Opgedragen aan mijn ouders, Eddy Wyns & Anita Leys

Universiteit Antwerpen

Convergence Analysis and Application of ADI Schemes for Partial Differential Equations from Financial Mathematics

Proefschrift voorgelegd op 26 juni 2017 tot het behalen van de graad van Doctor in de Wetenschappen – Wiskunde,
bij de faculteit Wetenschappen aan de Universiteit Antwerpen.

Promotor:
Prof. dr. Karel in 't Hout

Maarten Wyns

Applied Mathematics
University of Antwerp

Doctoral committee

Chairman:

Prof. dr. Wim Vanroose	(University of Antwerp)

Supervisor:

Prof. dr. Karel in 't Hout	(University of Antwerp)

Other members:

Prof. dr. Ann De Schepper	(University of Antwerp)
Prof. dr. ir. Florence Guillaume	(University of Antwerp)
Prof. dr. Willem Hundsdorfer	(CWI Amsterdam)
Prof. dr. ir. Cornelis W. Oosterlee	(CWI Amsterdam)
Prof. dr. Michèle Vanmaele	(Ghent University)

Original title: *Convergence Analysis and Application of ADI Schemes for Partial Differential Equations from Financial Mathematics*

Nederlandse titel: *Convergentie-Analyse en Toepassing van ADI Schema's voor Partiële Differentiaalvergelijkingen uit de Financiële Wiskunde*

Keywords: ADI schemes, Douglas scheme, Modified Craig–Sneyd scheme, Craig–Sneyd scheme, Hundsdorfer–Verwer scheme, convection-diffusion equation, mixed spatial derivative, convergence, non-smooth initial data, Dirac delta, backward Kolmogorov equation, forward Kolmogorov equation, finite difference methods, finite volume methods, adjoint spatial discretization.

Published and distributed by Maarten Wyns. The research presented in this thesis was supported by the University Research Fund (BOF) of the University of Antwerp and by a PhD fellowship of the Research Foundation–Flanders.

Contact information

✉ Applied Mathematics, Dept. Mathematics & Computer Science
University of Antwerp (CMI), Building G3.07
Middelheimlaan 1, B–2020 Antwerp, Belgium

☎ +32 (0)3 265 38 54

wyns.maarten@gmail.com

https://www.linkedin.com/in/maarten-wyns-525b3352

Samenvatting

Convergentie-Analyse en Toepassing van ADI Schema's voor Partiële Differentiaalvergelijkingen uit de Financiële Wiskunde

In de huidige internationale financiële markten zijn opties producten die vaak verhandeld worden. Gevorderde wiskundige modellen worden gebruikt voor het bepalen van de eerlijke waarde van deze contracten, evenals hun afhankelijkheid van onderliggende variabelen en parameters. Dit leidt tot meerdimensionale tijdsafhankelijke partiële differentiaalvergelijkingen (PDVen). Voor de meerderheid van deze PDVen is er geen analytische oplossing beschikbaar en dient men gebruik te maken van numerieke methoden om hun exacte oplossing te benaderen. De methode-der-lijnen is een welgekende en veelzijdige aanpak voor het bepalen van een numerieke oplossing. Hierbij discretiseert men eerst de plaatsvariabelen, met bvb. eindige differentiemethoden, hetgeen leidt tot een groot stelsel van gewone differentiaalvergelijkingen (GDVen). In een tweede stap wordt dit zogenaamde semidiscreet-systeem numeriek opgelost aan de hand van een geschikte impliciete tijdstapmethode. Indien de PDV meerdimensionaal is, dan kan deze tweede stap erg rekenintensief zijn bij het gebruik van klassieke impliciete tijdstapmethodes.

In dit proefschrift beschouwen we de convergentie en de toepassing van vier alternerende richting (Engels: direction) impliciete (ADI) schema's voor de numerieke oplossing van semidiscrete tweedimensionale convectie-diffusievergelijkingen. Meer bepaald onderzoeken we het Douglas (Do) schema, het Craig–Sneyd (CS) schema, het aangepaste (Engels: Modified) Craig–Sneyd (MCS) schema en het Hundsdorfer–Verwer (HV) schema. ADI schema's maken gebruik van een opsplitsing van het semidiscreet-systeem in de verschillende plaatsrichtingen. Dit kan zorgen voor een aanzienlijk computationeel voordeel in iedere tijdstap aangezien het makkelijker is om de suboperatoren achtereenvolgens impliciet te behandelen, in plaats van de gehele operator in één keer. De vier ADI schema's zijn aangepast aan PDVen met gemengde afgeleiden en worden vaak gebruikt in de financiële wiskunde. In dit gebied zijn gemengde plaatsafgeleiden alomtegenwoordig vanwege correlatie tussen de onderliggende stochastische processen.

In het eerste inleidende hoofdstuk worden de niet-uniforme Cartesische roosters en de tweede-orde eindige differentieformules voorgesteld die doorheen de thesis gebruikt worden voor de plaatsdiscretisatie van tijdsafhankelijke PDVen.

In het tweede inleidende hoofdstuk worden de vier ADI schema's geïntroduceerd en wordt er een overzicht gegeven van de bestaande stabiliteits- en convergentieresultaten. Voor het Do schema leidt dit reeds tot een algemeen eerste-orde convergentieresultaat.

Het beschouwen van een verstoorde versie van het (M)CS schema, respectievelijk het HV schema, leidt tot een recursie voor de totale discretisatiefout van de verschillende schema's. Door de lokale discretisatiefout spitsvondig op te splitsen, en een lemma van Hundsdorfer toe te passen, bekomen we een tweede-orde convergentieresultaat voor het (M)CS schema en het HV schema onder enkele natuurlijke stabiliteits- en gladheidsaannames.

Bij niet-gladde beginfuncties kan toepassing van de ADI schema's leiden tot foutieve oscillaties in de numerieke oplossing. We geven een voorbeeld dat het positieve effect van Rannacher tijdstappen illustreert, dit betekent het vervangen van de eerste N_0 ADI stappen door $2N_0$ halve tijdstappen met het achterwaartse Euler schema. Voor het uitvoeren van een theoretische analyse wordt een tweedimensionale convectie-diffusievergelijking beschouwd die voorzien is van Dirac-delta begindata. We passen een Fouriertransformatie toe om aan te tonen dat, als de tijdsdiscretisatie gebeurt met het (M)CS schema, dan $N_0 = 2$ een ondergrens is voor N_0 om te verzekeren dat de numerieke oplossing convergeert naar de exacte oplossing. Uitgebreide numerieke experimenten suggereren gelijkaardige resultaten voor het Do schema en het HV schema.

Onze convergentieresultaten verantwoorden het gebruik van de ADI schema's in praktische toepassingen. In de laatste hoofdstukken introduceren we twee methoden voor het kalibreren van stochastische lokale volatiliteitsmodellen (SLV) aan hun onderliggende lokale volatiliteitsmodel (LV). De ADI schema's zijn hierbij belangrijk voor de efficiëntie van de kalibratiemethoden. De eerste kalibratiemethode is geïnspireerd door een verband tussen de voorwaartse en achterwaartse Kolmogorov-vergelijking. De plaatsdiscretisatie van de achterwaartse Kolmogorov-vergelijking wordt gebruikt voor het invoeren van een toegevoegde plaatsdiscretisatie van de voorwaartse Kolmogorov-vergelijking. Onder enkele voorwaarden kan deze toegevoegde plaatsdiscretisatie gebruikt worden om het semidiscrete SLV model exact te kalibreren aan het semidiscrete LV model. Om deze kalibratie uit te voeren dient er nog een stelsel niet-lineaire GDVen opgelost te worden. We maken gebruik van het MCS schema voor de tijdsdiscretisatie en behandelen de niet-lineariteit door middel van een iteratieprocedure. De tweede kalibratieprocedure is gebaseerd op een nieuwe eindige volume (EV) methode voor de plaatsdiscretisatie van algemene ééndimensionale en tweedimensionale voorwaartse Kolmogorov-vergelijkingen. De EV methode is massabehoudend en kan op een natuurlijke manier omgaan met de randvoorwaarden. Bovendien is er voor de nieuwe EV methode geen transformatie van de voorwaartse Kolmogorov vergelijking nodig. Dit vormt een belangrijk voordeel in vergelijking met bestaande EV methoden. Het gebruik van de EV methode voor de kalibratie van SLV modellen leidt tot een groot stelsel van niet-lineaire GDVen. Discretisatie in de tijd gebeurt aan de hand van het HV schema en de niet-lineariteit wordt opnieuw behandeld met een iteratieprocedure. Uitgebreide numerieke experimenten tonen aan dat beide kalibratieprocedures leiden tot een snelle, stabiele, en accurate kalibratie van SLV modellen aan hun onderliggende LV model.

Summary

Convergence Analysis and Application of ADI Schemes for Partial Differential Equations from Financial Mathematics

In the contemporary international financial markets option products are widely traded. Advanced mathematical models are employed for determining the fair values of these contracts as well as their sensitivities to underlying variables and parameters. This leads to multidimensional time-dependent partial differential equations (PDEs). For the majority of these PDEs there is no analytical solution available and one resorts to numerical methods for their approximate solution. A well-known and versatile approach to the numerical solution is given by the method-of-lines. The PDE is then first discretized in the spatial variables, e.g. by finite differences, leading to a large system of ordinary differential equations (ODEs). In a second step this so-called semidiscrete system is numerically solved by applying a suitable implicit time discretization method. If the PDE is multidimensional, then the latter task can be computationally intensive when classical implicit time stepping methods are used.

In this thesis we consider the convergence and application of four alternating direction implicit (ADI) time stepping schemes in the numerical solution of semidiscretized two-dimensional convection-diffusion equations. More precisely, we consider the Douglas (Do) scheme, the Craig–Sneyd (CS) scheme, the Modified Craig–Sneyd (MCS) scheme and the Hundsdorfer–Verwer (HV) scheme. ADI schemes employ a splitting of the semidiscrete PDE operator in the different spatial directions. This can lead to a major computational advantage in each time step as it turns out that the implicitness is often much easier to deal with when the suboperators are handled successively, instead of treating the full operator all at once. The four ADI schemes are adapted to mixed spatial derivative terms and widely used in financial mathematics. Mixed spatial derivatives are prominent in computational finance due to correlation between the underlying stochastic processes.

The first preliminary chapter presents the non-uniform Cartesian grids and second order finite difference schemes that are used in the thesis for the spatial discretization of time-dependent PDEs. The second preliminary chapter introduces the four ADI time stepping schemes under consideration and gives an overview of their existing stability and consistency results. For the Do scheme this already leads to a general first order convergence result.

A recursion formula for the total discretization error of the (M)CS scheme, respectively the HV scheme, has been obtained by considering a perturbed version of the respective schemes. We perform an ingenious splitting of the local discretization errors and apply a key lemma from Hundsdorfer to arrive at a second order convergence result for the (M)CS scheme and the HV scheme under natural stability and smoothness assumptions.

If the initial function is non-smooth, then application of the ADI schemes can lead to spurious erratic behaviour of the numerical solution. A motivating example illustrates the positive effect of Rannacher time stepping, i.e. replacing the first N_0 ADI time steps by $2N_0$ half-time steps of the implicit Euler scheme. We consider a model two-dimensional convection-diffusion equation with Dirac delta initial data to perform a theoretical analysis. Application of a two-dimensional Fourier transformation leads to the result that, if temporal discretization is performed with the (M)CS scheme, then $N_0 = 2$ is a lower bound on N_0 for the Rannacher time stepping in order to ensure convergence of the numerical solution to the exact solution. Based on ample numerical experiments, similar convergence results are conjectured for the Do scheme and the HV scheme.

Our convergence results provide a basis for the schemes being used in practice. We introduce two methods for the calibration of state-of-the-art stochastic local volatility (SLV) models to their underlying local volatility (LV) model. Here, the ADI schemes are important for the efficiency of the calibration algorithms. The first calibration method makes use of a relationship between the corresponding forward and backward Kolmogorov equation. The spatial discretization of the backward Kolmogorov equation is used to adopt an adjoint semidiscretization for the forward Kolmogorov equation. The latter spatial discretization allows, under some natural assumptions, to create an exact match between the semidiscrete LV model and the semidiscrete SLV model. In order to perform this calibration, a large system of non-linear ODEs needs to be solved. Time stepping is performed with the MCS scheme and an inner iteration is introduced to handle the non-linearity. For the second calibration procedure, we propose a new finite volume (FV) method for the spatial discretization of general one-dimensional and two-dimensional forward Kolmogorov equations. The FV method is mass-conservative and handles the boundary conditions in a natural way. Moreover, it does not require a transformation of the forward Kolmogorov equation, which is a major advantage in comparison with existing FV methods. Using the FV spatial discretization for the calibration of SLV models leads to a large system of non-linear ODEs. Time stepping is performed with the HV scheme and, as before, an inner iteration is used to handle the non-linearity. Ample numerical experiments show that both calibration procedures lead to a fast, stable and accurate calibration of SLV models to their underlying LV model.

Dankwoord

Het schrijven van dit dankwoord bezorgt mij een dubbel gevoel. Enerzijds ben ik blij dat het onderzoek volledig is afgerond en dat alle resultaten verwerkt zijn in dit proefschrift. Anderzijds overvalt mij het besef dat een mooie, leerrijke periode van vier jaar bijna ten einde loopt. Graag zou ik iedereen willen bedanken die dit onderzoek mogelijk heeft gemaakt en die heeft bijgedragen tot deze fantastische periode. Enkele personen waren hierbij in het bijzonder heel belangrijk. Hen zou ik dan ook expliciet willen bedanken:

Eerst en vooral zou ik mijn ouders enorm willen bedanken. Mama en papa, jullie hebben mijn broer en mij steeds op de eerste plaats gezet, ons alle mogelijke kansen gegeven, ons onvoorwaardelijk gesteund in alles wat we deden en doen, en nog zoveel meer. Maar bovenal zorgen jullie voor een warme thuis waar ik altijd graag naartoe ga, zowel in goede als mindere momenten. Bedankt om me de capaciteiten en de opvoeding te geven die het mogelijk maakten om dit proefschrift af te ronden.

Niet enkel ik, maar ook mijn broer Kenneth heeft het geluk gehad bij deze ouders te mogen opgroeien (en bijgevolg het geluk gehad om mijn broer te mogen zijn). Kenneth, bedankt om me te helpen als ik weer eens iets niet begreep op school, om me naar een hoger niveau te brengen omdat ik niet wou onderdoen voor mijn grote broer, maar vooral om steeds opnieuw klaar te staan met goede raad. Met andere woorden, bedankt om een broer te zijn die er mede voor gezorgd heeft dat ik al zo ver ben geraakt.

Misschien is dit ook meteen het gepaste moment om mij eens te excuseren aan mijn ouders en broer voor uitspraken als “Ik ga niet alles gestudeerd krijgen tegen morgen.”en “Het examen ging niet zo goed, maar waarschijnlijk ben ik er wel door.”die jullie vijf jaar moesten trotseren tijdens mijn studies aan de universiteit.

Daarnaast wil ik graag mijn grootouders en de hele familie bedanken voor hun steun en interesse in mijn onderzoek gedurende deze vier jaar.

Een zeer speciale vermelding is hier tevens op zijn plaats voor mijn vriendin, mijn verloofde, Sarah. Schattie, ondanks het feit dat ik je pas tijdens het tweede deel van mijn onderzoek leerde kennen, ben je hierbij zeer belangrijk voor mij geweest. Je zorgt niet alleen voor de leuke momenten buiten het werk, je bent er voor me als het even minder gaat, je helpt me bij het maken van

moeilijke keuzes, je steunt mij in alles wat ik doe, jij maakt me gelukkig. Je bent fantastisch en je bent al snel een van de belangrijkste personen in mijn leven geworden!

Verder zou ik graag mijn vrienden willen bedanken. Jullie zorgden vaak voor ontspanning tijdens dit onderzoek, en dat kan in sommige gevallen letterlijk genomen worden wanneer ik weeral mijn lach niet kon inhouden tijdens het werk omdat er bepaalde berichten op het scherm verschenen. Bedankt om regelmatig, al dan niet oprecht geïnteresseerd, te vragen wat het onderwerp van mijn onderzoek was en hoe het vorderde.

Uiteraard dienen er ook op de Universiteit Antwerpen enkele mensen bedankt te worden. Allereerst mijn promotor Karel in 't Hout, die me zowel begeleidde voor mijn Masterthesis als voor dit proefschrift. Bedankt om me in contact te brengen met dit onderwerp en dit onderzoek mogelijk te maken. Zonder u was de projectaanvraag nooit aanvaard en was dit proefschrift er dus nooit gekomen. Later heeft u me laten kennismaken met toponderzoek(ers) van over de hele wereld. Deze ervaringen hebben samen met uw uitgebreide kennis dit onderzoek naar een hoger niveau getild. Verder stond u altijd klaar voor een gesprek over onderzoek, maar ook over andere onderwerpen zoals sport, reizen of privé zaken. Bedankt om me gedurende deze vijf jaar, na onze eerste afspraak voor de Masterthesis, te voorzien van een ruime bagage die ongetwijfeld nog vaak van pas zal komen in de toekomst.

Daarnaast zou ik alle collega's van het departement Wiskunde–Informatica willen bedanken. Door hen ben ik steeds met veel plezier naar bureau gekomen. Graag zou ik hierbij mijn bureaugenoten expliciet willen vermelden. Radoslav, Lynn en Koen, bedankt om M.G.307 met mij te delen en steeds klaar te staan voor een ontspannende babbel.

Tot slot welgemeende dank aan alle belastingbetalers die al dan niet met veel plezier een deel van hun inkomen afstonden aan de overheid om dit onderzoek te financieren.

Maarten Wyns
Antwerpen, 2017

Whakawhetai ki a koutou nui atu!

Contents

List of Abbreviations

PDE	Partial Differential Equation
MOL	Method-Of-Lines
FD	Finite Difference
ODE	Ordinary Differential Equation
ADI	Alternating Direction Implicit
Do	Douglas
CS	Craig–Sneyd
MCS	Modified Craig–Sneyd
HV	Hundsdorfer–Verwer
SLV	Stochastic Local Volatility
FX	Foreign Exchange
LV	Local Volatility
SV	Stochastic Volatility
SDE	Stochastic Differential Equation
FaV	Fair Value
FV	Finite Volume
CIR	Cox–Ingersoll–Ross

List of Tables

List of Figures

CHAPTER 1

Introduction and Outline of the Thesis

In the contemporary international financial markets option products are widely traded. In addition to standard European call and put options, a broad range of exotic derivatives exists. The primary goal of financial mathematics consists of determining the fair values of these financial contracts as well as their sensitivities to underlying variables and parameters, which are crucial for hedging. To this purpose advanced mathematical models are employed nowadays, yielding initial-boundary value problems for multidimensional time-dependent *partial differential equations* (PDEs), see e.g. [24, 27, 32, 49, 50, 64]. These PDEs are in general of the convection-diffusion kind. An example of a time-dependent convection-diffusion equation in one spatial dimension is given by

$$u_t(x,t) + a(x,t)u_x(x,t) = d(x,t)u_{xx}(x,t), \tag{1.0.1}$$

for $x \in \Omega \subset \mathbb{R}$, $t > 0$. Here a and d are assumed to be given real functions with d positive. In some cases closed-form analytical formulas for the exact solutions have been obtained in the literature. For the majority of option valuation PDEs, however, such formulas are not available. In view of this, one resorts to numerical methods for their approximate solution. For banks and other financial institutions, the fast, accurate and stable numerical approximation of option values and their sensitivities is of paramount importance.

A well-known and versatile approach to the effective numerical solution of time-dependent convection-diffusion equations is given by the *method-of-lines* (MOL), cf. e.g. [35]. It consists of two general, consecutive steps. In the first step, the PDE is discretized in the spatial variables, e.g. by finite difference, finite volume or (Galerkin) finite element methods. This leads to a so-called *semidiscrete system* of *ordinary differential equations* (ODEs). In the second step the obtained semidiscrete system is numerically solved by applying a suitable, implicit time-discretization method. If the PDE is multidimensional, then the latter task can be computationally very intensive when classical implicit methods, such as the Crank-Nicolson scheme, are applied. In the recent years, a variety of operator splitting methods have been developed that enable, in principle, a highly efficient and stable numerical solution of semidiscretized multidimensional PDEs that arise in financial mathematics, see e.g. [35–37].

A prominent class of operator splitting methods are the *Alternating Direction Implicit* (ADI) schemes. ADI schemes employ a splitting of the semidiscrete PDE operator in the different spatial dimensions. This can lead to a

major computational advantage in each time step as it turns out that the implicitness is often much easier to deal with when the suboperators are handled successively, instead of treating the full operator all at once. ADI schemes form state of the art time-discretization methods in contemporary financial mathematics, see e.g. [3, 8, 37, 47, 49]. They were successfully used already in several other application areas in science and engineering. However, the PDEs of financial mathematics are often of new types or have features that were studied only marginally in other areas. This raises new, important questions about their applicability and their fundamental properties such as stability and convergence. Notably, the Brownian motions in the underlying asset price processes are almost always correlated. This gives rise to *mixed spatial derivative terms*. Only recently have ADI schemes been adapted to such situations, see [11,41,42], and has a stability analysis been performed in a structured way. Before the start of this thesis, however, there were little or no theoretical convergence results available for the pertinent ADI methods. Next to mixed derivative terms, *non-smooth initial data* constitute a common feature in financial applications. For example, this data can be given by the option's payoff functions, which are in general only piecewise differentiable, or by the Dirac delta function. It is well-known that convergence can then be seriously impaired, cf. e.g. [22, 56].

In this thesis a convergence analysis is presented for four ADI methods adapted to mixed spatial derivative terms that are widely used in computational finance: the Douglas (Do) scheme, the Craig–Sneyd (CS) scheme, the Modified Craig–Sneyd (MCS) scheme and the Hundsdorfer–Verwer (HV) scheme. The results are directly relevant to semidiscretized two-dimensional convection-diffusion equations from financial mathematics. Next to this, we show the applicability of ADI schemes in the calibration of state-of-the-art stochastic local volatility (SLV) models. Two different PDE-based numerical methods are proposed where the ADI schemes contribute to a fast, stable and accurate calibration.

Spatial discretization by finite differences on non-uniform Cartesian grids is widely considered for the numerical solution of initial-boundary value problems for PDEs stemming from financial mathematics. In the preliminary Chapter 2 we present a short overview of the Cartesian grids and finite difference schemes that are used in the thesis. The pertinent smooth, non-uniform meshes and second order finite difference formulas are well-known in computational finance and have already been studied extensively in the literature, see e.g. [35, 37, 65]. Spatial discretization leads to a large system of ODEs. We show that, if the semidiscrete system is stemming from a two-dimensional PDE, then application of standard implicit time stepping methods to this system of ODEs can lead to a number of operations in each time step that grows faster than the total number of spatial grid points.

ADI time stepping methods employ a splitting of the semidiscrete operator into suboperators that correspond with the spatial derivatives in the different spatial directions. In the preliminary Chapter 3 we illustrate the favourable result that implicit time steps with the suboperators, that do not correspond with a mixed spatial derivative term, often require a number of operations that is directly proportional to the total number of grid points. We introduce the

four ADI schemes that are considered in this thesis, and give an overview of the existing stability and consistency results relevant to semidiscretized two-dimensional convection-diffusion equations with mixed derivative term. For the Do scheme this already leads to a convergence result, cf. [34].

Starting from Chapter 4, we present our novel research results on the convergence and application of ADI schemes. The majority of these results has been published in the international scientific literature, see [43, 44, 68–70].

Chapter 4 deals with the convergence of the MCS scheme when it is applied to a general semidiscrete system stemming from spatial discretization of a two-dimensional convection-diffusion equation with mixed derivative term provided with smooth initial and boundary data. We consider a perturbed version of the MCS scheme and obtain a recursion formula for the resulting total error. Under natural smoothness assumptions, a Taylor expansion yields useful expressions for the local errors in the perturbed scheme. By inserting the latter expressions in the recursion formula, and by applying a key lemma from Hundsdorfer [33], we arrive at a second order convergence theorem for the MCS scheme. Application of the lemma requires several stability assumptions. Positive theoretical results on these stability assumptions are obtained in the von Neumann framework.

Our convergence analysis for the HV scheme in Chapter 5 is completely analogous to that for the MCS scheme in Chapter 4. We prove that, under natural stability and smoothness assumptions, the HV scheme is second order convergent in the application to two-dimensional time-dependent convection-diffusion equations with mixed derivative term. As before, the stability assumptions are analysed theoretically in a model framework and ample numerical experiments confirm our convergence result.

PDEs from financial mathematics are often provided with non-smooth initial functions. It is well-known that application of ADI time stepping methods can then lead to spurious erratic behaviour of the numerical solution. Chapter 6 deals with the influence of Rannacher time stepping, i.e. replacing the first steps of the ADI scheme by several (sub)steps of the implicit Euler scheme, on the order of convergence of the ADI schemes when they are applied to a model two-dimensional convection-diffusion equation with mixed derivative term, provided with Dirac delta initial data. We introduce a discrete/continuous Fourier transformation pair and show that the total discretization error can be written as the sum of a low-wavenumber error and a high-wavenumber error. By performing an asymptotic analysis in Fourier space, we prove a theoretical convergence theorem for the MCS scheme and the CS scheme. Extensive numerical experiments lead to a conjecture for the Do scheme, respectively the HV scheme. In general, for this model PDE provided with Dirac delta initial data, we observe that the first two ADI time steps have to be replaced by four half-time steps of the implicit Euler scheme in order to ensure convergence of the numerical solution to the exact solution.

In Chapter 7 and Chapter 8 the ADI schemes are applied for the calibration of SLV models. Although the ADI schemes are important for the efficiency of the calibration procedures, their application is not the main contribution of the pertinent chapters.

Calibration of SLV models to the underlying local volatility (LV) model is

a highly non-trivial task. In general, there is no closed-form solution available for the fair value of vanilla options under the (S)LV model and it is difficult to check whether the SLV model is calibrated "perfectly" to the LV model. Our calibration method in Chapter 7 aims at matching the numerical approximation of the fair value of non-path-dependent options under the LV model with the corresponding approximation of the fair value under the SLV model. By introducing an adjoint spatial discretization, we prove that the semidiscretized SLV model can be calibrated exactly to the semidiscretized LV model whenever similar spatial discretization methods are used. The calibration technique leads to a large system of non-linear ODEs. We employ the MCS scheme for the efficient temporal discretization of this system of ODEs and describe an iteration procedure for handling the non-linearity.

Numerically solving the forward Kolmogorov equation corresponding to an SLV model is a common approach for its calibration to the underlying LV model. The solutions of forward Kolmogorov equations represent density functions and conservation of mass is a key property. In Chapter 8 we propose a new finite volume (FV) spatial discretization method that is mass-conservative and that does not require a transformation of the PDE. The latter property forms a major advantage in comparison with existing FV methods, since in practical applications the PDE coefficients are often non-smooth. Applying the FV discretization for the calibration of SLV models results in a large systems of non-linear ODEs. The HV scheme is employed for the efficient temporal discretization and, as in Chapter 7, an inner iteration is used for handling the non-linearity. We conclude Chapter 8 by comparing our adjoint technique and the FV method for the calibration of SLV models.

The final Chapter 9 summarizes our main results and conclusions, and gives an outlook for future research.

CHAPTER 2

Spatial Discretization of PDEs from Finance

2.1. Introduction

In the first step of the MOL, the pertinent PDE is discretized in the spatial variables. The spatial variables of time-dependent PDEs from finance often represent asset prices, volatilities or they are the result of a continuous transformation of such quantities. The resulting spatial domains are rectangular and easy to discretize. Moreover, in financial applications the important regions, i.e. *regions of interest*, where the solution is of main interest or where it is non-smooth, are often known in advance. Semidiscretization by *finite differences* (FD) on fixed *Cartesian grids* then forms a natural candidate for the spatial discretization of such PDEs. This approach is widely used and generally adopted in computational finance, see e.g. [49, 65]. In this chapter we present a short overview of the Cartesian grids and FD schemes that are used in the thesis.

The chapter is structured as follows. In Section 2.2 it is shown that uniform Cartesian grids often lead to an excess of grid points. The use of non-uniform meshes can be beneficial. We introduce a general approach for the construction of smooth non-uniform grids. In Section 2.3 several well-known FD formulas are presented for the approximation of the first and second derivative. We illustrate that semidiscretization of one-dimensional and two-dimensional time-dependent convection-diffusion equations by the pertinent FD schemes, leads to large systems of ODEs. An analysis of the sparsity structure of the semidiscretization matrix reveals that, if the semidiscrete system is stemming from a two-dimensional PDE, then application of standard implicit time stepping methods to this system of ODEs can be computationally very intensive.

2.2. Cartesian Grids

Since the spatial domains of PDEs from financial mathematics are often rectangular, the use of Cartesian spatial grids is very natural [35]. This type of grid is constructed by defining meshes in each of the spatial directions. The multidimensional spatial grid consists of the Cartesian product of the one-dimensional meshes. For example, let $\Omega = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}] \subset \mathbb{R}^2$ be a rectangular

domain and let

$$x_1 < x_2 < \cdots < x_{m_1}, \quad \text{respectively} \quad y_1 < y_2 < \cdots < y_{m_2},$$

be a mesh in the interval $[x_{\min}, x_{\max}]$, respectively $[y_{\min}, y_{\max}]$. The corresponding Cartesian grid representing the full domain Ω is then given by

$$(x_j, y_k) \quad \text{for } 1 \leq j \leq m_1, 1 \leq k \leq m_2.$$

Uniform Cartesian grids are the most simple grids. Here, the underlying one-dimensional meshes all have a uniform mesh width. When working with uniform Cartesian grids, the uniform mesh widths are limited in the different spatial directions by the accuracy requirements in each part of the spatial domain. In practical applications, however, one often wants a higher accuracy in a region of interest than in the rest of the domain. Moreover, non-smooth behaviour of the solution can lead to a loss of accuracy of spatial discretization methods, see e.g. [56, 65]. The use of uniform Cartesian grids then leads to an excess of grid points, e.g. in regions where the solution is smooth and a low accuracy is sufficient. By introducing non-uniform underlying meshes, the number of grid points can be reduced whilst the accuracy is as required in each part of the spatial domain. For spatial discretization methods (cf. Section 2.3) it is often important that the *meshes* are *smooth*. Let

$$\Delta x_j = x_j - x_{j-1}, \quad \text{for } 2 \leq j \leq m_1$$

denote spatial mesh widths. We say that the corresponding one-dimensional mesh is smooth if there exist strictly positive constants C_0, C_1, C_2 such that

$$C_0 \Delta x \leq \Delta x_j \leq C_1 \Delta x \quad \text{and} \quad |\Delta x_{j+1} - \Delta x_j| \leq C_2 (\Delta x)^2 \tag{2.2.1}$$

uniformly in j and m_1, and where Δx denotes a maximal mesh width, cf. [35, 37]. In the next subsection a technique is introduced to create smooth meshes that have a small uniform mesh width inside a given interval and larger mesh widths outside this interval, cf. [24, 25].

2.2.1. Construction of Smooth Non-Uniform Grids

In general, non-uniform meshes are not smooth in the sense of (2.2.1). This property, however, can be important for the performance of spatial discretization methods. A well-known technique to construct smooth non-uniform grids consists of defining the mesh by a smooth transformation of a uniform underlying grid, see e.g. [37, 65]. Based on the meshes defined in [24, 25] we introduce a smooth grid

$$x_{\min} = x_1 < x_2 < \cdots < x_{m_1} = x_{\max} \tag{2.2.2}$$

in the interval $[x_{\min}, x_{\max}]$ that has a small uniform mesh width within a subinterval $[x_{\text{left}}, x_{\text{right}}] \subset [x_{\min}, x_{\max}]$ and increasing mesh widths outside the subinterval.

Let integer $m_1 \geq 2$ denote the number of mesh points for the discretization of the interval $[x_{\min}, x_{\max}]$ and let $d_1 > 0$ be a parameter that controls the

fraction of points that lie inside the subinterval $[x_{\text{left}}, x_{\text{right}}]$. Next, define equidistant points $\zeta_{\min} = \zeta_1 < \zeta_2 < \cdots < \zeta_{m_1} = \zeta_{\max}$ where

$$\zeta_{\min} = \sinh^{-1}\left(\frac{x_{\min} - x_{\text{left}}}{d_1}\right),$$

$$\zeta_{\text{int}} = \frac{x_{\text{right}} - x_{\text{left}}}{d_1},$$

$$\zeta_{\max} = \zeta_{\text{int}} + \sinh^{-1}\left(\frac{x_{\max} - x_{\text{right}}}{d_1}\right).$$

It follows that $\zeta_{\min} \leq 0 \leq \zeta_{\text{int}} \leq \zeta_{\max}$. A mesh (2.2.2) that satisfies the above properties is then defined through the transformation

$$x_j = \Psi(\zeta_j), \quad \text{for } 1 \leq j \leq m_1, \tag{2.2.3}$$

where

$$\Psi(\zeta) = \begin{cases} x_{\text{left}} + d_1 \sinh(\zeta), & \text{for } \zeta_{\min} \leq \zeta \leq 0, \\ x_{\text{left}} + d_1 \zeta, & \text{for } 0 \leq \zeta \leq \zeta_{\text{int}}, \\ x_{\text{right}} + d_1 \sinh(\zeta - \zeta_{\text{int}}), & \text{for } \zeta_{\text{int}} \leq \zeta \leq \zeta_{\max}. \end{cases}$$

It is readily seen that the above mesh is smooth in the sense of (2.2.1).

For some applications, e.g. the discretization of a *Dirac delta* function, it is important to have a specific point included in the mesh. This can be obtained easily by slightly changing the underlying uniform grid, cf. [8]. Suppose for example that the point $X_0 \in (x_{\min}, x_{\max})$ needs to coincide with a grid point. Let j_0 be the index such that ζ_{j_0} is the grid point closest to $\Psi^{-1}(X_0)$, define $\tilde{\zeta}_{j_0}$ by $\Psi^{-1}(X_0)$ and let f be a piecewise linear interpolant corresponding to the points $\{(\zeta_{\min}, \zeta_{\min}), (\zeta_{j_0}, \tilde{\zeta}_{j_0}), (\zeta_{\max}, \zeta_{\max})\}$. If we define $\tilde{x}_j$ by $\Psi(f(\zeta_j))$, then the new grid $\tilde{x}_j, 1 \leq j \leq m_1$, is similar to (2.2.3) and contains the value X_0. Since the point X_0 is assumed not to be at the boundary, it can be shown that the new grid is also smooth in the sense of (2.2.1).

2.3. Finite Difference Discretization

In this thesis, discretization of the spatial derivatives is mainly performed by finite differences. Once the Cartesian grid is defined, the exact solution of the PDE is approximated at the grid points by replacing all the spatial derivatives at the nodes by finite differences. Spatial discretization by FD on Cartesian grids has already been studied extensively in the literature, see e.g. [35, 62], and is widely considered for the numerical solution of initial-boundary value problems for time-dependent PDEs stemming from financial mathematics, see e.g. [8, 37, 49, 64, 65]. Dependent on the properties of the PDE and the desired properties of the numerical approximations, different FD formulas can be applied. For financial applications it is common to consider first and second order FD schemes, cf. the above references.

2.3.1. Finite Difference Formulas

Semidiscretization on uniform grids leads to standard first and second order FD formulas. Recently, however, non-uniform spatial grids have been introduced in computational finance, see e.g. [24,25,37,65], and the second order FD formulas then need to be generalised. Let $f : \mathbb{R} \to \mathbb{R}$ be any given function, let x_j, $j \in \mathbb{Z}$, be any given increasing sequence of mesh points and let $\Delta x_j = x_j - x_{j-1}$ for all j. In order to approximate the first derivative of f, we consider three FD formulas:

$$f'(x_j) \approx \alpha_{j,-2} f(x_{j-2}) + \alpha_{j,-1} f(x_{j-1}) + \alpha_{j,0} f(x_j), \tag{2.3.1a}$$
$$f'(x_j) \approx \beta_{j,-1} f(x_{j-1}) + \beta_{j,0} f(x_j) + \beta_{j,1} f(x_{j+1}), \tag{2.3.1b}$$
$$f'(x_j) \approx \gamma_{j,0} f(x_j) + \gamma_{j,1} f(x_{j+1}) + \gamma_{j,2} f(x_{j+2}), \tag{2.3.1c}$$

with coefficients given by

$$\alpha_{j,-2} = \tfrac{\Delta x_j}{\Delta x_{j-1}(\Delta x_{j-1}+\Delta x_j)},\ \alpha_{j,-1} = \tfrac{-\Delta x_{j-1}-\Delta x_j}{\Delta x_{j-1}\Delta x_j},\ \alpha_{j,0} = \tfrac{\Delta x_{j-1}+2\Delta x_j}{\Delta x_j(\Delta x_{j-1}+\Delta x_j)},$$

$$\beta_{j,-1} = \tfrac{-\Delta x_{j+1}}{\Delta x_j(\Delta x_j+\Delta x_{j+1})},\quad \beta_{j,0} = \tfrac{\Delta x_{j+1}-\Delta x_j}{\Delta x_j\Delta x_{j+1}},\quad \beta_{j,1} = \tfrac{\Delta x_j}{\Delta x_{j+1}(\Delta x_j+\Delta x_{j+1})},$$

$$\gamma_{j,0} = \tfrac{-2\Delta x_{j+1}-\Delta x_{j+2}}{\Delta x_{j+1}(\Delta x_{j+1}+\Delta x_{j+2})},\ \gamma_{j,1} = \tfrac{\Delta x_{j+1}+\Delta x_{j+2}}{\Delta x_{j+1}\Delta x_{j+2}},\quad \gamma_{j,2} = \tfrac{-\Delta x_{j+1}}{\Delta x_{j+2}(\Delta x_{j+1}+\Delta x_{j+2})}.$$

To approximate the second derivative $f''(x_j)$, we employ the central finite difference scheme

$$f''(x_j) \approx \delta_{j,-1} f(x_{j-1}) + \delta_{j,0} f(x_j) + \delta_{j,1} f(x_{j+1}), \tag{2.3.2}$$

where

$$\delta_{j,-1} = \tfrac{2}{\Delta x_j(\Delta x_j+\Delta x_{j+1})},\quad \delta_{j,0} = \tfrac{-2}{\Delta x_j\Delta x_{j+1}},\quad \delta_{j,1} = \tfrac{2}{\Delta x_{j+1}(\Delta x_j+\Delta x_{j+1})}.$$

For a function of two variables $f : \mathbb{R}^2 \to \mathbb{R}$, the mixed derivative f_{xy} is approximated by application of (2.3.1b) successively in the two directions.

The finite difference schemes above are all well-known in the literature. Formula (2.3.1a), respectively (2.3.1b), (2.3.1c), is called the second order backward, respectively second order central, second order forward, formula for the first derivative. Finite difference scheme (2.3.2) is called the second order central formula for the second derivative. Through Taylor expansion it can be verified that each of the finite difference approximations above has a second order truncation error, provided that the function f is sufficiently often continuously differentiable and the mesh is smooth in the sense of (2.2.1).

2.3.2. Spatial Discretization of One-Dimensional PDEs

The non-uniform grids from Subsection 2.2.1 can be used in combination with the FD schemes (2.3.1), (2.3.2) to define approximations of the exact solution of time-dependent convection-diffusion equations. In this subsection we illustrate FD discretization for the general one-dimensional PDE (1.0.1) where the spatial domain $\Omega = (x_{\min}, x_{\max})$ is assumed to be a finite interval. It is assumed

that $u(x,0) = u_0(x)$ with given initial function u_0, and that conditions on the boundary of Ω are defined. For some (financial) applications the spatial domain can be unbounded, e.g. if x represents an asset price (or the logarithm of an asset price) and u represents the fair value of a non-path-dependent option. Then, to make the semidiscretization feasible, the spatial domain needs to be truncated and complimentary boundary conditions have to be imposed.

In a first step, the spatial domain is discretized by choosing grid points

$$x_{\min} = x_1 < x_2 < \cdots < x_{m_1} = x_{\max}.$$

Next, using the FD formulas from the previous subsection leads to approximations $U_j(t)$ of the exact solution $u(x_j,t)$, $1 \le j \le m_1$. For example, if the second order central FD scheme is considered for both the convection and diffusion part, then the approximations are defined by the ordinary differential equation

$$\begin{aligned} U_j'(t) = {} & d(x_j,t)\left(\delta_{j,-1}U_{j-1}(t) + \delta_{j,0}U_j(t) + \delta_{j,1}U_{j+1}(t)\right) \\ & - a(x_j,t)\left(\beta_{j,-1}U_{j-1}(t) + \beta_{j,0}U_j(t) + \beta_{j,1}U_{j+1}(t)\right), \end{aligned} \tag{2.3.3}$$

for $1 < j < m_1$, $t > 0$. The initial values $U_j(0)$ are given by the function values $u_0(x_j)$. In order to complete the above system of ODEs, the boundary conditions are used to define $U_1(t)$ and $U_{m_1}(t)$. For example, if a Dirichlet boundary condition $u(x_{\min},t) = u_{\min}(t)$ holds at the lower boundary, with given function $u_{\min}$, then one can put $U_1(t) = u_{\min}(t)$. In financial applications, the *linear boundary condition* is widely considered. Suppose that the linear boundary condition $u_{xx}(x_{\max},t) = 0$ is imposed at the upper boundary. Let $x_{m_1+1} = x_{m_1} + \Delta x_{m_1}$ be a virtual point and define the value $U_{m_1+1}(t)$ via

$$\delta_{m_1,-1}U_{m_1-1}(t) + \delta_{m_1,0}U_{m_1}(t) + \delta_{m_1,1}U_{m_1+1}(t) = 0,$$

which can be viewed as a discrete equivalent of the linear boundary condition. Then, one can use the standard second order FD schemes at x_{m_1} such that

$$\begin{aligned} U_{m_1}'(t) &= -a(x_{m_1},t)\left(\beta_{m_1,-1}U_{m_1-1}(t) + \beta_{m_1,0}U_{m_1}(t) + \beta_{m_1,1}U_{m_1+1}(t)\right) \\ &\quad + d(x_{m_1},t)\left(\delta_{m_1,-1}U_{m_1-1}(t) + \delta_{m_1,0}U_{m_1}(t) + \delta_{m_1,1}U_{m_1+1}(t)\right) \\ &= -a(x_{m_1},t)\left(-\tfrac{1}{\Delta x_{m_1}}U_{m_1-1}(t) + \tfrac{1}{\Delta x_{m_1}}U_{m_1}(t)\right). \end{aligned}$$

It is readily seen that this discretization of the linear boundary condition corresponds with putting the second derivative u_{xx} equal to zero and using the first-order backward FD formula for the first derivative u_x.

Let $U(t)$ be the vector containing all the approximations $U_j(t)$, $1 \le j \le m_1$. The FD discretization can then be written as a system of ODEs

$$U'(t) = A(t)U(t) + g(t), \tag{2.3.4}$$

for $t > 0$, with given matrices $A(t)$ and vectors $g(t)$ where the latter ones contain the information about the boundary conditions. The initial vector $U(0)$ is defined via the initial function u_0. Let $\boldsymbol{x}$ be the vector containing all the grid points x_j and denote by diag$[\cdot]$ the operator that turns a vector into

a diagonal matrix with diagonal entries given by the elements of the vector. For a function $f : \mathbb{R}^2 \to \mathbb{R}$, define by $f(\boldsymbol{x}, t)$ a vector with elements $f(x_j, t)$, $1 \le j \le m_1$. The matrix $A(t)$ can then be written as

$$A(t) = \mathrm{diag}[d(\boldsymbol{x}, t)]D_{xx} - \mathrm{diag}[a(\boldsymbol{x}, t)]D_x,$$

where D_x, respectively D_{xx}, denotes the matrix that corresponds to the discretization of the first, respectively second derivative. This type of expression is useful for the FD discretization of multidimensional convection-diffusion equations.

