the other hand, the problem could be avoided in the first place, if we abandoned the hat functions alltogether and turned to already smooth basis functions, instead. This will be the motivation for the next section on splines.

3.4.3.2 Splines

After our analysis of the hat functions we now investigate a second, well-established class of finite element basis function candidates by considering cubic splines. Spline theory is a well-investigated field that applies to a much broader context than we consider here. We refer the reader to Schumaker (2007) for thorough introduction and overview. In this section, we focus on the following facts. Splines are smooth basis functions. Their Fourier transform is accessible and the theory of function spaces they span is well-established. As such, they offer a very interesting alternative to non-differentiable hat functions by avoiding theoretical challenges regarding their deployment in the algorithm while maintaining the promise of numerical feasibility at the same time.

We give the definition of the Irwin-Hall cubic spline that inherits its name from the related probability distribution.

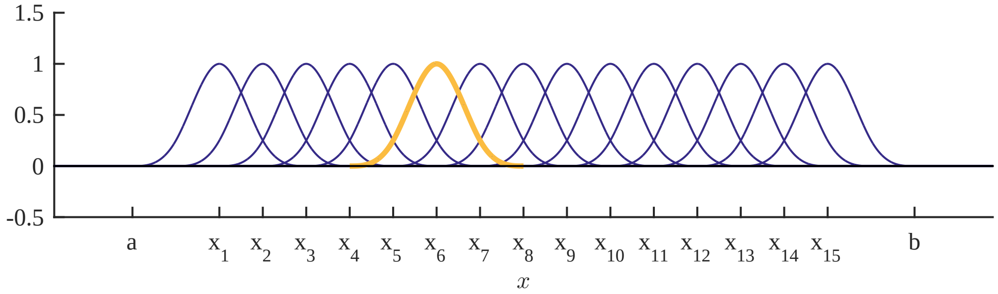

Figure 3.15 A plot of $N = 15$ spline functions φ_i, $i \in \{1, \dots, N\}$, as given by Definition 3.48 on an equidistant grid. For convenience, φ_6 is depicted in orange. Note that in contrast to hat functions, the support of an inner spline function does not only overlap with the supports of two but six neighboring splines.

Definition 3.47 (Irwin-Hall cubic spline)
We define the univariate Irwin-Hall spline $\varphi : \mathbb{R} \to \mathbb{R}^+$ *by*

$$\varphi(x) = \frac{1}{4} \begin{cases} (x+2)^3 & , \ -2 \leq x < -1 \\ 3|x|^3 - 6x^2 + 4 & , \ -1 \leq x < 1 \\ (2-x)^3 & , \ 1 \leq x \leq 2 \\ 0 & , \ \textit{elsewhere} \end{cases} \tag{3.162}$$

for all $x \in \mathbb{R}$. The spline φ has compact support on $[-2, 2]$ and is a cubic spline.

Definition 3.48 (Spline basis functions on an equidistant grid)
Choose $N \in \mathbb{N}$. Assume an equidistant grid $\Omega = \{x_1, \dots, x_N\}$, $x_i \in \mathbb{R}$ for all $i \in \{1, \dots, N\}$, with mesh fineness $h > 0$. Let φ be the Irwin-Hall spline of Definition 3.47. For $i \in \{1, \dots, N\}$ define

$$\varphi_i(x) = \varphi((x - x_i)/h), \qquad \forall x \in \mathbb{R}.$$

We call φ_i the spline basis function associated to node i.

Figure 3.15 displays a set of Irwin-Hall spline basis functions as defined by Definition 3.48. The functions cover a real domain $[a, b] \subset \mathbb{R}$ equidistantly.

For a given equidistant grid consisting of $N \in \mathbb{N}$ grid nodes, the set of associated splines $\varphi_1, \dots, \varphi_N$ given by Definition 3.48 and illustrated by Figure 3.15 constitutes the complete basis which our approximate solution relies on. We are well aware that in the literature often the set of Irwin-Hall basis function splines contains additional functions associated with the fringes of the domain, that the discrete grid spans, for the purpose of providing more flexibility concerning boundary conditions. Yet, this flexibility comes with the numerical cost that those additional basis function again lack elementary smoothness in terms of differentiability and even continuity which disqualifies their deployment for our purposes. Furthermore, this additional flexibility could not even be appreciated in our setup, as we will again transform the PDEs we consider to zero boundary problems, anyway. The issue of omitting spline basis functions that do not belong to the set described by Definition 3.48 has also been investigated theoretically and numerically in Zimmermann (2016). The numerical studies therein confirm that flexibility regarding boundary conditions of Dirichlet or Neumann type or with respect to higher derivatives can be neglected for the options we consider here and thus validate our approach. Thirdly, constraining the set of basis functions in such a way that each function can be transformed into another one by a mere horizontal shift preserves advantageous properties regarding the derivation of the associated Fourier transforms as the following two results demonstrate.

Lemma 3.49 (Fourier transform of the Irwin-Hall spline)
Let φ be the Irwin-Hall cubic spline of Definition 3.47. Then its Fourier transform $\widehat{\varphi}$ is given by

$$\widehat{\varphi}(\xi) = \frac{3}{\xi^4}\left(\cos(2\xi) - 4\cos(\xi) + 3\right) \tag{3.163}$$

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

Proof
Elementary calculations yield

$$\begin{aligned}
4\,\widehat{\varphi}(\xi) &= 4\int_{\mathbb{R}} e^{i\xi x}\varphi(x)\,\mathrm{d}x \\
&= \int_{-2}^{-1}(x+2)^3 e^{i\xi x}\,\mathrm{d}x + \int_{-1}^{1}(3|x|^3 - 6x^2 + 4)e^{i\xi x}\,\mathrm{d}x + \int_{1}^{2}(2-x)^3 e^{i\xi x}\,\mathrm{d}x \\
&= 2\int_{-2}^{-1}(x+2)^3\cos(\xi x)\,\mathrm{d}x + 2\int_{0}^{1}(3x^3 - 6x^2 + 4)\cos(\xi x)\,\mathrm{d}x.
\end{aligned}$$

Standard integration rules lead to

$$\begin{aligned}
4\,\widehat{\varphi}(\xi) &= \frac{2}{\xi^4}\left[\xi(x+2)\left(\xi^2(x+2)^2 - 6\right)\sin(\xi x) + 3\left(\xi^2(x+2)^2 - 2\right)\cos(\xi x)\right]_{x=-2}^{x=-1} \\
&\quad + \frac{2}{\xi^4}\left[\xi\left(\xi^2\left(3x^3 - 6x^2 + 4\right) - 18x + 12\right)\sin(\xi x) + 3\left(\xi^2 x(3x-4) - 6\right)\cos(\xi x)\right]_{x=0}^{x=1} \\
&= \frac{2}{\xi^4}\left(3(\xi^2 - 2)\cos(\xi) + 6\cos(2\xi) - 3(\xi^2 + 6)\cos(\xi) + 18\right) \\
&= \frac{2}{\xi^4}\left(-6\cos(\xi) + 6\cos(2\xi) - 18\cos(\xi) + 18\right) \\
&= \frac{12}{\xi^4}\left(\cos(2\xi) - 4\cos(\xi) + 3\right).
\end{aligned}$$

Consequently,

$$\widehat{\varphi}(\xi) = \frac{3}{\xi^4}\left(\cos(2\xi) - 4\cos(\xi) + 3\right),$$

which finishes the proof. □

Corollary 3.50 (Fourier transform of spline basis functions)
Choose $N \in \mathbb{N}$. Assume an equidistant grid $\Omega = \{x_1, \ldots, x_N\}$, $x_i \in \mathbb{R}$ for all $i \in \{1, \ldots, N\}$, with mesh fineness $h > 0$ and let φ_i be the spline basis function associated with node i as defined in Definition 3.48. Its Fourier transform is given by

$$\widehat{\varphi}_i(\xi) = e^{i\xi x_i} \frac{3}{h^3 \xi^4} (\cos(2\xi h) - 4\cos(\xi h) + 3)$$

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

Proof
Denote by φ_0 the scaled spline function centered over the origin,

$$\varphi_0(x) = \varphi(x/h), \tag{3.164}$$

where φ is the Irwin-Hall spline of Definition 3.47. With property ii) of Lemma 2.4 we compute

$$\begin{aligned}\widehat{\varphi_0}(\xi) =& h\widehat{\varphi}(\xi h) \\ =& \frac{3h}{(\xi h)^4} (\cos(2\xi h) - 4\cos(\xi h) + 3) \\ =& \frac{3}{h^3 \xi^4} (\cos(2\xi h) - 4\cos(\xi h) + 3).\end{aligned}$$

Exploiting property i) of Lemma 2.4 shows the claim. □

Figure 3.16 illustrates the decay of the Fourier transform derived by Lemma 3.49 or Corollary 3.50, respectively. Recalling the respective Figure 3.7 where the analogous situation for Fourier transform of the classic hat function had been shown together with Figure 3.14 that display the oscillatory decay of the Fourier transform of the hat function after mollification we observe that the Fourier transform of the Irwin-Hall spline falls in between those two.

Finally, Figure 3.17 provides a visual overview over the Fourier transforms of all three basis function candidates that are the classic hat functions, the mollified hat functions and the cubic splines of Irwin-Hall type. When all three Fourier transforms are displayed together, those of the mollified hat function and the Irwin-Hall splines can hardly be distinguished and appear to attain zero value very quickly, while the oscillations of the Fourier transform of the classic hat function endure over the whole displayed domain. In Remark 2.8 we established a connection between smoothness of a function and the speed of decay of its Fourier transform. Figure 3.17 indeed serves as an impressive reminder.

In the previous section, Corollary 3.46 presented the formula for the stiffness matrix entries in the Black&Scholes model with mollified hat functions as basis functions. The following corollary translates that result to the situation when splines are used as basis functions, instead.

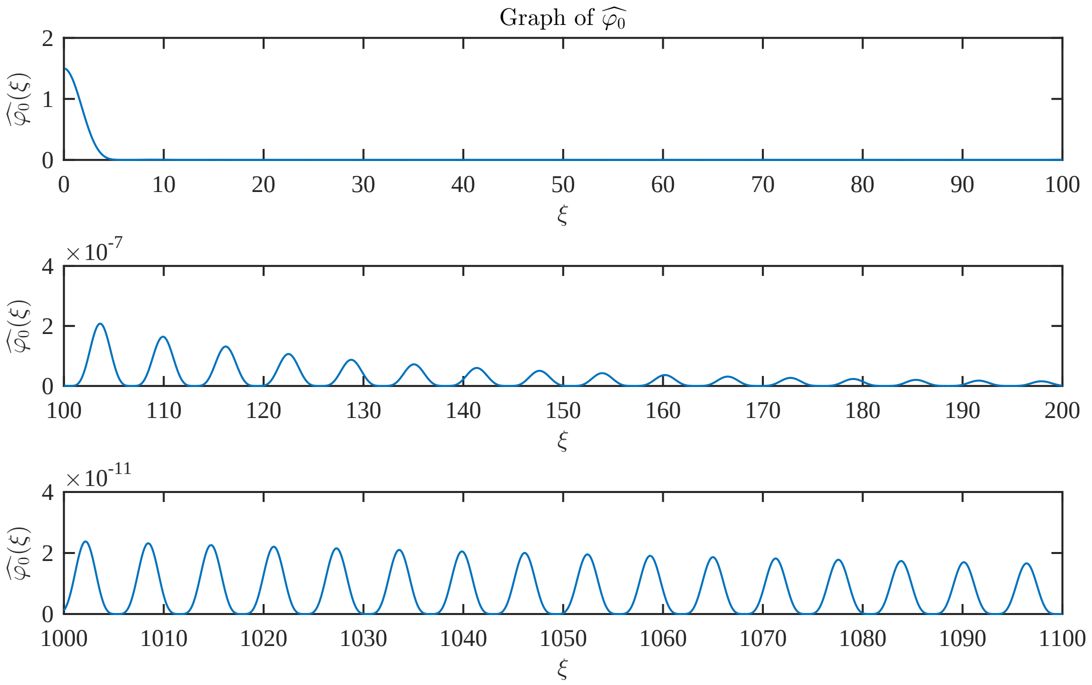

Figure 3.16 Graph of $\widehat{\varphi_0}$, the Fourier transform of the Irwin-Hall spline function φ_0 centered over the origin, evaluated over three subintervals of $\mathbb{R}^+$. The mesh is chosen with $h = 1$. Oscillations and decay rate of the function lie inbetween those displayed in Figure 3.7 and Figure 3.14.

Corollary 3.51 (Black&Scholes mass and stiffness matrix with splines) *Consider the pricing PDE of the univariate Black&Scholes model, that is a PDE of form* (3.52) *wherein the operator $\mathcal{A}$ is parametrized following Example 3.11, with $r \geq 0$ and $\sigma > 0$. Consider a numerical FEM solver and assume $N > 0$ Irwin-Hall spline functions on an equidistant grid with grid fineness $h > 0$ as defined in Definition 3.48 as basis functions. Then the respective mass matrix $M \in \mathbb{R}^{N\times N}$ is given by*

$$M_{lk} = \frac{9}{\pi h^6}\int_0^\infty \cos(\xi(x_k - x_l))\frac{1}{\xi^8}(\cos(2\xi h) - 4\cos(\xi h) + 3)^2\,\mathrm{d}\xi \tag{3.165}$$

and the stiffness matrix $A \in \mathbb{R}^{N\times N}$ computes to

$$\begin{aligned}A_{lk} = {} & \frac{9\sigma^2}{2\pi h^6}\int_0^\infty \frac{1}{\xi^6}\cos(\xi h(k-l))(\cos(2\xi h) - 4\cos(\xi h) + 3)^2\,\mathrm{d}\xi \\ & - \frac{9}{\pi h^6}(r - \frac{1}{2}\sigma^2)\int_0^\infty \frac{1}{\xi^7}\sin(\xi h(k-l))(\cos(2\xi h) - 4\cos(\xi h) + 3)^2\,\mathrm{d}\xi \\ & + rM_{lk},\end{aligned} \tag{3.166}$$

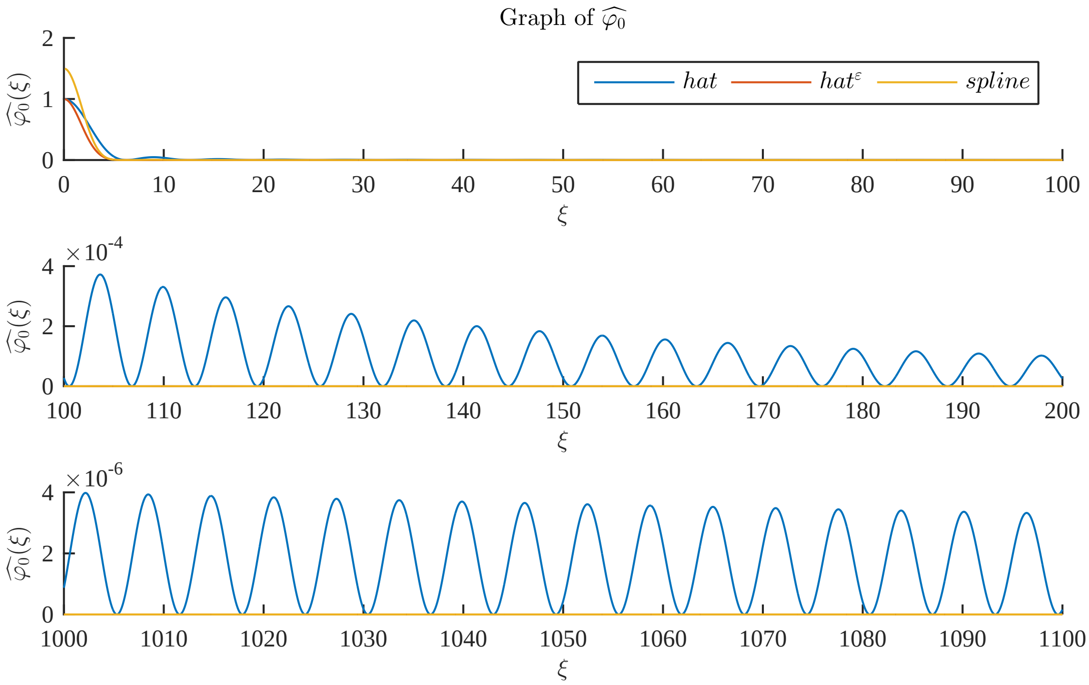

Figure 3.17 Graphs of the Fourier transforms of all basis function candidates presented in this section, evaluated over three subintervals of $\mathbb{R}^+$. The mesh is chosen with $h = 1$, the mollification parameter is again set to $\varepsilon = 0.3h$.

for all $i, j \in \{1, \ldots, N\}$.

Proof
The mass matrix is derived by applying Parseval's identity of Theorem 2.7 and then using the characteristic function of the Irwin-Hall spline derived in Lemma 3.49. The expression for the stiffness matrix entries is derived analogously to the proof of Corollary 3.34. □

We have implemented a symbol method based FEM solver using Irwin-Hall spline functions as basis functions and conducted an empirical order of convergence study. The results are presented in the next section.

3.5 Implementation and numerical results

The previous sections have outlined the necessary consecutive phases in setting up a Finite Element solver for option pricing. In a first step, using the Merton model as an example, the key ingredients of such a solver have been analytically calculated. During the derivation we faced serious limitations regarding the generalizability of that approach. Therefore, in a second step, we introduced the symbol method which considers all components of the FEM solver in Fourier space, instead. There, components are based on the symbol instead of the Lévy measure and become numerically accessible. Many examples of asset models for which the associated symbols exist in analytically closed form have deemed this alternative approach being worthwhile to pursue. At the same time, however, smoothness of the FEM basis functions became a critical issue which ruled out further working with the classic hat functions that we had considered, before. In a third step, we therefore investigated two examples of basis functions that manage to combine smoothness and numerical accessibility. Mollified hat functions and splines were introduced as promising examples to construct a symbol method based FEM solver with.

This section will put that promise to the test. In addition to the hat function based FEM solver for the Merton model we implemented the symbol method for both mollified hats and splines. The FEM solver with hat functions is tailored to the Merton model and can not easily be generalized to other asset models. In stark contrast, the symbol method enjoys the flexibility of being able to easily plug in the symbol of any Lévy model for which it is available in analytically closed form. The model restriction of that first implementation thus disappears. Instead of having to restrict ourselves to the Merton model, we could therefore enhance the model scope of our symbol method based implementation to additionally comprise the NIG and the CGMY model with virtually no additional implementation effort. In this regard, the method impressively underlines its appeal for applications in practice where the suitability of a model might depend on the asset class it is employed for. An institution that needs to maintain pricing routines for several asset classes will thus cherish the flexibility that the symbol method offers, recall Algorithm 1 in this regard which sketches the implementation of a general, symbol method based FEM solver that easily adapts to various models.

Finally, we conduct an empirical order of convergence study. We consider the univariate Merton, CGMY and NIG model and investigate the empirical rates of convergence for the different implementations as Table 3.1 summarizes.

For each model and each implemented basis function type enlisted in Table 3.1 we conduct an empirical order of convergence study using the pricing problem of a call option with strike $K = 1$ as an example, thus considering the payoff function

$$g(x) = \max(e^x - 1, 0). \tag{3.167}$$

In each study we compute FEM prices for N_k basis functions, with

$$N_k = 1 + 2^k, \qquad k \in \{4, \dots, 9\} \tag{3.168}$$

resulting in $N_4 = 17$ basis functions in the most coarse and $N_9 = 513$ basis functions in the most granular case. On each grid, the nodes that basis functions are associated with

Model	**Symbol**	**Parameter choices**		**Implemented basis functions**		
				Hats	Mollified hats	Splines
Merton	Example 3.22	$\sigma = 0.15$, $\beta = 0.2$,	$\alpha = -0.04$, $\lambda = 3$	✓	✓	✓
CGMY	Example 3.32	$C = 0.5$, $M = 27.24$,	$G = 23.78$, $Y = 1.1$		✓	✓
NIG	Example 3.33	$\alpha = 12.26$, $\delta = 0.52$	$\beta = -5.77$,		✓	✓

Table 3.1 An overview of the models considered in the empirical order of convergence analysis and their parametrization. For these models, the symbol method is implemented and tested for both mollified hat functions and splines. In addition, we investigate the empirical convergence rate for the Merton model using classic hat functions as basis functions in a classic implementation disregarding the symbol method. In all models, the constant risk-less interest rate has been set to $r = 0.03$.

are equidistantly spaced from another and the basis functions always span the space interval $\Omega = [-5, 5]$. The time discretization is kept constant with $N_{\text{time grid}} = 2000$ equidistantly spaced time nodes spanning a grid range of two years up until maturity, thus covering a time to maturity interval of

$$[T_1, T_{N_{\text{time}}}], \qquad \text{with } T_1 = 0 \text{ and } T_{N_{\text{time}}} = 2. \tag{3.169}$$

For each $k \in \{4, \dots, 9\}$, the resulting price surface constructed by N_k basis functions in space and $N_{\text{time}} = 2000$ grid points in time is computed. A comparison of these surfaces is drawn to a price surface of most granular structure based on the same type of basis function. We call this most granular surface *true* price surface. It rests on $N_{\text{true}} = N_{11} = 1 + 2^{11} = 2049$ basis functions in space and N_{time} grid points in time spanning the same grid intervals as above, that is $\Omega = [-5, 5]$ in space and $[0, 2]$ in time, respectively. The underlying FEM implementation is thus based on distances h_{true} between grid nodes that basis function are associated with of

$$\begin{aligned} h_{\text{true}}^{\text{(mollified) hat}} &= (5 - (-5))/(2 + 2^{11}) \approx 0.0049, \\ h_{\text{true}}^{\text{splines}} &= (5 - (-5))/(4 + 2^{11}) \approx 0.0049, \\ \Delta t_{\text{true}} &= 2/(2000 - 1) \approx 0.001 \end{aligned} \tag{3.170}$$

in space and time, respectively. Note that all space grids are designed in such a way that the log-strike $\log(K) = 0$ is one of the space nodes. For each model and method and

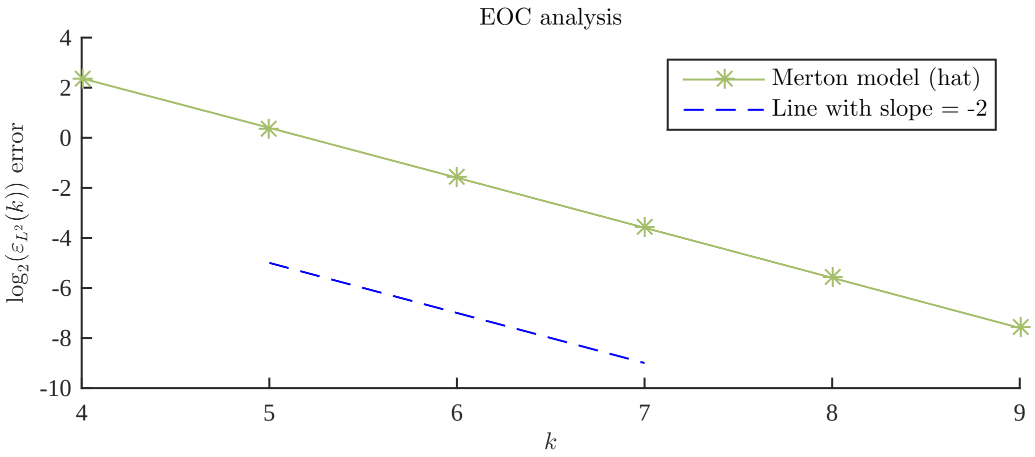

Figure 3.18 Results of the empirical order of convergence study for the Merton model with classic hat functions. Refer to Table 3.1 for the chosen parametrization of the model. Additionally, part of a straight line with (absolute) slope of 2 is depicted and serves as a comparison.

each $k \in \{4, \dots, 9\}$, the (discrete) L^2 error ε_{L^2} is calculated as

$$\varepsilon_{L^2}(k) = \sqrt{\Delta t_{\text{true}} \cdot h_{\text{true}} \cdot \sum_{i=1}^{N_{\text{time}}} \sum_{j=1}^{N_{\text{true}}} \Big(Price_{\text{true}}(i,j) - Price_k(i,j) \Big)^2},$$

wherein $Price_{\text{true}}(i,j)$ is the value of the true pricing surface at space node $j \in \{1, \dots, 1+2^{11}\}$ and time node $i \in \{1, \dots, 2000\}$ and $Price_k(i,j)$ is the respective, linearly interpolated value of the coarser pricing surface with only N_k basis function nodes. Figure 3.18 illustrates the results for the first implementation, the taylormade approach for the Merton model using the classic hat functions as basis functions. Similarly, Figure 3.19 summarizes the results of the six studies of empirical order of convergence in the Merton, the NIG and the CGMY model in a symbol based implementation once using mollified hats and once using splines as basis functions.

In each implementation and for all considered models, the (discrete) L^2 error decays exponentially with rate 2. We thus precisely achieve an empirical rate of convergence that we would theoretically expect, as the upcoming Section 3.6 will explain in detail. This is especially remarkable for the mollified hat functions that are not FEM basis functions in a strict theoretical sense. Our numerical results may thus motivate further research on these functions and their appealing numerical features in Fourier space.

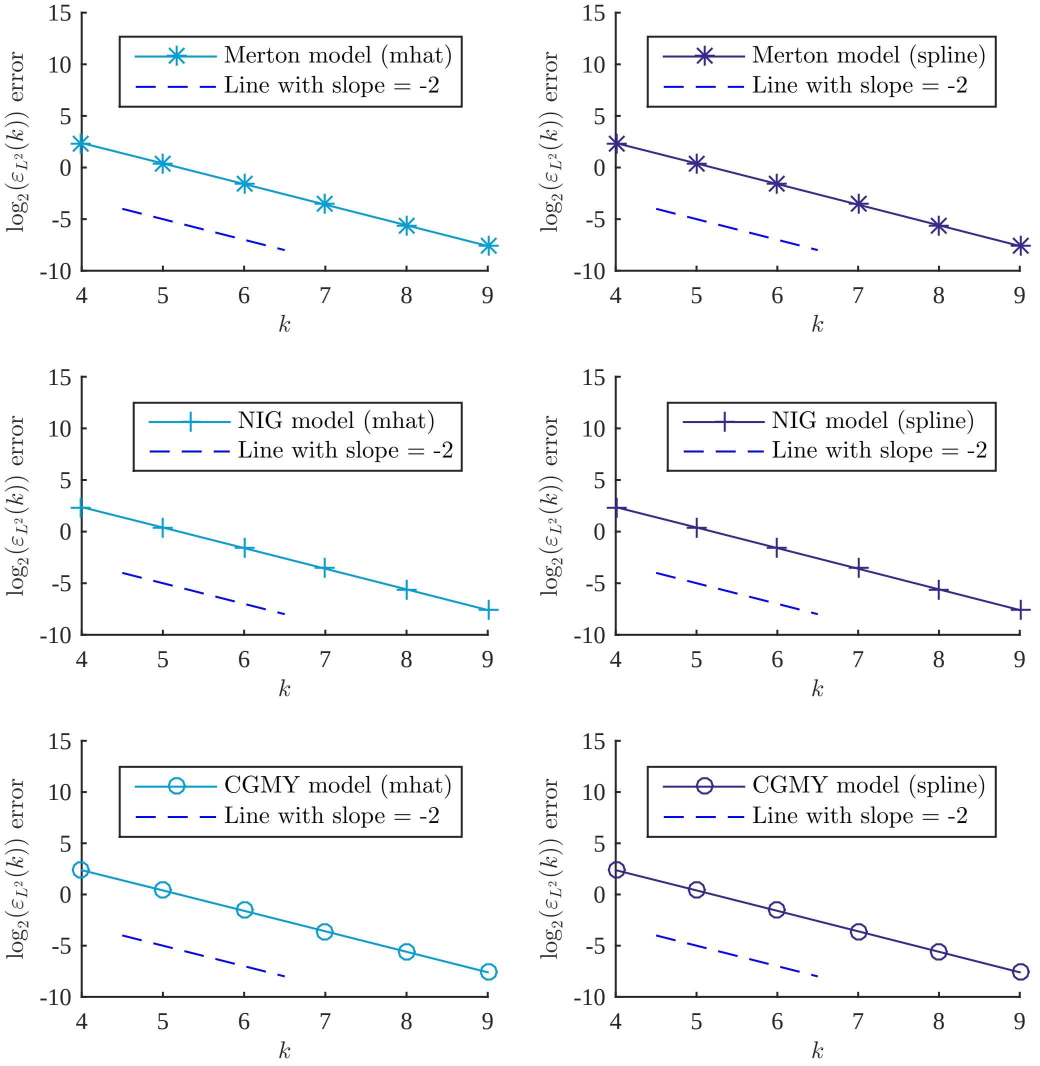

Figure 3.19 Results of the empirical order of convergence study for the Merton, the NIG and the CGMY model using mollified hats (left pictures) and splines (right pictures) as basis functions. All models are parametrized as stated in Table 3.1. Additionally, part of a straight line with (absolute) slope of 2 is depicted in each figure serving as a comparison.

3.6 Stability and convergence analysis

When an approximate finite dimensional solution to a PDE shall be obtained, two questions naturally arise.

i) **Is the numerical scheme deriving the solution stable?**
A numerical scheme is said to be stable, if its solution normed in a certain way is bounded by the equivalently normed right hand side of the scheme and its initial condition up to multiplication with a constant that is independent of the discretization itself. In other words, the solution to a numerically stable scheme is bounded by the input data.

ii) **Does the finite dimensional solution converge to the true solution?**
The precision of a solution to the numerical scheme should increase when the underlying mesh grids in space and time become finer. Only then can we expect error control. In fact, the larger topic of convergence separates into several individual questions. Does the solution converge polynomially or even exponentially? Which rate of convergence does it exhibit? How can the normed difference between the true solution and its approximation be expressed as a function of the mesh parameters?

In this final section of the chapter we want to assess these two questions. The framework that we consider for this task is kept very general. Not only do we consider PDEs with operators independent of time like the Black&Scholes PDE. Instead, our analysis comprises the time-inhomogeneous case, as well, and thus allows the stability and convergence analysis of approximate solutions to time-dependent problems. In this regard, the analysis below extends the work done by von Petersdorff and Schwab (2003) to the time-inhomogeneous case.

The group of PDE problems, however, can not only be separated along their dependence on time. Additionally, all PDE problems can be segregated along a different characteristic of the operator. In the classic existence and uniqueness result on weak solutions to PDEs that Theorem 3.7 presented, the bilinear form $a_t(\cdot,\cdot)$ associated with the operator $\mathcal{A}_t$ needed to satisfy, among other requirements, that there exist constants $\beta > 0$ and $\lambda \geq 0$ independent of t such that

$$a_t(\varphi,\psi) \geq \beta\|\varphi\|_V^2 - \lambda\|\psi\|_H^2, \qquad \forall t \in [0,T] \text{ and } \forall \varphi,\psi \in V. \tag{3.171}$$

Concerning the sophistication of stability and convergence analysis it will mean a significant difference, whether the constant $\lambda \geq 0$ in (3.171) is actually zero or not. For vanishing λ, the associated PDE is called *coercive.* For this case, the results in von Petersdorff and Schwab (2003) provide stability and convergence results for time-homogeneous problems. For nonnegative λ values, the PDE problem is called of *Gårding type.* When PDEs with $\lambda > 0$ are concerned, however, standard approaches to stability and convergence analysis fail and the proofs of these claims become a lot more involved. In finance, this is especially unsatisfactory, since PDE problems in the realm of option pricing are usually of Gårding type and thus require the consideration of the general case of $\lambda > 0$. A popular shortcut to avoid this issue is to transform the original PDE problem of Gårding type into a PDE problem that is coercive and then apply the discretization steps on the basis of the transformed problem. By this approach, however, the link between the original problem and the discrete scheme is lost. Claims regarding stability and convergence only apply to the transformed problem and do not extend to the original pricing PDE, as such. We will illustrate this issue in more detail, later.

In this section, we derive stability and convergence results that apply to PDEs with time-inhomogeneous operator of Gårding type. We begin by extending the results of von Petersdorff and Schwab (2003) to time-inhomogeneous problems focusing on coercive PDEs exlusively. On the basis of these results, we extend the scope of our findings in a second major step to fully general time-inhomogeneous problems of Gårding type.

Consider again the problem of finding solutions $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ to a problem of the form

$$\begin{aligned} \partial_t u + \mathcal{A}_t u &= f, \qquad \text{for almost all } t \in (0,T) \\ u(0) &= g, \end{aligned} \tag{3.172}$$

with $\mathcal{A} = (\mathcal{A}_t)_{t\in[0,T]}$ a time-inhomogeneous operator of order $\alpha_{\mathcal{A}} \in (0,2]$ as introduced in Definition 2.18, a *source term* or *right hand side* $f : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ and an *initial condition* $g : \mathbb{R}^d \to \mathbb{R}$.

The next few definitions introduce the notation that we use throughout the rest of the section.

Definition 3.52 (Semi-discrete weak solution)
Let V, H be separable Hilbert spaces and the dual V^* of V be given that form a Gelfand triplet,

$$V \hookrightarrow H \cong H^* \hookrightarrow V^*$$

and let $V_h \subset V$ be a finite dimensional subspace of V. Let $f \in L^2(0,T;V^*)$. Then we call $u_h \in W^1(0,T;V_h,H)$ a semi-discrete weak solution *to problem* (3.172)*, if for almost every $t \in (0,T)$*

$$(\partial_t u_h(t), v_h)_H + a_t(u_h(t), v_h) = \langle f(t), v_h\rangle_{V^*\times V} \tag{3.173}$$

holds for all $v_h \in V_h$ where the time derivative is understood in the weak sense and a is the bilinear form associated with operator $\mathcal{A}$ and

$$u_h(0) = g_h \tag{3.174}$$

wherein $g_h \in H$ is an approximation of g of problem (3.172)*.*

Note that in the literature, the notion of $g_h \in H$ approximating the initial data g is sometimes interpreted in a stricter and more precise sense thus affecting the initial condition of the semi-discrete weak solution of (3.174). Clearly, when convergence is concerned the interpretation of g_h approximating g becomes more crucial. For our notion of a semi-discrete weak solution as outlined by Definition 3.52, however, we do not yet focus on this issue.

To arrive at a fully discrete problem formulation that is numerically accessible we need to discretize the time horizon $[0, T]$, as well. Instead of accessing time via a continuous variable $t \in [0, T]$, we replace the continuum by $M + 1$ discrete points for some $M \in \mathbb{N}$. The following definition establishes the notion of an equidistant discretization of the time domain $[0, T]$.

Definition 3.53 (Equidistant time grid)
Let $T > 0$. Choose $M \in \mathbb{N}$ and define $\Delta t = T/M$ and $t^m = \Delta t\, m$ for all $m \in \{0, \ldots, M\}$. We call $(T, M, \Delta t)$ an equidistant time discretization*, the set $\{t^0, t^1, \ldots, t^M\}$ the associated* equidistant time grid *and we call Δt the* time stepping size*. Throughout the following, the number of time steps will always be denoted by M and Δt will always be defined as above.*

Definition 3.54 (Fully discrete weak solution)
Let V, H be separable Hilbert spaces and the dual V^* of V be given that form a Gelfand triplet,

$$V \hookrightarrow H \cong H^* \hookrightarrow V^*$$

and let $V_h \subset V$ be a finite dimensional subspace of V. Let $f \in L^2(0, T; V^*)$. Further choose $M \in \mathbb{N}$ and let $\{t^0, \ldots, t^M\}$ be an equidistant time grid with time stepping size Δt. Finally choose $\theta \in [0, 1]$. Then we call $(u_h^m)_{m\in\{0,\ldots,M\}}$, $u_h^m \in V_h$, the fully discrete weak solution *to problem* (3.172)*, if*

$$\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, v_h \right)_H + a^{m+\theta}(u_h^{m+\theta}(t), v_h) = \langle f^{m+\theta}, v_h \rangle_{V^* \times V} \tag{3.175}$$

holds for all $v_h \in V_h$ and for all $m \in \{0, \ldots, M-1\}$ and if

$$u_h^0 = g_h \tag{3.176}$$

wherein $g_h \in H$ is an approximation of g of problem (3.172) *and where we set*

$$u_h^{m+\theta} = \theta u_h^{m+1} + (1-\theta) u_h^m, \tag{3.177}$$

$$f^{m+\theta} = \theta f^{m+1} + (1-\theta) f^m, \tag{3.178}$$

wherein $f^m = f(t^m)$. With a being the bilinear form associated with operator $\mathcal{A}$ we have set

$$a^{m+\theta}(\cdot, \cdot) = a_{\theta t^{m+1} + (1-\theta)t^m}(\cdot, \cdot). \tag{3.179}$$

An iterative relation between u_h^m and u_h^{m+1} for all $m \in \{0, \ldots, M-1\}$ as arising from (3.175) *and* (3.176) *is also called θ* scheme.

Remark 3.55 (On the notation in θ schemes)
In (3.177), (3.178) and (3.179) of the previous definition we introduced a convention that we will repeatedly apply in the following. Beyond the scope of Definition 3.54 we therefore fix the following notation. Let $(T, M, \Delta t)$ be an equidistant time discretization, let $\mathcal{H}$ be a Hilbert space and choose $\theta \in [0, 1]$. With $u^m \in \mathcal{H}$ for all $m \in \{0, \dots, M\}$ we set

$$u^{m+\theta} = \theta u^{m+1} + (1-\theta) u^m, \qquad \forall m \in \{0, \dots, M-1\}. \tag{3.180}$$

With $f \in L^2(0, T; \mathcal{H})$ we set

$$f^m = f(t^m), \qquad \forall m \in \{0, \dots, M\}, \tag{3.181}$$

$$f^{m+\theta} = \theta f^{m+1} + (1-\theta) f^m, \qquad \forall m \in \{0, \dots, M-1\}. \tag{3.182}$$

And with $a : [0, T] \times \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ a time dependent bilinear form we set

$$a^{m+\theta}(v_{\mathcal{H}}, w_{\mathcal{H}}) = a_{\theta t^{m+1} + (1-\theta) t^m}(v_{\mathcal{H}}, w_{\mathcal{H}}), \qquad \forall v_{\mathcal{H}}, w_{\mathcal{H}} \in \mathcal{H}, \tag{3.183}$$

for all $m \in \{0, \dots, M-1\}$.

In contrast to the actual weak solution u to problem (3.172), its fully discretized counterpart $(u_h^m)_{m\in\{0,\dots,M\}}$ is numerically accessible.

With the notion of the fully discrete weak solution of Definition 3.54 we are able to restate the two initial questions from the beginning of this section more precisely. Considering the approximation $(u_h^m)_{m\in\{0,\dots,M\}}$ of u, we ask

1. under which conditions is the approximation $(u_h^m)_{m\in\{0,\dots,M\}}$ numerically stable? And,
2. under which conditions does the approximation $(u_h^m)_{m\in\{0,\dots,M\}}$ converge to u and if so in which sense and how fast?

3.6.1 Assumptions

The answers to these questions depend on the spaces that the weak solution u and its finite-dimensional approximation live in. The necessary conditions for stability and convergence are given below. In the way they are stated, they generalize the set of assumptions required by von Petersdorff and Schwab (2003) for their stability and convergence analysis.

Remark 3.56 (Adding a superscript s to V)
For the error analysis and the derivation of convergence results we assume that the space V that the solution space of the weak solution $u \in W^1(0, T; V, H)$ is built on provides a certain smoothness, denoted by a positive real value $s \in \mathbb{R}^+$. More precisely, the space V will always be a Sobolev space with index $s \in \mathbb{R}^+$, see Definitions 2.25 or 2.26 for two exemplary definitions of such spaces. From here on, we therefore add the superscript s to V, thus writing V^s instead of V, and its finite dimensional subspaces, by writing analogously V_h^s instead of V_h, to represent the smoothness of the respective space.

Remark 3.57 (Space discretization by polynomials)
We discretize the space V^s to receive a finite dimensional subspace $V_h^s \subset V^s$ using for example piecewise polynomials of degree $p \geq 0$. Using spaces spanned by the classic hat functions as in our implementation in Section 3.3 we have $p = 1$, while the Irwin-Hall splines implemented thereafter result in $p = 3$. The higher $p \geq 0$, the more smoothness the spanned space V_h^s provides.

Assumption 3.A (General approximation property)
Let $V_h^s \subset V^s$ be a finite dimensional subspace. Let $s \leq t$ with $0 \leq s \leq \alpha_{\mathcal{A}}/2 \leq t$. We assume that for all $u \in V^t$ there exists $u_h \in V_h^s$ such that

$$\|u - u_h\|_{V^s} \leq C_\Upsilon \, \Upsilon(h, t, s, u) \tag{3.184}$$

for some positive constant $C_\Upsilon > 0$ and some function Υ both independent of s and t with $\Upsilon(h, t, s, u) \to 0$ in $h \to 0$ when $t > s$.

Assumption 3.B (Inverse property)
We assume that there is a constant $C_{IP} > 0$ independent of $h > 0$ such that with $0 \leq s \leq \alpha_{\mathcal{A}}/2$ we have

$$\|u_h\|_{V^s} \leq C_{IP} \, h^{-s} \|u_h\|_H \tag{3.185}$$

for all $u_h \in V_h^s$.

Assumption 3.C (Approximation property of the projector)
We assume that there exists a bounded linear projector

$$P_h : V^s \to V_h^s \tag{3.186}$$

for which the approximation property (3.184) *of Assumption 3.A holds when u_h is replaced by $P_h(u)$ for all $u \in V^s$.*

Example 3.58 (The setting of von Petersdorff and Schwab (2003))
We present a first example of a specific instance for Assumption 3.A. In von Petersdorff and Schwab (2003), the authors consider the space

$$\mathcal{H}^s(\Omega) = \begin{cases} V = \tilde{H}^{\alpha_{\mathcal{A}}/2}(\Omega), & s = \alpha_{\mathcal{A}}/2 \\ V \cap H^s(\Omega), & s > \alpha_{\mathcal{A}}/2, \end{cases}$$

for $\Omega \subset \mathbb{R}^d$ a bounded domain with Lipschitz boundary $\Gamma = \partial\Omega$ and $\alpha_{\mathcal{A}} \in [0, 2]$ the order of the possibly nonlocal operator $\mathcal{A}$ in problem (3.172) *with the space $\tilde{H}^s(\Omega)$ defined as*

$$\tilde{H}^s(\Omega) = \{u|_\Omega \,\big|\, u \in H^s(\mathbb{R}^d), \, u|_{\mathbb{R}^d \setminus \Omega} = 0\}.$$

The discrete approximation $(u_h^m)_{m \in \{0,\dots,M\}}$ lives in $V_h \in \{V_h\}_{h>0} \subset V$, a finite dimensional subspace based on piecewise polynomials of degree $p \geq 0$, see Section 3.4.1 in von Petersdorff and Schwab (2003) for details.

In this setting, Assumption 3.A assumes that for all $u \in \mathcal{H}^t(\Omega)$ with $t \geq \alpha_{\mathcal{A}}/2$ there exists a $u_h \in V_h$ such that for $0 \leq s \leq \alpha_{\mathcal{A}}/2$ and $\alpha_{\mathcal{A}}/2 \leq t \leq p+1$

$$\|u - u_h\|_{\tilde{H}^s(\Omega)} \leq c\, h^{t-s} \|u\|_{\mathcal{H}^t(\Omega)} \tag{3.187}$$

for some $c > 0$. The general function Υ of Assumption 3.A is thus defined as

$$\Upsilon(h,t,s,u) = h^{t-s}\|u\|_{\mathcal{H}^t(\Omega)} \tag{3.188}$$

for all $u \in \mathcal{H}^t(\Omega)$ and $C_\Upsilon = c$.

Example 3.59 (The setting of Hilber et al. (2008))
In Hilber et al. (2008), the authors conduct a convergence analysis for the time-homogeneous case for coercive operators. They implicitly assume Assumption 3.A by taking V to be a Sobolev-type space with smoothness index r, in the sense that

$$V = \widetilde{H}^r, \qquad \widetilde{H}^0 = H = L^2. \tag{3.189}$$

As they state, r depends on the order of the operator. They assume the solution u to problem (3.172) to possess higher regularity in space, $u(t) \in \mathcal{H}^s \subset \widetilde{H}^r$ for $t \in (0,T]$, where they assume $\mathcal{H}^s$ again to be a Sobolev-type space with smoothness index s.

We present a final example for Sobolev spaces with integer index that originates from the results of da Veiga et al. (2014) as presented in Zimmermann (2016).

Example 3.60 (Results from da Veiga et al. (2014))
Let $\Omega \subset \mathbb{R}$ be a bounded domain. Let further $s,t \in \mathbb{N}_0$ with $0 \leq s \leq t \leq p+1$ with $p \in \mathbb{N}$. Consider the space of B-splines with degree p spanned over a partition $\widetilde{\Delta}$, confer da Veiga et al. (2014) for details. Then there exists a projector

$$\Pi_{p,\widetilde{\Delta}} : H^{p+1}(\Omega) \to S_p(\widetilde{\Delta}) \tag{3.190}$$

from the Sobolev space $H^{p+1}(\Omega)$ onto $S_p(\widetilde{\Delta})$, the space spanned by B-splines with degree p such that with

$$V = H^s(\Omega) \tag{3.191}$$

there exists a constant $C(p) > 0$ such that for all $u \in H^t(\Omega)$

$$\left\| u - \Pi_{p,\widetilde{\Delta}} u \right\|_{H^s(\Omega)} \leq C h^{t-s} \|u\|_{H^t(\Omega)} \tag{3.192}$$

holds.

The results of Example 3.60 also extend to non-integer Sobolev spaces. Consider for example Theorem 2.3.2 in Roop (2006) or similarly Theorem 7.2 in Ervin and Roop (2007) for a verification of the approximation property in a fractional Sobolev space setting. Moreover we refer to Takacs and Takacs (2015), Karkulik and Melenk (2015) and Du et al. (2013) for further results and examples on the abstract approximation property of Assumption 3.A.

Additionally, consider Definition 1.9 of the Ph.D. thesis of Schötzau (1999), where a projector Π_l^r is defined. In Theorem 1.19 and Corollary 1.20, the author then derives approximation results for that projector. These results present themselves in the spirit of Assumption 3.A and hold for integer and non-integer Sobolev spaces, respectively.

Attached to the spaces H, V^s and V_h^s that the (approximate) solutions live in we consider the norms

$$\begin{aligned} \|u\| &:= \|u\|_H, && \text{for } u \in H, \\ \|f\|_{V_h^{s*}} &:= \sup_{\substack{v_h \in V_h^s \\ v_h \neq 0}} \frac{(f, v_h)}{\|v_h\|_{V^s}}, && \text{for } f \in V^{s*}, \end{aligned} \tag{3.193}$$

Remark 3.61 (An estimate for $\|\cdot\|_{V_h^{s*}}$)
From the definition of $\|\cdot\|_{V_h^{s*}}$ in (3.193), we immediately get for $f \in V^{s*}$ the estimate

$$\|f\|_{V_h^{s*}} = \sup_{\substack{v_h \in V_h^s \\ v_h \neq 0}} \frac{(f, v_h)}{\|v_h\|_{V^s}} \leq \sup_{\substack{v_h \in V_h^s \\ v_h \neq 0}} \frac{\sqrt{(f,f)}\sqrt{(v_h, v_h)}}{\|v_h\|_{V^s}} = \|f\|_H \sup_{\substack{v_h \in V_h^s \\ v_h \neq 0}} \frac{\|v_h\|_H}{\|v_h\|_{V^s}} \leq \|f\|_H, \tag{3.194}$$

which we state here for later use.

We will also need the constant Λ, defined by

$$\Lambda = \sup_{\substack{v_h \in V_h^s \\ v_h \neq 0}} \frac{\|v_h\|_H^2}{\|v_h\|_{V_h^{s*}}^2}. \tag{3.195}$$

Remark 3.62 (On Λ)
Given $h > 0$ and the respective finite dimensional space $V_h^s \subset V^s$, the constant Λ defined in (3.195) is finite due to the fact that all norms involved are norms restricted to finite dimensional spaces and in finite dimensional spaces all norms are equivalent. From this, Λ being finite follows immediately. As a consequence, however, Λ depends on h and thus on the dimension of the spaces involved,

$$\Lambda = \Lambda(h).$$

Yet, Λ is not necessarily bounded in h and thus in the limit, $h \to 0$, not necessarily finite, anymore.

In Definition 3.54 we introduced the notion of a Theta scheme as an iterative relation between approximate solutions to the original problem 3.172 that live on an (equidistant) time grid in finite dimensional space. We restate this structure here for later reference.

Theta Scheme 3.63 (Fully discretized θ scheme)
Let $(u_h^m)_{m\in\{0,\ldots,M\}}$ be the fully discretized solution to problem (3.172)*, $u_h^m \in V_h^s \subset V^s$. The $u_h^m \in V_h^s$, $m \in \{0,\ldots,M\}$, solve the θ scheme*

$$\begin{aligned}\left(\frac{u_h^{m+1}-u_h^m}{\Delta t}, v_h\right) + a^{m+\theta}(u_h^{m+\theta}, v_h) &= (f^{m+\theta}, v_h),\\ u_h^0 &= g_h,\end{aligned} \tag{3.196}$$

for all $v_h \in V_h^s$.

In general, $\theta \in [0,1]$ and $M \in \mathbb{N}$ or rather $\Delta t = \Delta t(M)$ of the time discretization $(T, M, \Delta t)$ that the solution of Theta Scheme 3.63 rests on can not be chosen independently from another. The variable M serves as a measure of the fineness of the discretization $(T, M, \Delta t)$ in time. The value of θ controls the degree of implicitness of the scheme (3.196). With $\theta = 1$, the element u_h^{m+1} appears twice in the scheme (3.196) which is then called *fully implicit.* With $\theta = 0$ the element u_h^{m+1} appears only once and thus the scheme is called *fully explicit.* So called *semi-explicit* schemes are those with $\theta \in (0,1)$ with the Crank-Nicolson scheme as the most prominent example ($\theta = \frac{1}{2}$). As we will see later, in case that $\theta \leq \frac{1}{2}$, convergence and stability lemmas and theorems only grant their claims if Δt is small enough. Conditions of that sort are always called *time stepping conditions.*

For the accuracy of an approximate solution $(u_h^m)_{m\in\{0,\ldots,M\}}$ to problem (3.172), the approximation quality of g_h, the approximate of the initial value g plays a vital role.

Assumption 3.D (Quasi-optimality of the initial condition)
The initial condition $u_h^0 = g_h$ in Theta Scheme 3.63 initiates the iterative relation of (3.196)*. This initial value $g_h \in H$ is in general only an approximation of the initial value $g \in H$ of the original problem* (3.172)*. We assume* quasi optimality *in H of the initial condition of the scheme in the sense that $\exists C_I > 0$ such that*

$$\|g - g_h\|_H \leq C_I \inf_{v_h \in V_h^s} \|g - v_h\|_H \tag{3.197}$$

holds.

The framework that we have presented in the previous section has prepared us for proving the theorems and lemmas that follow. Therein, properties of the original problem (3.172) will be specified and consequent features of the approximate fully discretized solution will be rigorously derived.

In this section, we prove stability and convergence results for θ schemes that yield solutions to fully discretized PIDEs in subspaces $V_h \subset V$. In the analysis of stability and convergence, we distinguish between two major classes. First, we consider PIDEs where the operator $\mathcal{A}$ induces a bilinear form that is both continuous and coercive.
Secondly, we consider the more general class of PIDEs where the operator $\mathcal{A}$ induces a

bilinear form that is still continuous but only of Gårding type. A bilinear form of Gårding type generalizes the notion of coercivity and complicates the derivation of stability and convergence results considerably.

3.6.2 Results for continuous and coercive bilinear forms

In this subsection we consider PIDEs with time dependent operator $\mathcal{A}$ that induces a family of bilinear forms $a_t(\cdot,\cdot) : V^s \times V^s \to \mathbb{R}$ for each $t \in [0,T]$ that is continuous and coercive uniformly in time. We will generalize the second restriction in the next subsection. Our results on PIDEs with operators of this first kind thus generalize the results of von Petersdorff and Schwab (2003) in that they allow for time-dependence of the operator and thus admit more flexibility in the model choice. Furthermore we keep track of the involved constants in all estimates and make them explicit wherever possible.

Definition 3.64 (Continuity)
A bilinear form $a_{\cdot}(\cdot,\cdot) : [0,T] \times V^s \times V^s \to \mathbb{R}$ is called continuous *uniformly in time with respect to V^s, if there exists $\alpha \in \mathbb{R}^+$ independent of t such that*

$$|a_t(u,v)| \leq \alpha \|u\|_{V^s} \|v\|_{V^s} \tag{3.198}$$

holds for all $u, v \in V^s$ and for all $t \in [0,T]$. We call such an α a continuity constant *of the bilinear form a.*

Definition 3.65 (Coercivity)
A bilinear form $a_{\cdot}(\cdot,\cdot) : [0,T] \times V^s \times V^s \to \mathbb{R}$ is called coercive *uniformly in time with respect to V^s, if there exists $\beta \in \mathbb{R}^+$ independent of t such that*

$$a_t(u,u) \geq \beta \|u\|_{V^s}^2 \tag{3.199}$$

holds for all $u \in V^s$ and for all $t \in [0,T]$. We call such a β a coercivity constant *of the bilinear form a.*

Remark 3.66 (Energy norm)
A bilinear form a that is both continuous uniformly in time in the sense of Definition 3.64 and coercive uniformly in time in the sense of Definition 3.65 induces a norm $\|\cdot\|_{a_t} = \sqrt{a_t(\cdot,\cdot)}$ on V^s for each $t \in [0,T]$ that is equivalent to the norm of V^s, since

$$\sqrt{\beta}\|u\|_{V^s} \leq \|u\|_{a_t} \leq \sqrt{\alpha}\|u\|_{V^s},$$

for all $u \in V^s$ wherein α and β are the time independent constants from Definition 3.64 and Definition 3.65, respectively. The norm $\|\cdot\|_{a_t}$ is called enery norm *of $a_t(\cdot,\cdot)$.*

Remark 3.67 (On the continuity and coercivity definition)
Note that the definition of continuity by inequality (3.198) *and the definition of coercivity by inequality* (3.199) *comply precisely with the requirements ii) and iii) of Theorem 3.7 for the existence and uniqueness of weak solutions to problems of form* (3.172).

3.6.2.1 Stability of coercive schemes

We derive a stability result regarding a solution to Theta Scheme 3.63 under the assumption of continuity and coercivity of the associated time dependent bilinear form.

Lemma 3.68 (Stability estimate for θ scheme)
Let $a.(\cdot,\cdot)$ be a time dependent bilinear form that is both continuous and coercive uniformly in time with respect to V^s and H. Let $\theta \in [0,1]$ and let $(u_h^m)_{m\in\{0,\dots,M\}}$ be a solution of the associated θ scheme 3.63 on an equidistant time grid $(T, M, \Delta t)$. For $\theta \in \left[\frac{1}{2}, 1\right]$ let

$$\begin{aligned} 0 < C_1 &< 2, \\ C_2 &\geq \frac{1}{\beta(2-C_1)}, \end{aligned}$$

with β the coercivity constant of bilinear form a. For $\theta \in \left[0, \frac{1}{2}\right)$ assume the time stepping size Δt to satisfy the time stepping condition

$$0 < \Delta t < \frac{2\beta}{(1-2\theta)\Lambda\alpha^2}, \tag{3.200}$$

with Λ defined in (3.195)*, define the constant*

$$\mu = (1-2\theta)\,\Lambda\Delta t > 0 \tag{3.201}$$

and let

$$C_1 \in \left(0, 2 - \frac{\mu\alpha^2}{\beta}\right), \tag{3.202}$$

$$C_2 \geq \max\left\{\mu, \frac{(1+\mu\alpha)^2}{(2-C_1)\beta - \mu\alpha^2} + \mu\right\}. \tag{3.203}$$

Then the stability estimate

$$\left\|u_h^M\right\|_H^2 + \Delta t\, C_1 \sum_{m=0}^{M-1} \left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \leq \left\|u_h^0\right\|_H^2 + \Delta t\, C_2 \sum_{m=0}^{M-1} \left\|f^{m+\theta}\right\|_{V_h^{s*}}^2,$$

is satisfied.

Before we prove Lemma 3.68, the following remark argues that the intervals for the constants C_1, C_2 introduced therein are indeed well-defined.

Remark 3.69 (On the constants of Lemma 3.68)
For $\theta \in [0, \frac{1}{2})$ the constant μ is well defined and indeed larger than zero and the set of possible values for C_1, C_2 is non-empty. With Δt chosen according to (3.200) *we have*

$$\frac{\mu\alpha^2}{\beta} = \frac{(1-2\theta)\Lambda\Delta t\alpha^2}{\beta} < \frac{(1-2\theta)\Lambda\frac{2\beta}{(1-2\theta)\Lambda\alpha^2}\alpha^2}{\beta} = 2$$

which admits a non-empty interval of choices for C_1. Since μ is finite and bounded, we have $C_2 < \infty$ if

$$(2 - C_1)\beta > \mu\alpha^2$$

which is the case if $(2 - C_1) > \mu\alpha^2/\beta$ which is true by the interval that C_1 is chosen from.

Proof (of Lemma 3.68)
The proof follows the structure of the proof of Proposition 4.1 by von Petersdorff and Schwab (2003) replacing their norm $\|\cdot\|_*$ by norm $\|\cdot\|_{V_h^{s*}}$ as defined in equation (3.193). At the core of the proof lies verifying that

$$X^m := \|u_h^m\|_H^2 - \left\|u_h^{m+1}\right\|_H^2 - \Delta t\, C_1 \left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + \Delta t\, C_2 \left\|f^{m+\theta}\right\|_{V_h^{s*}}^2 \geq 0, \tag{3.204}$$

for all $m \in \{0, \ldots, M-1\}$, since summing up the X^m, $m \in \{0, \ldots, M-1\}$, then yields

$$\sum_{m=0}^{M-1} X^m = \left\|u_h^0\right\|_H^2 - \left\|u_h^M\right\|_H^2 - \Delta t\, C_1 \sum_{m=0}^{M-1} \left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + \Delta t\, C_2 \sum_{m=0}^{M-1} \left\|f^{m+\theta}\right\|_{V_h^{s*}}^2 \geq 0,$$

from which by simple rearrangement of terms if follows that

$$\left\|u_h^M\right\|_H^2 + \Delta t\, C_1 \sum_{m=0}^{M-1} \left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \leq \left\|u_h^0\right\|_H^2 + \Delta t\, C_2 \sum_{m=0}^{M-1} \left\|f^{m+\theta}\right\|_{V_h^{s*}}^2,$$

which shows the claim.

Fix $m \in \{0, \ldots, M-1\}$ and define

$$\bar{w} = u_h^{m+1} - u_h^m. \tag{3.205}$$

With this definition of $\bar{w}$ the θ scheme 3.63 yields

$$\begin{aligned}(\bar{w}, u_h^{m+\theta}) &= \Delta t \left(-a^{m+\theta}\left(u_h^{m+\theta}, u_h^{m+\theta}\right) + \left(f^{m+\theta}, u_h^{m+\theta}\right)\right) \\ &= \Delta t \left(-\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + \left(f^{m+\theta}, u_h^{m+\theta}\right)\right),\end{aligned} \tag{3.206}$$

The definition of the norm $\|\cdot\|_{V_h^{s*}}$ in (3.193) directly gives the estimate

$$\left(f^{m+\theta}, u_h^{m+\theta}\right) \leq \left\|f^{m+\theta}\right\|_{V_h^{s*}} \left\|u_h^{m+\theta}\right\|_{V^s}. \tag{3.207}$$

Combining (3.206) and (3.207) gives the estimate

$$(\bar{w}, u_h^{m+\theta}) \leq \Delta t \left(-\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + \left\|f^{m+\theta}\right\|_{V_h^{s*}} \left\|u_h^{m+\theta}\right\|_{V^s}\right). \tag{3.208}$$

Using the definition of $\bar{w}$ in (3.205) we get

$$u_h^{m+\theta} = \left(u_h^m + u_h^{m+1}\right)/2 + \left(\theta - \frac{1}{2}\right)\bar{w},$$

which is equivalent to the relation

$$u_h^{m+1} + u_h^m = 2u_h^{m+\theta} - (2\theta - 1)\bar{w}. \tag{3.209}$$

With the definition of $\bar{w}$ and (3.209) we see that

$$\left\|u_h^{m+1}\right\|_H^2 - \|u_h^m\|_H^2 = (u_h^{m+1} - u_h^m, u_h^{m+1} + u_h^m) = \left(\bar{w}, 2u_h^{m+\theta} - (2\theta - 1)\bar{w}\right),$$

such that by changing signs we arrive at

$$\|u_h^m\|_H^2 - \left\|u_h^{m+1}\right\|_H^2 = -2(\bar{w}, u_h^{m+\theta}) + (2\theta - 1)(\bar{w}, \bar{w}). \tag{3.210}$$

Continuing in (3.210), we get by invoking the upper boundary of (3.208) in the first summand that

$$\begin{aligned} -2(\bar{w}, u_h^{m+\theta}) &+ (2\theta - 1)(\bar{w}, \bar{w}) \\ &\geq 2\Delta t\left(\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 - \left\|f^{m+\theta}\right\|_{V_h^{s*}}\left\|u_h^{m+\theta}\right\|_{V^s}\right) + (2\theta - 1)\|\bar{w}\|_H^2. \end{aligned} \tag{3.211}$$

Thus the final estimate for the difference between $\|u_h^m\|_H^2$ and $\left\|u_h^{m+1}\right\|_H^2$ is given by combining (3.210) and (3.211) as

$$\begin{aligned} \|u_h^m\|_H^2 &- \left\|u_h^{m+1}\right\|_H^2 \\ &\geq 2\Delta t\left(\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 - \left\|f^{m+\theta}\right\|_{V_h^{s*}}\left\|u_h^{m+\theta}\right\|_{V^s}\right) + (2\theta - 1)\|\bar{w}\|_H^2. \end{aligned} \tag{3.212}$$

Now we have collected all prerequisites for analyzing X^m of (3.204). Taking its definition and the estimate (3.212) we deduce

$$\begin{aligned} X^m &= \|u_h^m\|_H^2 - \left\|u_h^{m+1}\right\|_H^2 - \Delta t\, C_1\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + \Delta t\, C_2\left\|f^{m+\theta}\right\|_{V_h^{s*}}^2 \\ &\geq 2\Delta t\left(\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 - \left\|f^{m+\theta}\right\|_{V_h^{s*}}\left\|u_h^{m+\theta}\right\|_{V^s}\right) + (2\theta - 1)\|\bar{w}\|_H^2 \\ &\qquad - \Delta t\, C_1\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + \Delta t\, C_2\left\|f^{m+\theta}\right\|_{V_h^{s*}}^2. \end{aligned} \tag{3.213}$$

Collecting terms gives

$$\begin{aligned} X^m &\geq \Delta t(2 - C_1)\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + (2\theta - 1)\|\bar{w}\|_H^2 \\ &\quad - 2\Delta t\left\|f^{m+\theta}\right\|_{V_h^{s*}}\left\|u_h^{m+\theta}\right\|_{V^s} + \Delta t\, C_2\left\|f^{m+\theta}\right\|_{V_h^{s*}}^2 \\ &= (2\theta - 1)\|\bar{w}\|_H^2 \\ &\quad + \Delta t\left[(2 - C_1)\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 + C_2\left\|f^{m+\theta}\right\|_{V_h^{s*}}^2 - 2\left\|f^{m+\theta}\right\|_{V_h^{s*}}\left\|u_h^{m+\theta}\right\|_{V^s}\right]. \end{aligned} \tag{3.214}$$

By assumption, the bilinear form $a_t(\cdot,\cdot)$ is uniformly coercive with coercivity constant β, so

$$\left\|u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 = a^{m+\theta}\left(u_h^{m+\theta}, u_h^{m+\theta}\right) \geq \beta\left\|u_h^{m+\theta}\right\|_{V^s}^2. \tag{3.215}$$

Using the assumption $C_1 < 2$ and inserting (3.215) into (3.214) gives

$$\begin{aligned} X^m \geq{}& (2\theta-1)\|\bar{w}\|_H^2 \\ &+ \Delta t\left[(2-C_1)\beta\left\|u_h^{m+\theta}\right\|_{V^s}^2 + C_2\left\|f^{m+\theta}\right\|_{V_h^{s*}}^2 - 2\left\|f^{m+\theta}\right\|_{V_h^{s*}}\left\|u_h^{m+\theta}\right\|_{V^s}\right]. \end{aligned} \tag{3.216}$$

To proceed we distinguish two cases for $\theta \in [0,1]$.