2

2.3.3. Spatial Discretization of Two-Dimensional PDEs

Finite difference discretization of multidimensional time-dependent convection-diffusion equations can be performed similarly to the FD discretization of one-dimensional PDEs. In view of the second step of the MOL, i.e. time discretization, it is important that the resulting semidiscrete system can be written in a convenient form. In the following this is illustrated for the general two-dimensional time-dependent convection-diffusion equation

$$u_t + a_1 u_x + a_2 u_y = d_{11} u_{xx} + 2d_{12} u_{xy} + d_{22} u_{yy}, \tag{2.3.5}$$

for $(x, y) \in \Omega \subset \mathbb{R}^2$, $t > 0$. The spatial domain $\Omega = (x_{\min}, x_{\max}) \times (y_{\min}, y_{\max})$ is assumed to be a finite rectangle and $a_1, a_2, d_{11}, d_{12}, d_{22}$ are given real functions of x, y and t such that

$$d_{11} \ge 0, \quad d_{22} \ge 0, \quad d_{12}^2 \le \gamma d_{11} d_{22}, \tag{2.3.6}$$

with $0 \le \gamma \le 1$. It is assumed that $u(x, y, 0) = u_0(x, y)$ with given initial function u_0, and that boundary conditions are defined.

The semidiscretization is initiated by constructing meshes in both spatial directions

$$x_{\min} = x_1 < x_2 < \cdots < x_{m_1} = x_{\max},$$
$$y_{\min} = y_1 < y_2 < \cdots < y_{m_2} = y_{\max},$$

and defining the full Cartesian grid as (x_j, y_k), $1 \le j \le m_1, 1 \le k \le m_2$. Application of the second order central FD schemes leads to approximations $\boldsymbol{U}_{j,k}(t)$ of $u(x_j, y_k, t)$ for $1 < j < m_1, 1 < k < m_2$. Recall that the mixed spatial derivative u_{xy} is approximated by applying the second order central FD scheme (2.3.1b) successively in the x- and y-direction. Approximations $\boldsymbol{U}_{j,k}(t)$ of $u(x_j, y_k, t)$ at the boundary points are obtained by using the boundary conditions.

Let $\boldsymbol{U}(t)$ be the matrix with entries $\boldsymbol{U}_{j,k}(t)$, and denote by $\boldsymbol{x}$, respectively $\boldsymbol{y}$, the vector with elements x_j, $1 \le j \le m_1$, respectively y_k, $1 \le k \le m_2$. For a function $f : \mathbb{R}^3 \to \mathbb{R}$, we define by $f(\boldsymbol{x}, \boldsymbol{y}, t)$ the $m_1 \times m_2$ matrix with entries $f(x_j, y_k, t)$ for $1 \le j \le m_1, 1 \le k \le m_2$. The semidiscretization can then be formulated as

$$\begin{aligned} \boldsymbol{U}'(t) = {} & d_{12}(\boldsymbol{x}, \boldsymbol{y}, t) \circ \left[D_x \boldsymbol{U}(t) D_y^T\right] + \boldsymbol{G}_0(t) \\ & + d_{11}(\boldsymbol{x}, \boldsymbol{y}, t) \circ [D_{xx} \boldsymbol{U}(t)] - a_1(\boldsymbol{x}, \boldsymbol{y}, t) \circ [D_x \boldsymbol{U}(t)] + \boldsymbol{G}_1(t) \\ & + d_{22}(\boldsymbol{x}, \boldsymbol{y}, t) \circ \left[\boldsymbol{U}(t) D_{yy}^T\right] - a_2(\boldsymbol{x}, \boldsymbol{y}, t) \circ \left[\boldsymbol{U}(t) D_y^T\right] + \boldsymbol{G}_2(t), \end{aligned} \tag{2.3.7}$$

for $t > 0$, and where $\circ$ denotes the *Hadamard product*. The given matrices $\boldsymbol{G}_0(t)$, $\boldsymbol{G}_1(t)$, $\boldsymbol{G}_2(t)$ contain the information about the boundary conditions corresponding to the mixed derivative, respectively derivatives in the x-direction and derivatives in the y-direction. The matrices D_x, D_{xx}, D_y, D_{yy} correspond to the discretization of the different spatial derivatives and are similar to those from the previous subsection. The initial matrix $\boldsymbol{U}(0)$ is defined via the initial condition u_0.

For the second step of the MOL we rewrite (2.3.7) in a more convenient form. Denote by vec[$\cdot$] the operator that turns any given matrix into a vector by putting its successive columns below each other. Let $U(t) = \text{vec}[\boldsymbol{U}(t)]$ and denote by I_x, respectively I_y, the identity matrix of size $m_1 \times m_1$, respectively $m_2 \times m_2$. Using some well-known properties of the *Kronecker product* $\otimes$, see e.g. [31], it follows that

$$\begin{aligned} U'(t) &= A(t)U(t) + g(t) \\ &= \left(A_0(t) + A_1(t) + A_2(t)\right) U(t) + g_0(t) + g_1(t) + g_2(t), \end{aligned} \tag{2.3.8}$$

for $t > 0$, with

$$\begin{aligned} A_0(t) &= \text{diag}[\text{vec}[d_{12}(\boldsymbol{x}, \boldsymbol{y}, t)]](D_y \otimes D_x), \\ A_1(t) &= \text{diag}[\text{vec}[d_{11}(\boldsymbol{x}, \boldsymbol{y}, t)]](I_y \otimes D_{xx}) - \text{diag}[\text{vec}[a_1(\boldsymbol{x}, \boldsymbol{y}, t)]](I_y \otimes D_x), \\ A_2(t) &= \text{diag}[\text{vec}[d_{22}(\boldsymbol{x}, \boldsymbol{y}, t)]](D_{yy} \otimes I_x) - \text{diag}[\text{vec}[a_2(\boldsymbol{x}, \boldsymbol{y}, t)]](D_y \otimes I_x), \end{aligned}$$

and

$$g_i(t) = \text{vec}[\boldsymbol{G}_i(t)], \quad \text{for } 0 \leq i \leq 2.$$

Suppose that d_{12} is the product of two functions d_{12}^x, d_{12}^y where the former one is only dependent on (x, t) and the latter one is only dependent on (y, t). The matrix $A_0(t)$ can then be rewritten as

$$\begin{aligned} A_0(t) &= \left(\text{diag}[d_{12}^y(\boldsymbol{y}, t)] \otimes \text{diag}[d_{12}^x(\boldsymbol{x}, t)]\right)(D_y \otimes D_x) \\ &= \left(\text{diag}[d_{12}^y(\boldsymbol{y}, t)] D_y\right) \otimes \left(\text{diag}[d_{12}^x(\boldsymbol{x}, t)] D_x\right). \end{aligned}$$

Similarly, if one of the other coefficient functions d_{11}, d_{22}, a_1, a_2 can be written as the product of a function that is only dependent on (x, t) and a function that is only dependent on (y, t), then the corresponding term in the semidiscretization (2.3.8) can be simplified in this way.

2.3.4. Sparsity Structure of the Semidiscrete System

In general, the exact solutions to semidiscrete systems of the type (2.3.4), (2.3.8) are not known in analytical form and one relies on numerical time stepping methods for their approximate solution. Let I be the identity matrix of the same size as the matrix $A(t)$. Classical implicit time stepping often requires solving linear systems of equations involving a matrix

$$B = I - \theta \Delta t A(t), \tag{2.3.9}$$

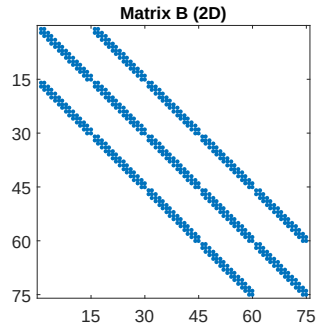


Figure 2.1: Sparsity structure of the matrix B corresponding to the semidiscrete system (2.3.8) *where $m_1 = 15$, $m_2 = 5$.*

for some real, strictly positive coefficient θ and temporal step size Δt, cf. e.g. [28, 35]. Hence, the sparsity structure of the semidiscretization matrix $A(t)$ plays an important role in the computational cost of the time stepping scheme.

Recall that the semidiscrete system (2.3.4) for the one-dimensional PDE is constructed by using second order central FD schemes. The corresponding matrix $A(t)$ is tridiagonal and solving linear systems involving the matrix B from (2.3.9) can then be performed very efficiently. Moreover, if the matrix B is independent of the time step, one can determine a LU factorization of it once, beforehand, and then use it in all steps. As the matrix B is tridiagonal, both matrices L and U are bidiagonal and the cost of solving one linear system with matrix B grows just linearly in m_1, which is very favourable.

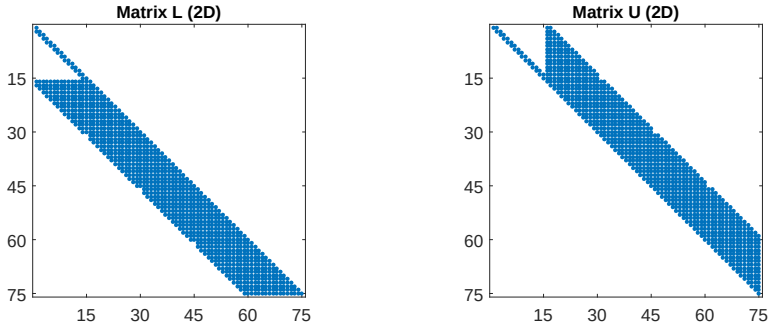


Figure 2.2: Sparsity structure of the matrix L (left) and U (right) corresponding to the semidiscrete system (2.3.8) *where $m_1 = 15$, $m_2 = 5$.*

The semidiscretization matrix A from (2.3.8) for the two-dimensional PDE, that involves Kronecker products of the FD matrices, has at most nine non-zero elements per row and column. The same holds for the corresponding matrix B from (2.3.9). As illustrated in Figure 2.1, however, the FD discretization of the two-dimensional problem gives rise to non-zero subdiagonals that lie

at a distance $m_1 + 1$ from the main diagonal. Therefore, solving linear systems with matrix B is more involved in the two-dimensional case than in the one-dimensional case. Suppose that the matrix B stemming from the two-dimensional case is independent of the time step so that the LU decomposition can be applied beforehand to increase the computational efficiency. As shown in Figure 2.2, the corresponding matrices L and U suffer from *fill-in*. In general, each row of L and U possesses $m_1 + 1$ non-zero entries and, consequently, the number of operations needed in each time step to solve a linear system with matrix B is directly proportional to $m_1^2 m_2$. Hence, this number of operations in each time step grows faster than the total number of grid points $m_1 m_2$, which is not favourable.

CHAPTER 3

ADI Time Discretization Methods

3.1. Introduction

Semidiscretization by FD formulas of initial-boundary value problems for time-dependent convection-diffusion equations leads to large systems of ODEs. In general there is no analytical solution available to these semidiscrete systems and one relies on numerical methods for their approximate solution. Since these systems of ODEs are usually stiff, implicit time stepping schemes form natural candidates. Let A be the semidiscretization matrix. Classical implicit methods such as the Crank–Nicolson scheme, see e.g. [12], require solving linear systems involving a matrix B as defined in (2.3.9). At the end of Chapter 2 it was shown that, if the PDE is two-dimensional, then the latter task can be computationally very intensive when standard LU-decomposition is applied. Similar results can be shown for higher-dimensional PDEs.

Starting from the 1950s, a variety of *Alternating Direction Implicit* (ADI) time stepping schemes have been developed, see e.g. [7, 13, 14, 55], that enable a highly efficient and stable numerical solution of semidiscretized multidimensional PDEs. ADI schemes employ a splitting of the semidiscrete operator in the different spatial directions, such as the splitting in (2.3.8). In each implicit stage of a given time step only one spatial dimension is handled, which can lead to a major computational advantage.

Consider for example the semidiscretization of the two-dimensional PDE in Subsection 2.3.3. Application of an ADI scheme requires solving linear systems of equations involving matrices

$$B_1 = I - \theta \Delta t A_1(t) \quad \text{and} \quad B_2 = I - \theta \Delta t A_2(t),$$

where $A_1(t), A_2(t)$ are defined by (2.3.8). As illustrated in Figure 3.1, the matrix B_1 is tridiagonal and B_2 is essentially tridiagonal such that the linear systems can be solved very efficiently. Moreover, if the matrices are independent of the time step, the computational efficiency can again be improved by determining a LU factorization of both matrices once, beforehand. Since the matrices B_1, B_2 are essentially tridiagonal, the corresponding matrices L and U are essentially bidiagonal. Once the factorization is performed, the number of operations needed in each time step to solve the linear systems of equations is directly proportional to the total number of grid points $m_1 m_2$. This yields

a large computational advantage in comparison with most classical implicit time stepping methods, which often require a number of operations directly proportional to $m_1^2 m_2$, cf. Subsection 2.3.4.

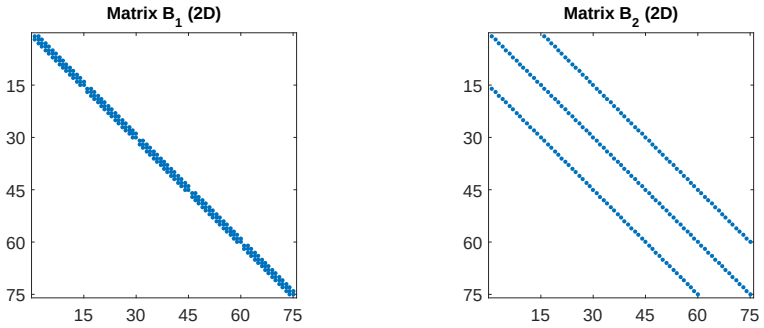


Figure 3.1: Sparsity structure of the matrix B_1 (left) and B_2 (right) corresponding to the semidiscrete system (2.3.8) *where $m_1 = 15$, $m_2 = 5$.*

The remainder of this chapter is organised as follows. In Section 3.2 we present four ADI schemes adapted to mixed spatial derivative terms that are widely used in computational finance. Section 3.3 gives an overview of the existing stability results for the pertinent ADI schemes relevant to two-dimensional convection-diffusion equations. In order to establish convergence of numerical processes for differential equations, the well-known, general approach consists of proving both stability and consistency of the scheme, see e.g. [28,35]. In Section 3.4 the classical orders of consistency of the ADI schemes are presented together with the uniform consistency bounds derived in [34].

3.2. ADI Methods Adapted to Mixed Spatial Derivatives

Directional splitting methods have already been used successfully in several application areas in science and engineering other than financial mathematics, cf. e.g. [7, 48, 67]. However, the PDEs of financial mathematics often have mixed spatial derivative terms due to correlated Brownian motions in the underlying stochastic processes. This feature was only studied marginally in other application areas. Only recently have ADI schemes been adapted to mixed spatial derivative terms and has a first analysis been performed in a structured way. In contemporary financial mathematics ADI schemes form state-of-the-art time discretization methods, see e.g. [8, 26, 37, 45, 64].

FD discretization of initial-boundary value problems for multidimensional time-dependent convection-diffusion equations leads to large systems of stiff ODEs

$$U'(t) = F(t, U(t)), \quad \text{for } 0 \leq t \leq T, \tag{3.2.1}$$

with given vector-valued function $F : [0, T] \times \mathbb{R}^m \to \mathbb{R}^m$, given end time T and given initial vector $U(0) = U_0 \in \mathbb{R}^m$. Here, m denotes the total number of

spatial grid points. Assume that the underlying PDE is l-dimensional and let the vector-valued function F be decomposed as

$$F(t, \boldsymbol{v}) = F_0(t, \boldsymbol{v}) + F_1(t, \boldsymbol{v}) + \cdots + F_l(t, \boldsymbol{v}), \quad \text{for } 0 \le t \le T,\ \boldsymbol{v} \in \mathbb{R}^m, \tag{3.2.2}$$

where F_0 represents the mixed spatial derivative terms and F_i, for $1 \le i \le l$, represents all spatial derivative terms in the ith direction. In this thesis, for the time discretization of (3.2.1) four prominent schemes of the ADI type are considered: the *Douglas* (Do) scheme, the *Craig–Sneyd* (CS) scheme, the *Modified Craig–Sneyd* (MCS) scheme and the *Hundsdorfer–Verwer* (HV) scheme. Let $\theta > 0$ be a given parameter, let $\Delta t_n > 0$ be given temporal step sizes and set $t_n = t_{n-1} + \Delta t_n$ for integers $n \ge 0$. For the ease of presentation we omit the dependency of the temporal step size Δt_n on the time step number n. Also, in computational practice uniform temporal grids are widely considered. The four ADI schemes define, in a one-step manner, approximations U_n to $U(t_n)$ successively for $n = 1, 2, 3, \ldots$ with $t_n \le T$ through:

Do scheme:

$$\begin{cases} Y_0 = U_{n-1} + \Delta t\, F(t_{n-1}, U_{n-1}), \\ Y_i = Y_{i-1} + \theta \Delta t \left(F_i(t_n, Y_i) - F_i(t_{n-1}, U_{n-1})\right), \text{ for } i = 1, 2, \ldots, l \\ U_n = Y_l, \end{cases} \tag{3.2.3}$$

CS scheme:

$$\begin{cases} Y_0 = U_{n-1} + \Delta t\, F(t_{n-1}, U_{n-1}), \\ Y_i = Y_{i-1} + \theta \Delta t \left(F_i(t_n, Y_i) - F_i(t_{n-1}, U_{n-1})\right), \text{ for } i = 1, 2, \ldots, l \\ \widetilde{Y}_0 = Y_0 + \tfrac{1}{2}\Delta t \left(F_0(t_n, Y_l) - F_0(t_{n-1}, U_{n-1})\right), \\ \widetilde{Y}_i = \widetilde{Y}_{i-1} + \theta \Delta t \left(F_i(t_n, \widetilde{Y}_i) - F_i(t_{n-1}, U_{n-1})\right), \text{ for } i = 1, 2, \ldots, l \\ U_n = \widetilde{Y}_l, \end{cases} \tag{3.2.4}$$

MCS scheme:

$$\begin{cases} Y_0 = U_{n-1} + \Delta t\, F(t_{n-1}, U_{n-1}), \\ Y_i = Y_{i-1} + \theta \Delta t \left(F_i(t_n, Y_i) - F_i(t_{n-1}, U_{n-1})\right), \text{ for } i = 1, 2, \ldots, l \\ \widehat{Y}_0 = Y_0 + \theta \Delta t \left(F_0(t_n, Y_l) - F_0(t_{n-1}, U_{n-1})\right), \\ \widetilde{Y}_0 = \widehat{Y}_0 + (\tfrac{1}{2} - \theta)\Delta t \left(F(t_n, Y_l) - F(t_{n-1}, U_{n-1})\right), \\ \widetilde{Y}_i = \widetilde{Y}_{i-1} + \theta \Delta t \left(F_i(t_n, \widetilde{Y}_i) - F_i(t_{n-1}, U_{n-1})\right), \text{ for } i = 1, 2, \ldots, l \\ U_n = \widetilde{Y}_l, \end{cases} \tag{3.2.5}$$

HV scheme:

$$\begin{cases} Y_0 = U_{n-1} + \Delta t\, F(t_{n-1}, U_{n-1}), \\ Y_i = Y_{i-1} + \theta\Delta t\,(F_i(t_n, Y_i) - F_i(t_{n-1}, U_{n-1}))\,, \ \text{for } i = 1, 2, \ldots, l \\ \widetilde{Y}_0 = Y_0 + \frac{1}{2}\Delta t\,(F(t_n, Y_l) - F(t_{n-1}, U_{n-1}))\,, \\ \widetilde{Y}_i = \widetilde{Y}_{i-1} + \theta\Delta t\,(F_i(t_n, \widetilde{Y}_i) - F_i(t_n, Y_l)), \ \text{for } i = 1, 2, \ldots, l \\ U_n = \widetilde{Y}_l. \end{cases} \tag{3.2.6}$$

Each of the ADI schemes above treats the F_0 part, which represents all mixed derivative terms, in an explicit manner. Each implicit substep only handles spatial derivatives in one spatial dimension, often leading to systems of equations involving essentially tridiagonal matrices, cf. Section 3.1.

The Do scheme (3.2.3) has been considered for example in [34,35], and can be regarded as a direct generalisation of the classical ADI schemes for diffusion equations by Douglas & Rachford [14] and Peaceman & Rachford [55] to the situation where mixed spatial derivative terms are present in the equation. To the best of our knowledge, this generalisation was first considered by McKee & Mitchell [51]. The Do scheme starts with an explicit Euler predictor stage, which is followed by l implicit but unidirectional corrector stages.

The CS scheme (3.2.4) can be viewed as an extension to the Do scheme. It was introduced by Craig & Sneyd [11] with the goal to arrive at a second order ADI scheme for diffusion equations with mixed derivative terms. The first two lines of (3.2.4) are exactly the same as in the Do scheme (3.2.3). Afterwards, the CS scheme performs another explicit update followed by l implicit unidirectional corrector stages. It is readily seen that the CS scheme reduces to the Do scheme if $F_0 = 0$.

The MCS scheme (3.2.5) has been introduced by In 't Hout & Welfert [42] and generalises the CS scheme by adding an explicit stage after the first set of implicit corrections. By doing so, the MCS scheme offers more flexibility than the CS scheme in the choice of θ if second order consistency is desired, cf. Section 3.4. For $\theta = \frac{1}{2}$ the MCS scheme reduces to the CS scheme.

The HV scheme (3.2.6) can also be regarded as an extension of the Do scheme and was introduced by Hundsdorfer [34] and Verwer et. al. [67] for the numerical solution of convection-diffusion-reaction equations from chemistry. In 't Hout & Welfert [41,42] studied the application of the HV scheme to equations containing mixed spatial derivative terms. The main difference between the HV scheme and the MCS scheme is the fact that the former one uses the approximations Y_l corresponding to time t_n (instead of U_{n-1}, corresponding to t_{n-1}) in the implicit corrector stages after the explicit update.

In this PhD thesis, the four ADI schemes are considered in application to semidiscretized two-dimensional time-dependent convection-diffusion equations. From now on, unless explicitly stated otherwise, it will be tacitly assumed that the spatial dimension l equals two when the ADI schemes are mentioned.

3.3. Stability of ADI Schemes

In order to establish *convergence* of numerical processes for differential equations, the well-known general approach consists of proving both *stability* and *consistency* of the scheme, see e.g. [28, 35]. Stability means that any errors, e.g. rounding errors or discretization errors, cannot grow excessively during the numerical process and forms a crucial property for every time stepping method. Since systems of ODEs stemming from semidiscretization of convection-diffusion equations are usually stiff, *unconditional stability* is a desirable property. This means that the time stepping method is stable without any restriction on the temporal step size.

Theoretical unconditional stability results in the von Neumann sense, relevant to FD discretizations of two-dimensional convection-diffusion equations with mixed derivative term, have been derived for all four ADI schemes from Section 3.2 in [11, 38–42, 51–54]. We briefly review the main conclusions from these references, where stability is always considered in the l_2-norm. Consider a two-dimensional time-dependent convection-diffusion equation (2.3.5) with constant coefficients and periodic boundary condition. Assume that spatial discretization is performed on uniform Cartesian grids with relevant second order FD schemes. Both the Do and CS schemes are then unconditionally stable whenever $\theta \geq \frac{1}{2}$. When applied to two-dimensional pure diffusion equations, i.e. if $a_1 = a_2 = 0$, the MCS scheme, respectively HV scheme, is unconditionally stable if θ satisfies

$$\theta \geq \max\left\{\tfrac{1}{4}, \tfrac{\gamma+1}{6}\right\}, \quad \text{respectively} \quad \theta \geq \max\left\{\tfrac{1}{4}, \tfrac{\gamma+1}{4+2\sqrt{2}}\right\}, \tag{3.3.1}$$

where $\gamma \in [0, 1]$ represents the relative size of the mixed derivative term, see (2.3.6). When convection terms are present, it holds that the MCS scheme is unconditionally stable if $\frac{1}{2} \leq \theta \leq 1$. For values $\frac{1}{4} \leq \theta < \frac{1}{2}$ it can be shown that the MCS scheme is unconditionally stable under a restriction on γ. An analytical expression for a sufficient restriction can be found in [54]. For the practically important case of $\theta = \frac{1}{3}$ this sufficient restriction is given by

$$\gamma \leq \tfrac{2+\sqrt{10}}{6} \approx 0.86.$$

Additional numerical experiments in [39] suggest that this restriction can be weakened to $\gamma \lesssim 0.96$. For the HV scheme, to the best of our knowledge, there are no theoretical stability results available if convection terms and a mixed spatial derivative are present. In [48] the HV scheme is used for an application from physics involving a two-dimensional convection-diffusion equation without mixed derivative term. In the article it is shown that if there is no mixed derivative term, i.e. if $\gamma = 0$, then the HV scheme is unconditionally stable if

$$\theta \geq \tfrac{1}{2} + \tfrac{1}{6}\sqrt{3} \approx 0.79.$$

In [41] it is conjectured that the latter bound on θ is also sufficient for unconditional stability of the HV scheme in the presence of mixed derivative terms.

Numerical experiments on the stability of the ADI schemes in application to semidiscretized two-dimensional convection-diffusion equations with mixed

derivative term from financial mathematics have been performed for example in [37, 53]. In the pertinent literature the ADI schemes have been applied for example to the well-known and challenging Heston PDE [33]. This PDE possesses non-constant coefficients, non-periodic boundary conditions and has special features such as a vanishing diffusion term near a boundary. The experiments in [37, 53] and ample additional references suggest that the theoretical stability results, for the situation where convection terms and a mixed derivative are present, are very useful outside the model framework.

3.4. Consistency of ADI Schemes

Next to stability, consistency of the ADI schemes forms a fundamental concept in proving convergence. Consistency is concerned with the so-called local discretization error, which is the error incurred in one, arbitrary, fictitious step of the numerical process if one would have started from the exact solution.

We first consider the classical order of consistency, that is the order of consistency for fixed non-stiff ODE systems. Recall that in this thesis the ADI schemes are applied to two-dimensional problems. Let

$$U'(t) = (A_0 + A_1 + A_2)U(t) + g_0(t) + g_1(t) + g_2(t),$$

be a fixed non-stiff ODE system with given constant matrices A_i, $0 \leq i \leq 2$, and given vectors $g_i(t)$, $0 \leq i \leq 2$. By Taylor expansion, and after some elaborate calculations, the order of consistency can be obtained for the four ADI schemes from Section 3.2. It follows that the classical order of consistency of the Do scheme is equal to one for all θ. This low order is a consequence of the fact that the A_0 part is treated in a simple, forward Euler fashion. If $A_0 = 0$, however, the order can be increased to two by choosing $\theta = \frac{1}{2}$. For the CS scheme it holds that the classical order of consistency is equal to two if and only if $\theta = \frac{1}{2}$, also if A_0 is non-zero. For all other values of θ the order of the CS scheme drops to one. The MCS scheme and the HV scheme always attain classical order of consistency equal to two for any given θ. The parameter can then be chosen such that the ADI schemes meet additional requirements. In view of the order of the CS scheme, from now on it will be tacitly assumed that the parameter value $\theta = \frac{1}{2}$ is used for this scheme.

For practical relevance it is crucial that the local error bounds hold uniformly in the arbitrarily large size of the semidiscrete system or, equivalently, in the arbitrarily small spatial mesh width. Since semidiscretized convection-diffusion equations are usually stiff, and the size of the systems depends on the spatial mesh widths, the classical order of consistency cannot be used to prove relevant convergence results. Analysing the uniform order of consistency (and convergence) of ADI schemes for stiff systems of ODEs is a non-trivial task.

The local discretization errors of the Do scheme and the HV scheme have been analysed in [34] for arbitrarily small spatial mesh widths. Although the F_0 term in [34] does not represent a mixed spatial derivative, the results are relevant to our problem setting. It is shown that the local discretization errors of the Do scheme and the HV scheme are of second order under some natural stability and smoothness conditions. Provided that the parameter θ is chosen

such that the Do scheme, respectively HV scheme, is stable, this leads to a uniform order of convergence of at least one. In general, one cannot expect a higher order of convergence for the Do scheme since it treats the F_0 term only in a forward Euler fashion. For the special case where $F_0 = 0$ and $\theta = \frac{1}{2}$, however, it is shown in [34] that if the temporal step sizes are uniform, then the Do scheme is second order convergent. In addition, a second order convergence result for the HV scheme is proved for the case $l = 1$, i.e. when only one part of the semidiscrete system is handled implicitly.

In the following chapters a detailed analysis of the discretization errors of the MCS scheme and the HV scheme is performed for the relevant case $l = 2$. The results are used to prove second order convergence of the pertinent ADI schemes, uniformly in the spatial mesh width.

CHAPTER 4

Convergence of the MCS Scheme

4.1. Introduction

Recall that semidiscretization by finite difference methods of initial-boundary value problems for multidimensional time-dependent convection-diffusion equations leads to large systems of stiff ODEs,

$$U'(t) = F(t, U(t)) \quad (0 \le t \le T),\ U(0) = U_0, \tag{4.1.1}$$

with given vector-valued function $F : [0, T] \times \mathbb{R}^m \to \mathbb{R}^m$ and given initial vector $U_0 \in \mathbb{R}^m$, where integer $m \ge 1$ is the number of spatial grid points. When the underlying PDE is multidimensional, the semidiscrete function F can often be decomposed as

$$F(t, \boldsymbol{v}) = F_0(t, \boldsymbol{v}) + F_1(t, \boldsymbol{v}) + \cdots + F_l(t, \boldsymbol{v}), \quad \text{for } 0 \le t \le T,\ \boldsymbol{v} \in \mathbb{R}^m, \tag{4.1.2}$$

where F_0 represents all mixed spatial derivative terms and F_i, for $1 \le i \le l$, represents all spatial derivative terms in the i-th direction, cf. Subsection 2.3.3.

In this chapter, for the effective time discretization of systems (4.1.1) we consider the prominent *MCS scheme* (3.2.5) with constant temporal step size Δt such that $t_n = n\Delta t$. The MCS scheme is often used in financial practice for the numerical approximation of option values and their sensitivities. *Speed, stability and accuracy* of the method thus form key concepts. The former two have already been studied extensively in the literature, see e.g. [37, 42, 54]. An overview is presented in Chapter 3. To the best of our knowledge, a relevant theoretical convergence analysis is still open in the literature. Nevertheless, in order for the MCS scheme to be useful in practice, convergence and hence accuracy is of paramount importance.

It can be verified by standard arguments that if natural stability and smoothness assumptions hold, then the MCS scheme is convergent of order two for fixed, nonstiff ODE systems, see Subsection 3.4. It is well-known in the literature, however, that this standard convergence analysis of time stepping schemes has limited relevance for the application to semidiscrete systems

This chapter is based on the article 'Convergence of the Modified Craig–Sneyd scheme for two-dimensional convection-diffusion equations with mixed derivative term', published in J. Comp. Appl. Math., 296:170–180, 2016 [44].

(4.1.1). In this analysis, the size of the error constant in the obtained bound for the global temporal discretization errors may become arbitrarily large as the spatial mesh width from the semidiscretization tends to zero ($m \to \infty$), which renders this bound impractical.

In the present chapter we shall derive a first useful convergence bound for the MCS scheme, with constant step size, that is directly relevant to semidiscretized two-dimensional convection-diffusion equations with mixed derivative term. Our analysis is inspired by that of Hundsdorfer [33, 34], cf. also [35], for operator splitting schemes applied to multidimensional convection-diffusion-reaction problems without mixed derivative terms.

The outline of the chapter is the following. In Section 4.2 a perturbed version of the MCS scheme is used to derive a recursion formula for the total error. Expressions for the local errors in the perturbed scheme are derived by Taylor expansion. We employ a subtle splitting of the resulting local discretization error such that each part meets the requirements for application of a key lemma from [33]. Under some stability assumptions, this leads to the main result of the chapter: a second order convergence theorem for the MCS scheme. Positive results on the stability assumptions are derived in the von Neumann framework. In Section 4.3 numerical experiments illustrate that the MCS scheme is second order convergent in application to a semidiscretized model convection-diffusion equation. This positive conclusion is in line with our theoretical convergence analysis. Section 4.4 gives concluding remarks.

4.2. Convergence Analysis

4.2.1. Preliminaries

Assume that

$$F(t, \boldsymbol{v}) = A\boldsymbol{v} + g(t), \quad F_i(t, \boldsymbol{v}) = A_i\boldsymbol{v} + g_i(t), \quad \text{for } 0 \le t \le T,\ \boldsymbol{v} \in \mathbb{R}^m, 0 \le i \le l,$$

where A, A_i, $0 \le i \le l$, are given real $m \times m$-matrices and g, g_i, $0 \le i \le l$, are given real m-vector valued functions. Recall from the previous chapter that I denotes the $m \times m$ identity matrix. For convenience, define the matrices

$$Z = \Delta t A, \quad Z_i = \Delta t A_i, \quad Q_i = I - \theta Z_i, \quad \text{for } 0 \le i \le l, \quad P = Q_1 Q_2 \cdots Q_l,$$

where we emphasize that the temporal step size Δt is assumed to be uniform. Consider the naturally scaled inner product $(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{m}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}$ for $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^m$ with induced vector and matrix norms $\|\cdot\|_2$. We shall assume that

$$(A_i\boldsymbol{v}, \boldsymbol{v}) \le 0 \quad \text{whenever } \boldsymbol{v} \in \mathbb{R}^m,\ 1 \le i \le l.$$

This assumption is often fulfilled when dealing with semidiscrete systems stemming from time-dependent convection-diffusion equations, cf. e.g. [33–35]. It implies that the Q_i and P are invertible and

$$\|Q_i^{-1}\|_2 \le 1, \quad \text{for } 1 \le i \le l, \quad \|P^{-1}\|_2 \le 1. \tag{4.2.1}$$

4.2.2. Error Recursion

For the convergence analysis, we consider along with (3.2.5) the perturbed scheme

$$\begin{cases} Y_0^* = U_{n-1}^* + \Delta t\, F(t_{n-1}, U_{n-1}^*) + \rho_0, \\ Y_i^* = Y_{i-1}^* + \theta\Delta t \left(F_i(t_n, Y_i^*) - F_i(t_{n-1}, U_{n-1}^*)\right) + \rho_i, \quad \text{for } i = 1, 2, \ldots, l, \\ \widehat{Y}_0^* = Y_0^* + \theta\Delta t \left(F_0(t_n, Y_l^*) - F_0(t_{n-1}, U_{n-1}^*)\right) + \widehat{\rho}_0, \\ \widetilde{Y}_0^* = \widehat{Y}_0^* + (\frac{1}{2} - \theta)\Delta t \left(F(t_n, Y_l^*) - F(t_{n-1}, U_{n-1}^*)\right) + \widetilde{\rho}_0, \\ \widetilde{Y}_i^* = \widetilde{Y}_{i-1}^* + \theta\Delta t \left(F_i(t_n, \widetilde{Y}_i^*) - F_i(t_{n-1}, U_{n-1}^*)\right) + \widetilde{\rho}_i, \quad \text{for } i = 1, 2, \ldots, l, \\ U_n^* = \widetilde{Y}_l^*. \end{cases} \tag{4.2.2}$$

Here $\rho_i, \widetilde{\rho}_i \in \mathbb{R}^m$ $(0 \le i \le l)$ and $\widehat{\rho}_0 \in \mathbb{R}^m$ denote arbitrary given perturbation vectors. These perturbations may depend on the step number n. For ease of presentation, this is omitted in the notation. In the following we derive a useful formula for the error

$$e_n = U_n^* - U_n.$$

Define the auxiliary variables

$$\varepsilon_i = Y_i^* - Y_i, \quad \widetilde{\varepsilon}_i = \widetilde{Y}_i^* - \widetilde{Y}_i, \quad \text{for } 0 \le i \le l \quad \text{and} \quad \widehat{\varepsilon}_0 = \widehat{Y}_0^* - \widehat{Y}_0.$$

From (3.2.5), (4.2.2) one directly obtains

$$\begin{aligned} \varepsilon_0 &= e_{n-1} + \Delta t A e_{n-1} + \rho_0 \\ &= (I + Z)e_{n-1} + \rho_0, \\ \varepsilon_i &= \varepsilon_{i-1} + \theta\Delta t(A_i\varepsilon_i - A_i e_{n-1}) + \rho_i, \quad 1 \le i \le l. \end{aligned}$$

The latter equation can readily be rewritten as

$$\varepsilon_i = e_{n-1} + Q_i^{-1}(\varepsilon_{i-1} - e_{n-1} + \rho_i), \quad 1 \le i \le l. \tag{4.2.3}$$

Next,

$$\begin{aligned} \widehat{\varepsilon}_0 &= \varepsilon_0 + \theta Z_0(\varepsilon_l - e_{n-1}) + \widehat{\rho}_0 \\ &= (I + Z - \theta Z_0)e_{n-1} + \theta Z_0\varepsilon_l + \rho_0 + \widehat{\rho}_0, \\ \widetilde{\varepsilon}_0 &= \widehat{\varepsilon}_0 + (\tfrac{1}{2} - \theta)Z(\varepsilon_l - e_{n-1}) + \widetilde{\rho}_0 \\ &= (I + (\tfrac{1}{2} + \theta)Z - \theta Z_0)e_{n-1} + (\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\varepsilon_l + \rho_0 + \widehat{\rho}_0 + \widetilde{\rho}_0 \end{aligned}$$

and analogously to (4.2.3) there holds

$$\widetilde{\varepsilon}_i = e_{n-1} + Q_i^{-1}(\widetilde{\varepsilon}_{i-1} - e_{n-1} + \widetilde{\rho}_i), \quad 1 \le i \le l. \tag{4.2.4}$$

Using (4.2.4) together with the obtained expression for $\widetilde{\varepsilon}_0$, it follows that

$$\begin{aligned} e_n = \widetilde{\varepsilon}_l &= e_{n-1} + Q_l^{-1}(\widetilde{\varepsilon}_{l-1} - e_{n-1} + \widetilde{\rho}_l) \\ &= e_{n-1} + Q_l^{-1}Q_{l-1}^{-1}(\widetilde{\varepsilon}_{l-2} - e_{n-1} + \widetilde{\rho}_{l-1}) + Q_l^{-1}\widetilde{\rho}_l \\ &\vdots \\ &= e_{n-1} + P^{-1}(\widetilde{\varepsilon}_0 - e_{n-1} + \widetilde{\rho}_1) + \sum_{i=2}^{l} Q_l^{-1}Q_{l-1}^{-1}\cdots Q_i^{-1}\widetilde{\rho}_i \\ &= e_{n-1} + P^{-1}(-\theta Z_0 + (\tfrac{1}{2}+\theta)Z)e_{n-1} + P^{-1}(\theta Z_0 + (\tfrac{1}{2}-\theta)Z)\varepsilon_l \\ &\quad + P^{-1}(\rho_0 + \widehat{\rho}_0 + \widetilde{\rho}_0) + \sum_{i=1}^{l} Q_l^{-1}Q_{l-1}^{-1}\cdots Q_i^{-1}\widetilde{\rho}_i. \end{aligned}$$

In a similar way, using (4.2.3), it is seen that

$$\begin{aligned} \varepsilon_l &= e_{n-1} + P^{-1}(\varepsilon_0 - e_{n-1} + \rho_1) + \sum_{i=2}^{l} Q_l^{-1}Q_{l-1}^{-1}\cdots Q_i^{-1}\rho_i \\ &= (I + P^{-1}Z)e_{n-1} + P^{-1}\rho_0 + \sum_{i=1}^{l} Q_l^{-1}Q_{l-1}^{-1}\cdots Q_i^{-1}\rho_i. \end{aligned}$$

Inserting the obtained expression for ε_l into that for e_n, we arrive at the useful recursion formula

$$e_n = Re_{n-1} + d_n \tag{4.2.5}$$

with *stability matrix*

$$R = I + P^{-1}Z + P^{-1}(\theta Z_0 + (\tfrac{1}{2}-\theta)Z)P^{-1}Z \tag{4.2.6}$$

and vector

$$\begin{aligned} d_n &= P^{-1}(\theta Z_0 + (\tfrac{1}{2}-\theta)Z)(P^{-1}\rho_0 + \sum_{i=1}^{l} Q_l^{-1}Q_{l-1}^{-1}\cdots Q_i^{-1}\rho_i) \\ &\quad + P^{-1}(\rho_0 + \widehat{\rho}_0 + \widetilde{\rho}_0) + \sum_{i=1}^{l} Q_l^{-1}Q_{l-1}^{-1}\cdots Q_i^{-1}\widetilde{\rho}_i. \end{aligned} \tag{4.2.7}$$

For a temporal index N with $N\Delta t \le T$ the recursion (4.2.5) implies

$$e_N = R^N e_0 + \sum_{n=1}^{N} R^{N-n} d_n. \tag{4.2.8}$$

4.2.3. Local Discretization Errors

We now consider the perturbed MCS scheme (4.2.2) with constant step size and where the perturbations are such that

$$U_{n-1}^* = U(t_{n-1}), \quad Y_i^* = \widetilde{Y}_i^* = U(t_n), \quad \text{for } 0 \le i \le l, \quad \text{and} \quad \widehat{Y}_0^* = U(t_n).$$

With this choice, d_n is the *local discretization error* and $e_n = U(t_n) - U_n$ the *global discretization error* in the n-th step.

For the convergence analysis of any given time stepping scheme applied to semidiscrete PDEs to be practical, it is imperative that the pertinent stability and error bounds are not adversely affected by the (arbitrarily small) spatial mesh width employed in the semidiscretization. Accordingly, in this chapter, *by the notation* $\mathcal{O}\left((\Delta t)^p\right)$ *we shall always mean that the norm* $\|\cdot\|_2$ *of the term under consideration is bounded by a positive constant times* $(\Delta t)^p$ *where the constant is independent of the spatial mesh width, the temporal step size* $\Delta t > 0$ *and the step number* $n \geq 1$ *with* $n\Delta t \leq T$. *If* $p = 0$, *then we write* $\mathcal{O}(1)$ *for short.*

Throughout this chapter we will assume that the MCS scheme is stable in the sense that there exists a constant M such that the stability matrix satisfies the inequality $\|R^n\|_2 \leq M$ uniformly in the spatial mesh width, $\Delta t > 0$ and integer $n \geq 1$. Thus, $R^n = \mathcal{O}(1)$.

To arrive at an optimal convergence order p, it turns out that a careful investigation of the local discretization errors d_n is required. Define

$$\varphi_i(t) = F_i(t, U(t)), \quad \text{for } 0 \leq t \leq T,\ 0 \leq i \leq l.$$

We assume that the vector functions φ_i are twice continuously differentiable and that their second derivatives are bounded on $[0, T]$ uniformly in the spatial mesh width. Notice that $U'(t) = \sum_{i=0}^{l} \varphi_i(t)$, so the above smoothness condition for the φ_i implies one for U too. By Taylor expansion of $U(t)$ about $t = t_{n-1}$ it directly follows that

$$\begin{aligned}
\rho_0 &= U''(t_{n-1})\ \tfrac{1}{2}(\Delta t)^2 + \mathcal{O}\left((\Delta t)^3\right), \\
\widehat{\rho}_0 &= -\varphi_0'(t_{n-1})\ \theta(\Delta t)^2 + \mathcal{O}\left((\Delta t)^3\right), \\
\widetilde{\rho}_0 &= -U''(t_{n-1})\ (\tfrac{1}{2} - \theta)(\Delta t)^2 + \mathcal{O}\left((\Delta t)^3\right), \\
\rho_i &= \widetilde{\rho}_i = -\varphi_i'(t_{n-1})\ \theta(\Delta t)^2 + \mathcal{O}\left((\Delta t)^3\right), \quad 1 \leq i \leq l.
\end{aligned} \tag{4.2.9}$$

Since $\rho_i = \widetilde{\rho}_i$ for all $1 \leq i \leq l$, the expression (4.2.7) for d_n becomes

$$\begin{aligned}
d_n = {} & P^{-1}(\widehat{\rho}_0 + \widetilde{\rho}_0) \\
& + \left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right)\left(P^{-1}\rho_0 + \sum_{i=1}^{l} Q_l^{-1} Q_{l-1}^{-1} \cdots Q_i^{-1} \rho_i\right).
\end{aligned}$$

Inserting the expansions from (4.2.9) into this and taking into account the uniform boundedness of the matrices Q_i^{-1}, $1 \leq i \leq l$, see (4.2.1), we obtain

$$\begin{aligned}
d_n = {} & \left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right) P^{-1} U''(t_{n-1})\ \tfrac{1}{2}(\Delta t)^2 \\
& - \left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right) \sum_{i=1}^{l} Q_l^{-1} Q_{l-1}^{-1} \cdots Q_i^{-1} \varphi_i'(t_{n-1})\ \theta(\Delta t)^2 \\
& - P^{-1}\varphi_0'(t_{n-1})\ \theta(\Delta t)^2 - P^{-1} U''(t_{n-1})\ (\tfrac{1}{2} - \theta)(\Delta t)^2 \\
& + \left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right) \mathcal{O}\left((\Delta t)^3\right) + \mathcal{O}\left((\Delta t)^3\right).
\end{aligned}$$

Using that $U''(t) = \sum_{i=0}^{l} \varphi_i'(t)$ the latter expression for d_n can be written in the following form, which will be employed in the next subsection,

$$\begin{aligned} d_n = {} & P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)P^{-1}(\tfrac{1}{2}U''(t_{n-1}) - \theta \sum_{i=1}^{l} \varphi_i'(t_{n-1}))\ (\Delta t)^2 \\ & + \Big[\left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right) P^{-1} \\ & \times \sum_{i=2}^{l} (I - Q_1 Q_2 \cdots Q_{i-1}) \varphi_i'(t_{n-1})\ \theta(\Delta t)^2 \Big] \\ & + \left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right) \mathcal{O}\left((\Delta t)^3\right) + \mathcal{O}\left((\Delta t)^3\right). \end{aligned} \tag{4.2.10}$$

4.2.4. Convergence Theorem for the MCS Scheme

From (4.2.10) and the specific dependence of the matrices Z_0, Z, Q_i, $1 \le i \le l$, on Δt it is readily seen that for any given *fixed* semidiscrete system the local errors are bounded by a constant times $(\Delta t)^3$. Next, formula (4.2.8) together with the stability of the MCS scheme directly imply a well-known estimate for the global discretization errors in terms of the local discretization errors (note that $e_0 = 0$),

$$\|e_N\|_2 \le M \sum_{n=1}^{N} \|d_n\|_2.$$

Hence, it follows that the global errors are bounded by a constant times $(\Delta t)^2$, that is second order convergence. However, as the spatial mesh width decreases (and the size m of the semidiscrete system increases), the pertinent error constant can become arbitrarily large due to negative powers of the spatial mesh width occurring in the matrices A_j, $0 \le j \le l$. Clearly, this renders the global error bound obtained in this way impractical.

In the following we shall present for the MCS scheme with $l = 2$ a useful second order convergence result, which is valid uniformly in the spatial mesh width. We apply a key lemma from Hundsdorfer [33], cf. also [34, 35]. For completeness, its (short) proof is included.

Lemma 4.2.1 (Hundsdorfer) *Let $\alpha > 0$. If the time stepping scheme is stable and the local discretization errors satisfy*

$$d_n = (R - I)\xi_n + \eta_n,$$

with

$$\xi_n = \mathcal{O}\left((\Delta t)^\alpha\right),\ \xi_n - \xi_{n-1} = \mathcal{O}\left((\Delta t)^{\alpha+1}\right),\ \eta_n = \mathcal{O}\left((\Delta t)^{\alpha+1}\right), \tag{4.2.11}$$

then for the global discretization errors one has that $e_N = \mathcal{O}\left((\Delta t)^\alpha\right)$.

Proof Consider the expression (4.2.8) for the global discretization error. Inserting $d_n = (R - I)\xi_n + \eta_n$ and $e_0 = 0$ gives

$$e_N = R^N \xi_1 - \xi_N + \sum_{n=2}^{N} R^{N-n+1}(\xi_n - \xi_{n-1}) + \sum_{n=1}^{N} R^{N-n} \eta_n.$$

By stability of the time stepping scheme (cf. Subsection 4.2.3), this leads to the bound

$$\|e_N\|_2 \leq M\|\xi_1\|_2 + \|\xi_N\|_2 + M\sum_{n=2}^{N}\|\xi_n - \xi_{n-1}\|_2 + M\sum_{n=1}^{N}\|\eta_n\|_2.$$

Using the properties of ξ_n and η_n in (4.2.11) and $N\Delta t \leq T$, the assertion of the lemma follows.

■

If $l = 2$, then the obtained expression (4.2.10) for the local error simplifies to $d_n = d_n^{(1)} + d_n^{(2)} + d_n^{(3)}$ with

$$\begin{aligned}
d_n^{(1)} &= P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)P^{-1}(\tfrac{1}{2}U''(t_{n-1}) - \theta\sum_{i=1}^{2}\varphi_i'(t_{n-1}))\ (\Delta t)^2,\\
d_n^{(2)} &= \left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right) P^{-1}Z_1\,\varphi_2'(t_{n-1})\ \theta^2(\Delta t)^2,\\
d_n^{(3)} &= \left(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)\right)\mathcal{O}\left((\Delta t)^3\right) + \mathcal{O}\left((\Delta t)^3\right).
\end{aligned}$$

Using formula (4.2.6) for the stability matrix, the first two components of d_n can be rewritten as (assuming the pertinent inverses exist),

$$\begin{aligned}
d_n^{(1)} &= (R - I)Z^{-1}P(I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z))^{-1}d_n^{(1)}\\
&= (R - I)Z^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z)(P + \theta Z_0 + (\tfrac{1}{2} - \theta)Z)^{-1}\\
&\quad\times(\tfrac{1}{2}U''(t_{n-1}) - \theta\sum_{i=1}^{2}\varphi_i'(t_{n-1}))\ (\Delta t)^2,\\
d_n^{(2)} &= (R - I)Z^{-1}Z_1\varphi_2'(t_{n-1})\ \theta^2(\Delta t)^2.
\end{aligned}$$

Upon invoking Lemma 4.2.1 with $\alpha = 2$, we then arrive at the main result of this chapter.

Theorem 4.2.2 *Let $l = 2$ and consider a uniform temporal step size. Assume that the φ_i ($i = 0, 1, 2$) are twice continuously differentiable and their second derivatives are bounded on $[0, T]$ uniformly in the spatial mesh width. Assume $(A_i\boldsymbol{v}, \boldsymbol{v}) \leq 0$ whenever $\boldsymbol{v} \in \mathbb{R}^m$ and $i = 1, 2$. Assume the MCS scheme is stable, the matrices A and $P + \theta Z_0 + (\frac{1}{2} - \theta)Z$ are invertible and the matrices*

$$A^{-1}A_1,\ A^{-1}A_2,\ I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z),\ (P + \theta Z_0 + (\tfrac{1}{2} - \theta)Z)^{-1} \quad (4.2.12)$$

are all $\mathcal{O}(1)$. Then the global discretization errors for the MCS scheme satisfy

$$e_N = \mathcal{O}\left((\Delta t)^2\right).$$

4.2.5. Boundedness Assumptions in Theorem 4.2.2

The assumptions concerning the uniform boundedness of the matrices (4.2.12) in Theorem 4.2.2 are similar to those made in [34] in order to prove convergence of the HV scheme with $l = 1$. The uniform boundedness of $A^{-1}A_1$ and $A^{-1}A_2$ was also assumed there and is often fulfilled in practical applications. If $l = 1$, the assumption $(P+\theta Z_0+(\frac{1}{2}-\theta)Z)^{-1} = \mathcal{O}(1)$ is closely related to the condition (29) in [34] for the HV scheme. The assumption $I+P^{-1}(\theta Z_0+(\frac{1}{2}-\theta)Z) = \mathcal{O}(1)$ can be viewed as a counterpart of the condition $I - P^{-1} + \frac{1}{2}P^{-1}Z = \mathcal{O}(1)$ which was tacitly assumed in [34].

For $l = 2$ the conditions

$$I + P^{-1}(\theta Z_0 + (\tfrac{1}{2} - \theta)Z) = \mathcal{O}(1) \quad \text{and} \quad (P + \theta Z_0 + (\tfrac{1}{2} - \theta)Z)^{-1} = \mathcal{O}(1)$$

are new in the literature. To gain insight into these conditions, we follow the well-known von Neumann framework and consider the two-dimensional convection-diffusion equation, cf. Subsection 2.3.3,

$$u_t = d_{11}u_{xx} + 2d_{12}u_{xy} + d_{22}u_{yy} + c_1u_x + c_2u_y \tag{4.2.13}$$

for $(x, y) \in (0, 1) \times (0, 1)$, $0 \le t \le T$ with periodic boundary condition. In this chapter $c_1, c_2, d_{11}, d_{12}, d_{22}$ denote given real constants that satisfy (2.3.6). After semidiscretization of (4.2.13) by standard finite difference schemes on uniform rectangular grids, the analysis reduces to bounding from above the two scalar terms

$$|R_1| := |1 + \tfrac{1}{2}\tfrac{z_0}{p} + (\tfrac{1}{2} - \theta)\tfrac{z_1+z_2}{p}| \quad \text{and} \quad |R_2| := |p + \tfrac{1}{2}z_0 + (\tfrac{1}{2} - \theta)(z_1 + z_2)|^{-1} \tag{4.2.14}$$

with $p = (1 - \theta z_1)(1 - \theta z_2)$ for all complex numbers z_0, z_1, z_2 satisfying

$$\mathcal{R}z_1 \le 0, \quad \mathcal{R}z_2 \le 0, \quad |z_0| \le 2\gamma\sqrt{\mathcal{R}z_1\mathcal{R}z_2}. \tag{4.2.15}$$

The condition (4.2.15) arises naturally in the von Neumann stability analysis of ADI schemes when a mixed derivative u_{xy} is present. It has been considered in [39, 41, 42] with $\gamma = 1$ and in [40, 53] for arbitrary $\gamma \in [0, 1]$.

For the first term in (4.2.14), we obtain the following positive result under the conditions in (4.2.15).

Theorem 4.2.3 *Assume* (4.2.15) *and* $0 \le \gamma \le 1$. *Then*

$$|R_1| \le \begin{cases} \frac{1}{2\theta} - \frac{3}{2} & \text{if } 0 < \theta < \frac{1}{6}\,, \\ \frac{3}{2} & \text{if } \frac{1}{6} \le \theta \le 1\,, \\ 2 - \frac{1}{2\theta} & \text{if } 1 < \theta\,. \end{cases}$$

Proof By [41, Lemma 2.3] it holds that

$$p \ne 0 \quad \text{and} \quad |\alpha| + |\beta| \le \frac{1}{2\theta} \quad \text{with} \quad \alpha = \frac{z_0}{p}\ ,\ \beta = \frac{1}{2\theta} + \frac{z_1 + z_2}{p}\,.$$

Using this we obtain

$$\begin{aligned}\left|1+\tfrac{1}{2}\frac{z_0}{p}+(\tfrac{1}{2}-\theta)\frac{z_1+z_2}{p}\right| &= \left|1+\tfrac{1}{2}\alpha+(\tfrac{1}{2}-\theta)(\beta-\tfrac{1}{2\theta})\right| \\ &= \left|\tfrac{3}{2}-\tfrac{1}{4\theta}+\tfrac{1}{2}\alpha+(\tfrac{1}{2}-\theta)\beta\right| \\ &\leq \left|\tfrac{3}{2}-\tfrac{1}{4\theta}\right|+\tfrac{1}{2}\left|\alpha\right|+\left|\tfrac{1}{2}-\theta\right|\left|\beta\right| \\ &\leq \left|\tfrac{3}{2}-\tfrac{1}{4\theta}\right|+\tfrac{1}{2\theta}\cdot\max\{\tfrac{1}{2},\left|\tfrac{1}{2}-\theta\right|\},\end{aligned}$$

which readily yields the result of the theorem.

■

For bounding the second term in (4.2.14), we make use of the following elementary lemma, where $\mathbb{R}^+ = [0,\infty)$.

Lemma 4.2.4 *Let* $0 \leq \Upsilon \leq 1$ *and*

$$f : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R} : (x,y) \to \sqrt{1+x^2}\sqrt{1+y^2} - \Upsilon(x+y).$$

Then

$$\min\{f(x,y) \,|\, (x,y) \in \mathbb{R}^+ \times \mathbb{R}^+\} = 1-\Upsilon^2.$$

The proof of Lemma 4.2.4 is given in Appendix 4.A.

Theorem 4.2.5 *Assume* (4.2.15) *and* $0 \leq \gamma \leq 1$. *Then*

$$|R_2|^{-1} \geq \begin{cases} (\theta-\frac{1}{4})/\theta^2 & \text{if } \frac{1}{4} \leq \theta < \frac{1}{2},\ 0 \leq \gamma < 2\theta, \\ -3(\theta-\frac{1+\gamma}{6})(\theta-\frac{1+\gamma}{2})/\theta^2 & \text{if } \frac{1}{4} \leq \theta < \frac{1}{2},\ 2\theta \leq \gamma \leq \min\{6\theta-1,1\}, \\ 1 & \text{if } \frac{1}{2} \leq \theta. \end{cases}$$

Proof First, consider the case $\theta \geq \frac{1}{2}$ and put $\digamma = 2\theta - \frac{1}{2}$. Then,

$$\begin{aligned}|p+\tfrac{1}{2}z_0+(\tfrac{1}{2}-\theta)(z_1+z_2)| &\geq |p+(\tfrac{1}{2}-\theta)(z_1+z_2)| - \sqrt{\mathcal{R}z_1\mathcal{R}z_2} \\ &= |(\tfrac{\digamma}{\theta}-\theta z_1)(\tfrac{\digamma}{\theta}-\theta z_2)+1-\tfrac{\digamma^2}{\theta^2}| - \sqrt{\mathcal{R}z_1\mathcal{R}z_2} \\ &\geq |\tfrac{\digamma}{\theta}-\theta z_1||\tfrac{\digamma}{\theta}-\theta z_2| - |1-\tfrac{\digamma^2}{\theta^2}| - \sqrt{\mathcal{R}z_1\mathcal{R}z_2}.\end{aligned}$$

Now, since

$$\theta^2 - \digamma^2 = -3\theta^2+2\theta-\tfrac{1}{4} = -3(\theta-\tfrac{1}{6})(\theta-\tfrac{1}{2}),$$

it holds that $1-\frac{\digamma^2}{\theta^2}$ is negative. Further, since $\mathcal{R}(\frac{\digamma}{\theta}-\theta z_i) \geq 0$, we have that

$$|\tfrac{\digamma}{\theta}-\theta z_i| \geq \tfrac{\digamma}{\theta}-\theta\mathcal{R}z_i \quad \text{for } i=1,2.$$

As a consequence

$$\begin{aligned}|R_2|^{-1} &\geq \tfrac{\digamma^2}{\theta^2} - \digamma(\mathcal{R}z_1+\mathcal{R}z_2) + \theta^2\mathcal{R}z_1\mathcal{R}z_2 + 1 - \tfrac{\digamma^2}{\theta^2} - \sqrt{\mathcal{R}z_1\mathcal{R}z_2} \\ &= 1+\theta^2\mathcal{R}z_1\mathcal{R}z_2 + \digamma(\sqrt{-\mathcal{R}z_1}-\sqrt{-\mathcal{R}z_2})^2 + (2\digamma-1)\sqrt{\mathcal{R}z_1\mathcal{R}z_2} \\ &\geq 1+2(2\theta-1)\sqrt{\mathcal{R}z_1\mathcal{R}z_2} \\ &\geq 1,\end{aligned}$$

which completes the proof for $\theta \geq \frac{1}{2}$.

Next consider the case $\frac{1}{4} \leq \theta < \frac{1}{2}$ and $2\theta \leq \gamma \leq \min\{6\theta - 1, 1\}$. Define vectors

$$\boldsymbol{v}_i = \begin{bmatrix} \sqrt{-2\theta\mathcal{R}z_i} \\ \\ \sqrt{1+\theta^2|z_i|^2} \end{bmatrix} \quad \text{for } i = 1, 2.$$

By the Cauchy–Schwarz inequality, we have

$$|p| = \sqrt{\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1}\sqrt{\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_2} \geq \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_2 = 2\theta\sqrt{\mathcal{R}z_1\mathcal{R}z_2} + \sqrt{1+\theta^2|z_1|^2}\sqrt{1+\theta^2|z_2|^2}. \tag{4.2.16}$$

Also,

$$\begin{aligned} |z_1 + z_2| &= \sqrt{(\mathcal{R}z_1 + \mathcal{R}z_2)^2 + (\mathcal{I}z_1 + \mathcal{I}z_2)^2} \\ &= \sqrt{(\mathcal{R}z_1 - \mathcal{R}z_2)^2 + 4\mathcal{R}z_1\mathcal{R}z_2 + (\mathcal{I}z_1 + \mathcal{I}z_2)^2} \\ &\geq 2\sqrt{\mathcal{R}z_1\mathcal{R}z_2} \\ &\geq \frac{1}{\gamma}|z_0|. \end{aligned} \tag{4.2.17}$$

4

Using (4.2.16), (4.2.17) and $\frac{1}{2} - \frac{\theta}{\gamma} \geq 0$ we find

$$\begin{aligned} |R_2|^{-1} &\geq |p| - \tfrac{1}{2}|z_0| - (\tfrac{1}{2} - \theta)|z_1 + z_2| \\ &\geq |p| - 2\theta\sqrt{\mathcal{R}z_1\mathcal{R}z_2} - (\tfrac{1}{2} - \tfrac{\theta}{\gamma})|z_0| - (\tfrac{1}{2} - \theta)|z_1 + z_2| \\ &\geq \sqrt{1+\theta^2|z_1|^2}\sqrt{1+\theta^2|z_2|^2} - (\tfrac{1+\gamma}{2} - 2\theta)|z_1 + z_2| \\ &\geq \sqrt{1+\theta^2|z_1|^2}\sqrt{1+\theta^2|z_2|^2} - (\tfrac{1+\gamma}{2} - 2\theta)(|z_1| + |z_2|). \end{aligned}$$

Define

$$\Upsilon = (\tfrac{1+\gamma}{2} - 2\theta)/\theta.$$

It is easily verified that, in the case under consideration, $0 < \Upsilon \leq 1$ and application of Lemma 4.2.4 yields

$$|R_2|^{-1} \geq 1 - (\tfrac{1+\gamma}{2} - 2\theta)^2/\theta^2 = -3(\theta - \tfrac{1+\gamma}{6})(\theta - \tfrac{1+\gamma}{2})/\theta^2.$$

Finally consider the case $\frac{1}{4} \leq \theta < \frac{1}{2}$ and $\gamma < 2\theta$. Analogously as above one finds

$$\begin{aligned} |R_2|^{-1} &\geq |p| - 2\theta\sqrt{\mathcal{R}z_1\mathcal{R}z_2} - (\tfrac{1}{2} - \tfrac{\theta}{\gamma})|z_0| - (\tfrac{1}{2} - \theta)|z_1 + z_2| \\ &\geq \sqrt{1+\theta^2|z_1|^2}\sqrt{1+\theta^2|z_2|^2} - (\tfrac{1}{2} - \theta)|z_1 + z_2| \\ &\geq \sqrt{1+\theta^2|z_1|^2}\sqrt{1+\theta^2|z_2|^2} - (\tfrac{1}{2} - \theta)(|z_1| + |z_2|). \end{aligned}$$

Applying Lemma 4.2.4 with $\Upsilon = (\frac{1}{2} - \theta)/\theta$, it then follows that

$$|R_2|^{-1} \geq 1 - (\tfrac{1}{2} - \theta)^2/\theta^2 = (\theta - \tfrac{1}{4})/\theta^2,$$

and this completes the proof.

■

Theorem 4.2.5 directly implies the positive result that the second term in (4.2.14) is also bounded from above whenever $\{\frac{1}{4} < \theta \leq \frac{1}{3}$ and $0 \leq \gamma < 6\theta - 1\}$ or $\{\theta > \frac{1}{3}\}$.

4.3. Numerical Experiments

We present numerical experiments for the model convection-diffusion equation (4.2.13) with $(x, y) \in \Omega = (0, 1) \times (0, 1)$, $0 \le t \le 2$ and parameters

$$d_{11} = d,\ d_{12} = -2\gamma d,\ d_{22} = 4d,\ c_1 = -2,\ c_2 = -3 \quad \text{and} \quad d = 0.025,\ \gamma = 0.7. \tag{4.3.1}$$

The requirement (2.3.6) is fulfilled for this choice of parameters. We consider the initial condition

$$u(x, y, 0) = e^{-4(\sin^2 \pi x + \cos^2 \pi y)} \quad \text{for } (x, y) \in \Omega,$$

and Dirichlet boundary condition

$$u(x, y, t) = e^{-rt} u(x, y, 0) \quad \text{for } (x, y) \in \partial\Omega,\ 0 < t \le 2,$$

with $r = 0.05$. Semidiscretization of the initial-boundary value problem is performed using standard second order central finite difference schemes on a rectangular grid in Ω with spatial mesh widths $\Delta x = 1/(m_1 + 1)$ and $\Delta y = 1/(m_2+1)$. In order to avoid spurious oscillations, the convection terms u_x and u_y are discretized by second order backward finite differencing near $x = 1$ and $y = 1$, respectively. The semidiscretization leads to an initial value problem (4.1.1) with $m = m_1 m_2$ and $F(t, \boldsymbol{v}) = A\boldsymbol{v} + e^{-rt} g$ with given $m \times m$-matrix A and m-vector g. Figure 4.1 shows the semidiscrete solutions $U(0)$ and $U(2)$ on the grid in Ω if $m_1 = m_2 = 50$.

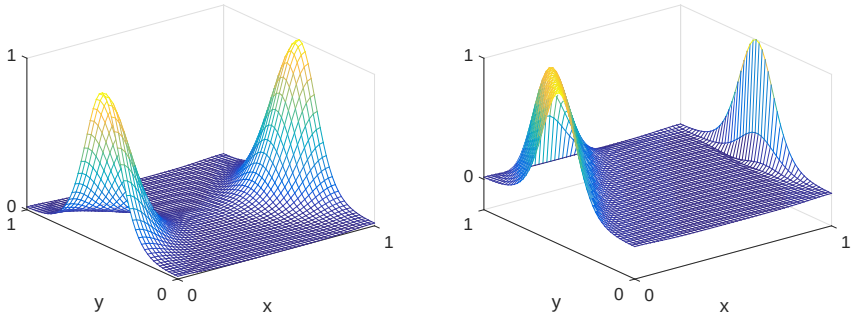


Figure 4.1: Semidiscrete solutions $U(t)$ on Ω for $t = 0, 2$ if $m_1 = m_2 = 50$.

We employ the splitting (4.1.2) of the function F with $l = 2$ and we consider application of the MCS scheme with interesting parameter values $\theta = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1$. Recall that for $\theta = \frac{1}{2}$ one recovers the CS scheme (3.2.4). Stability of the MCS scheme pertinent to two-dimensional convection-diffusion equations with mixed derivative term has been analysed in [39, 53]. Applying the results from these references to the situation at hand, we expect unconditional stability whenever $\theta \ge \frac{1}{3}$ and a lack thereof when $\theta = \frac{1}{4}$.

Figure 4.2 displays for $m_1 = m_2 \in \{50, 100, 150, 200\}$ the norms of the global discretization errors at $t = 2$ as a function of N,

$$e(N, m_1, m_2) = \|U(2) - U_{2N}\|_2,$$

where $N\Delta t = 1$ and $\|\cdot\|_2$ denotes the scaled Euclidean vector norm from Section 4.1. Here we applied the HV scheme (3.2.6) with $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ and $N = 10^4$ to obtain a reference solution $U(2)$.

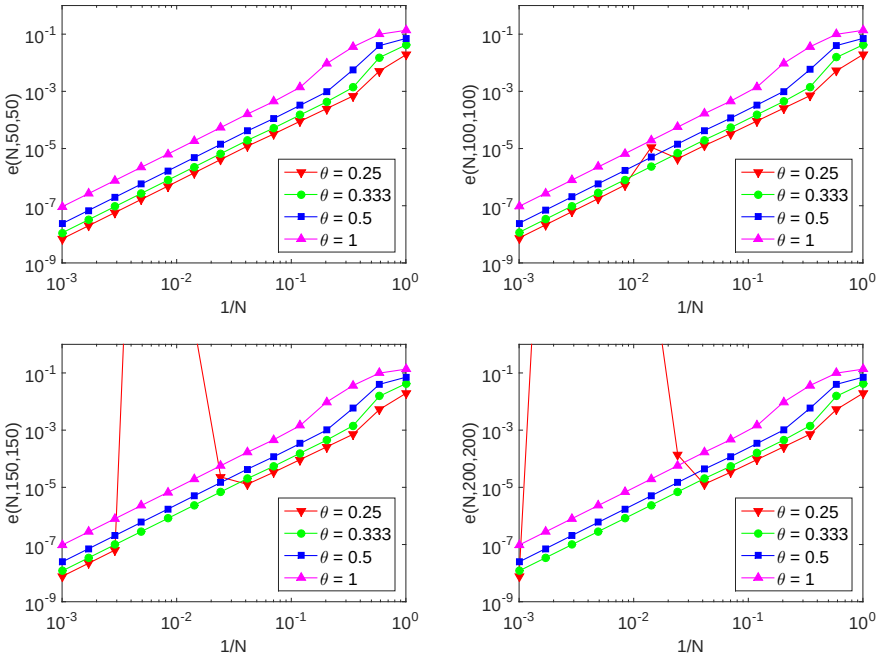


Figure 4.2: Global discretization errors $e(N, m_1, m_2)$ versus $1/N$ for $m_1 = m_2 = 50$ (top left), $m_1 = m_2 = 100$ (top right), $m_1 = m_2 = 150$ (bottom left) and $m_1 = m_2 = 200$ (bottom right) with $\theta = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1$.

For $\theta = \frac{1}{3}, \frac{1}{2}, 1$ the results of Figure 4.2 clearly reveal a second order convergence behaviour in Δt, uniformly in the spatial mesh width. This positive conclusion is in line with our theory of Section 4.2. For $\theta = \frac{1}{4}$, a second order convergence behaviour uniformly in the spatial mesh width is clearly absent. This conclusion also agrees with the theory of Section 4.2. The observed strong increase in the global discretization errors as the spatial mesh width decreases corresponds to a lack of unconditional stability when $\theta = \frac{1}{4}$ (cf. above).

We have performed numerical experiments for other convection-diffusion parameter sets than (4.3.1) which satisfy the condition (2.3.6) as well as for the celebrated two-dimensional Heston model [30] from financial mathematics. Semidiscretization of the latter model was performed as described in [37] on a non-uniform spatial grid, also leading to initial value problems of type (4.1.1) with semidiscrete function $F(t, \boldsymbol{v}) = A\boldsymbol{v} + g(t)$. In all of the experiments, we found the obtained conclusions concerning the temporal convergence behaviour of the MCS scheme to be in line with the theory of Section 4.2.

4.4. Conclusion

Under some natural stability and smoothness assumptions we have proved a useful second order convergence result for the MCS scheme in the application to two-dimensional time-dependent convection-diffusion equations with mixed derivative term. Here the obtained bound on the global temporal discretization errors has the important property that it is independent of the (arbitrarily small) spatial mesh width from the semidiscretization. Based on the convergence analysis in the present chapter and the stability results from [39, 53], we recommend to select, in the application to these equations, the parameter of the MCS scheme such that $\theta \geq \frac{1}{3}$. Numerical experiments further indicate that a smaller parameter value often yields a smaller error constant. In future research we wish to extend our convergence results, among others, to higher-dimensional problems.

4.A. Proof of Lemma 4.2.4

For the case $\Upsilon = 0$ the result is trivial. Next, consider the case $0 < \Upsilon < 1$. In order to find the minimum of f on its domain we determine first its stationary points in $(0, \infty) \times (0, \infty)$. These are given by

$$\begin{cases} f_x(x,y) = x\frac{\sqrt{1+y^2}}{\sqrt{1+x^2}} - \Upsilon = 0, \\ f_y(x,y) = y\frac{\sqrt{1+x^2}}{\sqrt{1+y^2}} - \Upsilon = 0. \end{cases} \tag{4.A.1}$$

From (4.A.1) it follows that

$$xy = \Upsilon^2,$$

which yields that x and y are nonzero and

$$y = \frac{\Upsilon^2}{x}. \tag{4.A.2}$$

From (4.A.1) it also follows that

$$\frac{x}{1+x^2} = \frac{y}{1+y^2}.$$

Inserting (4.A.2) into this yields

$$\frac{x}{1+x^2} = \frac{x\Upsilon^2}{x^2+\Upsilon^4},$$

which simplifies to

$$\left(1-\Upsilon^2\right)x^2 = \Upsilon^2(1-\Upsilon^2).$$

Because

$$\nu := 1 - \Upsilon^2 > 0,$$

this factor can be divided out. We thus conclude that

$$x = \Upsilon > 0,$$

and by (4.A.2),

$$y = \Upsilon = x.$$

Hence, the system (4.A.1) has precisely one solution, given by

$$(x, y) = (\Upsilon, \Upsilon)\,. \tag{4.A.3}$$

We next prove that f possesses a relative minimum in its stationary point (4.A.3) by showing that $f_{xx} > 0$, $f_{yy} > 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ in this point. For arbitrary (x, y) there holds

$$\begin{aligned} f_{xx}(x,y) &= \tfrac{\sqrt{1+y^2}}{(1+x^2)^{3/2}} > 0, \\ f_{yy}(x,y) &= \tfrac{\sqrt{1+x^2}}{(1+y^2)^{3/2}} > 0, \\ f_{xy}(x,y) &= \tfrac{xy}{\sqrt{1+x^2}\sqrt{1+y^2}} \end{aligned}$$

and

$$(f_{xx}f_{yy} - f_{xy}^2)(x,y) = \frac{1 - x^2y^2}{(1+x^2)(1+y^2)} = \frac{(1-xy)(1+xy)}{(1+x^2)(1+y^2)}.$$

4

It is therefore sufficient to prove that $1 - xy$ is strictly positive in the point (4.A.3) and indeed $1 - \Upsilon^2 = \nu > 0$. Hence, f has a relative minimum in (4.A.3), where it takes the value

$$1 + \Upsilon^2 - \Upsilon(\Upsilon + \Upsilon) = \nu > 0.$$

It remains to prove that on the boundary of its domain f is greater than the value ν. First,

$$\begin{aligned} f(x,y) &\geq \sqrt{1+x^2} - \Upsilon(x+y) \\ &\geq x - \Upsilon(x+y) \\ &= (1-\Upsilon)x - \Upsilon y. \end{aligned}$$

Thus for any given fixed $y \in \mathbb{R}^+$ there holds

$$\lim_{x\to\infty} f(x,y) = \infty.$$

Since $f(x, y) = f(y, x)$ for all (x, y) in the domain of f, it also holds for any given fixed $x \in \mathbb{R}^+$ that

$$\lim_{y\to\infty} f(x,y) = \infty.$$

We finally show that f is always greater than ν whenever $x = 0$ or $y = 0$. By the same symmetry argument as above, it suffices to consider only $y = 0$. Define

$$g : \mathbb{R}^+ \to \mathbb{R} : x \to \sqrt{1+x^2} - \Upsilon x,$$

so that $g(x) = f(x, 0)$. Then

$$g(0) = 1 > \nu \quad \text{and} \quad \lim_{x\to\infty} g(x) \geq \lim_{x\to\infty} (1-\Upsilon)x = \infty.$$

Next,

$$\begin{cases} g_x(x) = \frac{x}{\sqrt{1+x^2}} - \Upsilon, \\ g_{xx}(x) = \frac{1}{(1+x^2)^{3/2}} > 0. \end{cases}$$

Putting $g_x(x) = 0$, it readily follows that g has one relative minimum, which is in the point

$$x = \sqrt{\frac{1-\nu}{\nu}},$$

where it takes the value

$$g(x) = \sqrt{\nu}.$$

Since

$$\min_{x \in \mathbb{R}^+} g(x) = \sqrt{\nu} > \nu$$

the proof is complete for $0 < \Upsilon < 1$. For the case $\Upsilon = 1$ the result of the lemma is easily obtained by a continuity argument.

■

CHAPTER 5

Convergence of the HV Scheme

5.1. Introduction

Next to the MCS scheme, the HV scheme (3.2.6) forms a second prominent ADI time stepping method in computational finance. A linearised version of the HV scheme was designed in [67] as Rosenbrock-type method for the numerical solution of initial-boundary value problems for PDEs from chemistry. The general method (3.2.6) is introduced in [34]. The application of the HV scheme to equations containing mixed derivative terms was first studied in [41, 42].

We consider the HV scheme with uniform temporal step size for the temporal discretization of initial value problems for large systems of stiff ODEs (4.1.1) that allow a splitting of the type (4.1.2). A stability analysis for the pertinent ADI scheme, in application to semidiscretized multidimensional convection-diffusion equations with mixed derivative terms, has been performed in [40–42]. An overview of the results is given in Chapter 3. A theoretical second order convergence result for the HV scheme is presented in [34] for the case $l = 1$. A rigorous convergence analysis for larger values of l appears to be lacking at this moment.

In this chapter, we derive a second order convergence theorem for the HV scheme that is directly relevant to two-dimensional convection-diffusion equations from financial mathematics. Our analysis is similar to that in Chapter 4, and inspired by that of Hundsdorfer [33, 34].

The chapter is organised as follows. In Section 5.2 a recursion formula for the total error is derived by considering a perturbed version of the HV scheme. Taylor expansions for the perturbations are taken from [34] and lead to an expression for the local discretization error. A subtle splitting of the latter error, and application of Lemma 4.2.1, eventually leads to a second order convergence result under some stability assumptions. Positive theoretical results on the stability assumptions are again obtained in the von Neumann framework. The numerical experiments in Section 5.3 confirm the relevance of our theoretical analysis and Section 5.4 concludes.

This chapter is based on the article 'Convergence of the Hundsdorfer–Verwer scheme for two-dimensional convection-diffusion equations with mixed derivative term', published in AIP Conf. Proc., 1648:850054-1–850054-5, 2015 [43].

5.2. Convergence Analysis

5.2.1. Preliminaries

Assume (4.1.1) stems from spatial discretization of a linear convection-diffusion problem with mixed derivative terms, the semidiscrete system can be decomposed as

$$F(t,\boldsymbol{v}) = F_0(t,\boldsymbol{v}) + F_1(t,\boldsymbol{v}) + \cdots + F_l(t,\boldsymbol{v}), \quad \text{for } 0 \le t \le T,\ \boldsymbol{v} \in \mathbb{R}^m, \tag{5.2.1}$$

and

$$F(t,\boldsymbol{v}) = A\boldsymbol{v}+g(t), \quad F_i(t,\boldsymbol{v}) = A_i\boldsymbol{v}+g_i(t), \quad \text{for } 0 \le t \le T,\ \boldsymbol{v} \in \mathbb{R}^m, 0 \le i \le l,$$

with given real $m \times m$-matrices A and A_i, $0 \le i \le l$, and given real m-vector valued functions g and g_i, $0 \le i \le l$. Here m denotes the number of spatial grid points. Analogously to Chapter 4, to simplify notation, define the matrices

$$Z = \Delta t A, \quad Z_i = \Delta t A_i, \quad Q_i = I - \theta Z_i, \quad P = Q_1 Q_2 \cdots Q_l, \quad \text{for } 1 \le i \le l,$$

where it is important to emphasize that the temporal step size Δt is assumed to be uniform. We consider the norm $\|\cdot\|_2$ induced by the (naturally scaled) inner product $(\boldsymbol{v},\boldsymbol{w}) = \frac{1}{m}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}$ on $\mathbb{R}^m$ and assume that the semidiscretization satisfies

$$(A_i\boldsymbol{v},\boldsymbol{v}) \le 0 \quad \text{whenever } \boldsymbol{v} \in \mathbb{R}^m,\ 1 \le i \le l. \tag{5.2.2}$$

This implies, cf. [34] and Subsection 4.2.1, that the Q_i and P are invertible and

$$\|Q_i^{-1}\|_2 \le 1, \quad \text{for } 1 \le i \le l, \quad \|P^{-1}\|_2 \le 1. \tag{5.2.3}$$

5.2.2. Error Recursion and Local Discretization Errors

Consider along with (3.2.6) the perturbed scheme

$$\begin{cases} Y_0^* = U_{n-1}^* + \Delta t\, F(t_{n-1}, U_{n-1}^*) + \rho_0, \\ Y_i^* = Y_{i-1}^* + \theta\Delta t \left(F_i(t_n, Y_i^*) - F_i(t_{n-1}, U_{n-1}^*)\right) + \rho_i, \quad \text{for } i = 1,2,\ldots,l, \\ \widetilde{Y}_0^* = Y_0^* + \frac{1}{2}\Delta t \left(F(t_n, Y_l^*) - F(t_{n-1}, U_{n-1}^*)\right) + \widetilde{\rho}_0, \\ \widetilde{Y}_i^* = \widetilde{Y}_{i-1}^* + \theta\Delta t \left(F_i(t_n, \widetilde{Y}_i^*) - F_i(t_n, Y_l^*) + \widetilde{\rho}_i, \quad \text{for } i = 1,2,\ldots,l, \right. \\ \left. U_n^* = \widetilde{Y}_l^*, \right. \end{cases} \tag{5.2.4}$$

and let the error be denoted by

$$e_n = U_n^* - U_n.$$

By performing an analysis similar to that in Chapter 4, see also [34], it follows that e_n satisfies

$$e_n = Re_{n-1} + d_n, \tag{5.2.5}$$

with *stability matrix*

$$R = I + 2P^{-1}Z - P^{-2}Z + \tfrac{1}{2}(P^{-1}Z)^2,$$

and vector

$$d_n = (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)(P^{-1}\rho_0 + \sum_{i=1}^{l} Q_l^{-1}Q_{l-1}^{-1}\dots Q_i^{-1}\rho_i)$$
$$+ P^{-1}(\rho_0 + \widetilde{\rho}_0) + \sum_{i=1}^{l} Q_l^{-1}Q_{l-1}^{-1}\dots Q_i^{-1}\widetilde{\rho}_i. \tag{5.2.6}$$

Next, consider the perturbed HV scheme (5.2.4) with constant step size and let the perturbations be defined by

$$U_{n-1}^* = U(t_{n-1}), \quad Y_i^* = \widetilde{Y}_i^* = U(t_n), \quad \text{for } 0 \le i \le l,$$

such that d_n is the *local discretization error* and $e_n = U(t_n) - U_n$ the *global discretization error* in the n-th step. Let

$$\varphi_i(t) = F_i(t, U(t)), \quad \text{for } 0 \le t \le T,\ 0 \le i \le l,$$

and assume that U and the φ_i are sufficiently often differentiable and that their derivatives are bounded on $[0, T]$ uniformly in the spatial mesh width. It can be verified, see e.g. [34], that the perturbations satisfy

$$\begin{aligned}
\rho_0 &= U''(t_{n-1})\ \tfrac{1}{2}(\Delta t)^2 + U'''(t_{n-1})\tfrac{1}{6}(\Delta t)^3 + \mathcal{O}\left((\Delta t)^4\right),\\
\widetilde{\rho}_0 &= -U''(t_{n-1})\ \tfrac{1}{2}(\Delta t)^2 - U'''(t_{n-1})\tfrac{1}{4}(\Delta t)^3 + \mathcal{O}\left((\Delta t)^4\right),\\
\rho_i &= -\varphi_i'(t_{n-1})\ \theta(\Delta t)^2 + \mathcal{O}\left((\Delta t)^3\right), \quad 1 \le i \le l,\\
\widetilde{\rho}_i &= 0, \quad 1 \le i \le l.
\end{aligned}$$

We consider the case $l = 2$. Inserting the expansions in (5.2.6), we obtain for the local discretization error that

$$\begin{aligned}
d_n = {} & (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)\\
& \times [\tfrac{1}{2}(\Delta t)^2 P^{-1}U''(t_{n-1}) - \theta(\Delta t)^2 P^{-1}\varphi_1'(t_{n-1}) - \theta(\Delta t)^2 Q_2^{-1}\varphi_2'(t_{n-1})]\\
& + (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)\mathcal{O}\left((\Delta t)^3\right) - \tfrac{1}{12}(\Delta t)^3 P^{-1}U'''(t_{n-1})\\
& + P^{-1}\mathcal{O}\left((\Delta t)^4\right).
\end{aligned}$$

5.2.3. Convergence Theorem for the HV Scheme

The recursion (5.2.5) yields for the global discretization error that

$$e_N = R^N e_0 + \sum_{n=1}^{N} R^{N-n} d_n,$$

provided that N is a temporal index such that $N\Delta t \le T$. From this it is clear that one can distinguish two important steps in proving convergence. First

one wishes to show stability, i.e. there exists a moderate constant M such that $\|R^n\|_2 \leq M$ uniformly in the spatial mesh width, the temporal step size Δt and n. Secondly, one wishes to prove consistency, i.e. the local discretization errors d_n tend to zero as the temporal step size Δt tends to zero, uniformly in the spatial mesh width and n.

For the analysis in the present chapter it will be assumed that the HV scheme is stable and, to simplify the analysis, that the matrices A_i, $0 \leq i \leq 2$, commute. All of the foregoing assumptions in this section were made in [34] as well. For convenience we also assume in the following that A is invertible.

In general, when the local discretization errors are of second order, one would only expect first-order convergence. Often, however, an order of convergence can be recovered through Lemma 4.2.1. As a first attempt, one could check whether the matrix $(R-I)^{-1}(I - P^{-1} + \frac{1}{2}P^{-1}Z)$ is bounded uniformly in the spatial mesh width. Unfortunately, as it turns out, this is not always the case. We consider a splitting of the local discretization error. First, rewrite d_n as

$$\begin{aligned} d_n = &(I - P^{-1} + \tfrac{1}{2}P^{-1}Z)P^{-1} \\ &\times [\tfrac{1}{2}(\Delta t)^2\varphi_0'(t_{n-1}) + (\tfrac{1}{2} - \theta)(\Delta t)^2\Sigma_{i=1}^2\varphi_i'(t_{n-1})] \\ &+ (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)P^{-1}\theta(\Delta t)^2(I - Q_1)\varphi_2'(t_{n-1}) \\ &- \tfrac{1}{12}(\Delta t)^3 P^{-1}U'''(t_{n-1}) + (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)\mathcal{O}\left((\Delta t)^3\right) \\ &+ P^{-1}\mathcal{O}\left((\Delta t)^4\right). \end{aligned}$$

Then, we split the error into four parts: $d_n = d_n^{(1)} + d_n^{(2)} + d_n^{(3)} + d_n^{(4)}$ with

$$\begin{aligned} d_n^{(1)} = &(R-I)(R-I)^{-1}(I - P^{-1} + \tfrac{1}{2}P^{-1}Z)P^{-1} \\ &\times [\tfrac{1}{2}(\Delta t)^2\varphi_0'(t_{n-1}) + (\tfrac{1}{2} - \theta)(\Delta t)^2\Sigma_{i=1}^2\varphi_i'(t_{n-1})], \\ d_n^{(2)} = &(R-I)Z^{-1}\theta^2(\Delta t)^2 Z_1\varphi_2'(t_{n-1}), \\ d_n^{(3)} = &-(R-I)(R-I)^{-1}P^{-1}\theta^2(\Delta t)^2 Z_1\varphi_2'(t_{n-1}), \\ d_n^{(4)} = &-\tfrac{1}{12}(\Delta t)^3 P^{-1}U'''(t_{n-1}) + (I - P^{-1} + \tfrac{1}{2}P^{-1}Z)\mathcal{O}\left((\Delta t)^3\right) \\ &+ P^{-1}\mathcal{O}\left((\Delta t)^4\right). \end{aligned}$$

It is clear that the second part of the local discretization error, $d_n^{(2)}$, meets the requirements for application of Lemma 4.2.1 if $A^{-1}A_1 = \mathcal{O}(1)$. Further, as P^{-1} is bounded in $\|\cdot\|_2$ by 1, the fourth part satisfies the requirements if $I - P^{-1} + \frac{1}{2}P^{-1}Z = \mathcal{O}(1)$. The third part fulfils the requirements if the matrix

$$Z^{-1}P(2I - P^{-1} + \tfrac{1}{2}P^{-1}Z)^{-1}P^{-1}Z_1, \tag{5.2.7}$$

is uniformly bounded. Using that the A_i commute, this matrix can be rewritten as

$$Z^{-1}Z_1(2I - P^{-1} + \tfrac{1}{2}P^{-1}Z)^{-1}.$$

Since it was already assumed that $A^{-1}A_1 = \mathcal{O}(1)$, we obtain that the matrix (5.2.7) is uniformly bounded if $(2I - P^{-1} + \frac{1}{2}P^{-1}Z)^{-1} = \mathcal{O}(1)$. For the first

part it is sufficient if the matrix

$$Z^{-1}P(2I - P^{-1} + \tfrac{1}{2}P^{-1}Z)^{-1}(I - P^{-1} + \tfrac{1}{2}P^{-1}Z)P^{-1} \tag{5.2.8}$$

is uniformly bounded. By using that the A_i commute, it follows that this is equivalent to the matrix

$$Z^{-1}(-\theta Z_1 - \theta Z_2 + \theta^2 Z_1 Z_2 + \tfrac{1}{2}Z)P^{-1}(2I - P^{-1} + \tfrac{1}{2}P^{-1}Z)^{-1} \tag{5.2.9}$$

being uniformly bounded. As for the second and third part, we assume that $A^{-1}A_1 = \mathcal{O}(1)$. Analogously, the assumption is made that $A^{-1}A_2 = \mathcal{O}(1)$. Next, we consider the term

$$Z^{-1}Z_1Z_2P^{-1} = Z^{-1}Z_1Z_2Q_2^{-1}Q_1^{-1}$$

in matrix (5.2.9). By (5.2.3) it holds that $\|Q_1^{-1}\|_2$ is bounded by 1 and from the assumptions above it holds that $\|Z^{-1}Z_1\|_2 = \mathcal{O}(1)$. For the remaining part we use the following theorem due to von Neumann, see e.g. [28].