$\theta \in [\frac{1}{2}, 1]$ So, assume first that $\theta \in \left[\frac{1}{2}, 1\right]$. Then, proceeding from (3.216) gives by the second binomial formula

$$\begin{aligned} X^m \geq (2\theta-1)\|\bar{w}\|_H^2 + \Delta t\Bigg[&\left(\sqrt{(2-C_1)\beta}\left\|u_h^{m+\theta}\right\|_{V^s} - \sqrt{C_2}\left\|f^{m+\theta}\right\|_{V_h^{s*}}\right)^2 \\ &+ 2\left(\sqrt{(2-C_1)\beta C_2} - 1\right)\left\|f^{m+\theta}\right\|_{V_h^{s*}}\left\|u_h^{m+\theta}\right\|_{V^s}\Bigg]. \end{aligned} \tag{3.217}$$

Now,

$$(2-C_1)\beta C_2 \geq 1 \Leftrightarrow C_2 \geq \frac{1}{\beta(2-C_1)}, \tag{3.218}$$

which is true by assumption. By the choice of θ, $(2\theta-1)\|\bar{w}\|_H^2 \geq 0$, and by (3.218) all other summands in (3.217) are nonnegative as well, so

$$X^m \geq 0, \tag{3.219}$$

which proves the claim of the lemma for $\theta \in \left[\frac{1}{2}, 1\right]$.

$\theta \in [0, \frac{1}{2})$ Here, $(2\theta-1) < 0$, which prohibits arguing like above. By θ scheme 3.63, we have

$$(\bar{w}, v_h) = \Delta t\left(-a^{m+\theta}\left(u_h^{m+\theta}, v_h\right) + \left(f^{m+\theta}, v_h\right)\right), \tag{3.220}$$

for all $v_h \in V_h^s$. Consequently,

$$\begin{aligned} \|\bar{w}\|_{V_h^{s*}} &= \sup_{v_h\in V_h}\frac{(\bar{w}, v_h)}{\|v_h\|_{V^s}} \\ &= \sup_{v_h\in V_h}\frac{\Delta t\left(-a^{m+\theta}\left(u_h^{m+\theta}, v_h\right) + \left(f^{m+\theta}, v_h\right)\right)}{\|v_h\|_{V^s}}, \end{aligned} \tag{3.221}$$

which gives

$$\begin{aligned}\|\bar{w}\|_{V_h^{s*}} &\leq \Delta t \left(\sup_{v_h \in V_h} \frac{-a^{m+\theta}\left(u_h^{m+\theta}, v_h\right)}{\|v_h\|_{V^s}} + \sup_{v_h \in V_h} \frac{(f^{m+\theta}, v_h)}{\|v_h\|_{V^s}} \right) \\ &= \Delta t \left(\left\| a^{m+\theta}\left(u_h^{m+\theta}, \cdot\right) \right\|_{V_h^{s*}} + \left\| f^{m+\theta} \right\|_{V_h^{s*}} \right).\end{aligned} \tag{3.222}$$

Clearly, by taking the uniform continuity of $a_t(\cdot,\cdot)$ with respect to $\|\cdot\|_{V^s}$ into account we deduce

$$\begin{aligned}\left\| a^{m+\theta}\left(u_h^{m+\theta}, \cdot\right) \right\|_{V_h^{s*}} &= \sup_{v_h \in V_h} \frac{a^{m+\theta}\left(u_h^{m+\theta}, v_h\right)}{\|v_h\|_{V^s}} \\ &\leq \left\| u_h^{m+\theta} \right\|_{V^s} \sup_{v_h \in V_h} \frac{\alpha \|v_h\|_{V^s}}{\|v_h\|_{V^s}} = \alpha \left\| u_h^{m+\theta} \right\|_{V^s},\end{aligned} \tag{3.223}$$

which we insert into (3.222) to get

$$\|\bar{w}\|_{V_h^{s*}} \leq \Delta t \left(\alpha \left\| u_h^{m+\theta} \right\|_{V^s} + \left\| f^{m+\theta} \right\|_{V_h^{s*}} \right). \tag{3.224}$$

By the definition of Λ in (3.195) we have the „inverse estimate“

$$\|\bar{w}\|_H \leq \sqrt{\Lambda} \|\bar{w}\|_{V_h^{s*}}. \tag{3.225}$$

Assembling our results by combining (3.224) with (3.225) gives

$$\|\bar{w}\|_H \leq \sqrt{\Lambda} \|\bar{w}\|_{V_h^{s*}} \leq \sqrt{\Lambda} \Delta t \left(\alpha \left\| u_h^{m+\theta} \right\|_{V^s} + \left\| f^{m+\theta} \right\|_{V_h^{s*}} \right). \tag{3.226}$$

Finally, we can continue in (3.216) under our assumption that $\theta \in [0, \frac{1}{2})$ by applying result (3.226) to compute

$$\begin{aligned}X^m &\geq (2\theta - 1)\|\bar{w}\|_H^2 \\ &\quad + \Delta t \left[(2 - C_1)\beta \left\| u_h^{m+\theta} \right\|_{V^s}^2 + C_2 \left\| f^{m+\theta} \right\|_{V_h^{s*}}^2 - 2 \left\| f^{m+\theta} \right\|_{V_h^{s*}} \left\| u_h^{m+\theta} \right\|_{V^s} \right] \\ &\geq (2\theta - 1) \left(\sqrt{\Lambda} \Delta t \left(\alpha \left\| u_h^{m+\theta} \right\|_{V^s} + \left\| f^{m+\theta} \right\|_{V_h^{s*}} \right) \right)^2 \\ &\quad + \Delta t \left[(2 - C_1)\beta \left\| u_h^{m+\theta} \right\|_{V^s}^2 + C_2 \left\| f^{m+\theta} \right\|_{V_h^{s*}}^2 - 2 \left\| f^{m+\theta} \right\|_{V_h^{s*}} \left\| u_h^{m+\theta} \right\|_{V^s} \right].\end{aligned} \tag{3.227}$$

Expanding the squared brackets in (3.227) gives

$$\begin{aligned}X^m \geq \Delta t \Bigg[&(2\theta - 1)\Delta t \Lambda \left(\alpha^2 \left\| u_h^{m+\theta} \right\|_{V^s}^2 + 2\alpha \left\| u_h^{m+\theta} \right\|_{V^s} \left\| f^{m+\theta} \right\|_{V_h^{s*}} + \left\| f^{m+\theta} \right\|_{V_h^{s*}}^2 \right) \\ &+ (2 - C_1)\beta \left\| u_h^{m+\theta} \right\|_{V^s}^2 + C_2 \left\| f^{m+\theta} \right\|_{V_h^{s*}}^2 - 2 \left\| f^{m+\theta} \right\|_{V_h^{s*}} \left\| u_h^{m+\theta} \right\|_{V^s} \Bigg].\end{aligned}$$

Collecting terms we derive

$$\begin{aligned} X^m \geq \Delta t \Bigg[& \left[(2-C_1)\beta - (1-2\theta)\Delta t \Lambda \alpha^2\right] \left\| u_h^{m+\theta} \right\|_{V^s}^2 \\ & - 2\left[1 + (1-2\theta)\Delta t \alpha \Lambda\right] \left\| f^{m+\theta} \right\|_{V_h^{s*}} \left\| u_h^{m+\theta} \right\|_{V^s} \\ & + \left[C_2 - (1-2\theta)\Delta t \Lambda\right] \left\| f^{m+\theta} \right\|_{V_h^{s*}}^2 \Bigg]. \end{aligned} \tag{3.228}$$

Recall the definition of μ in (3.201) as

$$\mu = (1-2\theta)\,\Lambda \Delta t \tag{3.229}$$

which turns (3.228) into

$$\begin{aligned} X^m \geq \Delta t \Bigg[& \left[(2-C_1)\beta - \mu \alpha^2\right] \left\| u_h^{m+\theta} \right\|_{V^s}^2 \\ & - 2\left[1 + \mu\alpha\right] \left\| f^{m+\theta} \right\|_{V_h^{s*}} \left\| u_h^{m+\theta} \right\|_{V^s} \\ & + \left[C_2 - \mu\right] \left\| f^{m+\theta} \right\|_{V_h^{s*}}^2 \Bigg]. \end{aligned} \tag{3.230}$$

Define constants

$$\gamma = 1 + \mu\alpha, \tag{3.231}$$

$$\delta = C_2 - \mu, \tag{3.232}$$

$$\kappa = (2-C_1)\beta - \mu\alpha^2. \tag{3.233}$$

Trivially, $\gamma > 0$. We also have $\delta \geq 0$, since $C_2 \geq \mu$ by the first condition for C_2 in (3.203). Furthermore, we have $\kappa > 0$, since

$$C_1 < 2 - \frac{\mu\alpha^2}{\beta}$$

by the upper bound for the open interval of possible values for C_1 according to (3.202). Inserting the definitions of the nonnegative δ and the positive γ and κ into (3.230) and applying the second binomial formula then gives

$$\begin{aligned} X^m \geq \Delta t \Bigg[& \left(\sqrt{\kappa} \left\| u_h^{m+\theta} \right\|_{V^s} - \sqrt{\delta} \left\| f^{m+\theta} \right\|_{V_h^{s*}} \right)^2 \\ & + 2\left(\sqrt{\kappa\delta} - \gamma\right) \left\| f^{m+\theta} \right\|_{V_h^{s*}} \left\| u_h^{m+\theta} \right\|_{V^s} \Bigg]. \end{aligned} \tag{3.234}$$

The lower bound for X^m stated in (3.234) is thus nonnegative, if

$$\sqrt{\kappa\delta} \geq \gamma,$$

which by the nonnegativity of the constants involved is equivalent to

$$\kappa\delta \geq \gamma^2, \tag{3.235}$$

Using the definitions of κ, δ and γ, (3.235) holds, if

$$\left((2 - C_1)\beta - \mu\alpha^2\right)(C_2 - \mu) \geq (1 + \mu\alpha)^2,$$

which is the case, since

$$C_2 \geq \frac{(1 + \mu\alpha)^2}{(2 - C_1)\beta - \mu\alpha^2} + \mu,$$

by the second condition for C_2 in (3.203). Therefore, $X^m \geq 0$ which finishes the proof. □

Remark 3.70 (On the time stepping condition and the inverse property)
Let us have a closer look at the time stepping condition (3.200) *for* $\theta \in \left[0, \frac{1}{2}\right)$ *in Lemma 3.68. Under the inverse property Assumption 3.B we have for all* $w_h \in V_h^s$*,*

$$\|w_h\|_{V_h^{s*}} = \sup_{v_h \in V_h^s} \frac{(w_h, v_h)}{\|v_h\|_{V^s}} \geq \frac{1}{C_{IP}} h^s \sup_{v_h \in V_h^s} \frac{(w_h, v_h)}{\|v_h\|_H} \geq \frac{1}{C_{IP}} h^{\alpha_{\mathcal{A}}/2} \|w_h\|_H \tag{3.236}$$

and hence

$$\Lambda = \sup_{v_h \in V_h^s} \frac{\|v_h\|_H^2}{\|v_h\|_{V_h^{s*}}^2} \leq C_{IP}^2 h^{-\alpha_{\mathcal{A}}}, \tag{3.237}$$

with $\Lambda = \Lambda(h)$ *defined in* (3.195) *at the beginning of Section 3.6.1. Consequently, under Assumption 3.B, for* $\theta \in \left[0, \frac{1}{2}\right)$ *the time stepping condition on* Δt *as required by* (3.200) *in Lemma 3.68 is satisfied if*

$$0 < \Delta t < \frac{2\beta}{(1 - 2\theta) C_{IP}^2 \alpha^2} h^{\alpha_{\mathcal{A}}} = C_\theta\, h^{\alpha_{\mathcal{A}}} \tag{3.238}$$

with $C_\theta = 2\beta / [(1 - 2\theta) C_{IP}^2 \alpha^2]$.

Corollary 3.71 (Stability estimate for $\boldsymbol{\theta}$ scheme)
Under the assumptions of Lemma 3.68 the stability estimate

$$\|u_h^M\|_H^2 + \Delta t\, C_1 \beta \sum_{m=0}^{M-1} \left\|u_h^{m+\theta}\right\|_{V^s}^2 \leq \|u_h^0\|_H^2 + \Delta t\, C_2 \sum_{m=0}^{M-1} \left\|f^{m+\theta}\right\|_{V_h^{s*}}^2, \tag{3.239}$$

holds with positive constants C_1*,* C_2 *and the same time stepping condition for* $\theta \in [0, \frac{1}{2})$ *as required by Lemma 3.68.*

Proof
The claim is a direct consequence of Lemma 3.68 and the uniform coercivity of the bilinear form with coercivity constant β. □

Note that the result (3.239) of Corollary 3.71 in a way describes that the solution of the discrete scheme is bounded by its initial data in a discrete $L^2(0,T,V^s)$ or $L^2(0,T,V_h^{s*})$ norm fashion, respectively.

3.6.2.2 Convergence of coercive schemes

Under the assumptions of Lemma 3.68, the solution to Theta Scheme 3.63 is stable. In this subsection we show that it also converges in the dimensionality $N(h)$ of the space and the fineness Δt of the time grid. For that matter we consider the residuals between each member of $(u^m)_{m\in\{0,\dots,M\}}$, the weak solution of (3.172) evaluated at time points t^m, $m = 0,\dots,M$, and the respective members of the sequence $(u_h^m)_{m\in\{0,\dots,M\}}$, the solution of Theta Scheme 3.63.

In order to ultimately prove convergence, we will show that (parts of) these residuals satisfy their own θ scheme with a new right hand side. Applying Lemma 3.68 to these very residuals will yield an upper bound for the sum of their norms from which convergence can be deduced.

We define for all $m \in \{0,\dots,M\}$ the difference e_h^m between the weak solution evaluated at time point t^m and its finite dimensional approximation affiliated with time point t^m as

$$\begin{aligned} e_h^m &= u^m - u_h^m \\ &= (u^m - P_h u^m) + (P_h u^m - u_h^m) \\ &= \eta^m + \xi_h^m, \end{aligned} \tag{3.240}$$

with

$$\eta^m = u^m - P_h u^m, \qquad \forall m \in \{0,\dots,M\}, \tag{3.241}$$

$$\xi_h^m = P_h u^m - u_h^m, \qquad \forall m \in \{0,\dots,M\}, \tag{3.242}$$

with a projector P_h adhering to Assumption 3.C. The quantity e_h^m thus consists of two parts. The first part, η^m, carries the discretization error, the second part, ξ_h^m, denotes the inaccuracy of the approximate solution with respect to the projection of the weak solution into the finite dimensional subspace.

In the end, convergence itself then depends on the specification of the function Υ of Assumption 3.A and its behavior when h tends to zero. This behavior of Υ in turn originates from the smoothness that the weak solution u admits. The more smoothness it provides, the faster the achieved rate of convergence will be. We will keep this issue determining the rate of convergence separate from the derivation of the convergence results as such wherever possible. After the formal convergence analysis has been completed, the assumptions on the space that the solution lives in will determine the actual rate of convergence that the θ scheme achieves.

To derive the convergence result we focus on the term ξ_h^m in (3.240), first. Being the part of the residual e_h^m that denotes the deviation of the solution of the θ scheme from the projection of the weak solution, it is of central interest for the whole analysis.

Lemma 3.72 ($\boldsymbol{\theta}$ scheme for the $\boldsymbol{\xi_h^m}$)
Let $u \in W^1(0,T;V^s,H)$ be the weak solution to problem (3.172) *with continuous and coercive bilinear form a and $(u_h^m)_{m\in\{0,\ldots,M\}}$ the solution to the associated Theta Scheme 3.63. Further, let ξ_h^m, $m \in \{1,\ldots,M\}$, be defined by* (3.242)*. If additionally $u \in C^1([0,T];H)$ and the bilinear form is continuous in t then we have*

$$\begin{aligned}\left(\frac{\xi_h^{m+1}-\xi_h^m}{\Delta t}, v_h\right) + a^{m+\theta}(\theta\xi_h^{m+1} + (1-\theta)\xi_h^m, v_h) = (r^m, v_h),\\ \xi_h^0 = P_h g - u_h^0,\end{aligned} \tag{3.243}$$

for all $m = 1,\ldots,M-1$ and for all $v_h \in V_h^s$, where the weak residuals $r^m : V_h^s \to \mathbb{R}$ have the form

$$r^m = r_1^m + r_2^m + r_3^m \tag{3.244}$$

with

$$\begin{aligned}(r_1^m, v_h) &= \left(\frac{u^{m+1}-u^m}{\Delta t} - \dot{u}^{m+\theta}, v_h\right),\\ (r_2^m, v_h) &= \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t} - \frac{u^{m+1}-u^m}{\Delta t}, v_h\right),\\ (r_3^m, v_h) &= a^{m+\theta}\left(P_h u^{m+\theta} - u^{m+\theta}, v_h\right).\end{aligned}$$

for all $m \in \{0,\ldots,M-1\}$.

Proof
By admitting time dependence of the bilinear form, this lemma generalizes Lemma 5.1 in von Petersdorff and Schwab (2003). The proof therein provides very reliable guidelines along which we now derive our result, as well.

Choose $v_h \in V_h^s \subset V^s$ arbitrary but fix and $m \in \{0, \dots, M-1\}$ and compute

$$\begin{aligned}
&\left(\frac{\xi_h^{m+1} - \xi_h^m}{\Delta t}, v_h\right) + a^{m+\theta}\left(\theta \xi_h^{m+1} + (1-\theta)\xi_h^m, v_h\right) \\
&\quad = \left(\frac{(P_h u^{m+1} - u_h^{m+1}) - (P_h u^m - u_h^m)}{\Delta t}, v_h\right) \\
&\qquad + a^{m+\theta}\left(\theta\left(P_h u^{m+1} - u_h^{m+1}\right) + (1-\theta)\left(P_h u^m - u_h^m\right), v_h\right) \\
&\quad = \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t}, v_h\right) - \left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, v_h\right) \\
&\qquad + a^{m+\theta}(P_h u^{m+\theta}, v_h) - a^{m+\theta}(u_h^{m+\theta}, v_h) \\
&\quad = \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t}, v_h\right) + a^{m+\theta}(P_h u^{m+\theta}, v_h) \\
&\qquad - \left(\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, v_h\right) + a^{m+\theta}(u_h^{m+\theta}, v_h)\right). \qquad (3.245)
\end{aligned}$$

We invoke the relation provided by the fully discretized θ scheme 3.63 to bring $f^{m+\theta}$ into the equation, then add a zero and thus continue from (3.245) with

$$\begin{aligned}
&\left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t}, v_h\right) + a^{m+\theta}(P_h u^{m+\theta}, v_h) - \left(\left(\frac{u_h^{m+1} - u_h^m}{\Delta t}, v_h\right) + a^{m+\theta}(u_h^{m+\theta}, v_h)\right) \\
&\quad = \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t}, v_h\right) + a^{m+\theta}(P_h u^{m+\theta}, v_h) - (f^{m+\theta}, v_h) \\
&\quad = \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t} - \dot{u}^{m+\theta}, v_h\right) + a^{m+\theta}(P_h u^{m+\theta}, v_h) \\
&\qquad + (\dot{u}^{m+\theta}, v_h) - (f^{m+\theta}, v_h). \qquad (3.246)
\end{aligned}$$

By Equation (3.11) and the assumption that $u \in C^1([0,T], H)$ together with the bilinear form a being continuous in t, the fundamental theorem of variational calculus implies that

$$(\dot{u}^{m+\theta}, v) + a^{m+\theta}\left(u^{m+\theta}, v\right) = (f^{m+\theta}, v), \qquad \forall v \in V^s. \qquad (3.247)$$

Recalling that $v_h \in V_h^s \subset V^s$ we combine (3.247) and (3.246) to get

$$\begin{aligned}
&\left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t} - \dot{u}^{m+\theta}, v_h\right) + a^{m+\theta}(P_h u^{m+\theta}, v_h) + (\dot{u}^{m+\theta}, v_h) - (f^{m+\theta}, v_h) \\
&\quad = \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t} - \dot{u}^{m+\theta}, v_h\right) + a^{m+\theta}(P_h u^{m+\theta}, v_h) - a^{m+\theta}(u^{m+\theta}, v_h) \quad (3.248)
\end{aligned}$$

Adding an artificial zero to (3.248) we arrive at

$$\begin{aligned}
&\left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t} - \dot{u}^{m+\theta}, v_h\right) + a^{m+\theta}(P_h u^{m+\theta}, v_h) - a^{m+\theta}(u^{m+\theta}, v_h) \\
&\quad = \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t} - \frac{u^{m+1} - u^m}{\Delta t}, v_h\right) + \left(\frac{u^{m+1} - u^m}{\Delta t} - \dot{u}^{m+\theta}, v_h\right) \\
&\qquad + a^{m+\theta}(P_h u^{m+\theta} - u^{m+\theta}, v_h), \\
&\quad = (r^m, v_h),
\end{aligned}$$

with

$$(r^m, v_h) := (r_1^m + r_2^m + r_3^m, v_h),$$

wherein

$$\begin{aligned}
(r_1^m, \cdot) &= \left(\frac{u^{m+1} - u^m}{\Delta t} - \dot{u}^{m+\theta}, \cdot\right), \\
(r_2^m, \cdot) &= \left(\frac{P_h u^{m+1} - P_h u^m}{\Delta t} - \frac{u^{m+1} - u^m}{\Delta t}, \cdot\right), \\
(r_3^m, \cdot) &= a^{m+\theta}(P_h u^{m+\theta} - u^{m+\theta}, \cdot),
\end{aligned}$$

which validates the decomposition of r^m claimed by the lemma. □

We have therefore derived a θ scheme for ξ_h^m which we state for later reference.

Theta Scheme 3.73 (θ scheme for the ξ_h^m)
Under the assumptions of Lemma 3.72 with $\theta \in [0,1]$, the ξ_h^m defined by (3.242) satisfy the θ scheme

$$\begin{aligned}
\left(\frac{\xi_h^{m+1} - \xi_h^m}{\Delta t}, v_h\right) + a^{m+\theta}(\theta \xi_h^{m+1} + (1-\theta)\xi_h^m, v_h) &= (r^m, v_h), \\
\xi_h^0 &= P_h g - u_h^0,
\end{aligned} \tag{3.249}$$

for all $m = 1, \ldots, M-1$ and for all $v_h \in V_h^s$, where the right hand side given by r^m is defined as in (3.244) of the Lemma.

For the solution $(\xi_h^m)_{m \in \{0,\ldots,M\}}$ of Scheme 3.73, the following stability estimate holds.

Corollary 3.74 (Stability estimate for ξ_h^m)
Let $(\xi_h^m)_{m \in \{0,\ldots,M\}}$ be the solution of the θ scheme 3.73 with $\theta \in [0,1]$ and let the assumptions of Lemma 3.68 be satisfied. Then there exist positive constants C_1, C_2 such that the stability estimate

$$\left\|\xi_h^M\right\|_H^2 + \Delta t\, C_1 \sum_{m=0}^{M-1} \left\|\xi_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \leq \left\|\xi_h^0\right\|_H^2 + \Delta t\, C_2 \sum_{m=0}^{M-1} \left\|r^m\right\|_{V_h^{s*}}^2 \tag{3.250}$$

holds.

Proof
By assumption, the bilinear form $a_t(\cdot,\cdot)$ is continuous and coercive uniformly in time. The ξ_h^m thus take the role of the u_h^m in the θ scheme 3.63 and the r^m take the role of the $f^{m+\theta}$ therein. Consequently, we can directly apply Lemma 3.68. The constants C_1, C_2 of the corollary are thus identical to the two constants of the lemma. □

Convergence of the (approximate) solution $(u_h^m)_{m\in\{0,\dots,M\}}$ will depend on convergence of the right hand side in (3.250). In that respect, Corollary 3.74 is the key ingredient to our convergence results for bilinear forms that are both continuous as well as coercive uniformly in time. In preparation of these results we shall now derive upper bounds for the individual residual parts r_1^m, r_2^m and r_3^m.

The following lemma provides upper bounds for the residuals individually. Each of those bounds depends on the grid parameters h and Δt. Both serve as determinants for the rate of convergence of the θ scheme, later.

Lemma 3.75 (Upper bounds for normed residuals)
Let the assumptions of Lemma 3.72 be satisfied and let $(r_i^m,\cdot)_H$ with $r_i^m : V_h^s \to \mathbb{R}$, $i \in \{1,2,3\}$, be the weak residuals derived by the lemma. We require additional smoothness of the weak solution u by assuming further that

- *i) Assumption 3.A holds for some function Υ and some constant C_Υ*
- *ii) Assumption 3.C on the projector P_h holds*
- *iii) $u \in W^1(0,T;V^t,H)$ for some $t \geq \alpha_{\mathcal{A}}/2$*
- *iv) $u \in C^2([0,T],H)$*

In case $\theta = \frac{1}{2}$ assume optionally

- *v) $u \in C^3([0,T],H)$*

Then there exist positive constants C_{r_1}, C_{r_2} and C_{r_3} such that

$$\|r_1^m\|_{V_h^{s*}} \leq C_{r_1} \begin{cases} \sqrt{\Delta t}\left(\int_{t^m}^{t^{m+1}} \|\ddot{u}(s)\|_{V_h^{s*}}^2 \,\mathrm{d}s\right)^{\frac{1}{2}}, & \theta \in [0,1] \\ \Delta t^{\frac{3}{2}}\left(\int_{t^m}^{t^{m+1}} \|\dddot{u}(s)\|_{V_h^{s*}}^2 \,\mathrm{d}s\right)^{\frac{1}{2}}, & \theta = \frac{1}{2} \text{ and given v) holds} \end{cases} \tag{3.251}$$

$$\|r_2^m\|_{V_h^{s*}} \leq C_{r_2} \frac{1}{\sqrt{\Delta t}} \left(\int_{t^m}^{t^{m+1}} \Upsilon^2\left(h,t,\alpha_{\mathcal{A}}/2,\dot{u}(\tau)\right) \mathrm{d}\tau\right)^{\frac{1}{2}}, \tag{3.252}$$

$$\|r_3^m\|_{V_h^{s*}} \leq C_{r_3}\,\Upsilon(h,t,\alpha_{\mathcal{A}}/2,u^{m+\theta}), \tag{3.253}$$

for all $m = 0,\dots,M-1$.

Proof
Choose $v_h \in V_h^s$ arbitrary but fix. We derive each upper bound individually.

Upper bound for $\|r_1^m\|_{V_h^{s*}}$*:*

Clearly,

$$|(r_1^m, v_h)| \leq \left\| \frac{u^{m+1} - u^m}{\Delta t} - \dot{u}^{m+\theta} \right\|_{V_h^{s*}} \|v_h\|_{V^s}, \tag{3.254}$$

by the definition of the norm $\|\cdot\|_{V_h^{s*}}$. Recall that our time grid is equidistantly spaced so $t^{m+1} = t^m + \Delta t$ for all $m \in \{0, \ldots, M-1\}$. Under the assumption that $u \in C^2([0,T], H)$ we represent $u^{m+1} = u(t^{m+1})$ by the Taylor expansion of u around t^m, evaluated at t^{m+1}. Thus, by applying Theorem 2.40, the Taylor theorem for Banach-valued functions, we have

$$u^{m+1} = u^m + \dot{u}^m \Delta t + \int_{t^m}^{t^{m+1}} (t^{m+1} - \tau) \ddot{u}(\tau) \, \mathrm{d}\tau. \tag{3.255}$$

Proceeding with elementary calculations we get

$$\begin{aligned}
\frac{u^{m+1} - u^m}{\Delta t} &- \dot{u}^{m+\theta} \\
&= \frac{\left(u^m + \dot{u}^m \Delta t + \int_{t^m}^{t^{m+1}} (t^{m+1} - \tau)\ddot{u}(\tau)\,\mathrm{d}\tau\right) - u^m}{\Delta t} - \dot{u}^{m+\theta} \\
&= \dot{u}^m + \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} (t^{m+1} - \tau)\ddot{u}(\tau)\,\mathrm{d}\tau - \left(\theta \dot{u}^{m+1} + (1-\theta)\dot{u}^m\right) \\
&= \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} (t^{m+1} - \tau)\ddot{u}(\tau)\,\mathrm{d}\tau - \left(\theta \dot{u}^{m+1} - \theta \dot{u}^m\right).
\end{aligned} \tag{3.256}$$

Since $u \in C^2([0,T], H)$, Lemma 2.38 together with Proposition 1.2.3 in Arendt et al. (2011) grant that $\ddot{u}$ is Bochner integrable and

$$\theta \dot{u}^{m+1} - \theta \dot{u}^m = \theta \int_{t^m}^{t^{m+1}} \ddot{u}(\tau) \, \mathrm{d}\tau. \tag{3.257}$$

Inserting (3.257) into (3.256) yields

$$\begin{aligned}
\frac{u^{m+1} - u^m}{\Delta t} - \dot{u}^{m+\theta} &= \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \left(t^{m+1} - \tau\right) \ddot{u}(\tau)\,\mathrm{d}\tau - \left(\theta \dot{u}^{m+1} - \theta \dot{u}^m\right) \\
&= \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \left(t^{m+1} - \tau - \theta \Delta t\right) \ddot{u}(\tau)\,\mathrm{d}\tau \\
&= -\frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \left(\tau - (1-\theta)t^{m+1} - \theta t^m\right) \ddot{u}(\tau)\,\mathrm{d}\tau.
\end{aligned} \tag{3.258}$$

Since $\ddot{u}$ is Bochner integrable, we can apply Theorem 24.7 in Wloka (2002) to get

$$\begin{aligned}
&\left\| \frac{u^{m+1}-u^m}{\Delta t} - \dot{u}^{m+\theta} \right\|_{V_h^{s*}} \\
&\qquad = \left\| \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \big(\tau - (1-\theta)t^{m+1} - \theta t^m\big)\, \ddot{u}(\tau)\,\mathrm{d}\tau \right\|_{V_h^{s*}} \\
&\qquad \le \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \big\| \big(\tau - (1-\theta)t^{m+1} - \theta t^m\big)\, \ddot{u}(\tau) \big\|_{V_h^{s*}}\,\mathrm{d}\tau
\end{aligned} \tag{3.259}$$

taking the norm into the integral. Considering the function

$$g_\theta(\tau) = \tau - (1-\theta)t^{m+1} - \theta t^m, \quad g_\theta : \big[t^m,\ t^{m+1}\big] \to \mathbb{R}$$

we find due to its strict monotonicity in τ that for $\tau \in [t^m, t^{m+1}]$

$$\begin{aligned}
|g_\theta(\tau)| \le\ & \max\left\{ \left| t^{m+1} - (1-\theta)t^{m+1} - \theta t^m \right|, \left| t^m - (1-\theta)t^{m+1} - \theta t^m \right| \right\} \\
=\ & \max\{ |\theta(t^{m+1} - t^m)|, |(1-\theta)(t^{m+1} - t^m)| \} \\
=\ & \Delta t \max\{\theta, (1-\theta)\} = \Delta t\, C_\theta,
\end{aligned} \tag{3.260}$$

with $C_\theta = \max\{\theta, (1-\theta)\}$. Using the estimate (3.260) we develop (3.259) into

$$\begin{aligned}
&\frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \big\| \big(\tau - (1-\theta)t^{m+1} - \theta t^m\big)\, \ddot{u}(\tau) \big\|_{V_h^{s*}}\,\mathrm{d}\tau \\
&\qquad \le C_\theta \int_{t^m}^{t^{m+1}} \|\ddot{u}(\tau)\|_{V_h^{s*}}\,\mathrm{d}\tau \\
&\qquad \le C_\theta \sqrt{\Delta t} \left(\int_{t^m}^{t^{m+1}} \|\ddot{u}(\tau)\|^2_{V_h^{s*}}\,\mathrm{d}\tau \right)^{\frac{1}{2}},
\end{aligned} \tag{3.261}$$

with the Hölder inequality of Theorem 2.42 being applied in the last step.

Special case $\theta = 1/2$:

For $\theta = 1/2$ and under the assumption of additional smoothness of u in the sense of v) being satisfied, further computations are possible. So let us assume that $\theta = \frac{1}{2}$.

Continuing in (3.258), we get by integration by parts and elementary calculations

$$\begin{aligned}
&\frac{u^{m+1}-u^m}{\Delta t} - \dot{u}^{m+\theta} \\
={}& -\frac{1}{\Delta t}\int_{t^m}^{t^{m+1}} \left(\tau - \frac{1}{2}t^{m+1} - \frac{1}{2}t^m\right) \ddot{u}(\tau)\,\mathrm{d}\tau \\
={}& -\frac{1}{2\Delta t}\Bigg(\left[\left(\tau^2 - \left(t^{m+1}+t^m\right)\tau\right)\ddot{u}(\tau)\right]_{t^m}^{t^{m+1}} \\
&\qquad - \int_{t^m}^{t^{m+1}} \left(\tau^2 - \left(t^{m+1}+t^m\right)\tau\right)\dddot{u}(\tau)\,\mathrm{d}\tau\Bigg) \\
={}& -\frac{1}{2\Delta t}\Bigg(- t^m t^{m+1}\ddot{u}^{m+1} + t^m t^{m+1}\ddot{u}^m \\
&\qquad - \int_{t^m}^{t^{m+1}} \left(\tau^2 - (t^{m+1}+t^m)\tau\right)\dddot{u}(\tau)\,\mathrm{d}\tau\Bigg) \qquad (3.262)\\
={}& -\frac{1}{2\Delta t}\left(-t^m t^{m+1}\left(\ddot{u}^{m+1}-\ddot{u}^m\right) - \int_{t^m}^{t^{m+1}} \left(\tau^2 - (t^{m+1}+t^m)\tau\right)\dddot{u}(\tau)\,\mathrm{d}\tau\right) \\
={}& \frac{1}{2\Delta t}\int_{t^m}^{t^{m+1}} \left(\tau^2 - (t^{m+1}+t^m)\tau + t^m t^{m+1}\right)\dddot{u}(\tau)\,\mathrm{d}\tau \\
={}& \frac{1}{2\Delta t}\int_{t^m}^{t^{m+1}} (\tau - t^{m+1})(\tau - t^m)\dddot{u}(\tau)\,\mathrm{d}\tau.
\end{aligned}$$

The absolute value of $\tau \mapsto (\tau - t^{m+1})(\tau - t^m)$ with $\tau \in \left[t^m, t^{m+1}\right]$ is bounded,

$$\left|(\tau - t^{m+1})(\tau - t^m)\right| \le \frac{1}{4}\Delta t^2, \qquad \tau \in [t^m, t^{m+1}]. \tag{3.263}$$

We take the norm $\|\cdot\|_{V_h^{s*}}$ of the result of (3.262), use the Bochner integrability of $\dddot{u}$ guaranteed by Proposition 1.2.3 in Arendt et al. (2011), apply again Theorem 24.7 of Wloka (2002) and then get by invoking estimate (3.263) that

$$\begin{aligned}
\left\|\frac{1}{\Delta t}\left(u^{m+1}-u^m\right) - \dot{u}^{m+\theta}\right\|_{V_h^{s*}} &= \frac{1}{2\Delta t}\left\|\int_{t^m}^{t^{m+1}} \left(t^{m+1}-\tau\right)\left(t^m-\tau\right)\dddot{u}(\tau)\,\mathrm{d}\tau\right\|_{V_h^{s*}} \\
&\le \frac{1}{2\Delta t}\int_{t^m}^{t^{m+1}} \frac{1}{4}\Delta t^2 \|\dddot{u}(\tau)\|_{V_h^{s*}}\,\mathrm{d}\tau \\
&= \frac{1}{8}\Delta t \int_{t^m}^{t^{m+1}} \|\dddot{u}(\tau)\|_{V_h^{s*}}\,\mathrm{d}\tau \\
&\le \frac{1}{8}\Delta t^{\frac{3}{2}}\left(\int_{t^m}^{t^{m+1}} \|\dddot{u}(\tau)\|_{V_h^{s*}}^2\,\mathrm{d}\tau\right)^{\frac{1}{2}}, \qquad (3.264)
\end{aligned}$$

with the Hölder inequality yielding the last step. Setting

$$C_{r_1} = \max\left\{C_\theta, \frac{1}{8}\right\} = C_\theta \in \left[\frac{1}{2}, 1\right]$$

defines the constant in (3.251) and finishes the treatment of r_1^m.

Upper bound for $\|r_2^m\|_{V_h^{s*}}$*:*

With assumption iv) we have again by combining Lemma 2.38 with Proposition 1.2.3 in Arendt et al. (2011) that $\dot{u}$ is Bochner integrable and

$$u^{m+1} - u^m = \int_{t^m}^{t^{m+1}} \dot{u}(\tau) \,\mathrm{d}\tau. \tag{3.265}$$

We begin the derivation of an upper bound for the norm of r_2^m by using (3.265) and compute

$$\begin{aligned} |(r_2^m, v_h)| &\leq \frac{1}{\Delta t} \left\| \left(u^{m+1} - u^m\right) - P_h \left(u^{m+1} - u^m\right) \right\|_{V_h^{s*}} \|v_h\|_{V^s} \\ &= \frac{1}{\Delta t} \left\| (I - P_h) \left(u^{m+1} - u^m\right) \right\|_{V_h^{s*}} \|v_h\|_{V^s} \\ &= \frac{1}{\Delta t} \left\| (I - P_h) \left(\int_{t^m}^{t^{m+1}} \dot{u}(\tau) \,\mathrm{d}\tau \right) \right\|_{V_h^{s*}} \|v_h\|_{V^s}, \end{aligned} \tag{3.266}$$

where I denotes the identity mapping. By Proposition 1.1.6 in Arendt et al. (2011) we may interchange integration with the $(I - P_h)$ operator to get

$$(I - P_h) \left(\int_{t^m}^{t^{m+1}} \dot{u}(\tau) \,\mathrm{d}\tau \right) = \int_{t^m}^{t^{m+1}} (I - P_h) \left(\dot{u}(\tau)\right) \mathrm{d}\tau. \tag{3.267}$$

Proposition 1.1.6 in Arendt et al. (2011) also grants that with $\dot{u}$ being Bochner integrable, $(I - P_h)(\dot{u})$ is Bochner integrable, as well. Consequently, we may combine expressions (3.266) and (3.267) and conclude again by Theorem 24.7 in Wloka (2002) that

$$\begin{aligned} |(r_2^m, v_h)| &\leq \frac{1}{\Delta t} \left\| (I - P_h) \left(\int_{t^m}^{t^{m+1}} \dot{u}(\tau) \,\mathrm{d}\tau \right) \right\|_{V_h^{s*}} \|v_h\|_{V^s} \\ &= \frac{1}{\Delta t} \left\| \int_{t^m}^{t^{m+1}} (I - P_h) \left(\dot{u}(\tau)\right) \mathrm{d}\tau \right\|_{V_h^{s*}} \|v_h\|_{V^s} \\ &\leq \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \|(I - P_h)\dot{u}(\tau)\|_{V_h^{s*}} \,\mathrm{d}\tau \|v_h\|_{V^s}. \end{aligned} \tag{3.268}$$

At this point we want to apply the approximation property of the projector P_h outlined in Assumption 3.C. Before we can do that we need to establish a relation between $\|\cdot\|_{V_h^{s*}}$ that we recognize in the last line of (3.268) and $\|\cdot\|_{V^s}$.

Keep $\tau \in \left[t^m, t^{m+1}\right]$ arbitrary but fix. Using the definition of norm $\|\cdot\|_{V_h^{s*}}$ we derive

$$\begin{aligned} \|(I-P_h)\dot{u}(\tau)\|_{V_h^{s*}} &= \sup_{v_h \in V_h} \frac{((I-P_h)\dot{u}(\tau), v_h)}{\|v_h\|_{V^s}} \\ &\leq \sup_{v_h \in V_h} \frac{\|(I-P_h)\dot{u}(\tau)\|_H \|v_h\|_H}{\|v_h\|_{V^s}} \\ &= \|(I-P_h)\dot{u}(\tau)\|_H \sup_{v_h \in V_h} \frac{\|v_h\|_H}{\|v_h\|_{V^s}} \\ &\leq \|(I-P_h)\dot{u}(\tau)\|_{V^s}, \end{aligned} \tag{3.269}$$

since $\|v\|_H \leq \|v\|_{V^s}$ for all $v \in V^s$. Inserting (3.269) into (3.268) and applying the approximation property of P_h pointwise in time we derive

$$\begin{aligned} |(r_2^m, v_h)| &\leq \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \|(I-P_h)\dot{u}(\tau)\|_{V_h^{s*}} \,\mathrm{d}\tau \|v_h\|_{V^s} \\ &\leq \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \|(I-P_h)\dot{u}(\tau)\|_{V^s} \,\mathrm{d}\tau \|v_h\|_{V^s} \\ &\leq C_\Upsilon \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \Upsilon\left(h, t, \alpha_{\mathcal{A}}/2, \dot{u}(\tau)\right) \mathrm{d}\tau \, \|v_h\|_{V^s} \\ &\leq C_{r_2} \frac{1}{\sqrt{\Delta t}} \left(\int_{t^m}^{t^{m+1}} \Upsilon^2\left(h, t, \alpha_{\mathcal{A}}/2, \dot{u}(\tau)\right) \mathrm{d}\tau \right)^{\frac{1}{2}} \|v_h\|_{V^s}, \end{aligned} \tag{3.270}$$

where the Hölder inequality grants the last step and where we used the additional smoothness in the sense of assumption iii) and where $C_{r_2} = C_\Upsilon$.

Upper bound for $\|r_3^m\|_{V_h^{s*}}$*:*

The bound for the norm of r_3^m is a direct consequence of the uniform continuity of $a_t(\cdot,\cdot)$. We compute for $v_h \in V_h^s$ that

$$\begin{aligned} |(r_3^m, v_h)| &= \left| a^{m+\theta}(P_h u^{m+\theta} - u^{m+\theta}, v_h) \right| \\ &\leq \alpha \left\| P_h u^{m+\theta} - u^{m+\theta} \right\|_{V^s} \|v_h\|_{V^s} \\ &\leq C_{r_3} \, \Upsilon(h, t, \alpha_{\mathcal{A}}/2, u^{m+\theta}) \|v_h\|_{V^s}, \end{aligned}$$

wherein $C_{r_3} = \alpha C_\Upsilon$, with α the continuity constant of $a_t(\cdot,\cdot)$ and C_Υ the constant stemming from the approximation property (3.184) of Assumption 3.A.

This finishes the derivation of upper bounds for the norms of the individual residuals r_1^m, r_2^m and r_3^m, $m = 0, \ldots, M-1$. □

We are now able to state the core theorem, granting convergence of the θ scheme 3.63 where the involved bilinear form is continuous and coercive uniformly in time.

Theorem 3.76 (Convergence of the coercive $\boldsymbol{\theta}$ scheme)
Let $u \in W^1(0,T;V^t,H)$, $t > \alpha_{\mathcal{A}}/2$, be the weak solution to problem (3.172) *where the operator is associated with a bilinear form a that is continuous and coercive uniformly in time. Further, assume*

i) *u to be smooth enough in the sense that $u \in C^2([0,T],H)$*

ii) *and for $\theta \in [0,1/2)$ let the time stepping condition* (3.200) *of Lemma 3.68 be satisfied.*

In case $\theta = \frac{1}{2}$ assume optionally

iii) $u \in C^3([0,T],H)$

Let $(u_h^m)_{m\in\{0,\dots,M\}}$ be the solution to the associated Theta Scheme 3.63 with $\theta \in [0,1]$ and assume further

iv) *The approximation property Assumption 3.A holds for some function Υ and some constant C_Υ*

v) *The inverse property Assumption 3.B is satisfied*

vi) *Assumption 3.C on the projector P_h holds*

vii) *Assumption 3.D on the initial condition is satisfied*

Then there exists a constant $\overline{C} > 0$ such that

$$
\begin{aligned}
\left\|u^M - u_h^M\right\|^2 + \Delta t \sum_{m=0}^{M-1} \left\|u^{m+\theta} - u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \\
&\le \overline{C} \max_{0\le\tau\le T} \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u(\tau)) \\
&+ \overline{C} \int_0^T \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,\dot{u}(\tau))\,\mathrm{d}\tau \\
&+ \overline{C} \begin{cases} (\Delta t)^2 \int_0^T \|\ddot{u}(s)\|_{V_h^{s*}}^2 \,\mathrm{d}s, & \forall\theta \in [0,1] \\ (\Delta t)^4 \int_0^T \|\dddot{u}(s)\|_{V_h^{s*}}^2 \,\mathrm{d}s, & \theta = \frac{1}{2} \textit{ and if iii)} \end{cases}
\end{aligned}
\tag{3.271}
$$

holds.

Proof
For $m \in \{0,\dots,M\}$ recall the definition

$$
e_h^m = u^m - u_h^m = \eta^m + \xi_h^m
$$

with

$$
\eta^m = u^m - P_h u^m, \qquad \forall m \in \{0,\dots,M\}, \tag{3.272}
$$

$$
\xi_h^m = P_h u^m - u_h^m, \qquad \forall m \in \{0,\dots,M\}, \tag{3.273}
$$

as introduced in (3.240). Additionally, we denote

$$\eta^{m+\theta} = \theta\eta^{m+1} + (1-\theta)u^m = u^{m+\theta} - P_h u^{m+\theta}, \qquad \forall m \in \{0,\dots,M-1\}, \tag{3.274}$$

$$\xi_h^{m+\theta} = \theta\xi_h^{m+1} + (1-\theta)\xi_h^m = P_h u^{m+\theta} - u_h^{m+\theta}, \qquad \forall m \in \{0,\dots,M-1\}. \tag{3.275}$$

By the third binomial formula we get

$$\begin{aligned}
\left\|u^M - u_h^M\right\|_H^2 &+ \Delta t \sum_{m=0}^{M-1} \left\|u^{m+\theta} - u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \\
&= \left\|e_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|e_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \\
&= \left\|\eta^M + \xi_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|\eta^{m+\theta} + \xi_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \\
&\le 2\left(\left\|u^M - P_h u^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|u^{m+\theta} - P_h u^{m+\theta}\right\|_{a^{m+\theta}}^2\right) \qquad (3.276) \\
&\qquad + 2\left(\left\|\xi_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|\xi_h^{m+\theta}\right\|_{a^{m+\theta}}^2\right). \qquad (3.277)
\end{aligned}$$

Considering the first main summand, that is (3.276), we simply exploit the continuity of $a_t(\cdot,\cdot)$ to get

$$\begin{aligned}
\left\|u^M - P_h u^M\right\|_H^2 &+ \Delta t \sum_{m=0}^{M-1} \left\|u^{m+\theta} - P_h u^{m+\theta}\right\|_{a^{m+\theta}}^2 \\
&\le \left\|u^M - P_h u^M\right\|_{V^s}^2 + \alpha \frac{T}{M} \sum_{m=0}^{M-1} \left\|u^{m+\theta} - P_h u^{m+\theta}\right\|_{V^s}^2.
\end{aligned} \tag{3.278}$$

Considering the term $\sum_{m=0}^{M-1} \left\|u^{m+\theta} - P_h u^{m+\theta}\right\|_{V^s}^2$ in (3.278) we see by the linearity of the projector P_h and elementary calculations that

$$\begin{aligned}
&\sum_{m=0}^{M-1} \left\|u^{m+\theta} - P_h u^{m+\theta}\right\|_{V^s}^2 \\
&= \sum_{m=0}^{M-1} \left\|\theta\left(u^{m+1} - P_h u^{m+1}\right) + (1-\theta)\left(u^m - P_h u^m\right)\right\|_{V^s}^2 \\
&\le 2\sum_{m=0}^{M-1} \left(\theta^2\left\|u^{m+1} - P_h u^{m+1}\right\|_{V^s}^2 + (1-\theta)^2\left\|u^m - P_h u^m\right\|_{V^s}^2\right). \qquad (3.279)
\end{aligned}$$

We split the sum in (3.279) and replace the individual summands by the maximum summand yielding the estimate

$$\begin{aligned}
&\sum_{m=0}^{M-1} \left\| u^{m+\theta} - P_h u^{m+\theta} \right\|_{V^s}^2 \\
&\quad \leq 2 \Bigg(M\theta^2 \max_{0\leq\tau\leq T} \left(\|u(\tau) - P_h u(\tau)\|_{V^s}^2 \right) \\
&\qquad\qquad + M(1-\theta)^2 \max_{0\leq\tau\leq T} \left(\|u(\tau) - P_h u(\tau)\|_{V^s}^2 \right) \Bigg) \\
&\quad = 2M \left(\theta^2 + (1-\theta)^2\right) \max_{0\leq\tau\leq T} \left(\|u(\tau) - P_h u(\tau)\|_{V^s}^2 \right) \\
&\quad \leq M \max_{0\leq\tau\leq T} \left(\|u(\tau) - P_h u(\tau)\|_{V^s}^2 \right).
\end{aligned} \tag{3.280}$$

Inserting (3.280) into (3.278) yields

$$\begin{aligned}
\left\| u^M - P_h u^M \right\|_H^2 &+ \Delta t \sum_{m=0}^{M-1} \left\| u^{m+\theta} - P_h u^{m+\theta} \right\|_{a^{m+\theta}}^2 \\
&\leq \left\| u^M - P_h u^M \right\|_{V^s}^2 + \alpha T \max_{0\leq\tau\leq T} \left(\|u(\tau) - P_h u(\tau)\|_{V^s}^2 \right) \\
&\leq \ (1+\alpha T) \max_{0\leq\tau\leq T} \left(\|u(\tau) - P_h u(\tau)\|_{V^s}^2 \right).
\end{aligned} \tag{3.281}$$

Finally, the approximation property of the projector of Assumption 3.C applied pointwise in time yields

$$\begin{aligned}
\left\| u^M - P_h u^M \right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\| u^{m+\theta} - P_h u^{m+\theta} \right\|_{a^{m+\theta}}^2 \\
\leq \overline{C}_1 \max_{0\leq\tau\leq T} \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u(\tau)),
\end{aligned} \tag{3.282}$$

with $\overline{C}_1 = C_\Upsilon^2(1+\alpha T)$.

Considering now the main summand in (3.277) we find applying Corollary 3.74 using the positive constants C_1 and C_2 therein that

$$\begin{aligned}
\left\| \xi_h^M \right\|_H^2 &+ \Delta t \sum_{m=0}^{M-1} \left\| \xi_h^{m+\theta} \right\|_{a^{m+\theta}}^2 \\
&\leq \ \max\left\{1, \frac{1}{C_1}\right\} \left(\left\| \xi_h^M \right\|_H^2 + \Delta t\, C_1 \sum_{m=0}^{M-1} \left\| \xi_h^{m+\theta} \right\|_{a^{m+\theta}}^2 \right) \\
&\leq \ \max\left\{1, \frac{1}{C_1}\right\} \left(\left\| \xi_h^0 \right\|_H^2 + \Delta t\, C_2 \sum_{m=0}^{M-1} \|r_h^m\|_{V_h^{s*}}^2 \right).
\end{aligned} \tag{3.283}$$

We investigate $\|\xi_h^0\|_H^2$, first. By definition of ξ_h^m for $m = 0$ we get

$$\|\xi_h^0\|_H = \|P_h u^0 - u_h^0\|_H. \tag{3.284}$$

Recall that

$$u^0 = u(t^0) = u(0) = g, \tag{3.285}$$

the initial condition of the original problem (3.172) and further

$$u_h^0 = g_h, \tag{3.286}$$

by the initial condition of the fully discretized θ scheme 3.63. With inserting both (3.285) and (3.286) into (3.284) and exploiting approximation property of the projector P_h of Assumption 3.C as well as the quasi-optimality of the initial condition as stated in Assumption 3.D, we find

$$\begin{aligned} \|\xi_h^0\|_H &\leq \|P_h u^0 - g\|_H + \|g - u_h^0\|_H \\ &= \|u(0) - P_h u(0)\|_H + \|g - g_h\|_H \\ &\leq \|u(0) - P_h u(0)\|_H + C_I \inf_{v_h \in V_h^s} \|g - v_h\|_H \\ &= \|u(0) - P_h u(0)\|_H + C_I \inf_{v_h \in V_h^s} \|u(0) - v_h\|_H \\ &\leq \max_{0 \leq \tau \leq T} \left(\|u(\tau) - P_h u(\tau)\|_H + C_I \inf_{v_h \in V_h^s} \|u(\tau) - v_h\|_H \right) \\ &\leq \max_{0 \leq \tau \leq T} C_\Upsilon (1 + C_I) \Upsilon(h, t, \alpha_{\mathcal{A}}/2, u(\tau)) \\ &= \max_{0 \leq \tau \leq T} \sqrt{\overline{C}_2} \Upsilon(h, t, \alpha_{\mathcal{A}}/2, u(\tau)) \end{aligned} \tag{3.287}$$

with $\overline{C}_2 = C_\Upsilon^2 (1 + C_I)^2$, having applied the approximation property 3.C of the projector P_h at the end of the derivation. Secondly, considering the sum of normed residuals in (3.283) we observe that

$$\begin{aligned} \|r_h^m\|_{V_h^{s*}}^2 &= \|r_1^m + r_2^m + r_3^m\|_{V_h^{s*}}^2 \\ &\leq 4 \left(\|r_1^m\|_{V_h^{s*}}^2 + \|r_2^m\|_{V_h^{s*}}^2 + \|r_3^m\|_{V_h^{s*}}^2 \right) \end{aligned} \tag{3.288}$$

where we insert the individual upper bounds for the normed residuals $\|r_1^m\|_{V_h^{s*}}$, $\|r_2^m\|_{V_h^{s*}}$

and $\|r_3^m\|_{V_h^{s*}}$ that we have derived in Lemma 3.75 to find

$$\begin{aligned}\frac{1}{4}\sum_{m=0}^{M-1}\|r_h^m\|^2_{V_h^{s*}} \leq C_{r_1}^2 \sum_{m=0}^{M-1} &\begin{cases}\Delta t \int_{t^m}^{t^{m+1}} \|\ddot{u}(s)\|^2_{V_h^{s*}}\,\mathrm{d}s, & \forall\theta\in[0,1]\\ (\Delta t)^3 \int_{t^m}^{t^{m+1}} \|\dddot{u}(s)\|^2_{V_h^{s*}}\,\mathrm{d}s, & \theta=\frac{1}{2}\end{cases}\\ &+C_{r_2}^2\sum_{m=0}^{M-1}\frac{1}{\Delta t}\int_{t_m}^{t_{m+1}}\Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,\dot{u}(\tau))\,\mathrm{d}\tau\\ &+C_{r_3}^2\sum_{m=0}^{M-1}\Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u^{m+\theta})\\ \leq C_{r_1}^2 &\begin{cases}\Delta t \int_0^T \|\ddot{u}(s)\|^2_{V_h^{s*}}\,\mathrm{d}s, & \forall\theta\in[0,1]\\ (\Delta t)^3 \int_0^T \|\dddot{u}(s)\|^2_{V_h^{s*}}\,\mathrm{d}s, & \theta=\frac{1}{2}\end{cases}\\ &+C_{r_2}^2\frac{1}{\Delta t}\int_0^T\Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,\dot{u}(\tau))\,\mathrm{d}\tau\\ &+C_{r_3}^2 M\max_{0\leq\tau\leq T}\Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u(\tau)),\end{aligned} \tag{3.289}$$

with positive constants $C_{r_1}, C_{r_2}, C_{r_3}$ defined in the Lemma. We return to (3.277) and invoke (3.282) and (3.283) to derive

$$\begin{aligned}\|u^M-u_h^M\|_H^2 &+ \Delta t\sum_{m=0}^{M-1}\left\|u^{m+\theta}-u_h^{m+\theta}\right\|^2_{a^{m+\theta}}\\ &\leq 2\left(\|u^M-P_hu^M\|_H^2+\Delta t\sum_{m=0}^{M-1}\left\|u^{m+\theta}-P_hu^{m+\theta}\right\|^2_{a^{m+\theta}}\right)\\ &\quad+2\left(\|\xi_h^M\|_H^2+\Delta t\sum_{m=0}^{M-1}\left\|\xi_h^{m+\theta}\right\|^2_{a^{m+\theta}}\right)\\ &\leq 2\overline{C}_1\max_{0\leq\tau\leq T}\Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u(\tau))\\ &\quad+2\max\left\{1,\frac{1}{C_1}\right\}\left(\|\xi_h^0\|_H^2+\Delta t\,C_2\sum_{m=0}^{M-1}\|r_h^m\|^2_{V_h^{s*}}\right).\end{aligned} \tag{3.290}$$

Invoking our considerations for ξ_h^0 and the sum of normed residuals r_h^m in (3.287) and

(3.289) to deduce

$$
\begin{aligned}
&\left\|u^M - u_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|u^{m+\theta} - u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \\
&\le 2\overline{C}_1 \max_{0\le\tau\le T} \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u(\tau)) \\
&\quad + 2\max\left\{1, \frac{1}{C_1}\right\} \Bigg(\overline{C}_2 \max_{0\le\tau\le T} \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u(\tau)) \\
&\qquad + 4C_2 \Bigg(C_{r_1}^2 \begin{cases} (\Delta t)^2 \int_0^T \|\ddot{u}(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau, & \forall\theta\in[0,1] \\ (\Delta t)^4 \int_0^T \|\dddot{u}(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau, & \theta=\frac{1}{2} \end{cases} \\
&\qquad\quad + C_{r_2}^2 \int_0^T \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,\dot{u}(\tau))\,\mathrm{d}\tau \\
&\qquad\quad + C_{r_3}^2 T \max_{0\le\tau\le T} \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u(\tau)) \Bigg)\Bigg).
\end{aligned}
\tag{3.291}
$$

For a notationally more satisfying result we define the constant

$$
\overline{C} = 2\max\left\{3\overline{C}_1,\ \max\left\{1,\frac{1}{C_1}\right\}\max\left\{3\overline{C}_2,\ 4C_2\max\left\{C_{r_1}^2,\ C_{r_2}^2,\ 3C_{r_3}^2 T\right\}\right\}\right\}. \tag{3.292}
$$

Clearly, $\|g\|_{V^s} = \|u(0)\|_{V^s} \le \max\limits_{0\le\tau\le T} \|u(\tau)\|_{V^s}$. Thus, using (3.292) in (3.291) we get the estimate

$$
\begin{aligned}
\left\|u^M - u_h^M\right\|^2 + \ \Delta t \sum_{m=0}^{M-1} &\left\|u^{m+\theta} - u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 \\
&\le \overline{C} \max_{0\le\tau\le T} \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u(\tau)) \\
&+ \overline{C} \begin{cases} (\Delta t)^2 \int_0^T \|\ddot{u}(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau, & \forall\theta\in[0,1] \\ (\Delta t)^4 \int_0^T \|\dddot{u}(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau, & \theta=\frac{1}{2} \text{ and with iii)} \end{cases} \\
&+ \overline{C} \int_0^T \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,\dot{u}(\tau))\,\mathrm{d}\tau
\end{aligned}
\tag{3.293}
$$

which finishes the proof. □

Corollary 3.77 (Convergence of the $\boldsymbol{\theta}$ scheme)
Under the assumptions of Theorem 3.76 there exists $\overline{C} > 0$ such that the estimate

$$
\begin{aligned}
\left\|u^M - u_h^M\right\|^2 + \ \Delta t \sum_{m=0}^{M-1} \left\|u^{m+\theta} - u_h^{m+\theta}\right\|_V^2 & \\
\leq \overline{C} \ \max_{0\leq\tau\leq T} \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u(\tau)) & \\
+ \overline{C} \begin{cases} (\Delta t)^2 \int_0^T \|\ddot{u}(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau, & \forall\theta \in [0,1] \\ (\Delta t)^4 \int_0^T \|\dddot{u}(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau, & \theta = \frac{1}{2} \ \textit{and with iii)} \end{cases} & \\
+ \overline{C} \int_0^T \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, \dot{u}(\tau)) \,\mathrm{d}\tau &
\end{aligned}
\tag{3.294}
$$

holds.

Proof
The result is an immediate consequence from Theorem 3.76 and the fact that $a_t(\cdot, \cdot)$ is coercive uniformly in time with coercivity constant β. □

In Theorem 3.76 and Corollary 3.77, respectively, we have derived abstract convergence results. The particular convergence now follows immediately, when the form of Υ of Assumption 3.A, the general approximation property, is specified. The following corollary combines the result of

Corollary 3.78 (Convergence with $\boldsymbol{\Upsilon}$ of von Petersdorff and Schwab (2003))
Under the assumptions of Theorem 3.76 and in the setting of von Petersdorff and Schwab (2003) outlined in Example 3.58 there exists a constant $\overline{C} > 0$ such that the convergence estimate

$$
\begin{aligned}
\left\|u^M - u_h^M\right\|^2 + \Delta t \sum_{m=0}^{M-1} \left\|u^{m+\theta} - u_h^{m+\theta}\right\|_{a^{m+\theta}}^2 & \\
\leq \overline{C}\, h^{2(p+1-\alpha_{\mathcal{A}}/2)} \max_{0\leq\tau\leq T} \|u(\tau)\|_{\mathcal{H}^{p+1}(\Omega)}^2 & \\
+ \overline{C}\, h^{2(p+1-\alpha_{\mathcal{A}}/2)} \int_0^T \|\dot{u}(\tau)\|_{\mathcal{H}^{p+1}(\Omega)}^2 \,\mathrm{d}\tau & \\
+ \overline{C} \begin{cases} (\Delta t)^2 \int_0^T \|\ddot{u}(s)\|_{V_h^{\alpha_{\mathcal{A}}/2*}}^2 \,\mathrm{d}s, & \forall\theta \in [0,1] \\ (\Delta t)^4 \int_0^T \|\dddot{u}(s)\|_{V_h^{\alpha_{\mathcal{A}}/2*}}^2 \,\mathrm{d}s, & \theta = \frac{1}{2} \ \textit{and if } u \in C^3([0,T], H) \end{cases} &
\end{aligned}
\tag{3.295}
$$

holds.

Proof
The result is a direct consequence from Theorem 3.76 with

$$
\Upsilon(h, t, s, u) = h^{t-s} \|u\|_{\mathcal{H}^t(\Omega)},
$$

taking $t \leq p+1$ equal to its maximal admissible value with p the polynomial degree that the basis functions of $V_h^{\alpha_{\mathcal{A}}/2}$ achieve piecewisely. □

The result of Corollary 3.78 confirms the order of convergence derived in Theorem 5.4 of von Petersdorff and Schwab (2003). In contrast to that former result which our analysis is based on, we allow for time-dependent bilinear forms and thus generalize their result to the time-inhomogeneous case.

3.6.3 Results for continuous bilinear forms of Gårding type

The stability estimate of Lemma 3.68 and the convergence result of Theorem 3.76 hold for continuous and coercive bilinear forms, only. In this subsection, we will derive equivalent results for continuous bilinear forms that only fulfill the weaker Gårding inequality in the sense of the following Definition 3.79.

Definition 3.79 (Gårding)
A bilinear form $a_\cdot(\cdot,\cdot) : [0,T] \times V^s \times V^s \to \mathbb{R}$, $(t,u,v) \mapsto a_t(u,v)$, is said to fulfill a Gårding inequality *uniformly in time with respect to V^s, if there exist constants $\beta > 0$, $\lambda \geq 0$ independent of t such that*

$$a_t(u,u) \geq \beta \|u\|_{V^s}^2 - \lambda \|u\|_H^2 \tag{3.296}$$

holds $\forall u \in V^s$, $\forall t \in [0,T]$. We call β the coercivity constant *and λ the* Gårding constant.

Clearly, every uniformly coercive bilinear form in the sense of Definition 3.65 is of Gårding type in the sense of Definition 3.79 as well with Gårding constant $\lambda = 0$.

Before we dive into the stability and convergence analysis of solutions of fully discretized PIDEs with bilinear forms of Gårding type, let us shed some light on the relation between coercive bilinear forms and their more general siblings.