Theorem 5.2.1 (von Neumann) *Let $f : \mathbb{C} \to \mathbb{C}$ be any given rational function. Suppose that f has no poles in $\mathcal{R}z \leq 0$ and let Z be a given real square matrix. If*

$$(Z\boldsymbol{v}, \boldsymbol{v}) \leq 0 \quad \text{for } \boldsymbol{v} \in \mathbb{R}^m,$$

then

$$\|f(Z)\|_2 \leq \sup\{|f(z)| : \mathcal{R}z \leq 0\}.$$

From (5.2.2) it is readily seen that

$$(Z_2\boldsymbol{v}, \boldsymbol{v}) \leq 0 \quad \text{for } \boldsymbol{v} \in \mathbb{R}^m.$$

Application of Theorem 5.2.1 with $f(z) = \frac{z}{1-\theta z}$ yields

$$\|Z_2Q_2^{-1}\|_2 \leq \tfrac{1}{\theta},$$

and uniform boundedness of (5.2.8) follows from the assumptions made above. Summarizing, we proved the next theorem.

Theorem 5.2.2 *Let $l = 2$ and consider a uniform temporal step size. Assume that U and the φ_i, $i = 0, 1, 2$, are sufficiently often differentiable and their derivatives are bounded on $[0, T]$ uniformly in the spatial mesh width. Assume A is invertible, the A_i, $i = 0, 1, 2$, commute and $(A_i\boldsymbol{v}, \boldsymbol{v}) \leq 0$ whenever $\boldsymbol{v} \in \mathbb{R}^m$ and $i = 1, 2$. Assume the HV scheme is stable, the matrix $2I - P^{-1} + \frac{1}{2}P^{-1}Z$ is invertible and the four matrices $A^{-1}A_1$, $A^{-1}A_2$, $I - P^{-1} + \frac{1}{2}P^{-1}Z$, $(2I - P^{-1} + \frac{1}{2}P^{-1}Z)^{-1}$ are all $\mathcal{O}(1)$. Then the global discretization errors satisfy $e_N = \mathcal{O}\big((\Delta t)^2\big)$.*

The above theorem extends the result of [34] from the one-dimensional ($l = 1$) to the two-dimensional case ($l = 2$) . The crucial step in our derivation is the splitting of the local discretization error into four convenient parts. The uniform boundedness of $A^{-1}A_1$ and $A^{-1}A_2$, which is often fulfilled, was also assumed in [34]. The uniform boundedness of $I - P^{-1} + \frac{1}{2}P^{-1}Z$ was tacitly

assumed in there as well. Our assumption $(2I - P^{-1} + \frac{1}{2}P^{-1}Z)^{-1} = \mathcal{O}(1)$ is new. It may be regarded as replacing the condition (29) in [34].

For a theoretical result on the condition that the matrices $I - P^{-1} + \frac{1}{2}P^{-1}Z$ and $(2I - P^{-1} + \frac{1}{2}P^{-1}Z)^{-1}$ are $\mathcal{O}(1)$, we consider a two-dimensional constant coefficient convection-diffusion equation, cf. (4.2.13),

$$u_t = d_{11}u_{xx} + 2d_{12}u_{xy} + d_{22}u_{yy} + c_1u_x + c_2u_y \tag{5.2.10}$$

for $(x, y) \in (0, 1) \times (0, 1)$ and $0 \le t \le T$, provided with periodic boundary condition. Assume the diffusion coefficients satisfy (2.3.6) with $\gamma = 1$, and semidiscretization of (5.2.10) is performed by standard finite difference schemes on uniform rectangular grids. In this case the analysis reduces, see [41], to bounding from above

$$|1 - \tfrac{1}{p} + \tfrac{1}{2}\tfrac{z_0+z_1+z_2}{p}| \quad \text{and} \quad |2 - \tfrac{1}{p} + \tfrac{1}{2}\tfrac{z_0+z_1+z_2}{p}|^{-1}$$

with $p = (1-\theta z_1)(1-\theta z_2)$ for complex numbers z_0, z_1, z_2 satisfying, cf. (4.2.15),

$$\mathcal{R}z_1 \le 0, \quad \mathcal{R}z_2 \le 0, \quad |z_0| \le 2\sqrt{\mathcal{R}z_1\mathcal{R}z_2}. \tag{5.2.11}$$

5

Theorem 5.2.3 *If* (5.2.11) *and* $\theta > 1/2$, *then*

$$|1 - \tfrac{1}{p} + \tfrac{1}{2}\tfrac{z_0+z_1+z_2}{p}| \le 2 \quad \textit{and} \quad |2 - \tfrac{1}{p} + \tfrac{1}{2}\tfrac{z_0+z_1+z_2}{p}|^{-1} \le \tfrac{2\theta}{2\theta-1}.$$

Proof By [41, Lemmas 2.1, 2.3] we have $|1 + \frac{z_0+z_1+z_2}{p}| \le 1$. From this and $|p| \ge 1$ it follows that

$$|1 - \tfrac{1}{p} + \tfrac{1}{2}\tfrac{z_0+z_1+z_2}{p}| \le \tfrac{1}{2}|1 + \tfrac{z_0+z_1+z_2}{p}| + \tfrac{1}{2} + |\tfrac{1}{p}| \le 2,$$

which proves the first part of the theorem. Next, by [41, Lemma 2.3] there holds

$$|\tfrac{z_0}{p}| + |\tfrac{1}{2\theta} + \tfrac{z_1+z_2}{p}| \le \tfrac{1}{2\theta}.$$

Consequently,

$$\begin{aligned} |2 - \tfrac{1}{p} + \tfrac{1}{2}\tfrac{z_0+z_1+z_2}{p}| &= |2 - \tfrac{1}{4\theta} - \tfrac{1}{p} + \tfrac{1}{2}\tfrac{z_0}{p} + \tfrac{1}{4\theta} + \tfrac{1}{2}\tfrac{z_1+z_2}{p}| \\ &\ge |2 - \tfrac{1}{4\theta} - \tfrac{1}{p}| - \tfrac{1}{2}\left(|\tfrac{z_0}{p}| + |\tfrac{1}{2\theta} + \tfrac{z_1+z_2}{p}|\right) \\ &\ge 1 - \tfrac{1}{4\theta} - \tfrac{1}{4\theta} \\ &= \tfrac{2\theta-1}{2\theta}, \end{aligned}$$

which yields the second part of the theorem.

■

5.3. Numerical Experiments

The assumption of commuting matrices from Theorem 5.2.2 is not always valid in practical applications. We present numerical experiments, similar to the ones in Section 4.3, to show that the convergence result is relevant for general semidiscretized two-dimensional convection-diffusion equations with mixed derivative term. Consider the model equation (5.2.10) with parameters (4.3.1) for $(x, y) \in \Omega = (0, 1) \times (0, 1)$ and $0 \leq t \leq 2$. The PDE is supplied with the same initial and boundary condition as in Section 4.3, i.e. with initial condition

$$u(x, y, 0) = e^{-4(\sin^2 \pi x + \cos^2 \pi y)} \quad \text{for } (x, y) \in \Omega,$$

and Dirichlet boundary condition

$$u(x, y, t) = e^{-rt} u(x, y, 0) \quad \text{for } (x, y) \in \partial\Omega,\ 0 < t \leq 2,$$

where $r = 0.05$.

Semidiscretization of the initial-boundary value problem is performed using standard second order central finite difference schemes on a rectangular grid in Ω with spatial mesh widths $\Delta x = 1/(m_1 + 1)$ and $\Delta y = 1/(m_2 + 1)$. The convection terms u_x and u_y are again discretized by second order backward finite differencing near $x = 1$ and $y = 1$, respectively.

We employ the splitting (5.2.1) of the function F with $l = 2$ and consider application of the HV scheme with three interesting parameter values, namely $\theta = 1/(2 + \sqrt{2}), \frac{1}{2} + \frac{1}{6}\sqrt{3}, 1$. Stability of the HV scheme pertinent to two-dimensional convection-diffusion equations with mixed derivative term has been analysed in [40, 41], cf. also Section 3.3. For pure diffusion equations unconditional stability is expected whenever $\theta \geq 1/(2 + \sqrt{2})$. There is, however, no theoretical stability result available if convection terms and a mixed derivative are present. In [41] it is conjectured that unconditional stability can be expected whenever $\theta \geq \frac{1}{2} + \frac{1}{6}\sqrt{3}$.

Figure 5.1 displays for $m_1 = m_2 \in \{50, 100, 150, 200\}$ the norms of the global discretization errors at $t = 2$ as a function of N,

$$e(N, m_1, m_2) = \|U(2) - U_{2N}\|_2,$$

where $N\Delta t = 1$ and $\|\cdot\|_2$ denotes the scaled Euclidean vector norm from Section 5.1. Here we applied the MCS scheme (3.2.5) with $\theta = 1/3$ and $N = 10^4$ to obtain a reference solution $U(2)$.

For $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}, 1$ the results of Figure 5.1 clearly reveal a second order convergence behaviour in Δt, uniformly in the spatial mesh width. Although the semidiscretization matrices do not commute, this positive conclusion confirms the relevance of our theory in Section 5.2. For $\theta = 1/(2 + \sqrt{2})$, a uniform convergence behaviour uniformly in the spatial mesh width is clearly absent. This conclusion also agrees with the theory of Section 5.2. The observed strong increase in the global discretization errors as the spatial mesh width decreases indicates a lack of unconditional stability when $\theta = 1/(2 + \sqrt{2})$. The latter observation is also made in [37] where the HV scheme is applied in numerical experiments for the well-known Heston PDE [33].

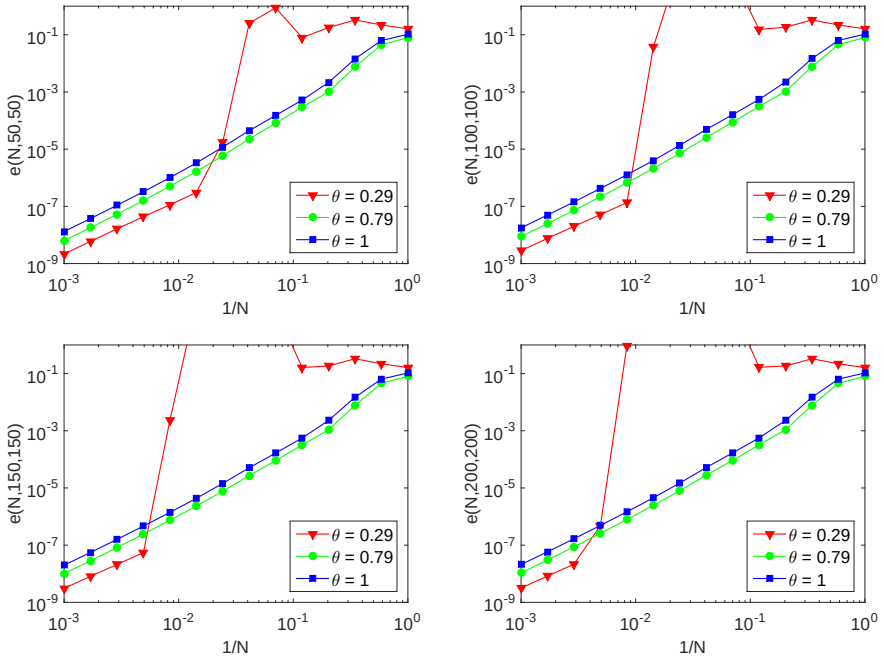


Figure 5.1: Global discretization errors $e(N, m_1, m_2)$ versus $1/N$ for $m_1 = m_2 = 50$ (top left), $m_1 = m_2 = 100$ (top right), $m_1 = m_2 = 150$ (bottom left) and $m_1 = m_2 = 200$ (bottom right) with $\theta = 1/(2+\sqrt{2}), \frac{1}{2} + \frac{1}{6}\sqrt{3}, 1$.

5.4. Conclusion

The HV scheme constitutes a popular ADI method for the effective numerical solution of multidimensional time-dependent convection-diffusion equations with mixed derivative terms. Various positive stability results have been derived in literature for this scheme. Also, a convergence result has been obtained pertinent to the special case of one-dimensional PDEs. Clearly, obtaining convergence results relevant to multidimensional PDEs is of much interest. In this chapter we studied the convergence of the HV scheme relevant to two-dimensional convection-diffusion equations with mixed derivative term. We proved that, under natural stability and smoothness conditions, the HV scheme is convergent of order two uniformly in the spatial mesh width. In future research we wish to extend the convergence analysis to, for example, higher-dimensional PDEs.

CHAPTER 6

ADI Schemes and Non-Smooth Initial Data

6.1. Introduction

In financial mathematics, the fair value $u(s_1, s_2, t)$ of a European style option on two underlying assets is modelled by the two-dimensional *Black–Scholes* PDE, see e.g. [5],

$$u_t = \tfrac{1}{2}\sigma_1^2 s_1^2 u_{s_1 s_1} + \rho\sigma_1\sigma_2 s_1 s_2 u_{s_1 s_2} + \tfrac{1}{2}\sigma_2^2 s_2^2 u_{s_2 s_2} + r s_1 u_{s_1} + r s_2 u_{s_2} - r u, \quad (6.1.1)$$

for $s_1, s_2 > 0,\ 0 < t \leq T$. Here, t denotes the time to maturity T and we assume real parameters $r, \sigma_1 > 0, \sigma_2 > 0, |\rho| < 1$. The PDE (6.1.1) is provided with an initial condition that is defined through the payoff of the option, which is often *non-smooth*.

The mixed spatial derivative term in (6.1.1) represents the correlation between the asset prices in the two-dimensional Black–Scholes model. Mixed spatial derivative terms are very important, notably, in the field of financial option valuation theory. Here they arise due to the correlation between underlying stochastic processes.

Recall that a well-known approach for determining the fair value $u(s_1, s_2, T)$ consists of numerically solving the PDE (6.1.1) by the MOL, whereby one first discretizes in space and subsequently in time. In this chapter we consider a uniform Cartesian grid and second order central finite difference schemes in space. This semidiscretization is second order convergent with respect to the spatial mesh width if the initial and boundary data is smooth, see e.g. [35]. For the effective time discretization of the resulting semidiscrete systems, we employ the operator splitting schemes of the ADI type from Chapter 3. In the past years various positive stability results for the ADI schemes have been derived relevant to multidimensional convection-diffusion equations with mixed derivative terms, cf. Section 3.3. From the analysis in [34] (cf. also Section 3.4) it follows that for the Do scheme one cannot expect an order of convergence higher than one if a mixed spatial derivative term is present. In Chapter 4 and Chapter 5, second order convergence results have been proven for the MCS scheme and the HV scheme under natural stability and smoothness

This chapter is mainly based on the article 'Convergence analysis of the Modified Craig–Sneyd scheme for two-dimensional convection-diffusion equations with nonsmooth initial data', published in IMA J. Numer. Anal., 37:798–831, 2017 [68].

assumptions. These temporal convergence results have the crucial property that they hold uniformly in the spatial mesh width. Hence, when the MCS scheme or HV scheme is used for the temporal discretization, the fully discrete numerical solution is second order convergent in space and time for smooth initial and boundary data.

A convergence analysis for the ADI schemes from Chapter 3 relevant to non-smooth data is still open in the literature. In financial applications, however, the initial function is in general non-smooth. It is well-known that convergence can then be seriously impaired. For the (sole) purpose of illustration, consider a two-asset cash-or-nothing option with strikes $K_1 > 0$ and $K_2 > 0$, so that

$$u(s_1, s_2, 0) = \mathbb{1}_{\{s_1 \geq K_1\}} \mathbb{1}_{\{s_2 \geq K_2\}},$$

where $\mathbb{1}$ denotes the indicator function. Let spatial discretization of (6.1.1) be performed by second order central finite difference schemes on a uniform Cartesian grid, and let temporal discretization be performed with the MCS scheme and MCS parameter $\theta = 1/3$. In the upper left plot in Figure 6.1, the numerical solution for $u(s_1, s_2, T)$ is shown for (natural) financial parameter values $r = 0.05$, $\sigma_1 = 0.2$, $\sigma_2 = 0.25$, $\rho = -0.7$, $K_1 = 1$, $K_2 = 1$, $T = 2$. Irregularities can be observed around the strikes, leading to a loss of accuracy in the maximum norm. For hedging reasons it is important to consider also the Greeks, for example the cross gamma $\Gamma = u_{s_1 s_2}$. The corresponding PDE is given by

$$\begin{aligned} \Gamma_t = {} & \tfrac{1}{2}\sigma_1^2 s_1^2 \Gamma_{s_1 s_1} + \rho\sigma_1\sigma_2 s_1 s_2 \Gamma_{s_1 s_2} + \tfrac{1}{2}\sigma_2^2 s_2^2 \Gamma_{s_2 s_2} \\ & + (r + \sigma_1^2 + \rho\sigma_1\sigma_2) s_1 \Gamma_{s_1} + (r + \sigma_2^2 + \rho\sigma_1\sigma_2) s_2 \Gamma_{s_2} + (r + \rho\sigma_1\sigma_2)\Gamma, \end{aligned} \tag{6.1.2}$$

for $s_1, s_2 > 0$, $0 < t \leq T$. This is supplemented with initial function

$$\Gamma(s_1, s_2, 0) = u_{s_1 s_2}(s_1, s_2, 0) = \delta(s_1 - K_1)\delta(s_2 - K_2),$$

where δ is the *Dirac delta* function. The lower left plot in Figure 6.1 shows the numerical solution for the cross gamma at maturity T for the same financial parameter values as above. Around the point $(s_1, s_2) = (K_1, K_2)$ strong, spurious erratic behaviour shows up and, hence, this approximation is useless in practice. If the cross gamma is approximated by applying finite difference schemes directly to the numerical solution for the option value, which is a common alternative technique in practice, the same observations are found. Additional numerical experiments reveal that similar erratic behaviour occurs when temporal discretization is performed with the Do scheme or HV scheme.

For one-dimensional applications in finance, the impact of non-smooth initial data on convergence has already been studied extensively and various techniques have been proposed in order to recover standard convergence results, see e.g. [22, 56]. A common technique consists of first applying several implicit Euler (sub)steps and then continue with the time stepping scheme under consideration, [57]. This is called *Rannacher time stepping* or *implicit Euler damping*.

Consider again PDEs (6.1.1) and (6.1.2) for the example of the two-asset cash-or-nothing option. Replacing the MCS scheme in the first two time steps

by four half-time steps of the implicit Euler scheme, the two right plots in Figure 6.1 are obtained. Clearly, there are no longer irregularities or oscillations present. In many other multidimensional applications, see e.g. [37], similar observations were made. To the best of our knowledge, however, there are no theoretical results available concerning the favourable effect of Rannacher time stepping on the convergence of ADI schemes if the initial data is non-smooth.

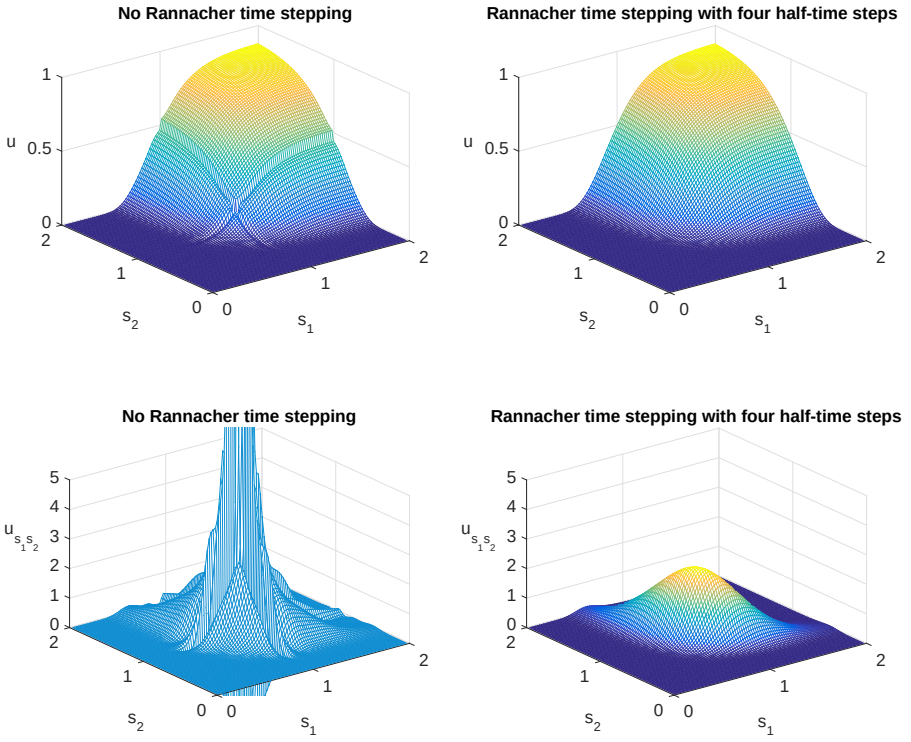


Figure 6.1: Numerical approximations of the cash-or-nothing option value (top) and of its cross gamma (bottom) without (left) and with (right) Rannacher time stepping with four half-time steps. The financial parameter values are $r = 0.05$, $\sigma_1 = 0.2$, $\sigma_2 = 0.25$, $\rho = -0.7$, $K_1 = 1$, $K_2 = 1$, $T = 2$.

In the present chapter we will prove a useful convergence bound for the MCS scheme when it is applied to a model two-dimensional convection-diffusion equation with mixed derivative term, provided with Dirac delta initial data. Based on numerical experiments, similar convergence results are conjectured for the Do scheme and the HV scheme. The precise influence of Rannacher time stepping on the order of convergence will be investigated. Our analysis in this chapter is inspired by that of Giles & Carter [22], who deal with the Crank-Nicolson scheme applied to a model one-dimensional convection-diffusion equation.

The outline of the chapter is as follows. In Section 6.2 a model two-dimensional convection-diffusion equation, provided with Dirac delta initial

data, is introduced. Making use of a classical Fourier transformation leads to a closed-form analytical solution. Section 6.3 describes a numerical discretization of the pertinent PDE. Spatial discretization is performed with second order central finite difference formulas on uniform Cartesian grids. Temporal discretization of the semidiscrete system with the ADI schemes leads to fully discrete approximations of the exact solution. In Section 6.4 we present a two-dimensional mixed discrete/continuous Fourier transform pair and we derive closed-form expressions for the discrete/continuous Fourier transformation of the numerical solutions. Studying the Fourier error, the difference between the Fourier transformation of the exact solution and of the numerical solution, reveals that this error has different properties in different parts of the Fourier domain. Section 6.5 analyses the asymptotic behaviour of the Fourier error corresponding to the numerical solution obtained with the MCS scheme. We partition the Fourier domain into five disjoint regions. By Taylor expansion we arrive at an expression for the Fourier error in each of the regions. In Section 6.6, application of the inverse transformation leads to an error bound in physical space for the MCS scheme and the CS scheme. The sharpness of the error bound is confirmed by ample numerical experiments. The main results of this chapter are formulated in Theorems 6.6.1 and 6.6.2. Section 6.7, respectively Section 6.8, deals with the convergence of the Do scheme, respectively the HV scheme. We perform numerical experiments and use our insights from the analysis for the MCS scheme to conjecture a convergence result for the Do scheme and the HV scheme. The final Section 6.9 gives concluding remarks.

6.2. Model Problem

Consider the coordinate transformation

$$x = \sqrt{2}\log(s_1)/\sigma_1 \quad \text{and} \quad y = \sqrt{2}\log(s_2)/\sigma_2.$$

The PDEs (6.1.1), (6.1.2) are then transformed into

$$\begin{aligned}
u_t &= u_{xx} + 2\rho u_{xy} + u_{yy} + (\tfrac{\sqrt{2}r}{\sigma_1} - \tfrac{\sigma_1}{\sqrt{2}})u_x + (\tfrac{\sqrt{2}r}{\sigma_2} - \tfrac{\sigma_2}{\sqrt{2}})u_y - ru,\\
\Gamma_t &= \Gamma_{xx} + 2\rho\Gamma_{xy} + \Gamma_{yy} + (r + \rho\sigma_1\sigma_2)\Gamma\\
&\quad + \left[(r + \sigma_1^2 + \rho\sigma_1\sigma_2)\tfrac{\sqrt{2}}{\sigma_1} - \tfrac{\sigma_1}{\sqrt{2}}\right]\Gamma_x + \left[(r + \sigma_2^2 + \rho\sigma_1\sigma_2)\tfrac{\sqrt{2}}{\sigma_2} - \tfrac{\sigma_2}{\sqrt{2}}\right]\Gamma_y,
\end{aligned}$$

for $-\infty < x, y < \infty,\ 0 < t \leq T$. This provides a motivation for considering a general constant coefficient convection-diffusion equation with mixed derivative term

$$u_t = u_{xx} + 2\rho u_{xy} + u_{yy} + a_1 u_x + a_2 u_y, \tag{6.2.1}$$

for $-\infty < x, y < \infty,\ 0 < t \leq T = 1$ and with $|\rho| < 1$. We supplement model equation (6.2.1) with the initial condition

$$u(x, y, 0) = \delta(x)\delta(y),$$

which arises for example in the case of the cross gamma of a two-asset cash-or-nothing option. The Dirac delta initial function, however, has other important

applications as well. For instance, it arises naturally in the adjoint equation for the joint density, cf. Chapter 7 and Chapter 8. By using the *Fourier transform* pair

$$\hat{u}(\omega_1,\omega_2,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x,y,t)\exp(-\mathbf{i}\omega_1 x)\exp(-\mathbf{i}\omega_2 y)dxdy,$$
$$u(x,y,t) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{u}(\omega_1,\omega_2,t)\exp(\mathbf{i}\omega_1 x)\exp(\mathbf{i}\omega_2 y)d\omega_1 d\omega_2,$$

an exact closed-form analytical solution will be derived. Here $\mathbf{i}$ denotes the imaginary unit. Taking the Fourier transformation of equation (6.2.1) yields the ODE

$$\hat{u}_t = -\omega_1^2\hat{u} - 2\rho\omega_1\omega_2\hat{u} - \omega_2^2\hat{u} + \mathbf{i}a_1\omega_1\hat{u} + \mathbf{i}a_2\omega_2\hat{u},$$

subject to initial condition $\hat{u}(\omega_1,\omega_2,0) = 1$. The solution of this transformed equation is given by

$$\hat{u}(\omega_1,\omega_2,t) = \exp(-(\omega_1^2 + 2\rho\omega_1\omega_2 + \omega_2^2 - \mathbf{i}a_1\omega_1 - \mathbf{i}a_2\omega_2)t). \tag{6.2.2}$$

Next, if (X_1, X_2) is a multivariate normally distributed random variable with mean (μ_1,μ_2) and covariance matrix Σ, its characteristic function is defined by

$$\mathbb{E}[\exp(\mathbf{i}\omega_1 X_1)\exp(\mathbf{i}\omega_2 X_2)] = \exp(\mathbf{i}\omega_1\mu_1 + \mathbf{i}\omega_2\mu_2 - \tfrac{1}{2}(\omega_1\ \omega_2)\Sigma(\omega_1\ \omega_2)^{\mathrm{T}}).$$

By exploring the connection between the characteristic function of a random variable and the Fourier transform of its density function, it follows that $u(x,y,t)$ can be seen as the density function of a two-dimensional normally distributed random variable with mean (μ_1,μ_2) and covariance matrix Σ given by

$$(\mu_1,\mu_2) = (-a_1 t, -a_2 t) \quad \text{and} \quad \Sigma = \begin{pmatrix} 2t & 2\rho t \\ 2\rho t & 2t \end{pmatrix}.$$

Since $|\rho| < 1$, this yields a closed-form analytical solution for $u(x,y,t)$:

$$\frac{1}{4\pi t\sqrt{1-\rho^2}}\exp\left(\frac{-1}{4t(1-\rho^2)}[(x+a_1t)^2 + (y+a_2t)^2 - 2\rho(x+a_1t)(y+a_2t)]\right).$$

6.3. Spatial and Temporal Discretization

As mentioned in Section 6.1, spatial discretization of (6.2.1) will be performed on a uniform Cartesian grid with second order central finite difference schemes. For the time integration the ADI schemes from Chapter 3 will be considered. The theoretical convergence analysis in this chapter is, however, restricted to the numerical approximations obtained with the MCS scheme. Let h_1 denote the spatial mesh width in the x-direction, h_2 the spatial mesh width in the y-direction and define spatial grid points $(x_j, y_k) = (jh_1, kh_2)$ for all $j,k \in \mathbb{Z}$. Semidiscretization of (6.2.1) with second order central finite difference schemes then gives rise to approximations $U_{j,k}(t)$ of the exact solution value $u(x_j,y_k,t)$ which are defined by the system

$$U'_{j,k}(t) = AU_{j,k}(t), \tag{6.3.1}$$

where $A = A_0 + A_1 + A_2$ and

$$A_0 = \tfrac{\rho}{2h_1h_2}\delta_{2x}\delta_{2y}, \quad A_1 = \tfrac{1}{h_1^2}\delta_x^2 + \tfrac{a_1}{2h_1}\delta_{2x}, \quad A_2 = \tfrac{1}{h_2^2}\delta_y^2 + \tfrac{a_2}{2h_2}\delta_{2y},$$

with $\delta_{2x},\ \delta_x^2,\ \delta_{2y},\ \delta_y^2$ the usual second order central finite difference operators. For example,

$$\begin{aligned}
\delta_{2x}U_{j,k}(t) &= U_{j+1,k}(t) - U_{j-1,k}(t),\\
\delta_x^2U_{j,k}(t) &= U_{j-1,k}(t) - 2U_{j,k}(t) + U_{j+1,k}(t),\\
\delta_{2x}\delta_{2y}U_{j,k}(t) &= U_{j+1,k+1}(t) + U_{j-1,k-1}(t) - U_{j+1,k-1} - U_{j-1,k+1}(t).
\end{aligned}$$

Semidiscrete system (6.3.1) is provided with initial data

$$U_{j,k}(0) = \begin{cases} \frac{1}{h_1h_2} & \text{if } j = k = 0,\\ 0 & \text{else,} \end{cases} \tag{6.3.2}$$

in order to approximate the Dirac delta initial function. Let $\theta > 0$ again denote the given parameter for the ADI scheme, $N \geq 1$ the number of time steps and set $t_n = n\Delta t$ with $\Delta t = T/N$. For convenience we define

$$Z = \Delta t A, \quad Z_i = \Delta t A_i, \quad \text{for } i = 0, 1, 2,$$

and we denote by I the identity operator. Then, starting from $U_{0,j,k} = U_{j,k}(0)$, application of the MCS scheme (3.2.5) to semidiscrete system (6.3.1) yields approximations $U_{n,j,k}$ of $U_{j,k}(t_n)$ successively for $n = 1, 2, \ldots, N$ through

$$\begin{cases}
Y_{0,j,k} = (I + Z)U_{n-1,j,k},\\
(I - \theta Z_i)Y_{i,j,k} = Y_{i-1,j,k} - \theta Z_i U_{n-1,j,k} \quad i = 1, 2,\\
\widehat{Y}_{0,j,k} = Y_{0,j,k} + \theta Z_0 Y_{2,j,k} - \theta Z_0 U_{n-1,j,k},\\
\widetilde{Y}_{0,j,k} = \widehat{Y}_{0,j,k} + (\frac{1}{2} - \theta)ZY_{2,j,k} - (\frac{1}{2} - \theta)ZU_{n-1,j,k},\\
(I - \theta Z_i)\widetilde{Y}_{i,j,k} = \widetilde{Y}_{i-1,j,k} - \theta Z_i U_{n-1,j,k} \quad i = 1, 2,\\
U_{n,j,k} = \widetilde{Y}_{2,j,k}.
\end{cases} \tag{6.3.3}$$

Application of the Do scheme or HV scheme to the semidiscrete system (6.3.1) can be performed analogously. The resulting approximations of $U_{j,k}(t_n)$ are denoted by $U_{n,j,k}^{\mathrm{Do}}$, respectively $U_{n,j,k}^{\mathrm{HV}}$.

Concerning the Rannacher time stepping, let N_0 denote the number of initial ADI time steps replaced by $2N_0$ half-time steps of implicit Euler integration. Consider for example temporal discretization with the MCS scheme. Whenever $N_0 > 0$ scheme (6.3.3) is replaced by

$$\begin{cases}
(I - \frac{1}{2}Z)U_{n-1/2,j,k} = U_{n-1,j,k},\\
(I - \frac{1}{2}Z)U_{n,j,k} = U_{n-1/2,j,k},
\end{cases} \tag{6.3.4}$$

for $n = 1, 2, \ldots, \min\{N_0, N\}$. This provides a numerical approximation U_N of the exact solution. The numerical approximation of the exact solution obtained with the Do scheme, respectively the HV scheme, is denoted by U_N^{Do}, respectively U_N^{HV}.

Consider a general semidiscrete system (6.3.1) and temporal discretization with the MCS scheme. If the operators A_0, A_1, A_2 commute and one replaces scheme (6.3.3) by scheme (6.3.4) for N_0 arbitrary different steps in $\{1, 2, \ldots, N\}$, one will always arrive at the same final approximation U_N. For example, this is the case for the spatial discretization defined in the current section and the corresponding A_0, A_1, A_2. If the implicit Euler steps are not performed directly at the beginning, however, then the (positive) effect of Rannacher time stepping will not be present in the $U_{n,j,k}$ for all $n < N$ when they are considered as approximations of the values $u(x_j, y_k, t_n)$. The same observation can be made if the temporal discretization is performed with the Do scheme or HV scheme.

The goal of our theoretical convergence analysis consists of quantifying the *total error*

$$U_{N,j,k} - u(x_j, y_k, 1), \tag{6.3.5}$$

i.e. the difference between the numerical solution and the exact solution of the model problem, given that the former one is obtained by applying second order central finite difference schemes for the spatial discretization and the MCS scheme for the temporal discretization. To do so, we analyse the asymptotic behaviour of a mixed discrete/continuous Fourier transform for $h_1, h_2, \Delta t$ simultaneously tending to zero. Applying the inverse Fourier transformation on the resulting error in Fourier space will yield a useful bound for the total error (6.3.5). Special attention will be paid to the influence of N_0, i.e. the influence of Rannacher time stepping, on the total error. The influence of Rannacher time stepping on the order of convergence of the Do scheme and HV scheme will be analysed numerically.

6.4. Discrete Fourier Transformation

We consider a mixed discrete/continuous Fourier transform pair, cf. e.g. [62],

$$\widehat{V}(\vartheta_1, \vartheta_2) = h_1 h_2 \sum_{j,k=-\infty}^{\infty} V_{j,k} \exp(-\mathbf{i} j \vartheta_1) \exp(-\mathbf{i} k \vartheta_2), \quad -\pi \leq \vartheta_1, \vartheta_2 \leq \pi,$$

$$V_{j,k} = \frac{1}{4\pi^2 h_1 h_2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \widehat{V}(\vartheta_1, \vartheta_2) \exp(\mathbf{i} j \vartheta_1) \exp(\mathbf{i} k \vartheta_2) d\vartheta_1 d\vartheta_2, \quad j, k \in \mathbb{Z}.$$

For ease of presentation, the dependency of the Fourier transform on ϑ_1 and ϑ_2 will be omitted in the notation. Fourier transformation of $U_{0,j,k}$ yields $\widehat{U}_0 = 1$. Concerning operator Z_0 it follows that

$$\begin{aligned}
\widehat{Z_0 V} &= h_1 h_2 \sum_{j,k=-\infty}^{\infty} Z_0 V_{j,k} e^{-\mathbf{i} j \vartheta_1} e^{-\mathbf{i} k \vartheta_2} \\
&= \tfrac{\rho \Delta t}{2} \sum_{j,k=-\infty}^{\infty} (V_{j+1,k+1} + V_{j-1,k-1} - V_{j+1,k-1} - V_{j-1,k+1}) e^{-\mathbf{i}(j\vartheta_1 + k\vartheta_2)} \\
&= \tfrac{\rho \Delta t}{2 h_1 h_2} \left[e^{\mathbf{i}\vartheta_1} e^{\mathbf{i}\vartheta_2} + e^{-\mathbf{i}\vartheta_1} e^{-\mathbf{i}\vartheta_2} - e^{\mathbf{i}\vartheta_1} e^{-\mathbf{i}\vartheta_2} - e^{-\mathbf{i}\vartheta_1} e^{\mathbf{i}\vartheta_2} \right] \widehat{V} \\
&= -\tfrac{2\rho \Delta t}{h_1 h_2} (\sin\vartheta_1 \sin\vartheta_2) \widehat{V}.
\end{aligned}$$

Analogously one finds

$$\widehat{Z_1 V} = \left(-\tfrac{4\Delta t}{h_1^2}\sin^2\tfrac{\vartheta_1}{2} + \mathbf{i}\,a_1\tfrac{\Delta t}{h_1}\sin\vartheta_1\right)\widehat{V},$$
$$\widehat{Z_2 V} = \left(-\tfrac{4\Delta t}{h_2^2}\sin^2\tfrac{\vartheta_2}{2} + \mathbf{i}\,a_2\tfrac{\Delta t}{h_2}\sin\vartheta_2\right)\widehat{V}.$$

Define functions

$$z_0 = z_0(\vartheta_1,\vartheta_2) = -\tfrac{2\rho\Delta t}{h_1 h_2}\sin\vartheta_1\sin\vartheta_2,$$
$$z_1 = z_1(\vartheta_1,\vartheta_2) = -\tfrac{4\Delta t}{h_1^2}\sin^2\tfrac{\vartheta_1}{2} + \mathbf{i}\,a_1\tfrac{\Delta t}{h_1}\sin\vartheta_1,$$
$$z_2 = z_2(\vartheta_1,\vartheta_2) = -\tfrac{4\Delta t}{h_2^2}\sin^2\tfrac{\vartheta_2}{2} + \mathbf{i}\,a_2\tfrac{\Delta t}{h_2}\sin\vartheta_2,$$

and $z = z_0 + z_1 + z_2$. Then, Fourier transformation of the implicit Euler scheme (6.3.4) gives

$$\widehat{U}_n = \left(\frac{1}{1-\frac{1}{2}z}\right)^2 \widehat{U}_{n-1}.$$

After some calculations, Fourier transformation of the MCS scheme (6.3.3) yields

$$\widehat{U}_n = R\widehat{U}_{n-1},$$

with

$$R = 1 + \frac{z}{p} + \frac{(\theta z_0 + (\frac{1}{2}-\theta)z)z}{p^2},$$

where

$$p = (1-\theta z_1)(1-\theta z_2). \tag{6.4.1}$$

Assume that $N_0 \le N$. Since $\widehat{U}_0 = 1$ it follows that

$$\widehat{U}_N = R^{N-N_0}\left(\frac{1}{1-\frac{1}{2}z}\right)^{2N_0}. \tag{6.4.2}$$

A similar expression is valid for the Fourier transformation of U_N^{Do} and U_N^{HV}. It is readily shown that

$$\widehat{U}_N^{\mathrm{Do}} = \left(1+\frac{z}{p}\right)^{N-N_0}\left(\frac{1}{1-\frac{1}{2}z}\right)^{2N_0}, \tag{6.4.3}$$

$$\widehat{U}_N^{\mathrm{HV}} = \left(1+2\frac{z}{p}-\frac{z}{p^2}+\frac{z^2}{p^2}\right)^{N-N_0}\left(\frac{1}{1-\frac{1}{2}z}\right)^{2N_0}. \tag{6.4.4}$$

6.5. Asymptotic Analysis in Fourier Space for the MCS Scheme

By applying the inverse Fourier transformation to (6.4.2), the numerical approximation at $t = T = 1$ can be written as

$$\begin{aligned} U_{N,j,k} &= \tfrac{1}{4\pi^2 h_1 h_2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\widehat{U}_N(\vartheta_1,\vartheta_2)\exp(\mathbf{i}\,j\vartheta_1)\exp(\mathbf{i}\,k\vartheta_2)d\vartheta_1 d\vartheta_2 \\ &= \tfrac{1}{4\pi^2}\int_{-\pi/h_2}^{\pi/h_2}\int_{-\pi/h_1}^{\pi/h_1}\widehat{U}_N(\omega_1 h_1,\omega_2 h_2)\exp(\mathbf{i}\,x_j\omega_1)\exp(\mathbf{i}\,y_k\omega_2)d\omega_1 d\omega_2, \end{aligned}$$

where we made use of the substitutions

$$\vartheta_1 = \omega_1 h_1, \quad \vartheta_2 = \omega_2 h_2.$$

From Section 6.2 it can be seen that the exact solution is given by

$$u(x,y,1) = \tfrac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \widehat{u}(\omega_1,\omega_2,1) \exp(\mathbf{i}x\omega_1) \exp(\mathbf{i}y\omega_2) d\omega_1 d\omega_2.$$

In our analysis in this section, we will examine the *Fourier error*

$$\widehat{E}(\omega_1,\omega_2) = \widehat{U}_N(\omega_1 h_1, \omega_2 h_2) - \widehat{u}(\omega_1,\omega_2,1), \quad \text{for } -\pi \le \omega_1 h_1, \omega_2 h_2 \le \pi. \tag{6.5.1}$$

For h_1, h_2 tending to zero, the *total error* (6.3.5) is approximated by

$$\tfrac{1}{4\pi^2} \int_{-\pi/h_2}^{\pi/h_2} \int_{-\pi/h_1}^{\pi/h_1} \widehat{E}(\omega_1,\omega_2) \exp(\mathbf{i}x_j\omega_1) \exp(\mathbf{i}y_k\omega_2) d\omega_1 d\omega_2. \tag{6.5.2}$$

Note that expression (6.5.2) can be viewed as the inverse mixed discrete/continuous Fourier transform of the Fourier error.

6.5.1. Partitioning of the Fourier Domain

It turns out that the Fourier error (6.5.1) has different properties in different parts of the Fourier domain. This is illustrated in Figure 6.2 and Figure 6.3. In the former one, the modulus $|\widehat{u}|$ is shown in the $(\vartheta_1, \vartheta_2)$-domain for parameter values $\rho = -0.7$, $a_1 = 2$, $a_2 = 3$. In the latter one $|\widehat{U}_N|$ is shown for the same parameter values and discretization is performed with $h_1 = h_2 = 1/6, \Delta t = 1/8$ and well-known MCS parameters $\theta = 1/3, 1/2, 1$. For the Rannacher time stepping we considered values $N_0 = 0, 2$. From Figure 6.2 and Figure 6.3 it follows that the Fourier domain can be partitioned into five regions where the difference $\widehat{U}_N - \widehat{u}$ behaves differently. These regions are illustrated in Figure 6.4.

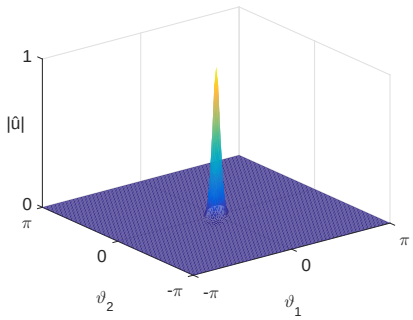


Figure 6.2: Magnitude of the Fourier transform of the exact solution $u(x,y,1)$ for parameter values $\rho = -0.7, a_1 = 2, a_2 = 3$.

First there is a *low-wavenumber region* ①, where both $|\vartheta_1|$ and $|\vartheta_2|$ are small, in which there is a good agreement between $\widehat{U}_N$ and $\widehat{u}$. Next, if either

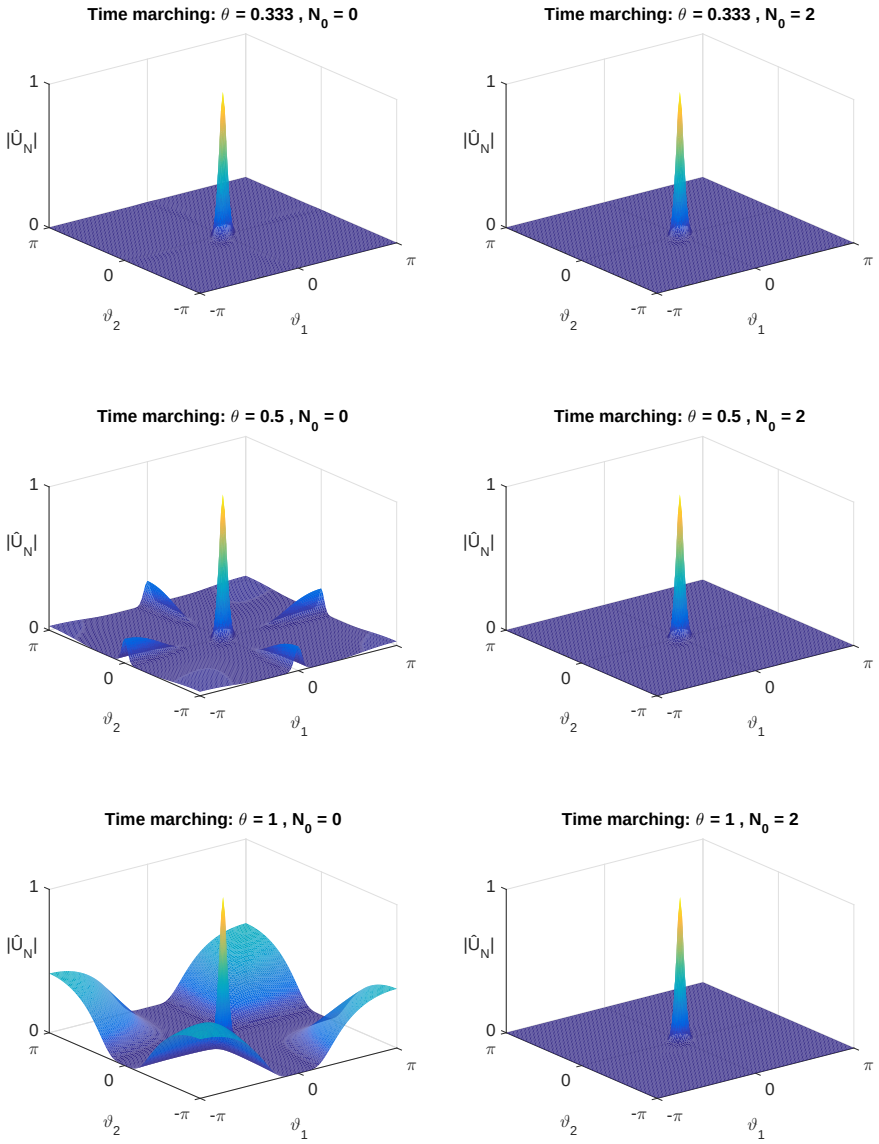


Figure 6.3: Magnitude of the Fourier transform $\widehat{U}_N$ with $N_0 = 0$ (left) and $N_0 = 2$ (right) for MCS parameter $\theta = 1/3$ (top), $\theta = 1/2$ (middle) and $\theta = 1$ (bottom). The other parameter values are: $\rho = -0.7, a_1 = 2, a_2 = 3, h_1 = h_2 = 1/6, \Delta t = 1/8$.

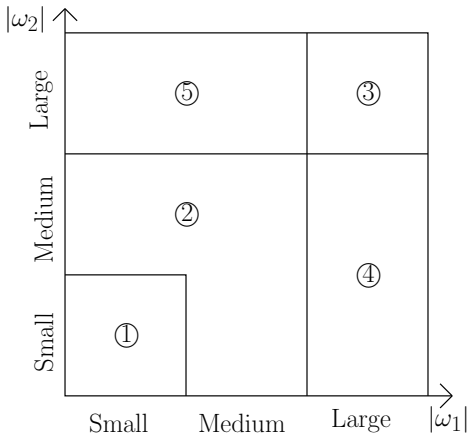


Figure 6.4: Illustration of the different disjoint regions of the Fourier domain.

$|\vartheta_1|$ or $|\vartheta_2|$ is medium and the other one is small or medium (region ②), then both the Fourier transforms of the numerical solution and analytical solution are negligible. In the *high-wavenumber region* ③, i.e. where both $|\vartheta_1|, |\vartheta_2|$ are large, we observe that the modulus of the Fourier transform $\widehat{u}$ is close to zero. The modulus $|\widehat{U}_N|$, however, is strongly dependent on N_0 and the MCS parameter θ. For larger values of θ we see that $\widehat{U}_N$ has a larger magnitude in the high-wavenumber region. Hence, a larger high-wavenumber error can be expected for larger values of θ. Further we observe that the modulus of $\widehat{U}_N$ in the high-wavenumber region is always damped whenever Rannacher time stepping is applied. This matches our observations from Figure 6.1 where unwanted erratic behaviour was avoided by using Rannacher time stepping. Finally, we have the case where either $|\vartheta_1|$ or $|\vartheta_2|$ is large but the other one is not. In our analysis, the region ④ where $|\vartheta_1|$ is large and the region ⑤ where $|\vartheta_2|$ is large will be treated separately. In both regions the Fourier transform $\widehat{u}$ is negligible but $\widehat{U}_N$ has to be further analysed. In particular, we will show that $\widehat{U}_N$ is not negligible if the MCS scheme reduces to the CS scheme.

Following Giles & Carter [22] we will perform an asymptotic analysis of the Fourier error $\widehat{U}_N - \widehat{u}$ in each of these five disjoint regions which form a partition of the Fourier domain. We consider the limit $h_1, h_2, \Delta t \to 0$ and since the same discretization is performed in both spatial directions,

$$c = h_2/h_1$$

is held fixed. For ease of presentation we denote $h = h_1$. Further, since both the semidiscretization and the time integration are convergent of order two for smooth initial data, it seems natural to keep

$$\lambda = \Delta t/h$$

constant. Substitutions $\vartheta_1 = \omega_1 h_1, \vartheta_2 = \omega_2 h_2$ yield

$$\begin{aligned} z_0 &= -\tfrac{2\rho\lambda}{ch}\sin\omega_1 h \sin c\omega_2 h \\ &= -\tfrac{\rho\lambda}{ch}(\cos((\omega_1 - c\omega_2)h) - \cos((\omega_1 + c\omega_2)h)), \end{aligned} \tag{6.5.3a}$$

$$\begin{aligned} z_1 &= -\tfrac{4\lambda}{h}\sin^2\tfrac{\omega_1 h}{2} + \mathbf{i}a_1\lambda\sin\omega_1 h \\ &= -\tfrac{2\lambda}{h}(1-\cos\omega_1 h) + \mathbf{i}a_1\lambda\sin\omega_1 h, \end{aligned} \tag{6.5.3b}$$

$$\begin{aligned} z_2 &= -\tfrac{4\lambda}{c^2 h}\sin^2\tfrac{c\omega_2 h}{2} + \mathbf{i}a_2\tfrac{\lambda}{c}\sin c\omega_2 h \\ &= -\tfrac{2\lambda}{c^2 h}(1-\cos c\omega_2 h) + \mathbf{i}a_2\tfrac{\lambda}{c}\sin c\omega_2 h. \end{aligned} \tag{6.5.3c}$$

The expressions in (6.5.3) will be used to analyse the asymptotic behaviour of (6.4.2) in every region as $h \to 0$. *Throughout the analysis in this chapter, by the notation* $\mathcal{O}\left(f(\omega_1, \omega_2, h)\right)$ *we shall always mean that the modulus* $|\cdot|$ *of the term under consideration is bounded by a positive constant times* $f(\omega_1, \omega_2, h)$ *where the constant is independent of* ω_1, ω_2 *and the mesh width* h. In order to deal with the powers in expression (6.4.2) a log-transformation of $\widehat{U}_N$ will be considered. Since $T = 1$, thus $N = 1/(\lambda h)$, it follows that

$$\log(\widehat{U}_N) = (N - N_0)\log(R) + 2N_0\log\left(\tfrac{1}{1-z/2}\right) \tag{6.5.4}$$

$$= \tfrac{1}{\lambda h}\left[\log(p^2 + pz + \theta z_0 z + (\tfrac{1}{2} - \theta)z^2) - 2\log(p)\right] \tag{6.5.5}$$

$$+ N_0\left[2\log(p) - \log(p^2 + pz + \theta z_0 z + (\tfrac{1}{2} - \theta)z^2) - 2\log(1 - \tfrac{1}{2}z)\right]. \tag{6.5.6}$$

6.5.2. Taylor Expansion of $\widehat{U}_N$

Multiple regions will encounter values $|\omega_1|, |c\omega_2| \le h^{-\mathfrak{q}}$ with certain $\mathfrak{q} \le 1/2$. By *Taylor expansion* of (6.5.3a) it directly follows that

$$\begin{aligned} z_0(h) &= -\tfrac{\rho\lambda}{ch}\left(\tfrac{(\omega_1+c\omega_2)^2h^2}{2} - \tfrac{(\omega_1-c\omega_2)^2h^2}{2} - \tfrac{(\omega_1+c\omega_2)^4h^4}{4!} + \tfrac{(\omega_1-c\omega_2)^4h^4}{4!} + \cdots\right) \\ &= z_0^{[1]}h + z_0^{[3]}h^3 + z_0^{[5]}h^5, \end{aligned}$$

where

$$z_0^{[1]} = -2\rho\lambda\omega_1\omega_2, \quad z_0^{[3]} = \tfrac{1}{3}\rho\lambda(\omega_1^2 + c^2\omega_2^2)\omega_1\omega_2, \quad |z_0^{[5]}| \le \tfrac{4}{6!}\tfrac{\rho\lambda}{c}(|\omega_1| + c|\omega_2|)^6.$$

Analogously as above, Taylor expansion of (6.5.3b) and (6.5.3c) yields

$$\begin{aligned} z_1(h) &= z_1^{[1]}h + z_1^{[3]}h^3 + z_1^{[5]}h^5, \\ z_2(h) &= z_2^{[1]}h + z_2^{[3]}h^3 + z_2^{[5]}h^5, \end{aligned}$$

where

$$\begin{aligned} &z_1^{[1]} = -\lambda\omega_1^2 + \mathbf{i}a_1\lambda\omega_1, \quad z_1^{[3]} = \tfrac{1}{12}\lambda\omega_1^4 - \tfrac{1}{6}\mathbf{i}a_1\lambda\omega_1^3, \\ &z_2^{[1]} = -\lambda\omega_2^2 + \mathbf{i}a_2\lambda\omega_2, \quad z_2^{[3]} = \tfrac{1}{12}\lambda c^2\omega_2^4 - \tfrac{1}{6}\mathbf{i}a_2\lambda c^2\omega_2^3, \\ &|z_1^{[5]}| \le \tfrac{2}{6!}\lambda\omega_1^6 + \tfrac{1}{5!}|a_1|\lambda|\omega_1|^5, \quad |z_2^{[5]}| \le \tfrac{2}{6!}\lambda c^4\omega_2^6 + \tfrac{1}{5!}|a_2|\lambda c^4|\omega_2|^5. \end{aligned}$$

Since $\mathfrak{q} \leq 1/2$, it is ensured that all terms in the above expansions stay bounded as h tends to zero. Using these expansions and the definition (6.4.1) of p it follows that

$$p(h) = 1 + p^{[1]}h + p^{[2]}h^2 + p^{[3]}h^3 + p^{[4]}h^4 + p^{[5]}h^5,$$

where

$$p^{[1]} = -\theta(z_1^{[1]} + z_2^{[1]}), \quad p^{[2]} = \theta^2 z_1^{[1]} z_2^{[1]}, \quad p^{[3]} = -\theta(z_1^{[3]} + z_2^{[3]}),$$
$$p^{[4]} = \theta^2(z_1^{[1]} z_2^{[3]} + z_1^{[3]} z_2^{[1]}), \quad p^{[5]} = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^3\right).$$

Under the condition $|\omega_1|, |c\omega_2| \leq h^{-\mathfrak{q}}$ with certain $\mathfrak{q} \leq 1/2$, the variables ω_1 and ω_2 can become very large as h tends to zero. The highest powers of ω_1, ω_2 will then dominate the order term in $p^{[5]}$. Under the same condition, however, ω_1 and ω_2 can both be very small and then the lowest powers of ω_1, ω_2 will dominate. By considering the sum of 1 and the highest powers of ω_1, ω_2 in the remaining order term, we ensure that both cases are covered.

As mentioned above we will make use of log-transformation (6.5.4) to analyse the asymptotic behaviour. Let f be a strictly positive and sufficiently smooth function and set

$$g(h) = \log(f(h)) \quad \text{for } h \geq 0.$$

Taylor expansion yields

$$g(h) = \log(f(0)) + g^{[1]}h + g^{[2]}h^2 + g^{[3]}h^3 + g^{[4]}h^4, \tag{6.5.7}$$

where

$$g^{[1]} = \tfrac{f'(0)}{f(0)},$$
$$g^{[2]} = \tfrac{1}{2}\left(\tfrac{f''(0)}{f(0)} - \tfrac{f'(0)^2}{f(0)^2}\right),$$
$$g^{[3]} = \tfrac{1}{6}\left(\tfrac{f'''(0)}{f(0)} - 3\tfrac{f'(0)f''(0)}{f(0)^2} + 2\tfrac{f'0)^3}{f(0)^3}\right),$$
$$g^{[4]} = \tfrac{1}{4!}\left(\tfrac{f^{(4)}(\xi)}{f(\xi)} - \tfrac{4f'(\xi)f'''(\xi)+3f''(\xi)^2}{f(\xi)^2} + 12\tfrac{f'(\xi)^2 f''(\xi)}{f(\xi)^3} - 6\tfrac{f'(\xi)^4}{f(\xi)^4}\right),$$

for certain $0 < \xi < h$. In order to analyse the argument of the first logarithm in (6.5.5) consider

$$f_M(h) = p(h)^2 + p(h)z(h) + \theta z_0(h)z(h) + (\tfrac{1}{2} - \theta)z(h)^2.$$

It readily follows that

$$f_M(0) = 1,$$
$$f_M'(0) = 2p'(0) + z'(0),$$
$$f_M''(0) = 2p'(0)^2 + 2p''(0) + 2p'(0)z'(0) + 2\theta z_0'(0)z'(0) + 2(\tfrac{1}{2} - \theta)z'(0)^2,$$
$$f_M'''(0) = 6p'(0)p''(0) + 2p'''(0) + 3p''(0)z'(0) + z'''(0),$$
$$f_M^4(h) = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^4\right).$$

Concerning the Rannacher time stepping, define

$$f_{N_0}(h) = 1 - \tfrac{1}{2}z(h),$$

such that $f_{N_0}(0) = 1$ and

$$f_{N_0}^{(i)}(0) = -\tfrac{1}{2}z^{(i)}(0) \quad \text{for } i = 1,2,3, \quad f_{N_0}^{(4)}(h) = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^2\right).$$

All of these expressions will be used in the forthcoming subsections, where we analyse the asymptotic behaviour of $\widehat{U}_N$ in five different regions of the Fourier domain, i.e. the (ω_1, ω_2)-domain with $|\omega_1|, |\omega_2| \leq \pi/h$.

6.5.3. Region 1: $|\boldsymbol{\omega_1}|, |\mathbf{c}\boldsymbol{\omega_2}| \leq \mathbf{h}^{-\mathfrak{q}}$ with $\mathfrak{q} < 1/3$

In order to analyse $\log \widehat{U}_N$ in this region, the parts stemming from the MCS scheme and Rannacher time stepping will be considered separately. Write (6.5.5) as

$$\tfrac{1}{\lambda h}\left[\log(f_M(h)) - 2\log(p(h))\right].$$

By using the expansions in Subsection 6.5.2 and after simplifying the resulting expressions it follows that

$$\tfrac{1}{\lambda h}\left[\log(f_M(h)) - 2\log(p(h))\right] = s^{[0]} + s^{[1]}h + s^{[2]}h^2 + s^{[3]}h^3, \tag{6.5.8}$$

where

$$\begin{aligned}
s^{[0]} &= -\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2,\\
s^{[1]} &= 0,\\
s^{[2]} &= \tfrac{1}{12}\omega_1^4 + \tfrac{1}{3}\rho(\omega_1^2 + c^2\omega_2^2)\omega_1\omega_2 + \tfrac{1}{12}c^2\omega_2^4 - \tfrac{1}{6}\mathbf{i}a_1\omega_1^3 - \tfrac{1}{6}\mathbf{i}a_2c^2\omega_2^3\\
&\quad - \lambda^2\theta^2(-\omega_1^2 + \mathbf{i}a_1\omega_1)(-\omega_2^2 + \mathbf{i}a_2\omega_2)\\
&\quad \times (-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2)\\
&\quad + \tfrac{\lambda^2}{12}(-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2)^3\\
&\quad - \lambda^2(-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2)\\
&\quad \times (-\rho\omega_1\omega_2 + (\tfrac{1}{2} - \theta)(-\omega_1^2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2))^2,\\
s^{[3]} &= \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^4\right).
\end{aligned}$$

Next, consider the part stemming from the Rannacher time stepping. Using the same analysis as above, one can rewrite (6.5.6) as

$$N_0[2\log(p) - \log(f_M(h)) - 2\log(f_{N_0}(h))] = N_0^{[1]}h + N_0^{[2]}h^2 + N_0^{[3]}h^3, \tag{6.5.9}$$

where

$$N_0^{[1]} = 0, \quad N_0^{[2]} = \tfrac{1}{4}N_0\lambda^2(-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2)^2,$$

and

$$N_0^{[3]} = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^3\right).$$

By combining (6.5.4), (6.5.8) and (6.5.9) it directly follows that $\widehat{U}_N$ can be written as

$$\exp(-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2)\exp((s^{[2]} + N_0^{[2]})h^2 + (s^{[3]} + N_0^{[3]})h^3).$$

Next, we will expand the second exponential in order to compare this expression with the Fourier transform $\widehat{u}$ from (6.2.2) at $t = 1$. Let

$$e(h) = \exp(c^{[2]}h^2 + c^{[3]}h^3),$$

where

$$c^{[2]} = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^3\right),\ c^{[3]} = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^4\right),$$

then

$$\begin{aligned}
e'(h) &= (2c^{[2]}h + 3c^{[3]}h^2)e(h),\\
e''(h) &= (2c^{[2]} + 6c^{[3]}h)e(h) + (2c^{[2]}h + 3c^{[3]}h^2)^2e(h),\\
e'''(h) &= 6c^{[3]}e(h) + 3(2c^{[2]} + 6c^{[3]}h)(2c^{[2]}h + 3c^{[3]}h^2)e(h)\\
&\quad + (2c^{[2]}h + 3c^{[3]}h^2)^3e(h).
\end{aligned}$$

Since $|\omega_1|, |c\omega_2| \leq h^{-\mathfrak{q}}$ with $\mathfrak{q} < 1/3$, we have that $e(0) = 1$ and $|e(h)| \leq \exp(1)$ whenever h is sufficiently small. Hence it follows that

$$e(h) = 1 + e^{[2]}h^2 + e^{[3]}h^3,$$

with

$$\begin{aligned}
e^{[2]} &= c^{[2]},\\
e^{[3]} &= \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^4\right) + \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^6\right)h\\
&\quad + \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^9\right)h^3\\
&= \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^4\right) + \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^6\right)h,
\end{aligned}$$

where the latter equality follows from the assumption $|\omega_1|, |c\omega_2| \leq h^{-\mathfrak{q}}$ with $\mathfrak{q} < 1/3$. Finally, for this region, one arrives at the following expression for the Fourier error (6.5.1):

$$\begin{aligned}
&\widehat{u}(\omega_1, \omega_2, 1)(s^{[2]} + N_0^{[2]})h^2\\
&\quad + \widehat{u}(\omega_1, \omega_2, 1)\left(\mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^4\right)h^3 + \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^6\right)h^4\right).
\end{aligned} \tag{6.5.10}$$

Here $s^{[2]} = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^3\right)$ and $N_0^{[2]} = \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^2\right)$. For ease of presentation, their dependency on ω_1, ω_2 is omitted in the notation.

6.5.4. Region 2: $|\boldsymbol{\omega_1}| \leq \mathbf{h}^{-\mathfrak{q}_1}, |\mathbf{c}\boldsymbol{\omega_2}| \leq \mathbf{h}^{-\mathfrak{q}_2}$ with $\mathfrak{q}_1, \mathfrak{q}_2 \leq \mathbf{1/2}$ and $|\boldsymbol{\omega_1}| \geq \mathbf{h}^{-\mathbf{1/3}}$ or $|\mathbf{c}\boldsymbol{\omega_2}| \geq \mathbf{h}^{-\mathbf{1/3}}$

First consider the case where both $\mathfrak{q}_1 < 1/2$ and $\mathfrak{q}_2 < 1/2$. Based on the analysis in Subsection 6.5.3, expression (6.5.5) can be rewritten as

$$N\log(R) = \tfrac{1}{\lambda h}\left[\log(p^2 + pz + \theta z_0 z + (\tfrac{1}{2} - \theta)z^2) - 2\log(p)\right] = s^{[0]} + s^{[2']}h^2,$$

where

$$s^{[0]} = -\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2,$$
$$s^{[2']} = \mathcal{O}\left((\omega_1^2 + c^2\omega_2^2)^3\right).$$

Since either ω_1 or ω_2 becomes large in this region as h tends to zero, only the highest powers of ω_1, ω_2 are taken into account in the order term in $s^{[2']}$. From $|\rho| < 1$ one gets

$$\omega_1^2 + 2\rho\omega_1\omega_2 + \omega_2^2 = (1 - |\rho|)(\omega_1^2 + \omega_2^2) + |\rho|(\omega_1 + \mathrm{sgn}(\rho)\omega_2)^2 > 0,$$

such that

$$\mathcal{R}(s^{[0]}) \leq -(1 - |\rho|)(\omega_1^2 + \omega_2^2) < 0.$$

Using that both $\mathfrak{q}_1 < 1/2$ and $\mathfrak{q}_2 < 1/2$, it directly follows that

$$\lim_{h\to 0}(\omega_1^2 + c^2\omega_2^2)^2 h^2 = 0,$$

and thus

$$\exists\ \delta > 0 \quad \exists\ h_0 > 0 \quad \forall\ h \leq h_0 :\ \mathcal{R}(N\log(R)) \leq -\delta(\omega_1^2 + \omega_2^2).$$

Hence, for $h \leq h_0$

$$|R^N| \leq \exp(-\delta(\omega_1^2 + \omega_2^2)),$$

and since $|\omega_1| \geq h^{-1/3}$ or $|c\omega_2| \geq h^{-1/3}$ we may conclude

$$|R^N| = \mathcal{O}\left(h^w\right) \quad \forall w > 0. \tag{6.5.11}$$

Next, we consider the case where at least one of the equalities, $\mathfrak{q}_1 = 1/2$ or $\mathfrak{q}_2 = 1/2$, holds. For analysing the asymptotic behaviour of R we make use of the following proposition. Its proof is a direct modification of the proof of one of the statements in [38, Theorem 1] and is therefore omitted.

Proposition 6.5.1 *Let $\widetilde{z_0}, \widetilde{z_1}, \widetilde{z_2}$ denote real numbers with*

$$\widetilde{z_1} \leq 0, \quad \widetilde{z_2} \leq 0, \quad |\widetilde{z_0}| \leq 2|\rho|\sqrt{\widetilde{z_1}\widetilde{z_2}}, \tag{6.5.12}$$

and $|\rho| < 1$. Set $\widetilde{z} := \widetilde{z_0} + \widetilde{z_1} + \widetilde{z_2}$ and $\widetilde{p} := (1 - \theta\widetilde{z_1})(1 - \theta\widetilde{z_2})$. If $\widetilde{z_1} < 0$ or $\widetilde{z_2} < 0$, then

$$\left|\frac{\widetilde{p}^2 + \widetilde{p}\widetilde{z} + \theta\widetilde{z_0}\widetilde{z} + (\frac{1}{2} - \theta)\widetilde{z}^2}{\widetilde{p}^2}\right| < 1,$$

whenever $\theta \geq \frac{1}{4}$ and $\theta > \frac{|\rho|+1}{6}$.

Recall that in the current region the assumptions $|\omega_1| \leq h^{-\mathfrak{q}_1}$ and $|c\omega_2| \leq h^{-\mathfrak{q}_2}$ with $\mathfrak{q}_1, \mathfrak{q}_2 \leq 1/2$ holds. This yields

$$\lim_{h\to 0} z_0(h) = \lim_{h\to 0} -2\rho\lambda\omega_1\omega_2 h =: \widetilde{z_0} \in \mathbb{R},$$
$$\lim_{h\to 0} z_1(h) = \lim_{h\to 0} -\lambda\omega_1^2 h =: \widetilde{z_1} \in \mathbb{R}^-,$$
$$\lim_{h\to 0} z_2(h) = \lim_{h\to 0} -\lambda\omega_2^2 h =: \widetilde{z_2} \in \mathbb{R}^-.$$

Since $|\omega_1| = h^{-1/2}$ or $|c\omega_2| = h^{-1/2}$ it follows that $\widetilde{z_1} < 0$ or $\widetilde{z_2} < 0$. Hence, all the assumptions on $\widetilde{z_0}, \widetilde{z_1}, \widetilde{z_2}$ in Proposition 6.5.1 are fulfilled such that

$$\lim_{h\to 0} |R| = \left| \frac{\widetilde{p}^2 + \widetilde{p}\widetilde{z} + \theta\widetilde{z_0}\widetilde{z} + (\frac{1}{2} - \theta)\widetilde{z}^2}{\widetilde{p}^2} \right| < 1,$$

and thus

$$|R^N| = |R|^{1/(\lambda h)} = \mathcal{O}(h^w) \quad \forall w > 0, \tag{6.5.13}$$

for

$$\theta \geq \tfrac{1}{4} \quad \text{and} \quad \theta > \tfrac{1+|\rho|}{6}. \tag{6.5.14}$$

Further, it always holds that $\mathcal{R}(z) \leq 0$ such that

$$|1 - \tfrac{1}{2}z|^{-1} \leq 1.$$

By combining this with (6.5.11) and (6.5.13), and by using that N_0 is independent from h, one may conclude that in this region it holds that

$$|\widehat{U}_N| = |R^N||R^{-N_0}||1 - \tfrac{1}{2}z|^{-2N_0} = \mathcal{O}\left(h^w\right) \quad \forall w > 0,$$

under restriction (6.5.14) on θ. This means that $|\widehat{U}_N|$ quickly becomes negligible as h tends to zero. It decays faster to zero than any polynomial in h.

6.5.5. Region 3: $|\boldsymbol{\omega_1}|, |\mathbf{c}\boldsymbol{\omega_2}| \geq \mathbf{h}^{-\mathfrak{q}}$ with $\mathfrak{q} > 1/2$

Here we reconsider the substitutions $\vartheta_1 = \omega_1 h_1 = \omega_1 h,\ \vartheta_2 = \omega_2 h_2 = \omega_2 ch$ in order to get

$$\begin{aligned}
z_0 &= -2\rho\tfrac{\lambda}{ch}\sin\vartheta_1 \sin\vartheta_2, \\
z_1 &= -4\tfrac{\lambda}{h}\sin^2\tfrac{\vartheta_1}{2} + \mathbf{i}a_1\lambda\sin\vartheta_1, \\
z_2 &= -4\tfrac{\lambda}{c^2h}\sin^2\tfrac{\vartheta_2}{2} + \mathbf{i}a_2\tfrac{\lambda}{c}\sin\vartheta_2.
\end{aligned}$$

Further, in this region ϑ_1, ϑ_2 are different from zero and since we consider values $-\pi \leq \vartheta_1, \vartheta_2 \leq \pi$, we may write

$$\begin{aligned}
\tfrac{c^2}{16\frac{\lambda^2}{h^2}\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}} z_0 &= -\rho\tfrac{c\cot\frac{\vartheta_1}{2}\cot\frac{\vartheta_2}{2}}{2\lambda}h, \\
\tfrac{c^2}{16\frac{\lambda^2}{h^2}\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}} z_1 &= -\tfrac{c^2}{4\lambda\sin^2\frac{\vartheta_2}{2}}h + \mathbf{i}a_1\tfrac{c^2\cot\frac{\vartheta_1}{2}}{8\lambda\sin^2\frac{\vartheta_2}{2}}h^2, \\
\tfrac{c^2}{16\frac{\lambda^2}{h^2}\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}} z_2 &= -\tfrac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}}h + \mathbf{i}a_2\tfrac{c\cot\frac{\vartheta_2}{2}}{8\lambda\sin^2\frac{\vartheta_1}{2}}h^2,
\end{aligned}$$

and

$$\tfrac{c^2}{16\frac{\lambda^2}{h^2}\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}} p$$

equals

$$\theta^2+\theta\left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}}-\theta\mathbf{i}a_1\frac{\cot\frac{\vartheta_1}{2}}{2}+\frac{c^2}{4\lambda\sin^2\frac{\vartheta_2}{2}}-\theta\mathbf{i}a_2\frac{c\cot\frac{\vartheta_2}{2}}{2}\right)h$$
$$+\left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}}-\theta\mathbf{i}a_1\frac{\cot\frac{\vartheta_1}{2}}{2}\right)\left(\frac{c^2}{4\lambda\sin^2\frac{\vartheta_2}{2}}-\theta\mathbf{i}a_2\frac{c\cot\frac{\vartheta_2}{2}}{2}\right)h^2.$$

Making use of an expansion similar to (6.5.7) it follows that

$$\log\left[\left(\frac{c^2}{16\frac{\lambda^2}{h^2}\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\right)^2\left(p^2+pz+\theta z_0 z+(\tfrac{1}{2}-\theta)z^2\right)\right]$$
$$=\log\theta^4$$
$$+\left[2\theta^3\left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}}-\theta\mathbf{i}a_1\frac{\cot\frac{\vartheta_1}{2}}{2}+\frac{c^2}{4\lambda\sin^2\frac{\vartheta_2}{2}}-\theta\mathbf{i}a_2\frac{c\cot\frac{\vartheta_2}{2}}{2}\right)\right.$$
$$\left.-\theta^2\rho\frac{c\cot\frac{\vartheta_1}{2}\cot\frac{\vartheta_2}{2}}{2\lambda}-\theta^2\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}}-\theta^2\frac{c^2}{4\lambda\sin^2\frac{\vartheta_2}{2}}\right]\frac{h}{\theta^4}$$
$$+\mathcal{O}\left(\left(\frac{1}{|\sin\frac{\vartheta_1}{2}|}+\frac{c}{|\sin\frac{\vartheta_2}{2}|}\right)^4h^2\right),$$

and

$$\log\left[\frac{c^2}{16\frac{\lambda^2}{h^2}\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}p\right]$$

can be written as

$$\log\theta^2+\left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}}-\theta\mathbf{i}a_1\frac{\cot\frac{\vartheta_1}{2}}{2}+\frac{c^2}{4\lambda\sin^2\frac{\vartheta_2}{2}}-\theta\mathbf{i}a_2\frac{c\cot\frac{\vartheta_2}{2}}{2}\right)\frac{h}{\theta}$$
$$+\mathcal{O}\left(\left(\frac{1}{|\sin\frac{\vartheta_1}{2}|}+\frac{c}{|\sin\frac{\vartheta_2}{2}|}\right)^4h^2\right).$$

Combining both expressions yields

$$\log(R)=-\frac{1}{4\lambda\theta^2}\left(2\rho c\cot\tfrac{\vartheta_1}{2}\cot\tfrac{\vartheta_2}{2}+\frac{1}{\sin^2\frac{\vartheta_1}{2}}+\frac{c^2}{\sin^2\frac{\vartheta_2}{2}}\right)h$$
$$+\mathcal{O}\left(\left(\frac{1}{|\sin\frac{\vartheta_1}{2}|}+\frac{c}{|\sin\frac{\vartheta_2}{2}|}\right)^4h^2\right)$$
$$=-\frac{1}{4\lambda\theta^2}\frac{c^2\sin^2\frac{\vartheta_1}{2}+2\rho c\cos\frac{\vartheta_1}{2}\sin\frac{\vartheta_1}{2}\cos\frac{\vartheta_2}{2}\sin\frac{\vartheta_2}{2}+\sin^2\frac{\vartheta_2}{2}}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}h$$
$$+\mathcal{O}\left(\left(\frac{1}{|\sin\frac{\vartheta_1}{2}|}+\frac{c}{|\sin\frac{\vartheta_2}{2}|}\right)^4h^2\right).$$

Further, recall that in this region ϑ_1,ϑ_2 are both different from zero such that

$$\iota(\vartheta_1,\vartheta_2):=c^2\sin^2\tfrac{\vartheta_1}{2}+2\rho c\cos\tfrac{\vartheta_1}{2}\sin\tfrac{\vartheta_1}{2}\cos\tfrac{\vartheta_2}{2}\sin\tfrac{\vartheta_2}{2}+\sin^2\tfrac{\vartheta_2}{2}>0,\quad(6.5.15)$$

and hence

$$\log(R^N) = -\tfrac{1}{4\lambda^2\theta^2}\tfrac{\iota(\vartheta_1,\vartheta_2)}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\left(1+\mathcal{O}\left(\left(\tfrac{1}{|\sin\frac{\vartheta_1}{2}|}+\tfrac{c}{|\sin\frac{\vartheta_2}{2}|}\right)^2 h\right)\right).$$

As for the implicit Euler time stepping scheme we note

$$\begin{aligned}\tfrac{c^2h}{2\lambda}\left(1-\tfrac{1}{2}z\right) &= c^2\sin^2\tfrac{\vartheta_1}{2}+\tfrac{1}{2}\rho c\sin\vartheta_1\sin\vartheta_2+\sin^2\tfrac{\vartheta_2}{2}\\ &\quad+\left(\tfrac{c^2}{2\lambda}-ia_1\tfrac{c^2}{4}\sin\vartheta_1-ia_2\tfrac{c}{4}\sin\vartheta_2\right)h.\end{aligned}$$

Using once again an expansion analogous to (6.5.7) it follows that

$$\log\left(1-\tfrac{1}{2}z\right)=\log\left(\tfrac{2\lambda}{c^2h}\right)+\log\left(\iota(\vartheta_1,\vartheta_2)\right)+\ \mathcal{O}\left(\left(\tfrac{1}{|\sin\frac{\vartheta_1}{2}|}+\tfrac{c}{|\sin\frac{\vartheta_2}{2}|}\right)^2 h\right), \tag{6.5.16}$$

which yields

$$\begin{aligned}\log\left(R^{-N_0}(1-\tfrac{1}{2}z)^{-2N_0}\right) &= -2N_0\log\left(\tfrac{2\lambda}{c^2h}\right)-2N_0\log\left(\iota(\vartheta_1,\vartheta_2)\right)\\ &\quad+\mathcal{O}\left(\left(\tfrac{1}{|\sin\frac{\vartheta_1}{2}|}+\tfrac{c}{|\sin\frac{\vartheta_2}{2}|}\right)^2 h\right).\end{aligned}$$

Making use of relationship (6.4.2) one becomes an expression for the Fourier transform $\widehat{U}_N$ in this region:

$$\widehat{U}_N=\tfrac{(c^2h)^{2N_0}}{[2\lambda\iota(\vartheta_1,\vartheta_2)]^{2N_0}}\exp\left(-\tfrac{1}{4\lambda^2\theta^2}\tfrac{\iota(\vartheta_1,\vartheta_2)}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\right)\left(1+\mathcal{O}\left(\tfrac{h}{(|\vartheta_1|+|\vartheta_2|)^2}\right)\right). \tag{6.5.17}$$

In Figure 6.3 we noticed that in the high-wavenumber region, i.e. where both $|\vartheta_1|$, $|\vartheta_2|$ are large, the norm $|\widehat{U}_N|$ is highly dependent on the MCS parameter θ. This is confirmed by (6.5.17) since inequality (6.5.15) holds in the high-wavenumber region. Hence, *for larger values of the MCS parameter θ one can expect a larger high-wavenumber error.*

6.5.6. Region 4: $|\boldsymbol{\omega_1}| \geq \mathbf{h}^{-\mathfrak{q}_1}, |\mathbf{c}\boldsymbol{\omega_2}| \leq \mathbf{h}^{-\mathfrak{q}_2}$ with $\mathfrak{q}_1 > 1/2$, $\mathfrak{q}_2 \leq 1/2$

Reconsider the substitution $\vartheta_1 = \omega_1 h$ and recall that $-\pi \leq \vartheta_1 \leq \pi$. Then, as ϑ_1 is non-zero in this region, one may write

$$\begin{aligned}\lim_{h\to0}\frac{1}{4\frac{\lambda}{h}\sin^2\frac{\vartheta_1}{2}}z_1 &= \lim_{h\to0}\left(-1+\tfrac{1}{2}\mathbf{i}a_1h\cot\tfrac{\vartheta_1}{2}\right)=-1,\\ \lim_{h\to0}\frac{1}{4\frac{\lambda}{h}\sin^2\frac{\vartheta_1}{2}}z_2 &= \lim_{h\to0}\left(-\frac{1}{c^2\sin^2\frac{\vartheta_1}{2}}\sin^2\tfrac{c\omega_2h}{2}+\mathbf{i}a_2\frac{h}{4c\sin^2\frac{\vartheta_1}{2}}\sin c\omega_2h\right)=0,\\ \lim_{h\to0}\frac{1}{4\frac{\lambda}{h}\sin^2\frac{\vartheta_1}{2}}z_0 &= \lim_{h\to0}\left(-\frac{\rho}{c}\cot\tfrac{\vartheta_1}{2}\sin c\omega_2h\right)=0,\\ \lim_{h\to0}\frac{1}{4\frac{\lambda}{h}\sin^2\frac{\vartheta_1}{2}}p &= \lim_{h\to0}\left(\theta+\left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}}-\tfrac{1}{2}\theta\mathbf{i}a_1\cot\tfrac{\vartheta_1}{2}\right)h\right)(1-\theta z_2)\\ &= \theta(1+\widetilde{\widetilde{z}}_2),\end{aligned}$$

where $\widetilde{\widetilde{z_2}}$ denotes a positive real number. Hence, concerning R it follows that

$$\lim_{h\to 0}|R| = \lim_{h\to 0}\left|\frac{p^2+pz+\theta z_0 z+(\frac{1}{2}-\theta)z^2}{p^2}\right| = \left|\frac{(1+\widetilde{\widetilde{z_2}})^2\theta^2-(2+\widetilde{\widetilde{z_2}})\theta+\frac{1}{2}}{(1+\widetilde{\widetilde{z_2}})^2\theta^2}\right|.$$

For the latter expression we obtain the following positive result.

Proposition 6.5.2 *If $\theta > 1/4$ and $\theta \neq 1/2$, then*

$$\left|\frac{(1+\widetilde{\widetilde{z_2}})^2\theta^2-(2+\widetilde{\widetilde{z_2}})\theta+\frac{1}{2}}{(1+\widetilde{\widetilde{z_2}})^2\theta^2}\right| < 1$$

for all real numbers $\widetilde{\widetilde{z_2}} \geq 0$. If $\theta = 1/4$ or $\theta = 1/2$, then the inequality holds for numbers $\widetilde{\widetilde{z_2}} > 0$.

Proof Let $\widetilde{\widetilde{z_2}}$ be a positive real number. First, it is clear that the inequality holds whenever both

$$-(2+\widetilde{\widetilde{z_2}})\theta+\tfrac{1}{2} < 0, \tag{6.5.18a}$$

$$2(1+\widetilde{\widetilde{z_2}})^2\theta^2-(2+\widetilde{\widetilde{z_2}})\theta+\tfrac{1}{2} > 0. \tag{6.5.18b}$$

It is readily seen that (6.5.18a) is satisfied for $\theta > 1/4$. For strictly positive $\widetilde{\widetilde{z_2}}$ the inequality is satisfied whenever $\theta \geq 1/4$. Regarding inequality (6.5.18b) we consider the left-hand side as a second order polynomial in θ with discriminant

$$\Delta = (2+\widetilde{\widetilde{z_2}})^2 - 4(1+\widetilde{\widetilde{z_2}})^2 = -\widetilde{\widetilde{z_2}}(3\widetilde{\widetilde{z_2}}+4).$$

If $\widetilde{\widetilde{z_2}} > 0$, then $\Delta < 0$ and the polynomial is strictly positive for all real numbers θ. If $\widetilde{\widetilde{z_2}} = 0$, the polynomial reduces to $2\theta^2-2\theta+1/2$ which reaches its minimum (zero) in $\theta = 1/2$.

■

Let $\theta > 1/4$ and $\theta \neq 1/2$. Then, applying Proposition 6.5.2 in this region yields $\lim_{h\to 0}|R| < 1$ and thus

$$|R^N| = |R|^{1/(\lambda h)} = \mathcal{O}(h^w) \quad \forall w > 0.$$

Since N_0 is independent from h and $|1/(1-\frac{1}{2}z)| \leq 1$ one may conclude that

$$|\widehat{U}_N| = \mathcal{O}(h^w) \quad \forall w > 0.$$

Next, consider the case $\theta = 1/2$. Recall that the MCS scheme then reduces to the original CS scheme. If $|c\omega_2| = h^{-1/2}$, it follows that

$$\lim_{h\to 0}\frac{1}{4\frac{\lambda}{h}\sin^2\frac{\vartheta_1}{2}}p = \theta(1+\widetilde{\widetilde{z_2}}),$$

with $\widetilde{\widetilde{z}}_2 > 0$ such that proposition 6.5.2 can be applied and $|R^N| = \mathcal{O}(h^w)$ for all $w > 0$. Now, assume $|c\omega_2| \leq h^{-\mathfrak{q}_2}$ with $\mathfrak{q}_2 < 1/2$. An expansion similar to (6.5.7) yields that

$$\log\left[\frac{-1}{16\frac{\lambda^2}{h^2}\sin^4\frac{\vartheta_1}{2}}(p^2 + pz + \theta z_0 z + (\tfrac{1}{2} - \theta)z^2)\right]$$

equals

$$\begin{aligned}
&\log(-\theta^2 + 2\theta - \tfrac{1}{2}) \\
&- \left[2\theta\left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}} - \tfrac{1}{2}\theta\mathbf{i}a_1\cot\tfrac{\vartheta_1}{2} - \theta^2 z_2^{[1]}\right) + \theta\left(-\rho\cot\tfrac{\vartheta_1}{2}\omega_2 + \tfrac{1}{2}\mathbf{i}a_1\cot\tfrac{\vartheta_1}{2}\right)\right. \\
&- \left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}} - \tfrac{1}{2}\theta\mathbf{i}a_1\cot\tfrac{\vartheta_1}{2} - \theta^2 z_2^{[1]}\right) + \theta\rho\cot\tfrac{\vartheta_1}{2}\omega_2 \\
&\left.- 2(\tfrac{1}{2} - \theta)\left(-\rho\cot\tfrac{\vartheta_1}{2}\omega_2 + \tfrac{1}{2}\mathbf{i}a_1\cot\tfrac{\vartheta_1}{2}\right)\right]\frac{h}{-\theta^2 + 2\theta - \frac{1}{2}} \\
&+ \mathcal{O}\left(\left(\frac{1}{\sin^2\frac{\vartheta_1}{2}} + \frac{|\omega_2|}{|\sin\frac{\vartheta_1}{2}|} + \omega_2^2\right)^2 h^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\log\left[\frac{1}{16\frac{\lambda^2}{h^2}\sin^4\frac{\vartheta_1}{2}}p^2\right] &= \log\theta^2 + 2\theta\left(\frac{1}{4\lambda\sin^2\frac{\vartheta_1}{2}} - \tfrac{1}{2}\theta\mathbf{i}a_1\cot\tfrac{\vartheta_1}{2} - \theta^2 z_2^{[1]}\right)\frac{h}{\theta^2} \\
&\quad + \mathcal{O}\left(\left(\frac{1}{\sin^2\frac{\vartheta_1}{2}} + \omega_2^2\right)^2 h^2\right).
\end{aligned}$$

Making use of $\theta = 1/2$ it follows that

$$\begin{aligned}
\log\left((-R)^N\right) &= \frac{1}{\lambda h}\left[\left(\frac{-1}{\lambda\sin^2\frac{\vartheta_1}{2}} + z_2^{[1]}\right)h + \mathcal{O}\left(\left(\frac{1}{\sin^2\frac{\vartheta_1}{2}} + \frac{\omega_2}{|\sin\frac{\vartheta_1}{2}|} + \omega_2^2\right)^2 h^2\right)\right] \\
&= \left(\frac{-1}{\lambda^2\sin^2\frac{\vartheta_1}{2}} - \omega_2^2 + \mathbf{i}a_2\omega_2\right)\left(1 + \mathcal{O}\left(\left(\frac{1}{\sin^2\frac{\vartheta_1}{2}} + \frac{|\omega_2|}{|\sin\frac{\vartheta_1}{2}|} + \omega_2^2\right)h\right)\right).
\end{aligned}$$

In order to analyse the Rannacher time stepping we note

$$\frac{1 - \frac{1}{2}z}{2\frac{\lambda}{h}\sin^2\frac{\vartheta_1}{2}} = 1 + \mathcal{O}\left(\left(\frac{1}{\sin^2\frac{\vartheta_1}{2}} + \frac{|\omega_2|}{|\sin\frac{\vartheta_1}{2}|} + \omega_2^2\right)h\right),$$

which yields

$$\begin{aligned}
\log\left((-R)^{-N_0}\left(1 - \tfrac{1}{2}z\right)^{-2N_0}\right) &= 2N_0\log\left(\frac{h}{2\lambda\sin^2\frac{\vartheta_1}{2}}\right) \\
&\quad + \mathcal{O}\left(\left(\frac{1}{\sin^2\frac{\vartheta_1}{2}} + \frac{|\omega_2|}{|\sin\frac{\vartheta_1}{2}|} + \omega_2^2\right)h\right).
\end{aligned}$$

By exploring relationship (6.4.2) one may conclude that

$$\widehat{U}_N = \frac{(-1)^{N-N_0}h^{2N_0}}{(2\lambda\sin^2\frac{\vartheta_1}{2})^{2N_0}}\exp\Big(\frac{-1}{\lambda^2\sin^2\frac{\vartheta_1}{2}}-\omega_2^2+\mathbf{i}a_2\omega_2\Big)\Big(1+\mathcal{O}\Big(\Big(\frac{1}{\vartheta_1^2}+\frac{|\omega_2|}{|\vartheta_1|}+\omega_2^2\Big)h\Big)\Big). \tag{6.5.19}$$

Whenever $|c\omega_2| = h^{-1/2}$, the right-hand side of (6.5.19) is $\mathcal{O}(h^w)$ for all $w > 0$ such that we can use expression (6.5.19) for the whole region in the case of $\theta = 1/2$.