3.6.3.1 On the relation between Coercivity and the Gårding property

A simple time transformation can transform a PIDE with operator of Gårding type into a PIDE with coercive operator. Consider the PIDE

$$\begin{aligned} \partial_t u + \mathcal{A}_t^{\text{Gårding}} u &= f, \\ u(0) &= g, \end{aligned} \tag{3.297}$$

with weak solution $u \in W^1(0,T;V^s,H)$, an operator $\mathcal{A}_t^{\text{Gårding}}$ that is assumed to be both continuous and of Gårding type uniformly in time and that is associated with a bilinear form

$$a_\cdot^{\text{Gårding}} : [0,T] \times V^s \times V^s \to \mathbb{R}, \qquad (t,u,v) \mapsto a_t(u,v), \tag{3.298}$$

that fulfills the Gårding inequality (3.296) of Definition 3.79 with $\beta > 0$ as well as $\lambda > 0$. Furthermore, let $a^{\text{Gårding}}$ be continuous with continuity constant denoted by α. Then, defining

$$\begin{aligned} u_\lambda(t,x) &= e^{-\lambda t}u(t,x), \qquad \forall (t,x) \in [0,T] \times \mathbb{R} \\ f_\lambda(t,x) &= e^{-\lambda t}f(t,x), \qquad \forall (t,x) \in [0,T] \times \mathbb{R} \end{aligned} \tag{3.299}$$

and inserting (3.299) into (3.297) yields

$$\begin{aligned} \partial_t \left(e^{\lambda t}u_\lambda(t,x)\right) + \mathcal{A}_t^{\text{Gårding}} \left(e^{\lambda t}u_\lambda(t,x)\right) &= e^{\lambda t}f_\lambda(t,x) \\ e^0 u_\lambda(0) &= g. \end{aligned} \tag{3.300}$$

Using the product rule on the time derivative gives

$$\partial_t \left(e^{\lambda \cdot}u_\lambda\right) = \lambda e^{\lambda \cdot}u_\lambda + e^{\lambda \cdot}\partial_t u_\lambda,$$

turning (3.300) into

$$\begin{aligned} e^{\lambda t}\partial_t u_\lambda(t,x) + e^{\lambda t}\mathcal{A}_t^{\text{Gårding}}u_\lambda(t,x) + \lambda e^{\lambda t}u_\lambda(t,x) &= e^{\lambda t}f_\lambda(t,x) \\ u_\lambda(0) &= g. \end{aligned} \tag{3.301}$$

Multiplying both sides in the first line of (3.301) by $e^{-\lambda t}$ gives

$$\begin{aligned} \partial_t u_\lambda(t,x) + \left(\mathcal{A}_t^{\text{Gårding}} + \lambda\right) u_\lambda(t,x) &= f_\lambda(t,x) \\ u_\lambda(0) &= g. \end{aligned}$$

which turns into

$$\begin{aligned} \partial_t u_\lambda(t,x) + \mathcal{A}_{\lambda t}u_\lambda(t,x) &= f_\lambda(t,x) \\ u_\lambda(0) &= g. \end{aligned} \tag{3.302}$$

when we define the operator $\mathcal{A}_{\lambda\cdot}$ by

$$\mathcal{A}_{\lambda\cdot} = \left(\mathcal{A}_\cdot^{\text{Gårding}} + \lambda\right). \tag{3.303}$$

In contrast to $a_\cdot^{\text{Gårding}}(\cdot,\cdot)$ of (3.298), the associated bilinear form

$$a_{\lambda\cdot}(\cdot,\cdot) : [0,T] \times V^s \times V^s \to \mathbb{R}, \qquad (t,u,v) \mapsto a_{\lambda t}(u,v), \tag{3.304}$$

is now coercive uniformly in time which we indicate by the subscript λ. The coercivity constant of a_λ is β, its continuity constant is given by $\alpha_\lambda = \alpha + \lambda$. The stability Lemma 3.68 and the convergence Theorem 3.76 apply when their other assumptions are fulfilled.

Given these considerations on the relation of bilinear forms between these two classes, the need for stability and convergence analysis for a Gårding type setting might seem questionable.

Yet, for the stability and convergence analysis there is no generic path from the Gårding to the coercive case. More precisely, while a solution to the "coercified problem" (3.302) allows for a transformation to the solution of the original PIDE (3.297) by

$$u_\lambda = e^{-\lambda \cdot} u, \tag{3.305}$$

on the whole space time domain, this is no longer possible once the PIDEs have been discretized,

$$u_{\lambda,h}^m \neq e^{-\lambda t^m} u_h^m, \tag{3.306}$$

in general. Stability and convergence results may still apply to solutions to the discrete version of the modified PIDE in (3.302). Due to Inequality (3.306), however, these results forbid an immediate conclusion with respect to stability and convergence of solutions to the related Gårding counterpart.

Despite the proximity between the two PIDEs, the respective stability and convergence analysis vary significantly in complexity. This increase in complexity is due to the fact that while a coercive (and continuous) bilinear form induces a norm that is equivalent to $\|\cdot\|_{V^s}$, a bilinear form of Gårding type loses this property. For that reason, stability and convergence analysis in the literature often focuses on coercive problems and disregards the more general yet more complex Gårding case.

In the following, we want to generalize our earlier stability and convergence analysis to the more general Gårding case. For the remaining part of this section we thus focus on the following problem. Let $\mathcal{A}$ be an operator of order $\alpha_{\mathcal{A}} \in [0,2]$ that induces a bilinear form $a : [0,T] \times V^s \times V^s$ that is continuous uniformly in time with continuity constant α and fulfills a Gårding inequality uniformly in time with Gårding constants $\beta, \lambda > 0$. Let the other requirements of Theorem 3.7 be satisfied such that there exists a unique weak solution $u \in W^1(0,T; V^s, H)$ to the problem

$$\begin{aligned} \dot{u} + \mathcal{A}u &= f \\ u(0) &= g \end{aligned} \tag{3.307}$$

where $f \in V^{s*}$, $g \in H$.

Let $u \in W^1(0,T; V^s, H)$ be the unique weak solution to problem (3.307). Now define $u_\lambda = e^{-\lambda \cdot} u$ and equivalently define $f_\lambda = e^{-\lambda \cdot} f$. We have seen above that u_λ solves

$$\begin{aligned} \dot{u_\lambda} + \mathcal{A}_\lambda u_\lambda &= f_\lambda \\ u_\lambda(0) &= g. \end{aligned} \tag{3.308}$$

Based on problems (3.307) and (3.308) and their discretizations, several θ schemes arise.

Theta Scheme 3.80 (Discretized Gårding scheme)
Let $\theta \in [0,1]$. We call $(u_h^m)_{m\in\{0,\dots,M\}}$ with $u_h^m \in V_h^s$ for all $m \in \{1,\dots,M\}$, $u_h^0 \in H$, the solution to the fully discretized version of problem (3.307) *if*

$$\begin{aligned}\left(\frac{u_h^{m+1}-u_h^m}{\Delta t}, v_h\right) + a_{m+\theta}\left(u_h^{m+\theta}, v_h\right) &= \left(f^{m+\theta}, v_h\right), \\ u_h^0 &= g_h,\end{aligned} \tag{3.309}$$

for all $v_h \in V_h^s$, for all $m \in \{0,\dots,M-1\}$, with $g_h \in H$ an approximation of g of problem (3.307) *of problem* (3.307)*.*

Theta Scheme 3.81 (Discretized coercified scheme)
Let $\theta \in [0,1]$. We denote by $(u_{\lambda,h}^m)_{m\in\{0,\dots,M\}}$, $u_{\lambda,h}^m \in V_h^s$ for all $m \in \{1,\dots,M\}$ and $u_{\lambda,h}^0 \in H$, the solution to the fully discretized version of problem (3.308)*. Then,*

$$\begin{aligned}\left(\frac{u_{\lambda,h}^{m+1}-u_{\lambda,h}^m}{\Delta t}, v_h\right) + a_{\lambda m+\theta}\left(u_{\lambda,h}^{m+\theta}, v_h\right) &= \left(f_\lambda^{m+\theta}, v_h\right), \\ u_{\lambda,h}^0 &= g_h,\end{aligned}$$

for all $v_h \in V_h^s$, for all $m \in \{0,\dots,M-1\}$, with $g_h \in H$ an approximation of g of problem (3.307)*.*

Let u_h^m, $m = 0,\dots,M$, be the solution to Scheme 3.80 that we just introduced. Define $\widetilde{u}_h^m = e^{-\lambda t^m} u_h^m$, and insert the result back into Scheme 3.80. Then we get

$$\begin{aligned}&\left(\frac{e^{\lambda(t^m+\Delta t)}\widetilde{u}_h^{m+1} - e^{\lambda t^m}\widetilde{u}_h^m}{\Delta t}, v_h\right) \\ &\qquad + a_{m+\theta}\left(\theta e^{\lambda(t^m+\Delta t)}\widetilde{u}_h^{m+1} + (1-\theta)e^{\lambda t^m}\widetilde{u}_h^m, v_h\right) = \left(f^{m+\theta}, v_h\right)\end{aligned}$$

for all $m \in \{0,\dots,M\}$ for all $v_h \in V_h^s$. Multiplying both sides by $e^{-\lambda t^m}$ shows that the $\widetilde{u}_h^m$ fulfill the following (degenerate) scheme.

Theta Scheme 3.82 (Degenerate Gårding scheme)
Choose $\theta \in [0,1]$ and let u_h^m, $m = 0,\dots,M$, be the solution to Scheme 3.80. Define

$$\widetilde{u}_h^m = e^{-\lambda t^m} u_h^m. \tag{3.310}$$

Then,

$$\begin{aligned}\left(\frac{e^{\lambda\Delta t}\widetilde{u}_h^{m+1} - \widetilde{u}_h^m}{\Delta t}, v_h\right) + a_{m+\theta}\left(\theta e^{\lambda\Delta t}\widetilde{u}_h^{m+1} + (1-\theta)\widetilde{u}_h^m, v_h\right) &= \left(e^{-\lambda t^m} f^{m+\theta}, v_h\right), \\ \widetilde{u}_h^0 &= g_h,\end{aligned}$$

for all $m \in \{0,\dots,M-1\}$.

Recall inequality (3.306) that stated

$$u^m_{\lambda,h} \neq \widetilde{u}^m_h,$$

in general. In other words, discretization and "coercification" do not permute. Nevertheless both quantities in inequality (3.306) are related after all. The strategy in the following derivations lies in showing that their difference vanishes when the dimensionality in space and the fineness in time increase. Yet that difference between $u^m_{\lambda,h}$ and $\widetilde{u}^m_h$ will cause significant complexity within the derivation of our analysis. For that reason, we simplify the involved theta schemes in return by assuming

$$\theta = 1 \tag{3.311}$$

throughout the whole derivation. While we do not expect assumption (3.311) to limit the generality of our results, it surely allows focusing on the key difficulties and serves the convenience of the reader. In order to quantify the difference between $u^m_{\lambda,h}$ and $\widetilde{u}^m_h$ we furthermore introduce the new quantities w^m, $m = 0, \dots, M$, defined as

$$w^m = \widetilde{u}^m_h - u^m_{\lambda,h} \tag{3.312}$$

and analyze their role in the validation of stability and convergence of the solutions to Gårding schemes. More precisely, we will realize that stability and convergence of solutions to Gårding schemes actually rely heavily on stability and convergence of the scheme that the w^m of (3.312) satisfy. We thus start our derivation by verifying those schemes and validating stability for these auxiliary quantities, themselves.

Lemma 3.83 (A scheme for w^m)
Set $\theta = 1$. Let $\widetilde{u}^m_h$, $m = 0, \dots, M$, be given by Scheme 3.82 and let $u^m_{\lambda,h}$, $m = 0, \dots, M$, be the solution to Scheme 3.81. Define

$$w^m = \widetilde{u}^m_h - u^m_{\lambda,h} \tag{3.313}$$

for all $m \in \{0, \dots, M\}$. Then we have

$$\begin{aligned} \left(\frac{w^{m+1} - w^m}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(w^{m+1}, v_h\right) &= \left(r^m_w, v_h\right), \\ w^0 &= 0, \end{aligned} \tag{3.314}$$

for all $m \in \{0, \dots, M-1\}$ and all $v_h \in V^s_h$ with

$$\left(r^m_w, v_h\right) = \lambda\left(w^{m+1} + u^{m+1}_{\lambda,h}, v_h\right) - \overline{\lambda}(\Delta t)\left(w^m + u^m_{\lambda,h}, v_h\right), \tag{3.315}$$

wherein the function $\overline{\lambda} : \mathbb{R}^+ \to \mathbb{R}$ is defined by

$$\overline{\lambda}(\Delta t) = \frac{1 - e^{-\lambda \Delta t}}{\Delta t}, \tag{3.316}$$

for all $\Delta t > 0$.

Proof
First, considering the initial condition we find

$$w^0 = \widetilde{u}_h^0 - u_{\lambda,h}^0 = g_h - g_h = 0.$$

Second, by the definition of w^m we have for $v_h \in V_h^s$ arbitrary but fix that

$$\begin{aligned}\left(\frac{w^{m+1}-w^m}{\Delta t}, v_h\right) &+ a_{\lambda m+1}\left(w^{m+1}, v_h\right) \\ &= \left(\frac{\widetilde{u}_h^{m+1}-\widetilde{u}_h^m}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(\widetilde{u}_h^{m+1}, v_h\right) \\ &\qquad - \left[\left(\frac{u_{\lambda,h}^{m+1}-u_{\lambda,h}^m}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(u_{\lambda,h}^{m+1}, v_h\right)\right].\end{aligned} \tag{3.317}$$

We artificially expand the first two summands in (3.317) to get

$$\begin{aligned}&\left(\frac{\widetilde{u}_h^{m+1}-\widetilde{u}_h^m}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(\widetilde{u}_h^{m+1}, v_h\right) \\ &= \left(\frac{-e^{\lambda\Delta t}\widetilde{u}_h^{m+1} + \widetilde{u}_h^{m+1} + e^{\lambda\Delta t}\widetilde{u}_h^{m+1} - \widetilde{u}_h^m}{\Delta t}, v_h\right) \\ &\qquad + a_{m+1}\left(-e^{\lambda\Delta t}\widetilde{u}_h^{m+1} + \widetilde{u}_h^{m+1} + e^{\lambda\Delta t}\widetilde{u}_h^{m+1}, v_h\right) + \lambda\left(\widetilde{u}_h^{m+1}, v_h\right) \\ &=: \left(\widetilde{r}_{\widetilde{u}_h}^m, v_h\right).\end{aligned} \tag{3.318}$$

From this we continue by isolating the left side of Scheme 3.82,

$$\begin{aligned}\left(\widetilde{r}_{\widetilde{u}_h}^m, v_h\right) = &\left(\frac{e^{\lambda\Delta t}\widetilde{u}_h^{m+1} - \widetilde{u}_h^m}{\Delta t}, v_h\right) + a_{m+1}\left(e^{\lambda\Delta t}\widetilde{u}_h^{m+1}, v_h\right) \\ &+ \left(1 - e^{\lambda\Delta t}\right)\left(\frac{\widetilde{u}_h^{m+1}}{\Delta t}, v_h\right) + \left(1 - e^{\lambda\Delta t}\right) a_{m+1}\left(\widetilde{u}_h^{m+1}, v_h\right) \\ &+ \lambda\left(\widetilde{u}_h^{m+1}, v_h\right).\end{aligned} \tag{3.319}$$

Next, we insert the right side of Scheme 3.82 for $\theta = 1$ into (3.319) and get

$$\begin{aligned}\left(\widetilde{r}_{\widetilde{u}_h}^m, v_h\right) = &\left(e^{-\lambda t^m} f^{m+1}, v_h\right) \\ &+ \left(1 - e^{\lambda\Delta t}\right)\left(\frac{\widetilde{u}_h^{m+1}}{\Delta t}, v_h\right) + \left(1 - e^{\lambda\Delta t}\right) a_{m+1}\left(\widetilde{u}_h^{m+1}, v_h\right) \\ &+ \lambda\left(\widetilde{u}_h^{m+1}, v_h\right).\end{aligned} \tag{3.320}$$

We artificially expand the part that just entered the equation and then use $t^{m+1} = t^m + \Delta t$ to get

$$\begin{aligned}\left(e^{-\lambda t^m} f^{m+1}, v_h\right) &= \left(e^{-\lambda t^m} f^{m+1} - e^{-\lambda t^{m+1}} f^{m+1}, v_h\right) + \left(e^{-\lambda t^{m+1}} f^{m+1}, v_h\right) \\ &= \left(1 - e^{-\lambda\Delta t}\right) e^{-\lambda t^m}\left(f^{m+1}, v_h\right) + \left(f_\lambda^{m+1}, v_h\right).\end{aligned} \tag{3.321}$$

We insert (3.321) into (3.320) and then use the equation from Scheme 3.81 to get

$$\begin{aligned}
&\left(r^m_{\widetilde{u}_h}, v_h\right)\\
&= \left(1 - e^{-\lambda\Delta t}\right) e^{-\lambda t^m} \left(f^{m+1}, v_h\right) + \left(f^{m+1}_\lambda, v_h\right)\\
&\quad + \left(1 - e^{\lambda\Delta t}\right) \left(\frac{\widetilde{u}^{m+1}_h}{\Delta t}, v_h\right) + \left(1 - e^{\lambda\Delta t}\right) a_{m+1}\left(\widetilde{u}^{m+1}_h, v_h\right)\\
&\quad + \lambda \left(\widetilde{u}^{m+1}_h, v_h\right)\\
&= \left(1 - e^{-\lambda\Delta t}\right) e^{-\lambda t^m} \left(f^{m+1}, v_h\right) + \left(\frac{u^{m+1}_{\lambda,h} - u^m_{\lambda,h}}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(u^{m+1}_{\lambda,h}, v_h\right)\\
&\quad + \left(1 - e^{\lambda\Delta t}\right) \left(\frac{\widetilde{u}^{m+1}_h}{\Delta t}, v_h\right) + \left(1 - e^{\lambda\Delta t}\right) a_{m+1}\left(\widetilde{u}^{m+1}_h, v_h\right)\\
&\quad + \lambda \left(\widetilde{u}^{m+1}_h, v_h\right).
\end{aligned} \tag{3.322}$$

Combining (3.322) with (3.320) we have thus rewritten the first part of (3.317). We take the resulting expression and insert it back into (3.317), to derive a right hand side for the w^m by

$$\begin{aligned}
&\left(\frac{w^{m+1} - w^m}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(w^{m+1}, v_h\right)\\
&= \left(\frac{\widetilde{u}^{m+1}_h - \widetilde{u}^m_h}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(\widetilde{u}^{m+1}_h, v_h\right)\\
&\quad - \left[\left(\frac{u^{m+1}_{\lambda,h} - u^m_{\lambda,h}}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(u^{m+1}_{\lambda,h}, v_h\right)\right]\\
&= \left(\frac{u^{m+1}_{\lambda,h} - u^m_{\lambda,h}}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(u^{m+1}_{\lambda,h}, v_h\right)\\
&\quad - \left[\left(\frac{u^{m+1}_{\lambda,h} - u^m_{\lambda,h}}{\Delta t}, v_h\right) + a_{\lambda m+1}\left(u^{m+1}_{\lambda,h}, v_h\right)\right]\\
&\quad + \left(1 - e^{-\lambda\Delta t}\right) e^{-\lambda t^m} \left(f^{m+1}, v_h\right)\\
&\quad + \left(1 - e^{\lambda\Delta t}\right) \left(\frac{\widetilde{u}^{m+1}_h}{\Delta t}, v_h\right) + \left(1 - e^{\lambda\Delta t}\right) a_{m+1}\left(\widetilde{u}^{m+1}_h, v_h\right) + \lambda \left(\widetilde{u}^{m+1}_h, v_h\right)\\
&= \left(1 - e^{-\lambda\Delta t}\right) e^{-\lambda t^m} \left(f^{m+1}, v_h\right)\\
&\quad + \left(1 - e^{\lambda\Delta t}\right) \left(\frac{\widetilde{u}^{m+1}_h}{\Delta t}, v_h\right) + \left(1 - e^{\lambda\Delta t}\right) a_{m+1}\left(\widetilde{u}^{m+1}_h, v_h\right) + \lambda \left(\widetilde{u}^{m+1}_h, v_h\right)\\
&=: \left(r^m_w, v_h\right).
\end{aligned} \tag{3.323}$$

Next we will further simplify this expression for $\left(r^m_w, v_h\right)$. We begin by eliminating the term containing f^{m+1} by invoking the relation provided by Scheme 3.80 for u^m_h. Using relation (3.310) between $\widetilde{u}_h^m$ and u_h^m we proceed from (3.323) by

$$\begin{aligned}(r_w^m, &v_h)\\ &= \left(1-e^{-\lambda\Delta t}\right)e^{-\lambda t^m}\left(f^{m+1},v_h\right)\\ &\quad+\left(1-e^{\lambda\Delta t}\right)\left(\frac{\widetilde{u}_h^{m+1}}{\Delta t},v_h\right)+\left(1-e^{\lambda\Delta t}\right)a_{m+1}\left(\widetilde{u}_h^{m+1},v_h\right)+\lambda\left(\widetilde{u}_h^{m+1},v_h\right)\\ &= \left(1-e^{-\lambda\Delta t}\right)e^{-\lambda t^m}\left(f^{m+1},v_h\right)+\lambda\left(\widetilde{u}_h^{m+1},v_h\right)\\ &\qquad+\left(1-e^{\lambda\Delta t}\right)e^{-\lambda t^{m+1}}\left[\left(\frac{u_h^{m+1}}{\Delta t},v_h\right)+a_{m+1}\left(u_h^{m+1},v_h\right)\right].\end{aligned}\tag{3.324}$$

As a next step, we artificially expand the brackets of (3.324) and use the equation given by Scheme 3.80 by

$$\begin{aligned}\left(\frac{u_h^{m+1}}{\Delta t},v_h\right)&+a_{m+1}\left(u_h^{m+1},v_h\right)\\ &=\left(\frac{u_h^{m+1}-u_h^m}{\Delta t},v_h\right)+a_{m+1}\left(u_h^{m+1},v_h\right)+\left(\frac{u_h^m}{\Delta t},v_h\right)\\ &=\left(f^{m+1},v_h\right)+\left(\frac{u_h^m}{\Delta t},v_h\right).\end{aligned}\tag{3.325}$$

Now we exploit the equidistant spacing of our time grid, $t^{m+1}=t^m+\Delta t$, leading to

$$\left(1-e^{\lambda\Delta t}\right)e^{-\lambda t^{m+1}}=-\left(1-e^{-\lambda\Delta t}\right)e^{-\lambda t^m}.\tag{3.326}$$

Inserting (3.325) and (3.326) into (3.324) yields

$$\begin{aligned}(r_w^m,v_h)=&\left(1-e^{-\lambda\Delta t}\right)e^{-\lambda t^m}\left(f^{m+1},v_h\right)+\lambda\left(\widetilde{u}_h^{m+1},v_h\right)\\ &-\left(1-e^{-\lambda\Delta t}\right)e^{-\lambda t^m}\left[\left(f^{m+1},v_h\right)+\left(\frac{u_h^m}{\Delta t},v_h\right)\right]\\ =&\,\lambda\left(\widetilde{u}_h^{m+1},v_h\right)-\left(1-e^{-\lambda\Delta t}\right)e^{-\lambda t^m}\left(\frac{u_h^m}{\Delta t},v_h\right)\\ =&\,\lambda\left(\widetilde{u}_h^{m+1},v_h\right)-\frac{\left(1-e^{-\lambda\Delta t}\right)}{\Delta t}\left(\widetilde{u}_h^m,v_h\right)\\ =&\,\lambda\left(\widetilde{u}_h^{m+1},v_h\right)-\overline{\lambda}(\Delta t)\left(\widetilde{u}_h^m,v_h\right),\end{aligned}\tag{3.327}$$

wherein the function $\overline{\lambda}:\mathbb{R}^+\to\mathbb{R}$ is defined as in (3.316). Consider the upcoming Lemma 3.85 for some properties of the function $\overline{\lambda}$.

We rewrite the expression of r_w^m in (3.327), the residuals of the w^m, by invoking the definition of w^m in (3.313) and get

$$\begin{aligned}(r_w^m,v_h)&=\lambda\left(\widetilde{u}_h^{m+1},v_h\right)-\overline{\lambda}(\Delta t)\left(\widetilde{u}_h^m,v_h\right)\\ &=\lambda\left(w^{m+1}+u_{\lambda,h}^{m+1},v_h\right)-\overline{\lambda}(\Delta t)\left(w^m+u_{\lambda,h}^m,v_h\right).\end{aligned}\tag{3.328}$$

We have thus derived an expression for the right hand side in (3.317). Note that the w^m reappear on the right side of their θ scheme that is a scheme with coercive bilinear form $a_{\lambda \cdot}(\cdot,\cdot)$. □

Collecting our results, for later reference we cast the claim of Lemma 3.83 in the following theta scheme framework that has been validated by the lemma.

Theta Scheme 3.84 (The special residuals $\boldsymbol{w^m}$)
Set $\theta = 1$. Let $\widetilde{u}_h^m$, $m = 0,\ldots,M$, be given by Scheme 3.82 and let $u_{\lambda,h}^m$, $m = 0,\ldots,M$, be the solution to Scheme 3.81. Let

$$w^m = \widetilde{u}_h^m - u_{\lambda,h}^m \tag{3.329}$$

for all $m \in \{0,\ldots,M\}$. The w^m fulfill the scheme

$$\begin{aligned}\left(\frac{w^{m+1}-w^m}{\Delta t}, v_h\right) + a_{\lambda^{m+1}}\left(w^{m+1}, v_h\right) &= \left(r_w^m, v_h\right),\\ w^0 &= 0\end{aligned} \tag{3.330}$$

with

$$\left(r_w^m, v_h\right) = \lambda\left(w^{m+1} + u_{\lambda,h}^{m+1}, v_h\right) - \overline{\lambda}(\Delta t)\left(w^m + u_{\lambda,h}^m, v_h\right), \tag{3.331}$$

where the function $\overline{\lambda}$ is defined by

$$\overline{\lambda}(\Delta t) = \frac{1 - e^{-\lambda\Delta t}}{\Delta t}, \tag{3.332}$$

for all $\Delta t > 0$.

The function $\overline{\lambda}$ that was defined in (3.316) of Lemma 3.85 will play a decisive role in the upcoming analysis. The following lemma derives some convergence results for this auxiliary function $\overline{\lambda}$ that we will need, later.

Lemma 3.85 (On the function $\overline{\boldsymbol{\lambda}}$)
Let $\lambda > 0$ and let the function $\overline{\lambda} : \mathbb{R}^+ \to \mathbb{R}$ be defined as in (3.316) of Lemma 3.85 by

$$\overline{\lambda} : \Delta t \mapsto \frac{1 - e^{-\lambda\Delta t}}{\Delta t}. \tag{3.333}$$

Then, for Δt approaching zero from above, $\overline{\lambda}$ satisfies

i) $\lim_{\Delta t \downarrow 0} \overline{\lambda}(\Delta t) = \lambda$ *from below, in the sense that* $\overline{\lambda}(\Delta t) \leq \lambda$, $\forall \Delta t > 0$,

ii) $\lim_{\Delta t \downarrow 0} \frac{\lambda - \overline{\lambda}(\Delta t)}{\Delta t} = \frac{\lambda^2}{2}$ *from below, in the sense that* $\frac{\lambda - \overline{\lambda}(\Delta t)}{\Delta t} \leq \frac{\lambda^2}{2}$, $\forall \Delta t > 0$.

Class	**Quantity**	**Solution to**	**Scheme reference**
coercive	u	Weak problem	–
	u_h	Weak problem discretized in space	3.63
Gårding	u	Weak problem	–
	u_h^m	Weak problem, fully discretized	3.80
	u_λ	Related "coercified" problem	–
	$u_{\lambda,h}^m$	Related "coercified" problem, fully discretized	3.81
	$\widetilde{u}_h^m$	$\widetilde{u}_h^m = e^{-\lambda t^m} u_h^m$ and degenerate Gårding scheme	3.82
	w^m	$w^m = \widetilde{u}_h^m - u_{\lambda,h}^m$ and auxiliary scheme	3.84

Table 3.2 An overview of all Schemes that have been derived so far. In the previous section where we considered coercive PIDEs exclusively, only the solutions to the weak formulation of the problem and its fully discretized counterpart were involved. For the analysis of problems with an operator $\mathcal{A}$ of Gårding type many auxiliary quantities and associated schemes will contribute.

Proof

i) The limit of the first claim follows from l'Hôpital's rule. A well known lower bound for the exponential function is given by

$$\exp(x) \geq 1 + x, \qquad \forall x \in \mathbb{R},$$

from which we deduce

$$\exp(-\lambda \Delta t) \geq 1 - \lambda \Delta t, \qquad \forall \Delta t > 0. \tag{3.334}$$

This is equivalent to

$$\frac{1 - e^{-\lambda \Delta t}}{\Delta t} \leq \lambda, \qquad \forall \Delta t > 0,$$

from which convergence of $\overline{\lambda}$ to its limit from below follows immediately. This proves *i)*.

ii) Similarly, the limit of the second claim follows from applying l'Hôpital's rule twice. Convergence of $\Delta t \mapsto (\lambda - \overline{\lambda}(\Delta t))/\Delta t$ to its limit from below is derived from

$$\frac{\lambda - \overline{\lambda}(\Delta t)}{\Delta t} \leq \frac{\lambda^2}{2}, \quad \forall \Delta t > 0, \tag{3.335}$$

which we prove in the following. Invoking the definition of $\overline{\lambda}$, (3.335) holds if

$$\frac{\lambda^2}{2} - \frac{\lambda \Delta t - (1 - e^{-\lambda \Delta t})}{(\Delta t)^2} \geq 0, \qquad \forall \Delta t > 0,$$

which is the case if

$$(\lambda \Delta t)^2 - 2\lambda \Delta t + 2(1 - e^{-\lambda \Delta t}) \geq 0, \qquad \forall \Delta t \geq 0. \tag{3.336}$$

Inequality (3.336) holds for $\Delta t = 0$. The function $f : \Delta t \mapsto (\lambda\Delta t)^2 - 2\lambda\Delta t + 2(1 - e^{-\lambda\Delta t})$ is continuously differentiable. It is thus non-negative for all $\Delta t \geq 0$ if $\frac{\partial}{\partial\Delta t} f \geq 0$ for all $\Delta t \geq 0$. Consequently, we consider the derivative of f and find

$$\frac{\partial}{\partial\Delta t} f(\Delta t) = 2\lambda\left(\lambda\Delta t - 1 + e^{-\lambda\Delta t}\right) \geq 0, \qquad \forall\Delta t \geq 0$$

if and only if

$$\lambda\Delta t - 1 + e^{-\lambda\Delta t} \geq 0, \qquad \forall\Delta t \geq 0,$$

which is equivalent to

$$e^{-\lambda\Delta t} \geq 1 - \lambda\Delta t, \qquad \forall\Delta t \geq 0,$$

which is again validated by (3.334) and thus proves claim *ii)*.

This finishes the proof of the lemma. □

3.6.3.2 Stability of Gårding schemes

We derive a stability estimate for the discrete solution to the Gårding problem. To this end, the Gårding scheme is split up into a coercive scheme for which the results of Section 3.6.2 can be applied and a scheme for the auxiliary quantities w^m, $m \in \{1, \dots, M\}$. For the latter, a stability estimate is derived, as well. From stability of these two parts, stability of the whole Gårding scheme follows.

Corollary 3.86 (A stability estimate for w^m)
Let the w^m, $m = 0, \dots, M-1$, be solutions to Scheme 3.84. Choose constants

$$0 < C_1 < 2, \qquad C_2 \geq \frac{1}{\beta(2 - C_1)} \tag{3.337}$$

with β being the coercivity constant from (3.296)*. Then, the w^m fulfill the stability estimate*

$$\left\|w^M\right\|_H^2 + C_1\beta\Delta t \sum_{m=0}^{M-1} \left\|w^{m+1}\right\|_{V^s}^2 \leq \Delta t C_2 \sum_{m=0}^{M-1} \|r_w^m\|_{V_h^{s*}}^2. \tag{3.338}$$

Proof
These estimates are a direct application of Corollary 3.71 with $\theta = 1$ based on the uniform coercivity of bilinear form a_λ with coercivity constant β. □

Theorem 3.87 (Stability estimate for the Gårding scheme)
Let u_h^m, $m = 0, \dots, M-1$, be the solution to Scheme 3.80 with $\theta = 1$, where the associated bilinear form a is uniformly continuous with continuity constant α and of Gårding type with coercivity constant $\beta > 0$ and Gårding constant $\lambda > 0$. Choose constants

$$0 < C_1 < 2, \qquad C_2 \geq \frac{1}{\beta(2 - C_1)} \tag{3.339}$$

and let Δt be small enough,

$$\Delta t \leq \min\left\{1,\ \sqrt{\frac{C_1\beta}{8\lambda^2 e^{\lambda} C_2\left(\frac{\lambda^2}{4}+(\alpha+\lambda)^2\right)}},\ \frac{C_1\beta+16\lambda^2 e^{\lambda}C_2}{8\lambda^2 e^{\lambda}C_1C_2\beta}\right\}. \tag{3.340}$$

Then there exist positive constants C_3, C_4 such that the stability estimate

$$\left\|u_h^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_h^{m+1}\right\|_{V^s}^2 \leq C_3\left\|u_h^0\right\|_H^2 + C_4\Delta t\sum_{m=0}^{M-1}\left\|f^{m+1}\right\|_{V_h^{s\star}}^2 \tag{3.341}$$

holds.

Proof

We expand the left hand side of (3.341) by elementary calculations,

$$\begin{aligned}
&\left\|u_h^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_h^{m+1}\right\|_{V^s}^2\\
&\leq e^{2\lambda T}\left(\left\|e^{-\lambda t^M}u_h^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|e^{-\lambda t^{m+1}}u_h^{m+1}\right\|_{V^s}^2\right)\\
&= e^{2\lambda T}\left(\left\|u_{\lambda,h}^M+\widetilde{u}_h^M-u_{\lambda,h}^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}+\widetilde{u}_h^{m+1}-u_{\lambda,h}^{m+1}\right\|_{V^s}^2\right)\\
&\leq 2e^{2\lambda T}\Bigg(\left\|u_{\lambda,h}^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_{V^s}^2\\
&\qquad\qquad + \left\|w^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|w^{m+1}\right\|_{V^s}^2\Bigg).
\end{aligned} \tag{3.342}$$

Consider the w^m terms in (3.342), first. By Corollary 3.86 and $w^0 = 0$ we have the estimate

$$\left\|w^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|w^{m+1}\right\|_V^2 \leq \Delta t C_2\sum_{m=0}^{M-1}\left\|r_w^m\right\|_{V_h^{s\star}}^2. \tag{3.343}$$

By Theta Scheme 3.84 we have for any $v_h \in V_h^s$ that

$$\begin{aligned}
(r_w^m, v_h) &= \lambda\left(w^{m+1}+u_{\lambda,h}^{m+1}, v_h\right) - \overline{\lambda}(\Delta t)\left(w^m+u_{\lambda,h}^m, v_h\right)\\
&= \left(\lambda w^{m+1}-\overline{\lambda}(\Delta t)w^m, v_h\right) + \left(\lambda u_{\lambda,h}^{m+1}-\overline{\lambda}(\Delta t)u_{\lambda,h}^m, v_h\right),
\end{aligned} \tag{3.344}$$

with $\overline{\lambda}$ as defined in (3.332). We continue by considering the first bracket in (3.344),

$$\begin{aligned}
&\left(\lambda w^{m+1}-\overline{\lambda}(\Delta t)w^m, v_h\right)\\
&\quad= \left(\lambda-\overline{\lambda}(\Delta t)\right)\left(w^{m+1}, v_h\right) + \overline{\lambda}(\Delta t)\left(w^{m+1}-w^m, v_h\right)\\
&\quad= \left(\lambda-\overline{\lambda}(\Delta t)\right)\left(w^{m+1}, v_h\right) + \overline{\lambda}(\Delta t)\Delta t\left(\frac{w^{m+1}-w^m}{\Delta t}, v_h\right)\\
&\quad= \left(\lambda-\overline{\lambda}(\Delta t)\right)\left(w^{m+1}, v_h\right) + \overline{\lambda}(\Delta t)\Delta t\left[(r_w^m, v_h) - a_{\lambda m+1}\left(w^{m+1}, v_h\right)\right],
\end{aligned} \tag{3.345}$$

where we used the relation provided by Scheme 3.84 for w^m in the last step. By definition of $\overline{\lambda}$, $\lambda - \overline{\lambda}(\Delta t) \neq 0$ for all $\Delta t > 0$. Thus, combining (3.344) and (3.345) we conclude

$$\begin{aligned} \left(r_w^m, v_h\right) = \frac{1}{1-\overline{\lambda}(\Delta t)\Delta t} \Big[&(\lambda - \overline{\lambda}(\Delta t)) \left(w^{m+1}, v_h\right) \\ &- \overline{\lambda}(\Delta t)\Delta t \, a_{\lambda m+1}\left(w^{m+1}, v_h\right) + \left(\lambda u_{\lambda,h}^{m+1} - \overline{\lambda}(\Delta t) u_{\lambda,h}^m, v_h\right) \Big]. \end{aligned} \tag{3.346}$$

Consequently,

$$\begin{aligned} \|r_w^m\|_{V_h^{s\star}} \leq \frac{1}{1-\overline{\lambda}(\Delta t)\Delta t} \Bigg[&\left(\frac{\lambda - \overline{\lambda}(\Delta t)}{\Delta t}\right) \Delta t \|w^{m+1}\|_{V^s} \\ &+ \overline{\lambda}(\Delta t)(\Delta t)\alpha_\lambda \|w^{m+1}\|_{V^s} + \lambda \left\|u_{\lambda,h}^{m+1}\right\|_H + \overline{\lambda}(\Delta t)\|u_{\lambda,h}^m\|_H \Bigg], \end{aligned} \tag{3.347}$$

with $\alpha_\lambda = \alpha + \lambda$ the continuity constant of bilinear form a_λ. Invoking the definition of $\overline{\lambda}$ and the first assumption on Δt in (3.340),

$$\frac{1}{1-\overline{\lambda}(\Delta t)\Delta t} = e^{\lambda \Delta t} \leq e^\lambda. \tag{3.348}$$

Thus, with (3.348) the convergence results of Lemma 3.85 imply the estimates

$$\begin{aligned} \|r_w^m\|_{V_h^{s\star}}^2 &\leq 4e^\lambda \Bigg[\left(\left(\frac{\lambda^2}{2}\right)^2 + \lambda^2 \alpha_\lambda^2 \right) (\Delta t)^2 \|w^{m+1}\|_{V^s}^2 \\ &\qquad + \lambda^2 \left\|u_{\lambda,h}^{m+1}\right\|_H^2 + \lambda^2 \|u_{\lambda,h}^m\|_H^2 \Bigg] \\ &= 4\lambda^2 e^\lambda \left[\left(\frac{\lambda^2}{4} + \alpha_\lambda^2\right) (\Delta t)^2 \|w^{m+1}\|_{V^s}^2 + \left\|u_{\lambda,h}^{m+1}\right\|_H^2 + \|u_{\lambda,h}^m\|_H^2 \right]. \end{aligned} \tag{3.349}$$

Combining (3.349) and (3.343) we get

$$\begin{aligned} \|w^M\|_H^2 + &\left(C_1\beta - C_2 4\lambda^2 e^\lambda \left(\frac{\lambda^2}{4} + \alpha_\lambda^2\right)(\Delta t)^2 \right) \Delta t \sum_{m=0}^{M-1} \|w^{m+1}\|_{V^s}^2 \\ &\leq 4\lambda^2 e^\lambda C_2 \Delta t \sum_{m=0}^{M-1} \left(\left\|u_{\lambda,h}^{m+1}\right\|_H^2 + \|u_{\lambda,h}^m\|_H^2 \right) \\ &\leq 4\lambda^2 e^\lambda C_2 \Delta t \|u_{\lambda,h}^0\|_H^2 + 8\lambda^2 e^\lambda C_2 \Delta t \sum_{m=0}^{M-1} \left\|u_{\lambda,h}^{m+1}\right\|_H^2. \end{aligned} \tag{3.350}$$

By the second assumption on Δt in (3.340), we have

$$C_1\beta - C_2 4\lambda^2 e^\lambda \left(\frac{\lambda^2}{4} + \alpha_\lambda^2\right)(\Delta t)^2 \geq \frac{C_1\beta}{2}$$

and thus, we deduce from (3.350),

$$\begin{aligned}\left\|w^M\right\|_H^2+&C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|w^{m+1}\right\|_{V^s}^2\\&\leq 8\lambda^2e^\lambda C_2\Delta t\left\|u_{\lambda,h}^0\right\|_H^2+16\lambda^2e^\lambda C_2\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_H^2.\end{aligned}\tag{3.351}$$

Returning to (3.342) and assembling our findings we get

$$\begin{aligned}&\left\|u_h^M\right\|_H^2+C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_h^{m+1}\right\|_{V^s}^2\\&\leq 2e^{2\lambda T}\left(\left\|u_{\lambda,h}^M\right\|_H^2+C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_{V^s}^2+\left\|w^M\right\|_H^2+C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|w^{m+1}\right\|_{V^s}^2\right)\\&\leq 2e^{2\lambda T}\left(\left\|u_{\lambda,h}^M\right\|_H^2+C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_{V^s}^2\right.\\&\qquad\left.+8\lambda^2e^\lambda C_2\Delta t\left\|u_{\lambda,h}^0\right\|_H^2+16\lambda^2e^\lambda C_2\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_H^2\right)\\&\leq 2e^{2\lambda T}\left(8\lambda^2e^\lambda C_2\Delta t\left\|u_{\lambda,h}^0\right\|_H^2+\left\|u_{\lambda,h}^M\right\|_H^2+\left(C_1\beta+16\lambda^2e^\lambda C_2\right)\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_{V^s}^2\right)\\&\leq 2e^{2\lambda T}\left(8\lambda^2e^\lambda C_2\Delta t\left\|u_{\lambda,h}^0\right\|_H^2+\left(1+\frac{16\lambda^2e^\lambda C_2}{C_1\beta}\right)\left(\left\|u_{\lambda,h}^M\right\|_H^2+C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_{V^s}^2\right)\right).\end{aligned}\tag{3.352}$$

We apply Corollary 3.71 which gives

$$\left\|u_{\lambda,h}^M\right\|_H^2+C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1}\right\|_{V^s}^2\leq\left\|u_{\lambda,h}^0\right\|_H^2+C_2\Delta t\sum_{m=0}^{M-1}\left\|f_\lambda^{m+1}\right\|_{V_h^{s\star}}^2.\tag{3.353}$$

Inserting (3.353) into (3.352) thus gives

$$\begin{aligned}&\left\|u_h^M\right\|_H^2+C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|u_h^{m+1}\right\|_{V^s}^2\\&\leq 2e^{2\lambda T}\left(8\lambda^2e^\lambda C_2\Delta t\left\|u_{\lambda,h}^0\right\|_H^2+\left(1+\frac{16\lambda^2e^\lambda C_2}{C_1\beta}\right)\left(\left\|u_{\lambda,h}^0\right\|_H^2+C_2\Delta t\sum_{m=0}^{M-1}\left\|f_\lambda^{m+1}\right\|_{V_h^{s\star}}^2\right)\right).\end{aligned}\tag{3.354}$$

By the third assumption on Δt in (3.340),

$$\frac{8\lambda^2e^\lambda C_2}{1+\frac{16\lambda^2e^\lambda C_2}{C_1\beta}}\Delta t\leq 1,$$

so (3.354) leads to

$$\begin{aligned}
\|u_h^M\|_H^2 &+ C_1\beta\Delta t\sum_{m=0}^{M-1}\|u_h^{m+1}\|_{V^s}^2\\
&\le 2e^{2\lambda T}\left(1+\frac{16\lambda^2e^\lambda C_2}{C_1\beta}\right)\left(2\|u_{\lambda,h}^0\|_H^2 + C_2\Delta t\sum_{m=0}^{M-1}\|f_\lambda^{m+1}\|_{V_h^{s\star}}^2\right).
\end{aligned}$$

Finally, since $f_\lambda^m = e^{-\lambda t^m}f^m$ and $u_{\lambda,h}^0 = u_h^0$, we arrive at

$$\begin{aligned}
\|u_h^M\|_H^2 &+ C_1\beta\Delta t\sum_{m=0}^{M-1}\|u_h^{m+1}\|_{V}^2\\
&\le 2e^{2\lambda T}\left(1+\frac{16\lambda^2e^\lambda C_2}{C_1\beta}\right)\left(2\|u_h^0\|_H^2 + C_2\Delta t\sum_{m=0}^{M-1}\|f^{m+1}\|_{V_h^{s\star}}^2\right).
\end{aligned}$$

Defining the constants

$$C_3 = 4e^{2\lambda T}\left(1+\frac{16\lambda^2e^\lambda C_2}{C_1\beta}\right), \qquad C_4 = 2e^{2\lambda T}\left(1+\frac{16\lambda^2e^\lambda C_2}{C_1\beta}\right)C_2 \tag{3.355}$$

yields the claim. □

Note that analogously to the coercive case of the previous section, the result (3.341) of Theorem 3.87 in a way describes that the solution of the discrete scheme is bounded by its initial data in a discrete $L^2(0,T,V^s)$ or $L^2(0,T,V_h^{s*})$ norm fashion, respectively.

3.6.3.3 Convergence of Gårding schemes

We prove convergence for time-inhomogeneous Gårding schemes. We begin by splitting the Gårding scheme into two parts. One is its "coercified" version for which the convergence result of Theorem 3.76 applies. The other is the scheme for the artificial quantities w^m, $m \in \{1,\dots,M\}$, as outlined by Theta Scheme 3.84, with residuals r_w^m, $m \in \{1,\dots,M\}$, defined therein. In Lemma 3.88, we derive an estimate for the sum over these normed residuals which depends on the sum of $\|w^m\|$, $m \in \{0,\dots,M\}$, and the sum over new quantities $\|U^m\|$, $m \in \{0,\dots,M\}$. In a second step we resolve this quasi-recurrence and reduce the estimate in Lemma 3.89 to the occurrence of terms $\|U^m\|$, $m \in \{0,\dots,M\}$, alone. In a third step, Lemma 3.93 derives an estimate for the sum over all $\|U^m\|$, $m \in \{0,\dots,M\}$, that splits into three parts. One is again the "coercified" scheme for which we already have a convergence result. The second is a sum of differences of the solution u_λ^m to the non-discretized "coercified" problem, to which Taylor's theorem is applied. The third part consists of a quantity depending on $\lambda - \overline{\lambda}(\Delta t)$, which converges by Lemma 3.85. Finally, Theorem 3.94 gathers all results and shows

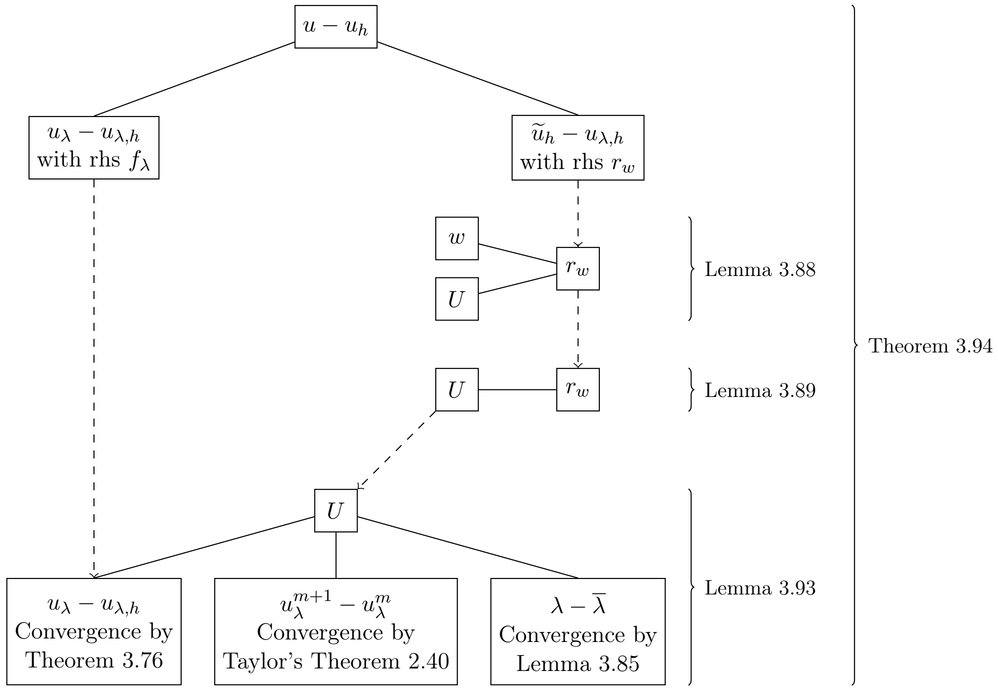
$u - u_h$
$u_\lambda - u_{\lambda,h}$ with rhs f_λ
$\widetilde{u}_h - u_{\lambda,h}$ with rhs r_w
w
U
r_w
Lemma 3.88
Theorem 3.94
U
r_w
Lemma 3.89
U
$u_\lambda - u_{\lambda,h}$ Convergence by Theorem 3.76
$u_\lambda^{m+1} - u_\lambda^m$ Convergence by Taylor's Theorem 2.40
$\lambda - \overline{\lambda}$ Convergence by Lemma 3.85
Lemma 3.93

Figure 3.20 A schematic overview over the convergence proof for Gårding schemes and the involved lemmas and quantities.

convergence of the Gårding scheme.

Figure 3.20 offers a schematic overview over the different lemmas and quantities that are involved in the proof of convergence. The right branch of the figure highlights the additional effort required to treat the Gårding case $\lambda > 0$.

Lemma 3.88 (Upper bounds for $\|r_w^m\|_H$)
Let $u_{\lambda,h}^m$, $m \in \{0, \dots, M\}$, be the solution of Scheme 3.81 and w^m, $m \in \{0, \dots, M\}$, be the solution to Scheme 3.84 and let $\overline{\lambda} : \mathbb{R}^+ \to \mathbb{R}$ be defined as in Lemma 3.85. Define

$$U^m = \lambda u_{\lambda,h}^{m+1} - \overline{\lambda}(\Delta t) u_{\lambda,h}^m, \qquad \forall m \in \{0, \dots, M-1\}, \tag{3.356}$$

and let r_w^m be the right hand side of Scheme 3.84 for the w^m. Then,

$$\|r_w^m\|_H^2 \le 4\lambda^2 \left(\left\|w^{m+1}\right\|_H^2 + \|w^m\|_H^2 \right) + 2\|U^m\|_H^2 \tag{3.357}$$

$$\le 4\lambda^2 \left(\left\|w^{m+1}\right\|_{V^s}^2 + \|w^m\|_{V^s}^2 \right) + 2\|U^m\|_H^2 \tag{3.358}$$

holds for all $m \in \{0, \dots, M-1\}$.

Proof
The proof is a straightforward calculation. We have

$$\begin{aligned}\|r_w^m\|_H^2 &= \left\|\lambda w^{m+1} - \overline{\lambda}(\Delta t)w^m + \lambda u_{\lambda,h}^{m+1} - \overline{\lambda}(\Delta t)u_{\lambda,h}^m\right\|_H^2\\ &= \left\|\lambda w^{m+1} - \overline{\lambda}(\Delta t)w^m + U^m\right\|_H^2\\ &\leq 4\lambda^2\left(\left\|w^{m+1}\right\|_H^2 + \|w^m\|_H^2\right) + 2\|U^m\|_H^2,\end{aligned}$$

since $\lim\limits_{\Delta t\downarrow 0}\overline{\lambda}(\Delta t) = \lambda$ from below by Lemma 3.85. Consequently, also

$$\|r_w^m\|_H^2 \leq 4\lambda^2\left(\left\|w^{m+1}\right\|_{V^s}^2 + \|w^m\|_{V^s}^2\right) + 2\|U^m\|_H^2, \tag{3.359}$$

which proves the claim. □

Lemma 3.88 provides us with an upper bound for the residuals of the artificial quantities w^m, $m \in \{0,\ldots,M\}$. Within that upper bound, however, those artificial quantities reappear. Using Gronwall's Lemma C.1, the following result resolves that recurrence and reduces the upper bound to an exclusive dependence on the new quantities U^m, $m \in \{0,\ldots,M-1\}$, as defined by (3.356).

Lemma 3.89 (Non-recursive bounds for $\left\|r_w^m\right\|_H$)
Let r_w^m, $m \in 0,\ldots,M-1$, be the right hand side of Scheme 3.84 wherein the bilinear form a_λ is coercive uniformly in time with coercivity constant $\beta > 0$. Let U^m be defined as in (3.356) *of Lemma 3.88 and $\lambda > 0$ as given therein. Assume*

$$\lambda > \frac{1}{\sqrt{8}}\beta. \tag{3.360}$$

Choose positive constants

$$0 < C_1 < 2, \qquad C_2 \geq \frac{1}{\beta(2-C_1)}$$

and assume further Δt to be small enough,

$$0 < \Delta t \leq \frac{1}{8C_2\lambda^2 - C_1\beta}. \tag{3.361}$$

Then $\exists C_5 > 0$ such that

$$\sum_{m=0}^{M-1}\|r_w^m\|_H^2 \leq C_5\sum_{m=0}^{M-1}\|U^m\|_H^2 \tag{3.362}$$

holds.

Before we give a proof for the claim of Lemma 3.89, the following remark comments on the time stepping condition (3.361) of the Lemma.

Remark 3.90 (On the time stepping condition in Lemma 3.89)
In Lemma 3.89, the prescribed interval for Δt*. In condition* (3.361)

$$8C_2\lambda^2 > C_1\beta \Leftrightarrow \lambda > \sqrt{\frac{C_1\beta}{8C_2}}, \tag{3.363}$$

which holds if

$$\sqrt{\frac{C_1\beta}{8C_2}} \geq \frac{1}{\sqrt{8}}\beta\sqrt{C_1(2-C_1)} \tag{3.364}$$

by the interval that the value of $C_2 > 0$ *is chosen from. Given the set of possible values for* C_1*, the expression* $\sqrt{C_1(2-C_1)}$ *is bounded by* 1*. Consequently, Inequality* (3.364) *and thus the inequalities in* (3.363) *hold if condition* (3.360) *is satisfied. In other words, by condition* (3.360) *the condition on* Δt *in* (3.361) *is well-posed. When*

$$8C_2\lambda^2 - C_1\beta > 0$$

then trivially also

$$4\lambda^2 - \frac{C_1\beta}{2C_2} > 0$$

which we state here for later use.

Proof (of Lemma 3.89)
We have by Corollary 3.86 and Remark 3.61 that

$$\left\|w^M\right\|_H^2 + C_1\beta\Delta t\sum_{m=0}^{M-1}\left\|w^{m+1}\right\|_{V^s}^2 \leq \Delta t C_2\sum_{m=0}^{M-1}\|r_w^m\|_H^2. \tag{3.365}$$

By Lemma 3.88 we have

$$\|r_w^m\|_H^2 \leq 4\lambda^2\left(\left\|w^{m+1}\right\|_H^2 + \|w^m\|_H^2\right) + 2\|U^m\|_H^2 \tag{3.366}$$

for all $m \in \{0,\dots,M-1\}$.

The two inequalities (3.365) and (3.366) are intertwined in the sense that combining them leaves us with w^m terms occurring on both sides of the resulting inequality. In order to confirm the claim (3.362), this entanglement has to be resolved. We take inequality (3.365) as a starting point.

By the fact that $\|v\|_H \leq \|v\|_{V^s}$ for all $v \in V^s$ and since $\left\|w^0\right\|_H = 0$ by Scheme 3.84 and further by (3.366) we thus have

$$\begin{aligned}\left\|w^M\right\|_H^2 + C_1\beta\frac{1}{2}\Delta t\sum_{m=0}^{M-1}&\left(\left\|w^{m+1}\right\|_H^2 + \|w^m\|_H^2\right)\\ &\leq \Delta t C_2\sum_{m=0}^{M-1}\left\{4\lambda^2\left(\left\|w^{m+1}\right\|_H^2 + \|w^m\|_H^2\right) + 2\|U^m\|_H^2\right\}\end{aligned} \tag{3.367}$$

from which we deduce

$$\begin{aligned}\|w^M\|_H^2 &\le \Delta t C_2 \sum_{m=0}^{M-1} \left\{ \left[4\lambda^2 - \frac{C_1\beta}{2C_2} \right] \left(\|w^{m+1}\|_H^2 + \|w^m\|_H^2 \right) + 2\|U^m\|_H^2 \right\} \\ &= \Delta t C_2 \sum_{m=0}^{M-1} \left\{ \widetilde{C}_1 \left(\|w^{m+1}\|_H^2 + \|w^m\|_H^2 \right) + 2\|U^m\|_H^2 \right\}\end{aligned} \tag{3.368}$$

wherein we defined

$$\widetilde{C}_1 = \left[4\lambda^2 - \frac{C_1\beta}{2C_2} \right]. \tag{3.369}$$

By assumption (3.360), $\widetilde{C}_1 > 0$ as outlined by Remark 3.90.

Inequality (3.368) does not only hold for w^M, as the result of the final step of our θ scheme, but for all $w^{\widetilde{M}}$, $1 \le \widetilde{M} \le M$. The function $w^{\widetilde{M}}$ can be interpreted as the solution associated with the final time step of Scheme 3.84 with $\widetilde{M} \le M$ time steps instead of M and with time horizon $\widetilde{T} = \Delta t \widetilde{M}$ instead of $T = \Delta t M$ with time discretization parameter $\Delta t = \widetilde{T}/\widetilde{M} = T/M$ yielding a solution on the equidistant time grid $(\widetilde{T}, \widetilde{M}, \Delta t)$ as defined in Definition 3.53. It is important to realize that this Δt is thus *the very same* for both the solution of the original scheme, $\{w^0, w^1, \dots, w^M\}$, as well as the solution of the second, shortened scheme, $\{w^0, w^1, \dots, w^{\widetilde{M}}\}$.

Relabeling the summation indices in (3.368) gives

$$\|w^m\|_H^2 \le \Delta t C_2 \sum_{k=0}^{m-1} \left\{ \widetilde{C}_1 \left(\left\|w^{k+1}\right\|_H^2 + \left\|w^k\right\|_H^2 \right) + 2\left\|U^k\right\|_H^2 \right\}. \tag{3.370}$$

Continuing with elementary calculations in (3.370) and recalling that $\left\|w^0\right\|_H = 0$ we get

$$\|w^m\|_H^2 \le 2\Delta t C_2 \sum_{k=0}^{m-1} \left\{ \widetilde{C}_1 \left\|w^k\right\|_H^2 + \left\|U^k\right\|_H^2 \right\} + \Delta t C_2 \widetilde{C}_1 \|w^m\|_H^2, \tag{3.371}$$

which is equivalent to

$$\|w^m\|_H^2 \le \frac{2\Delta t C_2}{1 - \Delta t C_2 \widetilde{C}_1} \sum_{k=0}^{m-1} \left\{ \widetilde{C}_1 \left\|w^k\right\|_H^2 + \left\|U^k\right\|_H^2 \right\}, \tag{3.372}$$

since

$$1 - \Delta t C_2 \widetilde{C}_1 > 0$$

by the condition on Δt in (3.361). Now, defining q by

$$q = q(\Delta t) = \frac{\Delta t C_2 \widetilde{C}_1}{1 - \Delta t C_2 \widetilde{C}_1} \tag{3.373}$$

gives

$$\|w^m\|_H^2 \le 2q \sum_{k=0}^{m-1} \left\|w^k\right\|_H^2 + \frac{2q}{\widetilde{C}_1} \sum_{k=0}^{m-1} \left\|U^k\right\|_H^2. \tag{3.374}$$

By Gronwall's Lemma C.1, setting for $m \geq 0$

$$\begin{aligned} y_m &= \|w^m\|_H^2, \\ f_m &= \sum_{k=0}^{m-1} \frac{2q}{\widetilde{C}_1} \left\|U^k\right\|_H^2, \\ g_m &\equiv g = 2q, \end{aligned}$$

we deduce

$$\begin{aligned} \|w^m\|_H^2 \le & \sum_{k=0}^{m-1} \frac{2q}{\widetilde{C}_1} \left\|U^k\right\|_H^2 + \sum_{0\le j<m} \left[\left(\sum_{k=0}^{j-1} \frac{2q}{\widetilde{C}_1} \left\|U^k\right\|_H^2 \right) (2q) \prod_{j<i<m} (1+2q) \right] \\ &= \frac{2q}{\widetilde{C}_1} \sum_{k=0}^{m-1} \left\|U^k\right\|_H^2 + 2q \frac{2q}{\widetilde{C}_1} \sum_{j=0}^{m-1} \left(\sum_{k=0}^{j-1} \left\|U^k\right\|_H^2 \right) (1+2q)^{m-j-1} \\ &\le \frac{2q}{\widetilde{C}_1} \left(\sum_{k=0}^{m-1} \left\|U^k\right\|_H^2 + 2q(1+2q)^m \sum_{j=0}^{m-1} \sum_{k=0}^{j-1} \left\|U^k\right\|_H^2 \right). \end{aligned} \tag{3.375}$$

An elementary induction shows that the equality in

$$\sum_{j=0}^{m-1} \sum_{k=0}^{j-1} \left\|U^k\right\|_H^2 = \sum_{k=0}^{m-2} (m-1-k) \left\|U^k\right\|_H^2 \le (m-1) \sum_{k=0}^{m-2} \left\|U^k\right\|_H^2 \tag{3.376}$$

holds for all $0 \le m \le M$. Using (3.376) in (3.375) yields

$$\|w^m\|_H^2 \le \frac{2q}{\widetilde{C}_1} \left(\sum_{k=0}^{m-1} \left\|U^k\right\|_H^2 + 2q(1+2q)^m (m-1) \sum_{k=0}^{m-2} \left\|U^k\right\|_H^2 \right). \tag{3.377}$$

Consequently,

$$\begin{aligned} \sum_{m=0}^{M} \|w^m\|_H^2 \le \frac{2q}{\widetilde{C}_1} \Bigg(& \sum_{m=0}^{M} \sum_{k=0}^{m-1} \left\|U^k\right\|_H^2 \\ & + 2q \sum_{m=0}^{M} (1+2q)^m (m-1) \sum_{k=0}^{m-2} \left\|U^k\right\|_H^2 \Bigg). \end{aligned} \tag{3.378}$$

Furthermore, the same induction as above shows that

$$\sum_{m=0}^{\widetilde{M}} \sum_{k=0}^{m-1} \left\|U^k\right\|_H^2 = \sum_{k=0}^{\widetilde{M}-1} (\widetilde{M}-k) \left\|U^k\right\|_H^2, \qquad \forall 1 \le \widetilde{M} \le M. \tag{3.379}$$

Using (3.379) we thus have

$$\sum_{m=0}^{M}\sum_{k=0}^{m-2}\left\|U^k\right\|_H^2 \le \sum_{m=0}^{M}\sum_{k=0}^{m-1}\left\|U^k\right\|_H^2 = \sum_{k=0}^{M-1}(M-k)\left\|U^k\right\|_H^2 \le M\sum_{k=0}^{M-1}\left\|U^k\right\|_H^2. \tag{3.380}$$

With

$$\big(1+2q\big)^m(m-1) \le \big(1+2q\big)^M M, \qquad \text{for } m\in\{0,\dots,M\} \tag{3.381}$$

and by applying (3.380) to both of its sums we develop (3.378) into

$$\begin{aligned}\sum_{m=0}^{M}\|w^m\|_H^2 &\le \frac{2q}{\widetilde{C}_1}\left(\sum_{m=0}^{M}\sum_{k=0}^{m-1}\left\|U^k\right\|_H^2 + 2q\sum_{m=0}^{M}\big(1+2q\big)^m(m-1)\sum_{k=0}^{m-2}\left\|U^k\right\|_H^2\right)\\ &\le \frac{2q}{\widetilde{C}_1}\left(M\sum_{k=0}^{M-1}\left\|U^k\right\|_H^2 + 2q\big(1+2q\big)^M M\,M\sum_{k=0}^{M-1}\left\|U^k\right\|_H^2\right)\\ &= \frac{2qM}{\widetilde{C}_1}\left(\left(1+2qM\big(1+2q\big)^M\right)\sum_{m=0}^{M-1}\|U^m\|_H^2\right).\end{aligned} \tag{3.382}$$

Define

$$\widetilde{c} = 8C_2\lambda^2 - C_1\beta. \tag{3.383}$$

By the time stepping condition (3.361) we have

$$\Delta t \le \frac{1}{\widetilde{c}}$$

and can write

$$\widetilde{C}_1 = \frac{\widetilde{c}}{2C_2}.$$

Recalling the definition of q in (3.373), the relation $\Delta t = T/M$ and above calculations, we deduce

$$qM = \frac{\Delta t C_2\widetilde{C}_1}{1-\Delta t C_2\widetilde{C}_1}M \le \frac{\frac{T}{M}C_2\frac{\widetilde{c}}{2C_2}}{1-\frac{1}{\widetilde{c}}C_2\frac{\widetilde{c}}{2C_2}}M = \frac{\frac{\widetilde{c}}{2}}{1-\frac{1}{2}}T = \widetilde{c}T. \tag{3.384}$$

Also, by the very same ingredients,

$$\big(1+2q\big)^M \le \left(1+\frac{2\widetilde{c}T}{M}\right)^M. \tag{3.385}$$

It is well known that

$$\lim_{\widetilde{M}\to\infty}\left(1+\frac{x}{\widetilde{M}}\right)^{\widetilde{M}} = \exp(x), \qquad \forall x\in\mathbb{R} \tag{3.386}$$

and that this convergence occurs from below. Consequently, incorporating (3.386) in (3.385) gives

$$\big(1+2q\big)^M \le \exp\left(2\widetilde{c}T\right). \tag{3.387}$$

Taking (3.384) and (3.387) and returning to (3.382) gives

$$\begin{aligned}\sum_{m=0}^{M} \|w^m\|_H^2 &\leq \frac{2qM}{\widetilde{C}_1}\left(\left(1+2qM\left(1+2q\right)^M\right)\sum_{m=0}^{M-1}\|U^m\|_H^2\right)\\ &\leq \frac{2\widetilde{c}T}{\widetilde{C}_1}\left(1+2\widetilde{c}T\exp\left(2\widetilde{c}T\right)\right)\sum_{m=0}^{M-1}\|U^m\|_H^2 \\ &= \widetilde{C}_2\sum_{m=0}^{M-1}\|U^m\|_H^2\end{aligned} \tag{3.388}$$

with

$$\widetilde{C}_2 = \frac{2\widetilde{c}T}{\widetilde{C}_1}\left(1+2\widetilde{c}T\exp\left(2\widetilde{c}T\right)\right) > 0. \tag{3.389}$$

We know by Lemma 3.88 that

$$\|r_w^m\|_H^2 \leq 4\lambda^2\left(\left\|w^{m+1}\right\|_H^2 + \|w^m\|_H^2\right) + 2\|U^m\|_H^2 \tag{3.390}$$

for all $m \in \{0,\dots,M-1\}$. Thus,

$$\begin{aligned}\sum_{m=0}^{M-1}\|r_w^m\|_H^2 &\leq 4\lambda^2\left(\sum_{m=0}^{M-1}\left\|w^{m+1}\right\|_H^2 + \sum_{m=0}^{M-1}\|w^m\|_H^2\right) + 2\sum_{m=0}^{M-1}\|U^m\|_H^2\\ &= 4\lambda^2\left(\sum_{m=1}^{M}\|w^m\|_H^2 + \sum_{m=0}^{M-1}\|w^m\|_H^2\right) + 2\sum_{m=0}^{M-1}\|U^m\|_H^2\\ &\leq 8\lambda^2\sum_{m=0}^{M}\|w^m\|_H^2 + 2\sum_{m=0}^{M-1}\|U^m\|_H^2.\end{aligned} \tag{3.391}$$

Now inserting (3.388) gives

$$\begin{aligned}\sum_{m=0}^{M-1}\|r_w^m\|_H^2 &\leq 8\lambda^2\widetilde{C}_2\sum_{m=0}^{M-1}\|U^m\|_H^2 + 2\sum_{m=0}^{M-1}\|U^m\|_H^2\\ &= 2\left(1+4\lambda^2\widetilde{C}_2\right)\sum_{m=0}^{M-1}\|U^m\|_H^2.\end{aligned} \tag{3.392}$$

Defining

$$C_5 = 2\left(1+4\lambda^2\widetilde{C}_2\right) \tag{3.393}$$

yields the claim. □

Remark 3.91 (Interpretation of restriction (3.360)**)**
Condition (3.360) *is only a mild restriction. Recall that a coercive bilinear form* $a^{coercive}(\cdot,\cdot) : V^s \times V^s \to \mathbb{R}$ *satisfies*

$$a^{coercive}(v,v) \geq \beta\|v\|_{V^s}^2, \qquad \forall v \in V^s, \tag{3.394}$$

for some $\beta > 0$ whereas a bilinear form $a^{Gårding}(\cdot,\cdot) : V^s \times V^s \to \mathbb{R}$ of Gårding type satisfies the weaker inequality

$$a^{Gårding}(v,v) \geq \beta \|v\|_{V^s}^2 - \lambda \|v\|_H^2, \qquad \forall v \in V^s, \tag{3.395}$$

for some $\beta > 0$, $\lambda \geq 0$. We know that $\|v\|_H \leq \|v\|_{V^s}$ for all $v \in V^s$, so if assumption (3.360) *in Lemma 3.89 is violated by $a^{Gårding}$ in the sense that $\lambda \leq \beta/\sqrt{8} < \beta$, we have*

$$\begin{aligned} a^{Gårding}(v,v) &\geq \beta \|v\|_{V^s}^2 - \lambda \|v\|_H^2 \\ &\geq \beta \|v\|_{V^s}^2 - \lambda \|v\|_{V^s}^2 = (\beta - \lambda)\|v\|_{V^s}^2, \qquad \forall v \in V^s. \end{aligned}$$

Consequently, $a^{Gårding}$ is not a genuine Gårding bilinear form but a coercive bilinear form with coercivity constant $\widetilde{\beta} = \beta - \lambda > 0$ and thus Theorem 3.76 directly applies for the confirmation of convergence of the associated scheme.

Remark 3.92 (Disregarding condition (3.360)**)**
Condition (3.360) *can be disregarded, if we define $\widetilde{C}_1$ in* (3.369) *in the proof of Lemma 3.89 differently via*

$$\widetilde{C}_1' = 4\lambda^2 > \widetilde{C}_1. \tag{3.396}$$

This basically means disregarding the large sum on the left hand side in (3.367) *in the proof of Lemma 3.89 making the upper estimate for $\left\|w^M\right\|_H^2$ in* (3.368) *weaker. As a consequence, constant C_5 defined in* (3.393) *at the end of the proof becomes larger and the time stepping condition for Δt in* (3.361) *stricter. Note, however, that constants C_1 and C_2 must be chosen differently, such that the time stepping condition is still well-posed.*

Lemma 3.93 (Upper bound for $\|U^m\|_H^2$)
Let U^m be given by

$$U^m = \lambda u_{\lambda,h}^{m+1} - \overline{\lambda}(\Delta t) u_{\lambda,h}^m, \qquad \forall m \in \{0,\dots,M-1\},$$

as introduced in (3.356) *of Lemma 3.88. Then there is an upper bound $\|U^m\|_H^2$ in the form of*

$$\begin{aligned} \|U^m\|_H^2 \leq 4\lambda^2 \bigg(&\left\|u_{\lambda,h}^{m+1} - u_\lambda^{m+1}\right\|_H^2 + \left\|u_{\lambda,h}^m - u_\lambda^m\right\|_H^2 \\ &+ \left\|u_\lambda^{m+1} - u_\lambda^m\right\|_H^2 \bigg) + 4\left(\lambda - \overline{\lambda}(\Delta t)\right)^2 \|u_\lambda^m\|_H^2 \end{aligned} \tag{3.397}$$

for all $m \in \{0,\dots,M-1\}$.

Proof
The proof is a straightforward calculation. By definition of U^m in (3.356) of Lemma 3.88,

we have with $m \in \{0, \ldots, M-1\}$ and by using Lemma 3.85

$$\begin{aligned}
\|U^m\|_H^2 &= \left\| \lambda u_{\lambda,h}^{m+1} - \overline{\lambda}(\Delta t) u_{\lambda,h}^m \right\|_H^2 \\
&= \left\| \lambda \left(u_{\lambda,h}^{m+1} - u_\lambda^{m+1} \right) - \overline{\lambda}(\Delta t) \left(u_{\lambda,h}^m - u_\lambda^m \right) + \lambda u_\lambda^{m+1} - \overline{\lambda}(\Delta t) u_\lambda^m \right\|_H^2 \\
&= \left\| \lambda \left(u_{\lambda,h}^{m+1} - u_\lambda^{m+1} \right) - \overline{\lambda}(\Delta t) \left(u_{\lambda,h}^m - u_\lambda^m \right) + \lambda \left(u_\lambda^{m+1} - u_\lambda^m \right) + \left(\lambda - \overline{\lambda}(\Delta t) \right) u_\lambda^m \right\|_H^2 \\
&\leq 4 \Big(\lambda^2 \left(\left\| u_{\lambda,h}^{m+1} - u_\lambda^{m+1} \right\|_H^2 + \left\| u_{\lambda,h}^m - u_\lambda^m \right\|_H^2 \right) \\
&\qquad\quad + \lambda^2 \left\| u_\lambda^{m+1} - u_\lambda^m \right\|_H^2 + \left(\lambda - \overline{\lambda}(\Delta t) \right)^2 \|u_\lambda^m\|_H^2 \Big),
\end{aligned}$$

which proves the claim. □

Theorem 3.94 (Convergence of the Gårding scheme)
Let $u \in W^1(0,T;V^t,H)$, $t > \alpha_{\mathcal{A}}/2$, be the weak solution to problem (3.172), where the operator is associated with a bilinear form a that is continuous and satisfies a Gårding inequality with respect to V^s uniformly in time. Let u be smooth enough in the sense that $u \in C^2([0,T],H)$. Let $(u_h^m)_{m\in\{0,\ldots,M\}}$ be the solution to the associated Theta Scheme 3.63 with $\theta = 1$ and assume further

i) The approximation property Assumption 3.A holds for some function Υ and some constant C_Υ

ii) The inverse property Assumption 3.B is satisfied

iii) Assumption 3.C on the projector P_h holds

iv) Assumption 3.D on the initial condition is satisfied

Then there exists a constant $C_6 > 0$ such that the convergence estimate

$$\begin{aligned}
&\left\| u^M - u_h^M \right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\| u^{m+1} - u_h^{m+1} \right\|_{V^s}^2 \\
&\leq \; C_6 (3 + \Delta t) \max_{0 \leq \tau \leq T} \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u_\lambda(\tau)) \\
&\quad + C_6 (\Delta t)^2 \left(\int_0^T \|\ddot{u}_\lambda(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau + \int_0^T \|\dot{u}_\lambda(\tau)\|_H^2 \,\mathrm{d}\tau + \max_{\tau \in [0,T]} \|u_\lambda(\tau)\|_H^2 \right) \\
&\quad + C_6 \int_0^T \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, \dot{u}_\lambda(\tau)) \,\mathrm{d}\tau
\end{aligned} \tag{3.398}$$

holds.

Proof
We derive

$$\begin{aligned}
\left\|u^M - u_h^M\right\|_H^2 &+ \Delta t \sum_{m=0}^{M-1} \left\|u^{m+1} - u_h^{m+1}\right\|_{V^s}^2 \\
&= e^{2\lambda t^M} \left\|u_\lambda^M - \widetilde{u}_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} e^{2\lambda t^{m+1}} \left\|u_\lambda^{m+1} - \widetilde{u}_h^{m+1}\right\|_{V^s}^2 \\
&\le 2e^{2\lambda T} \Bigg(\left\|u_\lambda^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|u_\lambda^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \\
&\qquad + \left\|\widetilde{u}_h^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|\widetilde{u}_h^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \Bigg),
\end{aligned} \tag{3.399}$$

wherein u_λ solves scheme (3.308), $u_{\lambda,h}^m$ its fully discretized counterpart in Scheme 3.81 and the $\widetilde{u}_h^m$ solve the degenerate Scheme 3.82.

Inequality (3.399) contains two separate groups of quantities that need to converge.

i) First, for $m \in \{0, \dots, M\}$, we recognize the weak solution u_λ^m of problem (3.308) and its approximation in a finite dimensional subspace, $u_{\lambda,h}^m$, that satisfies Theta Scheme 3.81.

ii) Second, for $m \in \{0, \dots, M\}$, we identify the differences between $\widetilde{u}_h^m$ and $u_{\lambda,h}^m$ that define the auxiliary terms w^m introduced in (3.313) of Lemma 3.83 or Theta Scheme 3.84, respectively.

We know from the previous Section 3.6.2.2 and Theorem 3.76 therein, that the norm of the difference between the quantities of group i) converges to zero, indeed. For convergence of the right hand side of inequality (3.399) it thus remains to show, that also the norm of the difference between the quantities of group ii) converges to zero, equivalently. The convergence of the normed w^m, $m \in \{0, \dots, M\}$, thus lies in the focus of this proof.

By Corollary 3.86 we have

$$\left\|\widetilde{u}_h^M - u_{\lambda,h}^M\right\|_H^2 + C_1\beta\Delta t \sum_{m=0}^{M-1} \left\|\widetilde{u}_h^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \le \Delta t C_2 \sum_{m=0}^{M-1} \left\|r_w^m\right\|_{V_h^{s*}}^2 \tag{3.400}$$

from which

$$\begin{aligned}
\min\{1, C_1\beta\} \Big(\left\|\widetilde{u}_h^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|\widetilde{u}_h^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \Big) \\
\le \Delta t C_2 \sum_{m=0}^{M-1} \left\|r_w^m\right\|_{V_h^{s*}}^2
\end{aligned} \tag{3.401}$$

immediately follows. Applying Lemma 3.89 to inequality (3.401) gives

$$\left\|\widetilde{u}_h^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|\widetilde{u}_h^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \leq \Delta t \frac{C_2 C_5}{\min\{1, C_1\beta\}} \sum_{m=0}^{M-1} \|U^m\|_H^2, \tag{3.402}$$

with C_5 and U^m as defined in Lemma 3.89. We define

$$\widetilde{C}_1 = \frac{C_2 C_5}{\min\{1, C_1\beta\}} \tag{3.403}$$

for later use. Lemma 3.93 provides bounds for $\|U^m\|_H^2$, $m \in \{0, \dots, M-1\}$, and develops the sum over the normed U^m into

$$\begin{aligned}\sum_{m=0}^{M-1} \|U^m\|_H^2 \leq 4 \sum_{m=0}^{M-1} \Bigg[\lambda^2 \bigg(\left\|u_{\lambda,h}^{m+1} - u_\lambda^{m+1}\right\|_H^2 + \left\|u_{\lambda,h}^m - u_\lambda^m\right\|_H^2 \\ + \left\|u_\lambda^{m+1} - u_\lambda^m\right\|_H^2\bigg) + \left(\lambda - \overline{\lambda}(\Delta t)\right)^2 \|u_\lambda^m\|_H^2\Bigg] \\ \leq 4\lambda^2 \bigg(\left\|u_{\lambda,h}^0 - u_\lambda^0\right\|_H^2 + 2\sum_{m=0}^{M-1} \left\|u_{\lambda,h}^{m+1} - u_\lambda^{m+1}\right\|_H^2 \\ + \sum_{m=0}^{M-1} \left\|u_\lambda^{m+1} - u_\lambda^m\right\|_H^2 + (\Delta t)^2 \frac{\lambda^2}{4} \sum_{m=0}^{M-1} \|u_\lambda^m\|_H^2\bigg),\end{aligned} \tag{3.404}$$

where we used in the last step that

$$\frac{\lambda - \overline{\lambda}(\Delta t)}{\Delta t} \leq \frac{\lambda^2}{2}$$

for all $\Delta t > 0$ as shown in Lemma 3.85. Considering the last two sums in (3.404), we find for the first one using the Bochner integrability of $\dot{u}_\lambda$ resulting from $u \in C^2([0, T,], H)$ and Lemma 2.38 with Proposition 1.2.3 of Arendt et al. (2011) and the Hölder inequality that

$$\begin{aligned}\sum_{m=0}^{M-1} \left\|u_\lambda^{m+1} - u_\lambda^m\right\|_H^2 &= \sum_{m=0}^{M-1} \left\|\int_{t^m}^{t^{m+1}} \dot{u}_\lambda(s)\,\mathrm{d}s\right\|_H^2 \\ &\leq \sum_{m=0}^{M-1} \left(\int_{t^m}^{t^{m+1}} \|\dot{u}_\lambda(s)\|_H\,\mathrm{d}s\right)^2 \\ &\leq \sum_{m=0}^{M-1} \left((\Delta t)^{\frac{1}{2}} \left[\int_{t^m}^{t^{m+1}} \|\dot{u}_\lambda(s)\|_H^2\,\mathrm{d}s\right]^{\frac{1}{2}}\right)^2 \\ &= \Delta t \sum_{m=0}^{M-1} \int_{t^m}^{t^{m+1}} \|\dot{u}_\lambda(s)\|_H^2\,\mathrm{d}s \\ &= \Delta t \int_0^T \|\dot{u}_\lambda(s)\|_H^2\,\mathrm{d}s,\end{aligned} \tag{3.405}$$

wherein we use Theorem 24.7 of Wloka (2002) in the second step. For the second sum we get

$$\Delta t \sum_{m=0}^{M-1} \|u_\lambda^m\|_H^2 \leq T \max_{0\leq\tau\leq T} \|u_\lambda(\tau)\|_H^2. \tag{3.406}$$

Now we take (3.405) and (3.406) and insert them back into (3.404) to get

$$\begin{aligned}\sum_{m=0}^{M-1} \|U^m\|_H^2 \leq 4\lambda^2\Bigg(&\left\|u_{\lambda,h}^0 - u_\lambda^0\right\|_H^2 + 2\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1} - u_\lambda^{m+1}\right\|_H^2 \\ &+ \Delta t\int_0^T \|\dot{u}_\lambda(s)\|_H^2 \,\mathrm{d}s + \Delta t\frac{\lambda^2}{4}T\max_{0\leq\tau\leq T}\|u_\lambda(\tau)\|_H^2\Bigg).\end{aligned} \tag{3.407}$$

Now combining (3.402) and (3.407) and inserting the result in (3.399) gives

$$\begin{aligned}&\left\|u^M - u_h^M\right\|_H^2 + \Delta t\sum_{m=0}^{M-1}\left\|u^{m+1} - u_h^{m+1}\right\|_{V^s}^2 \\ &\leq 2e^{2\lambda T}\Bigg(\left\|u_\lambda^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t\sum_{m=0}^{M-1}\left\|u_\lambda^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 + \Delta t\widetilde{C}_1\sum_{m=0}^{M-1}\|U^m\|_H^2\Bigg) \\ &\leq 2e^{2\lambda T}\Bigg(\left\|u_\lambda^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t\sum_{m=0}^{M-1}\left\|u_\lambda^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \\ &\qquad + \Delta t 4\widetilde{C}_1\lambda^2\Bigg(\left\|u_{\lambda,h}^0 - u_\lambda^0\right\|_H^2 + 2\sum_{m=0}^{M-1}\left\|u_{\lambda,h}^{m+1} - u_\lambda^{m+1}\right\|_H^2 \\ &\qquad\qquad + \Delta t\int_0^T\|\dot{u}_\lambda(s)\|_H^2\,\mathrm{d}s + \Delta t\frac{\lambda^2}{4}T\max_{0\leq\tau\leq T}\|u_\lambda(\tau)\|_H^2\Bigg)\Bigg) \\ &\leq 2e^{2\lambda T}\Bigg(\left\|u_\lambda^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t(1+8\widetilde{C}_1\lambda^2)\sum_{m=0}^{M-1}\left\|u_\lambda^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \\ &+ \Delta t 4\widetilde{C}_1\lambda^2\Bigg(\left\|u_\lambda^0 - u_{\lambda,h}^0\right\|_H^2 + \Delta t\int_0^T\|\dot{u}_\lambda(s)\|_H^2\,\mathrm{d}s + \Delta t\frac{\lambda^2}{4}T\max_{0\leq\tau\leq T}\|u_\lambda(\tau)\|_H^2\Bigg)\Bigg)\end{aligned} \tag{3.408}$$

with $\widetilde{C}_1$ as defined in (3.403). The upper estimate in (3.408) now depends only on terms $u_{\lambda,h}^m$ and u_λ^m. These solve the related "coercified" scheme. While the u_λ^m solve problem (3.308), their discrete counterparts $u_{\lambda,h}^m$ solve Scheme 3.81. We can thus apply the convergence results that we have derived in Section 3.6.2.2, earlier.

Considering the term $\left\|u_{\lambda,h}^0 - u_\lambda^0\right\|_H^2$ in (3.408) we recognize

$$\begin{aligned}u_\lambda^0 &= u_\lambda(0) = g, \\ u_{\lambda,h}^0 &= u_{\lambda,h}(0) = g_h,\end{aligned} \tag{3.409}$$

the initial conditions. We use the quasi-optimality of the initial condition of Assumption 3.D in (3.409) and then the approximation property of Assumption 3.A yielding

$$\begin{aligned}\left\|u_\lambda^0 - u_{\lambda,h}^0\right\|_H &= \|g - g_h\|_H \leq C_I \inf_{v_h \in V_h^s} \|g - g_h\|_H \\ &\leq C_I \inf_{v_h \in V_h^s} \|g - g_h\|_{V^s} \leq \widetilde{C}_2 \Upsilon(h, t, \alpha_{\mathcal{A}}/2, u_\lambda(0)),\end{aligned} \tag{3.410}$$

where $\widetilde{C}_2 = C_\Upsilon C_I > 0$ with C_I being the constant from Assumption 3.D and C_Υ the constant from Assumption 3.A. For the remaining normed residual terms in (3.408) we have trivially

$$\begin{aligned}\left\|u_\lambda^M - u_{\lambda,h}^M\right\|_H^2 &+ \Delta t(1 + 8\widetilde{C}_1\lambda^2) \sum_{m=0}^{M-1} \left\|u_\lambda^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \\ &\leq \widetilde{C}_3 \left(\left\|u_\lambda^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|u_\lambda^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2 \right),\end{aligned} \tag{3.411}$$

with

$$\widetilde{C}_3 = 1 + 8\widetilde{C}_1\lambda^2. \tag{3.412}$$

Now we assemble our findings. Applying Corollary 3.77 for $\theta = 1$ to (3.411) and inserting the result together with (3.410) into (3.408) gives

$$
\begin{aligned}
&\left\|u^M - u_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|u^{m+1} - u_h^{m+1}\right\|_{V^s}^2 \\
&\leq 2e^{2\lambda T}\biggl(\widetilde{C}_3\left(\left\|u_\lambda^M - u_{\lambda,h}^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|u_\lambda^{m+1} - u_{\lambda,h}^{m+1}\right\|_{V^s}^2\right) \\
&+\Delta t 4\widetilde{C}_1\lambda^2\biggl(\left\|u_\lambda^0 - u_{\lambda,h}^0\right\|_H^2 + \Delta t \int_0^T \|\dot{u}_\lambda(s)\|_H^2 \,\mathrm{d}s + \Delta t \frac{\lambda^2}{4} T \max_{0\leq\tau\leq T} \|u_\lambda(\tau)\|_H^2\biggr)\biggr) \\
&\leq 2e^{2\lambda T}\biggl(\widetilde{C}_3\biggl(\overline{C} \max_{0\leq t\leq T} \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u_\lambda(t)) \\
&\qquad\qquad + \overline{C}\,\Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u_\lambda(0)) \\
&\qquad\qquad + \overline{C}(\Delta t)^2 \int_0^T \|\ddot{u}_\lambda(s)\|_{V_h^{s*}}^2 \,\mathrm{d}s \\
&\qquad\qquad + \overline{C} \int_0^T \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,\dot{u}_\lambda(s)) \,\mathrm{d}s \\
&\qquad\qquad + \overline{C} \max_{0\leq t\leq T} \Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u_\lambda(t))\biggr) \\
&\qquad + \Delta t 4\widetilde{C}_1\lambda^2\biggl(\widetilde{C}_2^2\Upsilon^2(h,t,\alpha_{\mathcal{A}}/2,u_\lambda(0)) \\
&\qquad\qquad + \Delta t \int_0^T \|\dot{u}_\lambda(s)\|_H^2 \,\mathrm{d}s \\
&\qquad\qquad + \Delta t \frac{\lambda^2}{4} T \max_{0\leq\tau\leq T} \|u_\lambda(\tau)\|_H^2\biggr)\biggr)
\end{aligned}
\tag{3.413}
$$

Defining

$$
C_6 = 2e^{2\lambda T} \max\left\{\widetilde{C}_3\overline{C},\, 4\widetilde{C}_1\lambda^2 \max\left\{\widetilde{C}_2^2, 1, \frac{\lambda^2}{4}T\right\}\right\} \tag{3.414}
$$

gives

$$
\begin{aligned}
\left\|u^M - u_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} &\left\|u^{m+1} - u_h^{m+1}\right\|_{V^s}^2 \\
&\le\ C_6 \max_{0\le\tau\le T} \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u_\lambda(\tau)) \\
&\quad + C_6 \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u_\lambda(0)) \\
&\quad + C_6(\Delta t)^2 \int_0^T \|\ddot{u}_\lambda(s)\|_{V_h^{s*}}^2 \,\mathrm{d}s \\
&\quad + C_6 \int_0^T \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, \dot{u}_\lambda(s)) \,\mathrm{d}s \\
&\quad + C_6 \max_{0\le\tau\le T} \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u_\lambda(\tau)) \\
&\quad + C_6 \Delta t \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u_\lambda(0)) \\
&\quad + C_6(\Delta t)^2 \int_0^T \|\dot{u}_\lambda(s)\|_H^2 \,\mathrm{d}s \\
&\quad + C_6(\Delta t)^2 \max_{0\le\tau\le T} \|u_\lambda(\tau)\|_H^2.
\end{aligned}
\tag{3.415}
$$

Collecting terms gives

$$
\begin{aligned}
&\left\|u^M - u_h^M\right\|_H^2 + \Delta t \sum_{m=0}^{M-1} \left\|u^{m+1} - u_h^{m+1}\right\|_{V^s}^2 \\
&\le\ C_6(3 + \Delta t) \max_{0\le\tau\le T} \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, u_\lambda(\tau)) \\
&\quad + C_6(\Delta t)^2 \left(\int_0^T \|\ddot{u}_\lambda(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau + \int_0^T \|\dot{u}_\lambda(\tau)\|_H^2 \,\mathrm{d}\tau + \max_{0\le\tau\le T} \|u_\lambda(\tau)\|_H^2 \right) \\
&\quad + C_6 \int_0^T \Upsilon^2(h, t, \alpha_{\mathcal{A}}/2, \dot{u}_\lambda(\tau)) \,\mathrm{d}\tau.
\end{aligned}
\tag{3.416}
$$

which proves the claim. □

Corollary 3.95 (Convergence with Υ of von Petersdorff and Schwab (2003)) *Let the assumptions of Theorem 3.94 be satisfied and assume further the setting of von Petersdorff and Schwab (2003) as outlined in Example 3.58. Then there exists a constant*

$\overline{C} > 0$ *such that the convergence estimate*

$$\begin{aligned}\left\|u^M - u_h^M\right\|_H^2 &+ \Delta t \sum_{m=0}^{M-1} \left\|u^{m+1} - u_h^{m+1}\right\|_{V^{\alpha_{\mathcal{A}}/2}}^2 \\ &\leq \overline{C}\, h^{2(p+1-\alpha_{\mathcal{A}}/2)} \Bigg((3+\Delta t) \max_{0\leq\tau\leq T} \|u(\tau)\|_{\mathcal{H}^{p+1}(\Omega)}^2 \qquad (3.417) \\ &\qquad + \int_0^T \|\dot{u}(\tau)\|_{\mathcal{H}^{p+1}(\Omega)}^2 + \|u(\tau)\|_{\mathcal{H}^{p+1}(\Omega)}^2 \,\mathrm{d}\tau \Bigg) \\ &\quad + \overline{C}(\Delta t)^2 \Bigg(\int_0^T \|\ddot{u}(\tau)\|_{V_h^{s*}}^2 + \|\dot{u}(\tau)\|_{V_h^{s*}}^2 + \|u(\tau)\|_{V_h^{s*}}^2 \,\mathrm{d}\tau \\ &\qquad + \int_0^T \|\dot{u}(\tau)\|_H^2 + \|u(\tau)\|_H^2 \,\mathrm{d}\tau + \max_{0\leq\tau\leq T} \|u(\tau)\|_H^2 \Bigg)\end{aligned}$$

holds.

Proof

The result follows from Theorem 3.94 with

$$\Upsilon(h,t,s,u) = h^{t-s}\|u\|_{\mathcal{H}^t(\Omega)},$$

as outlined by Example 3.58 taking $t \leq p+1$ equal to its maximal admissible value with p the polynomial degree that the basis functions of $V_h^{\alpha_{\mathcal{A}}/2}$ achieve piecewisely. Additionally, one uses the relations

$$u_\lambda = e^{-\lambda\cdot}u, \qquad \dot{u}_\lambda = e^{-\lambda\cdot}(\dot{u} - \lambda u), \qquad \ddot{u}_\lambda = e^{-\lambda\cdot}(\ddot{u} - 2\lambda \cdot u + \lambda^2 u)$$

to derive the result. □

Theorem 3.94 that served as the foundation for the above Corollary 3.95 has generalized the claim of Theorem 5.4 of von Petersdorff and Schwab (2003) in several ways. First, we allow for a time-inhomogeneous PIDE and consequentially a time-dependent bilinear form. Second, the convergence result now applies to bilinear forms of Gårding type instead of being restricted to the special case of coercive bilinear forms. Additionally, we framed the approximation property of Assumption 3.A in very general terms.
Within this generalized theoretical framework Corollary 3.95 expresses the fact that the order of convergence derived in the initial result of von Petersdorff and Schwab (2003) has remained unchanged. We thus achieve the same order of convergence in the generalized framework that we observed in the special case, before.

We close by noting that the rate of convergence critically depends on the regularity of the initial condition. Some numerical schemes have been designed with that requirement in mind. In this regard we refer to the literature on *hp* discontinuous Galerkin schemes, for example Schötzau and Schwab (2001) wherein a time stepping scheme is proposed that manages to resolve non-smooth initial data at exponential rates of convergence. Confer also more generally Schötzau (1999) in this regard.

4 Chebyshev polynomial interpolation

The numerical PIDE solver that we implemented in the previous chapter enables us to derive prices for European options in Lévy models. The symbol method equips this tools with a rich flexibility regarding the model choice and the numerical results at the end of the previous chapter verify the numerical feasibility of its implementation. At the same time, a theoretical framework for error control proved convergence and stability of the approach under even more general theoretical assumptions. With a working numerical environment for option pricing at our disposal, we now focus on improving computational runtimes of option pricing routines.

The complexity of today's model universe reflects in the sophistication of numerical implementations. Stochastic volatilities, time dependent Lévy jump models or pricing in higher dimensions place challenging demands to numerics. While complexity in the models is justified by the desire to reproduce the observed complex behavior of financial markets, the industry insists on feasible runtimes that match practical needs. Buyers and sellers of financial products alike expect quotations that reflect the market situation while still being issued live. Risk managers rely on a model capable of capturing all relevant sources of risk but depend on risk assessment before that risk materializes. Similarly, a bank aims at maintaining a rich model environment but needs to be able to recalibrate it steadily to markets that keep on moving constantly.

Industry thus faces a seeming contradiction. While numerical complexity continuously grows, fast runtimes are expected to be maintained. In this chapter, we introduce a method that aims at resolving this contradiction. It is based on an interpolation technique that has been known in other contexts before but has not been applied in finance, yet. This is truly surprising as the theory behind the method connects to finance smoothly and its numerical implementation yields highly appealing results. The method is called Chebyshev polynomial interpolation. It is an interpolation technique that uses prices at prespecified points in the parameter space to interpolate prices for parameters inbetween with Chebyshev polynomials.

The first section of this chapter introduces the method mathematically. Building on its original one-dimensional form, a multivariate extension by tensorization is presented and investigated. In Section 4.2, the so called online/offline decomposition being the key element of the algorithm and responsible for its fast runtimes is explored more thoroughly. The subsequent Section 4.3 derives conditions that prices interpreted as functions of the model and option parameters need to fulfill for exponential convergence of the algorithm and verifies these conditions for several Lévy models and option types. Section 4.4

describes an implementation of the Chebyshev algorithm and presents empirical accuracy and convergence studies for several models and options. The numerical results indicate that the approximative power of the algorithm even exceeds the scope of applicability that the theoretical findings suggest.

The results of this chapter are taken from the article Gaß et al. (2016) where they have been jointly developed, first. The proofs of those theoretical results that were developed by coauthors will only be referenced in this thesis. We refer the interested reader to the article in these cases. Explanatory descriptions of the method and the accompanying results have been rewritten in parts but clearly cannot deny their close relation to the paper source. The numerical experiments have been repeated based on a different parametrization, thereby validating the theoretical results from a different perspective, again.

4.1 An algorithmic introduction of the method

In introducing Chebyshev interpolation we distinguish between the univariate case and its multivariate extension. In both cases we present the method in an algorithmic fashion already adapted to option pricing.

4.1.1 The univariate interpolation method

We present the Chebyshev interpolation method. We begin by introducing the method in its one-dimensional form that was originally outlined in Trefethen (2013). In a second step, we extend the uni-variate framework to the multivariate case and present results of error convergence analysis both theoretically as well as empirically.

Assume a one-dimensional parameter space $\mathcal{P}$ given by $\mathcal{P} = [-1, 1]$ and an option price depending on a single varying parameter taking values in that space,

$$\mathit{Price}^{p}, \qquad p \in \mathcal{P}. \tag{4.1}$$

We define the Chebyshev interpolator I_N which will be the driving quantity in the interpolation of Price^p based on Chebyshev polynomials of degree N. It is given by

$$I_N(\mathit{Price}^{(\cdot)})(p) = \sum_{j=0}^{N} c_j T_j(p), \tag{4.2}$$

wherein the coefficients c_j, $j \in \{0, \ldots, N\}$, are defined by

$$c_j = \frac{2^{\mathbb{1}_{0<j<N}}}{N} \sum_{l=0}^{N}{}'' \mathit{Price}^{p_l} \cos\Big(j\pi \frac{l}{N}\Big), \qquad j \in \{0, \ldots, N\}, \tag{4.3}$$

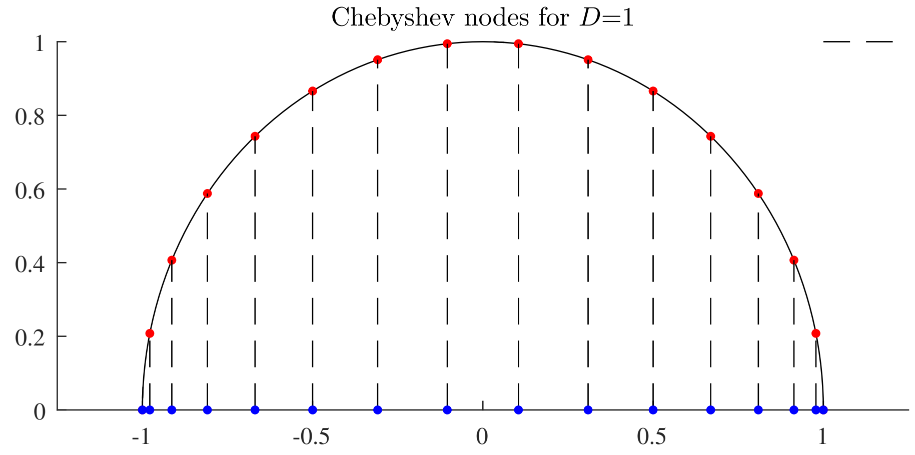

Figure 4.1 A set of Chebyshev points $p_l \in [-1,1]$ (blue) for degree $N = 15$ constructed by equidistantly spaced auxiliary construction points (red) on the semi-circle. The MATLAB code for this construction of the Chebyshev nodes is taken from the book Trefethen (2013).

where the notation of the doubly primed sum indicates that its first and last summand are multiplied by 1/2. The basis functions T_j in (4.2) are given by

$$T_j(p) = \cos\big(j \arccos(p)\big), \qquad j \in \{0, \ldots, N\}, \text{ defined on } [-1,1]. \tag{4.4}$$

The Chebyshev polynomial interpolation method inherits its name from the fact that the basis functions T_j in (4.4) allow for a polynomial representation, as well. Chebyshev nodes

$$p_l = \cos\left(\pi \frac{l}{N}\right), \qquad l \in \{0, \ldots, N\}, \tag{4.5}$$

mark locations in the parameter space where the method interpolates perfectly. Their construction admits a beautiful geometric interpretation as illustrated by Figure 4.1. There, $N = 15$ Chebyshev nodes in one dimension are depicted and their geometric construction is emphasized. Self-evidently, univariate Chebyshev interpolation is not restricted to the generic parameter space $[-1,1]$. Instead, invoking an appropriate linear transformation opens the method to any parameter space that can be cast in the form of a real parameter interval $[\underline{p}, \overline{p}]$. The interpolation operator (4.2) is then easily adjusted accordingly.

4.1.2 A multivariate extension

The scope of Chebyshev polynomial interpolation is not limited to univariate applications. Instead, a tensor based extension captures the multivariate case. Assume a parameter space

$$\mathcal{P} = \mathcal{K} \times \mathcal{T} \times \mathcal{Q} = [-1,1]^{D_1} \times [-1,1]^{D_2} \times [-1,1]^{D_3} = [-1,1]^D \tag{4.6}$$

for the interpolation of prices

$$\mathit{Price}^{(K,T,q)}, \qquad (K,T,q) \in \mathcal{P}, \tag{4.7}$$

where $D_1 = \dim(\mathcal{K}) \in \mathbb{N}$, $D_2 = \dim(\mathcal{T}) \in \mathbb{N}$, $D_3 = \dim(\mathcal{Q}) \in \mathbb{N}$ and $D = D_1 + D_2 + D_3$. As above, more general hyperrectangular parameter spaces are not excluded. Parameter spaces given by

$$\begin{aligned}\mathcal{P} = {} & [\underline{K}_1, \overline{K}_1] \times \cdots \times [\underline{K}_{D_1}, \overline{K}_{D_1}] \\ & \times [\underline{T}_1, \overline{T}_1] \times \cdots \times [\underline{T}_{D_2}, \overline{T}_{D_2}] \\ & \times [\underline{Q}_1, \overline{Q}_1] \times \cdots \times [\underline{Q}_{D_3}, \overline{Q}_{D_3}],\end{aligned} \tag{4.8}$$

become admissible by appropriate linear transformations. With $\overline{N} = (N_1, \ldots, N_D)$ and $N_i \in \mathbb{N}_0$ for $i \in \{1, \ldots, D\}$, the univariate interpolator as defined in (4.2) then extends rather naturally to the multivariate case. With $\prod_{i=1}^{D}(N_i + 1)$ summands it is given by

$$I_{\overline{N}}(\mathit{Price}^{(\cdot)})(p) = \sum_{j \in J} c_j T_j(p), \qquad p \in \mathcal{P}, \tag{4.9}$$

where the summation index j is a multiindex with values in

$$J = \{(j_1, \ldots, j_D) \in \mathbb{N}_0^D, \text{ where } j_i \in \{0, \ldots, N_i\} \text{ for } i \in \{1, \ldots, D\}\}. \tag{4.10}$$

Thus, being fully explicit, equation (4.9) indeed turns into

$$I_{\overline{N}}(\mathit{Price}^{(\cdot)})(p) = \sum_{j_1=0}^{N_1} \cdots \sum_{j_D=0}^{N_D} c_{(j_1,\ldots,j_D)} T_{(j_1,\ldots,j_D)}(p), \qquad p \in \mathcal{P}. \tag{4.11}$$

In the multivariate case, the basis functions T_j for $j = (j_1, \ldots, j_D) \in J$ are defined by

$$T_j(p_1, \ldots, p_D) = \prod_{i=1}^{D} T_{j_i}(p_i), \qquad p \in \mathcal{P}, \tag{4.12}$$

and the associated coefficients c_j with $j = (j_1, \ldots, j_D) \in J$ are given by

$$c_j = \Big(\prod_{i=1}^{D} \frac{2^{\mathbb{1}_{\{0<j_i<N_i\}}}}{N_i}\Big) \sum_{l_1=0}^{N_1}{}'' \cdots \sum_{l_D=0}^{N_D}{}'' \mathit{Price}^{p^{(l_1,\ldots,l_D)}} \prod_{i=1}^{D} \cos\left(j_i \pi \frac{l_i}{N_i}\right). \tag{4.13}$$

Similarly, the Chebyshev nodes p^l are now defined for a multiindex $l = (l_1, \ldots, l_D) \in J$ and distributed accordingly,

$$p^l = (p_{l_1}, \ldots, p_{l_D}), \tag{4.14}$$

inheriting their actual values from their univariate counterparts,

$$p_{l_i} = \cos\left(\pi \frac{l_i}{N_i}\right), \qquad \text{for } l_i \in \{0, \ldots, N_i\} \text{ and } i \in \{1, \ldots, D\}. \tag{4.15}$$

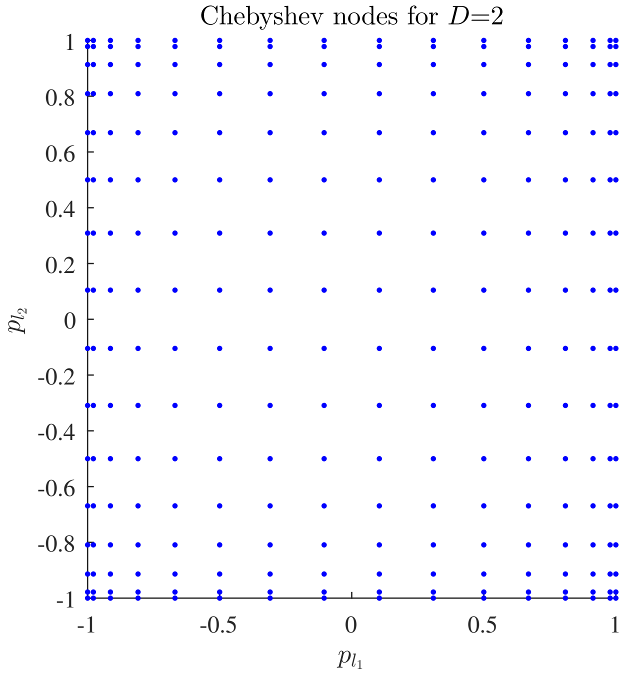

Figure 4.2 A set of D-variate Chebyshev points $p^l \in [-1,1]^D$ for $D = 2$ and $N_1 = N_2 = 15$. Their arrangement adheres to the rule (4.15).

Figure 4.2 displays a set of D-variate Chebyshev nodes $p^{(l_1,\dots,l_D)}$ for $D = 2$ and $N_1 = N_2 = 15$.

In the univariate case, convergence of interpolation by the Chebyshev method is well known. We cite the according remark on convergence of the Chebyshev method from Gaß et al. (2016).

Remark 4.1 (Convergence of multivariate Chebyshev interpolation) *It is well known that the error of approximation with Chebyshev polynomials decays polynomially for differentiable functions and exponentially for analytic functions. More precisely, going back to Mastroianni and Szabados (1995), it is shown by Theorem 7.2 in Trefethen (2013) that the error $\|f - I_N(f)\|_{L^\infty([-1,1])}$ decays at rate $O(N^\nu)$ if $x \mapsto f(x)$ is ν times differentiable with νth derivative of finite variation and $f^{(1)}, f^{(2)}, \dots, f^{(\nu-1)}$ are absolutely continuous. Let additionally f be analytic in $[-1,1]$ then it is analytic in some Bernstein ellipse $B([-1,1],\varrho)$ with parameter $\varrho > 1$, as defined in Definition 2.43. Theorem 8.2 in Trefethen (2013), that traces back to the seminal work of Bernstein (1912), shows that if f has an analytic extension to some Bernstein ellipse $B([-1,1],\varrho)$ with parameter $\varrho > 1$ then the error decay is of exponential rate $O(\varrho^{-N})$.*

The rest of this chapter extends the result presented in Remark 4.1 for the univariate case to its multivariate extension having the application of parametric option pricing in

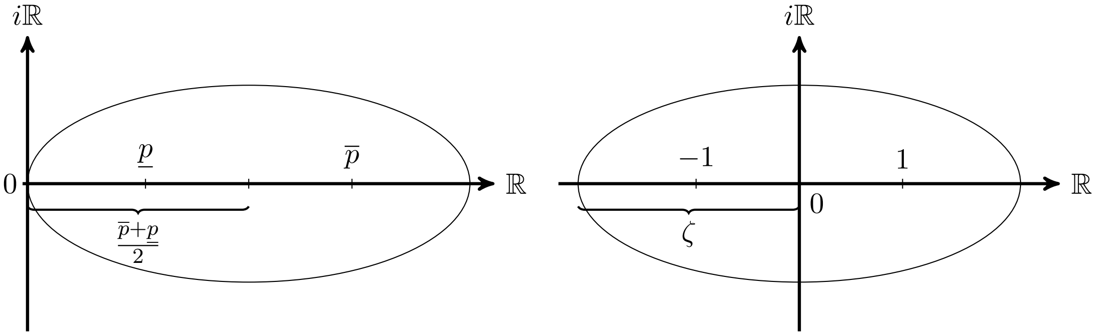

Figure 4.3 Left: Generalized Bernstein ellipse E_1 with foci $\underline{p}$, $\overline{p}$ and semimajor $\frac{\overline{p}+\underline{p}}{2}$. Right: Bernstein ellipse E_2 with foci ± 1 and semimajor ζ.

mind. We consider these parametric option prices to be given by

$$\mathit{Price}^{(K,T,q)}, \qquad \text{for } (K,T,q) \in \mathcal{P} = \mathcal{K} \times \mathcal{T} \times \mathcal{Q}, \tag{4.16}$$

wherein $\mathcal{K} \subset \mathbb{R}^{D_1}$, $\mathcal{T} \subset \mathbb{R}^{D_2}$, $\mathcal{Q} \subset \mathbb{R}^{D_3}$ and therefore $\mathcal{P} \subset \mathbb{R}^{D}$ with dimensionality $D = D_1 + D_2 + D_3$. The underlying parameter space $\mathcal{P}$ is assumed to be of hyperrectangular structure, in the sense that

$$\mathcal{P} = [\underline{p}_1, \overline{p}_1] \times \ldots \times [\underline{p}_D, \overline{p}_D] \tag{4.17}$$

with real $\underline{p}_i \leq \overline{p}_i$ for all $i \in \{1, \ldots, D\}$.

As Remark 4.1 indicates for the univariate case, exponential convergence of interpolation by the Chebyshev method relies on regularity assumptions to be met by the interpolated function. More precisely, for exponential convergence in the univariate case the interpolated function is required to be analytic on a Bernstein ellipse $B([-1,1], \varrho)$ with a certain ellipse parameter $\varrho > 1$. This ellipse parameter directly determines the rate of convergence. Before we can generalize the univarate convergence results to the multivariate case, the concept of a Bernstein ellipse must be extended accordingly. To this end we define the D-variate and transformed analogon of a Bernstein ellipse around the hyperrectangle $\mathcal{P}$ with parameter vector $\varrho \in (1, \infty)^D$ as

$$B(\mathcal{P}, \varrho) = B([\underline{p}_1, \overline{p}_1], \varrho_1) \times \ldots \times B([\underline{p}_D, \overline{p}_D], \varrho_D) \tag{4.18}$$

based on D generalized univariate Bernstein ellipses

$$B([\underline{p}, \overline{p}], \varrho) = \tau_{[\underline{p}, \overline{p}]} \circ B([-1,1], \varrho), \tag{4.19}$$

as given by Definition 2.43 in the preliminary chapter. Analogously to the univariate case, the ellipse parameter vector $\varrho \in (1, \infty)^D$ will determine the rate of convergence. Its value corresponds with the extension of the parameter space $\mathcal{P}$ and is determined by the following remark.

Remark 4.2 (How to derive ϱ)
We comment on the derivation of the ellipse parameter $\varrho > 1$ of the generalized Bernstein ellipse. Assume a Chebyshev approximation of a function in the parameter $p \in [\underline{p}, \overline{p}]$. We denote the generalized Bernstein ellipse with $\underline{p}$, $\overline{p}$ as foci and with the origin at its boundary by E_1,

$$E_1 = B([\underline{p}, \overline{p}], \varrho). \tag{4.20}$$

Furthermore, we denote by E_2 the Bernstein ellipse that originates from linearly mapping the foci of E_1 to the Bernstein ellipse with foci ± 1 using the transformation $\tau_{[\underline{p},\overline{p}]}$ from identity (4.19), that is

$$E_2 = B([-1, 1], \varrho) = \tau^{-1}_{[\underline{p},\overline{p}]} \circ E_1. \tag{4.21}$$

A schematic illustration of E_1 and E_2 is provided by Figure 4.3. In our notation, the ellipse parameter of the generalized Bernstein ellipse E_1 is defined as the ellipse parameter ρ of the Bernstein ellipse E_2. We determine the ellipse parameter $\varrho > 1$ of E_2 using the mapping $\tau_{[\underline{p},\overline{p}]}$, or rather its inverse. To this extent recall that for a Bernstein ellipse with semimajor a_ϱ and semiminor b_ϱ the relations

$$a_\varrho = \frac{\varrho + \frac{1}{\varrho}}{2}, \qquad b_\varrho = \frac{\varrho - \frac{1}{\varrho}}{2}, \qquad \varrho = a_\varrho + b_\varrho > 1 \tag{4.22}$$

hold. Evidently, E_1 has a semimajor value of $\frac{\overline{p}+\underline{p}}{2}$ and thus E_2 has a semimajor value of

$$\zeta = a^{E_2}_\varrho = \tau^{-1}_{[\underline{p},\overline{p}]}\left(\frac{\overline{p} + \underline{p}}{2}\right) - \tau^{-1}_{[\underline{p},\overline{p}]}(0) = \frac{\overline{p} + \underline{p}}{\overline{p} - \underline{p}}. \tag{4.23}$$

Using the relations (4.22), we derive

$$\varrho = \zeta + \sqrt{\zeta^2 - 1}. \tag{4.24}$$

The value in (4.24) determines the ellipse parameter of E_2 and will provide the exponential decay rate in our theoretical results, later.

We are thus prepared to cite the core theorem granting exponential error decay of the Chebyshev interpolation in the multivariate case.

Theorem 4.3 (Asymptotic error decay with tensorized Chebyshev interpolation)
Let $\mathcal{P} \ni p \mapsto Price^p$ be a real valued function that has an analytic extension to some generalized Bernstein ellipse $B(\mathcal{P}, \varrho)$ for some parameter vector $\varrho \in (1, \infty)^D$ and assume $\max_{p \in B(\mathcal{P},\varrho)} |Price^p| \leq V$. Then

$$\max_{p \in \mathcal{P}} \left| Price^p - I_{\overline{N}}(Price^{(\cdot)})(p) \right| \leq 2^{\frac{D}{2}+1} \cdot V \cdot \left(\sum_{i=1}^{D} \varrho_i^{-2N_i} \prod_{j=1}^{D} \frac{1}{1 - \varrho_j^{-2}} \right)^{\frac{1}{2}}. \tag{4.25}$$

The proof of the theorem is provided in Gaß et al. (2016). As an immediate consequence of Theorem 4.3 we obtain the following corollary.

Corollary 4.4 (Asymptotic error decay with tensorized Chebyshev interpolation)
Under the assumptions of Theorem 4.3 there exists a constant $C > 0$ *such that*

$$\max_{p \in \mathcal{P}} \left| Price^p - I_{\overline{N}}(Price^{(\cdot)})(p) \right| \leq C \underline{\varrho}^{-\underline{N}}, \tag{4.26}$$

where $\underline{\varrho} = \min_{1 \leq i \leq D} \varrho_i$ *and* $\underline{N} = \min_{1 \leq i \leq D} N_i$.

Citing the following remark from Gaß et al. (2016) we obtain exponential error convergence when the same number of Chebyshev nodes N is chosen in each dimension of the parameter space.

Remark 4.5 (Exponential error decay in N)
In particular, for the same number of nodes N *in each dimension of the parameter space Corollary 4.4 shows that the error decay is of exponential order* $O\big(\varrho^{-N}\big) = O\big(\varrho^{-\sqrt[D]{M}}\big)$ *for some* $\varrho > 1$ *with* M *denoting the number of degrees of freedom of the interpolation method and* D *the dimension of the parameter space under the assumptions of Theorem 4.3.*

4.2 The online/offline decomposition feature

In the introduction of the chapter we highlighted our goal of accelerating numerical runtimes while at the same time maintaining flexibility regarding the choice of the model and its complexity. Let us emphasize how the Chebyshev interpolation approach achieves this goal. To this extent recall the interpolation operator in D dimensions as presented by (4.11) as

$$Price^p \approx I_{\overline{N}}(Price^{(\cdot)})(p) = \sum_{j_1=0}^{N_1} \cdots \sum_{j_D=0}^{N_D} \underbrace{c_{(j_1,\ldots,j_D)}}_{\text{i) offline phase}} \underbrace{T_{(j_1,\ldots,j_D)}(p)}_{\text{ii) online phase}}, \tag{4.27}$$

for a parameter $p \in \mathcal{P}$ from the parameter space. The computation of $Price^p$ based on an arbitrarily complex model and possibly suffering from a lengthy numerical derivation has thus turned into the evaluation of a finite sum with known coefficients. This development is by no means trivial since it allowed for a separation of the complex model and the attached model pricing routine from the actual parameter $p \in \mathcal{P}$ that the price is evaluated or rather approximated for. The overall pricing procedure has thus split into two separate stages, which are called *offline phase* and *online phase*. Both labels have their origin in the more general theory of model reduction techniques, yet their meaning applies to the Chebyshev method equally.

i) **Offline phase**
In the first phase, the algorithm is set up and prepared for pricing or related applications. The model is chosen and model prices are computed for all Chebyshev points in the parameter space in order to determine the coefficients $c_{(j_1,\ldots,j_D)}$ using

a pricing method of choice. Depending on the complexity of the model and the runtime of the pricing method which can be Monte Carlo, Fourier pricing, PIDE techniques or other algorithms, this offline phase possibly consumes a considerable amount of time. Yet it is crucial for an understanding of the performance of the Chebyshev algorithm to keep in mind, that this offline phase is only conducted once.

ii) **Online phase**
Now that the algorithm is prepared, the pricing for model parameters of interest takes place. The pricing routine used for the derivation of prices at the Chebyshev points during the first phase has become unnecessary. Instead, pricing now consists in the evaluation of the Chebyshev polynomials $T_{(j_1,\dots,j_D)}$ at the parameter $p \in \mathcal{P}$ of interest and an assembling of the weighted sum in (4.27) with known coefficients $c_{(j_1,\dots,j_D)}$ that are independent of the parameter $p \in \mathcal{P}$.

The splitting of the original pricing routine into those two phases results in a tremendous increase in pricing runtime. With the computationally intense derivations being shifted into the offline phase, only numerically cheap evaluations of polynomials remain for online pricing. The questions remain whether the resulting approximate price is accurate and how far it converges to its true value. The next section discovers conditions under which exponential convergence is obtained before the numerical sections investigate both accuracy and convergence empirically.

4.3 Exponential convergence of Chebyshev interpolation for parametric option pricing

In this section we embed the multivariate Chebyshev interpolation into the option pricing framework. First, as in Gaß et al. (2016), we provide sufficient conditions under which option prices analytically depend on the parameters. Second, these are verified for payoff profiles and asset models individually. As an example, we investigate the interpolation of call option prices in Lévy models in more detail.

Analytic properties of option prices can be conveniently studied in terms of Fourier transforms. First, Fourier representations of option prices are explicitly available for a large class of both option types and asset models. Second, Fourier transformation unveils the analytic properties of both the payoff structure and the distribution of the underlying stochastic quantity in a beautiful way. By contrast, if option prices are represented as expectations, their analyticity in the parameters is hidden. For example the function $K \mapsto (S_T - K)^+$ is not even differentiable, whereas the Fourier transform of the dampened call payoff function evidently is analytic in the strike, compare Table 4.1.

4.3.1 Conditions for exponential convergence

In Gaß et al. (2016), we introduce a general option pricing framework. We consider option prices of the form

$$Price^{p=(K,T,q)} = \mathbb{E}\big[f_K(X_T^q)\big], \qquad p \in \mathcal{P}, \tag{4.28}$$

where f_K is a parametrized family of measurable payoff functions $f_K : \mathbb{R}^d \to \mathbb{R}_+$ with payoff parameters $K \in \mathcal{K}$ and X_T^q is a family of $\mathbb{R}^d$-valued random variables with model parameters $(T, q) \in \mathcal{T} \times \mathcal{Q}$. The parameter set

$$p = (K, T, q) \in \mathcal{P} = \mathcal{K} \times \mathcal{T} \times \mathcal{Q} \subset \mathbb{R}^D \tag{4.29}$$

is again of hyperrectangular structure, that is

$$\begin{aligned} \mathcal{K} &= [\underline{p}_1, \overline{p}_1] \times \ldots \times [\underline{p}_{D_1}, \overline{p}_{D_1}] \\ \mathcal{T} \times \mathcal{Q} &= [\underline{p}_{D_1+1}, \overline{p}_{D_1+1}] \times \ldots \times [\underline{p}_D, \overline{p}_D] \end{aligned} \tag{4.30}$$

for some $1 \leq D_1 \leq D$ and real $\underline{p}_i \leq \overline{p}_i$ for all $i \in \{1, \ldots, D\}$.

Option price representations of form (4.28) capture a large variety of option types including plain vanilla European as well as American and other path dependent options. In Gaß et al. (2016), all of these options types are considered. Here, we focus on the case that the price (4.28) can be represented in Fourier terms. Focusing on these representations, the following paragraphs derive sufficient conditions under which the parametrized prices possess an analytic extension to an appropriate Bernstein ellipsoid such that the Chebyshev approximation method applies.

For most relevant options, the payoff profile f_K is not integrable and its Fourier transform is not well-defined. The European call and put options are prominent examples. In these cases, however, the notion of the generalized Fourier transform of Definition 2.6 applies. The following set of conditions establishes the foundation for employing the Chebyshev method for Fourier pricing.

Conditions 4.6 (Chebyshev method on Fourier prices)
Let the parameter set $\mathcal{P} = \mathcal{K} \times \mathcal{T} \times \mathcal{Q} \subset \mathbb{R}^D$ possess a hyperrectangular structure as in (4.30). Let $\varrho \in (1, \infty)^D$ and denote $\varrho^{\mathcal{K}} = (\varrho_1, \ldots, \varrho_{D_1})$ and $\varrho^{\mathcal{T}\mathcal{Q}} = (\varrho_{D_1+1}, \ldots, \varrho_D)$ and let weight $\eta \in \mathbb{R}^d$.

(4.A) For every $K \in \mathcal{K}$ the mapping $x \mapsto e^{\langle \eta, x\rangle} f_K(x)$ is in $L^1(\mathbb{R}^d)$.

(4.B) For every $z \in \mathbb{R}^d$ the mapping $K \mapsto \widehat{f_K}(z - i\eta)$ is analytic in the generalized Bernstein ellipse $B(\mathcal{K}, \varrho^{\mathcal{K}})$ and there are constants c_1, $c_2 > 0$ such that

$$\sup_{K \in B(\mathcal{K}, \varrho^{\mathcal{K}})} |\widehat{f_K}(-z - i\eta)| \leq c_1 e^{c_2 |z|} \tag{4.31}$$

for all $z \in \mathbb{R}^d$.

(4.C) For every $(T, q) \in \mathcal{T} \times \mathcal{Q}$ the exponential moment condition $\mathbb{E}\big[e^{-\langle \eta, X_T^q \rangle}\big] < \infty$ holds.

(4.D) For every $z \in \mathbb{R}^d$ the mapping $(T, q) \mapsto \varphi_{T,q}(z + i\eta)$ is analytic in the generalized Bernstein ellipse $B(\mathcal{T} \times \mathcal{Q}, \varrho^{\mathcal{T}\mathcal{Q}})$ and there are constants $\alpha \in (1, 2]$ and c_1, $c_2 > 0$ such that

$$\sup_{(T,q) \in B(\mathcal{T} \times \mathcal{Q})} |\varphi_{T,q}(z + i\eta)| \leq c_1 e^{-c_2 |z|^\alpha} \tag{4.32}$$

for all $z \in \mathbb{R}^d$.

Conditions (4.A)–(4.D) are satisfied for a large class of payoff functions and asset models, see Sections 2.4 and 2.3. More precisely, there are examples of (time-inhomogeneous) Lévy processes that fall in the scope of Conditions 4.6 indeed and we refer the interested reader to Glau (2016) for an overview and the article Gaß et al. (2016) that this chapter is based on for more details.

Theorem 4.7 (Convergence of prices)
Let $\varrho \in (1, \infty)^D$ and weight $\eta \in \mathbb{R}^d$. Under conditions (4.A)–(4.D) *we have*

$$\begin{aligned}
&\max_{p \in \mathcal{P}} |Price^p - I_{\overline{N}}(Price^{(\cdot)})(p)| \\
&\qquad \leq \sum_{i=1}^{D} 4V \frac{\varrho_i^{-N_i}}{\varrho_i^{-N_i} - 1} + \sum_{l=2}^{D} 4V \frac{\varrho_-^{-N_l}}{\varrho_l - 1} \cdot 2^{l-1} \frac{(l-1) + 2^{k-l} - 1}{\prod_{j=1}^{l-1}(1 - \varrho_j^{-1})}.
\end{aligned} \tag{4.33}$$

Proof
This is Theorem 3.2 in Gaß et al. (2016) where a proof is provided. □

4.3.2 Selected option prices

In the previous Section, Conditions 4.6 introduced a framework in which the Chebyshev approximation achieves (sub)exponential error decay. This abstract framework can be related to two concrete option pricing settings in connection with Fourier pricing as introduced by Proposition 2.20 from the preliminary chapter.

First, we assess European options in univariate Lévy models. Let r be the deterministic and constant interest rate. We consider the parametrized family of asset prices,

$$S_t^q = S_0 e^{L_t^{q'}} \tag{4.34}$$

with $t \geq 0$. For fixed $q = (S_0, r, \sigma) \in \mathcal{Q} = [\underline{S_0}, \overline{S_0}] \times [\underline{r}, \overline{r}] \times [\underline{\sigma}, \overline{\sigma}]$ we denote $q' = (r, \sigma)$ and assume $L^{q'}$ to be a Lévy process and special semimartingale with characteristics

(b, σ, F) and parametric Lévy measure F. As we have seen in Section 2.2 of the introductory Chapter 2, the characteristic function of the parametrized Lévy process can be represented by

$$\begin{aligned} \varphi_{t,q'}(z) &= \mathbb{E}\big[e^{izL_t^{q'}}\big] = e^{t\psi^{q'}(z)}, \\ \psi^{q'}(z) &= ibz + \frac{\sigma^2 z}{2} + \int_{\mathbb{R}} \big(e^{izx} - 1 - izh(x)\big)F(\mathrm{d}x). \end{aligned} \tag{4.35}$$

Additionally, we separately denote the jump part of the cumulant generating function by

$$\widetilde{\psi}(z) = \int_{\mathbb{R}} \big(e^{izx} - 1 - izx\big)F(\mathrm{d}x) \tag{4.36}$$

for later reference. We assume L is defined under a risk neutral measure. Therefore, for every $q \in \mathcal{Q}$ we assume $\mathbb{E}[e^{L_t^{q'}}] < \infty$ for some and equivalently all $t > 0$ and the drift condition

$$b = b(r, \sigma) = r - \frac{\sigma^2}{2} - \int_{\mathbb{R}} \big(e^{x} - 1 - h(x)\big)F(\mathrm{d}x), \tag{4.37}$$

to ensure that the discounted asset price process is a martingale, as already outlined by identity (2.31) in Section 2.3. In asset model S^q the fair value at time $t = 0$ of a European option with payoff written as function f_K for $K \in \mathcal{K} = [\underline{K}, \overline{K}] \subset \mathbb{R}$ with maturity $T \in \mathcal{T} = [\underline{T}, \overline{T}] \subset (0, \infty)$ is given by

$$\mathit{Price}^{(K,T,q)} = e^{-rT}\mathbb{E}\big[f_K(S_0\, e^{L_T^{q'}})\big]. \tag{4.38}$$

In order to guarantee (sub)exponential convergence of the Chebyshev interpolation, we translate condition (4.C) on exponential moments and condition (4.D) on analyticity and the upper bound into conditions on the cumulant function $\psi^{q'}$. Then the following corollary applies.