6.5.7. Region 5: $|\boldsymbol{\omega_1}| \leq \mathbf{h^{-q_1}}, |\mathbf{c\boldsymbol{\omega_2}}| \geq \mathbf{h^{-q_2}}$ with $\mathbf{q_1 \leq 1/2}$, $\mathbf{q_2 > 1/2}$

The analysis for this region is completely analogous to the one in Subsection 6.5.6. Hence, for $\theta > 1/4$ and $\theta \neq 1/2$ it follows that

$$|\widehat{U}_N| = \mathcal{O}(h^w) \quad \forall w > 0.$$

Whenever the CS scheme is considered, i.e. $\theta = 1/2$, one gets the expression

$$\widehat{U}_N = \frac{(-1)^{N-N_0}h^{2N_0}}{(\frac{2\lambda}{c}\sin^2\frac{\vartheta_2}{2})^{2N_0}}\exp\Big(\frac{-c^2}{\lambda^2\sin^2\frac{\vartheta_2}{2}}-\omega_1^2+\mathbf{i}a_1\omega_1\Big)\Big(1+\mathcal{O}\Big(\Big(\omega_1^2+\frac{|\omega_1|}{|\vartheta_2|}+\frac{1}{\vartheta_2^2}\Big)h\Big)\Big). \tag{6.5.20}$$

6.5.8. Connection with Stability of the MCS Scheme

In the above analysis natural bounds on the MCS parameter θ arise under which the asymptotic results are valid. These bounds can be interpreted as stability bounds. In particular, the conditions $\theta \geq \frac{1}{4},\ \theta > \frac{1+|\rho|}{6}$ are needed to ensure that the Fourier transform $\widehat{U}_N$ is negligible in the second region. This restriction is only slightly stronger than the lower bound on θ derived in [38], cf. also equation (3.3.1), guaranteeing unconditional stability of the MCS scheme in the von Neumann sense pertinent to two-dimensional diffusion equations with mixed derivative term. This is, indeed, not very surprising. In [38] it is stated that the stability analysis of the MCS scheme in this case reduces to bounding by one the modulus of the scalar expression

$$1 + \frac{\widetilde{z}}{\widetilde{p}} + \frac{(\theta\widetilde{z}_0 + (\frac{1}{2}-\theta)\widetilde{z})\widetilde{z}}{\widetilde{p}^2},$$

where $\widetilde{z} = \widetilde{z}_0+\widetilde{z}_1+\widetilde{z}_2,\ \widetilde{p} = (1-\theta\widetilde{z}_1)(1-\theta\widetilde{z}_2)$ and $\widetilde{z}_0, \widetilde{z}_1, \widetilde{z}_2$ denote real numbers satisfying the condition (6.5.12). This explains why Proposition 6.5.1 is just a slight modification of one of the statements in [38, Theorem 1].

6.6. Asymptotic Analysis in Physical Space for the MCS Scheme

In this section we will use the asymptotic results in Fourier space for the MCS scheme from Section 6.5 to perform an error analysis in physical space. First note that the Fourier transform $\widehat{u}$ is only sizeable in region 1 of the Fourier

domain. In the other regions it holds that $\omega_1 \geq h^{-1/3}$ or $c\omega_2 \geq h^{-1/3}$ and hence

$$\widehat{u}(\omega_1, \omega_2, 1) = \mathcal{O}\left(h^w\right) \quad \forall w > 0.$$

Based on equalities (6.2.2), (6.5.10) and (6.5.17) we define

$$\widehat{E}^{\text{low}} = h^2 \exp(-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + ia_1\omega_1 + ia_2\omega_2)(s^{[2]}(\omega_1, \omega_2) + N_0^{[2]}(\omega_1, \omega_2))$$

and

$$\widehat{E}^{\text{high}} = \frac{(c^2 h)^{2N_0}}{[2\lambda\iota(\vartheta_1, \vartheta_2)]^{2N_0}} \exp\left(-\frac{1}{4\lambda^2\theta^2} \frac{\iota(\vartheta_1, \vartheta_2)}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\right).$$

Recall that $\vartheta_1 = \omega_1 h_1$, $\vartheta_2 = \omega_2 h_2$ and $h_2 = ch_1 = ch$. As a consequence, $\widehat{E}^{\text{low}}$ is only sizeable in region 1 and $\widehat{E}^{\text{high}}$ is only sizeable in region 3. In the other regions of the Fourier domain $\widehat{U}_N$ is negligible whenever $\theta > \max\{\frac{1}{4}, \frac{1+|\rho|}{6}\}$ and $\theta \neq 1/2$. Hence, for these values of θ, the results can be combined to

$$\widehat{U}_N(\omega_1 h_1, \omega_2 h_2) - \widehat{u}(\omega_1, \omega_2, 1) \approx \widehat{E}^{\text{low}} + \widehat{E}^{\text{high}}, \quad \text{for } |\omega_1|, |c\omega_2| \leq \pi/h. \tag{6.6.1}$$

When $\theta = 1/2$, i.e. when the MCS scheme reduces to the CS scheme, $\widehat{U}_N$ is also sizeable in region 4 and region 5 of the Fourier domain. This case will be treated separately.

6.6.1. MCS Scheme with $\theta \neq 1/2$

Consider the case where the MCS scheme is different from the CS scheme, $\theta \neq 1/2$, and suppose that the restriction $\theta > \max\{\frac{1}{4}, \frac{1+|\rho|}{6}\}$ is satisfied. Approximation (6.6.1) is then valid and based on (6.5.2) we have for the total error:

$$U_{N,j,k} - u(x_j, y_k, 1) \approx E_{j,k}^{\text{low}} + E_{j,k}^{\text{high}},$$

where the low-wavenumber error $E_{j,k}^{\text{low}}$ is given by

$$\frac{h^2}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widehat{u}(\omega_1, \omega_2, 1)(s^{[2]}(\omega_1, \omega_2) + N_0^{[2]}(\omega_1, \omega_2)) \exp(\mathbf{i}(\omega_1 x_j + \omega_2 y_k)) d\omega_1 d\omega_2 \tag{6.6.2}$$

and the high-wavenumber error $E_{j,k}^{\text{high}}$ is given by

$$\frac{h^{2N_0} c^{4N_0}}{4\pi^2}\int_{-\pi/h_2}^{\pi/h_2}\int_{-\pi/h_1}^{\pi/h_1} \frac{\exp(\mathbf{i}\omega_1 x_j)\exp(\mathbf{i}\omega_2 y_k)}{[2\lambda\iota(\vartheta_1,\vartheta_2)]^{2N_0}} \exp\left(-\frac{1}{4\lambda^2\theta^2}\frac{\iota(\vartheta_1,\vartheta_2)}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\right) d\omega_1 d\omega_2,$$

which can be rewritten as

$$\frac{h^{2N_0-2} c^{4N_0-1}}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{\exp(\mathbf{i} j\vartheta_1)\exp(\mathbf{i} k\vartheta_2)}{[2\lambda\iota(\vartheta_1,\vartheta_2)]^{2N_0}} \exp\left(-\frac{1}{4\lambda^2\theta^2}\frac{\iota(\vartheta_1,\vartheta_2)}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\right) d\vartheta_1 d\vartheta_2.$$

First, consider the low-wavenumber error. The inverse mixed discrete/continuous Fourier transform of $\widehat{E}^{\text{low}}$ is given by

$$\frac{h^2}{4\pi^2 h_1 h_2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \widehat{u}(\tfrac{\vartheta_1}{h_1}, \tfrac{\vartheta_2}{h_2}, 1)(s^{[2]}(\tfrac{\vartheta_1}{h_1}, \tfrac{\vartheta_2}{h_2}) + N_0^{[2]}(\tfrac{\vartheta_1}{h_1}, \tfrac{\vartheta_2}{h_2})) \exp(\mathbf{i}(j\vartheta_1 + k\vartheta_2) d\vartheta_1 d\vartheta_2$$

which can be approximated by (6.6.2) as h_1, h_2 tend to zero. Let

$$\phi_\rho(x,y) = \tfrac{1}{\sqrt{4\pi^2(1-\rho^2)}} \exp\left(-\tfrac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right),$$

the density function of a two-dimensional standard normally distributed random variable with correlation ρ. For all positive integers n_1, n_2 the Fourier transform of $\frac{\partial^{n_1+n_2}}{\partial x^{n_1}\partial y^{n_2}}\phi_\rho\left(\frac{x+a_1}{\sqrt{2}}, \frac{y+a_2}{\sqrt{2}}\right)$ is

$$2(\mathbf{i}\sqrt{2}\omega_1)^{n_1}(\mathbf{i}\sqrt{2}\omega_2)^{n_2}\exp(-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2), \quad (6.6.3)$$

such that the inverse Fourier transform of

$$h^2\exp(-\omega_1^2 - 2\rho\omega_1\omega_2 - \omega_2^2 + \mathbf{i}a_1\omega_1 + \mathbf{i}a_2\omega_2)\omega_1^{n_1}\omega_2^{n_2}$$

is given by

$$\tfrac{h^2}{2}\tfrac{1}{(\mathbf{i}\sqrt{2})^{n_1+n_2}}\tfrac{\partial^{n_1+n_2}}{\partial x^{n_1}\partial y^{n^2}}\phi_\rho\left(\tfrac{x+a_1}{\sqrt{2}}, \tfrac{y+a_2}{\sqrt{2}}\right).$$

Recalling the formulas for $s^{[2]}$ and $R^{[2]}$ from Subsection 6.5.3, this leads to the following expression for the low-wavenumber error:

$$E_{j,k}^{\text{low}} = h^2 C_{x_j,y_k}^{\text{low}}, \quad (6.6.4)$$

with

$$\begin{aligned}
C_{x_j,y_k}^{\text{low}} = \tfrac{1}{2}\Bigg[&\tfrac{1}{48}\tfrac{\partial^4}{\partial x^4} + \tfrac{\rho}{12}\left(\tfrac{\partial^4}{\partial x^3\partial y} + \tfrac{c^2\partial^4}{\partial x\partial y^3}\right) + \tfrac{c^2}{48}\tfrac{\partial^4}{\partial y^4} + \tfrac{a_1}{12\sqrt{2}}\tfrac{\partial^3}{\partial x^3} + \tfrac{a_2c^2}{12\sqrt{2}}\tfrac{\partial^3}{\partial y^3}\\
&- \lambda^2\theta^2\left(\tfrac{1}{2}\tfrac{\partial^2}{\partial x^2} + \tfrac{a_1}{\sqrt{2}}\tfrac{\partial}{\partial x}\right)\left(\tfrac{1}{2}\tfrac{\partial^2}{\partial y^2} + \tfrac{a_2}{\sqrt{2}}\tfrac{\partial}{\partial y}\right)\\
&\times\left(\tfrac{1}{2}\tfrac{\partial^2}{\partial x^2} + \rho\tfrac{\partial^2}{\partial x\partial y} + \tfrac{1}{2}\tfrac{\partial^2}{\partial y^2} + \tfrac{a_1}{\sqrt{2}}\tfrac{\partial}{\partial x} + \tfrac{a_2}{\sqrt{2}}\tfrac{\partial}{\partial y}\right)\\
&+ \tfrac{\lambda^2}{12}\left(\tfrac{1}{2}\tfrac{\partial^2}{\partial x^2} + \rho\tfrac{\partial^2}{\partial x\partial y} + \tfrac{1}{2}\tfrac{\partial^2}{\partial y^2} + \tfrac{a_1}{\sqrt{2}}\tfrac{\partial}{\partial x} + \tfrac{a_2}{\sqrt{2}}\tfrac{\partial}{\partial y}\right)^3\\
&- \lambda^2\left(\tfrac{1}{2}\tfrac{\partial^2}{\partial x^2} + \rho\tfrac{\partial^2}{\partial x\partial y} + \tfrac{1}{2}\tfrac{\partial^2}{\partial y^2} + \tfrac{a_1}{\sqrt{2}}\tfrac{\partial}{\partial x} + \tfrac{a_2}{\sqrt{2}}\tfrac{\partial}{\partial y}\right)\\
&\times\left(\tfrac{\rho}{2}\tfrac{\partial^2}{\partial x\partial y} + \left(\tfrac{1}{2}-\theta\right)\left(\tfrac{1}{2}\tfrac{\partial^2}{\partial x^2} + \tfrac{1}{2}\tfrac{\partial^2}{\partial y^2} + \tfrac{a_1}{\sqrt{2}}\tfrac{\partial}{\partial x} + \tfrac{a_2}{\sqrt{2}}\tfrac{\partial}{\partial y}\right)\right)^2\\
&+ \tfrac{N_0\lambda^2}{4}\left(\tfrac{1}{2}\tfrac{\partial^2}{\partial x^2} + \rho\tfrac{\partial^2}{\partial x\partial y} + \tfrac{1}{2}\tfrac{\partial^2}{\partial y^2} + \tfrac{a_1}{\sqrt{2}}\tfrac{\partial}{\partial x} + \tfrac{a_2}{\sqrt{2}}\tfrac{\partial}{\partial y}\right)^2\Bigg]\phi_\rho\left(\tfrac{x_j+a_1}{\sqrt{2}}, \tfrac{y_k+a_2}{\sqrt{2}}\right).
\end{aligned}$$

Next, consider the high-wavenumber error and note that

$$\exp(\mathbf{i}j\vartheta_1)\exp(\mathbf{i}k\vartheta_2) = \cos(j\vartheta_1 + k\vartheta_2) + \mathbf{i}\sin(j\vartheta_1 + k\vartheta_2).$$

Symmetry yields

$$E_{j,k}^{\text{high}} = h^{2N_0-2}C_{j,k}^{\text{high}}, \quad (6.6.5)$$

where

$$\begin{aligned}
C_{j,k}^{\text{high}} = &\tfrac{c^{4N_0-1}}{2\pi^2}\int_0^\pi\int_0^\pi \tfrac{\cos(j\vartheta_1+k\vartheta_2)}{[2\lambda\iota(\vartheta_1,\vartheta_2)]^{2N_0}}\exp\left(-\tfrac{1}{4\lambda^2\theta^2}\tfrac{\iota(\vartheta_1,\vartheta_2)}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\right)d\vartheta_1 d\vartheta_2\\
&+ \tfrac{c^{4N_0-1}}{2\pi^2}\int_{-\pi}^0\int_0^\pi \tfrac{\cos(j\vartheta_1+k\vartheta_2)}{[2\lambda\iota(\vartheta_1,\vartheta_2)]^{2N_0}}\exp\left(-\tfrac{1}{4\lambda^2\theta^2}\tfrac{\iota(\vartheta_1,\vartheta_2)}{\sin^2\frac{\vartheta_1}{2}\sin^2\frac{\vartheta_2}{2}}\right)d\vartheta_1 d\vartheta_2.
\end{aligned}$$

Combining the expression (6.6.4) for the low-wavenumber error and expression (6.6.5) for the high-wavenumber error yields the main result of the chapter:

Theorem 6.6.1 *Consider the model PDE* (6.2.1). *Assume spatial discretization is performed by standard second order central finite differences on a uniform Cartesian grid with mesh width $h_1 = h$ in the x-direction and mesh width h_2 in the y-direction. Assume the obtained semidiscrete system is discretized in time by using the MCS scheme, with parameter $\theta \geq 1/3$ and $\theta \neq 1/2$, on a uniform temporal grid with temporal step size Δt. Let $N_0 \geq 0$ denote the number of initial MCS time steps that are replaced by $2N_0$ half-time steps of the implicit Euler scheme. If $c = h_2/h_1$ and $\lambda = \Delta t/h$ are kept constant, then as h tends to zero the total error is approximated by*

$$U_{N,j,k} - u(x_j, y_k, 1) \approx h^2 C^{\text{low}}_{x_j,y_k} + h^{2N_0-2} C^{\text{high}}_{j,k}.$$

The values $C^{\text{low}}_{x_j,y_k}$ are only dependent on the position $(x_j, y_k) = (jh_1, kh_2)$, the parameter values of the problem and the ratios c and λ. The constants $C^{\text{high}}_{j,k}$ only depend on the index (j,k), the correlation parameter ρ and the ratios c, λ. For the numerical experiments, cf. infra, the values $C^{\text{low}}_{x_j,y_k}$ are calculated by determining all the partial derivatives. The integrals in $C^{\text{high}}_{j,k}$ are approximated by numerical integration. It is readily seen that

$$\max_{j,k} |C^{\text{high}}_{j,k}| = |C^{\text{high}}_{0,0}|,$$

so $E^{\text{high}}_{j,k}$ has a maximum magnitude where $(x_j, y_k) = (0,0)$. *This is exactly at the position of the discontinuity of the initial function.* At the end of Subsection 6.5.5 it was conjectured that for larger values of the MCS parameter θ one can expect a larger high-wavenumber error. This conjecture is confirmed by the above analysis given that $\iota(\vartheta_1, \vartheta_2)$ is always positive. In order to avoid spurious erratic behaviour in the numerical solution, *it is therefore recommended to use smaller values of the parameter θ.* However, one has to take into account the lower bound on θ described in Subsection 6.5.8.

Theorem 6.6.1 shows that the total error is $\mathcal{O}(h^{\min\{2,2N_0-2\}})$ so that $N_0 = 2$ is a lower bound on N_0 for the Rannacher time stepping in order to ensure convergence of the numerical solution to the exact solution. This is confirmed by the plots in Figure 6.5 which display total errors (in the maximum norm) in actual numerical experiments for model problem (6.2.1) as a function of $1/h$, with parameter values $\rho = -0.7, a_1 = 2, a_2 = 3$, MCS parameter $\theta = 1/3$ and with $c = 1$, $0.2 \leq \lambda \leq 0.8$. Since it is not possible to handle infinite domains in numerical experiments, the computational domain is restricted to spatial grid points $(x_j, y_k) \in [-10, 10] \times [-10, 10]$. At the boundaries, homogeneous Dirichlet boundary conditions are applied. In the left plots the case $N_0 = 0$ is considered, whereas the right plots show the corresponding results for $N_0 = 2$. In the upper plots the maximum error between our numerical solution and the exact solution is shown as a function of $1/h$ for different values of λ. In the lower plots we show the same maximum error for one value of λ, together with our theoretical estimates for the corresponding low-wavenumber error and high-wavenumber error. In these lower plots it is clearly seen that our theoretical estimates for the total error are sharp.

For the case where no Rannacher time stepping is applied, the left plots in Figure 6.5 reveal second order convergence behaviour until h reaches a critical

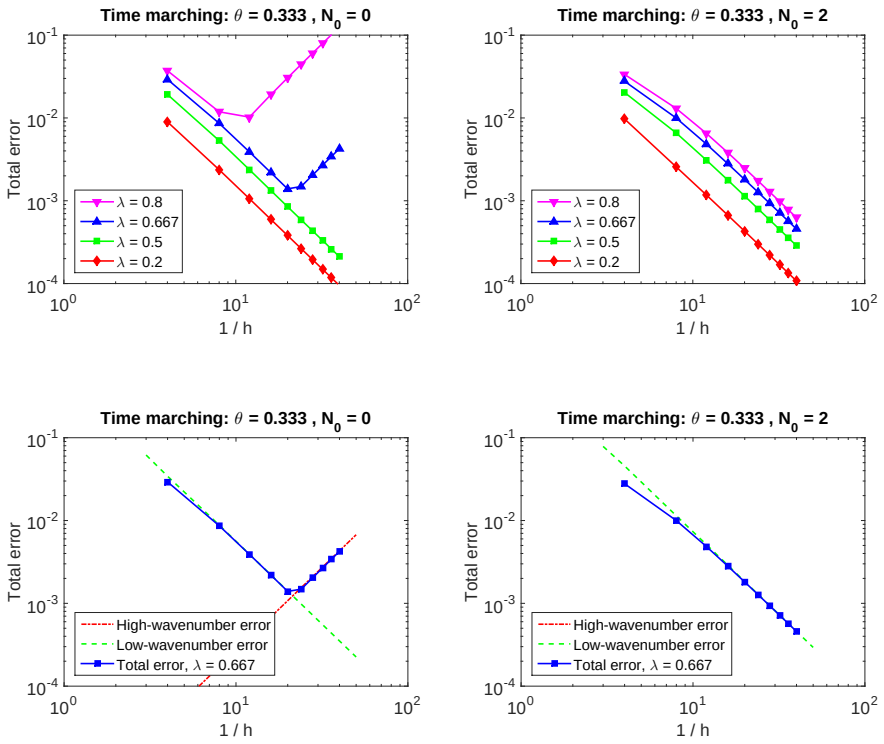


Figure 6.5: Convergence of the numerical solution for $N_0 = 0$ (left) and $N_0 = 2$ (right). The parameter values are: $\rho = -0.7, a_1 = 2, a_2 = 3, \theta = 1/3$.

6

value where the high-wavenumber error starts exceeding the low-wavenumber error. It can be observed that this value of h, and thus the high-wavenumber error, is highly dependent on the ratio $\lambda = \Delta t/h$. For smaller values of λ, $E_{j,k}^{\text{high}}$ is only sizeable whenever h is very small, whereas for larger values of λ, $E_{j,k}^{\text{high}}$ already dominates the total error for larger values of h. Moreover, the error constant for the low-wavenumber error is also dependent on λ. However, this is much less pronounced than for the high-wavenumber error.

The right plots in Figure 6.5 show the corresponding results in the case where the first two MCS time steps are replaced by four backward Euler half-time steps, thus $N_0 = 2$. One observes that the numerical approximations now exhibit second order convergence for all values of the ratio λ. In the bottom right plot the high-wavenumber error is not visible since it is strongly dominated by the low-wavenumber error. The same observation is made for other values of λ. Hence, whenever Rannacher time stepping is applied with $N_0 = 2$, the total error can be approximated by $E_{j,k}^{\text{low}}$, which is of second order in h. We find that the error constant for the low-wavenumber error is mildly dependent on the ratio $\lambda = \Delta t/h$. This can be explained through the fact that for a fixed value of h but smaller value of λ the same semidiscrete system is solved with a smaller temporal step size Δt. Finally, we notice that the

latter error constant is slightly larger than for the case where $N_0 = 0$. Thus, by applying Rannacher time stepping with $N_0 = 2$, second order convergence can be recovered at the small cost of a marginally larger error constant for the low-wavenumber error.

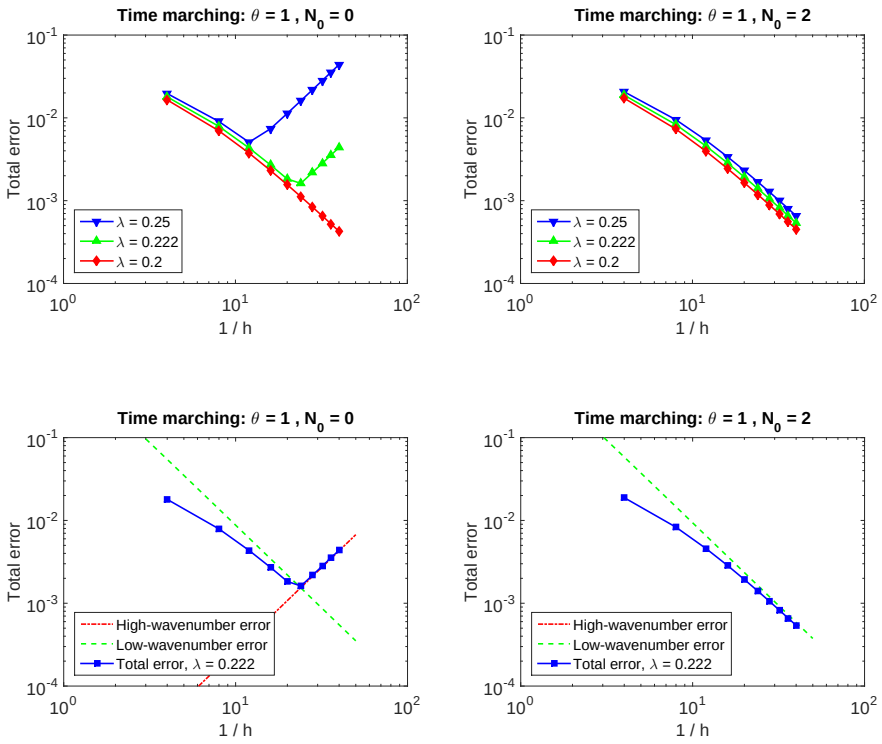


Figure 6.6: Convergence of the numerical solution for $N_0 = 0$ (left) and $N_0 = 2$ (right). The parameter values are: $\rho = -0.7, a_1 = 2, a_2 = 3, \theta = 1$.

As stated above, the high-wavenumber error is very sensitive to the MCS parameter θ. To illustrate this, Figure 6.6 shows the same plots as in Figure 6.5 but with the MCS parameter replaced by $\theta = 1$. It can be seen that all the conclusions from Figure 6.5 remain valid. In order to get decent plots, however, it is necessary to consider smaller values for λ. This confirms that, for fixed λ, $E_{j,k}^{\text{high}}$ is strongly increasing as a function of θ. Figure 6.7 reveals that, for the considered parameter values, the low-wavenumber error is also sensitive to the MCS parameter θ. Note, however, that this is less pronounced than for the high-wavenumber error. We conjecture that for fixed λ and fixed h, the low-wavenumber error is also increasing as a function of θ. Therefore, regardless of the number of Rannacher time steps N_0, it seems more favourable to consider smaller values of θ. In particular, the lowest value of θ which satisfies the restrictions from Subsection 6.5.8 for all values $|\rho| < 1$ is given by $\theta = 1/3$.

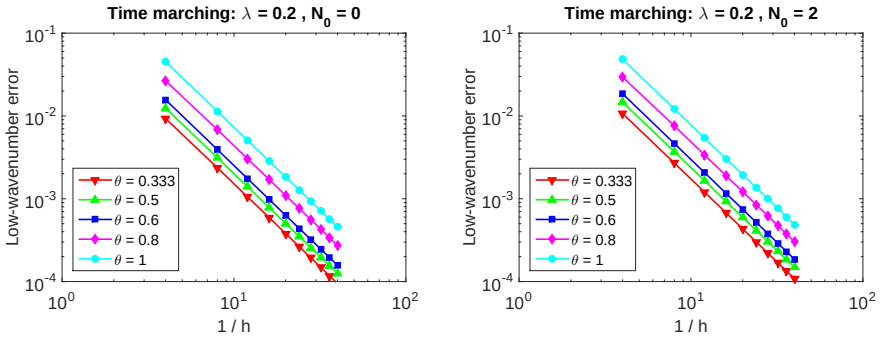


Figure 6.7: Convergence of the low-wavenumber error for different values of θ, and for $N_0 = 0$ (left) and $N_0 = 2$ (right). The parameter values are: $\rho = -0.7$, $a_1 = 2$, $a_2 = 3$, $\lambda = 1/5$.

6.6.2. MCS Scheme with $\theta = 1/2$

For $\theta = 1/2$, the MCS scheme reduces to the CS scheme and $\widehat{U}_N$ is not negligible in region 4 and region 5. Based on equalities (6.5.19) and (6.5.20) we define

$$\widehat{E}^{\mathrm{CS},4} = (-1)^{N-N_0} \frac{h^{2N_0}}{(2\lambda \sin^2 \frac{\vartheta_1}{2})^{2N_0}} \exp\left(\frac{-1}{\lambda^2 \sin^2 \frac{\vartheta_1}{2}} - \omega_2^2 + \mathbf{i} a_2 \omega_2 \right)$$

and

$$\widehat{E}^{\mathrm{CS},5} = (-1)^{N-N_0} \frac{(ch)^{2N_0}}{(2\lambda \sin^2 \frac{\vartheta_2}{2})^{2N_0}} \exp\left(\frac{-c^2}{\lambda^2 \sin^2 \frac{\vartheta_2}{2}} - \omega_1^2 + \mathbf{i} a_1 \omega_1 \right).$$

Since $\vartheta_1 = \omega_1 h_1$, $\vartheta_2 = \omega_2 h_2$ and $h_2 = c h_1 = ch$, $\widehat{E}^{\mathrm{CS},4}$ only has to be considered in region 4 and $\widehat{E}^{\mathrm{CS},5}$ is only not negligible in region 5. Hence, the Fourier error (6.5.1) can be approximated by

$$\widehat{U}_N(\omega_1 h_1, \omega_2 h_2) - \widehat{u}(\omega_1, \omega_2, 1) \approx \widehat{E}^{\mathrm{low}} + \widehat{E}^{\mathrm{high}} + \widehat{E}^{\mathrm{CS},4} + \widehat{E}^{\mathrm{CS},5},$$

for $|\omega_1|, |c\omega_2| \leq \pi/h$. The inverse mixed discrete/continuous Fourier transform of the term $(-1)^{N-N_0}(\widehat{E}^{\mathrm{CS},4} + \widehat{E}^{\mathrm{CS},5})$ is given by

$$\frac{h^{2N_0}}{4\pi^2} \int_{-\pi/h_2}^{\pi/h_2} \int_{-\pi/h_1}^{\pi/h_1} \frac{\exp(\mathbf{i}\omega_1 x_j)\exp(\mathbf{i}\omega_2 y_k)}{(2\lambda \sin^2 \frac{\vartheta_1}{2})^{2N_0}} \exp\left(\frac{-1}{\lambda^2 \sin^2 \frac{\vartheta_1}{2}} - \omega_2^2 + \mathbf{i} a_2 \omega_2 \right) d\omega_1 d\omega_2$$
$$+ \frac{h^{2N_0}}{4\pi^2} \int_{-\pi/h_2}^{\pi/h_2} \int_{-\pi/h_1}^{\pi/h_1} \frac{\exp(\mathbf{i}\omega_1 x_j)\exp(\mathbf{i}\omega_2 y_k)}{(2\frac{\lambda}{c} \sin^2 \frac{\vartheta_2}{2})^{2N_0}} \exp\left(\frac{-c^2}{\lambda^2 \sin^2 \frac{\vartheta_2}{2}} - \omega_1^2 + \mathbf{i} a_1 \omega_1 \right) d\omega_1 d\omega_2.$$

As h_1, h_2 tend to zero this can be approximated by

$$\frac{h^{2N_0-1}}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\pi}^{\pi} \frac{\exp(\mathbf{i} j \vartheta_1)\exp(\mathbf{i}\omega_2 y_k)}{(2\lambda \sin^2 \frac{\vartheta_1}{2})^{2N_0}} \exp\left(\frac{-1}{\lambda^2 \sin^2 \frac{\vartheta_1}{2}} - \omega_2^2 + \mathbf{i} a_2 \omega_2 \right) d\vartheta_1 d\omega_2$$
$$+ \frac{h^{2N_0-1}}{4c\pi^2} \int_{-\pi}^{\pi} \int_{-\infty}^{\infty} \frac{\exp(\mathbf{i}\omega_1 x_j)\exp(\mathbf{i} k \vartheta_2)}{(2\frac{\lambda}{c} \sin^2 \frac{\vartheta_2}{2})^{2N_0}} \exp\left(\frac{-c^2}{\lambda^2 \sin^2 \frac{\vartheta_2}{2}} - \omega_1^2 + \mathbf{i} a_1 \omega_1 \right) d\omega_1 d\vartheta_2,$$

and we denote this expression by $(-1)^{N-N_0}E_{j,k}^{\mathrm{CS}}$. Making use of a symmetry argument and a one-dimensional inverse Fourier transformation, $E_{j,k}^{\mathrm{CS}}$ can be rewritten as

$$E_{j,k}^{\mathrm{CS}} = h^{2N_0-1}(-1)^{N-N_0}(C_{j,y_k}^{\mathrm{CS}} + C_{x_j,k}^{\mathrm{CS}}), \tag{6.6.6}$$

with

$$C_{j,y_k}^{\mathrm{CS}} = \tfrac{1}{2\sqrt{2}\pi}\phi\left(\tfrac{y_k+a_2}{\sqrt{2}}\right)\int_{-\pi}^{\pi}\frac{\cos(j\vartheta_1)}{(2\lambda\sin^2\frac{\vartheta_1}{2})^{2N_0}}\exp\left(\frac{-1}{\lambda^2\sin^2\frac{\vartheta_1}{2}}\right)d\vartheta_1,$$

$$C_{x_j,k}^{\mathrm{CS}} = \tfrac{1}{2\sqrt{2}c\pi}\phi\left(\tfrac{x_j+a_1}{\sqrt{2}}\right)\int_{-\pi}^{\pi}\frac{\cos(k\vartheta_2)}{(2\frac{\lambda}{c}\sin^2\frac{\vartheta_2}{2})^{2N_0}}\exp\left(\frac{-c^2}{\lambda^2\sin^2\frac{\vartheta_2}{2}}\right)d\vartheta_2,$$

where ϕ denotes the density function of a standard normally distributed random variable. It is readily seen that C_{j,y_k}^{CS}, respectively $C_{x_j,k}^{\mathrm{CS}}$, reaches its highest magnitude near the points (j,k) where $(x_j,y_k) \approx (0,-a_2)$, respectively $(x_j,y_k) \approx (-a_1,0)$. For the numerical experiments, the integrals in C_{j,y_k}^{CS} and $C_{x_j,k}^{\mathrm{CS}}$ are approximated by numerical integration. Combining the expressions (6.6.4), (6.6.5) and (6.6.6) leads to the following theorem:

Theorem 6.6.2 *Consider the model PDE* (6.2.1). *Assume spatial discretization is performed by standard second order central finite differences on a uniform Cartesian grid with mesh width $h_1 = h$ in the x-direction and mesh width h_2 in the y-direction. Assume the obtained semidiscrete system is discretized in time by using the CS scheme on a uniform temporal grid with temporal step size Δt. Let $N_0 \geq 0$ denote the number of initial CS time steps that are replaced by $2N_0$ half-time steps of the implicit Euler scheme. If $c = h_2/h_1$ and $\lambda = \Delta t/h$ are kept constant, then as h tends to zero the total error is approximated by*

$$U_{N,j,k} - u(x_j,y_k,1) \approx h^2 C_{x_j,y_k}^{\mathrm{low}} + h^{2N_0-2}C_{j,k}^{\mathrm{high}} + h^{2N_0-1}(-1)^{N-N_0}(C_{j,y_k}^{\mathrm{CS}} + C_{x_j,k}^{\mathrm{CS}}).$$

From Theorem 6.6.2 it can be concluded that when CS time stepping is considered, the total error is also $\mathcal{O}\left(h^{\min\{2,2N_0-2\}}\right)$. This matches the observations from the plots in Figure 6.8 which show convergence results for the same problem as in Subsection 6.6.1 but with MCS parameter $\theta = 1/2$. The lower plots indicate again that our theoretical estimates for the total error are sharp. Without Rannacher time stepping, i.e. $N_0 = 0$, the results in Figure 6.8 show second order convergence in h until $E_{j,k}^{\mathrm{CS}}$ starts exceeding the low-wavenumber error. Then the total error increases in a first order way until the high-wavenumber error starts dominating. From there the total error is $\mathcal{O}\left(h^{-2}\right)$. In case the CS scheme is replaced in the first two time steps by four half-time steps of the implicit Euler scheme, i.e. $N_0 = 2$, Figure 6.8 reveals unconditional second order convergence in h. Note that both $E_{j,k}^{\mathrm{CS}}$ and $E_{j,k}^{\mathrm{high}}$ are not visible in the lower-right plot because they are strongly dominated by the low-wavenumber error. The same observation as in Subsection 6.6.1 can be made concerning the dependency of the low- and high-wavenumber errors on the parameter λ.

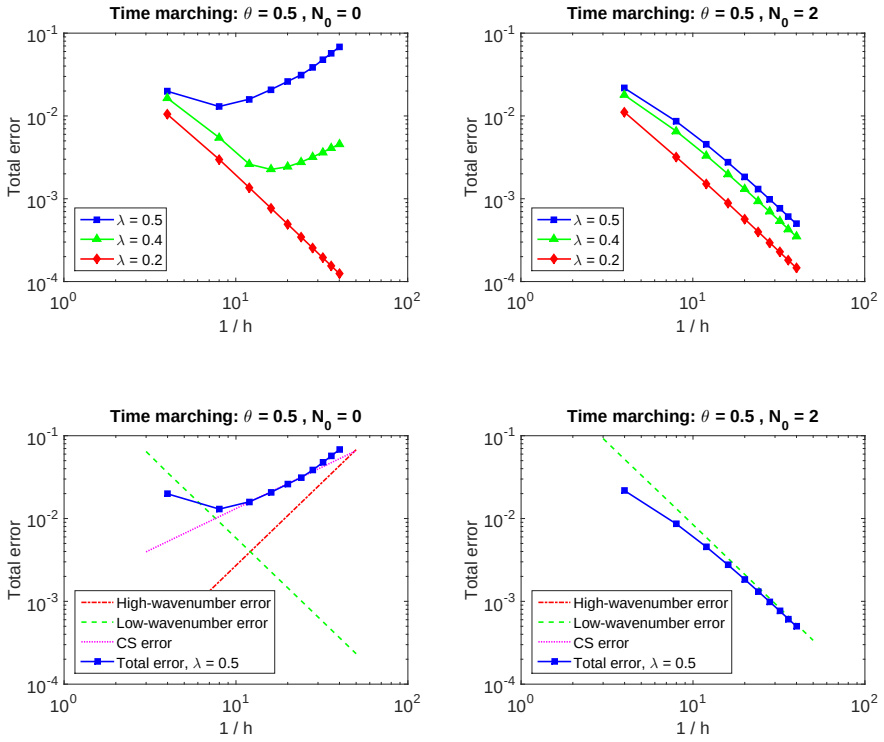


Figure 6.8: Convergence of the numerical solution for $N_0 = 0$ (left) and $N_0 = 2$ (right). The parameter values are: $\rho = -0.7, a_1 = 2, a_2 = 3, \theta = 1/2$.

6

6.6.3. Alternative Initial Data

Until now, the focus of our analysis has been on Dirac delta initial data. If the Dirac delta is approximated on the spatial grid by (6.3.2), then both the Fourier transforms $\widehat{u}(\omega_1, \omega_2, 0)$ and $\widehat{U}_0(\omega_1 h_1, \omega_2 h_2)$ are identically equal to one, which facilitates the analysis. However, also alternative approximations of the Dirac delta or other (non-smooth) initial functions can be considered. Assume again that $c = h_2/h_1$ and $\lambda = \Delta t/h_1$ are kept constant, and write $h_1 = h$.

A possible alternative discretization of the Dirac delta is given by

$$U_{0,j,k} = \begin{cases} \frac{1}{4h_1h_2} & \text{if } j = k = 0, \\ \frac{1}{8h_1h_2} & \text{if } j = 0,\ |k| = 1 \text{ or } |j| = 1,\ k = 0, \\ \frac{1}{16h_1h_2} & \text{if } |j| = |k| = 1, \\ 0 & \text{else.} \end{cases}$$

An asymptotic expansion of its discrete Fourier transform yields

$$\widehat{U}_0 = 1 + \mathcal{O}\left(\omega_1^2 h_1^2 + \omega_2^2 h_2^2\right) = 1 + \mathcal{O}\left(\vartheta_1^2 + \vartheta_2^2\right). \tag{6.6.7}$$

Hence, the orders of the low-wavenumber error and the high-wavenumber error will be unaffected. In general, if one considers an alternative discretization

of the Dirac delta function such that (6.6.7) is valid, then Theorem 6.6.1 and Theorem 6.6.2 will remain valid, except with other error constants.

As a natural example of another non-smooth initial function, consider the payoff of a two-asset cash-or-nothing option, i.e. $u(x,y,0) = \mathbb{1}_{\{x\geq 0\}}\mathbb{1}_{\{y\geq 0\}}$, such that

$$\widehat{u}(\omega_1,\omega_2,0) = (\tfrac{1}{\mathbf{i}\omega_1} + \pi\delta(\omega_1))(\tfrac{1}{\mathbf{i}\omega_2} + \pi\delta(\omega_2)). \tag{6.6.8}$$

A naive approximation of this initial function on the spatial grid is

$$U_{0,j,k} = \begin{cases} 1 & \text{if } j \geq 0,\ k \geq 0, \\ 0 & \text{else.} \end{cases}$$

It can be shown that discrete Fourier transformation of this discretization gives

$$\begin{aligned}\widehat{U}_0(\omega_1 h_1, \omega_2 h_2) &= \Big(\tfrac{h_1}{1-\exp(-\mathbf{i}\omega_1 h_1)} + h_1\pi\delta(\omega_1 h_1)\Big)\Big(\tfrac{h_2}{1-\exp(-\mathbf{i}\omega_2 h_2)} + h_2\pi\delta(\omega_2 h_2)\Big) \\ &= \Big(\tfrac{h_1}{1-\exp(-\mathbf{i}\vartheta_1)} + h_1\pi\delta(\vartheta_1)\Big)\Big(\tfrac{h_2}{1-\exp(-\mathbf{i}\vartheta_2)} + h_2\pi\delta(\vartheta_2)\Big).\end{aligned}$$

If the Dirac delta is regarded as a distribution, then the expression $h_1\pi\delta(\omega_1 h_1)$ can be replaced by $\pi\delta(\omega_1)$ and the similarity with (6.6.8) readily becomes clear. Note, however, that for example

$$(1-\exp(-\mathbf{i}\omega_1 h_1))/h_1 - \mathbf{i}\omega_1 = \mathcal{O}\left(\omega_1^2 h_1\right).$$

This can be seen to imply the unfavourable result that the low-wavenumber error is only of first order. Concerning the high-wavenumber error it can be shown that this will be $\mathcal{O}\left(h^{2N_0}\right)$. Here, if the CS scheme is used, we also mean the unidirectional high-wavenumber errors, i.e. the errors corresponding to $\widehat{E}^{\mathrm{CS},4}$ and $\widehat{E}^{\mathrm{CS},5}$. In view of the foregoing, we propose a different discretization of the initial function under consideration:

$$U_{0,j,k} = \begin{cases} 1/4 & \text{if } j = 0,\ k = 0, \\ 1/2 & \text{if } j > 0,\ k = 0 \text{ or } \ j = 0,\ k > 0, \\ 1 & \text{if } j > 0,\ k > 0, \\ 0 & \text{else,} \end{cases}$$

such that

$$\begin{aligned}\widehat{U}_0(\omega_1 h_1, \omega_2 h_2) &= \Big(\tfrac{2h_1}{\exp(\mathbf{i}\omega_1 h_1)-\exp(-\mathbf{i}\omega_1 h_1)} + h_1\pi\delta(\omega_1 h_1)\Big) \\ &\quad\times\Big(\tfrac{2h_2}{\exp(\mathbf{i}\omega_2 h_2)-\exp(-\mathbf{i}\omega_2 h_2)} + h_2\pi\delta(\omega_2 h_2)\Big) \\ &= \Big(\tfrac{2h_1}{\exp(\mathbf{i}\vartheta_1)-\exp(-\mathbf{i}\vartheta_1)} + h_1\pi\delta(\vartheta_1)\Big) \\ &\quad\times\Big(\tfrac{2h_2}{\exp(\mathbf{i}\vartheta_2)-\exp(-\mathbf{i}\vartheta_2)} + h_2\pi\delta(\vartheta_2)\Big).\end{aligned}$$

With this discretization it can be proven that the low-wavenumber error is again of second order and the high-wavenumber error will be $\mathcal{O}\left(h^{2N_0}\right)$ as before.