Corollary 4.8 (Exponential convergence of Fourier prices in N)
Let Conditions (4.A) and (4.B) be satisfied for weight $\eta \in \mathbb{R}$ and $\varrho^{\mathcal{K}} > 1$ and set $\mathcal{P}^{\mathcal{K}} = [\underline{K}, \overline{K}]$. Moreover, let $\mathcal{P}^{\mathcal{T}\mathcal{Q}} = [\underline{T}, \overline{T}] \times [\underline{S_0}, \overline{S_0}] \times [\underline{r}, \overline{r}] \times [\underline{\sigma}, \overline{\sigma}] \subset \mathbb{R}^4$ with $\underline{T}, \underline{S_0} > 0$ and $\underline{\sigma} \geq 0$. Assume

$$\int_{|x|>1} (e^{-\eta x} \vee e^{x})F(\mathrm{d}x) < \infty. \tag{4.39}$$

If additionally one of the following conditions is satisfied,

i) $\underline{\sigma} > 0$,

ii) there exist $\alpha \in (1, 2]$ and constants $C_1, C_2 > 0$ such that

$$\Re\big(\widetilde{\psi}\big)(z + i\eta) \leq C_1 - C_2|z|^{\alpha} \qquad \textit{for all } z \in \mathbb{R},$$

then there exist constants $C > 0$ and $\underline{\varrho} > 1$ such that

$$\max_{p \in \mathcal{K} \times \mathcal{T} \times \mathcal{Q}} \left| Price^p - I_{\overline{N}}(Price^{(\cdot)})(p) \right| \leq C \underline{\varrho}^{-\underline{N}}, \tag{4.40}$$

where $\underline{N} = \min_{1 \leq i \leq 5} N_i$.

Proof
This is Corollary 3.4 in Gaß et al. (2016) where a proof is provided. □

For an application of the Chebyshev method to the pricing of basket options in affine models both in theoretical and numerical terms, we refer the interested reader to Gaß et al. (2016).

4.3.3 Examples of payoff profiles

We enlist in Table 4.1 a selection of payoff profiles f_K for option parameter K as function of the logarithm of the underlying. By Proposition 2.20, we can represent option prices under certain conditions in Fourier terms. Therefore, the table provides the generalized Fourier transform $\widehat{f_K}$ of the respective option payoff, as well.

Type	**Payoff** $f(x)$	**Weight** η	**Fourier transform** $\widehat{f_K}(z - i\eta)$	$\widehat{f_K}$ **holomorphic in** $\log(K)$
Call	$(e^x - K)^+$	< -1	$\frac{K^{iz+1+\eta}}{(iz+\eta)(iz+1+\eta)}$	✓
Put	$(K - e^x)^+$	> 0	$\frac{K^{iz+1+\eta}}{(iz+\eta)(iz+1+\eta)}$	✓
Digital down&out	$\mathbb{1}_{x>\log(K)}$	< 0	$-\frac{K^{iz+\eta}}{iz+\eta}$	✓
Asset-or-nothing down&out	$e^x \mathbb{1}_{x>\log(K)}$	< -1	$-\frac{K^{iz+1+\eta}}{iz+1+\eta}$	✓

Table 4.1 Examples of payout profiles of a single underlying and the respective (generalized) Fourier transforms.

4.3.4 Chebyshev conditions and asset models

In this section, we shortly introduce some analyticity properties of the Fourier transforms of the Lévy models introduced in Section 2.3. For some models and some parameters, the domain of analyticity is immediately observable. For some non-trivial cases, we state the domain briefly. Throughout the section, $T > 0$ denotes the time to maturity of the option while $r \geq 0$ refers to the constant risk-free interest rate.

We define

$$\begin{aligned} \mathbb{C}^+ &= \{z \in \mathbb{C} \,|\, \Re(z) > 0\}, \\ \mathbb{C}^+_0 &= \{z \in \mathbb{C} \,|\, \Re(z) \geq 0\}, \end{aligned} \tag{4.41}$$

for later reference.

In the multivariate Black&Scholes model of Section 2.3.1, analyticity in the parameters is immediately confirmed, that is $(T, q') \mapsto \varphi_{T,q'}(z)$ is holomorphic for every $z \in \mathbb{R}^d$. The admissible parameter domain, however, is restricted to parameter constellations that encode a covariance matrix.

Remark 4.9 (Analyticity in the Black&Scholes model)
Let $\eta \in \mathbb{R}^d$ be the chosen weight in Conditions 4.6 and let the open set U be given by

$$U \subseteq \mathbb{C}^+ \times \mathbb{C}^+_0 \times \left\{\vec{\sigma} \in \mathbb{C}^{d(d+1)/2} \mid \sigma(\Re(\vec{\sigma})) \text{ positive definite}\right\}, \tag{4.42}$$

where $\sigma : \mathbb{R}^{d(d+1)/2} \to \mathbb{R}^{d\times d}$ is defined by $\sigma(\vec{\sigma})_{ij} = \sigma_{(\max\{i,j\}-1)\max\{i,j\}/2+\min\{i,j\}}$, $i, j \in \{1, \dots, d\}$, for $\vec{\sigma} \in \mathbb{R}^{d(d+1)/2}$. By construction, $\sigma(\vec{\sigma})$ is symmetric for any $\vec{\sigma} \in \mathbb{R}^{d(d+1)/2}$.

Then for every $z \in \mathbb{R}^d$, $(T, r, \vec{\sigma}) \mapsto \varphi_{T,(r,\sigma(\vec{\sigma}))}(z + i\eta)$ is analytic on U. Note that U does not depend on η.

The Merton model of Merton (1976) has been introduced in Section 2.3.2.

Remark 4.10 (Analyticity in the univariate Merton model)
In the Merton model, we find ourselves in the same situation, since the characteristic function for the Merton jump diffusion model is itself composed of analytic functions. Let $\eta \in \mathbb{R}$ be the chosen weight in Conditions 4.6 and choose the complex parameter space U open according to

$$U \subseteq \mathbb{C}^+ \times \mathbb{C}^+_0 \times \left\{(\sigma, \alpha, \beta, \lambda) \in \mathbb{C}^+ \times \mathbb{C} \times \mathbb{C}^+_0 \times \mathbb{C}^+\right\}. \tag{4.43}$$

Then for every $z \in \mathbb{R}^d$, the mapping $(T, r, \sigma, \alpha, \beta, \lambda) \mapsto \varphi_{T,(r,\sigma,\alpha,\beta,\lambda)}(z + i\eta)$ is analytic on U. Again, the domain of analyticity in the parameters does not depend on the weight η.

Recall the Normal Inverse Gaussian (NIG) model of Section 2.3.4.

Remark 4.11 (Analyticity in the univariate NIG model)
For weight $\eta \in \mathbb{R}$ from Conditions 4.6, choose an open set $U(\eta)$ with

$$U(\eta) \subseteq \mathbb{C}^+ \times \mathbb{C}_0^+ \times \mathbb{C}_0^+ \times \big\{(\alpha, \beta) \in \mathbb{C}^+ \times \mathbb{C} \;\big|\; \Re(\alpha)^2 - \Re(\beta)^2 > \Im(\alpha)^2 - \Im(\beta)^2, \quad (4.44)$$
$$\Re(\alpha)^2 - (\Re(\beta) - \eta)^2 > \Im(\alpha)^2 - \Im(\beta)^2\big\}.$$

Then for every $z \in \mathbb{R}$, $(T, r, \delta, \alpha, \beta) \mapsto \varphi_{T,(r,\delta,\alpha,\beta)}(z + i\eta)$ is analytic on $U(\eta)$.

In Section 2.3.3 we introduced the CGMY model of Carr et al. (2002).

Remark 4.12 (Analyticity in the univariate CGMY model)
The gamma function Γ that is part of the characteristic function in the CGMY model has an analytic extension to the complex semispace $\mathbb{C}^+$. Consequently, with weight $\eta \in \mathbb{R}$ from Conditions 4.6, we can choose an open set $U(\eta)$ with

$$U(\eta) \subseteq \mathbb{C}^+ \times \mathbb{C}_0^+ \times \mathbb{C}^+ \times \big\{(G, M) \in \mathbb{C}^+ \times \mathbb{C}^+ \;\big|\; \Re(G) - \eta > 0, \, \Re(M) + \eta > 0\big\}$$
$$\times \{Y \in \mathbb{C}^+ \mid \Re(Y) \in (1, 2)\}. \quad (4.45)$$

Then for every $z \in \mathbb{R}^d$, $(T, r, C, G, M, Y) \mapsto \varphi_{T,(r,C,G,M,Y)}(z + i\eta)$ for the characteristic function $\varphi_{T,q}$ of the CGMY model is analytic on $U(\eta)$.

Table 4.2 taken from Gaß et al. (2016) displays for selected Lévy models conditions on the weight $\eta \in \mathbb{R}^d$ and the index $\alpha \in (1, 2]$ that guarantee $(4.C)$ and $(4.D)$.

Class	**Conditions for** $(4.C)$, $(4.D)$ **to hold** on η	 on α
Brownian Motion with drift		$\alpha = 2$
Merton Jump Diffusion		$\alpha = 2$
Lévy jump diffusion with characteristics (b, σ, F)	$\int_{\lvert x\rvert>1} e^{\lvert\eta\rvert\lvert x\rvert} F(\mathrm{d}x) < \infty$	$\alpha = 2$
univariate CGMY with parameters (C, G, M, Y) with $Y > 1$	$\eta \in (-\min\{G, M\}, \max\{G, M\})$	$\alpha = Y$

Table 4.2 Conditions on η and α for $(4.C)$ and $(4.D)$ to hold for a fixed model parameter constellation. The selected Lévy models are described in more detail in Section 2.3.

4.3.4.1 Heston model for two assets

Here we state the two asset version of the multivariate Heston model in the special case of having a single, univariate driving volatility process $(v_t)_{t\geq 0}$. The two asset price processes are modeled as

$$S_t^1 = S_0^1 e^{H_t^1} \quad \text{and} \quad S_t^2 = S_0^2 e^{H_t^2}, \quad \text{for } t \geq 0, \tag{4.46}$$

where $H = (H^1, H^2)$ solves the following system of SDEs,

$$\begin{aligned} \mathrm{d}H_t^1 &= \left(r - \frac{1}{2}\sigma_1^2\right)\mathrm{d}t + \sigma_1\sqrt{v_t}\,\mathrm{d}W_t^1, \\ \mathrm{d}H_t^2 &= \left(r - \frac{1}{2}\sigma_2^2\right)\mathrm{d}t + \sigma_2\sqrt{v_t}\,\mathrm{d}W_t^2, \\ \mathrm{d}v_t &= \kappa(\theta - v_t)\,\mathrm{d}t + \sigma_3\sqrt{v_t}\,\mathrm{d}W_t^3, \end{aligned}$$

where the Brownian motions W_i, $i \in \{1,2,3\}$, are correlated according to $\langle W^1, W^2\rangle = \rho_{12}$, $\langle W^1, W^3\rangle = \rho_{13}$, $\langle W^2, W^3\rangle = \rho_{23}$. Following Eberlein et al. (2010), the characteristic function of H_T in this framework is

$$\begin{aligned} &\varphi_{T,(r,v_0,\kappa,\theta,\sigma_1,\sigma_2,\sigma_3,\rho_{12},\rho_{13},\rho_{23})}(z) \\ &\quad = \exp\left(Ti\left\langle\begin{pmatrix} r \\ r\end{pmatrix}, z\right\rangle\right)\exp\left(\frac{v_0}{\sigma_3^2}\frac{(a-c)(1-\exp(-cT))}{1-g\exp(-cT)}\right. \\ &\qquad\quad \left. + \frac{\kappa\theta}{\sigma_3^2}\left[(a-c)T - 2\log\left(\frac{1-g\exp(-cT)}{1-g}\right)\right]\right), \end{aligned} \tag{4.47}$$

with auxiliary functions

$$\begin{aligned} \zeta = \zeta(z) &= -\left(\left\langle z, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} z\right\rangle + \left\langle \begin{pmatrix}\sigma_1 \\ \sigma_2\end{pmatrix}, iz\right\rangle\right) \\ &\quad - \left(\sigma_1^2 z_1^2 + \sigma_2^2 z_2^2 + 2\rho_{12}\sigma_1\sigma_2 z_1 z_2 + i\sigma_1^2 z_1 + i\sigma_2^2 z_2\right), \\ a = a(z) &= \kappa - i\rho_{13}\sigma_1\sigma_3 z_1 - i\rho_{23}\sigma_2\sigma_3 z_2, \\ c = c(z) &= \sqrt{a(z)^2 - \sigma_3^2\zeta(z)}, \\ g = g(z) &= \frac{a(z) - c(z)}{a(z) + c(z)}, \end{aligned} \tag{4.48}$$

and positive parameters v_0, κ, θ and σ_3 fulfilling the Feller condition

$$\sigma_3^2 \leq 2\kappa\theta \tag{4.49}$$

ensuring an almost surely non-negative volatility process $(v_t)_{t\geq 0}$. Obviously, for each $z \in \mathbb{R}^2$, the characteristic function $\varphi_{T,(r,v_0,\kappa,\theta,\sigma_1,\sigma_2,\sigma_3,\rho_{12},\rho_{13},\rho_{23})}(z)$ of (4.47) is analytic in v_0 and θ. For further analysis of the domain of analyticity in the Heston model confer Levendorskiĭ (2012).

4.3.5 Example: Call options in Lévy models

The fair price at $t = 0$ of a call option with strike K and maturity T in a geometric Lévy model with deterministic interest rate $r \geq 0$ is given by

$$Call_T^{S_0,K} = e^{-rT}\mathbb{E}\big[S_0 e^{L_T} - K\big]^+ \tag{4.50}$$

under a risk-neutral probability measure. Noticing that

$$Call_T^{S_0,K} = e^{-rT} K\, \mathbb{E}\big[(S_0/K)e^{L_t} - 1\big]^+, \tag{4.51}$$

it suffices to interpolate the function

$$(T,S) \mapsto Call_T^{S,1} \tag{4.52}$$

on $[\underline{T},\overline{T}] \times [\underline{S_0}/\overline{K}, \overline{S_0}/\underline{K}]$ in order to approximate the prices $Call_T^{S_0,K}$ on the range $(T,S_0,K) \in [\underline{T},\overline{T}] \times [\underline{S_0},\overline{S_0}] \times [\underline{K},\overline{K}] \subset (0,\infty)^3$.

Let us fix some $\eta < -1$. From Table 4.1 we see that for every $z \in \mathbb{R}$ the mapping $K \mapsto \widehat{f^K}(z - i\eta)$ is analytic on $(0,\infty)$.

Let (b,σ,F) be the characteristics of L and recall

$$\widetilde{\psi}(z) = \int_{\mathbb{R}} \big(e^{izx} - 1 - izx\big) F(\mathrm{d}x) \tag{4.53}$$

from (4.36). Now Corollary 4.8 yields the following

Corollary 4.13 (Convergence for call options in Lévy models)
Assume one of the following conditions.

i) L is a jump diffusion Lévy process, that is it possesses a non-zero Brownian component $\sigma > 0$.

ii) There exist $\alpha \in (1,2]$ and constants $C_1, C_2 > 0$ such that

$$\Re\big(\widetilde{\psi}\big)(z + i\eta) \leq C_1 - C_2|z|^\alpha \qquad \textit{for all } z \in \mathbb{R}.$$

Let $\mathcal{P} = [\underline{T},\overline{T}] \times [\underline{S_0}/\overline{K}, \overline{S_0}/\underline{K}]$, $\zeta^1 = \frac{\overline{S_0}\underline{K} + \underline{S_0}\overline{K}}{\overline{S_0}\underline{K} - \underline{S_0}\overline{K}}$, $\zeta^2 = \frac{\overline{T}+\underline{T}}{\overline{T}-\underline{T}}$ and $\tilde{\varrho}_j = \frac{(\varrho^j)^{-N_j}}{\varrho^j - 1}$, then for every $\varrho^j \in (1, \zeta^j + \sqrt{(\zeta^j)^2 - 1})$ for $j = 1,2$, there exists a constant $V > 0$ such that

$$\max_{(T,S_0)\in\mathcal{P}} \big|Call_T^{S_0,1} - I_{N_1,N_2}(Call_{(\cdot)}^{(\cdot),1})(T,S_0)\big| \leq 4V\big(\tilde{\varrho}_1 + \tilde{\varrho}_2 + \tilde{\varrho}_2 \frac{4}{1 - \varrho_1^{-1}}\big).$$

In particular, under the assumptions of Corollary 4.13 there exists a constant $C > 0$ such that

$$\max_{(T,S_0)\in\mathcal{P}} \big|Call_T^{S_0,1} - I_{N_1,N_2}(Call_{(\cdot)}^{(\cdot),1})(T,S_0)\big| \leq C\underline{\varrho}^{-\underline{N}}, \tag{4.54}$$

where $\underline{\varrho} = \min\{\varrho_1, \varrho_2\}$ and $\underline{N} = \min\{N_1, N_2\}$.

Remark 4.14
Under the assumptions of Corollary 4.13, when fixing the maturity T, letting $\zeta = \frac{\overline{S_0}\underline{K}+\underline{S_0}\overline{K}}{\overline{S_0}\underline{K}-\underline{S_0}\overline{K}}$, we obtain the exponential error decay

$$\max_{\underline{S_0}/\overline{K}\leq S_0\leq\overline{S_0}/\underline{K}} \left| Call_T^{S_0,1} - I_N(Call_T^{(\cdot),1})(S_0)\right| \leq 4V\frac{\varrho^{-N}}{\varrho-1}, \tag{4.55}$$

for some $\varrho \in (1, \zeta + \sqrt{\zeta^2-1})$ and $V = \max_{S_0\in B([\underline{S_0}/\overline{K},\overline{S_0}/\underline{K}],\varrho)} \left| Call_T^{S_0,\overline{K}} \right|$.

Examples of Lévy jump diffusion models that satisfy condition i) of Corollary 4.13 are for example the Black&Scholes and Merton model. Examples of pure jump Lévy models satisfying condition ii) of Corollary 4.13 are provided in Glau (2016) also compare Table 4.2.

4.4 Numerical experiments

We apply the Chebyshev interpolation method to parametric option pricing considering a variety of option types in different well known option pricing models. Moreover, we conduct an error analysis, a convergence study as well as a demonstration of the gain in efficiency realized by the method. The first focuses on the accuracy that can be achieved with a reasonable number of Chebyshev interpolation points. The second confirms the theoretical order of convergence derived in Section 4.3, when the number of Chebyshev points increases. The latter visualizes the gain in efficiency in terms of improved runtimes for pricing procedures. We measure the numerical accuracy of the Chebyshev method by comparing derived prices with prices coming from a reference method. We employ the reference method not only for computing reference prices but also for computing prices at Chebyshev nodes $Price^{p^{(l_1,\ldots,l_D)}}$ with $(l_1,\ldots,l_D) \in J$ in the precomputing phase of the Chebyshev coefficients c_j, $j \in J$, in (4.13). Thereby, a comparability between Chebyshev prices and reference prices is maintained. In this section we price plain vanilla European products and use Fourier pricing by numerical integration as reference method. In Gaß et al. (2016) we also consider exotic and higher dimensional derivatives and apply the Monte Carlo method for measuring the accuracy of prices from the Chebyshev approximation.

We implemented the Chebyshev method for applications with two parameters. To that extent we pick two free parameters p_{i_1}, p_{i_2} out of (4.30), $1 \leq i_1 < i_2 \leq D$, in each model setup and fix all other parameters at reasonable constant values. We then evaluate option prices for different products on a discrete parameter grid $\overline{\mathcal{P}} \subseteq [\underline{p}_{i_1}, \overline{p}_{i_1}] \times [\underline{p}_{i_2}, \overline{p}_{i_2}]$ defined by

$$\begin{aligned} \overline{\mathcal{P}} &= \left\{ \left(p_{i_1}^{l_{i_1}}, p_{i_2}^{l_{i_2}} \right),\ l_{i_1}, l_{i_2} \in \{0,\ldots,100\} \right\}, \\ p_{i_j}^{l_{i_j}} &= \underline{p}_{i_j} + \frac{l_{i_j}}{100}\left(\overline{p}_{i_j} - \underline{p}_{i_j}\right),\ l_{i_j} \in \{0,\ldots,100\},\ j \in \{1,2\}. \end{aligned} \tag{4.56}$$

Once the prices have been derived on $\overline{\mathcal{P}}$, we compute the discrete $L^\infty(\overline{\mathcal{P}})$ and $L^2(\overline{\mathcal{P}})$ error measures,

$$\begin{aligned}\varepsilon_{L^\infty}(\overline{N}) &= \max_{p\in\overline{\mathcal{P}}}\left|Price^p - I_{\overline{N}}(Price^{(\cdot)})(p)\right|,\\ \varepsilon_{L^2}(\overline{N}) &= \sqrt{\Delta_{\overline{\mathcal{P}}}\sum_{p\in\overline{\mathcal{P}}}\left|Price^p - I_{\overline{N}}(Price^{(\cdot)})(p)\right|^2},\end{aligned} \tag{4.57}$$

where $\Delta_{\overline{\mathcal{P}}} = \frac{(\overline{p}_{i_1}-\underline{p}_{i_1})}{100}\frac{(\overline{p}_{i_2}-\underline{p}_{i_2})}{100}$, to interpret the accuracy of our implementation and of the Chebyshev method as such.

4.4.1 European options

We consider a plain vanilla European call option on one asset. The payoff profile of the derivative and its generalized Fourier transform are enlisted in Table 4.1. For these products we investigate the performance of the Chebyshev interpolation method for the Heston model and the Lévy models of Black and Scholes (1973), Merton (1976) and Carr et al. (2002). We keep the strike parameter constant at $K = 1$, and disregard interest rates, setting $r = 0$. For the three Lévy models we vary the maturity T (in years) as well as the option moneyness S_0/K whereas for the Heston model we let v_0 as one of the model parameters float. A detailed overview of the chosen parametrization is given by Table 4.3. For numerical integration in Fourier pricing we use Matlab's `quadgk` routine over the interval $[0, \infty)$ with absolute precision bound of $\varepsilon < 10^{-14}$.

The first question we address concerns the achievable accuracy with a fixed number of Chebyshev polynomials. We set $N_1 = N_2 = 7$ and precompute the Chebyshev coefficients as defined in (4.13) with $D = 2$ using the parametrization of Table 4.3 for the models therein. We evaluate the resulting polynomial over a parameter grid of dimension $D = 2$ and compute the approximate European option prices in each node. As a comparison, we also compute the respective Fourier price via numerical integration of the accordingly parametrized integrand, see Proposition (2.20). Figure 4.4 shows the results for the European call option. The Chebyshev method achieves rather homogeneous accuracy results over the four different models for $N = N_1 = N_2 = 7$ and reaches a very satisfying error level of order 10^{-6}. Increasing the number of Chebyshev points further improves the accuracy. Since at its core the implementation of the Chebyshev method consists of summing up matrices, this refinement comes at virtually no additional cost.

In Gaß et al. (2016), we perform the same analysis for a European digital down&out option. While a call payoff profile is not differentiable but at least continuous, the digital payoff function is not even continuous, compare Table 4.1. This reduction in smoothness of the payoff function reduces the accuracy of the interpolation $p \mapsto Price^{(p)}$, as well. The analysis performed in Gaß et al. (2016) empirically demonstrates, however, that the

Model	fixed parameters			free parameters	
	K	T	q	q	T
BS	$K = 1$		$\sigma = 0.25$	$S_0/K \in [0.8,\ 1.2]$	$T \in [0.5, 2]$
Merton	$K = 1$		$\sigma = 0.2,$ $\alpha = -0.02,$ $\beta = 0.1,$ $\lambda = 2.5$	$S_0/K \in [0.8,\ 1.2]$	$T \in [0.5, 2]$
CGMY	$K = 1$		$C = 0.6,$ $G = 14,$ $M = 25,$ $Y = 1.2$	$S_0/K \in [0.8,\ 1.2]$	$T \in [0.5, 2]$
Heston	$K = 1$	$T = 2$	$\kappa = 1.3,$ $\theta = 0.2^2,$ $\sigma = 0.2,$ $\rho = 0.6$	$S_0/K \in [0.8,\ 1.2]$ $v_0 \in [0.1^2,\ 0.5^2]$	

Table 4.3 Parametrization of models and the European call option for the accuracy and convergence study of the Chebyshev interpolation method.

reduced smoothness of the payoff profile effects the accuracy of the Chebyshev method only marginally.

Coming back to the European call option, we conduct an empirical convergence study for this very same setting of option and model parametrization. For an increasing degree $N = N_1 = N_2$, the Chebyshev polynomial is set up and prices over a parameter grid of structure (4.56) are computed. Again, Fourier pricing serves as a comparison. For each $N \in \{1, \dots, 35\}$, the error measures ε_{L^∞} and ε_{L^2}, defined by (4.57) on the discrete parameter grid $\overline{\mathcal{P}}$ defined in (4.56), are evaluated. We observe exponential convergence for all four models in both error measures. Figure 4.5 shows the L^∞ decay for the European call option while Figure 4.6 displays the L^2 error of the same option prices.

Following Remark 4.2, we define $\zeta = \frac{\overline{T}+\underline{T}}{\overline{T}-\underline{T}}$ and set $\varrho = \zeta + \sqrt{\zeta^2 - 1}$. The theoretical convergence analysis predicts a slope of the error decays in Figure 4.5 of at least

$$\mathcal{S} = \log_{10}(\varrho) \approx -0.47$$

or steeper. Empirically, we observe a slope for the Black&Scholes model of about $\mathcal{S}_{\text{BS}} = -0.64$, for the Merton model of $\mathcal{S}_{\text{Merton}} = -0.64$ and for the CGMY model of $\mathcal{S}_{\text{CGMY}} = -0.62$. Thus, the error in each Lévy model empirically confirms our theoretical claims.

This analysis has also been performed for a European digital down&out option. We refer the interested reader to Gaß et al. (2016) where the results of this additional study are discussed.

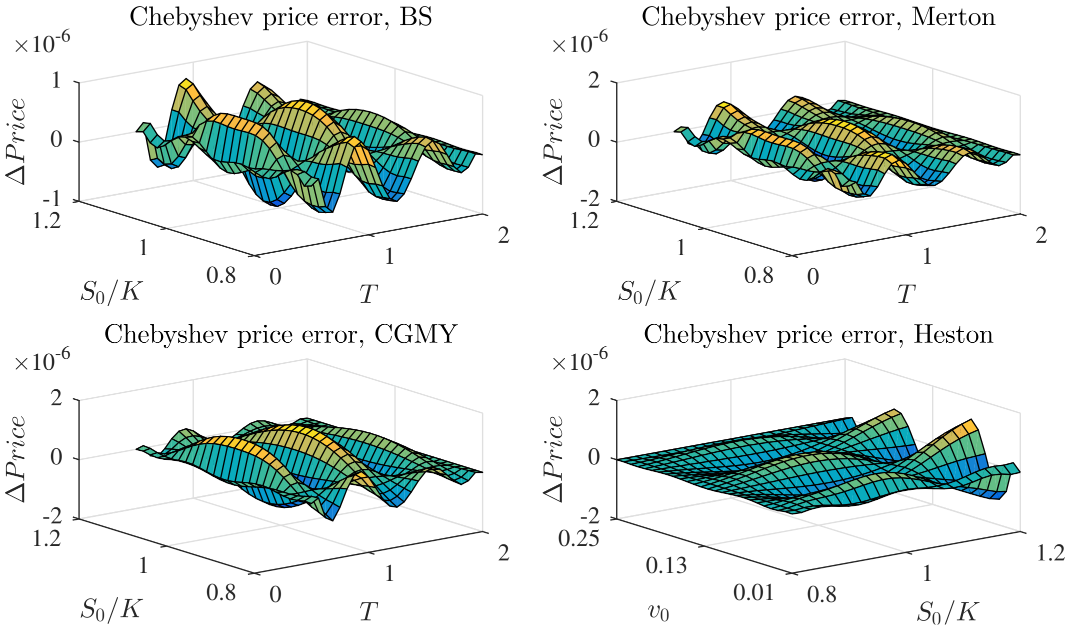

Figure 4.4 Absolute pricing error for a European call option with strike $K = 1$ in various models. We compare the Chebyshev interpolation with $N_1 = N_2 = 7$ to classic Fourier pricing by numerical integration. The parametrization of the models and the option has been chosen according to Table 4.3. We observe homogeneous accuracy results over all considered models. The structure of each surface reveals the location of Chebyshev nodes in the bounded tensorized parameter space.

4.4.2 Basket and path-dependent options

In the paper, we also consider basket and path-dependent options, see Section 3.2.2 in Gaß et al. (2016) for theoretical background on basket options in affine models and Section 4.2 in Gaß et al. (2016) for numerical results on the Chebyshev method applied to these options.

4.4.3 Study of the gain of efficiency

Finally, we investigate the gain in efficiency achieved by the method in comparison to Fourier pricing. We choose the pricing problem of a call option on the minimum of two assets as an example. The generalized Fourier transform of this option is given by Lemma 2.22 in the preliminaries. Today's values of the underlying two assets are fixed at

$$S_0^{(1)} = 1, \qquad S_0^{(2)} = 1.2. \tag{4.58}$$

Modeling the future development of the underlyings, $(S_t^{(j)})_{t\geq 0}$, $j \in \{1, 2\}$, we consider two bivariate models, separately. First, the two assets will be driven by the bi-

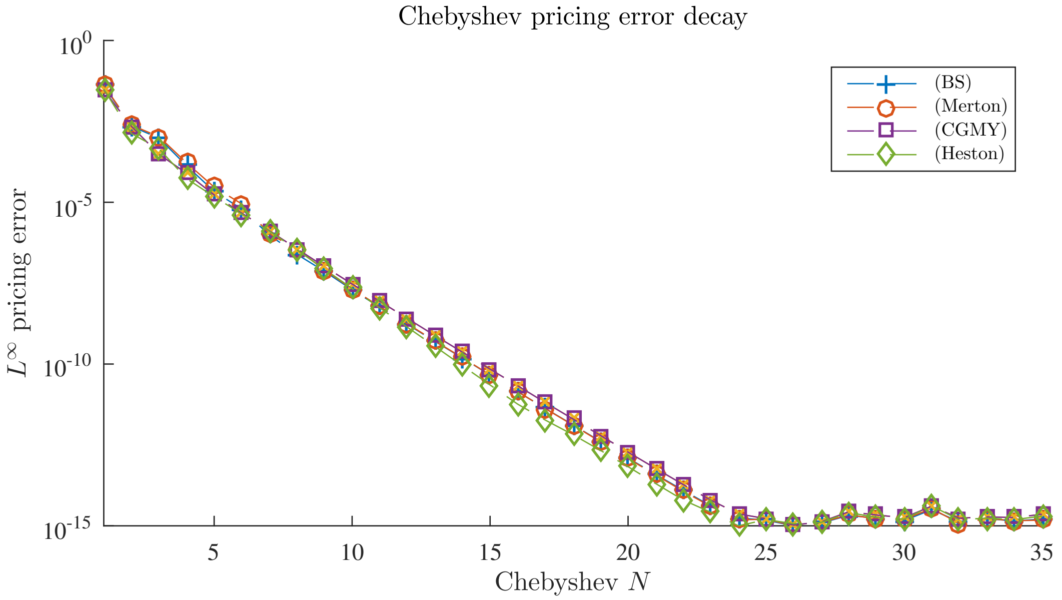

Figure 4.5 Convergence study for the Black&Scholes model, Merton, CGMY and the Heston model for prices of a European call option parametrized as stated in Table 4.3. The reference price is derived by Fourier pricing and numerical integration with an absolute accuracy of 10^{-14}, which is reached by all models for $N = N_1 = N_2 \approx 25$ the latest. The L^∞ error is depicted. The error decays in all considered very precisely coincide and thereby extend the findings illustrated by Figure 4.4.

variate Black&Scholes model of Section 2.3.1. The bivariate Black&Scholes model is parametrized by a covariance matrix $\sigma \in \mathbb{R}^{2\times 2}$ that we choose to be given by

$$\sigma_{11} = 0.2^2, \qquad \sigma_{12} = 0.01, \qquad \sigma_{22} = 0.25^2. \tag{4.59}$$

In a second efficiency study, asset movements follow the more involved bivariate Heston model in the version of Section 4.3.4.1 above for which we choose the parametrization

$$\begin{aligned} v_0 &= 0.05, & \sigma_1 &= 0.15, & \rho_{13} &= 0.01,\\ \kappa &= 0.4963, & \sigma_2 &= 0.2, & \rho_{12} &= 0,\\ \theta &= 0.2286, & \sigma_3 &= 0.1, & \rho_{23} &= 0.02. \end{aligned} \tag{4.60}$$

In both cases we neglect interest rates, thus setting $r = 0$. The benchmark method, that is Fourier pricing, is evaluated using Matlab's `quad2d` routine. We prescribe an absolute and relative accuracy of at least 10^{-8} from the integration result and integrate the Fourier integrand over the domain $\Omega = [-50,\ 50] \times [0, 50]$, prescribing a maximum number of 4000 function evaluations. The Chebyshev method is set up for pricing based on strike K and maturity T as the two free parameters taking values in the intervals

$$\begin{aligned} K &\in [K_{\min},\ K_{\max}], & K_{\min} &= 0.8, & K_{\max} &= 1.2,\\ T &\in [T_{\min},\ T_{\max}], & T_{\min} &= 0.5, & T_{\max} &= 2. \end{aligned} \tag{4.61}$$

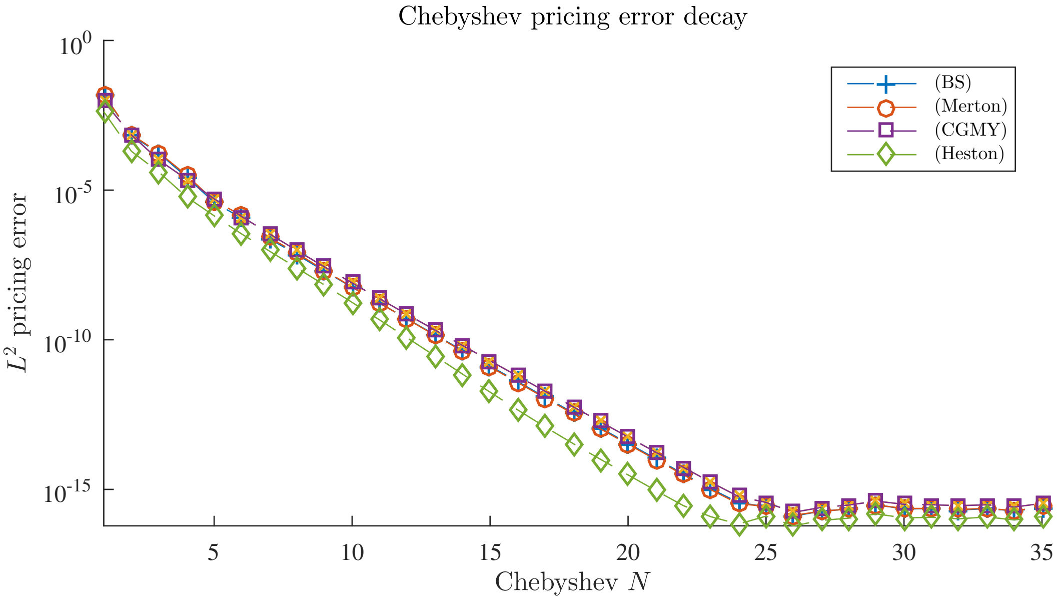

Figure 4.6 Convergence study for the Black&Scholes model, Merton, CGMY and the Heston model for prices of a European call option parametrized as stated in Table 4.3. Contrary to Figure 4.5, now the L^2 error is displayed. Here, the error decay appears slightly more nuanced. While the decays of the three Lévy models still coincide, the Heston model achieves a marginally faster convergence rate. In all considered models, the Chebyshev approximation reaches the accuracy of the reference method at $N \approx 25$.

For a fair comparison, the number of Chebyshev polynomials is chosen such that Chebyshev interpolation prices yield an accuracy that matches the accuracy of the benchmark method resulting in

$$N^{\text{BS}}_{\text{Cheby}} = 11 \quad \text{and} \quad N^{\text{Heston}}_{\text{Cheby}} = 23, \tag{4.62}$$

for the bivariate Black&Scholes model and the bivariate Heston model, respectively. Figure 4.7 illustrates the absolute errors over the whole $K \times T$ domain of interest between Fourier pricing and the Chebyshev method for both models, with the Chebyshev interpolator being based on $N^{\text{BS}}_{\text{Cheby}} + 1$ polynomials in the Black&Scholes model case and $N^{\text{Heston}}_{\text{Cheby}} + 1$ polynomials in the Heston model case.

When the offline phase of the Chebyshev method has been completed we compute 98 pricing surfaces, that is for each $M \in \{3, \dots, 100\}$ we compute prices for all parameter tuples from Θ_M defined by

$$\begin{aligned}\Theta_M = \Big\{(K^M_i, T^M_j) \;\Big|\; K^M_i &= K_{\min} + \frac{i-1}{M-1}(K_{\max} - K_{\min}),\\ T^M_j &= T_{\min} + \frac{j-1}{M-1}(T_{\max} - T_{\min}),\ \text{for } 1 \le i, j \le M\Big\}.\end{aligned} \tag{4.63}$$

The computation time consumed by the Chebyshev offline phase is measured and stored.

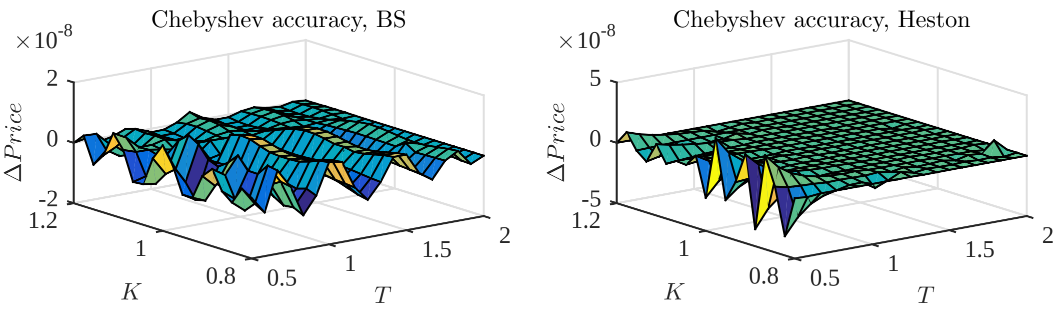

Figure 4.7 Left: Difference between prices from the Fourier method and Chebyshev interpolation in the bivariate Black&Scholes model over the whole parameter domain of interest. The model is parametrized as indicated by (4.59). Chebyshev interpolation is based on $N_{\text{Cheby}}^{\text{BS}} + 1 = 12$ Chebyshev polynomials. Right: The respective plot for the Heston model parametrized as in (4.60). Here, Chebyshev interpolation is based on $N_{\text{Cheby}}^{\text{Heston}} + 1 = 24$ Chebyshev polynomials. We achieve an absolute accuracy of order 10^{-8} in both cases, thus matching the accuracy that the benchmark method Fourier pricing provides.

Also, for each $M \in \{3, \dots, 100\}$, runtimes for deriving all $|\Theta_M| = M^2$ prices are measured and stored for both routines, the Fourier pricing method and the Chebyshev interpolation algorithm. Figure 4.8 depicts these runtime measurements visually while Table 4.4 provides a second perspective in numbers.

In the Black&Scholes model case, the offline phase required $T_{\text{offline}}^{\text{BS}} = 8$ seconds for deriving option prices at all $(N_{\text{Cheby}}^{\text{BS}} + 1)^2 = 144$ Chebyshev nodes. The more involved Heston model required $T_{\text{offline}}^{\text{Heston}} = 101$ seconds for the $(N_{\text{Cheby}}^{\text{Heston}} + 1)^2 = 576$ supporting prices. Taking this initial investment into account deems pricing with the Chebyshev method rather costly when only few option prices are derived after the offline phase has been completed. Yet, as Figure 4.8 shows and Table 4.4 quantifies, the increase in pricing speed that is achieved once the Chebyshev algorithm has been set up eventually outpaces Fourier pricing as far as (combined) pricing runtimes are concerned. From our experiments we conclude that the Chebyshev method outruns Fourier pricing in terms of total runtimes when the number of prices to be computed exceeds $(N_{\text{Cheby}}^{\text{BS}} + 1)^2$ or $(N_{\text{Cheby}}^{\text{Heston}} + 1)^2$, respectively.

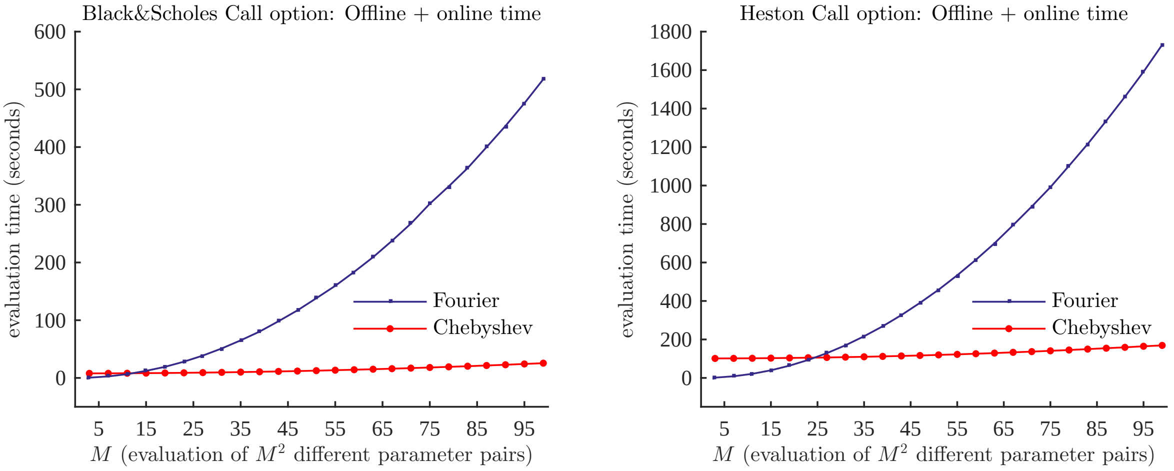

Figure 4.8 Comparison of pricing times between Fourier pricing and the Chebyshev method for a call option on the minimum of two assets in the Black&Scholes model (left) and the Heston model (right). For each $M \in \{3, \ldots, 100\}$, runtimes for deriving option prices for all M^2 parameter tupels from Θ_M defined by (4.63) are depicted. In both model cases, computation times for the Chebyshev method contain the duration of the offline phase that has to be conducted once in the beginning. The Fourier and the Chebyshev curves roughly intersect when $M = N^{\text{BS}}_{\text{Cheby}} + 1$ for the Black&Scholes model and when $M = N^{\text{Heston}}_{\text{Cheby}} + 1$ for the Heston model, respectively.

			BS				**Heston**	
M	10	50	75	100	10	50	75	100
$T^{\text{Cheby}}_{\text{online}}$ (in s)	0.18	4.54	10.20	18.11	0.70	17.58	39.66	69.82
$T^{\text{Cheby}}_{\text{offline+online}}$ (in s)	8.06	12.42	18.07	25.98	101.96	118.85	140.92	171.08
T^{Fourier} (in s)	5.34	131.96	301.82	528.74	17.60	442.62	991.33	1788.08
$\frac{T^{\text{Cheby}}_{\text{offline+online}}}{T^{\text{Fourier}}}$	151%	9.41%	5.99%	4.91%	579.27%	26.85%	14.22%	9.57%

Table 4.4 Selected results of the Chebyshev efficiency study for the bivariate Black&Scholes model and the bivariate Heston model. With increasing number of derived prices, the Chebyshev algorithm increasingly benefits from the initial investment of the offline phase. The complete record of the study is illustrated by Figure 4.8.

5 Empirical interpolation with magic points

The Chebyshev interpolation method that we considered in the previous chapter has succeeded in tremendously accelerating computational runtimes in option pricing and related tasks by separating parameter dependence from model complexity. Interestingly, during both the offline and online phase, the Chebyshev interpolation method was blind to the underlying pricing algorithm. It prescribed fixed nodes in the parameter space called Chebyshev nodes and demanded option prices at these parameter nodes disregarding the method they were computed with. On the basis of these given prices, the method interpolated inbetween with Chebyshev polynomials, thus taking a black box stance with respect to the approximated pricing routine.

Cleary, the Chebyshev algorithm conveys a very elegant appeal. Numerically, it seamlessly connects to arbitrary pricing methods and is thus not restricted in this regard. At the same time, Chebyshev approximation comes at a significant cost. Due to its tensorized extension for multivariate applications, the Chebyshev approach suffers from the curse of dimensionality, rendering it inapt for models with a large number of free parameters. There, the number of Chebyshev nodes grows exponentially in the dimensionality of the pricing problem and thus increases the number of Chebyshev nodes that need to be computed during the offline phase rather unpleasantly. The online phase is affected similarly as the number of evaluated polynomials that need to be summed up grows at the same unfavourable rate.

In this chapter we address the issue of dimensionality by introducing a different approximation method for option pricing and related tasks and we call this method *empirical interpolation for parametric option pricing*. Its name is inherited from Barrault et al. (2004) where the method has been originally developed in the context of parametric nonlinear partial differential equations. Here, we apply the concept to option pricing or rather Fourier pricing, more concisely.

Contrary to the Chebyshev method, we do not approximate prices directly but rather represent them in terms of Fourier integrals and then approximate the associated parametric integrands, instead. Tayloring the algorithm to Fourier pricing enables us to exploit the structure of the model specific Fourier integrands. We thus open the black box that the Chebyshev method left sealed and use this additional knowledge to our advantage.

Similarly to the Chebyshev method, the empirical interpolation approach separates into an offline phase and an online phase. Yet, instead of forcing the algorithm to consider prescribed locations in the parameter space, the empirical interpolation routine may decide for itself which regions in the parameter space it needs to explore more deeply and which ones it may disregard to achieve optimal global approximation results. The empirical interpolation algorithm can thus afford to spare parts of the parameter space that the Chebyshev method is obliged to examine. Precisely due to this feature, empirical interpolation is not affected by the curse of dimensionality in the same way as the Chebyshev method is.

In Section 5.1, we present the algorithm in detail. Then, Section 5.2 explores the online/offline decomposition which is more involved than the respective phases of the Chebyshev method. In Section 5.3, we derive conditions that grant exponential convergence of the algorithm before we investigate examples of asset models and payoff profiles in Section 5.4, for which these conditions are satisfied. We numerically implemented the algorithm and present the results of a thorough numerical survey in Section 5.5, containing an empirical convergence study both in and out of sample and indepth studies for several models individually. We investigate the interpolation operator of the algorithm in Section 5.6 more closely and describe a structural inconvenience of it, that we finally resolve in Section 5.7.

The results of this chapter are taken from the articles Gaß et al. (2015) and Gaß and Glau (2015) where they have been jointly developed, first. The proofs of those theoretical results that were developed by coauthors will only be referenced in this thesis. We refer the interested reader to the articles in these cases. Explanatory descriptions of the method and the accompanying results have been rewritten in parts but clearly cannot deny their close relation to the paper sources. The numerical experiments have been repeated based on a different parametrization, thereby validating the theoretical results from a different perspective, again.

5.1 Magic point interpolation for integration

We introduce the *Empirical Magic Point Interpolation method for parametric integration* presented in Gaß et al. (2015) to approximate parametric integrals of the form

$$\mathcal{I}(h_p) := \int_\Omega h_p(z)\,\mathrm{d}z, \qquad p \in \mathcal{P}, \tag{5.1}$$

with the parametric integrands

$$h_p(z) = h_{(K,T,q)}(z) := \Re\big(\widehat{f_K}(-z)\varphi_{T,q}(z)\big) \tag{5.2}$$

for every $p = (K, T, q)$ in a given parameter set $\mathcal{P}$. With $\mathcal{P}$ we associate

$$\mathcal{U} := \big\{h_p : \Omega \to \mathbb{R} \,|\, p \in \mathcal{P}\big\}, \tag{5.3}$$

the set of all parametric integrands.

Within this chapter, the integrands h_p, $p \in \mathcal{P}$, will be Fourier pricing integrands. The method, however, not only applies to integrals with integrands of this kind. Instead, integrands of more general type can be considered and their integral value can be approximately derived. We investigate this more general integration approximation routine in Gaß and Glau (2015).

The approximation of functions using the Empirical Interpolation method that our integration approximation rests on has been originally introduced by Maday et al. (2009). Before we present their algorithm, let us cite some basic assumptions from Gaß et al. (2015) that ensure the well-definedness of the iterative procedure.

Assumptions 5.1 (Approximation framework)
Let $(\Omega, \|.\|_\infty)$ and $(\mathcal{P}, \|.\|_\infty)$ be compact, $\mathcal{P} \times \Omega \ni (p, z) \mapsto h_p(z)$ bounded and $p \mapsto h_p$ be sequentially continuous, i.e. for every sequence $p_i \to p$ we have $\|h_{p_i} - h_p\|_\infty \to 0$. Moreover, $\mathcal{U}$ is nontrivial in the sense that the set contains elements other than the function that is constantly zero.

For $M \in \mathbb{N}$ we define a mapping I_M from $\mathcal{U}$ to a tensor specified by

$$I_M(h)(p,z) := \sum_{m=1}^{M} h_p(z_m^*)\theta_m^M(z) \tag{5.4}$$

and the *Magic Point Integration* with M points by

$$\mathcal{I}_M(h)(p) := \sum_{m=1}^{M} h_p(z_m^*) \int_\Omega \theta_m^M(z)\,\mathrm{d}z \tag{5.5}$$

with

$$\theta_m^M(z) := \sum_{j=1}^{M} (B^M)^{-1}_{jm} q_j(z), \qquad B^M_{jm} := q_m(z_j^*), \tag{5.6}$$

where we denote by $(B^M)^{-1}_{jm}$ the entry in the jth line and mth column of the inverse of matrix B^M. By definition, B^M is a lower triangular matrix with unity diagonal and is thus invertible. We could call I_M the interpolator of integrands and $\mathcal{I}_M$ the interpolator of integrals. The *magic points* $z_1^*, \dots, z_M^* \in \Omega$ and the *basis functions* $q_1, \dots, q_M$ are recursively defined in the following way:

In the first step, let

$$u_1 := \arg\max_{u \in \mathcal{U}} \|u\|_\infty, \quad z_1^* := \arg\max_{z \in \Omega} |u_1(z)|, \quad q_1(\cdot) := \frac{u_1(\cdot)}{u_1(z_1^*)}. \tag{5.7}$$

Note that thanks to Assumptions 5.1, these operations are well-defined. Then, recursively, as long as there are at least M linearly independent functions in $\mathcal{U}$, u_M is chosen according to a greedy procedure: The algorithm chooses u_M as the function in the set

$\mathcal{U}$ which is worst represented by the approximation with the previously identified $M-1$ magic points and basis functions,

$$u_M := \arg\max_{u\in\mathcal{U}} \|u - I_{M-1}(u)\|_\infty. \tag{5.8}$$

Since every $u \in \mathcal{U}$ is a parametric function, $u = h_p$ for some $p \in \mathcal{P}$, it can be identified by the associated parameter p. We call $p^*_M \in \mathcal{P}$ identifying u_M in (5.8) the Mth *magic parameter.* In the same spirit, let

$$z^*_M := \arg\max_{z\in\Omega} \big|u_M(z) - I_{M-1}(u_M)(z)\big|, \tag{5.9}$$

and we call z^*_M the Mth *magic point.* The Mth *basis function* is the residual, normed to 1, when evaluated at the new magic point z^*_M,

$$q_M(\cdot) := \frac{u_M(\cdot) - I_{M-1}(u_M)(\cdot)}{u_M(z^*_M) - I_{M-1}(u_M)(z^*_M)}. \tag{5.10}$$

The functionality of the empirical integration operator $\mathcal{I}_M$ is thus based on magic points and taylored to Fourier integrands and integrals. To emphasize those building blocks, we also call the whole algorithm *MagicFT* method sometimes. Some general features and properties of the algorithm outlined above are summarized by Appendix B.

Note the well-definedness of the operations in the iterative step thanks to Assumptions 5.1 and the fact that the denominator in (5.10) is only zero, if all functions in $\mathcal{U}$ are perfectly represented by the interpolation I_{M-1}. In that case, however, they span a linear space of dimension $M-1$ or less and the procedure would have stopped already.

5.2 The online/offline decomposition of the algorithm

The magic point integration operator $\mathcal{I}_M$ of the empirical interpolation algorithm approximates prices for a given parameter $p \in \mathcal{P}$ from the parameter space. The operator consists of components some of which depend on this parameter p and others that do not. Naturally, the whole algorithm thus splits into an offline and an online phase. During the offline phase, numerically intense computations are conducted and the respective results are stored from which the online phase later benefits. Recall the integration operator $\mathcal{I}_M$ as given in (5.5) by

$$\mathit{Price}^p \approx \mathcal{I}_M(h)(p) = \sum_{m=1}^{M} \underbrace{h_p(z^*_m)}_{\text{ii) online phase}} \underbrace{\int_\Omega \theta^M_m(z)\,\mathrm{d}z}_{\text{i) offline phase}}. \tag{5.11}$$

The two phases the algorithm relies on can be summarized as follows.

i) **Offline phase**
In the offline phase, the algorithm explores the parameter space $\mathcal{P}$ iteratively and identifies in each iteration associated integrands that are approximated worst by solving the optimization problems (5.8) and (5.9). The solutions to these problems consist of integrands u_m that are worst approximated and locations z^*_m on the integration domain of the respective Fourier integral where this poor approximation shows. The algorithm iteratively solves these optimization problems until it reaches a prescribed global approximation precision. Then, basis functions q_m are assembled from all these identified integrands resulting in functions θ^M_m which are integrated and stored. The offline phase is only conducted once.

ii) **Online phase**
For a given parameter $p \in \mathcal{P}$ of interest, the online phase serves the only purpose of determining the coefficients for the integrated θ^M_m functions such that the weighted sum approximates $Price^p$. To this end, the respective Fourier integrand h_p is evaluated at the magic points z^*_m and thus those coefficients are found.

The following section provides the theoretical requirements for exponential convergence of the algorithm.

5.3 Convergence analysis of magic point integration

A general convergence result for Magic Point Interpolation has been originally derived in Maday et al. (2009). In Gaß et al. (2015), we link their convergence result to the best linear n-term approximation that is formally expressed by the Kolmogorov n-width. For a real or complex normed linear space $\big(\mathcal{X}, \|\cdot\|\big)$ and $\mathcal{U} \subset \mathcal{X}$, the *Kolmogorov n-width* is defined as

$$d_n(\mathcal{U},\mathcal{X}) = \inf_{\mathcal{U}_n \in \mathcal{E}(\mathcal{X},n)} \sup_{g \in \mathcal{U}} \inf_{f \in \mathcal{U}_n} \|g - f\|, \tag{5.12}$$

where $\mathcal{E}(\mathcal{X}, n)$ is the set of all n dimensional subspaces of $\mathcal{X}$. Denoting by $\big(L^\infty(\Omega,\mathbb{C}), \|\cdot\|_\infty\big)$ the Banach space of functions mapping from $\Omega \subset \mathbb{C}^d$ to $\mathbb{C}$ that are bounded in the supremum norm we cite the following proposition from Gaß et al. (2015).

Proposition 5.2 (Convergence of the empirical interpolation method)
For the set $\mathcal{U}$ from (5.3) and $M \in \mathbb{N}$

(5.*A*) *assume $\Omega \subset \mathbb{C}^d$ and Assumptions 5.1,*

(5.*B*) *assume there exist constants $\alpha > \log(4)$ and $c > 0$ such that*

$$d_M\big(\mathcal{U}, L^\infty(\Omega,\mathbb{C})\big) \le c e^{-\alpha M}.$$

Then for arbitrary $\varepsilon > 0$ and $C := \frac{c}{4}e^\alpha + \varepsilon$ we have for all $u \in \mathcal{U}$ that

$$\big\|u - I_M(u)\big\|_\infty \le C M e^{-(\alpha - \log(4))M}. \tag{5.13}$$

Proof
The proposition directly follows from Theorem 2.4 in Maday et al. (2009), where a slightly different version that does not explicitly use the Kolmogorov n-width is provided. We therefore presented a detailed version of the proof of the proposition in Appendix A of Gaß et al. (2015). □

5.3.1 Exponential convergence of magic point integration

As in the previous Chapter 4, we formulate our analyticity assumptions in terms of (generalized) Bernstein ellipses. Recall the definition of a (generalized) Bernstein ellipse $B([\underline{b},\overline{b}],\varrho)$ for $\underline{b} < \overline{b} \in \mathbb{R}$ and ellipse parameter $\varrho > 1$ from Definition 2.43. Using this concept, we formulate two analyticity conditions in order to estimate the error resulting from the Magic Point Interpolation method.

Conditions 5.3 (Analyticity conditions)

(5.A) *The function $(p,z) \mapsto h_p(z)$ is continuous on $\mathcal{P}\times\Omega$ and there exist functions $H_1 : \mathcal{P}\times\Omega \to \mathbb{C}$ and $H_2 : \mathcal{P} \to \mathbb{C}$ such that for all $(p,z) \in \mathcal{P}\times\Omega$,*

$$h_p(z) = H_1(p,z)H_2(p)$$

and $H_1(p,z)$ has an extension $H_1 : \mathcal{P}\times B(\Omega,\varrho) \to \mathbb{C}$ such that, for all fixed $p \in \mathcal{P}$ the mapping $z \mapsto H_1(p,z)$ is analytic in the interior of the generalized Bernstein ellipse $B(\Omega,\varrho)$.

(5.B) *The function $(p,z) \mapsto h_p(z)$ is continuous on $\mathcal{P}\times\Omega$ and there exist functions $H_1 : \mathcal{P}\times\Omega \to \mathbb{C}$ and $H_2 : \Omega \to \mathbb{C}$ such that for all $(p,z) \in \mathcal{P}\times\Omega$,*

$$h_p(z) = H_1(p,z)H_2(z)$$

and $H_1(p,z)$ has an extension $H_1 : B(\mathcal{P},\varrho)\times\Omega \to \mathbb{C}$ such that, for all fixed $z \in \Omega$ the mapping $p \mapsto H_1(p,z)$ is analytic in the interior of the generalized Bernstein ellipse $B(\mathcal{P},\varrho)$.

Condition (5.A) is tailored to the case of univariate integration domains and (5.B) to the case of univariate parameter spaces.

5.3.1.1 Parametric European options, generalized moments and other univariate integrals

In the generic situation where option prices have to be evaluated for a large set of different parameter constellations, a parametric integral of form (5.1) for a high dimensional parameter space and a univariate integration domain needs to be computed. This comprises many well known examples such as prices of European and exotic options and sensitivities of these prices as expressed by the Greeks for different option and model parameters. Also risk measures like VaR and ES and other generalized moments or parametric univariate integrals fall into the scope of this paragraph.

Theorem 5.4
Let $\Omega \subset \mathbb{R}$ and $\mathcal{P} \subset \mathbb{R}^D$ be compact. Fix some $\eta \in \mathbb{R}$, some $\varrho > 4$ and assume that integrability conditions (Exp) *and* (Int) *as well as analyticity condition* (5.A) *are satisfied. Then for all $p \in \mathcal{P}$ and $M \in \mathbb{N}$,*

$$\begin{aligned}\big\|h_p - I_M(h_p)\big\|_\infty &\leq CM(\varrho/4)^{-M},\\ \big|\mathcal{I}(h_p) - \mathcal{I}_M(h_p)\big| &\leq C|\Omega|M(\varrho/4)^{-M},\end{aligned}$$

where

$$C = \frac{\varrho}{\varrho - 1} \max_{(p,z)\in\mathcal{P}\times B(\Omega,\varrho)} \big|H_1(p,z)\big| \max_{p\in\mathcal{P}} |H_2(p)|. \tag{5.14}$$

The proof is provided in Gaß and Glau (2015) and in the appendix of Gaß et al. (2016).

5.3.1.2 Basket options, multivariate generalized moments and other multivariate integrals

A similar result applies to the error analysis of Magic Point Integration for basket options for a single free parameter. Hence, real-time pricing of basket options with either varying strikes or varying maturities in a fixed calibrated asset model could benefit. Additionally, the computation of generalized moments such as covariances, and general multivariate integrals with a single varying parameter in the integrand can be approximated with the method. The result is stated in Section 4.1.2 of Gaß et al. (2015).

5.4 Examples of payoff profiles and asset models

We apply the MagicFT method to the pricing problem using univariate payoff profiles and some well known Lévy asset models.

5.4.1 Examples of univariate payoff profiles

In Table 5.1 we summarize a selection of payoff profiles f_K for option parameter K as function of the logarithm of the underlying asset. We state the range of possible weight values η such that $x \mapsto e^{\eta x} f_K(x) \in L^1(\mathbb{R})$ and the respective generalized Fourier transform exists.

Examining the generalized Fourier transforms of the payoff profiles f_K in Table 5.1, we realize that all of them admit a factorization in the spirit of condition (5.A) as

$$\widehat{f_K}(z + i\eta) = K^{iz+c} H_2(z) \tag{5.15}$$

for some $c \in \mathbb{R}$. While all of the payoff profiles f_K of Table 5.1 either are not differentiable or even discontinuous, the mapping $z \mapsto K^{iz+c}$ is a holomorphic function and thus perfectly fits the requirements of Theorem 5.4.

Type	**Payoff** $f_K(x)$	**Weight** η	**Fourier transform** $\widehat{f_K}(z-i\eta)$
Call	$(e^x-K)^+$	<-1	$\frac{K^{iz+1+\eta}}{(iz+\eta)(iz+1+\eta)}$
Put	$(K-e^x)^+$	>0	$\frac{K^{iz+1+\eta}}{(iz+\eta)(iz+1+\eta)}$
Digital down&out	$\mathbb{1}_{x>\log(K)}$	<0	$-\frac{K^{iz+\eta}}{iz+\eta}$
Asset-or-nothing down&out	$e^x\mathbb{1}_{x>\log(K)}$	<-1	$-\frac{K^{iz+1+\eta}}{iz+1+\eta}$

Table 5.1 Typical payoff profiles for single stock options and the respective generalized Fourier transform.

5.4.2 Examples of asset models

We present a selection of asset models that we use for pricing options in the numerical experiments in section 5.5 below. The MagicFT algorithm, as we apply it, operates on Fourier integrands that consist of the generalized Fourier transform of the option profile, $\widehat{f_K}$, as well as the Fourier transform of the process that drives the underlying asset at maturity, $\varphi_{T,q}$. Theoretically, Theorem 5.4 requires the analytic property from the characteristic function $\varphi_{T,q}$ of the model in the sense of condition $(5.A)$. Yet, for some models fulfilling this requirement means strongly restricting the parameter space. This would leave us with parameter spaces that are too limited for practical purposes. Empirically, however, we observe that condition $(5.A)$ may be replaced by a much weaker condition while still maintaining exponential convergence. The existence of a shared strip of analyticity $S_R(\eta)$ of width $R\in(0,\infty)^d$ given by

$$S_R(\eta)=\mathbb{R}^d+i(\eta-R,\eta+R)\subset\mathbb{C}^d, \tag{5.16}$$

where all $\xi\mapsto\varphi_{T,q}(\xi)$, $T\in\mathcal{T}$, $q\in\mathcal{Q}$, are analytic on, grants exponential convergence of the algorithm, already. Enforcing such a shared strip means imposing conditions on the model parameter space $\mathcal{Q}$, too. Yet these restrictions turn out to be rather mild compared to the stronger condition $(5.A)$ of Theorem 5.4.

In the following model presentations we denote by $\widetilde{\mathcal{Q}}$ the parameter space that the model as such is defined on. From this we derive admissible parameter sets $\mathcal{Q}$ such that condition $(5.A)$ is satisfied. If this is not possible, they are chosen to guarantee the existence of a shared strip of analyticity according to (5.16). Throughout the following model introductions, constant $r > 0$ denotes the risk-free interest rate.

5.4.2.1 Multivariate Black&Scholes model

Recall the d-variate Black&Scholes model introduced in Section 2.3.1. The parameter space determines the underlying covariance matrix $\sigma \in \mathbb{R}^{d\times d}$ exclusively. Thus, $\widetilde{\mathcal{Q}}$ is defined as

$$\widetilde{\mathcal{Q}} = \{q \in \mathbb{R}^{d(d+1)/2} \mid \det(\sigma(q)) > 0\} \subset \mathbb{R}^{d(d+1)/2} \tag{5.17}$$

with the function $\sigma : \mathbb{R}^{d(d+1)/2} \to \mathbb{R}^{d\times d}$ defined by

$$\sigma(q)_{ij} = q_{(\max\{i,j\}-1)\max\{i,j\}/2+\min\{i,j\}}, \qquad i, j \in \{1, \dots, d\}. \tag{5.18}$$

For each $q \in \widetilde{\mathcal{Q}}$ given by (5.17), the characteristic function of the d-variate Black&Scholes model is analytic in z on the whole of $\mathbb{C}^d$. We thus may choose the parameter set $\mathcal{Q}$ for the MagicFT algorithm according to the following remark.

Remark 5.5 ($\mathcal{Q}$ for the multivariate Black&Scholes model)
Let $\underline{\sigma}_i \le \overline{\sigma}_i \in \mathbb{R}^+$ for all $i \in \{1, \dots, d(d+1)/2\}$. Define

$$\mathcal{Q} = \{q \in \mathbb{R}^{d(d+1)/2} \,|\, \underline{\sigma}_i \le q_i \le \overline{\sigma}_i \textit{ such that } \det(\sigma(q)) > 0\} \tag{5.19}$$

with the function σ given by (5.18)*. With the parameter set $\mathcal{Q}$ defined as above and compact $\mathcal{T} \subset \mathbb{R}^+$, the characteristic function of the Black&Scholes model satisfies condition* (5.A) *of Theorem 5.4.*

5.4.2.2 Univariate Merton jump diffusion model

We introduced the univariate Merton jump diffusion model by Merton (1976) in Section 2.3.2. As we have seen there, the model parameter space is given by

$$\widetilde{\mathcal{Q}} = \{(\sigma, \alpha, \beta, \lambda) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}_0^+ \times \mathbb{R}^+\} \subset \mathbb{R}^4 \tag{5.20}$$

and the characteristic function of X_T^q with $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$ computes to

$$\varphi_{T,q}(z) = \exp\left(T\left(ibz - \frac{\sigma^2}{2}z^2 + \lambda\left(e^{iz\alpha - \frac{\beta^2}{2}z^2} - 1\right)\right)\right), \tag{5.21}$$

for all $z \in \mathbb{R}$, with no-arbitrage condition

$$b = r - \frac{\sigma^2}{2} - \lambda\left(e^{\alpha + \frac{\beta^2}{2}} - 1\right). \tag{5.22}$$

As in the univariate Black&Scholes model, for each $q \in \mathcal{Q}$ and $T > 0$, the characteristic function $\varphi_{T,q}$ of the Merton model is holomorphic and the set $\mathcal{Q}$ for an application of the MagicFT algorithm to the univariate Merton model is defined by the following remark.

Remark 5.6 ($\mathcal{Q}$ for the Merton model)
Let $\underline{\sigma} \leq \overline{\sigma} \in \mathbb{R}^+$, $\underline{\alpha} \leq \overline{\alpha} \in \mathbb{R}$, $\underline{\beta} \leq \overline{\beta} \in \mathbb{R}_0^+$ and $\underline{\lambda} \leq \overline{\lambda} \in \mathbb{R}^+$. Define

$$\begin{aligned}\mathcal{Q} = \{(\sigma, \alpha, \beta, \lambda) \in \mathbb{R}^4 \mid \underline{\sigma} \leq \sigma \leq \overline{\sigma}, \quad &\underline{\alpha} \leq \alpha \leq \overline{\alpha}, \\ \underline{\beta} \leq \beta \leq \overline{\beta}, \quad &\underline{\lambda} \leq \lambda \leq \overline{\lambda}\}.\end{aligned} \tag{5.23}$$

With the parameter set $\mathcal{Q}$ defined as above and compact $\mathcal{T} \subset \mathbb{R}^+$, the characteristic function of the Merton model satisfies condition (5.*A*) *of Theorem 5.4.*

5.4.2.3 Univariate CGMY model

Details on the CGMY model can be found in Section 2.3.3. With the model parameter space given by

$$\widetilde{\mathcal{Q}} = \{(C, G, M, Y) \in \mathbb{R}^+ \times \mathbb{R}_0^+ \times \mathbb{R}_0^+ \times (1, 2) \,|\, (M-1)^Y \in \mathbb{R}\} \subset \mathbb{R}^4, \tag{5.24}$$

the associated characteristic function of X_T^q with $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$ computes to

$$\begin{aligned}\varphi_{T,q}(z) = \exp \big(T \big(&ibz + C\Gamma(-Y) \\ &\big[(M-iz)^Y - M^Y + (G+iz)^Y - G^Y\big]\big)\big),\end{aligned} \tag{5.25}$$

for all $z \in \mathbb{R}$, where $\Gamma(\cdot)$ denotes the Gamma function. For no-arbitrage pricing we set the drift $b \in \mathbb{R}$ to

$$b = r - C\Gamma(-Y)\left[(M-1)^Y - M^Y + (G+1)^Y - G^Y\right]. \tag{5.26}$$

The condition $(M-1)^Y \in \mathbb{R}$ in (5.24) guarantees $b \in \mathbb{R}$. Contrary to the models of Black and Scholes (1973) and Merton (1976), the domain in $\mathbb{C}$ that the characteristic function of the CGMY model is analytic on does not exist independently of its parametrization. Consequently, Theorem 5.4 does not apply to pricing in the CGMY model unless the parameter set that the algorithm may choose from is unreasonably restricted. Yet, empirically we maintain exponential convergence in the CGMY model case when $\mathcal{Q}$ and η are chosen such that all $\xi \mapsto \varphi_{T,q}(\xi)$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$, share a common strip of analyticity $S_R(\eta)$ as introduced in (5.16) depending on $\eta \in \mathbb{R}$ and $R > 0$, the desired strip width. In the following, we derive conditions which guarantee the existence of such a strip. The result of our analysis will consist in a combined suggestion for the weight value η that complies with the restriction posed by the option choice as outlined by Table 5.1 and a set of restrictions on the parameter space. These restrictions guarantee a shared strip of analyticity as described above achieving a certain prescribed width $R > 0$.

Strip of analyticity for CGMY Before we are able to derive conditions on the parameter space that originate a shared strip of analyticity, let us first determine the strip of maximal width $R > 0$ that an individually parameterized characteristic function of the CGMY model $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, is analytic on.

This strip in $\mathbb{C}$ is derived by analyzing the characteristic function $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, of the CGMY process on the domain of integration in (2.51) of Proposition 2.20 for different weight values. Let $\widetilde{\eta} \in \mathbb{R}$ and consider the characteristic function $\varphi_{T,q}$ on the line

$$z_{\widetilde{\eta}}(\xi) = \xi + i\widetilde{\eta}, \qquad \xi \in \mathbb{R}. \tag{5.27}$$

The values of $\widetilde{\eta}$ for which $\varphi_{T,q}$ is analytic on the associated line (5.27) determine the width of the strip of analyticity of $\varphi_{T,q}$. For these values of $\widetilde{\eta} \in \mathbb{R}$, both mappings

$$\begin{aligned} &\xi \mapsto (M - iz_{\widetilde{\eta}}(\xi))^Y, \\ &\xi \mapsto (G + iz_{\widetilde{\eta}}(\xi))^Y \end{aligned}$$

need to be analytic on $\mathbb{R}$. By (5.27), we have

$$\xi \mapsto (M - iz_{\widetilde{\eta}}(\xi))^Y = (M + \widetilde{\eta} - i\xi)^Y,$$

and

$$\xi \mapsto (G + iz_{\widetilde{\eta}}(\xi))^Y = (G - \widetilde{\eta} - i\xi)^Y.$$

For analyticity of these two quantities on $\mathbb{R}$ we need to ensure that both

$$M + \widetilde{\eta} > 0, \tag{5.28}$$

$$G - \widetilde{\eta} > 0, \tag{5.29}$$

hold. Inequalities (5.28) and (5.29) yield bounds η^-, η^+ given by

$$\begin{aligned} \eta^+ &= G, \\ \eta^- &= \ -M. \end{aligned} \tag{5.30}$$

These two bounds span the strip of analyticity $S_R(\eta)$ for an individually parametrized characteristic function of the CGMY model, wherein $\eta = (\eta^+ + \eta^-)/2 = (G - M)/2$ and diameter $2R = G + M$, as shown in Figure 5.1.

Now we can translate these findings to conditions on the model parameter set to derive a compact set $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ and a value for $\eta \in \mathbb{R}$ that ensure a common strip of analyticity $S_R(\eta)$ for all mappings $\xi \mapsto \varphi_{T,q}(\xi)$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$. From our considerations during the derivation above and in particular by (5.30) we conclude that such a $\mathcal{Q}$ and η need to satisfy

$$\max_{(C,G,M,Y) \in \mathcal{Q}} -M < \eta < \min_{(C,G,M,Y) \in \mathcal{Q}} G. \tag{5.31}$$

We limit the rest of this analysis to the case of a call option where we necessarily have

$$\eta < -1 \tag{5.32}$$

by Table 5.1. With $G \geq 0$ due to the model parametrization (5.24), the second inequality in (5.31) trivially holds automatically. Combining (5.31) and (5.32) thus yields condition

$$\max_{(C,G,M,Y) \in \mathcal{Q}} -M < \eta < -1. \tag{5.33}$$

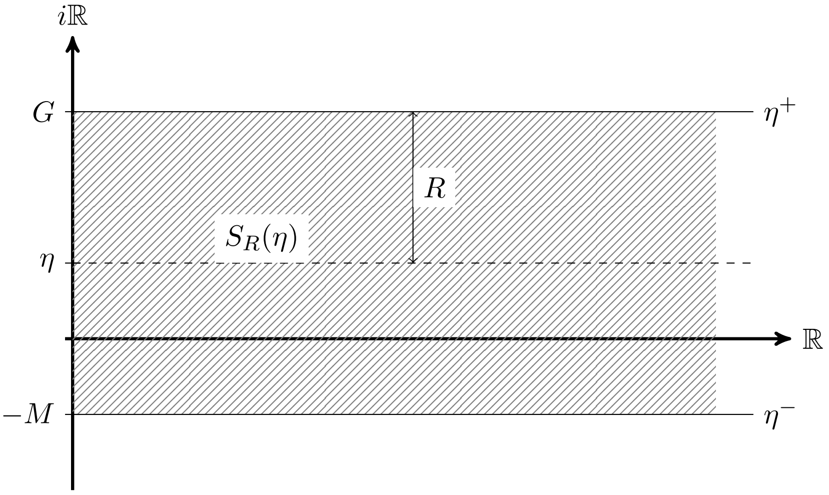

Figure 5.1 For fixed parametrization $q \in \widetilde{\mathcal{Q}}$, the hatched area visualizes the strip of analyticity of the characteristic function of the CGMY process at $T \in \mathcal{T}$, X_T^q. Its bounds are determined by $G \geq 0$ and $M \geq 0$.

A strip width of $R > 0$ consequently follows if the final strip condition

$$\min_{(C,G,M,Y)\in\mathcal{Q}} M > 1 + 2R \tag{5.34}$$

is satisfied. In other words, choosing $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ satisfying condition (5.34) and setting

$$\eta = -\min_{(C,G,M,Y)\in\mathcal{Q}} (M+1)/2 \tag{5.35}$$

yields a strip of analyticity $S_R(\eta)$ with diameter $2R$ that all of the mappings $\xi \mapsto \varphi_{T,q}(\xi)$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$, share. We collect and summarize these results in the following remark.

Remark 5.7 ($\mathcal{Q}$ for the CGMY model)
Let $\underline{C} \leq \overline{C} \in \mathbb{R}^+$, $\underline{G} \leq \overline{G} \in \mathbb{R}_0^+$, $1 \leq \underline{M} \leq \overline{M} \in \mathbb{R}_0^+$ and $\underline{Y} \leq \overline{Y} \in (1,2)$. Let $R > 0$ and define

$$\begin{aligned}\mathcal{Q} = \{(C,G,M,Y) \in \mathbb{R}^4 \,|\, & \underline{C} \leq C \leq \overline{C}, \quad \underline{G} \leq G \leq \overline{G},\\ & \underline{M} \leq M \leq \overline{M}, \quad \underline{Y} \leq Y \leq \overline{Y},\\ & (M-1)^Y \in \mathbb{R},\\ & M + 2R > 1\}.\end{aligned} \tag{5.36}$$

All $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$, share a common strip of analyticity $S_R(\eta)$ with

$$\eta = -\frac{\left(\min_{(C,G,M,Y)\in\mathcal{Q}} M\right) + 1}{2}. \tag{5.37}$$

While the characteristic function of the CGMY model parametrized by $\mathcal{Q}$ of (5.36) *in general does not satisfy condition* (5.*A*) *of Theorem 5.4, empirically we still observe exponential convergence of the MagicFT algorithm.*

Additionally, to avoid forcing the algorithm to support unrealistic parameter constellations, impose the following additional plausibility restriction.

Remark 5.8 (Plausibility constraint on $\mathcal{Q}$ in the CGMY model)
The implied variance σ^2_{CGMY} of a CGMY process $(X^q_t)_{t\geq 0}$, $q = (C, G, M, Y) \in \widetilde{\mathcal{Q}}$, at $t = 1$ is given by

$$\sigma^2_{CGMY} = C\Gamma(2-Y)\left(\frac{1}{M^{2-Y}} + \frac{1}{G^{2-Y}}\right),$$

see Carr et al. (2002). For appropriate constants $0 < \sigma_- < \sigma_+$ consider imposing the additional condition

$$\sigma^2_- \leq C\Gamma(2-Y)\left(\frac{1}{M^{2-Y}} + \frac{1}{G^{2-Y}}\right) \leq \sigma^2_+$$

for all $(C, G, M, Y) \in \mathcal{Q}$ of Remark 5.7 thus keeping supported variance levels within reasonable bounds.

5.4.2.4 Univariate Normal Inverse Gaussian model

The Normal Inverse Gaussian (NIG) model has been introduced in Section 2.3.4. The parameterization of the univariate version consists of $\delta, \alpha > 0$, $\beta \in \mathbb{R}$, with $\alpha^2 > \beta^2$. The model parameter set $\widetilde{\mathcal{Q}}$ is thus given by

$$\widetilde{\mathcal{Q}} = \left\{(\delta, \alpha, \beta) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R} \,|\, \alpha^2 > \beta^2, \alpha^2 \geq (\beta+1)^2\right\} \subset \mathbb{R}^3. \tag{5.38}$$

The characteristic function of X^q_T for this model is given by

$$\varphi_{T,q}(z) = \exp\left(T\left(ibz + \delta\left(\sqrt{\alpha^2-\beta^2} - \sqrt{\alpha^2-(\beta+iz)^2}\right)\right)\right) \tag{5.39}$$

for $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, wherein the no-arbitrage condition requires

$$b = r - \delta\left(\sqrt{\alpha^2-\beta^2} - \sqrt{\alpha^2-(\beta+1)^2}\right). \tag{5.40}$$

The second condition in (5.38), $\alpha^2 \geq (\beta+1)^2$, guarantees $b \in \mathbb{R}$.

As in the CGMY model, the analyticity condition (5.A) posed by Theorem 5.4 is not satisfied by all realistic parameter choices $q \in \widetilde{\mathcal{Q}}$. We therefore, analogously to the CGMY case, derive a common strip of analyticity. Yet again, empirically, exponential convergence is still observed when a strip of analyticity is shared among all parametrized characteristic functions $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$, of interest.

Strip of analyticity for univariate NIG We derive additional conditions that the MagicFT parameter set $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ for the NIG model needs to satisfy for the existence of a shared strip of analyticity $S_R(\eta)$ of a certain width $R > 0$.

We begin by deriving the domain in $\mathbb{C}$ which a characteristic function $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, of the Normal Inverse Gaussian model in the one-dimensional case, is analytic on. From (5.39) we observe that the characteristic function is analytic, if and only if the mapping

$$z \mapsto \sqrt{\alpha^2 - (\beta + iz)^2} \tag{5.41}$$

is analytic. Let $\widetilde{\eta} \in \mathbb{R}$ and let us denote the complex line $z \in \mathbb{C}$ by

$$z = z_{\widetilde{\eta}}(\xi) = \xi + i\widetilde{\eta}, \qquad \xi \in \mathbb{R} \tag{5.42}$$

and determine the set of possible values for $\widetilde{\eta} \in \mathbb{R}$ such that

$$\xi \mapsto \sqrt{\alpha^2 - (\beta + iz(\xi))^2} \tag{5.43}$$

is analytic on $\mathbb{R}$. The function of (5.43) is analytic, if the radicand of the square root lies in $\mathbb{C}^- = \{z \in \mathbb{C} \,|\, \Re(z) > 0\}$ for all $\xi \in \mathbb{R}$,

$$\Re\left(\alpha^2 - (\beta + iz_{\widetilde{\eta}}(\xi))^2\right) > 0, \qquad \forall \xi \in \mathbb{R}. \tag{5.44}$$

Since

$$\alpha^2 - (\beta + i(\xi + i\widetilde{\eta}))^2 = \alpha^2 - (\beta - \widetilde{\eta})^2 + \xi^2 - 2i\xi(\beta - \widetilde{\eta}),$$

the function in (5.43) is analytic on $\mathbb{R}$ whenever

$$\alpha^2 - (\beta - \widetilde{\eta})^2 > 0. \tag{5.45}$$

Since by definition $\alpha > 0$, the expression on the left hand side of (5.45) equals zero if

$$\widetilde{\eta} = \beta \pm \alpha. \tag{5.46}$$

Equation (5.46) thus yields bounds η^+, η^- given by

$$\begin{aligned} \eta^+ &= \beta + \alpha, \\ \eta^- &= \beta - \alpha \end{aligned} \tag{5.47}$$

that determine the strip $S_R(\eta)$, $\eta = (\eta^+ + \eta^-)/2 = \beta$, $R = \alpha$ that an individually parametrized characteristic function $\varphi_{T,q}$ of the Normal Inverse Gaussian model is analytic on as shown in Figure 5.2.

Understood as a function on $\mathbb{R}$, $\xi \mapsto \varphi_{T,q}(\xi + i\eta)$, the characteristic function of the one-dimensional NIG model, parameterized by $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$ is thus analytic on $\mathbb{R}$, if

$$\beta - \alpha = \eta^- < \eta < \eta^+ = \beta + \alpha. \tag{5.48}$$

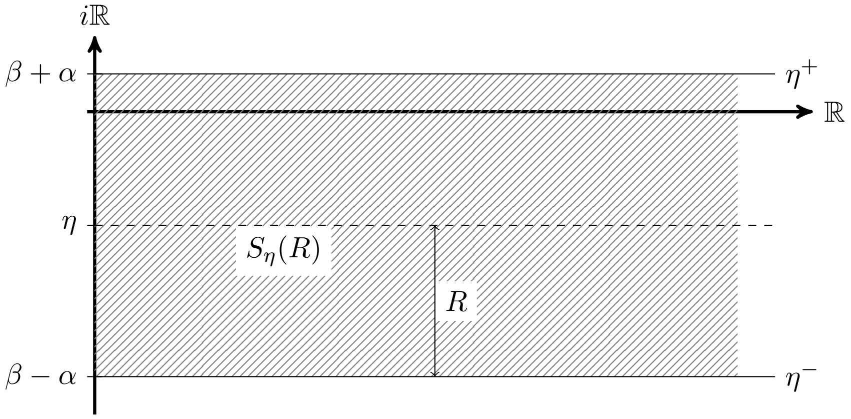

Figure 5.2 The strip of analyticity, $S_R(\eta)$, for a fix parametrization $q \in \widetilde{\mathcal{Q}}$ in the one-dimensional NIG model.

We can now in turn derive additional restrictions on the parametrization of the model that guarantee the existence of a shared strip of analyticity $S_R(\eta)$ with strip width $R > 0$ for all functions $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$ for some $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ incorporating these restrictions. Due to (5.48), the weight parameter η and the parameter set $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ that the strip $S_R(\eta)$ rests on need to satisfy

$$\max_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \beta - \alpha < \eta < \min_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \beta + \alpha. \tag{5.49}$$

We focus on the case of a call option again. There, we already know that $\eta < -1$, so condition (5.49) transforms into

$$\max_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \beta - \alpha < -1 \tag{5.50}$$

and

$$-1 < \min_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \beta + \alpha. \tag{5.51}$$

The distance between the maximum value on the left hand side of (5.50) and -1 determines the width of the strip of analyticity. Condition (5.51) prevents parameter choices that narrow the common strip of analyticity or even prohibit its appearance altogether, as Figure 5.3 shows.

We enforce a certain width R of the resulting strip. Choose $R > 0$. Based on condition (5.50), we require $\mathcal{Q}$ to be chosen such that

$$\beta - \alpha < -1 - 2R \qquad \text{and} \qquad \beta + \alpha > -1, \qquad \forall(\delta,\alpha,\beta,\Lambda) \in \mathcal{Q} \tag{5.52}$$

which is equivalent to

$$\min_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \alpha - \beta > 2R + 1 \qquad \text{and} \qquad \min_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \alpha + \beta > -1. \tag{5.53}$$

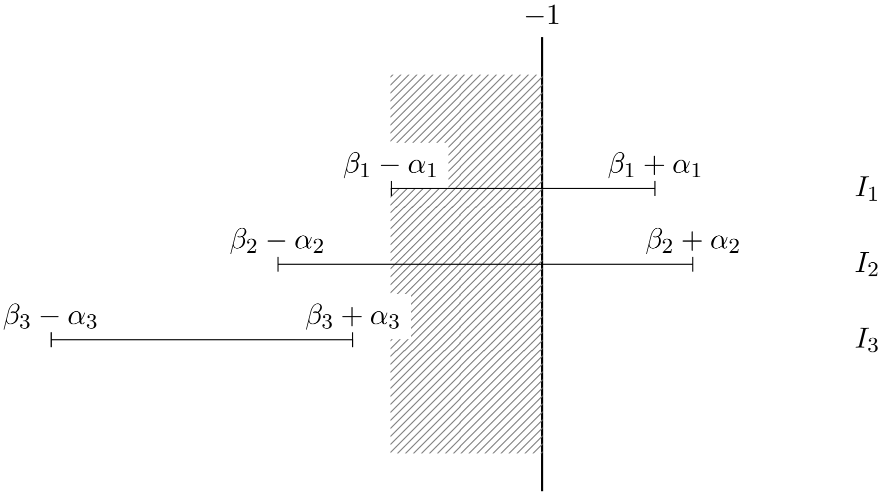

Figure 5.3 Illustration of the necessity of condition (5.51) for the existence of a common strip of analyticity for the NIG model. The hatched area indicates the common strip of analyticity for three parameter sets $q_i = (\delta_i, \alpha_i, \beta_i, 1) \in \widetilde{\mathcal{Q}}$, $i \in \{1, 2, 3\}$, in the call option case, $\eta < -1$. Since q_3 violates condition (5.51), $I_1 \cap I_2 \cap I_3 = \emptyset$ and a common strip of analyticity does not exist.

Using the derived condition (5.53), we define $\mathcal{Q}$ for pricing call options in the the one-dimensional NIG model along the following remark.

Remark 5.9 ($\mathcal{Q}$ for the univariate NIG model)
Let $\underline{\delta} \leq \overline{\delta} \in \mathbb{R}^+$, $\underline{\alpha} \leq \overline{\alpha} \in \mathbb{R}^+$ and $\underline{\beta} \leq \overline{\beta} \in \mathbb{R}$. Let $R > 0$ and define

$$\begin{aligned} \mathcal{Q} = \{(\delta, \alpha, \beta) \in \mathbb{R}^3 \mid\ & \underline{\delta} \leq \delta \leq \overline{\delta}, \quad \underline{\alpha} \leq \alpha \leq \overline{\alpha}, \\ & \underline{\beta} \leq \beta \leq \overline{\beta}, \\ & \alpha^2 > \beta^2, \quad \alpha^2 \geq (\beta+1)^2, \\ & \alpha - \beta > 2R + 1, \quad \alpha + \beta > -1\}. \end{aligned} \tag{5.54}$$

All $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$, share a common strip of analyticity $S_R(\eta)$ with

$$\eta = \frac{\Big(\max\limits_{(\delta,\alpha,\beta)\in\mathcal{Q}} \beta - \alpha\Big) - 1}{2} < -1. \tag{5.55}$$

While $\mathcal{Q}$ of (5.54) in general does not satisfy (5.A) of Theorem 5.4, empirically we still observe exponential convergence of the MagicFT algorithm.

Remark 5.10 (Plausibility constraint on $\mathcal{Q}$ in the univariate NIG model)
Let $q \in \widetilde{\mathcal{Q}}$ of (5.38). The implied variance σ^2_{NIG} of a univariate NIG process at $t = 1$, X^q_1, is given by

$$\sigma^2_{NIG}(\delta, \alpha, \beta) = \frac{\delta\alpha^2}{(\alpha^2 - \beta^2)^{\frac{3}{2}}}, \tag{5.56}$$

confer Prause (1999). To keep volatilities supported by the MagicFT algorithm within reasonable bounds $0 < \sigma_- < \sigma_+$ *add the final restriction*

$$\sigma_-^2 \leq \sigma_{NIG}^2(q) \leq \sigma_+^2, \tag{5.57}$$

for all $q \in \mathcal{Q}$ *of* (5.54).

5.4.2.5 Multivariate Normal Inverse Gaussian model

For the parameter space $\widetilde{\mathcal{Q}}$ of the d variate Normal Inverse Gaussian model consider Section 2.3.5. Again as in the univariate case, Condition (5.A) of Theorem 5.4 is not fullfilled for all $q \in \widetilde{\mathcal{Q}}$.

Strip of analyticity for d variate NIG We extend our analyticity analysis to the d-variate case. Similarly to the derivation in one dimension we identify the strip in $\mathbb{C}^d$ that a d-variate characteristic function $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, of the d-variate NIG model is analytic on.

In a second step, we will derive conditions from our observations that the MagicFT algorithm parameter set $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ must satisfy such that all $\varphi_{T,q}$, $q \in \mathcal{Q}$, share a common strip of analyticity. Empirically this suffices for exponential error decay during the offline phase of the MagicFT algorithm.

The analyticity of the characteristic function as given by (2.47) hinges on the second radicand therein. More precisely, analogously to the computations from (5.43) in one dimension, the analyticity of the characteristic function of the NIG model in d dimensions depends on the analyticity of

$$z \mapsto \sqrt{\alpha^2 - \langle \beta + iz, \Lambda(\beta + iz)\rangle}. \tag{5.58}$$

Let $\widetilde{\eta} \in \mathbb{R}^d$ and define

$$z = z_{\widetilde{\eta}}(\xi) = \xi + i\widetilde{\eta}, \qquad \xi \in \mathbb{R}^d. \tag{5.59}$$

The existence of a strip of analyticity then requires the existence of bounds $\eta^- < \widetilde{\eta} < \eta^+ \in \mathbb{R}^d$ where this inequality is to be understood component-wise such that

$$\begin{aligned} \xi \mapsto & \sqrt{\alpha^2 - \langle \beta + iz_{\widetilde{\eta}}(\xi), \Lambda(\beta + iz_{\widetilde{\eta}}(\xi))\rangle} \\ & = \sqrt{\alpha^2 - \langle \beta + i(\xi + i\widetilde{\eta}), \Lambda(\beta + i(\xi + i\widetilde{\eta}))\rangle} \end{aligned} \tag{5.60}$$

is analytic on $\mathbb{R}^d$ whenever $\eta^- < \widetilde{\eta} < \eta^+$. Analogously to the one dimensional case, analyticity of (5.60) translates into positivity of the real part of the radicand for all $\xi \in \mathbb{R}^d$. The subsequent condition

$$\Re\left(\alpha^2 - \langle \beta + i(\xi + i\widetilde{\eta}), \Lambda(\beta + i(\xi + i\widetilde{\eta}))\rangle\right) > 0, \qquad \forall \xi \in \mathbb{R}^d \tag{5.61}$$

is equivalent to

$$\alpha^2 - \langle \beta - \widetilde{\eta}, \Lambda(\beta - \widetilde{\eta})\rangle > 0. \tag{5.62}$$

Due to the symmetry of $\Lambda \in \mathbb{R}^{d\times d}$, condition (5.62) is equivalent to

$$\alpha^2 - (\langle \beta, \Lambda\beta\rangle - 2\langle \beta, \Lambda\widetilde{\eta}\rangle + \langle \widetilde{\eta}, \Lambda\widetilde{\eta}\rangle) > 0. \tag{5.63}$$

Identifying for a given $q \in \widetilde{\mathcal{Q}}$ all $\widetilde{\eta} \in \mathbb{R}^d$ that satisfy (5.63) is a highly complex problem. We reduce that complexity by assuming $\widetilde{\eta}$ to be given by

$$\widetilde{\eta} = \overline{\eta} \cdot \mathbb{1}^d, \qquad \text{with } \mathbb{1}^d = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^d, \quad \overline{\eta} \in \mathbb{R}, \tag{5.64}$$

thus reducing the degrees of freedom in choosing $\widetilde{\eta}$ to one. This restricts the generality of possible choices for $\widetilde{\eta} \in \mathbb{R}^d$. In return, however, it simplifies the matter considerably. Using (5.64), condition (5.63) turns into identifying all $\overline{\eta} \in \mathbb{R}$ such that

$$\alpha^2 - \Big(\langle \beta, \Lambda\beta\rangle - 2\overline{\eta}\langle \beta, \Lambda\mathbb{1}^d\rangle + \overline{\eta}^2\langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle\Big) > 0. \tag{5.65}$$

The values of $\overline{\eta} \in \mathbb{R}$ setting the left hand side in (5.65) to zero are

$$\overline{\eta}_{1/2} = \frac{\langle \beta, \Lambda\mathbb{1}^d\rangle \pm \sqrt{\langle \beta, \Lambda\mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle\,(\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle}. \tag{5.66}$$

Note that for $d = 1$ and thus $\Lambda = 1$, (5.66) reduces to (5.46). From (5.66) we infer that the strip of analyticity of the d variate $\varphi_{T,q}$, $T \in \mathcal{T}$ and $q \in \widetilde{\mathcal{Q}}$, is spanned by the two extreme values η^-, $\eta^+ \in \mathbb{R}^d$ where

$$\eta^- = \overline{\eta}_1\mathbb{1}^d, \qquad \eta^+ = \overline{\eta}_2\mathbb{1}^d \tag{5.67}$$

with coefficients $\overline{\eta}_1$, $\overline{\eta}_2 \in \mathbb{R}$ given by

$$\overline{\eta}_1 = \frac{\langle \beta, \Lambda\mathbb{1}^d\rangle - \sqrt{\langle \beta, \Lambda\mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle\,(\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle}, \tag{5.68}$$

$$\overline{\eta}_2 = \frac{\langle \beta, \Lambda\mathbb{1}^d\rangle + \sqrt{\langle \beta, \Lambda\mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle\,(\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle}. \tag{5.69}$$

Let us now determine the conditions that $\mathcal{Q}$ has to satisfy and derive the value of $\eta \in \mathbb{R}^d$ such that all d-variate NIG characteristic functions $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$, share a common strip of analyticity $S_R(\eta)$, with width $R > 0$.

We focus on pricing call-type options for instance a call option on the minimum of d assets, such that $\eta < -1$ component-wise. We thus pose onto $\mathcal{Q}$ the conditions that

$$\begin{aligned} \max_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \frac{\langle \beta, \Lambda\mathbb{1}^d\rangle - \sqrt{\langle \beta, \Lambda\mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle\,(\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle} &< -1 - 2R, \\ \min_{(\delta,\alpha,\beta,\Lambda)\in\mathcal{Q}} \frac{\langle \beta, \Lambda\mathbb{1}^d\rangle + \sqrt{\langle \beta, \Lambda\mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle\,(\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda\mathbb{1}^d\rangle} &> -1. \end{aligned} \tag{5.70}$$

Again, this pair of conditions (5.70) comprises its one dimensional equivalent of (5.53) as a special case. Consequently, fixing $R > 0$, d variate characteristic functions of the NIG model parametrized by $q \in \mathcal{Q}$ with $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ satisfying strip conditions (5.70) share a common strip of analyticity $S_R(\eta)$ of width $R > 0$ with $\eta \in \mathbb{R}^d$ given by

$$\eta = \frac{\left(\max\limits_{(\alpha,\beta,\delta,\Lambda)\in\mathcal{Q}} \frac{\langle \beta, \Lambda \mathbb{1}^d\rangle - \sqrt{\langle \beta, \Lambda \mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda \mathbb{1}^d\rangle (\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda \mathbb{1}^d\rangle} \right) - 1}{2} \mathbb{1}^d \in \mathbb{R}^d. \tag{5.71}$$

This derivation is summarized by the following remark.

Remark 5.11 ($\mathcal{Q}$ for the d-dimensional NIG model)
Choose $\underline{\delta} \leq \overline{\delta} \in \mathbb{R}^+$, $\underline{\alpha} \leq \overline{\alpha} \in \mathbb{R}^+$, $\underline{\beta}_i \leq \overline{\beta}_i \in \mathbb{R}$, $i \in \{1, \dots, d\}$, and $\underline{\lambda}_j \leq \overline{\lambda}_j$, $j \in \{1, \dots, d(d+1)/2\}$. Let $R > 0$ and define

$$\begin{aligned} \mathcal{Q} = \Big\{ (\delta, \alpha, \beta, \lambda) \in \mathbb{R}^{2+d+d(d+1)/2} \mid\ & \underline{\delta} \leq \delta \leq \overline{\delta}, \quad \underline{\alpha} \leq \alpha \leq \overline{\alpha}, \\ & \underline{\beta} \leq \beta \leq \overline{\beta} \text{ componentwise}, \quad \underline{\lambda} \leq \lambda \leq \overline{\lambda} \text{ componentwise}, \\ & \det(\Lambda(\lambda)) = 1, \\ & \alpha^2 > \langle \beta, \Lambda(\lambda)\beta\rangle, \\ & \alpha^2 \geq \langle (\beta + e_i), \Lambda(\lambda)(\beta + e_i)\rangle, \ \forall i \in \{1, \dots, d\}, \\ & \frac{\langle \beta, \Lambda \mathbb{1}^d\rangle - \sqrt{\langle \beta, \Lambda \mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda \mathbb{1}^d\rangle (\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda \mathbb{1}^d\rangle} < -1 - 2R, \\ & \frac{\langle \beta, \Lambda \mathbb{1}^d\rangle + \sqrt{\langle \beta, \Lambda \mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda \mathbb{1}^d\rangle (\alpha^2 - \langle \beta, \Lambda\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda \mathbb{1}^d\rangle} > -1 \Big\}. \end{aligned} \tag{5.72}$$

All $\varphi_{T,q}$, $T \in \mathcal{T}$, $q \in \mathcal{Q}$, share a common strip of analyticity $S_R(\eta)$ with

$$\eta = \frac{\left(\max\limits_{(\alpha,\beta,\delta,\lambda)\in\mathcal{Q}} \frac{\langle \beta, \Lambda(\lambda) \mathbb{1}^d\rangle - \sqrt{\langle \beta, \Lambda(\lambda) \mathbb{1}^d\rangle^2 + \langle \mathbb{1}^d, \Lambda(\lambda) \mathbb{1}^d\rangle (\alpha^2 - \langle \beta, \Lambda(\lambda)\beta\rangle)}}{\langle \mathbb{1}^d, \Lambda(\lambda) \mathbb{1}^d\rangle} \right) - 1}{2} \mathbb{1}^d \in \mathbb{R}^d. \tag{5.73}$$

and thus fulfill the empirically required condition for pricing d variate call type options in the NIG model using the MagicFT algorithm.

Remark 5.12 (Plausibility constraint on $\mathcal{Q}$ in the d-variate NIG model)
In Prause (1999), the covariance matrix of a d dimensional NIG process at $t = 1$, X_1^q, $q \in \widetilde{\mathcal{Q}}$, is computed to

$$\Sigma^{NIG} = \delta \left(\alpha^2 - \langle \beta, \Lambda\beta\rangle\right)^{-\frac{1}{2}} \left(\Lambda + \left(\alpha^2 - \langle \beta, \Lambda\beta\rangle\right)^{-1} \Lambda\beta\beta^T\Lambda \right). \tag{5.74}$$

So additionally to the condition satisfied by (5.72) of Remark 5.11, one might use (5.74) to impose additional restrictions regarding Σ^{NIG} to keep supported implicit (co-)variances within realistic bounds.

5.4.2.6 The univariate Heston model

The models considered so far are all Lévy models. We now introduce the model by Heston (1993) that does not fall into this class but is an affine stochastic volatility model, instead. In the univariate Heston model, the asset price process $(S_t^q)_{t\geq 0}$ follows the stochastic differential equation

$$\begin{aligned} \mathrm{d}S_t^{q=(v_0,\kappa,\theta,\sigma,\rho)} &= rS_t\,\mathrm{d}t + \sqrt{v_t^q}S_t\,\mathrm{d}W_t^1, \\ \mathrm{d}v_t^{q=(v_0,\kappa,\theta,\sigma,\rho)} &= \kappa(\theta - v_t)\,\mathrm{d}t + \sigma\sqrt{v_t^q}\,\mathrm{d}W_t^2, \end{aligned} \tag{5.75}$$

with the two Brownian motions W^1, W^2 correlated by $\rho \in [-1,1]$ and with $q \in \widetilde{\mathcal{Q}}$ defined by

$$\widetilde{\mathcal{Q}} = \big\{(v_0,\kappa,\theta,\sigma,\rho) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \times [-1,1], \sigma^2 \leq 2\kappa\theta\big\}. \tag{5.76}$$

The Feller condition

$$\sigma^2 \leq 2\kappa\theta$$

in $\widetilde{\mathcal{Q}}$ of (5.76) ensures an almost surely non-negative volatility process $(v_t)_{t\geq 0}$. With $T \in \mathcal{T}$, $q \in \widetilde{\mathcal{Q}}$, the characteristic function $\varphi_{T,q}$ of the log-asset price process $(\log(S_t/S_0))_{t\geq 0}$ at T is given by

$$\begin{aligned} \varphi_{T,q}(z) =& \exp\left(T\,irz\right)\exp\Bigg(\frac{v_0}{\sigma^2}\frac{(a-c)(1-\exp(-cT))}{1-g\exp(-cT))} \\ &+ \frac{\kappa\theta}{\sigma^2}\left[(a-c)T - 2\log\left(\frac{1-g\exp(-cT)}{1-g}\right)\right]\Bigg), \end{aligned} \tag{5.77}$$

for all $z \in \mathbb{R}$, with supporting functions defined by

$$\begin{aligned} a = a(z) &= \kappa - i\rho\sigma z, \\ c = c(z) &= \sqrt{a(z)^2 - \sigma^2(-zi - z^2)}, \\ g = g(z) &= \frac{a(z)-c(z)}{a(z)+c(z)}, \end{aligned}$$

confer Schoutens et al. (2004). We simply choose $\mathcal{Q} \subset \widetilde{\mathcal{Q}}$ to be a bounded subset of the parameter space.

Remark 5.13 ($\mathcal{Q}$ for the univariate Heston model)
Choose bounds for the initial value of the volatility process, $0 < \underline{v_0} \leq \overline{v_0}$, for its speed of mean reversion, $0 < \underline{\kappa} \leq \overline{\kappa}$, the long-term volatility mean, $0 < \underline{\theta} \leq \overline{\theta}$, and the volatility of the volatility process itself, $0 < \underline{\sigma} \leq \overline{\sigma}$, and a domain for the correlation parameter, $-1 \leq \underline{\rho} \leq \overline{\rho} \leq 1$. Define

$$\begin{aligned} \mathcal{Q} = \big\{(v_0,\kappa,\theta,\sigma,\rho) \mid\ & \underline{v_0} \leq v_0 \leq \overline{v_0},\ \underline{\kappa} \leq \kappa \leq \overline{\kappa}, \\ & \underline{\theta} \leq \theta \leq \overline{\theta},\ \underline{\sigma} \leq \sigma \leq \overline{\sigma}, \underline{\rho} \leq \rho \leq \overline{\rho}, \\ & \sigma^2 \leq 2\kappa\theta\big\}. \end{aligned} \tag{5.78}$$

Despite the fact that $\mathcal{Q}$ defined above in general might not satisfy condition (5.*A*) *of Theorem 5.4, we still observe exponential convergence of the MagicFT algorithm.*

For an analysis of the strip of analyticity in the Heston model, see Levendorskiĭ (2012).

5.5 Numerical experiments

In the previous sections of this chapter we introduced the MagicFT algorithm for option pricing and presented several asset models and option types. We also proved theoretical claims for option pricing with the MagicFT algorithm. In this section we numerically validate these theoretical claims and provide empirical indication that the scope of the algorithm extends to a much wider class of pricing applications than suggested by the theorems earlier.

5.5.1 Implementation

The following description is partially taken from Gaß et al. (2015) where an equivalent implementation was used for the numerical experiments. This implementation of the algorithm in Matlab introduces some simplifications. The continuous parameter space $\mathcal{P}$ is replaced by a discrete parameter cloud randomly sampled. Each magic parameter that the algorithm selects is a member of this discrete set. Consequently, the set $\mathcal{U}$ that the algorithm is trained on is replaced by a discrete set, as well. Additionally, we take Ω to be a discrete set with a finite number of points in each spacial dimension distributed along a logarithmic allocation. Each function $u \in \mathcal{U}$ is then represented by its evaluation on this discrete Ω and is thus replaced by a finite-dimensional vector, numerically. The optimization steps from (5.7)–(5.9) thus reduce to a search on finite sets. When all $h_{p_m^*} \in \mathcal{U}$ for $m = 1, \dots, M$ are identified, they are integrated using Matlab's `quadgk` routine (with an absolute tolerance requirement of 10^{-14}, a relative tolerance requirement of 10^{-12}, a maximum number of intervals of 200000) and linearly assembled to derive the quantities $\int_\Omega \theta_m^M(z)\,\mathrm{d}z$ for $m = 1, \dots, M$.

5.5.2 Empirical convergence

We study the empirical convergence of our implementation of the MagicFT pricing algorithm. A plain vanilla European call option on one asset serves as an example. We investigate the convergence in several models. For each model we set up a pool $\mathcal{U}$ of parametrized Fourier integrands that the algorithm picks from. For each model, the discrete parameter pool is chosen as a uniform sample of magnitude $|\mathcal{P}| = 6000$ from the free parameter ranges enlisted in Table 5.2.

Model	**Fixed parameters**		**Free parameters**	
BS	$K = 1$		$S_0/K \in [0.5,\ 2]$, $\sigma \in [0.1,\ 0.9]$	$T \in [0.1,\ 1.5]$,
Merton	$K = 1$		$S_0/K \in [0.5,\ 2]$, $\sigma \in [0.1,\ 0.7]$, $\beta \in [0.01,\ 0.3]$,	$T \in [0.1,\ 1.5]$, $\alpha \in [-0.2,\ 0.2]$, $\lambda \in [10^{-5},\ 3]$
NIG	$K = 1$		$S_0/K \in [0.5,\ 2]$, $\alpha \in [10^{-5},\ 3]$, $\delta \in [0.2,\ 1]$	$T \in [0.1,\ 1.5]$, $\beta \in [-3,\ 3]$,
CGMY	$K = 1$,	$Y = 1.1$	$S_0/K \in [0.5,\ 2]$, $C \in [10^{-5},\ 1]$, $M \in [0,\ 30]$	$T \in [0.1,\ 1.5]$, $G \in [0,\ 25]$,
Heston	$K = 1$, $\sigma = 0.15$	$\kappa = 2$,	$S_0/K \in [0.5,\ 2]$, $v_0 \in [0.2^2,\ 0.3^2]$, $\rho \in [-1, 1]$	$T \in [0.1,\ 1.5]$, $\theta \in [0.15^2,\ 0.35^2]$,

Table 5.2 In the numerical experiments, we price European call options as an example. Various models have been selected. In the implementation, the Fourier integrands that the algorithm constructs the basis functions q_m with are parametrized according to the intervals above. For each model investigated, $\mathcal{U}$ consists of a pool of $|\mathcal{U}| = 6000$ Fourier integrands.

Additionally, for the NIG and CGMY model, a shared strip of analyticity of width $R = 1/2$ is enforced such that for all investigated models, the dampening factor η could be set to $\eta = -1.5$. Furthermore, all model restrictions stated in Section 5.4.2 are respected. Also, implied variances are kept in the interval $[0.01^2, 0.8^2]$. Each Fourier integrand is evaluated on a discrete $\Omega \subset [0, 75]$ with $|\Omega| = 1750$. The individual $\omega_i \in \Omega$, $i \in \{1, \dots, 1750\}$, are distributed on a log scale.

Figure 5.4 shows the empirically observed error decay during the offline phase of the algorithm for all five considered models in the number of basis functions M. For each model, the quantity $\max_{z\in\Omega}\big|u_M(z) - I_{M-1}(u_M)(z)\big|$ is shown for increasing values of M. The algorithm has been instructed to construct basis functions q_m until an error threshold of 10^{-10} has been reached in step (5.8) or until M has reached the value 50. We observe exponential error decay in all considered models. Recall that Theorem 5.4 predicts this behavior only for the Black&Scholes and the Merton model where analyticity of the associated Fourier integrands is parameter independent. For the other two Lévy models, however, the existence of a shared strip of analyticity results in exponential error decay, as well. In case of the Heston model, the issue of analyticity of the Fourier integrands in $\mathcal{U}$ has not been investigated here. Still, we observe exponential error decay too. The empirical results depicted in Figure 5.4 thus indicate that it might be promising to

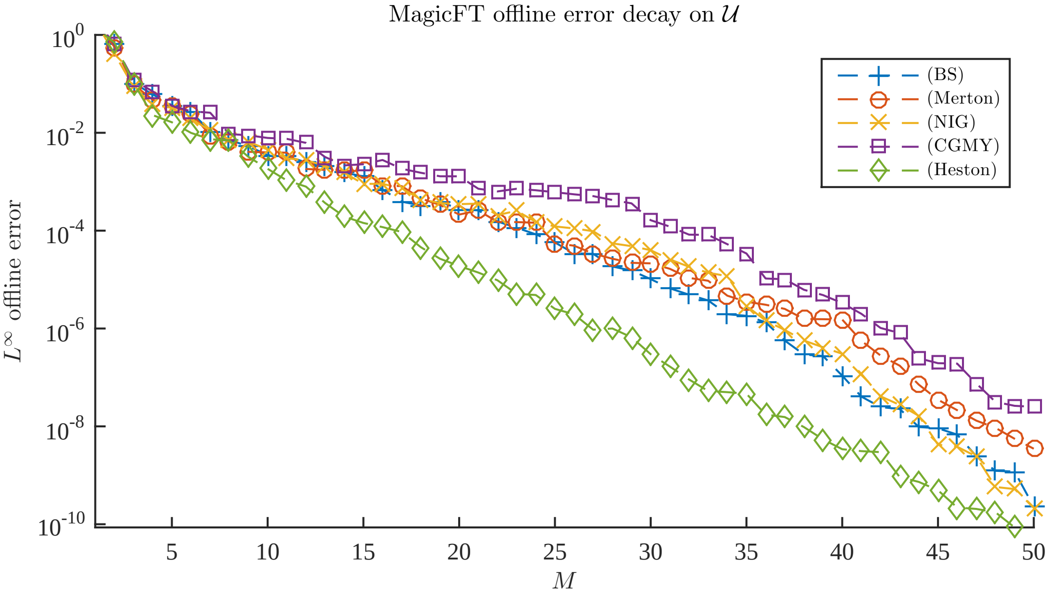

Figure 5.4 A study of the empirical order of convergence of the error in step (5.9) during the offline phase of the MagicFT algorithm. Five different models and European call options are considered. Both the models and the option are parametrized according to Table 5.2. The convergence result is theoretically backed by Theorem 5.4 for the Black&Scholes and the Merton model. A shared strip of analyticity of the respective Fourier integrands of width $R = 1/2$ has been enforced for the NIG and CGMY model.

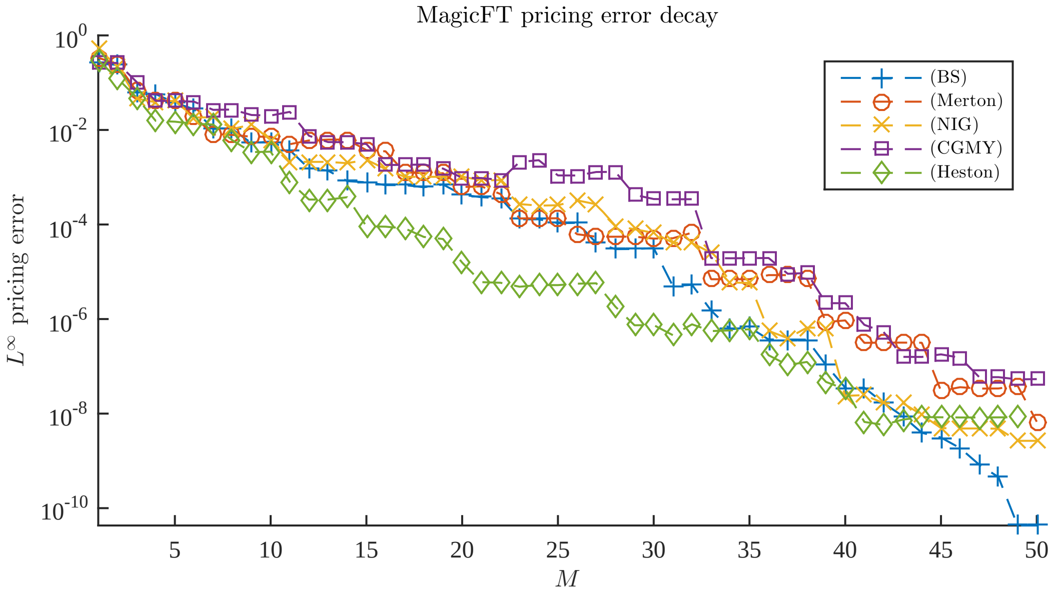

Figure 5.5 Pricing error decay study on 1000 out of sample parameter constellations for different models. In each model, for increasing values of M, the L^∞ error over the randomly drawn parameter sets is evaluated. The parameter sets have been drawn from the intervals given by Table 5.2.

investigate a theoretical result providing exponential error decay beyond the scope of Theorem 5.4.

5.5.3 Out of sample pricing study

In the previous paragraph we studied empirical convergence during the offline phase of the algorithm. More precisely, we investigated for several models how accurately all Fourier integrands in the given pool $\mathcal{U}$ could be approximated on their integration interval Ω by the M selected integrands or rather by the basis functions q_m, $m = 1, \dots, M$, constructed thereof. Now we analyze, how the observed accuracy on the level of in sample integrands translates to the accuracy in an out of sample call option pricing exercise.

To this extent we randomly draw 1000 parameter constellations for each model according to the same rules as in the offline phase. For each such sample we compute the respective Fourier price by numerical integration on $[0, 75]$ thus containing the discrete Ω that the MagicFT algorithm has been trained on. We integrate using Matlab's `quadgk` with absolute tolerance of 10^{-12} and $200,000$ integration intervals. Additionally, in each model we approximate all prices associated with the randomly drawn parameters for increasing values of M, evaluate the L^∞ error and study its decay in M as depicted in Figure 5.5.

We observe exponential rates for all considered models. Curiously, the error decay attains plateau-like shapes, especially for higher values of M. We explain this decay structure by assuming that each plateau is associated with a certain single parameter realization from the random sample that dominates the L^∞ error until a magic parameter close to it or rather the respective basis function contributes to the approximation of the belonging price. Due to such outliers, the order in which the offline phase errors were decaying in Figure 5.4 has changed.

In Figure 5.6, we depict evaluations of the absolute as well as the relative pricing errors for all out of sample parameter sets, individually. Here, relative errors have been computed only for prices larger than 10^{-3} to exclude numerical noise. In each model, M is set to its final value assigned during the respective model's offline phase and can be read off from Figure 5.4.

Pricing accuracy in this out of sample pricing exercise reaches very satisfactory levels albeit the achieved accuracies vary between the considered models. For all models, average absolute pricing accuracy reaches levels between $\text{avg}_{\text{min}} \approx 10^{-12}$ in the Black&Scholes model and $\text{avg}_{\text{max}} \approx 10^{-10}$ for the CGMY model. Average relative pricing accuracy ranges between 10^{-11} and 10^{-9}. We observe individual outliers for all models. The ten worst (largest) absolute errors together with the ten best (smallest) absolute errors in each model are further addressed in the next section.

5.5.4 Individual case studies

We take a closer look into the numerical results for each model individually. We are interested in the distribution of the magic points as well as the distribution of the magic parameters that the algorithm picked. Figure 5.7 shows some basis functions q_m from the Black&Scholes model and the Merton model that the algorithm constructed evaluated over the domain Ω.

Figure 5.8 displays basis functions for the approximation of prices in the NIG and the CGMY model and finally Figure 5.9 depicts basis functions for approximation of prices in the Heston model.

Intersections of basis functions q_m with the Ω axis correspond to the location of magic points. An accumulation of such magic points at the origin reveals that the Fourier integrands of the respective model possess the largest variation there. Differences between the five models for example with respect to the distribution of magic points reflect the different structure of the underlying Fourier integrands that all seem to possess a certain model specific nature.

After this assessment of the magic points let us now analyze the distribution of magic parameters in each model.

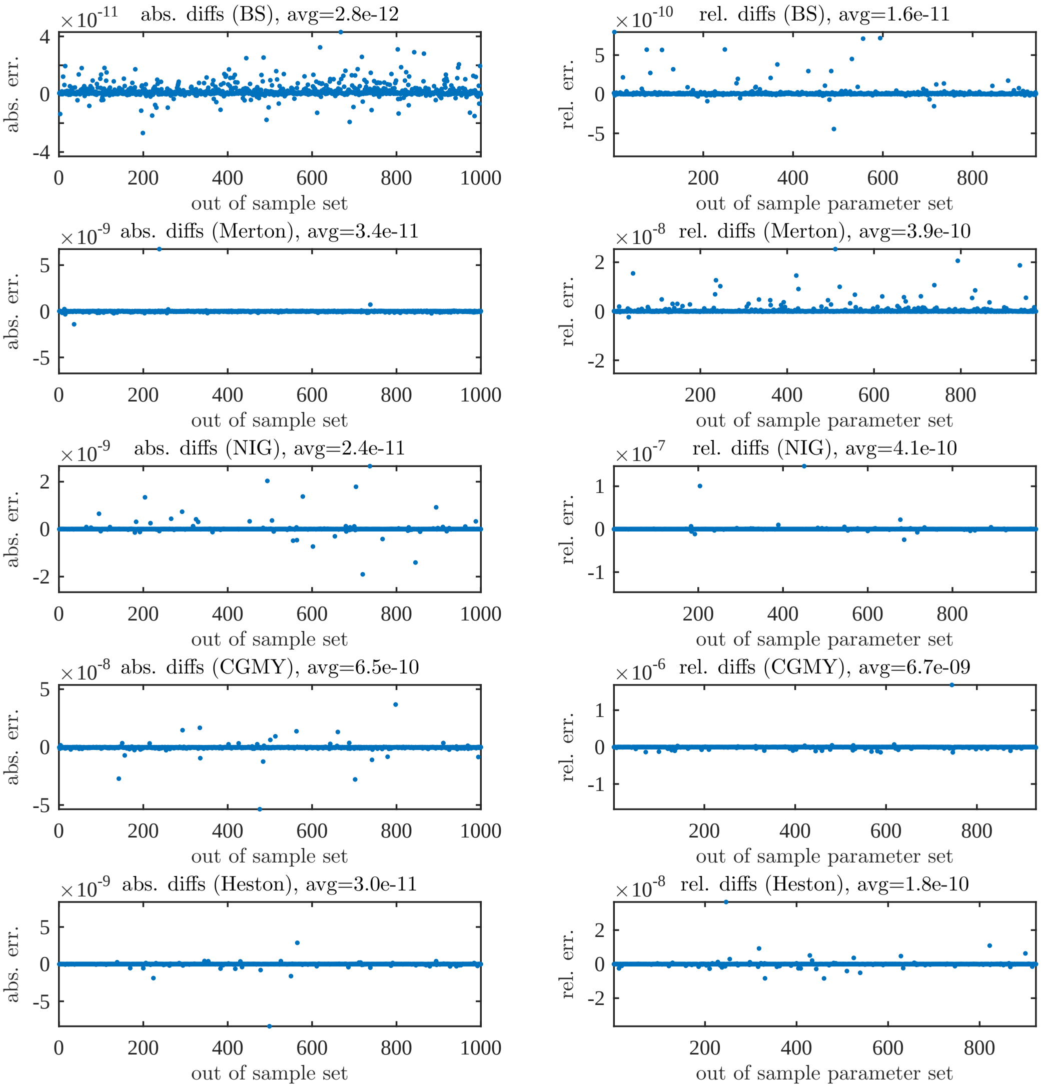

Figure 5.6 Results of the out of sample pricing exercise. For each of the five considered models, 1000 parameter sets have been drawn from the intervals given by Table 5.2. For each set, the Fourier price as well as the MagicFT price have been calculated. On the left column, all absolute errors are depicted. On the right, the relative errors are shown.

Black&Scholes During the offline phase of the algorithm for the Black&Scholes model only the option strike K has been fixed, $K = 1$. The model parameter σ as well as the

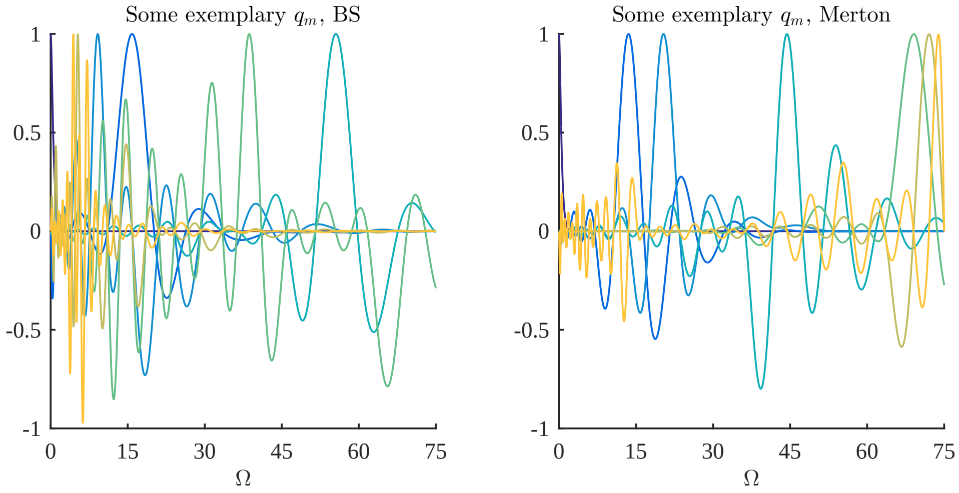

Figure 5.7 Some exemplary q_m basis functions in the Black&Scholes model and the Merton model. Intersections with the Ω-axis mark the location of magic points. In both cases, magic points accumulate close to the origin.

two other parameters S_0/K and maturity T were allowed to vary within the bounds assigned by Table 5.2. In the Black&Scholes case, the individual parameter intervals tensorize meaning that any combination of parameter values respecting the individual bounds can be picked by the algorithm. As Figure 5.10 demonstrates for the magic parameter choices for S_0/K and T, however, rather extreme constellations have been selected. Figure 5.11 provides a complete overview over all parameter combinations selected in the offline phase of the algorithm for the Black&Scholes model. With the exception of T and σ combinations, rather extreme parameter pairs have been selected. This special behavior is not surprising, since T and σ always appear together as a product in the Fourier integrands of the Black&Scholes model, compare the definition of the characteristic function in the Black&Scholes model in (2.34). The even distribution of the (T, σ) parameter pairs thus reflects the even distribution of all individual parameters over their domain, observable on the elements on the main diagonal of the figure. Additionally, the paper illustrates parameter areas that are particularly challenging for the MagicFT algorithm to approximate together with those that the algorithm is well prepared for. Counterintuitively at first, orange parameter sets that resulted in the largest absolute errors are often to be found in close proximity to selected magic parameters. And green parameter sets the associated prices of which could be best approximated by the algorithm lie in fallow fields. On second thought, however, this result corresponds to the rule according to which magic parameters have been selected during the offline phase. Parameter areas densely populated by magic parameters are precisely those that the algorithm is facing the largest challenges in. Empty areas by contrast can already be sufficiently approximated be the previously selected magic parameters.

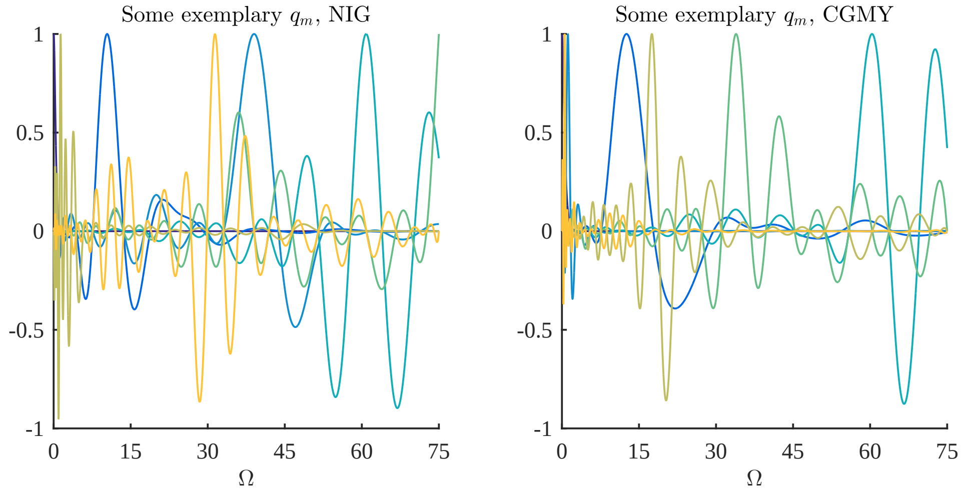

Figure 5.8 Some exemplary q_m basis functions in the NIG and CGMY model. Intersections with the Ω-axis mark the location of magic points.

Merton We perform the same analysis for the Merton model. Apart from the European call option strike parameter $K = 1$, all other parameters were allowed to vary within the intervals of Table 5.2. Figure 5.12 displays the distribution of the magic parameters together with the randomly drawn parameter constellations of the out of sample pricing accuracy study that resulted in the ten largest absolute pricing errors. Again, orange parameter constellations in Figure 5.12 which indicate large pricing inaccuracies, seem to particularly occur in areas densely populated by magic parameters – areas which we would thus expect a rather high accuracy in pricing from. Yet, again we see from the definition of u_M in (5.8), during the offline phase, magic parameters are chosen precisely where the approximation of the algorithm is worst. An accumulation of magic parameters at one location indicates rather diverse shapes of the Fourier integrands parametrized in this very location. In other words, in subsets of the parameter space where magic parameters accumulate, pricing is especially challenging for the MagicFT algorithm. This interpretation is confirmed by the location of the green parameter sets marking those constellations that the algorithm approximated best.

NIG, CGMY & Heston Figure 5.13 depicts the parameter clouds for the free parameters in our parametrization of the NIG model. Note in the (α, β) combinations the effect of model restriction $\alpha^2 > \beta^2$. Figure 5.14 illustrates the magic parameter distribution for the CGMY model and Figure 5.15 finally visualizes the magic parameter choices for the Heston model together with those out of sample draws that lead to the ten worst and the ten best pricing results.

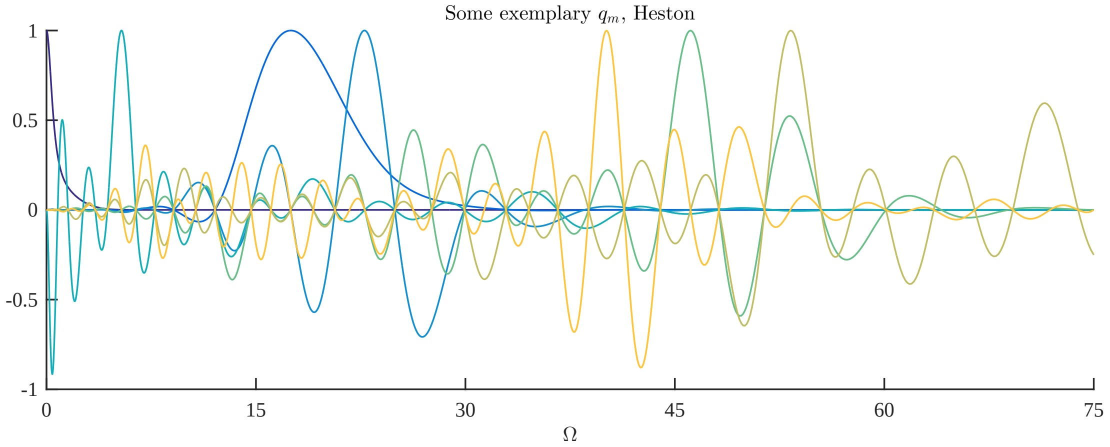

Figure 5.9 Some exemplary q_m basis functions in the Heston model. In contrast to the four Lévy models displayed in Figure 5.7 and Figure 5.8, the magic points are rather equally spread over the whole domain Ω.