As a second example of another non-smooth initial function, consider that corresponding to the delta Greek of a two-asset cash-or-nothing option with respect to the first asset, i.e. $u(x,y,0) = \delta(x)\mathbb{1}_{\{y\geq 0\}}$. The results for the delta Greek with respect to the second asset are completely similar. Continuous Fourier transformation of $u(x,y,0)$ gives

$$\widehat{u}(\omega_1,\omega_2,0) = \tfrac{1}{\mathbf{i}\omega_2} + \pi\delta(\omega_2).$$

Based on the foregoing insights, we discretize the initial function by

$$U_{0,j,k} = \begin{cases} 1/(2h_1) & \text{if } j=0,\ k=0,\\ 1/h_1 & \text{if } j=0,\ k>0,\\ 0 & \text{else.}\end{cases}$$

Discrete Fourier transformation yields

$$\begin{aligned}\widehat{U}_0(\omega_1 h_1,\omega_2 h_2) &= \tfrac{2h_2}{\exp(\mathbf{i}\omega_2 h_2)-\exp(-\mathbf{i}\omega_2 h_2)} + h_2\pi\delta(\omega_2 h_2)\\ &= \tfrac{2h_2}{\exp(\mathbf{i}\vartheta_2)-\exp(-\mathbf{i}\vartheta_2)} + h_2\pi\delta(\vartheta_2).\end{aligned}$$

It can be shown that the low-wavenumber error is then of second order, and the high-wavenumber error is $\mathcal{O}\left(h^{2N_0-1}\right)$.

As a general remark, the order of the low-wavenumber error is dependent on the quality of the discrete approximation of the initial function, whereas the order of the high-wavenumber error is mainly dependent on the smoothness of the initial data and the number N_0 of Rannacher time steps.

6

6.7. Asymptotic Analysis for the Do Scheme

The analysis above for the MCS scheme shows that the Fourier transformation of the numerical solution can provide important insight in the convergence behaviour of the discretization method. Consider again the model PDE (6.2.1) and assume that spatial discretization is performed with second order central finite difference schemes on a uniform Cartesian grid as described in Section 6.3. Temporal discretization with the Do scheme leads to approximations $U^{\mathrm{Do}}_{N,j,k}$ of $u(x_j,y_k,1)$ and its discrete Fourier transformation is given by (6.4.3). The Fourier transformation of the exact solution is given by (6.2.2). Recall the substitutions

$$\vartheta_1 = \omega_1 h_1, \quad \vartheta_2 = \omega_2 h_2.$$

In this section, the Fourier error

$$\widehat{E}^{\mathrm{Do}}(\omega_1,\omega_2) = \widehat{U}^{\mathrm{Do}}_N(\omega_1 h_1,\omega_2 h_2) - \widehat{u}(\omega_1,\omega_2,1), \quad \text{for } -\pi \leq \omega_1 h_1,\omega_2 h_2 \leq \pi, \tag{6.7.1}$$

is analysed numerically. The observations are used to conjecture a convergence result, i.e. a result on the total error

$$U^{\mathrm{Do}}_{N,j,k} - u(x_j,y_k,1), \tag{6.7.2}$$

which can be approximated by, cf. Section 6.5,

$$\frac{1}{4\pi^2}\int_{-\pi/h_2}^{\pi/h_2}\int_{-\pi/h_1}^{\pi/h_1}\widehat{E}^{\mathrm{Do}}(\omega_1,\omega_2)\exp(\mathbf{i}x_j\omega_1)\exp(\mathbf{i}y_k\omega_2)d\omega_1 d\omega_2,$$

or equivalently

$$\frac{1}{4\pi^2 h_1 h_2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\widehat{E}^{\mathrm{Do}}(\vartheta_1/h_1,\vartheta_2/h_2)\exp(\mathbf{i}j\vartheta_1)\exp(\mathbf{i}k\vartheta_2)d\vartheta_1 d\vartheta_2.$$

In Figure 6.9 the norm $|\widehat{U}_N^{\mathrm{Do}}|$ is shown in the $(\vartheta_1,\vartheta_2)$-domain for the parameter values $\rho=-0.7$, $a_1=2$, $a_2=3$. Discretization is performed with $h_1=h_2=1/6$, $\Delta t=1/8$ and well-known Do parameters $\theta=1/2,1$. For the Rannacher time stepping we considered values $N_0=0,2$. The plots have to be compared with the modulus $|\widehat{u}|$ shown in Figure 6.2. It readily follows that the Fourier domain can again be partitioned into five regions where the Fourier error behaves differently. The regions correspond with the ones illustrated in Figure 6.4.

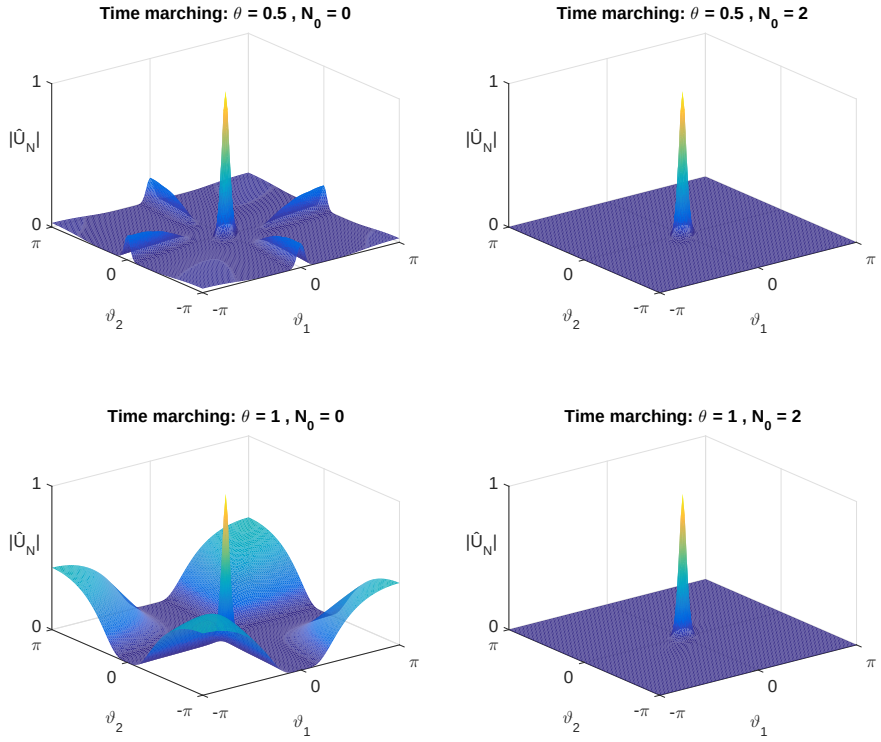


Figure 6.9: Magnitude of the Fourier transform $\widehat{U}_N^{\mathrm{Do}}$ with $N_0=0$ (left) and $N_0=2$ (right) for Do parameter $\theta=1/2$ (top) and $\theta=1$ (bottom). The other parameter values are: $\rho=-0.7, a_1=2, a_2=3, h_1=h_2=1/6, \Delta t=1/8$.

In the low-wavenumber region ① there is a good agreement between $|\widehat{U}_N^{\mathrm{Do}}|$ and $|\widehat{u}|$. Let $c=h_2/h_1$ again be constant, denote $h=h_1$ and assume that

$\lambda = \Delta t/h$ is kept constant. Experiments in region ① for h tending to zero indicate that the Fourier error (6.7.1) decreases in a first-order way as a function of h. A Taylor expansion within Mathematica yields that in this region the pertinent Fourier error can be written as

$$\widehat{u}(\omega_1, \omega_2, 1) s^{[\mathrm{Do},1]} h + \widehat{u}(\omega_1, \omega_2, 1) \mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^3\right) h^2,$$

where $s^{[\mathrm{Do},1]}$ is a polynomial of degree four in $(\mathbf{i}\omega_1,\ \mathbf{i}\omega_2)$, and with coefficients defined by $a_1, a_2, \rho,\ c, \lambda, \theta$. Denote

$$\widehat{E}^{\mathrm{low,Do}} = \widehat{u}(\omega_1, \omega_2, 1) s^{[\mathrm{Do},1]} h.$$

It is readily seen that $\widehat{E}^{\mathrm{low,Do}}$ is only sizeable in the low-wavenumber region ①. By using the Fourier transformations from (6.6.3), it follows that its inverse discrete/continuous Fourier transformation can be approximated by

$$E_{j,k}^{\mathrm{low,Do}} = h C_{x_j, y_k}^{\mathrm{low,Do}}, \tag{6.7.3}$$

for h_1, h_2 simultaneously tending to zero. Here, the coefficient $C_{x_j,y_k}^{\mathrm{low,Do}}$ is only dependent on the point (x_j, y_k) and parameters $a_1, a_2, \rho,\ c, \lambda, \theta$. Note that this error term is only of first order in h, which corresponds to the order of convergence for the Do scheme when it is applied to semidiscretized two-dimensional convection-diffusion equations with smooth initial and boundary data.

If either $|\vartheta_1|$ or $|\vartheta_2|$ is medium and the other one is small or medium, then both the Fourier transforms of the numerical solution and analytical solution seem to be negligible. This suggests that the Fourier error from region ② has no significant contribution to the total error (6.7.2).

In the high-wavenumber region ③, observations similar to the ones in Subsection 6.5.1 can be made. The modulus of the Fourier transform $|\widehat{u}|$ is close to zero, whereas the modulus $|\widehat{U}_N^{\mathrm{Do}}|$ is strongly dependent on N_0 and the Do parameter θ. In this region, the magnitude of $\widehat{U}_N^{\mathrm{Do}}$ is increasing as a function of θ and its modulus is always damped whenever Rannacher time stepping is applied. Additional numerical experiments reveal that if $N_0 = 0$ and λ is fixed, then for h tending to zero the Fourier error (6.7.1) remains constant as a function of $(\vartheta_1, \vartheta_2)$. Given these observations and the expansion from (6.5.16), we *conjecture* that in the high-wavenumber region the Fourier error (6.7.1) can be approximated by

$$\widehat{E}^{\mathrm{high,Do}} = h^{2N_0} \widehat{C}^{\mathrm{high,Do}},$$

where $\widehat{C}^{\mathrm{high,Do}}$ is a function of ϑ_1, ϑ_2 and parameters $\rho, c, \lambda, \theta, N_0$, that is only sizeable in the pertinent region of the Fourier domain. Applying the inverse Fourier transformation (with integrals over ϑ_1, ϑ_2) yields

$$E_{j,k}^{\mathrm{high,Do}} = h^{2N_0 - 2} C_{j,k}^{\mathrm{high,Do}},$$

where $C_{j,k}^{\mathrm{high,Do}}$ is a constant that only depends on the index (j, k) and parameters $\rho, c, \lambda, \theta, N_0$.

Finally, we have region ④ and region ⑤ where either $|\vartheta_1|$ or $|\vartheta_2|$ is large and the other one is not. Extensive numerical experiments reveal that in both

regions of the Fourier domain $\widehat{u}$ is negligible but $\widehat{U}_N^{\mathrm{Do}}$ is sizeable if $\theta = 1/2$ and $N_0 = 0$. For other values of θ or integer N_0, the Fourier transformation of the numerical solution is always negligible. Let U_N, respectively U_N^{Do}, be the numerical solution obtained with the MCS scheme, respectively the Do scheme, and the same parameter values $\rho, c, \lambda, h, N_0 = 0$ and same ADI parameter $\theta = 1/2$. In the pertinent regions, the magnitude of $\widehat{U}_N^{\mathrm{Do}}$ is very similar to the magnitude of $\widehat{U}_N$. Hence, the error in physical space corresponding to the Fourier error in region ④ and region ⑤ can be expected to be similar to (6.6.6). Let

$$E_{j,k}^{\frac{1}{2}\mathrm{Do}} = h^{2N_0-1}(C_{j,y_k}^{\frac{1}{2}\mathrm{Do}} + C_{x_j,k}^{\frac{1}{2}\mathrm{Do}}),$$

with certain coefficients $C_{j,y_k}^{\frac{1}{2}\mathrm{Do}}, C_{x_j,k}^{\frac{1}{2}\mathrm{Do}}$ that are independent of h. By combining the observations above we arrive at the following conjecture:

Conjecture 6.7.1 *Consider the model PDE* (6.2.1). *Assume that spatial discretization is performed by standard second order central finite differences on a uniform Cartesian grid with mesh width $h_1 = h$ in the x-direction and mesh width h_2 in the y-direction. Assume the obtained semidiscrete system is discretized in time by using the Do scheme with parameter θ on a uniform temporal grid with temporal step size Δt. Let $N_0 \geq 0$ denote the number of initial Do time steps that are replaced by $2N_0$ half-time steps of the implicit Euler scheme. If $c = h_2/h_1$ and $\lambda = \Delta t/h$ are kept constant, then as h tends to zero the total error is approximated by*

$$\begin{aligned} U_{N,j,k}^{\mathrm{Do}} - u(x_j, y_k, 1) \approx\ & hC_{x_j,y_k}^{\mathrm{low,Do}} + h^{2N_0-2}C_{j,k}^{\mathrm{high,Do}} \\ & + \mathbb{1}_{\{\theta=\frac{1}{2}\}} h^{2N_0-1}(C_{j,y_k}^{\frac{1}{2}\mathrm{Do}} + C_{x_j,k}^{\frac{1}{2}\mathrm{Do}}), \end{aligned}$$

with coefficients $C_{x_j,y_k}^{\mathrm{low,Do}}, C_{j,k}^{\mathrm{high,Do}}, C_{j,y_k}^{\frac{1}{2}\mathrm{Do}}, C_{x_j,k}^{\frac{1}{2}\mathrm{Do}}$ that are independent of h.

Conjecture 6.7.1 states the total error (6.7.2) is $\mathcal{O}\left(h^{\min\{1,2N_0-2\}}\right)$ and the value $N_0 = 2$ is again a lower bound on N_0 for the Rannacher time stepping in order to ensure convergence of the numerical solution to the exact solution. The conjecture is confirmed by the plots in Figure 6.10 which display total errors (in the maximum norm) in actual numerical experiments for model problem (6.2.1) as a function of $1/h$, with parameter values $\rho = -0.7, a_1 = 2, a_2 = 3$, Do parameter $\theta = 1/2$ (top), $\theta = 1$ (bottom) and with $c = 1$, $0.2 \leq \lambda \leq 0.5$. The computational domain is restricted to grid points $(x_j, y_k) \in [-10, 10] \times [-10, 10]$ and at the boundaries homogeneous Dirichlet boundary conditions are applied, cf. Section 6.6. In the left plots the case $N_0 = 0$ is considered, whereas the right plots show the corresponding results for $N_0 = 2$.

The convergence plots in Figure 6.10 are similar to the ones in Section 6.6. The main difference is that the low-wavenumber error is now only of first order. When no Rannacher time stepping is applied ($N_0 = 0$), the total error decreases in a first order way until a high-wavenumber error exceeds the low-wavenumber error. If $\theta = 1/2$, the total error then increases in a first order way. Eventually, for every Do parameter θ, the total error becomes $\mathcal{O}\left(h^{-2}\right)$ when h tends to

zero. The right plots in Figure 6.10 reveal that if Rannacher time stepping is applied with $N_0 = 2$, then the total error is always $\mathcal{O}(h)$. It is readily seen that the error constant in the right plots is similar to the error constant associated with the low-wavenumber error in the left plots. Thus, by applying Rannacher time stepping with $N_0 = 2$, convergence can be recovered whilst the error constant remains unaffected.

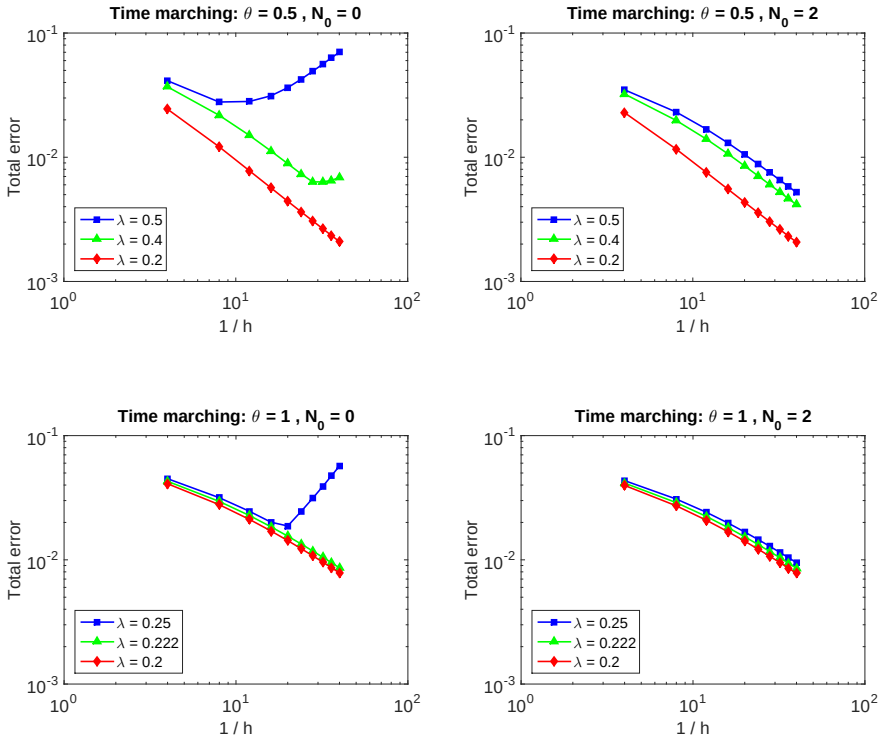


Figure 6.10: Convergence of the numerical solution for Do parameter $\theta = 1/2$ (top), $\theta = 1$ (bottom) and for $N_0 = 0$ (left), $N_0 = 2$ (right). The parameter values are: $\rho = -0.7, a_1 = 2, a_2 = 3$.

The analysis of the Fourier error (6.7.1) in the high-wavenumber region revealed that it is highly dependent on the Do parameter θ. This observation is confirmed by the left plots in Figure 6.10. It is readily seen that for larger values of θ, a much lower value of λ is needed in order for the total error to have the same magnitude. We conjecture that the high-wavenumber error constant $C_{j,k}^{\text{high,Do}}$ is strongly increasing as a function of θ. Additional numerical experiments show that, although less pronounced, the low-wavenumber error is also increasing as a function of θ. Therefore, regardless of the number of Rannacher time steps N_0, it seems more favourable to consider smaller values of θ. In particular, the lowest value of θ which satisfies the stability restrictions from Section 3.3 is given by $\theta = 1/2$.

If one considers an alternative discretization of the Dirac delta, or another

non-smooth initial function, then an analysis similar to that in Subsection 6.6.3 can be performed. Since the expression (6.7.3) is only of first order in h, we expect that the low-wavenumber error is always $\mathcal{O}(h)$ if the initial function is discretized consistently. For the high-wavenumber error, we conjecture that the orders of convergence from Subsection 6.6.3 are also valid if the temporal discretization is performed with the Do scheme. This conjecture is based on the fact that the Fourier errors (6.5.1) and (6.7.1) are very similar in the high-wavenumber region.

6.8. Asymptotic Analysis for the HV Scheme

In this section, we use the discrete/continuous Fourier transformation introduced in Section 6.4 to gain insight in the convergence properties of the HV scheme when it is applied to a semidiscretized model PDE provided with Dirac delta initial data. Let spatial discretization of the convection-diffusion equation (6.2.1) once more be performed with second order central finite difference schemes on a uniform Cartesian grid as described in Section 6.3. Temporal discretization with the HV scheme leads to approximations $U^{\mathrm{HV}}_{N,j,k}$ of $u(x_j, y_k, 1)$. Its discrete Fourier transformation is given by (6.4.4) and, recalling the substitutions

$$\vartheta_1 = \omega_1 h_1, \quad \vartheta_2 = \omega_2 h_2,$$

this leads to the Fourier error

$$\widehat{E}^{\mathrm{HV}}(\omega_1, \omega_2) = \widehat{U}^{\mathrm{HV}}_N(\omega_1 h_1, \omega_2 h_2) - \widehat{u}(\omega_1, \omega_2, 1), \quad \text{for } -\pi \leq \omega_1 h_1, \omega_2 h_2 \leq \pi. \tag{6.8.1}$$

As before, $\widehat{u}$ denotes the Fourier transformation (6.2.2) of the exact solution. Analogously to the previous section, we analyse the Fourier error experimentally to make a conjecture about the total error, which is now defined by

$$U^{\mathrm{HV}}_{N,j,k} - u(x_j, y_k, 1). \tag{6.8.2}$$

Recall that the total error can be approximated by,

$$\tfrac{1}{4\pi^2} \int_{-\pi/h_2}^{\pi/h_2} \int_{-\pi/h_1}^{\pi/h_1} \widehat{E}^{\mathrm{HV}}(\omega_1, \omega_2) \exp(\mathbf{i} x_j \omega_1) \exp(\mathbf{i} y_k \omega_2) d\omega_1 d\omega_2,$$

or equivalently

$$\tfrac{1}{4\pi^2 h_1 h_2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \widehat{E}^{\mathrm{HV}}(\vartheta_1/h_1, \vartheta_2/h_2) \exp(\mathbf{i} j \vartheta_1) \exp(\mathbf{i} k \vartheta_2) d\vartheta_1 d\vartheta_2,$$

for h_1, h_2 simultaneously tending to zero.

The analysis for the MCS scheme and the Do scheme revealed that the Fourier error can have different properties in different parts of the domain. In Figure 6.11 this is illustrated for the numerical solution obtained with the HV scheme, i.e. for (6.8.1). The modulus $|\widehat{U}^{\mathrm{HV}}_N|$ is shown in the $(\vartheta_1, \vartheta_2)$-domain for the parameter values $\rho = -0.7$, $a_1 = 2$, $a_2 = 3$. The spatial and temporal discretization is performed with $h_1 = h_2 = 1/6$, $\Delta t = 1/8$ and well-known HV

parameters $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}, 1$. Note that we do not consider $\theta = 1/2$, since this parameter value does not guarantee unconditional stability of the HV scheme if convection terms are present, cf. Section 3.3. For the Rannacher time stepping we considered values $N_0 = 0, 2$. By comparing these plots with the modulus $|\widehat{u}|$, shown in Figure 6.2, it readily follows that the Fourier error (6.8.1) has different properties in the five regions illustrated in Figure 6.4. Additional experiments with smaller values of $h_1, h_2, \Delta t$ reveal, however, that for all parameter values $\theta \geq \frac{1}{2} + \frac{1}{6}\sqrt{3}$ it sufficient to distinguish three different regions where the Fourier error behaves differently. These regions are illustrated in Figure 6.12.

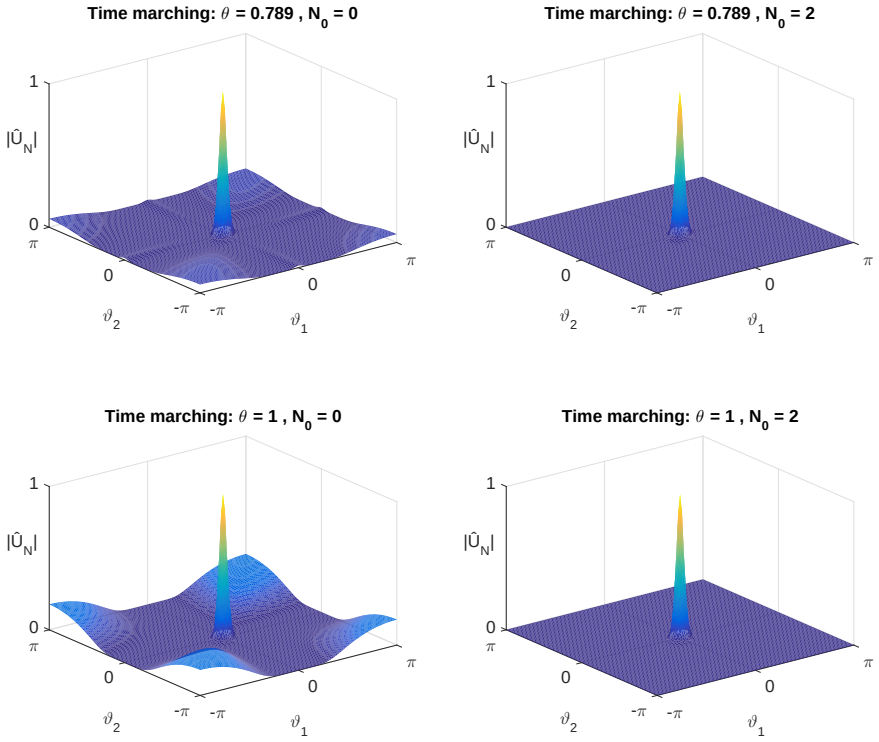


Figure 6.11: Magnitude of the Fourier transform $\widehat{U}_N^{\mathrm{HV}}$ with $N_0 = 0$ (left) and $N_0 = 2$ (right) for HV parameter $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ (top) and $\theta = 1$ (bottom). The other parameter values are: $\rho = -0.7, a_1 = 2, a_2 = 3, h_1 = h_2 = 1/6, \Delta t = 1/8$.

In the low-wavenumber region ①, where both $|\vartheta_1|$, $|\vartheta_2|$ are small, there is a good agreement between $|\widehat{u}|$ and $|\widehat{U}_N^{\mathrm{HV}}|$. For the convergence analysis we denote $h = h_1$ and assume that the ratios $c = h_2/h_1$, $\lambda = \Delta t/h$ are kept constant. Extensive numerical experiments show that in region ① the Fourier error (6.8.1) decreases in a second order fashion as h tends to zero. This is confirmed by a Taylor expansion within Mathematica, which shows that in this region the pertinent Fourier error can be written as

$$\widehat{u}(\omega_1, \omega_2, 1)(s^{[\mathrm{HV},2]} + N_0^{[\mathrm{HV},2]})h^2 + \widehat{u}(\omega_1, \omega_2, 1)\mathcal{O}\left(1 + (\omega_1^2 + c^2\omega_2^2)^4\right) h^3.$$

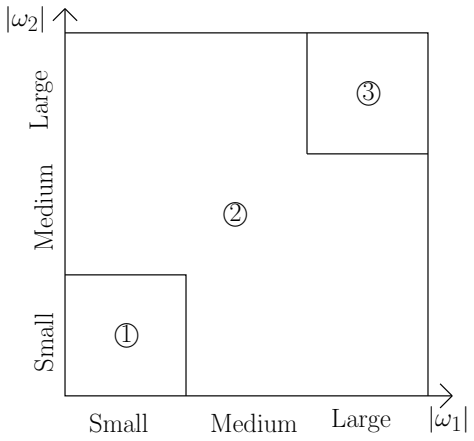


Figure 6.12: Illustration of the different disjoint regions of the Fourier domain in case of the HV scheme.

Here $s^{[\mathrm{HV},2]}$, respectively $N_0^{[\mathrm{HV},2]}$, is a polynomial of degree six, respectively four, in $(\mathbf{i}\omega_1, \mathbf{i}\omega_2)$ and with coefficients defined by $a_1, a_2, \rho, c, \lambda, \theta$. The latter polynomial is, of course, also dependent on N_0. Let

$$\widehat{E}^{\mathrm{low,HV}} = \widehat{u}(\omega_1, \omega_2, 1)(s^{[\mathrm{HV},2]} + N_0^{[\mathrm{HV},2]})h^2.$$

This expression corresponds with the leading order term in the Taylor expansion. It is clear that $\widehat{E}^{\mathrm{low,HV}}$ is only sizeable in the low-wavenumber region and by using the Fourier transformations from (6.6.3) it readily follows that its inverse discrete/continuous Fourier transformation can be approximated by

$$E_{j,k}^{\mathrm{low,HV}} = h^2 C_{x_j,y_k}^{\mathrm{low,HV}},$$

for h_1, h_2 simultaneously tending to zero. The magnitude of the coefficient $C_{x_j,y_k}^{\mathrm{low,HV}}$ is only dependent on the point (x_j, y_k) and parameters a_1,a_2,ρ,c,λ,θ, N_0. The error in physical space, corresponding to the Fourier error (6.8.1), is clearly of second order. This is in line with the observations from Section 6.6 and Section 6.7 for the MCS scheme and the Do scheme, where it is shown that the order of the low-wavenumber error equals the classical order of consistency of the pertinent ADI scheme.

In region ② of Figure 6.12, $|\vartheta_1|$ and $|\vartheta_2|$ are not both small nor both large. The Fourier transform of the analytical solution is then negligible and additional numerical experiments show that the magnitude of $|\widehat{U}_N^{\mathrm{HV}}|$ readily becomes negligible as h tends to zero. This suggests that the Fourier error in this region has no significant contribution to the total error (6.8.2).

Finally, we consider the high-wavenumber region ③. When $|\vartheta_1|$, $|\vartheta_2|$ are both large, the modulus of the Fourier transform $\widehat{u}$ is close to zero. The magnitude of $\widehat{U}_N^{\mathrm{HV}}$ is, however, strongly dependent on N_0 and the HV parameter θ. If $N_0 = 0$, then the modulus $|\widehat{U}_N^{\mathrm{HV}}|$ is increasing as a function of θ. If Rannacher time stepping is applied, the pertinent modulus is always damped. Consider

the case $N_0 = 0$ and let λ be fixed. Extensive numerical experiments show that the Fourier error (6.8.1) remains constant as a function of $(\vartheta_1, \vartheta_2)$ when h tends to zero. Similarly to the previous section, and based on the expansion from (6.5.16), we *conjecture* that in the high-wavenumber region the Fourier error can be approximated by

$$\widehat{E}^{\text{high,HV}} = h^{2N_0} \widehat{C}^{\text{high,HV}}, \tag{6.8.3}$$

with $\widehat{C}^{\text{high,HV}}$ a function of ϑ_1,ϑ_2 and parameters $\rho, c, \lambda, \theta, N_0$ that is only sizeable in region ③. If the latter property holds, then the inverse Fourier transformation of (6.8.3) can be written as

$$E^{\text{high,HV}} = h^{2N_0-2} C_{j,k}^{\text{high,HV}},$$

where the coefficients $C_{j,k}^{\text{high,HV}}$ only depend on the index (j,k) and parameters $\rho, c, \lambda, \theta, N_0$. By combining our observations for the different regions, we arrive at the following conjecture:

Conjecture 6.8.1 *Consider the model PDE* (6.2.1)*. Assume that spatial discretization is performed by standard second order central finite differences on a uniform Cartesian grid with mesh width $h_1 = h$ in the x-direction and mesh width h_2 in the y-direction. Assume the obtained semidiscrete system is discretized in time by using the HV scheme with parameter θ on a uniform temporal grid with temporal step size Δt. Let $N_0 \geq 0$ denote the number of initial HV time steps that are replaced by $2N_0$ half-time steps of the implicit Euler scheme. If $c = h_2/h_1$ and $\lambda = \Delta t/h$ are kept constant, then as h tends to zero the total error is approximated by*

$$U_{N,j,k}^{\text{HV}} - u(x_j, y_k, 1) \approx h^2 C_{x_j,y_k}^{\text{low,HV}} + h^{2N_0-2} C_{j,k}^{\text{high,HV}},$$

with coefficients $C_{x_j,y_k}^{\text{low,HV}}, C_{j,k}^{\text{high,HV}}$ that are independent of h.

According to Conjecture 6.8.1, the total error (6.8.2) is $\mathcal{O}\left(h^{\min\{2,2N_0-2\}}\right)$ and $N_0 = 2$ is once more a lower bound on N_0 for the Rannacher time stepping in order to ensure convergence of the numerical solution to the exact solution. Our conjecture is confirmed by the plots in Figure 6.13. Here, the total errors are displayed (in the maximum norm) in actual numerical experiments for model problem (6.2.1) as a function of $1/h$, with parameter values $\rho = -0.7$, $a_1 = 2$, $a_2 = 3$, HV parameter $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ (top), $\theta = 1$ (bottom) and with $c = 1$, $0.2 \leq \lambda \leq 0.5$. As before, the spatial domain is restricted to $[-10, 10] \times [-10, 10]$ and at the boundaries homogeneous boundary conditions are applied. The left plots display the results for the case $N_0 = 0$, and in the right plots the value $N_0 = 2$ is considered.

The plots in Figure 6.13 are very similar to the ones in Subsection 6.6.1. The left plots, where $N_0 = 0$, reveal second order convergence behaviour until h reaches a critical value where the high-wavenumber error starts exceeding the low-wavenumber error. It can be observed that the high-wavenumber error is highly dependent on the ratio λ and the HV parameter θ. For the case where

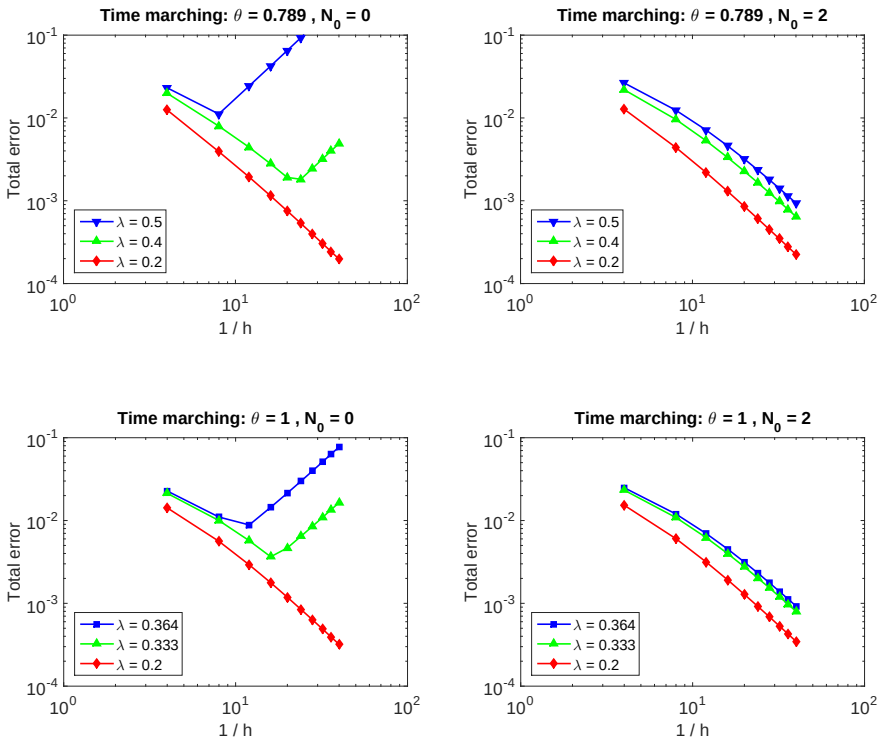


Figure 6.13: Convergence of the numerical solution for HV parameter $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ *(top),* $\theta = 1$ *(bottom) and for* $N_0 = 0$ *(left),* $N_0 = 2$ *(right). The parameter values are:* $\rho = -0.7, a_1 = 2, a_2 = 3$.

6

Rannacher time stepping is applied with $N_0 = 2$, the right plots in Figure 6.13 show that the total error is always $\mathcal{O}\left(h^2\right)$. We notice that the error constant in the right plots is only slightly larger than the error constant corresponding to the low-wavenumber error in the left plots. Thus, by applying Rannacher time stepping with $N_0 = 2$, second order convergence can be recovered at the small cost of a marginally larger error constant. Recall that the same observation is made in Subsection 6.6.1 for the MCS scheme.

From Figure 6.11 it can be observed that in the high-wavenumber region the Fourier error (6.8.1) is highly dependent on the HV parameter θ. In order to get decent plots in Figure 6.13 for larger values of θ, it is necessary to consider smaller values of λ. By combining these observations, we conjecture that the constant $C_{j,k}^{\text{high,HV}}$ is strongly increasing as a function of θ. A similar, although less pronounced, observation can be made for the low-wavenumber error constant $C_{x_j,y_k}^{\text{low,HV}}$. Regardless of the number of Rannacher time steps N_0, it seems more favourable to consider smaller values of θ. In particular, the lowest value of θ which satisfies the stability restrictions from Section 3.3 is given by $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$.

Finally, we consider alternative discretizations of the Dirac delta and other

non-smooth initial functions. An analysis similar to that in Subsection 6.6.3 yields that the order of the low-wavenumber error is mainly dependent on the quality of the discrete approximation of the initial function. If an appropriate discretization of the initial data is used, then the low-wavenumber error is of second order. The order of the high-wavenumber error is dependent on the smoothness of the initial data and the number N_0 of Rannacher time steps.

6.9. Conclusion

If the initial data is non-smooth, application of the ADI schemes for multi-dimensional time-dependent convection-diffusion equations with mixed spatial derivative terms can cause spurious erratic behaviour in the numerical solution. A motivating example, with the two-dimensional Black–Scholes equation for a two-asset cash-or-nothing option, shows that this undesirable feature can be resolved by replacing the very first N_0 ADI time steps by $2N_0$ half-time steps of the implicit Euler scheme, with $N_0 = 2$.

We proved by Fourier analysis that for a model two-dimensional convection-diffusion equation with mixed-derivative term and with Dirac delta initial data, the total error stemming from temporal discretization with the MCS scheme can be approximated by the sum of a low-wavenumber error of $\mathcal{O}(h^2)$ and a high-wavenumber error of $\mathcal{O}(h^{2N_0-2})$. In case the MCS scheme reduces to the CS scheme, i.e. when $\theta = 1/2$, this has to be augmented with an extra error term of $\mathcal{O}(h^{2N_0-1})$. Numerical experiments indicate that a similar decomposition of the total error can be performed if temporal discretization is performed with the Do scheme or the HV scheme. We conjecture that the error in physical space corresponding to the Do scheme can be written as the sum of a low-wavenumber error of $\mathcal{O}(h)$, a high-wavenumber error of $\mathcal{O}(h^{2N_0-2})$, and an additional error term of $\mathcal{O}(h^{2N_0-1})$ if $\theta = 1/2$. For the HV scheme we conjecture that the total error consists of a low-wavenumber error of $\mathcal{O}(h^2)$ and a high-wavenumber error of $\mathcal{O}(h^{2N_0-2})$.

For all ADI schemes considered, $N_0 = 2$ is the minimum on N_0 in order to guarantee convergence of the numerical solution to the exact solution, in the maximum norm. In general this choice for N_0 is optimal since larger values will increase the low-wavenumber error. Our convergence analysis and numerical experiments further indicate that it is favourable to consider small values of the ADI parameter θ. However, it is necessary to take into account the lower bounds on θ in order for our asymptotic analysis to be valid. The smallest value which satisfies all the restrictions for the MCS scheme, independent of the parameters of the model, is given by $\theta = 1/3$. For the Do scheme, respectively the HV scheme, stability restrictions lead to a lower bound of $\theta = 1/2$, respectively $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$. These lower bounds are also the most common values of θ for the respective ADI schemes considered in the literature.

CHAPTER 7

Adjoint Calibration of SLV Models

7.1. Introduction

In contemporary financial mathematics, *stochastic local volatility* (SLV) *models* are state-of-the-art for describing asset price processes, notably foreign exchange (FX) rates, see e.g. [50, 64]. They constitute a natural combination of *local volatility* (LV) and *stochastic volatility* (SV) models. Let S_τ represent the exchange rate at time $\tau \geq 0$ and let the spot value S_0 be given. For modelling the exchange rate we consider the transformation $X_\tau = \log(S_\tau/S_0)$ since the transformed variable X_τ reflects better the duality between the exchange rate S_τ and $1/S_\tau$, where the latter one is the exchange rate when the role of the domestic currency and the foreign currency are swapped. We deal in this chapter with general SLV models defined by the *stochastic differential equation* (SDE)

$$\begin{cases} dX_\tau = (r_d - r_f - \frac{1}{2}\sigma^2_{\mathrm{SLV}}(X_\tau,\tau)\psi^2(V_\tau))d\tau + \sigma_{\mathrm{SLV}}(X_\tau,\tau)\psi(V_\tau)dW^{(1)}_\tau, \\ \\ dV_\tau = \kappa(\eta - V_\tau)d\tau + \xi V^\alpha_\tau dW^{(2)}_\tau, \end{cases} \tag{7.1.1}$$

with $\psi(v)$ a non-negative function, α a non-negative parameter, κ, η, ξ strictly positive parameters, $W^{(1)}_\tau, W^{(2)}_\tau$ Brownian motions with $dW^{(1)}_\tau \cdot dW^{(2)}_\tau = \rho d\tau$, $-1 \leq \rho \leq 1$, and given spot values S_0, V_0. The function $\sigma_{\mathrm{SLV}}(x,\tau)$ is often called *the leverage function* and r_d, respectively r_f, denotes the risk-free interest rate in the domestic currency, respectively foreign currency.

The SLV model (7.1.1) can be viewed as obtained from a mixture of the LV model

$$dX_{\mathrm{LV},\tau} = (r_d - r_f - \tfrac{1}{2}\sigma^2_{\mathrm{LV}}(X_{\mathrm{LV},\tau},\tau))d\tau + \sigma_{\mathrm{LV}}(X_{\mathrm{LV},\tau},\tau)dW_\tau, \tag{7.1.2}$$

with LV function $\sigma_{\mathrm{LV}}(x,\tau)$, and the SV model

$$\begin{cases} dX_{\mathrm{SV},\tau} = (r_d - r_f - \psi^2(V_{\mathrm{SV},\tau}))d\tau + \psi(V_{\mathrm{SV},\tau})dW^{(1)}_{\mathrm{SV},\tau}, \\ \\ dV_{\mathrm{SV},\tau} = \kappa(\eta - V_{\mathrm{SV},\tau})d\tau + \xi V^\alpha_{\mathrm{SV},\tau}dW^{(2)}_{\mathrm{SV},\tau}. \end{cases} \tag{7.1.3}$$

This chapter is based on the article 'An adjoint method for the exact calibration of stochastic local volatility models', published in J. Comp. Sci., doi:10.1016/j.jocs.2017.02.004, 2017 [70].

Clearly, if $\sigma_{\mathrm{SLV}}(x,\tau)$ is identically equal to one, then the SLV model reduces to a SV model. Next, if the stochastic volatility parameter ξ is equal to zero, then the SLV model reduces to a LV model. The LV function $\sigma_{\mathrm{LV}}(x,\tau)$ can be determined by the Dupire formula [15] such that the LV model reproduces the known market prices for European call and put options. Since the LV model is completely determined by the LV function, it offers no flexibility in matching the market dynamics. SV models are typically well-suited to reflect forward volatilities, but they are often unable to capture the volatility smile exactly, see e.g. [66]. By combining features of the LV model with features of the SV models, SLV models are able to match the market dynamics and to reproduce the market prices for European call and put options.

The choice $\psi(v) = \sqrt{v}$, $\alpha = 1/2$ corresponds to the well-known Heston-based S(L)V model, the choice $\psi(v) = v$, $\alpha = 1$ to the S(L)V model considered in [64] and the choice $\psi(v) = \exp(v)$, $\alpha = 0$ corresponds to the S(L)V model based on the exponential Gaussian Ornstein–Uhlenbeck model described in [61].

If α is strictly positive, we assume that $\psi(0) = 0$ and the processes $V_\tau, V_{\mathrm{SV},\tau}$ are non-negative. For $0 < \alpha < 1/2$ it holds that $V_\tau = 0$ is attainable, for $\alpha > 1/2$ it holds that $V_\tau = 0$ is unattainable, and for $\alpha = 1/2$ one has that $V_\tau = 0$ is attainable if $2\kappa\eta < \xi^2$, see e.g. [2]. The analogous result is true for the pure SV model (7.1.3).