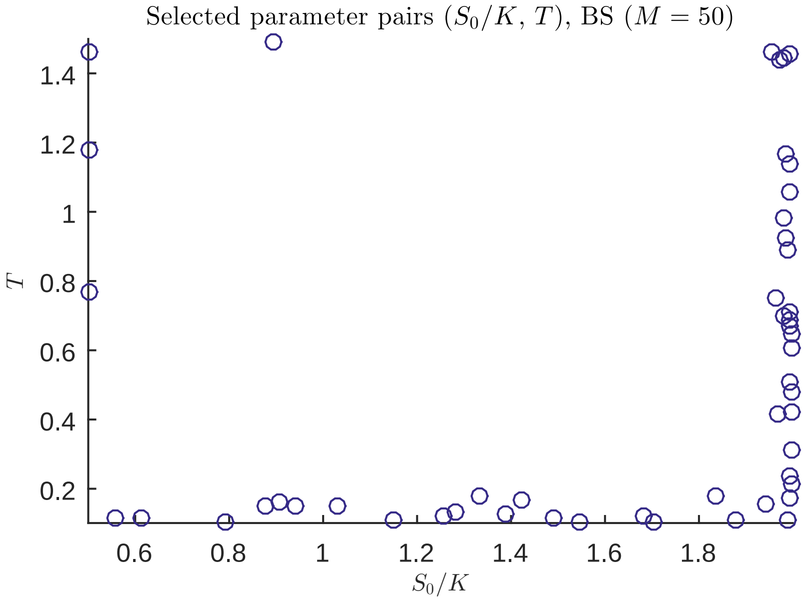

Figure 5.10 Parameter pairs $(S_0/K, T)$ selected by the MagicFT algorithm in the offline phase of the Black&Scholes model.

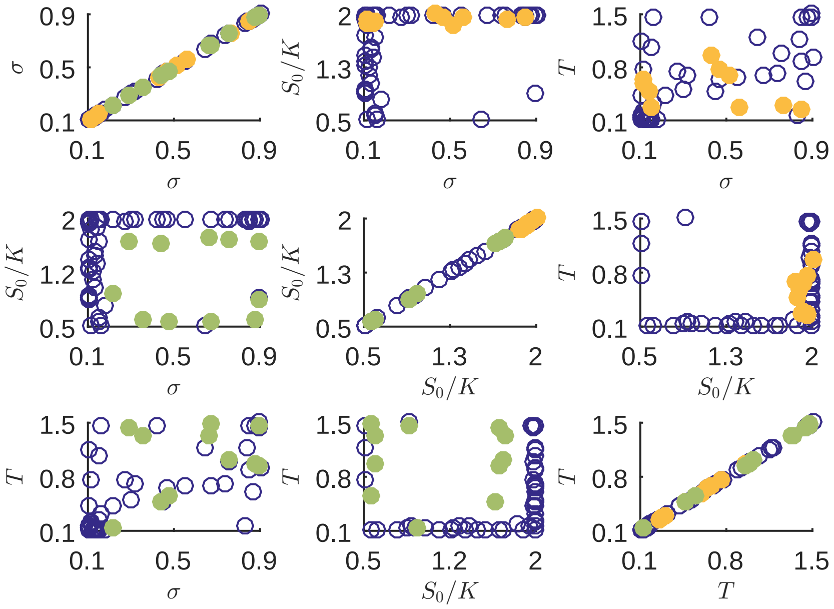

Figure 5.11 All magic parameters selected during the offline phase of the algorithm for the Black&Scholes model (empty blue circles). The filled orange circles denote the ten parameter constellations that resulted in the maximal absolute pricing errors during the out of sample pricing exercise. In contrast, the filled green circles mark the location of the ten parameter constellations that yielded the best approximate pricing results.

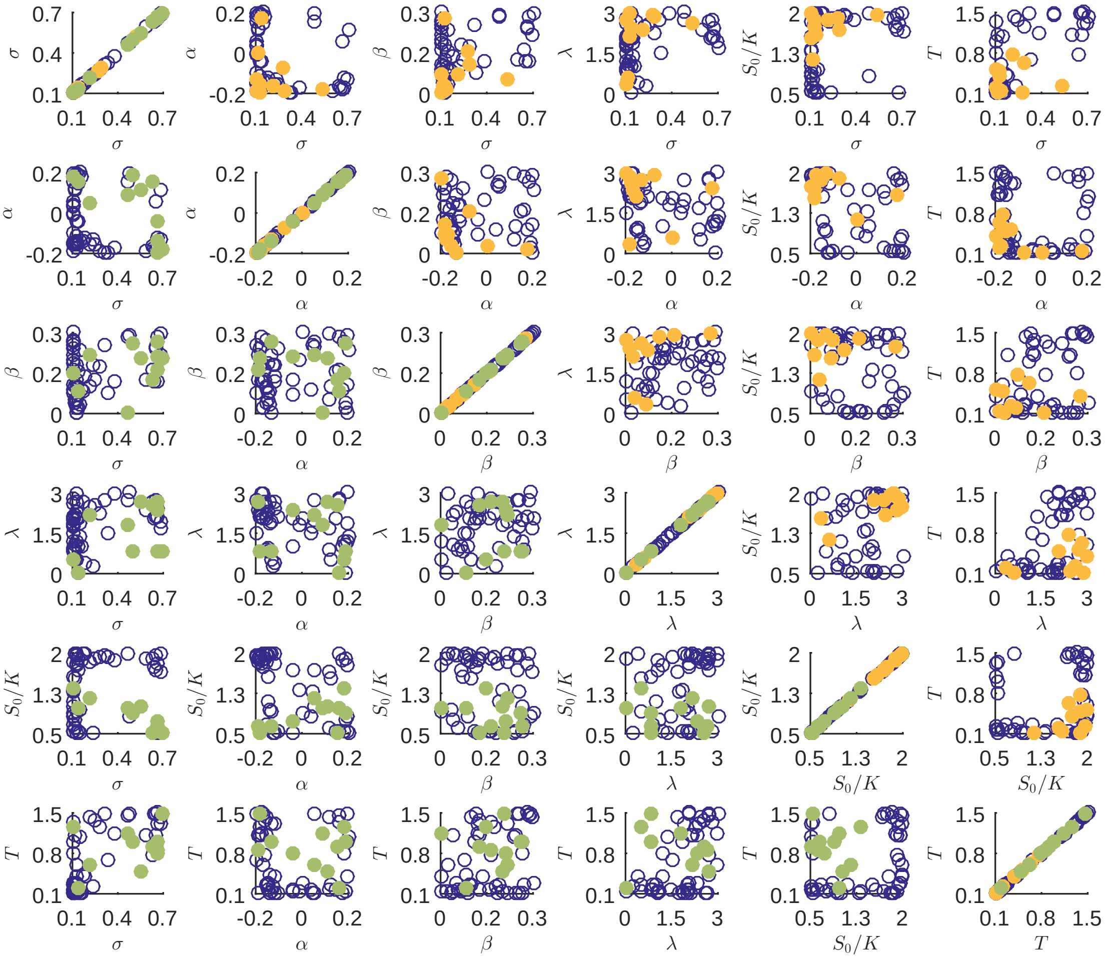

Figure 5.12 The distribution of magic parameters in the Merton model (blue empty circles) together with those randomly drawn parameter samples that resulted in the ten largest absolute pricing errors in the out of sample pricing exercise (filled orange circles). The filled green circles, by comparison, mark the location of parameter constellations that the algorithm could handle best.

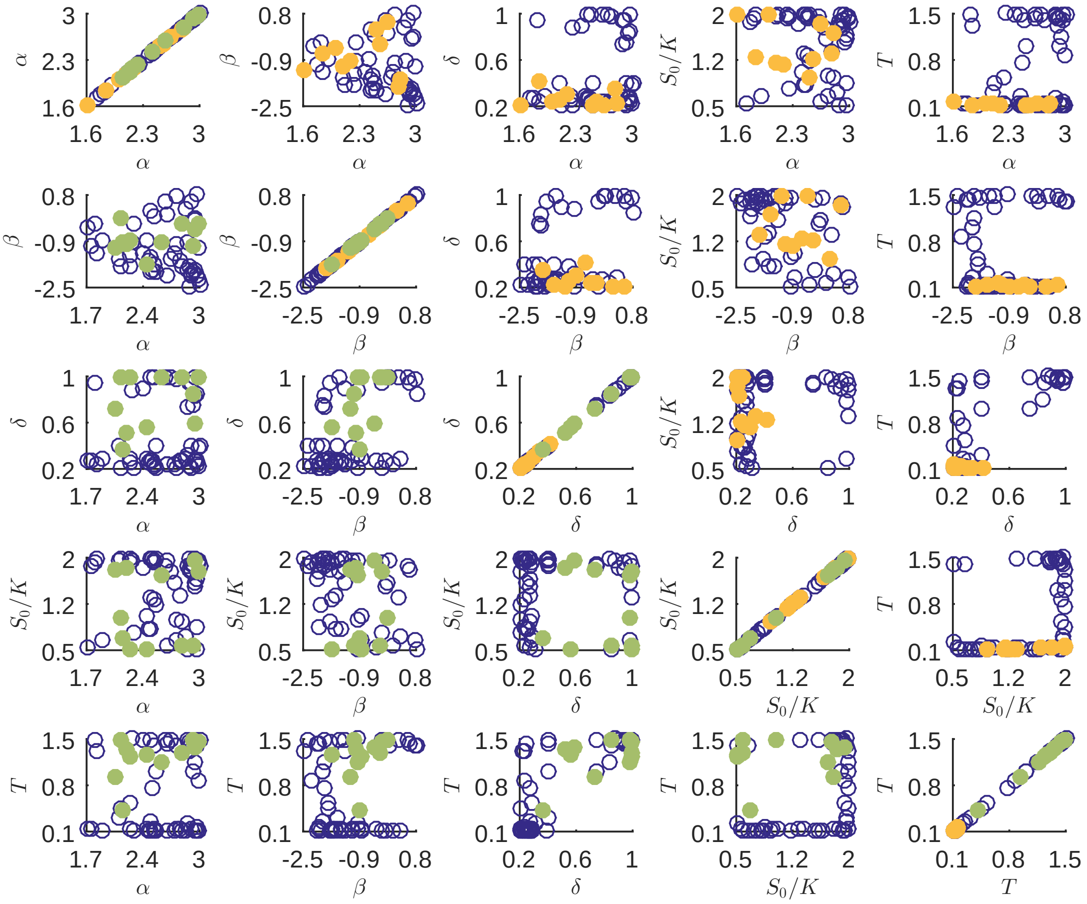

Figure 5.13 Distribution of NIG magic parameters (blue empty circles) and randomly drawn parameter constellations resulting in the ten largest (orange) and the ten smallest (green) absolute pricing errors during the out of sample pricing study.

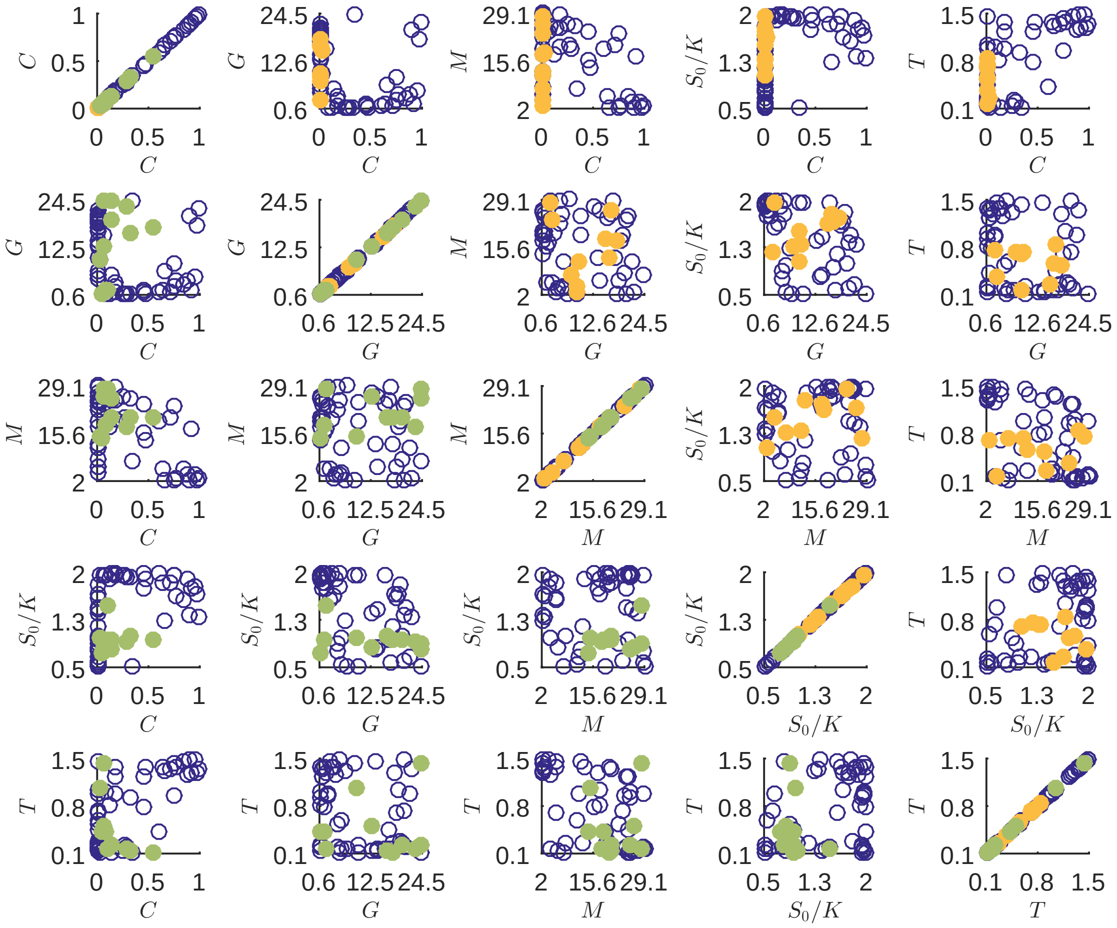

Figure 5.14 Magic parameters (blue empty circles) and randomly drawn parameter constellations resulting in the ten largest absolute pricing errors marked in orange and the ten smallest absolute pricing errors colored in green during the out of sample pricing study in the CGMY model case.

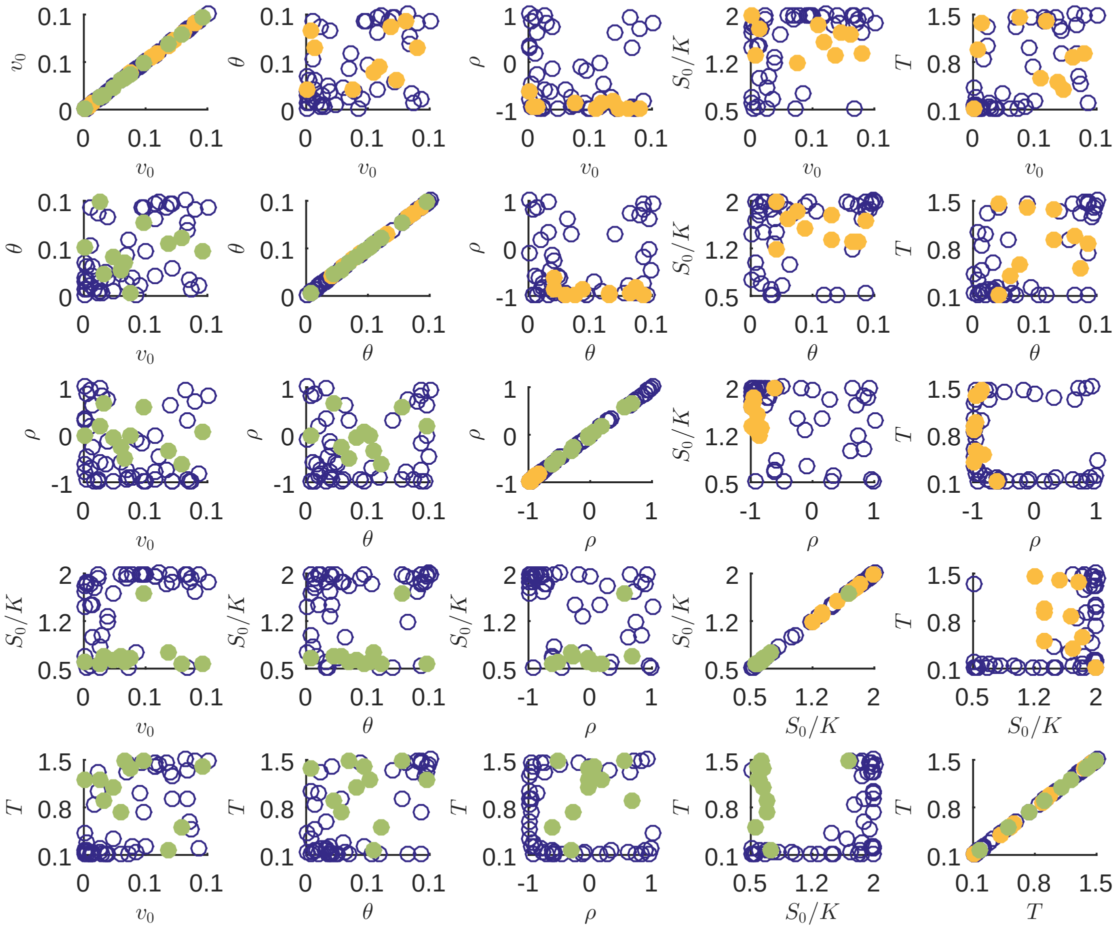

Figure 5.15 An overview over the distribution of magic parameters (blue empty circles) and the randomly drawn parameter combinations resulting in the ten largest absolute pricing errors (orange) in the Heston model. Equivalently, those random parameter constellations the prices of which could be best approximated are depicted (colored in green). Note the especially extreme combinations of selected T and ρ values.

5.5.5 Comparison with Chebyshev interpolation

Finally, we compare the numerical performance of the MagicFT method to a different pricing method using another interpolation. In Chapter 4, we presented a pricing method based on the Chebyshev polynomials. There, prices are interpreted as functions of a set p of model and option parameters and approximated by a linear combination of Chebyshev coefficients c_j, $j \in J$, independent of p and associated Chebyshev polynomials T_j, $j \in J$, depending on p,

$$Price^p \approx \sum_{j\in J} c_j T_j(p) \tag{5.79}$$

for a certain index set J. In the univariate case where $J = \{0, \dots, N\}$ for some $N \in \mathbb{N}$, the Chebyshev polynomials T_j, $j \in J$, are given by

$$T_j(x) = \cos(j \arccos(x)), \qquad x \in [-1, 1]. \tag{5.80}$$

Consequently, they are not adapted to the problem that the approximation method is applied to. The coefficients c_j, $j \in J$, are defined by a sum of precomputed prices $Price^{p_k}$ for certain parameter sets p_k, $k \in \{0, \dots, N\}$, in the parameter space.

Both algorithms thus resemble each other in the sense that they consist of an offline phase where prices for certain parameter constellations are precomputed and stored, and an online phase during which these precomputed quantities are added with weights depending on the parameter set of interest. Yet, while the MagicFT algorithm decides for itself which parameters to pick, the Chebyshev method fixes them in advance. Additionally, while the MagicFT algorithm iteratively constructs its basis functions, the Chebyshev method relies on the given Chebyshev polynomials of (5.80). And finally, while the MagicFT algorithm approximates Fourier integrands, the Chebyshev method approximates prices directly.

With these given similarities and differences in mind we compare the Chebyshev approximation method to the MagicFT algorithm in three aspects:

i) How are the parameters that are selected during the offline phase distributed in the parameter space in both methods?

ii) How do the basis functions of both algorithms compare?

iii) How accurately are prices approximated by both approaches in a comparable setting?

We study these questions in an elementary setting by applying both algorithms to the pricing of European call options on one asset in the Black&Scholes model with the Black&Scholes volatility $\sigma > 0$ being the only free parameter. More precisely, we fix a maturity $T > 0$, a strike value $K > 0$ and the current value of the underlying stock $S_0 > 0$, disregard interest rates, $r = 0$, and interpret call option prices in the Black&Scholes model as a function of $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, $0 < \sigma_{\min} < \sigma_{\max} < 1$. Since the (univariate)

Chebyshev method is defined for normed parameter intervals, $p \in [-1, 1]$, we introduce the transformation $\tau : [-1, 1] \to [\sigma_{\min}, \sigma_{\max}]$,

$$\tau(\sigma) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \left(\frac{1}{2} + \frac{1}{2}\sigma \right), \tag{5.81}$$

and approximate the Black&Scholes price $[\sigma_{\min}, \sigma_{\max}] \ni \sigma \mapsto \mathit{Price}^{K,T,\sigma}$ by the Chebyshev method using

$$[-1, 1] \ni p \mapsto I_N^{\text{Cheby}}(\mathit{Price}^{K,T,\sigma=\tau(\cdot)})(p) \tag{5.82}$$

wherein I_N^{Cheby} is the Chebyshev interpolator of (2.2) in Gaß et al. (2016), respectively (4.2) in Chapter 4, and by the MagicFT algorithm using

$$[\sigma_{\min}, \sigma_{\max}] \ni \sigma \mapsto \frac{1}{2\pi} \sum_{m=1}^{M} \widehat{f_K}(-z_m^*) \varphi_{T,q=\sigma}(z_m^*) \int_\Omega \theta_m^M(z)\,\mathrm{d}z. \tag{5.83}$$

To keep the two approximations roughly comparable, we provide both methods with a similar number of Chebyshev polynomials or magic points by choosing $N = M$ throughout this study.

In defining the parameter space we choose $\sigma_{\min} = 0.1$ and $\sigma_{\max} = 0.7$ and fix today's value of the underlying at $S_0 = 2.2$. The call option strike $K = 2$ and the time to maturity $T = 1$ are kept constant.
We run the offline phase of the MagicFT algorithm until $M = 10$ basis functions q_m out of a pool $\mathcal{U}$ with $|\mathcal{U}| = 6000$ are identified. The pool $\mathcal{U}$ is parameterized by a randomly drawn sample of uniformly distributed σ values, $\sigma_{\min} \leq \sigma \leq \sigma_{\max}$. Associated with these basis functions are 10 pairs of magic points and magic parameters (z_k^*, p_k^*), $1 \leq k \leq 10$.

Equivalently, we prepare the Chebyshev method setting $N = M = 10$ and run the precomputational offline phase deriving the coefficients c_j, $0 \leq j \leq 10$, by computing European Black&Scholes call option prices at prespecified and application independent Chebyshev nodes $p_k \in [\tau^{-1}(\sigma_{\min}), \tau^{-1}(\sigma_{\max})]$, $0 \leq k \leq N$. In both offline phases, all precomputed prices are derived using numerical integration of the respective Black&Scholes Fourier integrand. Consequently, the influence of numerical integration is the same for both methods. Figure 5.16 depicts the set of magic parameters chosen by the MagicFT algorithm and the set of Chebyshev nodes. Associated with these parameter sets are the two sets of basis functions. Again, the set of interpolands q_m, $1 \leq m \leq M$, constructed by the MagicFT algorithm is model adapted. Each q_m consists of a linear combination of Fourier integrands parametrized by the associated magic parameter. The set of Chebyshev polynomials on the other hand is application independently defined by (5.80).

Figure 5.17 shows the first five MagicFT basis functions $q_1, \ldots, q_5$ as well as the first five Chebyshev polynomials $T_1, \ldots, T_5$. The explanatory power of this comparison is of course limited. While linear combinations of the MagicFT basis functions approximate *Fourier integrands* on their integration domain Ω, linear combinations of the Chebyshev polynomials approximate *prices* on the (normed) parameter domain $[\tau^{-1}(\sigma_{\min}), \tau^{-1}(\sigma_{\max})]$.

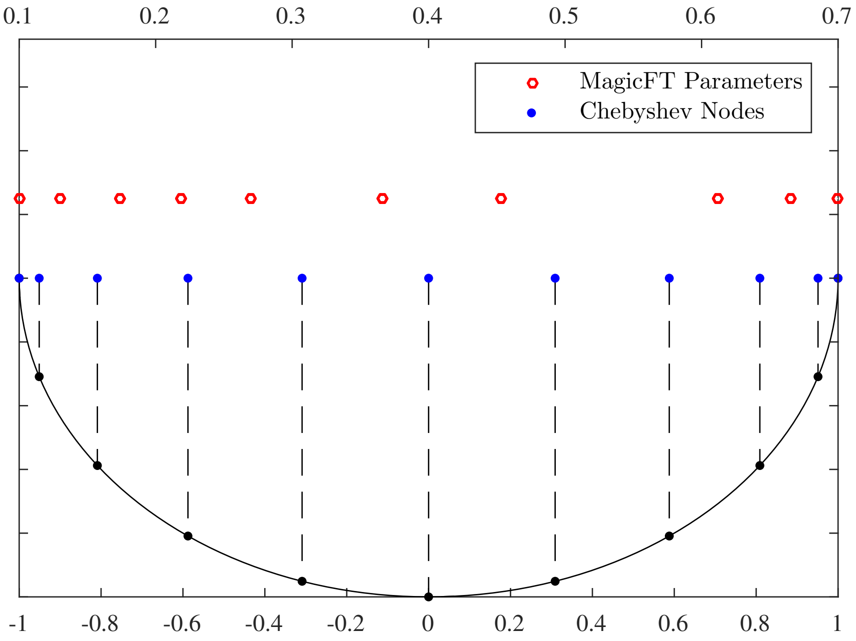

Figure 5.16 Comparison between the distribution of the $M = 10$ magic parameters $0.1 = \sigma_{\min} \leq p_k^* \leq \sigma_{\max} = 0.7$, $1 \leq k \leq M$, and the $N+1$ Chebyshev nodes $-1 \leq p_k \leq 1$, $0 \leq k \leq N$, where $N = M$. While the magic parameters have been selected by the MagicFT algorithm, the Chebyshev nodes are given by a model independent construction using a set of construction points equidistantly spaced on the semicircle as indicated in the figure. Interestingly, both sets are similarly distributed over the (normed) parameter space.

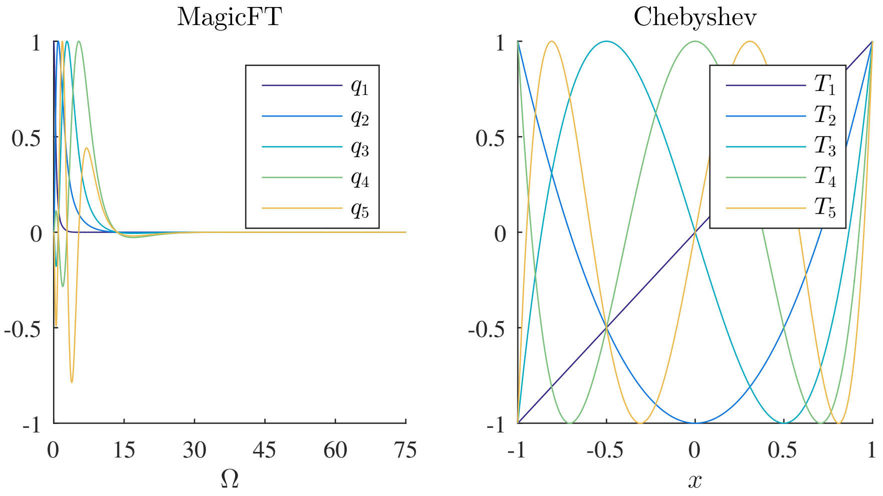

Figure 5.17 Left: The first five basis functions $q_1, \ldots, q_5$ as constructed by the MagicFT algorithm in the Black&Scholes call option setting of this study. Each q_i consist of a linear combination of Fourier integrands evaluated over $\Omega = [0, 75]$. The intersections with the Ω axis mark the location of magic points z_m^*. The magic points accumulate close to the origin indicating the strongest variations among all Fourier integrands in the set $\mathcal{U}$ there. Right: The first five Chebyshev polynomials $T_1, \ldots, T_5$ evaluated over their domain $[-1, 1]$.

Finally, we apply both methods to compute prices for a large discrete set of volatility values σ_k, $k \in \{0, \ldots, 1500\}$, on an equidistant grid,

$$\sigma_k = \sigma_{\min} + \frac{k}{1500} \left(\sigma_{\max} - \sigma_{\min}\right), \qquad k \in \{0, \ldots, 1500\}, \tag{5.84}$$

spanning the whole parameter space. Figure 5.18 depicts the results of the pricing accuracy study. Prices of both the MagicFT algorithm and the Chebyshev method are compared to Matlab's `blsprice` routine. The accuracy of both methods is similar. Intersections of both error curves with the σ axis or, put differently, points of perfect pricing results identify the position of magic parameters or Chebyshev nodes, respectively.

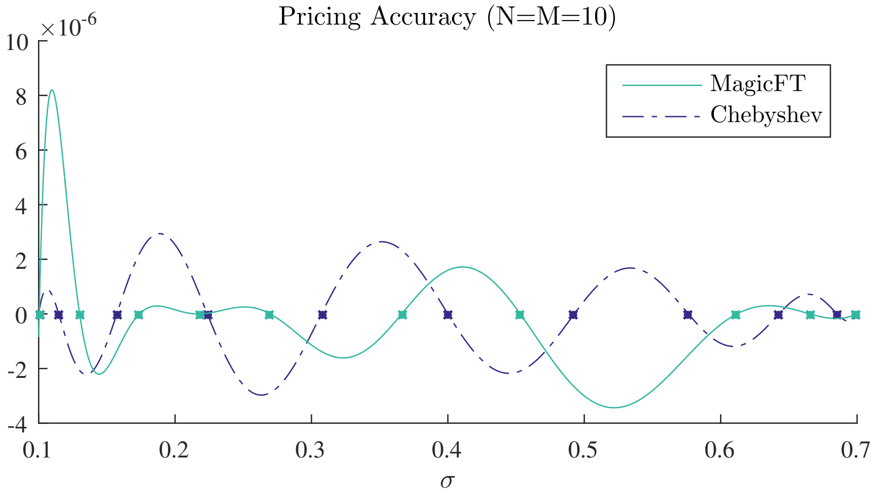

Figure 5.18 Pricing errors in both methods for 1501 volatility values. The magnitude of the error is the same for both methods and of order 10^{-6}. Intersections of the solid curve with the σ axis mark the location of a magic parameter. Similarly, σ values at the intersection of the dashed curve with the σ axis are associated with the location of the respective Chebyshev node in the normed parameter space. Both magic parameters and (appropriately scaled) Chebyshev nodes have been additionally highlighted.

5.6 A review of the interpolation operator

We take a closer look at the interpolation operator I_M that we used for the interpolation of Fourier integrands above. During the offline phase it iteratively interpolates basis function candidates to identify the one which is worst represented by the basis functions that have already been constructed. During the online phase it allows the approximate evaluation of functions on the domain Ω.

To meet both expectations satisfactorily, the interpolation operator must be evaluable quickly for each admissible function and for arbitrary points on the domain. At first sight, the linear dependence of the operator on the individual basis functions grants this feature immediately. On second thought, however, we recognize that each basis function itself depends recursively on the basis functions previously generated. The interpolation operator I_M inherits this structure which in general prevents fast evaluation calls especially for large values of M. In this section we illustrate the problem sketched above in more detail and comment on the fact why this topic was not an issue in our implementation above. We explain why nevertheless it becomes an issue when $\dim(\Omega) > 1$. Therefore we resolve the recursive dependence of the operator I_M on the basis functions in the final section of this chapter.

For convenience, Algorithm 2 restates the Empirical Interpolation method of Maday et al. (2009) that we already described in detail in Section 5.1 above.

Algorithm 2 Empirical Interpolation (EI)

```
Let $\Omega \subset \mathbb{R}^d$ be a bounded domain
Let $\mathcal{P}$ be some parameter space
Let further $\mathcal{U}$ be a set of parametric functions
    $\mathcal{U} = \{u(p) : \Omega \to \mathbb{R},\ p \in \mathcal{P}\}$
function INTERPOLATION OPERATOR $I_M(u)(\xi)$
    return $I_M(u)(\xi) = \sum_{j=1}^{M} u(\xi_j)\theta_j^M(\xi)$
    with
        $\theta_j^M(\xi) = \sum_{i=1}^{M} \left(B^M\right)_{ij}^{-1} q_i(\xi), \qquad B_{ij}^M = q_j(\xi_i),$
    where the set of *magic points* $T_M = \{\xi_1, \dots, \xi_M\} \subset \overline{\Omega}$ and the set of *basis
    functions* $\{q_1, \dots, q_M\}$ are recursively defined by
        $u_1 = \arg\max_{u \in \mathcal{U}} \|u\|_{L^\infty}$
        $\xi_1 = \arg\max_{\xi \in \overline{\Omega}} |u_1(\xi)|$
        $q_1(\cdot) = \frac{u_1(\cdot)}{u_1(\xi_1)}$
    and for $M > 1$ by
        $u_M = \arg\max_{u \in \mathcal{U}} \|u - I_{M-1}(u)\|_{L^\infty}$
        $\xi_M = \arg\max_{\xi \in \overline{\Omega}} |u_M(\xi) - I_{M-1}(u_M)(\xi)|$
        $q_M(\cdot) = \frac{u_M(\cdot) - I_{M-1}(u_M)(\cdot)}{u_M(\xi_M) - I_{M-1}(u_M)(\xi_M)}$
```

Algorithm 2 is stated continuously. In numerical applications, however, it is implemented discretely, instead. For that matter, several simplifications are introduced. We state these simplifications in the univariate case, $d = 1$. Then, instead of a continuous domain Ω we consider a discrete subset $\Omega_{\text{discr.}} = \{\omega_1, \dots, \omega_N\} \subset \Omega$ and instead of the continuous parameter set $\mathcal{P}$ we introduce a discrete subset $\mathcal{P}_{\text{discr}} = \{p_1, \dots, p_K\} \subset \mathcal{P}$. As a consequence of these changes, $\mathcal{U}$ is replaced by $\mathcal{U}_{\text{discr}} = \{\vec{u}_i = (u(p_i)(\omega_1), \dots, u(p_i)(\omega_N)) \mid p_i \in \mathcal{P}_{\text{discr}},\ i \in \{1, \dots, K\}\} \subset \mathbb{R}^N$ with $|\mathcal{U}_{\text{discr}}| = |\mathcal{P}_{\text{discr}}| = K$. Consequently, all resources of the algorithm become discrete and finite, as Algorithm 3 shows.

The optimization steps in line 14 and line 15 of Algorithm 3 fully consist of finding maxima on discrete sets and thus do not rely on special optimization routines. For $N, K \in \mathbb{N}$, the set of basis functions $\vec{q}_i$, $i \in \{1, \dots, M\}$, can be constructed by iteratively considering all K basis function candidate vectors $\vec{u}_i \in \mathcal{U}^{\text{discr}}$, $i \in \{1, \dots, K\}$, and all N magic point candidates $\omega_i \in \Omega^{\text{discr}}$, $i \in \{1, \dots, N\}$, or all N components of the involved vectors, respectively.

Remark 5.14 (Advantages of the discrete implementation)
For $d = 1$ using the discrete Algorithm 3 to approximate its continuous analogon, yields satisfying results for N and K large enough such that the parameter domain Ω and the parameter set $\mathcal{P}$ are represented reasonably well by their discrete counterparts, see the numerical results in Section 5.5 above where we present the results of applying the

Algorithm 3 Discrete EI algorithm, $d = 1$

```
1: Let $\Omega_{\text{discr.}}$ be a finite, discrete set in $\mathbb{R}$, $|\Omega_{\text{discr.}}| = N \in \mathbb{N}$, $\Omega = \{\omega_1, \dots, \omega_N\}$
2: Let $\mathcal{P}_{\text{discr.}}$ be some finite parameter set in $\mathbb{R}$, $|\mathcal{P}_{\text{discr.}}| = K \in \mathbb{N}$
3: Let further $\mathcal{U}_{\text{discr.}}$ be a finite set of parametrized vectors on $\Omega_{\text{discr.}}$, $|\mathcal{U}_{\text{discr.}}| = K \in \mathbb{N}$
4: function DISCRETE INTERPOLATION OPERATOR $I_M^{\text{DISCR}}(\vec{u})$
5:     return $I_M^{\text{discr}}(\vec{u}) = \sum_{i=1}^{M} \alpha_i(\vec{u})\vec{q}_i$
6:     with $\alpha_i \in \mathbb{R}$, $i \in \{1, \dots, M\}$, depending on $\vec{u}$ and given by
7:         $Q\vec{\alpha} = (\vec{u}^{(\iota_1)}, \dots, \vec{u}^{(\iota_M)}), \quad Q \in \mathbb{R}^{M \times M}, \quad Q_{ij} = \vec{q}_j^{(\iota_i)}$
8:     where the set of magic indices $\{\iota_1, \dots, \iota_M\} \subset \{1, \dots, N\}$ and the set of basis
       vectors $\{\vec{q}_1, \dots, \vec{q}_M\}$ are recursively defined by
9:         $\vec{u}_1 = \underset{\vec{u}_i \in \mathcal{U}_{\text{discr}},\ i=1,\dots,K}{\arg\max} \ \max_{j=1,\dots,N} \left|\vec{u}_i^{(j)}\right|$
10:        $\iota_1 = \arg\max_{j=1,\dots,N} \left|\vec{u}_1^{(j)}\right|$
11:        $\xi_1 = \omega_{\iota_1}$
12:        $\vec{q}_1 = \frac{1}{\vec{u}_1^{(\iota_1)}}\vec{u}_1$
13:    and for $M > 1$ with $\vec{r}_i = \vec{u}_i - I_{M-1}^{\text{discr}}(\vec{u}_i)$, $i \in \{1, \dots, N\}$, by
14:        $\vec{u}_M = \underset{\vec{u}_i \in \mathcal{U}_{\text{discr}},\ i=1,\dots,K}{\arg\max} \ \max_{j \in \{1,\dots,N\}} \left|\vec{r}_i^{(j)}\right|$
15:        $\iota_M = \underset{i=1,\dots,N}{\arg\max} \left|\vec{r}_M^{(i)}\right|$
16:        $\xi_M = \omega_{\iota_M}$
17:        $\vec{q}_M = \frac{1}{\vec{r}_M^{(\iota_M)}} \left(\vec{u}_M - I_{M-1}^{\text{discr}}(\vec{u}_i)\right)$
```

(discrete) EI algorithm to Fourier pricing. In the application we considered, the main advantages of implementing the EI algorithm discretely consisted in a fast offline phase, low storage costs and in avoiding the numerical misidentification of global maxima in line 14 and line 15 by considering the whole (discrete) domain in line 14, instead. In our experiments for the one-dimensional case, $d = 1$, we thus observe good results regarding both approximation accuracy and numerical cost.

The advantages sketched in Remark 5.14 are empirically validated by our numerical pricing experiments in Section 5.5, above. Unfortunately, these advantages vanish, when the dimensionality of the problem increases to $d > 1$.

Remark 5.15 (Disadvantages of the discrete implementation for $d > 1$)
The discrete Algorithm 3 can be naturally extended for the multivariate case, $d > 1$, by replacing vectors with matrices. For $d > 1$, however, the tradeoff between N and K large enough to provide Ω_{discr} and $\mathcal{U}_{discr}$ with enough richness and N and K small enough such that the numerical cost remains bearable can in general hardly be maintained. The antagonism between approximation precision and acceptable numerical complexity arises. This antagonism consists of two aspects, storage and computational speed. The requirements to the physical storage in the discrete setting outlined above are of order $N^d\cdot$

K, when the pool of basis function candidates $\mathcal{U}_{discr}$ is explicitly constructed and stored to select the basis function candidates $\vec{u}_i$, $i \in \{1, \dots, M\}$, from. Assuming reasonable values for N and K, this threshold is exceeded in our experiments for d $= 2$, by far. One can avoid these demands to physical storage by evaluating in step M each of the $K-(M-1)$ remaining basis function candidates in $\mathcal{U}^{discr}$ and their interpolation iteratively over Ω^{discr} and storing the respective maximum absolute value $\max_{j\in\{1,\dots,N\}} |\vec{r}_{\cdot}^{(j)}|$ of all components for each residual $\vec{r}_{\cdot}$. This approach results in a number of function evaluations of order $N^d \cdot K$ in each step of identifying the next basis function $\vec{q}_M$. As a consequence, the offline phase is prolonged considerably with respect to computational time.

We conclude that the discrete implementation provided by Algorithm 3 for approximating the continuous Empirical Interpolation method described in Algorithm 2 reaches its limits of feasibility when $d > 1$. To avoid both, the indicated storage requirements and the alternative of processing a large amount of function evaluations during the offline phase one needs to implement the Empirical Interpolation method not discretely but continuously, instead.

5.7 Non-recursive empirical interpolation

From the description of the Empirical Interpolation in Algorithm 2 we understand, that the definition of the interpolation operator I_M introduces a recursive pattern into the definition of the interpolating basis functions q_i, $1 \leq i \leq M$, as defined in line 16.

In a discrete implementation of the algorithm, this recursion is seamlessly adopted. As we see in line 17, each basis function vector $\vec{q}_M$ depends on $\vec{u}_M$ and all previously selected basis function vectors $\vec{q}_i$, $i \in \{1, \dots, M-1\}$. Thus, recursively, each basis function vector $\vec{q}_M$ depends on all previously selected basis function candidates $\vec{u}_i$, $i \in \{1, \dots, M\}$, but the precise design of that recursive dependence is hidden and of no relevance to a proper functioning of the discrete implementation. Once the basis function vectors $\vec{q}_i$, $i \in \{1, \dots, M\}$, have been computed, they can be stored and each of their components can be accessed, directly. In other words, each basis function vector $\vec{q}_i$, $i \in \{1, \dots, M\}$, can be evaluated over Ω^{discr} immediately. An application of the discrete Interpolation Operator I_M^{discr} of Algorithm 3 thus consists of solving an equation system and adding a (weighted) sum of vectors.

In a continuous implementation, this feature is lost. A numerical evaluation of a (weighted) sum of $u_i \in \mathcal{U}$, $i \in \{1, \dots, N\}$, for some $N \in \mathbb{N}$ at some $\xi \in \Omega$ on a continuous Ω relies on the evaluation of each individual u_i, $i \in \{1, \dots, N\}$, at $\xi \in \Omega$ and the subsequent composition of the (weighted) sum. The interpolation operator I_M of Algorithm 2 consists of such a weighted sum of u_i, $i \in \{1, \dots, M\}$. By its definition in line 6, the evaluation of $I_M(u)$ at $\xi \in \Omega$ relies on the evaluation of θ_j^M at ξ for all $j \in \{1, \dots, M\}$. By their definition in line 8, each of these θ_j^M relies on the evaluation of q_i, $i \in \{1, \dots, M\}$. While by line 12, q_1 only depends on u_1, each of the other basis functions q_i with $2 \leq i \leq M$ is defined via an evaluation of both u_i and an additional call to the interpolation operator I_{i-1}.

This recursion therefore numerically complicates the process of identifying u_M during the offline phase, especially for large values of M. Then, a naive call of I_M triggers a recursion that comes at an immense numerical cost.

Remark 5.16 (Complexity of the recursive interpolation operator $\boldsymbol{I_M(\cdot)}$)
Let $M \in \mathbb{N}$ and consider the recursive interpolation operator of Algorithm 2. Let $u \in \mathcal{U}$ and $\xi \in \Omega$. For an evaluation of $I_M(u)(\xi)$, the interpolation operator calls the chosen basis functions q_i, $i \in \{1, \dots, M\}$. By its definition in line 12, q_1 depends on u_1 only, while each of the other q_i, $i \in \{2, \dots, M\}$, depends on u_i and $I_{i-1}(u_i)$, letting the interpolation operator reappear. Thus, each q_i, $i \in \{2, \dots, M\}$, depends multiply on all q_j, $j \in \{1, \dots, i-1\}$. The scheme of this recursive dependence of I_M on the basis functions q_i, $i \in \{1, \dots, M\}$, is visualized in Figure 5.19 for the case of $M = 4$. Ultimately, an evaluation of $I_M(u)(\xi)$ translates into evaluations $u_i(\xi)$, $i \in \{1, \dots, M\}$, where due to the recursive definition of the operator, each u_i will be evaluated several times at ξ and the results will be weighted and summed up. In total, a naive, recursive call of $I_M(u)(\xi)$ results in

$$\sum_{k=1}^{M} \#\{\textit{elementary function evaluations triggered by } q_k\} = \sum_{k=1}^{M} 2^{k-1} = 2^M - 1$$

elementary function evaluations of the u_i, $i \in \{1, \dots, M\}$, which were chosen during the offline phase of the algorithm.

As Remark 5.16 underlines, the fully recursive Empirical Interpolation operator $I_M(\cdot)$ is numerically unfeasible. Due to its recursive structure, naive evaluations of the operator result in computation times that increase exponentially in M. Especially in a non-discrete implementation of the algorithm in combination with optimization routines for lines 14 and 15, this runtime behavior diminishes the practical value of the algorithm in its current form.

A closer look at the definition of the interpolation operator $I_M(\cdot)$ reveals, that for given $u \in \mathcal{U}$ and $\xi \in \Omega$ we do not have to expand each q_k, $k \in \{1, \dots, M\}$, *multiple times* until the level of the elementary function evaluations u_i, $i \in \{1, \dots, k\}$, in order to evaluate $I_M(u)(\xi)$. Instead, once $q_k(\xi)$, $k \in \{1, \dots, M\}$, is computed, we store this function value. For the evaluation of $q_{k+1}(\xi)$ we then only need to compute $u_{k+1}(\xi)$, access all previously stored values $q_1(\xi), \dots, q_k(\xi)$ and add them in the correct way,

$$\begin{aligned} q_{k+1}(\xi) &= c\left(u_{k+1}(\xi) - I_k(u_{k+1})(\xi)\right) \\ &= c\left(u_{k+1}(\xi) - \sum_{i=1}^{k}\sum_{j=1}^{k} (B^k)^{-1}_{ij} u_k(\xi_j) q_i(\xi)\right) \\ &= c\left(u_{k+1}(\xi) - \sum_{i=1}^{k} \alpha_i q_i(\xi)\right), \end{aligned} \tag{5.85}$$

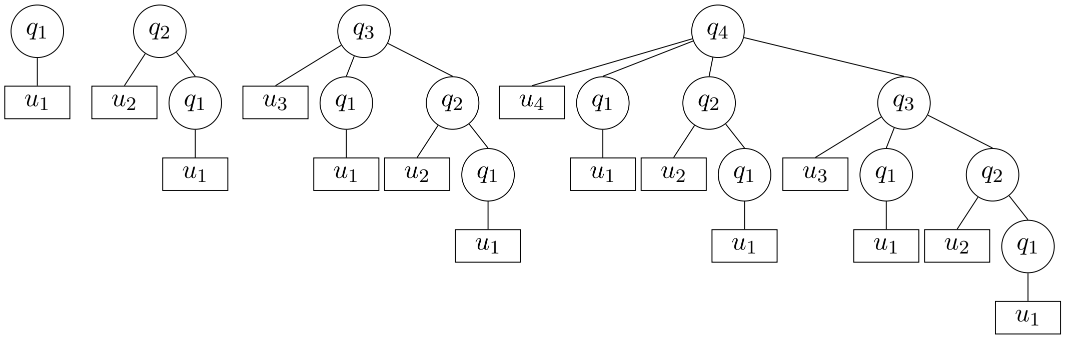

Figure 5.19 Visualization of the complexity of the (fully recursive) Empirical Interpolation operator $I_M(\cdot)$ of Algorithm 2 for $M = 4$. Evaluating $I_M(u)(\xi)$ for some $u \in \mathcal{U}$ and some $\xi \in \Omega$ results in function calls $q_i(\xi)$, $i \in \{1, \dots, M\}$, which ultimately translate into elementary function evaluations $u_i(\xi)$, $i \in \{1, \dots, M\}$. A rectangularly boxed u_i, $i \in \{1, \dots, M\}$, denotes such a function call of u_i at ξ. Due to the special kind of recursive dependence of each basis function q_k, $k \in \{2, \dots, M\}$, on all previous basis functions q_l, $l \in \{1, \dots, k-1\}$, the number of elementary function calls generated by q_k increases exponentially in k.

with $c = 1/\left(u_{k+1}(\xi_{k+1}) - I_k(u_{k+1})(\xi_{k+1})\right)$ and $\alpha_i = \sum_{j=1}^{k} (B^k)^{-1}_{ij} u_k(\xi_j)$, $i \in \{1, \dots, k\}$. By this approach, the recursive structure of 5.19 is reduced to a linear relation shown in 5.20.

Remark 5.17 (Complexity of the semi-recursive interpolation operator $I_M(\cdot)$) *Let $M \in \mathbb{N}$, $u \in \mathcal{U}$ and $\xi \in \Omega$. For the computation of $I_M(u)(\xi)$ we need only*

$$\sum_{k=1}^{M} \#\{\textit{elementary operations to derive } q_k(\xi)\} = \sum_{k=1}^{M} k = \frac{M(M+1)}{2}$$

elementary operations when intermediary results $q_k(\xi)$ are saved, once they are computed. This reduced the complexity of evaluating the Empirical Interpolation operator significantly, as 5.20 demonstrates. Note, however, that these intermediary results all depend on a $\xi \in \Omega$ that needs to be fixed beforehand. For a repeated evaluation of $I_M(u)(\xi)$ for different values $\xi \in \Omega$, the semi-recursion depicted in 5.20 needs to be resolved, repeatedly, for each such $\xi \in \Omega$ individually.

Saving intermediary results in the evaluation of $I_M(u)(\xi)$ for given $u \in \mathcal{U}$ and $\xi \in \Omega$ reduces the complexity of recursion of the interpolation operator significantly, as Remark 5.17 and Figure 5.20 demonstrate. This already allows continuous implementations of the algorithm with optimization routines looking for optimal $u \in \mathcal{U}$ and $\xi \in \Omega$ that maximize the quantities in lines 14 and 15 in iteration M of the offline phase. By such a continuous implementation, storage limitations as mentioned in Remark 5.15 are avoided. Yet, for given $u \in \mathcal{U}$, the quantities in this semi-recursive conception of the interpolation operator $I_M(\cdot)$, depend on $\xi \in \Omega$, see Equation (5.85). Thus, a repeated

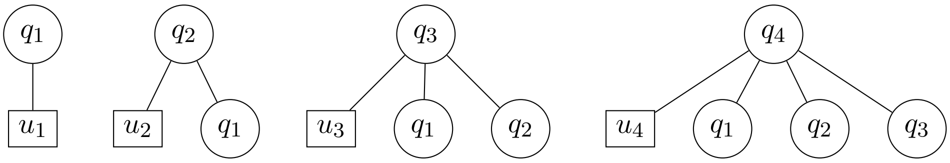

Figure 5.20 Visualization of the complexity of the (semi-recursive) Empirical Interpolation operator $I_M(\cdot)$ when intermediate results are saved. Again we chose $M = 4$. The recursive definition of each $q_k(\xi)$ is resolved only once and the result is stored. As a consequence, the exponentially recursive scheme of 5.19 reduces dramatically. Each u_i, $i \in \{1, \ldots, M\}$, is evaluated only once.

evaluation of $I_M(u)(\xi)$, the interpolated version of some $u \in \mathcal{U}$ at different $\xi \in \Omega$ can become numerically costly, as well. This is the case for example, when the numerical evaluation of an integral of a function containing $I_M(u)$ is concerned.

Therefore, we are interested in a non-recursive representation of the basis functions q_k, $1 \leq k \leq M$. By this we mean a representation of each q_k, $1 \leq k \leq M$, in terms of a weighted sum of u_i, $1 \leq i \leq k$, with explicitly given coefficients. If we were able to evaluate the interpolation operator I_M without relying on a recursive evaluation of all q_k, $1 \leq k \leq M$, in a numerically stable way, we could further speed up the offline phase considerably.

We introduce some quantities that will play the key roles in deriving a non-recursive expression for the interpolation operator I_M and discuss some of their properties. Let $M \in \mathbb{N}$ and define

$$\begin{aligned} r_1 &= u_1(\xi_1), \\ r_M &= u_M(\xi_M) - I_{M-1}(u_M)(\xi_M), \qquad M > 1, \end{aligned} \tag{5.86}$$

and

$$\widetilde{w}_i^{(j)} = \frac{w_i^{(j)}}{r_i}, \qquad 1 \leq i, j \leq M, \tag{5.87}$$

with

$$w_i^{(j)} = \left(\left(B^M\right)^{-1} u_j(\vec{\xi}^M) \right)_i, \qquad 1 \leq i, j \leq M, \tag{5.88}$$

the i-th component of the vector generated by multiplying the inverse of B^M with the j-th function u_j to be selected from the pool $\mathcal{U}$ in line 14 evaluated at all magic points. Note that $w_i^{(j)}$ is in a sense independent of M. This property is inherited from B^M being a lower triangular matrix (and thus $(B^M)^{-1}$ being a lower triangular matrix, as well). Additionally, with $M' \geq M$, by definition, $B_{ij}^M = B_{ij}^{M'}$ for $i, j \leq M$ and $B_{ij}^{M'} = 0$ for

$i \leq M$, $M < j \leq M'$,

$$B^{M'} = \begin{pmatrix} & & & 0 & \dots & 0 \\ & B^M & & 0 & \dots & 0 \\ & & & 0 & \dots & 0 \\ q_1(\xi_{M+1}) & \dots & q_M(\xi_{M+1}) & 1 & \ddots & 0 \\ \vdots & & \vdots & & \ddots & 0 \\ q_1(\xi_{M'}) & \dots & q_M(\xi_{M'}) & \dots & q_{M'-1}(\xi_{M'}) & 1 \end{pmatrix},$$

and since B^M is a lower triangular matrix we equivalently have $(B^M)^{-1}_{ij} = (B^{M'})^{-1}_{ij}$ for $i, j \leq M$ and $(B^{M'})^{-1}_{ij} = 0$ for $i \leq M$, $M < j \leq M'$, as well.

Consequently, for $i, j \leq M \leq M'$ we have

$$w_i^{(j)} = \left(\left(B^M\right)^{-1} u_j(\vec{\xi}^M) \right)_i = \left(\left(B^{M'}\right)^{-1} u_j(\vec{\xi}^{M'}) \right)_i . \tag{5.89}$$

The $w_i^{(j)}$, $i, j \leq M$, play a key role in the non-recursive representation of the basis functions q_i, $i \leq M$. Due to property (5.89) each $w_i^{(j)}$, $i, j \leq M$, needs to be computed only once during the offline phase and can then be stored and reused as B^M grows during the iterative process of finding the basis functions q_k, $k \geq 1$.

Lemma 5.18 (Representing $\boldsymbol{q_k}$ using the $\boldsymbol{w_i^{(j)}}$)
Using the definition of the $w_i^{(j)}$, $i, j \leq M$, that we just introduced and defined in (5.88)*, we can rewrite the definition of the Empirical Interpolation basis functions q_k, $1 \leq k \leq M$, to*

$$q_k(\xi) = \frac{1}{r_k} \left(u_k(\xi) - \sum_{i=1}^{k-1} w_i^{(k)} q_i(\xi) \right), \tag{5.90}$$

for all $\xi \in \Omega$.

Proof
Let $\xi \in \Omega$. For $k = 1$, the sum in (5.90) equals zero and the claim obviously holds by definition of q_1 in line 12. Let now $2 \leq k \leq M$ and $\widetilde{q}_k(\xi) = r_k q_k(\xi)$ with r_k as given in the algorithm. By the relation in line 16 we have

$$\begin{aligned} \widetilde{q}_k(\xi) &= u_k(\xi) - I_{k-1}(u_k)(\xi) \\ &= u_k(\xi) - \sum_{j=1}^{k-1} u_k(\xi_j) \theta_j^{k-1}(\xi) \\ &= u_k(\xi) - \sum_{j=1}^{k-1} u_k(\xi_j) \sum_{i=1}^{k-1} (B^{k-1})^{-1}_{ij} q_i(\xi) \\ &= u_k(\xi) - \sum_{i=1}^{k-1} \left(\sum_{j=1}^{k-1} (B^{k-1})^{-1}_{ij} u_k(\xi_j) \right) q_i(\xi), \end{aligned} \tag{5.91}$$

where we simply inserted the quantities defined in Algorithm 2. The expression given by the sum $\sum_{j=1}^{k-1}(B^{k-1})^{-1}_{ij}u_k(\xi_j)$ for some $1 \leq i \leq k-1$ in (5.91) can be interpreted as component i of the result of a matrix-vector multiplication,

$$\sum_{j=1}^{k-1}(B^{k-1})^{-1}_{ij}u_k(\xi_j) = \Big((B^{k-1})^{-1}u_k((\xi_1,\dots,\xi_{k-1}))\Big)_i. \tag{5.92}$$

By (5.89), the result of (5.92) is independent of k,

$$\Big((B^{k-1})^{-1}u_k((\xi_1,\dots,\xi_{k-1}))\Big)_i = \Big(\left(B^M\right)^{-1}u_k(\vec{\xi}^M)\Big)_i = w_i^{(k)}, \tag{5.93}$$

by the definition of $w_i^{(k)}$ in (5.87). Inserting (5.93) into (5.92) and the result into (5.91) proves the claim. □

Lemma 5.18 provides us with the starting point for resolving the recursive definition of the Empirical Interpolation basis functions q_k, $1 \leq k \leq M$, of line 16. Before we are able to state and prove a non-recursive representation of the basis functions we prove the following auxiliary Lemma 5.19.

Lemma 5.19 (Auxiliary lemma)
Let $M \in \mathbb{N}$. Let $\widetilde{w}_i^{(j)}$, $1 \leq i,j \leq M$ be given by (5.87). Define further for all $1 \leq j < k \leq M$

$$c_j^{(k)} = \widetilde{w}_{k-j}^{(k)} - \sum_{i=1}^{j-1}\widetilde{w}_{k-j}^{(k-i)}c_i^{(k)} \tag{5.94}$$

recursively. Then for $k < M$ the relation

$$\sum_{j=1}^{i-1}c_j^{(k+1-i+j)}\widetilde{w}_{k+1-i+j}^{(k+1)} = \sum_{j=1}^{i-1}\widetilde{w}_{k+1-i}^{(k+1-j)}c_j^{(k+1)} \tag{5.95}$$

holds for all $i \in \{1,\dots,k\}$.

Proof
We prove (5.95) by induction over $i \in \{1,\dots,k\}$.

For $i = 1$ there is nothing to show since both sides of (5.95) turn into empty sums with value 0. To provide an intuition about the structural relation between the two sums in (5.95), we additionally validate the equation for $i = 2$. Then, (5.95) holds if

$$c_1^{(k+1-2+1)}\widetilde{w}_{k+1-2+1}^{(k+1)} = \widetilde{w}_{k+1-2}^{(k+1-1)}c_1^{(k+1)}$$

which is equivalent to

$$c_1^{(k)}\widetilde{w}_k^{(k+1)} = \widetilde{w}_{k-1}^{(k)}c_1^{(k+1)}. \tag{5.96}$$

By the definition of $c_j^{(k)}$ in (5.94) we have $c_1^{(k)} = \widetilde{w}_{k-1}^{(k)}$ and $c_1^{(k+1)} = \widetilde{w}_k^{(k+1)}$ which proves the induction assumption.

For the induction step we assume the claim (5.95) to hold for all $1 \leq i' \leq i$ for some $1 \leq i < k$ and prove it for $i+1$, that is we show

$$\sum_{j=1}^{i} c_j^{(k-i+j)} \widetilde{w}_{k-i+j}^{(k+1)} = \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} c_j^{(k+1)}. \tag{5.97}$$

By invoking on both sides the definition of $c_j^{(k-i+j)}$ or $c_j^{(k+1)}$, respectively, as given by (5.94), we conclude that (5.97) is equivalent to

$$\begin{aligned} &\sum_{j=1}^{i} \left[\widetilde{w}_{k-i}^{(k-i+j)} - \sum_{l=1}^{j-1} \widetilde{w}_{k-i}^{(k-i+j-l)} c_l^{(k-i+j)} \right] \widetilde{w}_{k-i+j}^{(k+1)} \\ &\qquad = \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} \left[\widetilde{w}_{k+1-j}^{(k+1)} - \sum_{l=1}^{j-1} \widetilde{w}_{k+1-j}^{(k+1-l)} c_l^{(k+1)} \right]. \end{aligned} \tag{5.98}$$

We expand the multiplication in the outer sums of (5.98) and separate them, transforming the equality to

$$\begin{aligned} &\sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k-i+j)} \widetilde{w}_{k-i+j}^{(k+1)} - \sum_{j=1}^{i} \widetilde{w}_{k-i+j}^{(k+1)} \sum_{l=1}^{j-1} \widetilde{w}_{k-i}^{(k-i+j-l)} c_l^{(k-i+j)} \\ &\qquad = \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} \widetilde{w}_{k+1-j}^{(k+1)} - \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} \sum_{l=1}^{j-1} \widetilde{w}_{k+1-j}^{(k+1-l)} c_l^{(k+1)}. \end{aligned} \tag{5.99}$$

Let us consider the first, single sums on both sides of (5.99). Inverting the order of summation in the sum on the left reveals that the two leading individual sums on boths sides of (5.99) are identical,

$$\sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k-i+j)} \widetilde{w}_{k-i+j}^{(k+1)} = \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k-i+[i+1-j])} \widetilde{w}_{k-i+[i+1-j]}^{(k+1)} = \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} \widetilde{w}_{k+1-j}^{(k+1)}. \tag{5.100}$$

Therefore, (5.99) holds if

$$\sum_{j=1}^{i} \widetilde{w}_{k-i+j}^{(k+1)} \sum_{l=1}^{j-1} \widetilde{w}_{k-i}^{(k-i+j-l)} c_l^{(k-i+j)} = \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} \sum_{l=1}^{j-1} \widetilde{w}_{k+1-j}^{(k+1-l)} c_l^{(k+1)}. \tag{5.101}$$

We apply the induction assumption to each summand of the inner sum on the right hand side of (5.101). We can do so, since by our induction assumption the claim

$$\sum_{j=1}^{i'-1} c_j^{(k+1-i'+j)} \widetilde{w}_{k+1-i'+j}^{(k+1)} = \sum_{j=1}^{i'-1} \widetilde{w}_{k+1-i'}^{(k+1-j)} c_j^{(k+1)} \tag{5.102}$$

is validated for all $i' \in \{1, \dots, i\}$. The summation variable j on the right hand side of (5.101) assumes values $j \in \{1, \dots i\}$ and thus takes the role of i' in (5.102). Consequently, we invoke the induction assumption and conclude, that the claim holds if

$$\sum_{j=1}^{i} \widetilde{w}_{k-i+j}^{(k+1)} \sum_{l=1}^{j-1} \widetilde{w}_{k-i}^{(k-i+j-l)} c_l^{(k-i+j)} = \sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} \sum_{l=1}^{j-1} c_l^{(k+1-j+l)} \widetilde{w}_{k+1-j+l}^{(k+1)}. \tag{5.103}$$

In order to finally prove equality (5.103), we rearrange the order of summation on the right hand side of (5.103). To that extent we introduce for $i \in \mathbb{N}$ the set

$$\mathcal{I}_i = \{(j,l) \in \mathbb{N}^2 \mid j \leq i,\ l \leq j-1\} \tag{5.104}$$

and the mapping m on $\mathcal{I}_i$ by

$$m : (j,l) \mapsto (i+1-j+l, l), \quad \forall (j,l) \in \mathcal{I}_i, \tag{5.105}$$

and denote by $m_n((j,l))$ the projection of $m((j,l))$ onto the n-th component, $n \in \{1,2\}$, for all $(j,l) \in \mathcal{I}_i$. We have $m(\mathcal{I}_i) = \mathcal{I}_i$, since with $(j,l) \in \mathcal{I}_i$ we have $i+1-j+l \leq i$ if and only if $l \leq j-1$ which is true by definition of $\mathcal{I}_i$. Thus, the mapping m maps $\mathcal{I}_i$ onto itself, $m(\mathcal{I}_i) = \mathcal{I}_i$, and it does so bijectively with $m^{-1} = m$. The effect that the mapping m has to the order of summands is visualized in Table 5.3 and Table 5.4.

j l =	1	2	3	$i-1$
2	1			
3	2	3		
4	4	5	6	
	⋮		⋱	
i	$(i-1)(i-2)/2+1$		…	$i(i-1)/2$

Table 5.3 Visualization of the queue of non-zero summands on the right hand side of (5.103). The number in each of the boxes denotes the position of the summand belonging to the respective (j,l)-tuple in the sum.

We prove equality (5.103) by continuing on the right hand side of (5.103) and applying

l $\overset{j}{=}$	1	2	3		$i-1$
2	$\boxed{(i-1)(i-2)/2+1}$				
	$\vdots$				
$i-2$	$\boxed{4}$		$\cdot$		
$i-1$	$\boxed{2}$	$\boxed{5}$			
i	$\boxed{1}$	$\boxed{3}$	$\boxed{6}$	$\dots$	$\boxed{i(i-1)/2}$

Table 5.4 Visualization of the queue of non-zero summands on the right hand side of (5.103) after the order of summation has been changed by applying mapping m. The numbers in the boxes denote the previous position of the respective summand in the sum, compare Table 5.3.

mapping m to change the order of summation,

$$\begin{aligned}
&\sum_{j=1}^{i} \widetilde{w}_{k-i}^{(k+1-j)} \sum_{l=1}^{j-1} c_l^{(k+1-j+l)} \widetilde{w}_{k+1-j+l}^{(k+1)} \\
&= \sum_{(j,l)\in\mathcal{I}_i} \widetilde{w}_{k-i}^{(k+1-j)} c_l^{(k+1-j+l)} \widetilde{w}_{k+1-j+l}^{(k+1)} \\
&= \sum_{(j,l)\in\mathcal{I}_i} \widetilde{w}_{k-i}^{(k+1-m_1((j,l)))} c_{m_2((j,l))}^{(k+1-m_1((j,l))+m_2((j,l)))} \widetilde{w}_{k+1-m_1((j,l))+m_2((j,l))}^{(k+1)} \\
&= \sum_{j=1}^{i} \sum_{l=1}^{j-1} \widetilde{w}_{k-i}^{(k+1-[i+1-j+l])} c_{[l]}^{(k+1-[i+1-j+l]+[l])} \widetilde{w}_{k+1-[i+1-j+l]+[l]}^{(k+1)} \\
&= \sum_{j=1}^{i} \sum_{l=1}^{j-1} \widetilde{w}_{k-i}^{(k-i+j-l)} c_l^{(k-i+j)} \widetilde{w}_{k-i+j}^{(k+1)} \\
&= \sum_{j=1}^{i} \widetilde{w}_{k-i+j}^{(k+1)} \sum_{l=1}^{j-1} \widetilde{w}_{k-i}^{(k-i+j-l)} c_l^{(k-i+j)},
\end{aligned}$$

which coincides with the left hand side of (5.103) and thereby finishes the proof of the lemma. □

With Lemma 5.19 we are able to derive the non-recursive representation for the Empirical Interpolation basis functions q_k, $1 \le k \le M$.

Lemma 5.20 (Non-recursive representation of q_k)
Let $M \in \mathbb{N}$. Let u_i, $i \leq M$, be given by line 14, $\widetilde{w}_i^{(j)}$, $i, j \leq M$, as defined in (5.87) and r_k, $1 \leq k \leq M$ as given by the algorithm. Then, each of the basis functions q_k, $k \leq M$, defined by line 16 and selected during the offline phase of the Empirical Interpolation method as described in Algorithm 2 depends non-recursively on u_i, $1 \leq i \leq k$, and follows the formula

$$q_k(\xi) = \frac{1}{r_k}\left(u_k(\xi) - \sum_{j=1}^{k-1} c_j^{(k)} u_{k-j}(\xi)\right), \qquad \forall \xi \in \Omega, \quad 1 \leq k \leq M, \tag{5.106}$$

with recursively given coefficients

$$c_j^{(k)} = \widetilde{w}_{k-j}^{(k)} - \sum_{i=1}^{j-1} \widetilde{w}_{k-j}^{(k-i)} c_i^{(k)}, \qquad j \in \{1, \dots, k-1\}. \tag{5.107}$$

Proof
Let $\xi \in \Omega$. We prove the claim by induction over k. For $k = 1$ we have

$$q_1(\xi) = \frac{u_1(\xi)}{r_1} \tag{5.108}$$

which is true by definition of q_1 in line 12.

For the induction step, we assume the claim (5.106) to hold for all k' with $1 \leq k' \leq k$ for some $k < M$. For the sake of notational convenience we omit the ξ in the following. We use the recursive formula for q_{k+1} provided by Lemma 5.18 and state

$$q_{k+1} = \frac{1}{r_{k+1}}\left(u_{k+1} - \sum_{i=1}^{k+1-1} w_i^{(k+1)} q_i\right). \tag{5.109}$$

In the sum of (5.109), the q_i, $i \leq k$, are summed up. For each q_i, $i \leq k$, however, the induction assumption holds. We thus replace the q_i terms in (5.109) by the appropriate term given by (5.106) and proceed with

$$\begin{aligned} q_{k+1} &= \frac{1}{r_{k+1}}\left(u_{k+1} - \sum_{i=1}^{k} w_i^{(k+1)} q_i\right) \\ &= \frac{1}{r_{k+1}}\left(u_{k+1} - \sum_{i=1}^{k}\left(w_i^{(k+1)}\left[\frac{1}{r_i}\left(u_i - \sum_{j=1}^{i-1} c_j^{(i)} u_{i-j}\right)\right]\right)\right) \\ &= \frac{1}{r_{k+1}}\left(u_{k+1} - \sum_{i=1}^{k} \widetilde{w}_i^{(k+1)}\left(u_i - \sum_{j=1}^{i-1} c_j^{(i)} u_{i-j}\right)\right), \end{aligned} \tag{5.110}$$

where we used the definition of $\widetilde{w}_i^{(k)} = w_i^{(k)}/r_i$, $\forall 1 \leq i, k \leq M$ in the last step. Comparing the result of (5.110) to the claim in (5.106) we conclude that we have to show that

$$\sum_{i=1}^{k} \widetilde{w}_i^{(k+1)} \left(u_i - \sum_{j=1}^{i-1} c_j^{(i)} u_{i-j} \right) = \sum_{j=1}^{k} c_j^{(k+1)} u_{k+1-j} \tag{5.111}$$

to finish the proof. In the following we rewrite the left hand side of (5.111) such that each u_i, $1 \leq i \leq k$, appears only once in the sum and then prove the claim by a comparison of the respective coefficients. For the left hand side in (5.111) we trivially have

$$\sum_{i=1}^{k} \widetilde{w}_i^{(k+1)} \left(u_i - \sum_{j=1}^{i-1} c_j^{(i)} u_{i-j} \right) = \sum_{i=1}^{k} \widetilde{w}_i^{(k+1)} u_i - \sum_{i=1}^{k} \widetilde{w}_i^{(k+1)} \sum_{j=1}^{i-1} c_j^{(i)} u_{i-j}. \tag{5.112}$$

Inverting the order of summation in the first sum of (5.112) gives

$$\sum_{i=1}^{k} \widetilde{w}_i^{(k+1)} u_i = \sum_{i=1}^{k} \widetilde{w}_{(k+1)-i}^{(k+1)} u_{(k+1)-i}. \tag{5.113}$$

The second sum in (5.112) can trivially be written as

$$\sum_{i=1}^{k} \widetilde{w}_i^{(k+1)} \sum_{j=1}^{i-1} c_j^{(i)} u_{i-j} = \sum_{(i,j)\in\mathcal{I}_k} \widetilde{w}_i^{(k+1)} c_j^{(i)} u_{i-j} \tag{5.114}$$

with

$$\mathcal{I}_k = \left\{ (i,j) \in \mathbb{N}^2 \mid 1 \leq i \leq k,\ 1 \leq j \leq i-1 \right\}, \tag{5.115}$$

as already defined in (5.104) earlier with only slight labeling differences. Therefore, we may change the order of summation in (5.114) by applying the mapping m of (5.105) to get

$$\begin{aligned} \sum_{i=1}^{k} \widetilde{w}_i^{(k+1)} \sum_{j=1}^{i-1} c_j^{(i)} u_{i-j} &= \sum_{(i,j)\in\mathcal{I}_k} \widetilde{w}_i^{(k+1)} c_j^{(i)} u_{i-j} \\ &= \sum_{(i,j)\in\mathcal{I}_k} \widetilde{w}_{m_1(i,j)}^{(k+1)} c_{m_2(i,j)}^{(m_1(i,j))} u_{m_1(i,j)-m_2(i,j)} \\ &= \sum_{i=1}^{k} \sum_{j=1}^{i-1} c_j^{(k+1-i+j)} \widetilde{w}_{k+1-i+j}^{(k+1)} u_{k+1-i}. \end{aligned} \tag{5.116}$$

Inserting (5.116) together with (5.113) into (5.112) we have

$$\begin{aligned}\sum_{i=1}^{k}\widetilde{w}_i^{(k+1)}&\left(u_i-\sum_{j=1}^{i-1}c_j^{(i)}u_{i-j}\right)\\ &=\sum_{i=1}^{k}\widetilde{w}_i^{(k+1)}u_i-\sum_{i=1}^{k}\widetilde{w}_i^{(k+1)}\sum_{j=1}^{i-1}c_j^{(i)}u_{i-j}\\ &=\sum_{i=1}^{k}\widetilde{w}_{k+1-i}^{(k+1)}u_{k+1-i}-\sum_{i=1}^{k}\sum_{j=1}^{i-1}c_j^{(k+1-i+j)}\widetilde{w}_{k+1-i+j}^{(k+1)}u_{k+1-i}\\ &=\sum_{i=1}^{k}\left(\widetilde{w}_{k+1-i}^{(k+1)}-\sum_{j=1}^{i-1}c_j^{(k+1-i+j)}\widetilde{w}_{k+1-i+j}^{(k+1)}\right)u_{k+1-i}.\end{aligned}\tag{5.117}$$

Thus, recalling (5.111), in order to finish the proof of claim (5.106), we need to show that

$$\sum_{i=1}^{k}\left(\widetilde{w}_{k+1-i}^{(k+1)}-\sum_{j=1}^{i-1}c_j^{(k+1-i+j)}\widetilde{w}_{k+1-i+j}^{(k+1)}\right)u_{k+1-i}=\sum_{i=1}^{k}c_i^{(k+1)}u_{k+1-i}\tag{5.118}$$

which holds if the coefficients of each u_l, $1\le l\le k$ coincide on both sides of (5.118). We thus prove equality (5.118) by validating

$$\widetilde{w}_{k+1-i}^{(k+1)}-\sum_{j=1}^{i-1}c_j^{(k+1-i+j)}\widetilde{w}_{k+1-i+j}^{(k+1)}=c_i^{(k+1)},\qquad\forall i\in\{1,\dots,k\}.\tag{5.119}$$

By the definition of $c_i^{(k+1)}$ in (5.94), (5.119) is equivalent to

$$\widetilde{w}_{k+1-i}^{(k+1)}-\sum_{j=1}^{i-1}c_j^{(k+1-i+j)}\widetilde{w}_{k+1-i+j}^{(k+1)}=\widetilde{w}_{k+1-i}^{(k+1)}-\sum_{j=1}^{i-1}\widetilde{w}_{k+1-i}^{(k+1-j)}c_j^{(k+1)},\qquad\forall i\in\{1,\dots,k\},$$

which trivially is equivalent to

$$\sum_{j=1}^{i-1}c_j^{(k+1-i+j)}\widetilde{w}_{k+1-i+j}^{(k+1)}=\sum_{j=1}^{i-1}\widetilde{w}_{k+1-i}^{(k+1-j)}c_j^{(k+1)},\qquad\forall i\in\{1,\dots,k\}.\tag{5.120}$$

Equation (5.120) holds by Lemma 5.19, which finishes the proof of the claim. □

Remark 5.21 (On the non-recursive representation of q_k)
Lemma 5.20 resolves the recursive dependence of each q_k, $1\le k\le M$, on u_j, $1\le j\le k$, and provides us with a weighted sum of u_j, $1\le j\le k$, with explicitly known coefficients, instead. These coefficients are defined recursively. The lemma thus shifts the recursive pattern in the definition of the basis functions q_k, $1\le k\le M$, from the functions u_j,

$1 \leq j \leq k$, onto coefficients consisting of $c_j^{(k)}$, $1 \leq j < k$. Theoretically, the formula for q_k in (5.90) provided by Lemma 5.18 coincides with the expression (5.106) provided by Lemma 5.20. Numerically, however, the latter is far more appealing. Each basis function q_k, $1 \leq k \leq M$, may now be evaluated continuously with each u_j, $1 \leq j \leq M$, appearing only once thus avoiding the necessity of exponentially many elementary function evaluations as outlined by Remark 5.16. The points on the domain Ω for which the basis function q_k shall be evaluated must not be known beforehand, rendering the result of Lemma 5.20 superior over the semi-non-recursive solution described by Remark 5.17.

Corollary 5.22 (Non-recursive interpolation operator $\boldsymbol{I_M}$)
Let $u \in \mathcal{U}$ and $M \in \mathbb{N}$. Then

$$I_M(u)(\xi) = \sum_{j=1}^{M} c_j u_{M+1-j}(\xi), \qquad \forall \xi \in \Omega, \tag{5.121}$$

with

$$c_j = \widetilde{w}_{M+1-j} - \sum_{i=1}^{j-1} \widetilde{w}_{M+1-j}^{(M+1-i)} c_i, \qquad j \in \{1, \ldots, M\}$$

where

$$\widetilde{w}_i = \frac{\Big((B^M)^{-1} u(\vec{\xi}^M)\Big)_i}{r_i}, \qquad 1 \leq i \leq M,$$

the analogon to the $\widetilde{w}_i^{(j)}$, $1 \leq i, j \leq M$, as defined in (5.87) with r_i, $1 \leq i \leq M$, given by (5.86).

Remark 5.23 (New Challenges for the optimization routine)
During the offline phase of the Empirical Interpolation algorithm, at iteration M in line 16, a basis function q_M is defined via the solution to the optimization problem of line 14. Since the algorithm represents functions $u \in \mathcal{U}$ at the magic points ξ_i, $1 \leq i \leq M-1$, perfectly, the domain $\{u(\xi) - I_{M-1}(u)(\xi) \mid \xi \in \Omega\}$ becomes increasingly uneven for each $u \in \mathcal{U}$ for increasing values of M. Consequently, the identification of both $u \in \mathcal{U}$ and $\xi \in \Omega$ that maximize $|u(\xi) - I_{M-1}(u)(\xi)|$ becomes more and more challenging. For increasing values of M, the risk of running into local minima during the optimization

$$\max_{u \in \mathcal{U}, \, \xi \in \Omega} |u(\xi) - I_{M-1}(u)(\xi)| \tag{5.122}$$

thus rises.

A Integration of special periodic functions

Integrals of univariate integrable functions $f : \mathbb{R}^+ \to \mathbb{R}$ over the semi-infinite integration range $[0, \infty)$ are usually numerically approximated by cutting of the integration range at a value $N \in \mathbb{R}$ beyond which the absolute value of the integrand becomes sufficiently small and the value of

$$\int_N^\infty f(x)\,\mathrm{d}x \approx 0$$

becomes negligible. In some cases, however, the value of N is very large. Then, cutting off the integral at N means either having to deal with a large numerical effort to compute $\int_0^N f(x)\,\mathrm{d}x$ or accepting a significant error in the integral value by choosing $\widetilde{N} \ll N$ as cut-off point, alternatively. In this appendix, we introduce an approximate integration procedure for functions f of a special kind that avoids disregarding the function beyond the cutoff point N. Instead, the approach takes knowledge of the structure of the function f into account in order to improve the approximation of the whole integral value as such.

A.1 An introduction of the integration method

The integration method we are about to present is taylormade for a special kind of functions. After a brief definition, this section states the main result and gives a proof of the lemma.

Definition A.1 (Periodic function)
Let $g : \mathbb{R}_0^+ \to \mathbb{R}$ such that there exists a constant $p \in \mathbb{R}^+$ such that $g(x) = g(x + p)$ for all $x \in \mathbb{R}_0^+$. Then we call g a periodic function *and we call p the* period *of g.*

Lemma A.2 (Approximate integration of special periodic functions)
Let $f : \mathbb{R}^+ \to \mathbb{R}$, $f \in L^1(\mathbb{R}^+)$, be given by

$$f(x) = \frac{g(x)}{x^k}, \qquad \forall x \in \mathbb{R}^+,$$

for some periodic function g with period $p \in \mathbb{R}^+$ and some $1 < k \in \mathbb{N}$. Assume further that $\exists C \in \mathbb{R}^+$ such that

$$|g(x)| \leq C, \qquad \forall x \in [0, p]. \tag{A.1}$$

Then, for any $j \in \mathbb{N}$,

$$\begin{aligned} V := \int_0^\infty f(x)\,\mathrm{d}x &= V_{cut}^{N(j)} + V_{series}^{N(j)} + E(j) \\ &\approx V_{cut}^{N(j)} + V_{series}^{N(j)}, \end{aligned} \tag{A.2}$$

where $N(j) = pj$ for $j \in \mathbb{N}$, $V_{cut}^{N(j)}$ denotes the integral of f up to $N(j)$,

$$V_{cut}^{N(j)} = \int_0^{N(j)} f(x)\,\mathrm{d}x, \tag{A.3}$$

and $V_{series}^{N(j)}$ is given by

$$V_{series}^{N(j)} = I(j) j^k \left(\zeta(k) - \sum_{m=1}^{j-1} \frac{1}{m^k} \right), \tag{A.4}$$

wherein $I(i)$, $i \in \mathbb{N}$, is the integral of f over period i,

$$I(i) = \int_{N(i)}^{N(i+1)} f(x)\,\mathrm{d}x, \tag{A.5}$$

and $\zeta(\cdot)$ denotes the Riemann zeta function,

$$\zeta(s) = \sum_{m=1}^{\infty} \frac{1}{m^s},$$

defined for $s \in \{z \in \mathbb{C} \,|\, \Re(z) > 1\}$, confer Laurinčikas (1996). Furthermore, the error term $E(j)$ in (A.2) *decays as fast to zero as $j \mapsto \sum_{m=1}^{j} \frac{1}{m^k}$ approaches its limit $\zeta(k)$.*

Proof
Fix $1 < j \in \mathbb{N}$. Define

$$\widetilde{I}(m) = I(j)\frac{j^k}{m^k}, \qquad m \in \mathbb{N}. \tag{A.6}$$

We compute

$$\begin{aligned}
V = \int_0^\infty f(x)\,\mathrm{d}x &= \int_0^{N(j)} f(x)\,\mathrm{d}x + \int_{N(j)}^\infty f(x)\,\mathrm{d}x \\
&= V_{\mathrm{cut}}^{N(j)} + \sum_{m=j}^{\infty} \int_{N(m)}^{N(m+1)} f(x)\,\mathrm{d}x \\
&= V_{\mathrm{cut}}^{N(j)} + \sum_{m=j}^{\infty} I(m) \\
&= V_{\mathrm{cut}}^{N(j)} + \sum_{m=j}^{\infty} \widetilde{I}(m) + \sum_{m=j}^{\infty} \Big(I(m) - \widetilde{I}(m)\Big) \\
&= V_{\mathrm{cut}}^{N(j)} + I(j) j^k \sum_{m=j}^{\infty} \frac{1}{m^k} + \sum_{m=j}^{\infty} \Big(I(m) - \widetilde{I}(m)\Big) \\
&= V_{\mathrm{cut}}^{N(j)} + I(j) j^k \left(\zeta(k) - \sum_{m=1}^{j-1} \frac{1}{m^k}\right) + E(j),
\end{aligned}$$

with error term

$$E(j) = \sum_{m=j}^{\infty} \Big(I(m) - \widetilde{I}(m)\Big)\,.$$

We show that $E(j) \to 0$ for $j \to \infty$. Clearly,

$$|E(j)| \leq \sum_{m=j}^{\infty} \Big|I(m) - \widetilde{I}(m)\Big|\,. \tag{A.7}$$

Further, for $m \in \mathbb{N}$ we have

$$\begin{aligned}
\Big|I(m) - \widetilde{I}(m)\Big| &= \left|I(m) - \frac{j^k}{m^k} I(j)\right| \\
&= \left|\int_{N(m)}^{N(m+1)} \frac{g(x)}{x^k}\,\mathrm{d}x - \frac{j^k}{m^k}\int_{N(j)}^{N(j+1)} \frac{g(x)}{x^k}\,\mathrm{d}x\right| \\
&= \left|\int_{N(j)}^{N(j+1)} \frac{g(x)}{(x+(m-j)p)^k}\,\mathrm{d}x - \frac{j^k}{m^k}\int_{N(j)}^{N(j+1)} \frac{g(x)}{x^k}\,\mathrm{d}x\right| \qquad \text{(A.8)} \\
&= \left|\int_{N(j)}^{N(j+1)} \frac{g(x)}{x^k}\left(\frac{x^k}{(x+(m-j)p)^k} - \frac{j^k}{m^k}\right)\mathrm{d}x\right| \\
&\leq \max_{x\in[N(j),N(j+1)]} \left|\frac{x^k}{(x+(m-j)p)^k} - \frac{j^k}{m^k}\right| \int_{N(j)}^{N(j+1)} \left|\frac{g(x)}{x^k}\right|\mathrm{d}x, \qquad \text{(A.9)}
\end{aligned}$$

where we used in (A.8) that g is a periodic function. We find

$$\begin{aligned}\max_{x\in[N(j),N(j+1)]}\left|\frac{x^k}{(x+(m-j)p)^k}-\frac{j^k}{m^k}\right| &= \max_{\delta\in[0,1]}\left|\frac{((j+\delta)p)^k}{((j+\delta)p+(m-j)p))^k}-\frac{j^k}{m^k}\right| \\ &= \max_{\delta\in[0,1]}\left|\frac{(j+\delta)^k}{(m+\delta)^k}-\frac{j^k}{m^k}\right| \qquad (A.10) \\ &= \left[\max_{\delta\in[0,1]}\frac{(j+\delta)^k}{(m+\delta)^k}\right]-\frac{j^k}{m^k}.\end{aligned}$$

Since $m \geq j$ and $k > 1$, $[0,1] \ni \delta \mapsto (j+\delta)/(m+\delta)$ is monotonically increasing and thus assumes its maximum value in $\delta = 1$. Continuing in (A.10) we thus get

$$\left[\max_{\delta\in[0,1]}\frac{(j+\delta)^k}{(m+\delta)^k}\right]-\frac{j^k}{m^k}=\frac{(j+1)^k}{(m+1)^k}-\frac{j^k}{m^k}\leq\frac{1}{m^k}\left((j+1)^k-j^k\right). \qquad (A.11)$$

Consequently, taking (A.9) – (A.11) into account,

$$\left|I(m)-\widetilde{I}(m)\right|\leq\frac{1}{m^k}\left((j+1)^k-j^k\right)\int_{N(j)}^{N(j+1)}\left|\frac{g(x)}{x^k}\right|\mathrm{d}x. \qquad (A.12)$$

Analyzing the integral in (A.12) we have by assumption (A.1) that

$$\begin{aligned}\int_{N(j)}^{N(j+1)}\left|\frac{g(x)}{x^k}\right|\mathrm{d}x &\leq \max_{x\in[0,p]}|g(x)|\int_{jp}^{(j+1)p}\frac{1}{x^k}\,\mathrm{d}x \\ &\leq C\int_{jp}^{(j+1)p}\frac{1}{x^k}\,\mathrm{d}x \\ &= -C\frac{1}{k-1}\left[\frac{1}{x^{k-1}}\right]_{jp}^{(j+1)p} \qquad (A.13) \\ &= -C\frac{1}{(k-1)p^{k-1}}\left[\frac{1}{(j+1)^{k-1}}-\frac{1}{j^{k-1}}\right],\end{aligned}$$

where we again used the periodicity of g in the first step. We combine the results from (A.12) and (A.13) to find

$$\left|I(m)-\widetilde{I}(m)\right|\leq -C\frac{1}{m^k}\frac{1}{(k-1)p^{k-1}}\left[(j+1)^k-j^k\right]\left[\frac{1}{(j+1)^{k-1}}-\frac{1}{j^{k-1}}\right]. \qquad (A.14)$$

Elementary calculations yield

$$\begin{aligned}
-\big[(j+1)^k - j^k\big] &\left[\frac{1}{(j+1)^{k-1}} - \frac{1}{j^{k-1}}\right] \\
&= \left[\frac{1}{j^{k-1}} - \frac{1}{(j+1)^{k-1}}\right]\left[(j+1)^k - j^k\right] \\
&= (j+1)\left(\frac{j+1}{j}\right)^{k-1} - j - (j+1) + j\left(\frac{j}{j+1}\right)^{k-1} \\
&= j\left(\left(\frac{j}{j+1}\right)^{k-1} + \left(\frac{j+1}{j}\right)^{k-1}\right) - (2j+1) + \left(\frac{j+1}{j}\right)^{k-1} \\
&= j\left[\left(\frac{j}{j+1}\right)^{k-1} + \left(\frac{j+1}{j}\right)^{k-1} - 2\right] + \left[\left(\frac{j+1}{j}\right)^{k-1} - 1\right] \qquad \text{(A.15)} \\
&= j\left[\left(\frac{j}{j+1}\right)^{k-1} + \left[\left(1+\frac{1}{j}\right)^j\right]^{\frac{k-1}{j}} - 2\right] + \left[\left[\left(1+\frac{1}{j}\right)^j\right]^{\frac{k-1}{j}} - 1\right] \\
&\le j\left[1 + \exp\left(\frac{k-1}{j}\right) - 2\right] + \exp\left(\frac{k-1}{j}\right) - 1 \\
&= \left(\exp\left(\frac{k-1}{j}\right) - 1\right)(j+1),
\end{aligned}$$

where we used that $(1+1/j)^j$ converges to e from below for $j \to \infty$. By l'Hôpital's rule,

$$j \mapsto \left(\exp\left(\frac{k-1}{j}\right) - 1\right)(j+1) \to (k-1) \qquad \text{(A.16)}$$

for $j \to \infty$ from above. Therefore, there exists $j \in \mathbb{N}$ such that

$$\left(\exp\left(\frac{k-1}{j}\right) - 1\right)(j+1) \le (k-1)\max\{2, 1/C\}. \qquad \text{(A.17)}$$

Let $j \in \mathbb{N}$ be large enough such that (A.17) holds. Then

$$\begin{aligned}
|E(j)| &\le \sum_{m=j}^{\infty}\left|I(m) - \widetilde{I}(m)\right| \\
&\le C\max\{2, 1/C\}\frac{1}{p^{k-1}}\sum_{m=j}^{\infty}\frac{1}{m^k} \\
&= \frac{\max\{1, 2C\}}{p^{k-1}}\left(\zeta(k) - \sum_{m=1}^{j-1}\frac{1}{m^k}\right),
\end{aligned}$$

which proves the lemma. □

A.2 Numerical experiments

In this section, we apply the approximation method presented in Lemma A.2 onto the integration of an example function. We will investigate the approximation of the associated integral in depth to provide an intuitive understanding of the approximation technique as well as demonstrate its approximation power. Let f be given by

$$f(x) := \frac{\cos(x)(1-\cos(x))^2}{x^2} \tag{A.18}$$

and thus consider the integral

$$V = \int_0^\infty f(x)\,\mathrm{d}x. \tag{A.19}$$

Note the resemblance of function f in (A.18) with the integrands we considered in Section 3.4 of Chapter 3, for example the first integrand in expression (3.143) in Corollary 3.34. Therein, we were confronted with the challenge of computing the entries of a stiffness matrix expressed as integrals of highly oscillating functions. The integration method of Lemma A.2 has been originally developed with that application in mind and shows the relevance of the result in the context of finance and beyond.

Recall that functions of form (A.18) can be approximately integrated over a finite domain using Filon's formula, see for example Abramowitz and Stegun (2014). Our integration range, however, is (semi-)infinite which is why Filon's formula does not apply, here. The graph of f is shown in Figure A.1. With the definition of the integrand f in expression (A.18) and the setting of Lemma A.2 in mind we set

$$g(x) = \cos(x)(1-\cos(x))^2, \tag{A.20}$$
$$k = 2 \tag{A.21}$$

and g is a periodic function in the sense of Definition A.1 with period $p = 2\pi$. The periodic influence of g on f is clearly visible in Figure A.1. The oscillations that endure infinitely and the relatively slow decay lead to unsatisfactory results when the integration range is simply cut off while a high accuracy of the resulting integral value shall be reached. The integration approximation method introduced by Lemma A.2 circumvents this issue. It integrates f up to a chosen cut off point $N = jp$, $j \in \mathbb{N}$, exactly and approximates the integral value beyond that point. To that extent it approximates f for $x \in I_m := [mp, (m+1)p]$, $j \leq m \in \mathbb{N}$, by

$$f_m^j : x \mapsto f(x-(m-j)p)\frac{j^k}{m^k} = \frac{g(x)}{(x-(m-j)p)^k}\frac{j^k}{m^k}, \tag{A.22}$$

which leads to the approximation of the associated integral of f over I_m, $I(m) =$

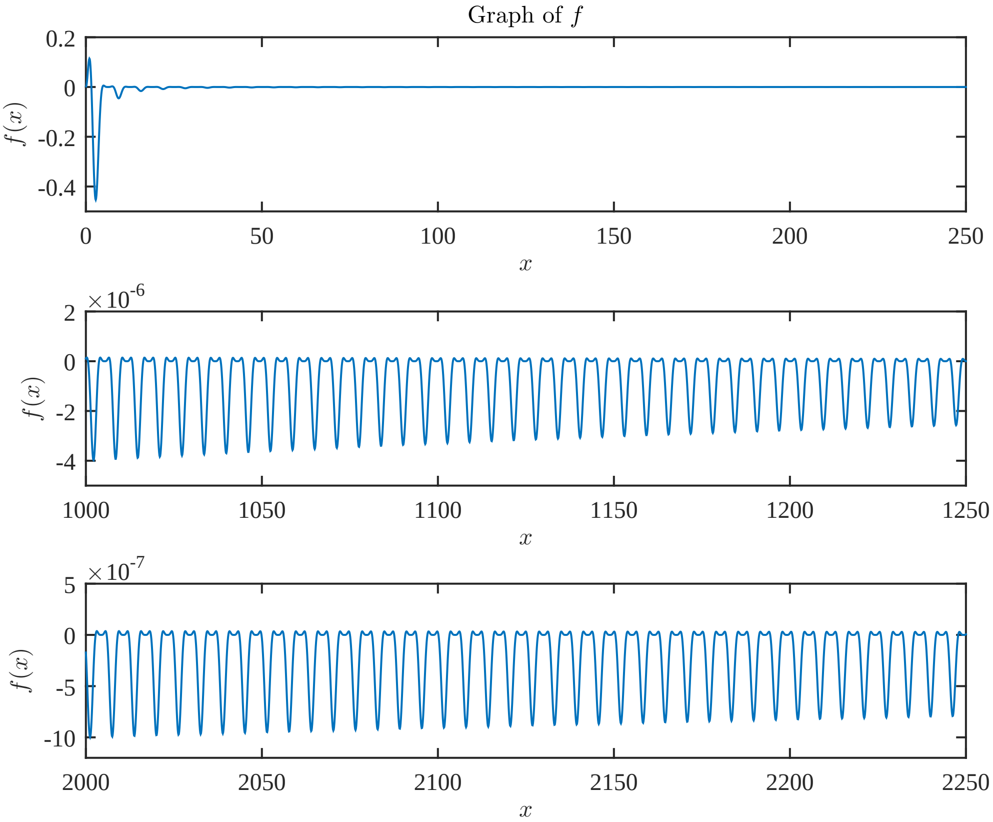

Figure A.1 Graph of example integrand f of Equation (A.18) evaluated over three different subintervals of the semi-infinite integration range. One observes the repeating structure that is exploited by the method of Lemma A.2 for the (approximative) integration of f over $\mathbb{R}^+$. The exact value of the integral is $\int_0^\infty f(x)\,\mathrm{d}x = -\frac{\pi}{4}$.

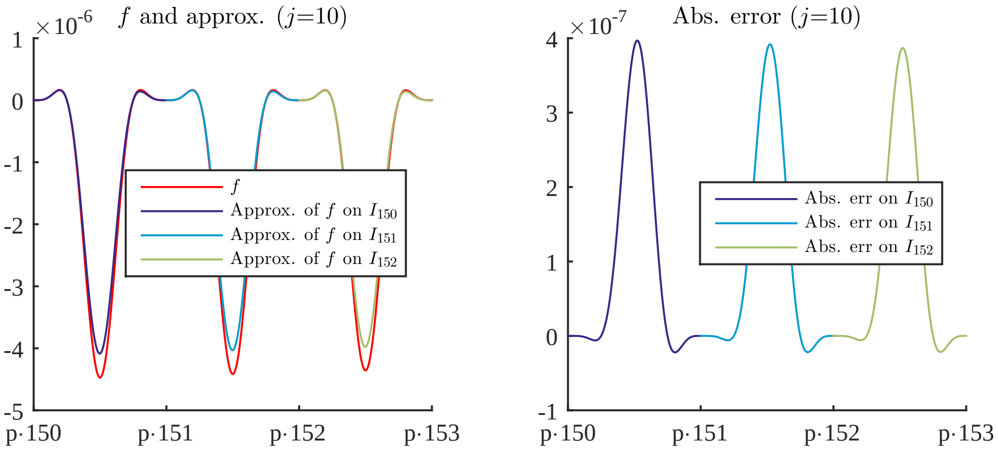

Figure A.2 Left: f evaluated over the interval $I = [p \cdot 150, p \cdot 153] = I_{150} \cup I_{151} \cup I_{152}$ and its piecewise approximation by f^j_{150}, f^j_{151} and f^j_{152}, respectively, as defined in (A.22), with a small value of $j = 10$. Right: The respective absolute error.

$\int_{I_m} f(x)\,\mathrm{d}x$, by $\widetilde{I}(m) = \int_{I_m} f^j_m(x)\,\mathrm{d}x$. Using the periodicity of g we get

$$\begin{aligned} \widetilde{I}(m) &= \int_{I_m} f^j_m(x)\,\mathrm{d}x \\ &= \int_{I_m} \frac{g(x)}{(x-(m-j)p)^k}\,\mathrm{d}x \frac{j^k}{m^k} \\ &= \int_{I_j} \frac{g(x)}{x^k}\,\mathrm{d}x \frac{j^k}{m^k} \\ &= I(j)\frac{j^k}{m^k}, \end{aligned} \tag{A.23}$$

compare the definition (A.6) at the beginning of the proof of Lemma A.2. The approximation outlined in (A.23) together with the involvement of the periodicity of g thus builds the cornerstone of the whole method. The relation between f and its approximation on I_m by f^j_m for some $j \leq m \in \mathbb{N}$ is illustrated in Figures A.2 and A.3 for both a small value of j and a large value of j. As expected, given $j_1 < j_2 \leq m$, the approximation of f on I_m by $f^{j_2}_m$ is more precise than the one provided by $f^{j_1}_m$.

Let us now turn back to the approximation of the integral $V = V(f)$ of (A.19). The exact value of this integral is $V = -\pi/4$. We choose increasing values of $j \in \mathbb{N}$ and compare the new approximation method to an integration with integration range cut off

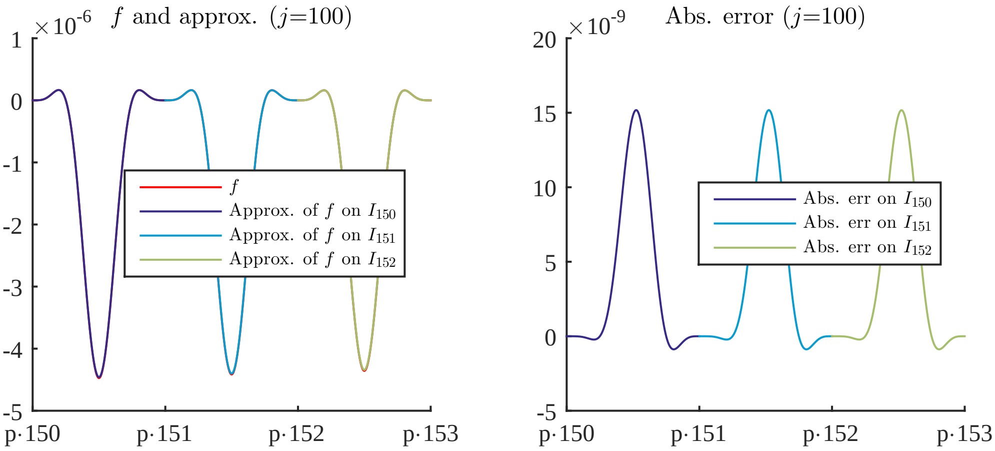

Figure A.3 Left: f evaluated over the interval $I = [p \cdot 150, p \cdot 153] = I_{150} \cup I_{151} \cup I_{152}$ and its piecewise approximation by f^j_{150}, f^j_{151} and f^j_{152}, respectively, as defined in (A.22), with a large value of $j = 100$. Right: The respective absolute error.

at the value $N(j) = pj$,

$$V^{N(j)}_{\text{cut}} := \int_0^{N(j)} f(x)\, \mathrm{d}x, \tag{A.24}$$

$$V^{N(j)}_{\text{method}} := V^{N(j)}_{\text{cut}} + V^{N(j)}_{\text{series}}, \tag{A.25}$$

with $V^{N(j)}_{\text{series}}$ defined by

$$V^{N(j)}_{\text{series}} = I(j) j^k \left(\zeta(k) - \sum_{m=1}^{j-1} \frac{1}{m^k} \right), \tag{A.26}$$

as in (A.4) of Lemma A.2.

We approximate the integral of f as given by (A.18) over $\mathbb{R}^+$ and compare the error decay of the new approximation method for increasing values of $j \in \mathbb{N}$ as described to the result yielded by simply cutting off the integration range. Figure A.4 illustrates the results.

The slow decay of f in x which we observed in Figure A.1 results in a slow decay of $V^{N(j)}_{\text{cut}}$ to the true value of $V = -\pi/4$. In contrast, exploiting the knowledge of the special structure of f containing a periodic function g allows faster convergence of the approximated integral value $V^{N(j)}_{\text{method}}$ in j.

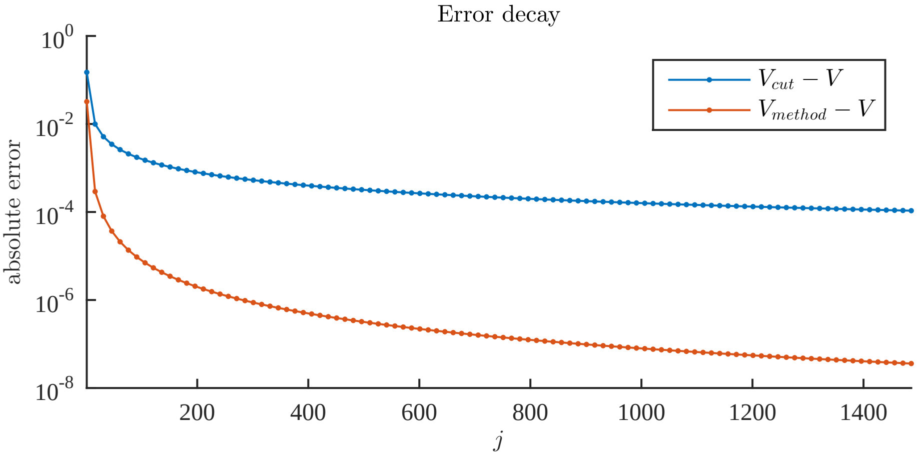

Figure A.4 Empirical study of the error decay for the approximative integration of f given by (A.18) over $\mathbb{R}^+$. We compare the results for two methods. First, we simply cut off the integration range as in (A.24) at $N(j) = pj$ for different values of j. Secondly, we approximate the function f or rather its integral *beyond* the cut off point $N(j)$ by the new approximation method as described by Lemma A.2. The results for both methods indicate a convergence to the true integral value that is $V = -\pi/4$.

B General features of magic point interpolation

We add some aspects of the magic point interpolation method from Chapter 5. The Magic Point Interpolation algorithm satisfies some immediate properties, which are identified by Barrault et al. (2004) and Maday et al. (2009) and summarised in the sequel. The content of this appendix has been presented in Gaß et al. (2015), already.

Exact interpolation at magic points For all functions $u \in \mathcal{U}$, the interpolation is exact at the magic points, in the sense that for every $m = 1, \dots, M$

$$I_m(u)(z_j^*) = u(z_j^*) \qquad \text{for all } j \leq m. \tag{B.1}$$

This property holds by construction of q_m. Note that $q_m(z_j^*) = 0$ for $j < m$.

Magic points as maxima The basis function q_m is maximal at the magic point z_m^* i.e.

$$q_m(z_m^*) = 1 = \sup_{z \in \Omega} |q_m(z)|. \tag{B.2}$$

The matrix B^M is invertible By construction, the quadratic matrix $B^M \in \mathbb{R}^{M \times M}$, introduced in (5.6) as

$$B_{jm}^M = q_m(z_j^*)$$

for all $j, m = 1, \dots, M$ is a lower triangular matrix with unity diagonal. Its inverse thus exists.

Coefficients of I_m equal to those of I_{m+1} The coefficients $\alpha_j^m = \alpha_j^m(u)$ of the interpolation $I_m(u) = \sum_{j=1}^m \alpha_j^m q_j$ of u do not depend on m, i.e. for all $i < m$ and $j \leq i$ it holds that

$$\alpha_j^m = \alpha_j^i. \tag{B.3}$$

This can be seen from the triangular structure of the defining linear system for $\alpha^m = (\alpha_j^m)_{j=1,\dots,m}$,

$$B^m \alpha^m = b^m \tag{B.4}$$

with $b_j^m = u(z_j^*)$. By this representation we also get the linearity of I_m, for all $u, v \in \mathcal{U}$,

$$I_m(u + v) = I_m(u) + I_m(v). \tag{B.5}$$

Idempotence Let $1 \leq m \leq M$. Since $I_m(v) = v$ for all $v \in \operatorname{span}\{q_1, \dots, q_m\}$ we have for all $u \in \mathcal{U}$,

$$I_m(I_{m-1}(u)) = I_{m-1}(u). \tag{B.6}$$

C Gronwall's Lemma

We provide a proof of a version of Gronwall's lemma. The lemma exists in surprisingly many shapes and forms. They all share the key idea of resolving an implicitly entangled estimate where the quantity that shall be estimated appears on both sides of the inequality. The first proof for resolving such inequalities is attributed to Thomas Hakon Gronwall who published his result in Gronwall (1919). The version of the lemma that we are interested in plays a key role in the convergence estimate at the end of Chapter 3 and guarantees the final step in the proof of the statement therein. In this appendix, we state and prove Gronwall's result in the version of Lemma C.1.

Lemma C.1 (Gronwall's Lemma)
Let $y = (y_m)_{m \geq 0}$, $f = (f_m)_{m \geq 0}$ and $g = (g_m)_{m \geq 0}$ be nonnegative sequences in $\mathbb{R}$ satisfying

$$y_m \leq f_m + \sum_{0 \leq j < m} g_j y_j, \qquad m \geq 0. \tag{C.1}$$

Then

$$y_m \leq f_m + \sum_{0 \leq j < m} \left(f_j g_j \prod_{j < i < m} (1 + g_i) \right) \tag{C.2}$$

holds for all $m \geq 0$.

We need the following auxiliary lemma for the proof of the lemma.

Lemma C.2 (Auxiliary lemma)
Let $(g_n)_{n \geq 0}$ be a nonnegative sequence in $\mathbb{R}_0^+$. Let $j \in \mathbb{N}_0$ arbitrary but fix. Then

$$1 + \sum_{j < k < m} g_k \prod_{j < i < k} (1 + g_i) = \prod_{j < i < m} (1 + g_i) \tag{C.3}$$

holds for all $m > j$.

Proof
We prove the Lemma C.2 by induction over m. The claim trivially holds for $m = j + 1$ due to an empty sum on the left and an empty product on the right. For the inductive

step we compute

$$1 + \sum_{j<k<m+1} g_k \prod_{j<i<k} (1+g_i) = \left[1 + \sum_{j<k<m} g_k \prod_{j<i<k} (1+g_i)\right] + \left[g_m \prod_{j<i<m} (1+g_i)\right]. \tag{C.4}$$

We insert the induction hypothesis into the left bracket of (C.4) to get

$$\begin{aligned}\left[1 + \sum_{j<k<m} g_k \prod_{j<i<k} (1+g_i)\right] + \left[g_m \prod_{j<i<m} (1+g_i)\right] \\ = \left[\prod_{j<i<m} (1+g_i)\right] + \left[g_m \prod_{j<i<m} (1+g_i)\right] \\ = (1+g_m) \prod_{j<i<m} (1+g_i) = \prod_{j<i<m+1} (1+g_i)\end{aligned}$$

which finishes the induction and proves the claim for all $m > j$. □

Proof (of Gronwall's Lemma C.1)
We follow the proof of Holte, J.M. (2009) and derive the claim of Lemma C.1 by induction over m. For $m = 0$ the claim trivially holds. Conducting the inductive step we get by assumption (C.1) and then by inserting the induction hypothesis that

$$\begin{aligned} y_{m+1} &\leq f_{m+1} + \sum_{0\leq j<m+1} g_j y_j \\ &\leq f_{m+1} + \sum_{0\leq j<m+1} g_j \left(f_j + \sum_{0\leq j^*<j} \left(f_{j^*} g_{j^*} \prod_{j^*<i<j} (1+g_i)\right)\right) \\ &= f_{m+1} + \sum_{0\leq j<m+1} (f_j g_j) + \sum_{0\leq j<m+1} g_j \sum_{0\leq j^*<j} \left(f_{j^*} g_{j^*} \prod_{j^*<i<j} (1+g_i)\right). \end{aligned} \tag{C.5}$$

The summands of the double sum in the last line of (C.5) are shown in Table C.1. We reorder summation in (C.5) to

$$\begin{aligned}\sum_{0\leq j<m+1} g_j \sum_{0\leq j^*<j} \left(f_{j^*} g_{j^*} \prod_{j^*<i<j} (1+g_i)\right) \\ = \sum_{0\leq j<m+1} (f_j g_j) \sum_{j<k<m+1} g_k \prod_{j<i<k} (1+g_i).\end{aligned} \tag{C.6}$$

The new summation order is illustrated by Table C.2. Consequently, combining (C.5)

j	**Summand**
0	$-$
1	$g_1[(f_0g_0)]$
2	$g_2[(f_0g_0)(1+g_1)+(f_1g_1)]$
3	$g_3[(f_0g_0)(1+g_1)(1+g_2)+(f_1g_1)(1+g_2)+(f_2g_2)]$
$\vdots$	$\vdots$
m	$g_m[(f_0g_0)(1+g_1)\dots(1+g_{m-1})+(f_1g_1)(1+g_2)\dots(1+g_{m-1})+\dots+(f_{m-1}g_{m-1})]$

Table C.1 The summands of the double sum in (C.5) before reordering.

j	**Summand**
0	$(f_0g_0)[g_1+g_2(1+g_1)+g_3(1+g_1)(1+g_2)+\dots+g_m(1+g_1)\dots(1+g_{m-1})]$
1	$(f_1g_1)[g_2+g_3(1+g_2)+\dots+g_m(1+g_2)\dots(1+g_{m-1})]$
$\vdots$	$\vdots$
$m-1$	$(f_{m-1}g_{m-1})g_m$
m	$-$

Table C.2 The summands of the double sum in (C.6) after reordering.

and (C.6) leads to

$$\begin{aligned}\sum_{0\le j<m+1}(f_jg_j)+&\sum_{0\le j<m+1}g_j\sum_{0\le j^*<j}\left(f_{j^*}g_{j^*}\prod_{j^*<i<j}(1+g_i)\right)\\&=\sum_{0\le j<m+1}(f_jg_j)\left(1+\sum_{j<k<m+1}g_k\prod_{j<i<k}(1+g_i)\right).\end{aligned}$$

Invoking the auxiliary Lemma C.2 finishes the induction and yields the claim. □

Bibliography

Abramowitz, M. and Stegun, I. (2014). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Martino Fine Books.

Achdou, Y. and Pironneau, O. (2005). *Computational Methods for Option Pricing.* Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia.

Alt, H. (2011). *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung.* Springer Lehrbuch. Springer, 6th edition.

Arendt, W., Batty, J., Hieber, M., and Neubrander, F. (2011). *Vector-valued Laplace Transforms and Cauchy Problems.* Birkhäuser, 2nd edition.

Bardi, M., Crandal, M., Evans, L., Soner, H., and Souganidis, P. (1997). *Viscosity Solutions and Applications.* Springer.

Barndorff-Nielsen, O. E. (1997). Processes of normal inverse Gaussian type. *Finance and Stochastics*, 2(1):41–68.

Barrault, M., Maday, Y., Nguyen, N. C., and Patera, A. T. (2004). An 'empirical interpolation' method: Application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathématique*, 339(9):667–672.

Bauer, H. (1992). *Maß- und Integrationstheorie.* de Gruyter.

Bernstein, S. N. (1912). Sur l'ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné. *Académie Royale de Belgique. Classe des Sciences. Mémoires*, 4:1–103.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Boyarchenko, S. I. and Levendorskiĭ, S. Z. (2002a). Barrier options and touch-and-out options under regular Lévy processes of exponential type. *Annals of Applied Probability*, 12(4):1261–1298.

Boyarchenko, S. I. and Levendorskiĭ, S. Z. (2002b). Pricing of perpetual Bermudan options. *Quantitative Finance*, 2:432–442.

Boyarchenko, S. I. and Levendorskiĭ, S. Z. (2014). Efficient variations of the Fourier transform in applications to option pricing. *Journal of Computational Finance*, 18(2):57–90.

Bracewell, R. (1999). *The Fourier Transform and its Applications.* McGraw-Hill Series in Electrical and Computer Engineering. McGraw-Hill Science/Engineering/Math, 3rd edition.

Burkovska, O., Glau, K., Gaß, M., Mahlstedt, M., Mair, M., Schoutens, W., and Wohlmuth, B. (2016). Calibration to American options: Numerical investigation of the de-Americanization method. working paper.

Cannon, J. and Browder, F. (2008). *The One-Dimensional Heat Equation*, volume 23 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press.

Carr, P., Geman, H., Madan, D. B., and Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business*, 75(2):305–332.

Carr, P. and Madan, D. B. (1999). Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2:61–73.

Chen, W. and Wang, S. (2015). A finite difference method for pricing European and American options under a geometric Lévy process. *Journal of Industrial and Management Optimization*, 11(1):241–264.

Christensen, O. (2013). *An Introduction to Frames and Riesz Bases.* Applied and Numerical Harmonic Analysis. Springer Science+Business Media.

Coclite, G., Reichmann, O., and Risebro, N. (2016). A convergent difference scheme for a class of partial integro-differential equations modeling pricing under uncertainty. *SIAM Journal on Numerical Analysis*, 54(2):588–605.

Company, R., Jódar, L., and Fakharany, M. (2013). Positive solutions of European option pricing with CGMY process models using double discretization difference schemes. *Abstract and Applied analysis*, 2013. 11 pages.

Cont, R., Lantos, N., and Pironneau, O. (2011). A reduced basis for option pricing. *SIAM Journal on Financial Mathematics*, 2(1):287–316.

Cont, R. and Voltchkova, E. (2005). A finite difference scheme for option pricing in jump diffusion and exponential Lévy models. *SIAM Journal on Numerical Analysis*, 43(4):1596–1626.

Cuchiero, C., Keller-Ressel, M., and Teichmann, J. (2012). Polynomial processes and their applications to mathematical finance. *Finance and Stochastics*, 4(16):711–740.

da Veiga, L. B., Buffa, A., Sangalli, G., and Vázquez, R. (2014). Mathematical analysis of variational isogeometric methods. *Acta Numerica*, 23:157–287.

Dang, D., Nguyen, D., and Sewell, G. (2016). Numerical schemes for pricing Asian options under state-dependent regime-switching jump-diffusion models. *Computers and Mathematics with Applications*, 71(1):443–458.

Du, Q., Tian, L., and Zhao, X. (2013). A convergent adaptive finite element algorithm for nonlocal diffusion and peridynamic models. *SIAM Journal on Numerical Analysis*, 51(2):1211–1234.

Duffie, D., Filipović, D., and Schachermayer, W. (2003). Affine processes and applications in finance. *Annals of Applied Probability*, 13(3):984–1053.

Duffie, D., Pan, J., and Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376.

Eberlein, E. and Glau, K. (2011). PIDEs for pricing European options in Lévy models – a Fourier approach. Technische Universität München, `https://mediatum.ub.tum.de/node?id=1091949`.

Eberlein, E. and Glau, K. (2014). Variational solutions of the pricing PIDEs for European options in Lévy models. *Applied Mathematical Finance*, 21(5):417–450.

Eberlein, E., Glau, K., and Papapantoleon, A. (2010). Analysis of Fourier transform valuation formulas and applications. *Applied Mathematical Finance*, 17(3):211–240.

Eberlein, E., Keller, U., and Prause, K. (1998). New insights into smile, mispricing and value at risk: The hyperbolic model. *Journal of Business*, 71(3):371–405.

Eberlein, E. and Özkan, F. (2005). The Lévy LIBOR model. *Finance and Stochastics*, 9(3):327–348.

Elstrodt, J. (2011). *Maß- und Integrationstheorie.* Springer.

Ervin, V. and Roop, J. (2007). Variational solution of fractional advection dispersion equations on bounded domains in $\mathbb{R}^d$. *Numerical Methods for Partial Differential Equations*, 23(2):256–281.

Eskin, G. I. (1981). *Boundary Value Problems for Elliptic Pseudodifferential Equations.* American Mathematical Society.

Fakharany, M., Company, R., and Jódar, L. (2016). Solving partial integro-differential option pricing problems for a wide class of infinite activity Lévy processes. *Journal of Computational and Applied Mathematics*, 296:739–752.

Fang, F. and Oosterlee, C. W. (2011). A Fourier-based valuation method for Bermudan and barrier options under Heston's model. *SIAM Journal on Financial Mathematics*, 2(1):439–463.

Feng, L. and Lin, X. (2013). Pricing Bermudan options in Lévy process models. *SIAM Journal on Financial Mathematics*, 4(1):474–493.

Feng, L. and Linetsky, V. (2008). Pricing discretely monitored barrier options and defaultable bonds in Lévy process models: A fast Hilbert transform approach. *Mathematical Finance*, 18(3):337–384.

Filipović, D., Larsson, M., and Trolle, A. (2016). Linear-rational term structure models. forthcoming in Journal of Finance; Swiss Finance Institute Research Paper No. 14-15.

Florescu, I., Mariani, M., and Sewell, G. (2014). Numerical solutions to an integro-differential parabolic problem arising in the pricing of financial options in a Lévy market. *Quantitative Finance*, 14(8):1445–1452.

Gaß, M. and Glau, K. (2014). Die PIDE-Methode. *RISIKO MANAGER*, 25:17–24.

Gaß, M. and Glau, K. (2015). Parametric integration by magic point empirical interpolation. working paper, `http://arxiv.org/abs/1511.08510`.

Gaß, M. and Glau, K. (2016). A flexible Galerkin scheme for option pricing in Lévy models. working paper, `http://arxiv.org/abs/1603.08216`.

Gaß, M., Glau, K., Mahlstedt, M., and Mair, M. (2016). Chebyshev interpolation for parametric option pricing. working paper, `http://arxiv.org/abs/1505.04648`.

Gaß, M., Glau, K., and Mair, M. (2015). Magic points in finance: Empirical interpolation for parametric option pricing. working paper, `http://arxiv.org/abs/1511.00884`.

Glau, K. (2010). *Feynman-Kac-Darstellung zur Optionspreisbewertung in Lévy-Modellen.* PhD thesis, Albert-Ludwigs-Universität Freiburg.

Glau, K. (2015). Classification of Lévy processes with parabolic Kolmogorov backward equations. forthcoming in SIAM journal Theory of Probability and its Applications.

Glau, K. (2016). Feynman-Kac formula for Lévy processes with discontinuous killing rate. *forthcoming in Finance and Stochastics.*

Gronwall, T. (1919). Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296.

Grossmann, C., Roos, H.-G., and Stynes, M. (2007). *Numerical Treatment of Partial Differential Equations.* Universitext. Springer.

Haasdonk, B., Salomon, J., and Wohlmuth, B. (2013). A reduced basis method for the simulation of American options. In *Numerical Mathematics and Advanced Applications 2011*, pages 821–829. Springer.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2):327–343.

Hilber, N., Reich, N., Schwab, C., and Winter, C. (2009). Numerical methods for Lévy processes. *Finance and Stochastics*, 13(4):471–500.

Hilber, N., Reichmann, O., Schwab, C., and Winter, C. (2013). *Computational Methods for Quantitative Finance.* Springer Finance. Springer.

Hilber, N., Schwab, C., and Winter, C. (2008). Variational sensitivity analysis of parametric Markovian market models. *Advances in Mathematics of Finance*, 83:85–106.

Holte, J.M. (2009). Discrete Gronwall lemma and applications. MAA-NCS meeting at the University of North Dakota, 24 October 2009.

Hull, J. (2015). *Options, Futures and other Derivatives.* Pearson, 9th edition.

Johnson, S. G. (2015). Saddle-point integration of C_∞ “bump” functions. working paper, `http://arxiv.org/abs/1508.04376`.

Kammler, D. W. (2007). *A first Course in Fourier Analysis.* Cambridge University Press, 2nd edition.

Karkulik, M. and Melenk, J. (2015). Local high-order regularization and applications to *hp*-methods. working paper, `http://arxiv.org/abs/1411.5209`.

Keller-Ressel, M., Papapantoleon, A., and Teichmann, J. (2013). The affine LIBOR models. *Mathematical Finance*, 23(4):627–658.

Klenke, A. (2008). *Probability Theory: A comprehensive Course.* Springer.

Kudryavtsev, O. and Levendorskiǐ, S. Z. (2009). Fast and accurate pricing of barrier options under Lévy processes. *Finance and Stochastics*, 13(4):531–562.

Laurinčikas, A. (1996). *Limit Theorems for the Riemann Zeta-Function.* Mathematics and Its Applications. Springer.

Lee, R. W. (2004). Option pricing by transform methods: Extensions, unification, and error control. *Journal of Computational Finance*, 7(3):51–86.

Levendorskiǐ, S. Z. (2012). Efficient pricing and reliable calibration in the Heston model. *International Journal of Theoretical and Applied Finance*, 15(7). 44 pages.

Levendorskiǐ, S. Z. and Xie, J. (2012). Pitfalls of the Fourier transform method in affine models, and remedies. *Journal of Computational Finance*, 15(3). 47 pages.

Lin, F. and Yang, H. (2012). A fast stationary iterative method for a partial integro-differential equation in pricing options. *Calcolo*, 50(4):313–327.

Loftin, J. (2010). Mollifiers and smooth functions. lecture notes. `http://andromeda.rutgers.edu/~loftin/ra1fal10/mollifier.pdf`.

Lord, R., Fang, F., Bervoets, F., and W., O. C. (2008). A fast and accurate FFT-based method for pricing early-exercise options under Lévy processes. *SIAM Journal on Scientific Computing*, 30(4):1678–1705.

Maday, Y., Nguyen, C. N., Patera, A. T., and Pau, G. S. H. (2009). A general multipurpose interpolation procedure: The magic points. *Communications on Pure and Applied Analysis*, 8(1):383–404.

Mastroianni, G. and Szabados, J. (1995). Jackson order of approximation by Lagrange interpolation. *Acta Mathematica Hungarica*, 69(1–2):73–82.

Matache, A.-M., Nitsche, P.-A., and Schwab, C. (2005a). Wavelet Galerkin pricing of American options on Lévy driven assets. *Quantitative Finance*, 5(4):403–424.

Matache, A.-M., Schwab, C., and Wihler, T. P. (2005b). Fast numerical solution of parabolic integrodifferential equations with applications in finance. *SIAM Journal on Scientific Computing*, 27(2):369–393.

Matache, A.-M., von Petersdorff, T., and Schwab, C. (2004). Fast deterministic pricing of options on Lévy driven assets. *Mathematical Modelling and Numerical Analysis*, 38(1):37–71.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3:125–144.

Papapantoleon, A. (2008). An introduction to Lévy processes with applications in finance. `http://arxiv.org/abs/0804.0482`. Lecture notes.

Platte, R. B. and Trefethen, N. L. (2008). Chebfun: A new kind of numerical computing. In *Progress in Industrial Mathematics at ECMI*, pages 69–86. Springer.

Prause, K. (1999). *The Generalized Hyperbolic Model: Estimation, Financial Derivatives, and Risk Measures.* PhD thesis, Albert-Ludwigs-Universität Freiburg.

Raible, S. (2000). *Lévy Processes in Finance: Theory, Numerics, and Empirical Facts.* PhD thesis, Albert-Ludwigs-Universität Freiburg.

Roop, J. (2006). Computational aspects of FEM approximation of fractional advection dispersion equations on bounded domains in $\mathbb{R}^2$. *Journal of Computational and Applied Mathematics*, 193(1):243–268.

Rudin, W. (1987). *Real and Complex Analysis.* McGraw-Hill, 3rd edition.

Rynne, B. and Youngson, M. (2000). *Linear Functional Analysis.* Springer.

Sachs, E. W. and Schu, M. (2010). Reduced order models in PIDE constrained optimization. *Control and Cybernetics*, 39(3):661–675.

Sato, K.-i. (2007). *Lévy Processes and Infinitely Divisible Distributions.* Cambridge University Press.

Schilling, R. (2005). *Measures, Integrals and Martingales.* Cambridge University Press.

Schötzau, D. (1999). *hp-DGFEM for Parabolic Evolution Problems.* PhD thesis, ETH Zürich.

Schötzau, D. and Schwab, C. (2001). *hp*-discontinuous Galerkin time-stepping for parabolic problems. *Comptes Rendues de l'Académie des Sciences - Series I - Mathematics*, 333:1121–1126.

Schoutens, W., Simons, E., and Tistaert, J. (2004). A perfect calibration! Now what? *Wilmott Magazin*, pages 66–78.

Schumaker, L. (2007). *Spline Functions: Basic Theory.* Cambridge Mathematical Library. Cambridge University Press, 3rd edition.

Seydel, R. (2012). *Tools for Computational Finance.* Universitext. Springer.

Showalter, R. E. (2010). *Hilbert Space Methods in Partial Differential Equations.* Courier Corporation.

Stein, E. and Shakarchi, R. (2003). *Fourier Analysis: An Introduction.* Princeton lectures in analysis. Princeton university press.

Stein, E. M. and Stein, J. C. (1991). Stock price distributions with stochastic volatility: An analytic approach. *Review of Financial Studies*, 4(4):727–752.

Takacs, S. and Takacs, T. (2015). Approximation error estimates and inverse inequalities for B-splines of maximum smoothness. working paper, `http://arxiv.org/abs/1502.03733`.

Trefethen, L. N. (2013). *Approximation Theory and Approximation Practice.* SIAM books.

von Petersdorff, T. and Schwab, C. (2003). Wavelet discretizations of parabolic integrodifferential equations. *SIAM Journal on Numerical Analysis*, 41(1):159–180.

Watson, G. (1995). *A Treatise on the Theory of Bessel Functions.* Cambridge University Press, 2nd edition.

Winter, C. (2009). *Wavelet Galerkin Schemes for Option Pricing in Multidimensional Lévy Models.* PhD thesis, ETH Zürich.

Wloka, J. (2002). *Partial Differential Equations.* Cambridge University Press.

Zeidler, E. (1990). *Nonlinear Functional Analysis and its Applications.* Springer. II/A. Linear Monotone Operators.

Zeng, P. and Kwok, Y. K. (2014). Pricing barrier and Bermudan style options under time-changed Lévy processes: Fast Hilbert transform approach. *SIAM Journal on Scientific Computing*, 36(3):B450–B485.

Zhylyevskyy, O. (2010). A fast Fourier transform technique for pricing American options under stochastic volatility. *Review of Derivatives Research*, 13:1–24.

Zimmermann, M. (2016). The Finite Element Method with Splines for Option Pricing. Master's thesis, Technische Universität München.