In financial practice, $\sigma_{\mathrm{LV}}(x,\tau)$ is determined such that the LV model (7.1.2) yields the exact market prices for vanilla options, see e.g. [4, 15], and the parameters κ, η, ξ are chosen such that the SV model (7.1.3) reflects the market dynamics of the underlying asset, see e.g. [64]. Next, the leverage function σ_{SLV} is calibrated such that the SLV model yields the exact market prices for European call and put options. In the literature, no closed-form analytical relationship appears to be available between the leverage function and the fair value of vanilla options within the SLV model. Accordingly, in financial practice the leverage function is calibrated by making use of a relationship between the SLV model and the LV model. It is well-known, see e.g. [23, 63], that these models yield the same marginal distribution for the exchange rate S_τ, and hence always define the same fair value for vanilla options, if the leverage function $\sigma_{\mathrm{SLV}}(x,\tau)$ satisfies

$$\sigma^2_{\mathrm{LV}}(x,\tau) = \mathbb{E}[\sigma^2_{\mathrm{SLV}}(X_\tau,\tau)\psi^2(V_\tau)|X_\tau = x] = \sigma^2_{\mathrm{SLV}}(x,\tau)\mathbb{E}[\psi^2(V_\tau)|X_\tau = x], \tag{7.1.4}$$

for all $x \in \mathbb{R}, \tau \geq 0$. The latter conditional expectation can be written as

$$\mathbb{E}[\psi^2(V_\tau)|X_\tau = x] = \frac{\int_{-\infty}^{\infty} \psi^2(v) p(x,v,\tau;X_0,V_0)dv}{\int_{-\infty}^{\infty} p(x,v,\tau;X_0,V_0)dv}, \tag{7.1.5}$$

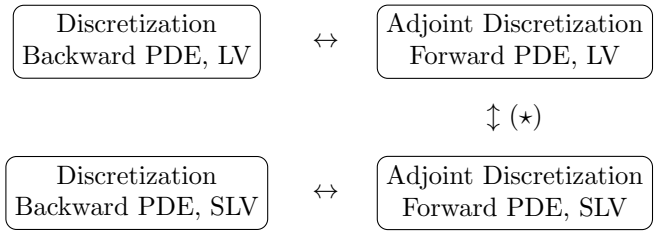
where $p(x,v,\tau;X_0,V_0)$ denotes the joint density of (X_τ, V_τ) given by the SLV model. Since the LV function is determined such that the LV model yields exactly the observed market prices for vanilla options, the SLV model will also exactly define the same fair value whenever one is able to determine the conditional expectation above and subsequently defines the leverage function by (7.1.4). This conditional expectation itself depends on $\sigma_{\mathrm{SLV}}(x,\tau)$, however, and determining it is a highly non-trivial task. Recently, a variety of numerical

techniques, see e.g. [8, 17, 29, 58, 66], has been proposed in order to approximate this conditional expectation and to approximate the appropriate leverage function.

The numerical techniques presented in the references above do not take into account explicitly that, even if the LV function is known analytically, it is often not possible to determine exactly the corresponding fair value of vanilla options. Even within the LV model one relies on numerical methods in order to approximate the fair option values. A common approach consists of numerically solving the corresponding backward PDE by for example finite difference or finite volume methods, see e.g. [65]. When calibrating the SLV model to the LV model, the best result one can thus aim for is that the numerical approximation of the fair value of vanilla options is the same for both models whenever similar numerical valuation methods are used.

In this chapter we assume that the fair option value (within the LV model, respectively SLV model) is approximated through numerically solving the backward PDE (corresponding to the LV model, respectively SLV model) by standard finite difference methods, see e.g. Chapter 2. Given such a spatial discretization for the backward PDE, an *adjoint spatial discretization* will be introduced for the corresponding forward PDE. This adjoint spatial discretization has the important property that it always defines *exactly the same approximation* for the fair value of non-path-dependent European options as the approximation given by the discretization of the backward equation. Moreover, if similar spatial discretizations are used for the backward PDE associated with the LV model and the backward PDE associated with the SLV model, then their adjoint spatial discretizations can be employed to *create an exact match between the approximations for the fair value of vanilla options within the LV model and the SLV model.*

The main contributions of this chapter can be visualized in the following scheme:



Here relationship (⋆) can only be achieved if similar discretizations are used for the backward PDEs stemming from the LV and SLV models.

In order to compute the leverage function that makes relationship (⋆) valid, a large system of non-linear ODEs needs to be solved. Since this system of ODEs is stemming from spatial discretization of a two-dimensional PDE, we employ the MCS scheme (3.2.5) to increase the computational efficiency in the numerical solution. The non-linearity is handled by an iteration procedure.

The outline of this chapter is as follows. In Section 7.2 a relationship between the forward PDE and backward PDE is introduced, both for the case

of the SLV model as for the case of the LV model. This relationship is preserved at the semidiscrete level in Section 7.3: given a spatial discretization of the backward PDE, an adjoint spatial discretization for the forward PDE is defined such that both discretizations yield identical approximations for the fair value of non-path-dependent European options. In Section 7.4 an actual spatial discretization, using second order central finite difference schemes, is constructed for the backward PDE stemming from the SLV model and subsequently the corresponding adjoint spatial discretization is stated. The main result of the chapter is derived in Section 7.5. It is shown that, under some assumptions, the adjoint spatial discretization can be employed to obtain an expression for the leverage function such that the approximation of the fair value of vanilla options is the same for the LV and SLV models. In order to effectively use this expression, one has to solve a large system of non-linear ODEs. In Section 7.6 we apply the MCS scheme for the numerical solution of this ODE system and in Section 7.7 an iteration procedure is described for handling the non-linearity. In Section 7.8 ample numerical experiments are presented to illustrate the performance of the obtained SLV calibration procedure. The final Section 7.9 gives concluding remarks.

7.2. Relationship Between the Forward and the Backward Kolmogorov Equation

Consider a European-style option with maturity T and payoff u_0. Denote by $u(x,v,t)$ the *non-discounted fair value* of the option under the SLV model (7.1.1) at time to maturity t, that is at time level $\tau = T-t$, if $S_\tau = S_0\exp(x)$ and $V_\tau = v$. It is well-known, see e.g. [8], that the function u satisfies the *backward Kolmogorov equation*

$$\begin{aligned}\frac{\partial}{\partial t}u = {} & \tfrac{1}{2}\sigma_{\mathrm{SLV}}^2(x,T-t)\psi^2(v)\tfrac{\partial^2}{\partial x^2}u + \rho\xi\sigma_{\mathrm{SLV}}(x,T-t)\psi(v)v^\alpha\tfrac{\partial^2}{\partial x\partial v}u + \tfrac{1}{2}\xi^2v^{2\alpha}\tfrac{\partial^2}{\partial v^2}u \\ & + (r_d - r_f - \tfrac{1}{2}\sigma_{\mathrm{SLV}}^2(x,T-t)\psi^2(v))\tfrac{\partial}{\partial x}u + \kappa(\eta - v)\tfrac{\partial}{\partial v}u,\end{aligned} \tag{7.2.1}$$

for $x,v \in \mathbb{R}$, $0 < t \leq T$. At maturity, i.e. at time level $\tau = T$, the initial condition $u(x,v,0)$ is defined by the payoff u_0 of the option. By solving PDE (7.2.1), the fair value $e^{-r_dT}u(X_0,V_0,T)$ of the option under the SLV model can be determined at the spot, i.e. at $\tau = 0$. For strictly positive values of the parameter α, the process V_τ is non-negative and the spatial domain in the v-direction reduces to $v \geq 0$.

If the option under consideration is non-path-dependent, then the payoff u_0 is only a function of (X_T,V_T), the initial condition for (7.2.1) is given by $u(x,v,0) = u_0(x,v)$ and the non-discounted fair value $u(x,v,t)$ of the option can be written as

$$u(x,v,t) = \mathbb{E}[u_0(X_T,V_T)|X_{T-t} = x, V_{T-t} = v],$$

for $0 \leq t \leq T$. By making use of the tower property for conditional expectations

it readily follows that

$$\begin{aligned} u(X_0, V_0, T) &= \mathbb{E}[u(X_{T-t}, V_{T-t}, t)|X_0, V_0] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x, v, t)p(x, v, T-t; X_0, V_0)dxdv, \end{aligned} \tag{7.2.2}$$

for $0 \leq t \leq T$. Recall that $p(x, v, \tau; X_0, V_0)$ denotes the joint density of (X_τ, V_τ) under the SLV model (7.1.1). If the parameter α is chosen strictly positive, then the integral with respect to v can be taken from $v = 0$. In particular, the fair value of non-path-dependent European options at the spot can also be computed by evaluating the integral

$$e^{-r_d T}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x, v, 0)p(x, v, T; X_0, V_0)dxdv, \tag{7.2.3}$$

where $u(x, v, 0)$ is defined by the payoff of the option.

It can be shown, see e.g. [59], that the joint density $p(x, v, \tau; X_0, V_0)$ satisfies the *forward Kolmogorov equation*

$$\begin{aligned} \tfrac{\partial}{\partial \tau}p = {} & \tfrac{1}{2}\tfrac{\partial^2}{\partial x^2}\left(\sigma_{\text{SLV}}^2\psi^2(v)p\right) + \tfrac{\partial^2}{\partial x \partial v}\left(\rho\xi\sigma_{\text{SLV}}\psi(v)v^\alpha p\right) + \tfrac{1}{2}\tfrac{\partial^2}{\partial v^2}\left(\xi^2 v^{2\alpha}p\right) \\ & - \tfrac{\partial}{\partial x}\left((r_d - r_f - \tfrac{1}{2}\sigma_{\text{SLV}}^2\psi^2(v))p\right) - \tfrac{\partial}{\partial v}\left(\kappa(\eta - v)p\right), \end{aligned} \tag{7.2.4}$$

for $x, v \in \mathbb{R}, \tau > 0$ and with initial condition

$$p(x, v, 0; X_0, V_0) = \delta(x - X_0)\delta(v - V_0),$$

where δ denotes the Dirac delta function. For ease of presentation, the dependency of σ_{SLV} on (x, τ) and the dependency of p on $(x, v, \tau; X_0, V_0)$ is omitted in (7.2.4). Recall that the process V_τ is non-negative whenever α is strictly positive. In this case the spatial domain of the PDE in the v-direction is naturally restricted to $v \geq 0$. The integrals in (7.1.5), (7.2.2), (7.2.3) with respect to the v-variable can then be taken from 0 to infinity.

Equation (7.2.2) establishes a *fundamental relationship between the forward and backward Kolmogorov equation*. It states that the fair value of non-path-dependent European options under the SLV model can be seen as the combination of the solution of two different PDEs. By considering the extreme time value $\tau = 0$ $(t = T)$, or $\tau = T$ $(t = 0)$, only one PDE has to be solved. In the forthcoming sections relationship (7.2.2) will be employed to define an adjoint spatial discretization for the forward equation.

Even if the functions u and p are known exactly, the integrals in (7.2.2) can often not be calculated analytically and one relies on numerical integration methods in order to approximate them. In this chapter we assume that the integrand is known on a Cartesian grid. Denote by m_1, respectively m_2, the number of spatial grid points in the x-direction, respectively v-direction. The Cartesian grid is given by

$$(x_j, v_k) \quad \text{for } 1 \leq j \leq m_1,\ 1 \leq k \leq m_2, \tag{7.2.5}$$

with

$$x_{\min} = x_1 < x_2 < \cdots < x_{m_1} = x_{\max}, \quad v_{\min} = v_1 < v_2 < \cdots < v_{m_2} = v_{\max},$$

and $x_{\min} < X_0 < x_{\max}$, $v_{\min} < V_0 < v_{\max}$. Define spatial mesh widths $\Delta x_j = x_j - x_{j-1}$ for $2 \leq j \leq m_1$, $\Delta v_k = v_k - v_{k-1}$ for $2 \leq k \leq m_2$ and put $\Delta x_1 = \Delta x_{m_1+1} = \Delta v_1 = \Delta v_{m_2+1} = 0$. When working with Cartesian grids, most numerical integration methods approximate the expression (7.2.2) by

$$u(X_0, V_0, T) \approx \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} p(x_j, v_k, T-t; X_0, V_0) u(x_j, v_k, t) w_{x,j} w_{v,k}, \qquad (7.2.6)$$

for certain weights $w_{x,j}, w_{v,k}$. If the numerical integration is performed with the trapezoidal rule, then the weights are given by

$$w_{x,j} = \tfrac{\Delta x_j + \Delta x_{j+1}}{2} \quad \text{for } 1 \leq j \leq m_1, \quad w_{v,k} = \tfrac{\Delta v_k + \Delta v_{k+1}}{2} \quad \text{for } 1 \leq k \leq m_2.$$

The values $x_{\min}, v_{\min}, x_{\max}, v_{\max}$ have to lie sufficiently far away from (X_0, V_0) such that the truncation error is negligible. If α is strictly positive, the value $v_{\min}$ can be set equal to zero. In order for (7.2.6) to be exact if $t = T$ ($\tau = 0$), it is assumed that there exist indices j_0, k_0 such that $(x_{j_0}, v_{k_0}) = (X_0, V_0)$ and the approximation $p(x, v, 0; X_0; V_0) \approx p_0(x, v)$ is used with

$$p_0(x,v) = \begin{cases} \frac{1}{w_{x,j_0} w_{v,k_0}} & \text{if } \begin{cases} x \in [x_{j_0} - \Delta x_{j_0}/2, x_{j_0} + \Delta x_{j_0+1}/2], \\ v \in [v_{k_0} - \Delta v_{k_0}/2, v_{k_0} + \Delta v_{k_0+1}/2], \end{cases} \\ 0 & \text{else.} \end{cases}$$

Analogously as above, consider within the LV model a European-style option with payoff $u_{\mathrm{LV},0}$ at maturity T and denote by $u_{\mathrm{LV}}(x,t)$ the non-discounted fair value of the option under the LV model (7.1.2) at time $\tau = T - t$ if $X_{\mathrm{LV},\tau} = x$. The function u_{LV} satisfies the backward Kolmogorov equation

$$\tfrac{\partial}{\partial t} u_{\mathrm{LV}} = \tfrac{1}{2}\sigma_{\mathrm{LV}}^2 \tfrac{\partial^2}{\partial x^2} u_{\mathrm{LV}} + (r_d - r_f - \tfrac{1}{2}\sigma_{\mathrm{LV}}^2) \tfrac{\partial}{\partial x} u_{\mathrm{LV}}, \qquad (7.2.7)$$

for $x \in \mathbb{R}$, $0 < t \leq T$ and with initial condition $u_{\mathrm{LV}}(x, 0)$ defined by the payoff $u_{\mathrm{LV},0}$ of the option. The non-discounted fair value of the option at the spot ($\tau = 0$) is then given by $u_{\mathrm{LV}}(X_0, T)$. For non-path-dependent options this fair value can also be formulated as

$$u(X_0, T) = \int_{-\infty}^{\infty} u_{\mathrm{LV}}(x,t) p_{\mathrm{LV}}(x, T-t; X_0) dx, \qquad (7.2.8)$$

for $0 \leq t \leq T$, where $p_{\mathrm{LV}}(x, \tau; X_0)$ denotes the density of the process $X_{\mathrm{LV},\tau}$ in the LV model (7.1.2). It can be shown, see e.g. [8], that this density function satisfies the forward Kolmogorov equation

$$\tfrac{\partial}{\partial \tau} p_{\mathrm{LV}} = \tfrac{1}{2} \tfrac{\partial^2}{\partial x^2} \left(\sigma_{\mathrm{LV}}^2 p_{\mathrm{LV}}\right) - \tfrac{\partial}{\partial x} \left((r_d - r_f - \tfrac{1}{2}\sigma_{\mathrm{LV}}^2) p_{\mathrm{LV}}\right), \qquad (7.2.9)$$

for $x \in \mathbb{R}, \tau > 0$, and with initial condition $p_{\mathrm{LV}}(x, 0; X_0) = \delta(x - X_0)$. Hence, the expression (7.2.8) establishes a fundamental relationship between the forward and backward Kolmogorov equation which is similar to (7.2.2). By applying the same numerical integration technique as above, the fair value from (7.2.8) can be approximated by

$$u(X_0, T) \approx \sum_{j=1}^{m_1} p_{\mathrm{LV}}(x_j, T-t; X_0) u(x_j, t) w_{x,j}.$$

Recall that the SLV model is calibrated perfectly to the LV model if the leverage function is defined by (7.1.4). It was shown by Gyöngy [23] that under this assumption both processes X_τ and $X_{\mathrm{LV},\tau}$ have the same marginal densities, i.e. that

$$\int_{-\infty}^{\infty} p(x, v, \tau; X_0, V_0)dv = p_{\mathrm{LV}}(x, \tau; X_0) \tag{7.2.10}$$

for $x \in \mathbb{R}, \tau \geq 0$. From now on, for the ease of presentation, the dependency of p and p_{LV} on the spot values X_0, V_0 is omitted.

7.3. Adjoint Spatial Discretization

In general the values $p(x_j, v_k, \tau)$ and $u(x_j, v_k, t)$ are not known exactly if $\tau > 0$ and $t > 0$, respectively, and one relies on numerical methods to approximate them. Extensive literature is available on numerical techniques to solve backward Kolmogorov equations, see e.g. [65]. Recall that in financial mathematics a common approach in order to approximate the fair value of options is given by numerically solving the pertinent PDE using the general MOL, cf. [35]. The PDE is first discretized in the spatial variables x and v, yielding a large system of stiff ordinary differential equations, cf. Chapter 2. This semidiscrete system is subsequently solved by applying a suitable implicit time stepping method, cf. Chapter 3.

Spatial discretization by FD methods of the backward Kolmogorov equation (7.2.1) on a Cartesian grid (7.2.5) yields approximations $\boldsymbol{U}_{j,k}(t)$ of the exact non-discounted option value $u(x_j, v_k, t)$. Denote by $\boldsymbol{U}(t)$ the $m_1 \times m_2$ matrix with entries $\boldsymbol{U}_{j,k}(t)$ and denote by $\boldsymbol{P}(\tau)$ a matrix with entries $\boldsymbol{P}_{j,k}(\tau)$ that represent approximations to the exact density values $p(x_j, v_k, \tau)$. In this section, for a general spatial discretization of the backward Kolmogorov equation, an adjoint spatial discretization of the corresponding forward equation is defined such that

$$\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} \boldsymbol{P}_{j,k}(T-t)\boldsymbol{U}_{j,k}(t)w_{x,j}w_{v,k} \tag{7.3.1}$$

is constant for $0 \leq t \leq T$. This can be viewed as a discrete version of relationship (7.2.2). Consistent *fully discrete* discretizations of the forward and backward Kolmogorov equation have already been considered in the literature, see e.g. [4, 45, 46, 49]. Here, we introduce an adjoint spatial discretization that is used in the subsequent sections for the fast, stable and accurate calibration of SLV models.

Let the vector

$$U(t) = \mathrm{vec}[\boldsymbol{U}(t)],$$

where we recall that vec[·] denotes the operator that turns any given matrix into a vector by putting its successive columns below each other. Spatial discretization of (7.2.1) leads to a large system of ODEs of the form

$$U'(t) = A^{(B)}(t)U(t), \tag{7.3.2}$$

for $0 < t \leq T$, with given matrix $A^{(B)}(t)$ and with given vector $U(0)$ that is defined by the payoff of the option. Let the vector

$$P(\tau) = \mathrm{vec}[\boldsymbol{P}(\tau)],$$

where the matrix $\boldsymbol{P}(0)$ is defined by the function p_0 from Section 7.2. Note that it has only one non-zero entry. Denote by M the diagonal matrix with diagonal entries

$$M_{\boldsymbol{l},\boldsymbol{l}} = w_{x,j} w_{v,k},$$

where j, k are the indices such that element $U_{\boldsymbol{l}}(t)$, respectively $P_{\boldsymbol{l}}(\tau)$, corresponds to $\boldsymbol{U}_{j,k}(t)$, respectively $\boldsymbol{P}_{j,k}(\tau)$. The semidiscrete analogue (7.3.1) of (7.2.6) can then be compactly written as

$$u(X_0, V_0, T) \approx P(T-t)^{\mathrm{T}} M U(t), \tag{7.3.3}$$

where $^{\mathrm{T}}$ denotes taking the transpose. If $\boldsymbol{l}_0$ is the index that corresponds to (j_0, k_0), then for $t = T$ the right-hand side of (7.3.3) equals $U_{\boldsymbol{l}_0}(T)$, and hence $\boldsymbol{U}_{j_0,k_0}(T)$. Now, consider the fair value of any given non-path-dependent European option with maturity T. It is readily seen that semidiscretization of the forward equation (7.2.4) and semidiscretization of the backward equation (7.2.1) define *the same approximation* (7.3.3) of the fair value for all $0 \leq t \leq T$, i.e. property (7.2.2) holds in the semidiscrete sense, if

$$U_{\boldsymbol{l}_0}(T) = P(0)^{\mathrm{T}} M U(T) = P(T-t)^{\mathrm{T}} M U(t) = P(\tau)^{\mathrm{T}} M U(T-\tau) \tag{7.3.4}$$

for all $0 \leq t, \tau \leq T$. This requirement is satisfied whenever

$$0 = P'(\tau)^{\mathrm{T}} M U(T-\tau) - P(\tau)^{\mathrm{T}} M A^{(B)}(T-\tau) U(T-\tau),$$

holds for all $0 \leq \tau \leq T$. Accordingly, we define the *adjoint spatial discretization* of the forward Kolmogorov equation (7.2.4) as

$$P'(\tau) = M^{-1}(A^{(B)}(T-\tau))^{\mathrm{T}} M P(\tau) \quad \text{for } 0 \leq \tau \leq T, \tag{7.3.5}$$

In this chapter we always employ the adjoint spatial discretization (7.3.5) of the forward equation. Thus, given any semidiscretization of the backward equation, the obtained approximated option values for non-path-dependent European options satisfy (7.3.4).

It is convenient to introduce

$$\overline{P}(\tau) = M P(\tau). \tag{7.3.6}$$

Denote by $\overline{\boldsymbol{P}}(\tau)$ the matrix corresponding to the vector $\overline{P}(\tau)$. The elements $\overline{\boldsymbol{P}}_{j,k}(\tau)$ can be viewed as approximations of

$$\int_{x_j - \frac{\Delta x_j}{2}}^{x_j + \frac{\Delta x_{j+1}}{2}} \int_{v_k - \frac{\Delta v_k}{2}}^{v_k + \frac{\Delta v_{k+1}}{2}} p(x, v, \tau) dx dv,$$

and hence, as an approximation of the probability that

$$(X_\tau, V_\tau) \in [x_j - \tfrac{\Delta x_j}{2}, x_j + \tfrac{\Delta x_{j+1}}{2}] \times [v_k - \tfrac{\Delta v_k}{2}, v_k + \tfrac{\Delta v_{k+1}}{2}].$$

Clearly,

$$\overline{P}'(\tau) = MM^{-1}(A^{(B)}(T-\tau))^{\mathrm{T}}MP(\tau) = (A^{(B)}(T-\tau))^{\mathrm{T}}\overline{P}(\tau),$$

for $0 \le \tau \le T$, with given vector $\overline{P}(0) = MP(0)$, i.e. with

$$\overline{P}_{\boldsymbol{l}}(0) = \begin{cases} 1 & \text{if } \boldsymbol{l} = \boldsymbol{l}_0, \\ 0 & \text{else.} \end{cases}$$

The approximation (7.3.3) of the non-discounted fair value of a non-path-dependent European option at the spot can then also be represented as

$$u(X_0, V_0, T) \approx U_{i_0}(T) = \overline{P}(T-t)^{\mathrm{T}}U(t).$$

7.4. Spatial Discretization by Finite Differences

In this section, a spatial discretization of the backward equation (7.2.1) by FD will be performed on a non-uniform Cartesian grid (7.2.5). This semidiscretization then defines the adjoint spatial discretization of the forward equation (7.2.4).

7.4.1. Spatial Discretization of the Backward Equation

To construct a spatial grid and a semidiscretization for (7.2.1), the spatial domain needs to be truncated to a bounded set $[x_{\min}, x_{\max}] \times [v_{\min}, v_{\max}]$. The boundaries have to lie sufficiently far away from (X_0, V_0) such that the truncation error incurred is negligible. If the parameter α is strictly positive, the process V_τ is non-negative and $v_{\min}$ is naturally set equal to zero. For non-path-dependent European options the following boundary conditions are imposed:

$$\begin{aligned} \tfrac{\partial^2}{\partial x^2}u(x_{\min}, v, t) &= \tfrac{\partial}{\partial x}u(x_{\min}, v, t) \quad \text{for } 0 \le v \le v_{\max},\ 0 < t \le T, \\ \tfrac{\partial^2}{\partial x^2}u(x_{\max}, v, t) &= \tfrac{\partial}{\partial x}u(x_{\max}, v, t) \quad \text{for } 0 \le v \le v_{\max},\ 0 < t \le T. \end{aligned} \tag{7.4.1}$$

The above conditions at $x = x_{\min}$ and $x = x_{\max}$ correspond to linear boundary conditions in the s-variable, where $s = S_0 \exp(x)$, cf. Section 7.1. If $\alpha = 0$, the process V_τ can take negative values and it is additionally assumed that

$$\begin{aligned} \tfrac{\partial^2}{\partial v^2}u(x, v_{\min}, t) &= 0 && \text{for } x_{\min} \le x \le x_{\max},\ 0 < t \le T, \\ \tfrac{\partial^2}{\partial v^2}u(x, v_{\max}, t) &= \tfrac{\partial}{\partial v}u(x, v_{\max}, t) && \text{for } x_{\min} \le x \le x_{\max},\ 0 < t \le T. \end{aligned}$$

Thus at $v = v_{\min}$ a linear boundary condition is taken. The boundary condition at $v = v_{\max}$ corresponds with a linear boundary condition in the variable $\exp(v)$. For values of α that are strictly positive, the process V_τ is non-negative. Moreover, $V_\tau = 0$ can be attained if $0 < \alpha \le 1/2$. In these cases the boundary $v = 0$ of the PDE requires special attention. It has been proved in [16] that setting $v = 0$ in the PDE (7.2.1) at this boundary then yields the correct condition here. At the boundary $v = v_{\max}$ it is then additionally assumed that

$$\tfrac{\partial^2}{\partial v^2}u(x, v_{\max}, t) = 0 \quad \text{for } x_{\min} \le x \le x_{\max},\ 0 < t \le T.$$

For this truncated domain, non-uniform meshes are applied in both the x- and v-direction such that relatively many mesh points lie in the neighbourhood of $x = X_0$ and $v = V_0$. The application of such non-uniform meshes improves the accuracy of the FD discretization compared to using uniform meshes. The type of non-uniform meshes that is employed is presented in Subsection 2.2.1. Recall that these meshes are smooth in the sense of (2.2.1). Further, the choice $m_1 = 2m_2$ is considered. As an illustration, the left plot in Figure 7.1 displays the spatial grid for the (small) sample values $m_1 = 30$, $m_2 = 15$, in the case $\alpha > 0$, $x_{\min} = -\log(30)$, $x_{\max} = \log(30)$, $v_{\min} = 0$, $v_{\max} = 15$ and $(X_0, V_0) = (0, 0.2)$. The right plot in Figure 7.1 displays a part of the spatial grid to show the local uniformity of the grid around (X_0, V_0).

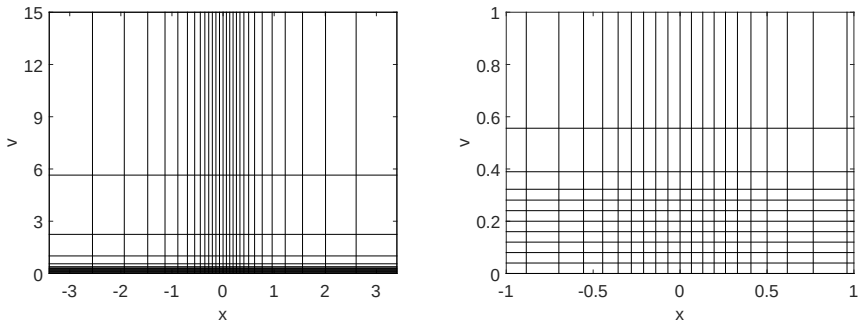


Figure 7.1: Sample grid for $m_1 = 30, m_2 = 15$, the case $\alpha > 0$, $x_{\min} = -\log(30)$, $x_{\max} = \log(30)$, $v_{\min} = 0$, $v_{\max} = 15$ and $(X_0, V_0) = (0, 0.2)$. The left plot displays the complete grid, the right plot shows the uniformity around (X_0, V_0).

Semidiscretization on the spatial grid is performed by finite differences as introduced in Section 2.3. Let $f : \mathbb{R} \to \mathbb{R}$ be any given smooth function. To approximate $f'(x_j)$ we consider the second order central formula (2.3.1b) and the second order forward formula (2.3.1c) for the first derivative. To approximate $f''(x_j)$, we apply the second order central formula (2.3.2) for the second derivative. The finite difference schemes in the v-direction are defined analogously. For a function of two variables $f : \mathbb{R}^2 \to \mathbb{R}$, the mixed derivative is approximated by application of (2.3.1b) successively in the two directions.

The actual semidiscretization of the backward PDE (7.2.1) is defined as follows. At all spatial grid points that do not lie on the boundary of the truncated domain, each spatial derivative appearing in (7.2.1) is replaced by its corresponding second order central finite difference scheme described above.

Concerning the boundaries in the x-direction, it is assumed that the pertinent conditions from (7.4.1) are valid for every x smaller than x_2 or larger than x_{m_1-1}. Thus for these extreme values x we assume that $u(\cdot, v, t)$ is an exponential function. For instance, considering the upper boundary

$$u_x = u_{xx} = C_{b,1} \exp(x) \quad \text{whenever } x > x_{m_1-1},$$

and hence

$$u = C_{b,1} \exp(x) + C_{b,2} \quad \text{whenever } x > x_{m_1-1},$$

for some constants $C_{b,1}$, $C_{b,2}$. Based on the function values of u at x_{m_1-1} and x_{m_1} the constants $C_{b,1}$, $C_{b,2}$ can be determined and this leads to the approximation for both the first and second derivatives in the x-direction at x_{m_1} given by

$$\begin{aligned} C_{b,1}\exp(x_{m_1}) = &\tfrac{-\exp(x_{m_1})}{\exp(x_{m_1})-\exp(x_{m_1-1})}u(x_{m_1-1},v,t) \\ &+\tfrac{\exp(x_{m_1})}{\exp(x_{m_1})-\exp(x_{m_1-1})}u(x_{m_1},v,t). \end{aligned}$$

The first and second derivatives in the x-direction at the lower boundary are approximated analogously.

If the parameter α is strictly positive, a linear boundary condition is applied at $v = v_{\max}$, i.e. the second derivative in the v-direction is equal to zero. The first derivative in the v-direction at this boundary is approximated by using the central scheme (2.3.1b) with the virtual point $v_{\max} + \Delta v_{m_2}$, where the value at this point is defined by extrapolation using the linear boundary condition. This discretization reduces to the first-order backward finite difference formula for the first derivative. Moreover, at the boundary $v_{\min} = 0$ all the second derivatives vanish and the first derivative $\partial u/\partial v$ is then approximated by using the forward scheme (2.3.1c). If α equals zero, a linear boundary condition is imposed at $v_{\min}$ and discretization of the spatial derivatives in the v-direction is performed as above. The discretization of the boundary condition at $v_{\max}$ is then performed analogously as the discretization of the boundary conditions in the x-direction.

Denote by D_x, respectively D_{xx}, the matrices corresponding to the first, respectively second, derivatives in the x-direction. Analogously, denote by D_v, D_{vv} the matrices corresponding to spatial derivatives in the v-direction. Denote by $L(\tau)$ the $m_1 \times m_1$ diagonal matrix with entries $\sigma_{\mathrm{SLV}}(x_j,\tau)$, let Λ be the $m_2 \times m_2$ diagonal matrix with entries v_k, and define for an arbitrary function $f : \mathbb{R} \to \mathbb{R}$ the matrix $f(\Lambda)$ as the diagonal matrix with entries $f(v_k)$. Further, denote by I_x, respectively I_v, the identity matrix of size $m_1 \times m_1$, respectively $m_2 \times m_2$. Then semidiscretization of (7.2.1) yields a system of differential equations given by

$$\begin{aligned} \boldsymbol{U}'(t) = &\tfrac{1}{2}L^2(T-t)(D_{xx}-D_x)\boldsymbol{U}(t)\psi^2(\Lambda) + \rho\xi L(T-t)D_x\boldsymbol{U}(t)D_v^{\mathrm{T}}\Lambda^\alpha\psi(\Lambda) \\ &+\tfrac{1}{2}\xi^2\boldsymbol{U}(t)D_{vv}^{\mathrm{T}}\Lambda^{2\alpha} + (r_d-r_f)D_x\boldsymbol{U}(t) + \boldsymbol{U}(t)D_v^{\mathrm{T}}\kappa(\eta I_v-\Lambda), \end{aligned}$$

for $0 < t \le T$. This can be written in the form (7.3.2),

$$U'(t) = A^{(B)}(t)U(t) = (A_0^{(B)}(t) + A_1^{(B)}(t) + A_2^{(B)}(t))U(t) \tag{7.4.2}$$

for $0 < t \le T$ where, using a well-known property of the Kronecker product,

$$\begin{aligned} A_0^{(B)}(t) &= (\rho\xi\psi(\Lambda)\Lambda^\alpha D_v)\otimes(L(T-t)D_x), \\ A_1^{(B)}(t) &= \tfrac{1}{2}\psi^2(\Lambda)\otimes(L^2(T-t)(D_{xx}-D_x)) + (r_d-r_f)I_v\otimes D_x \\ A_2^{(B)}(t) &= (\tfrac{1}{2}\xi^2\Lambda^{2\alpha}D_{vv} + \kappa(\eta I_v-\Lambda)D_v)\otimes I_x. \end{aligned}$$

The initial vector $U(0)$ is defined by the payoff of the option.

7.4.2. Spatial Discretization of the Forward Equation

As indicated in Section 7.3, semidiscretization of the forward equation (7.2.4) is performed by the adjoint spatial discretization (7.3.5). Since the transpose of the Kronecker product of two matrices is equal to the Kronecker product of the transposed matrices, and recalling that $t = T - \tau$, it follows that $\overline{P}(\tau)$ defined by (7.3.6) is given by the system of ODEs

$$\overline{P}'(\tau) = \overline{A}^{(F)}(\tau)\overline{P}(\tau) = (\overline{A}_0^{(F)}(\tau) + \overline{A}_1^{(F)}(\tau) + \overline{A}_2^{(F)}(\tau))\overline{P}(\tau), \tag{7.4.3}$$

for $\tau > 0$, with

$$\begin{aligned}
\overline{A}_0^{(F)}(\tau) &= (\rho\xi D_v^{\mathrm{T}}\Lambda^{\alpha}\psi(\Lambda)) \otimes (D_x^{\mathrm{T}}L(\tau)),\\
\overline{A}_1^{(F)}(\tau) &= \tfrac{1}{2}\psi^2(\Lambda) \otimes ((D_{xx}^{\mathrm{T}} - D_x^{\mathrm{T}})L^2(\tau)) + (r_d - r_f)I_v \otimes D_x^{\mathrm{T}},\\
\overline{A}_2^{(F)}(\tau) &= (\tfrac{1}{2}\xi^2 D_{vv}^{\mathrm{T}}\Lambda^{2\alpha} + D_v^{\mathrm{T}}\kappa(\eta I_v - \Lambda)) \otimes I_x,
\end{aligned}$$

and given initial vector $\overline{P}(0)$. This in turn corresponds to the system of differential equations

$$\begin{aligned}
\overline{\boldsymbol{P}}'(\tau) = {} & \tfrac{1}{2}(D_{xx}^{\mathrm{T}} - D_x^{\mathrm{T}})L^2(\tau)\overline{\boldsymbol{P}}(\tau)\psi^2(\Lambda) + \rho\xi D_x^{\mathrm{T}}L(\tau)\overline{\boldsymbol{P}}(\tau)\psi(\Lambda)\Lambda^{\alpha}D_v\\
& + \tfrac{1}{2}\xi^2\overline{\boldsymbol{P}}(\tau)\Lambda^{2\alpha}D_{vv} + (r_d - r_f)D_x^{\mathrm{T}}\overline{\boldsymbol{P}}(\tau) + \overline{\boldsymbol{P}}(\tau)\kappa(\eta I_v - \Lambda)D_v,
\end{aligned} \tag{7.4.4}$$

for $\tau > 0$. The expression (7.4.4) shall be employed to calibrate the SLV model to the LV model.

The total integral of a density function is equal to one. For a natural adjoint spatial discretization (7.4.3) one would expect that the total numerical integral of $\boldsymbol{P}$, corresponding to $\overline{\boldsymbol{P}}$, is close to one. Let e_x and e_v denote the vectors consisting of all ones with lengths m_1 and m_2, respectively. By construction of the finite difference discretization and the chosen boundary conditions for the SLV model (7.2.1) there holds

$$D_{xx}e_x = D_x e_x = 0 \quad \text{and} \quad D_{vv}e_v = D_v e_v = 0 \tag{7.4.5}$$

and it directly follows that

$$e_x^{\mathrm{T}}\,\overline{\boldsymbol{P}}'(\tau)e_v = 0$$

for all $\tau > 0$. Since further $e_x^{\mathrm{T}}\,\overline{\boldsymbol{P}}(0)e_v = 1$, this yields

$$\sum_{j=1}^{m_1}\sum_{k=1}^{m_2}\boldsymbol{P}_{j,k}(\tau)w_{x,j}w_{v,k} = \sum_{j=1}^{m_1}\sum_{k=1}^{m_2}\overline{\boldsymbol{P}}_{j,k}(\tau) = 1$$

for all $\tau \geq 0$. It can be concluded that the adjoint spatial discretization of the forward Kolmogorov equation keeps the total numerical integral of the density identically equal to one, which is a favourable property.

7.5. Matching the Semidiscrete LV and SLV Models

In this section the main result of the chapter is presented. It is shown that, under some assumptions, one can calibrate the semidiscrete SLV model exactly to the corresponding semidiscrete LV model.

A priori the leverage function σ_{SLV}, and hence the matrix function L, are unknown and one wishes to determine them in such a way that the LV model and the SLV model define identical values for European call and put options. In practice, however, even the LV function σ_{LV} is not known analytically in general and one relies on numerical methods to approximate the option values defined by the LV model. Accordingly, it is unrealistic to require an algorithm to produce a leverage function σ_{SLV} such that the SLV model yields the same exact European call and put values as the LV model. One rather wants to construct the leverage function in such a way that the two models yield *identical approximate values* for European call and put options whenever similar semidiscretizations of these models are used.

A common approach to approximate the fair value of a European-style option under the LV model is by discretizing the backward PDE (7.2.7) with finite differences. Since the region of interest in the x-direction in the LV model is the same as that in the SLV model, the same spatial mesh can be used in this spatial direction. Denote by $U_{\mathrm{LV}}(t)$ the vector with approximations $U_{\mathrm{LV},j}(t)$ to $u_{\mathrm{LV}}(x_j,t)$ for $1 \le j \le m_1$ such that the component $U_{\mathrm{LV},j_0}(T)$ is the approximation of the non-discounted fair value at the spot. Semidiscretization by FD then leads to a system of ODEs

$$U'_{\mathrm{LV}}(t) = A^{(B)}_{\mathrm{LV}}(t)U_{\mathrm{LV}}(t) \tag{7.5.1}$$

for $0 < t \le T$, with initial vector $U_{\mathrm{LV}}(0)$ defined by the payoff $u_{\mathrm{LV},0}$. Since the spatial derivatives $\partial/\partial x$ and $\partial^2/\partial x^2$ in (7.2.7) also occur in the backward equation (7.2.1), and since also the same boundary conditions from Section 7.4 can be applied for non-path-dependent European options, the same FD matrices D_x and D_{xx} can be used to perform semidiscretization, and hence

$$A^{(B)}_{\mathrm{LV}}(t) = \tfrac{1}{2}L^2_{\mathrm{LV}}(T-t)(D_{xx} - D_x) + (r_d - r_f)D_x,$$

where $L_{\mathrm{LV}}(\tau)$ is the $m_1 \times m_1$ diagonal matrix with entries $\sigma_{\mathrm{LV}}(x_j,\tau)$.

Denote by $P_{\mathrm{LV}}(\tau)$ a vector with approximations $P_{\mathrm{LV},j}(\tau)$ of $p_{\mathrm{LV}}(x_j,\tau)$, where p_{LV} is given by the forward equation (7.2.9), and let $\overline{P}_{\mathrm{LV}}(\tau)$ be defined by $M_{\mathrm{LV}}P_{\mathrm{LV}}(\tau)$ where M_{LV} is the diagonal matrix with entries

$$(M_{\mathrm{LV}})_{j,j} = w_{x,j}, \quad \text{for } 1 \le j \le m_1.$$

Analogously to Section 7.3, we define an adjoint forward discretization by

$$\overline{P}'_{\mathrm{LV}}(\tau) = (A^{(B)}_{\mathrm{LV}}(T-\tau))^{\mathrm{T}}\,\overline{P}_{\mathrm{LV}}(\tau) = \overline{A}^{(F)}_{\mathrm{LV}}(\tau)\overline{P}_{\mathrm{LV}}(\tau) \tag{7.5.2}$$

for $\tau > 0$, with

$$\overline{A}^{(F)}_{\mathrm{LV}}(\tau) = \tfrac{1}{2}(D^{\mathrm{T}}_{xx} - D^{\mathrm{T}}_x)L^2_{\mathrm{LV}}(\tau) + (r_d - r_f)D^{\mathrm{T}}_x,$$

and

$$\overline{P}_{\mathrm{LV},j}(0) = \begin{cases} 1 & \text{if } j = j_0, \\ 0 & \text{else,} \end{cases}$$

so that

$$U_{\mathrm{LV},j_0}(T) = \overline{P}_{\mathrm{LV}}(T-t)^{\mathrm{T}}U_{\mathrm{LV}}(t)$$

for all $0 \leq t \leq T$. Especially, for non-path-dependent European options one can just solve the forward problem and approximate the non-discounted fair value at the spot by $\overline{P}_{\mathrm{LV}}(T)^{\mathrm{T}} U_{\mathrm{LV}}(0)$.

Now, consider a non-path-dependent option whose payoff is only dependent on the exchange rate S_T. Then

$$U_{\mathrm{LV},j}(0) = \boldsymbol{U}_{j,k}(0)$$

whenever $1 \leq j \leq m_1, 1 \leq k \leq m_2$. It is readily verified that the semidiscretizations of the LV model and the SLV model define the same value at the spot, i.e. $U_{\mathrm{LV},j_0}(T) = \boldsymbol{U}_{j_0,k_0}(T)$, if

$$\overline{\boldsymbol{P}}(T) e_v = \overline{P}_{\mathrm{LV}}(T).$$

This property is desirable for every maturity. Hence, one would like to have

$$\overline{\boldsymbol{P}}(\tau) e_v = \overline{P}_{\mathrm{LV}}(\tau) \tag{7.5.3}$$

for all $\tau \geq 0$. Notice that (7.5.3) is equivalent to

$$\sum_{k=1}^{m_2} \boldsymbol{P}_{j,k}(\tau) w_{v,k} = P_{\mathrm{LV},j}(\tau) \quad \text{for } 1 \leq j \leq m_1,\ \tau \geq 0,$$

which can be viewed as a semidiscrete analogue of (7.2.10). Since the equality $\overline{\boldsymbol{P}}(0) e_v = \overline{P}_{\mathrm{LV}}(0)$ holds, the condition (7.5.3) is satisfied if

$$\overline{\boldsymbol{P}}'(\tau) e_v = \overline{P}'_{\mathrm{LV}}(\tau) \tag{7.5.4}$$

for all $\tau > 0$. From (7.4.4), (7.4.5) we directly obtain

$$\overline{\boldsymbol{P}}'(\tau) e_v = \tfrac{1}{2}(D_{xx}^{\mathrm{T}} - D_x^{\mathrm{T}}) L^2(\tau) \overline{\boldsymbol{P}}(\tau) \psi^2(\Lambda) e_v + (r_d - r_f) D_x^{\mathrm{T}}\, \overline{\boldsymbol{P}}(\tau) e_v.$$

If the (initially unspecified) diagonal matrix $L(\tau)$ is now defined through

$$L^2(\tau) \overline{\boldsymbol{P}}(\tau) \psi^2(\Lambda) e_v = L^2_{\mathrm{LV}}(\tau) \overline{\boldsymbol{P}}(\tau) e_v, \tag{7.5.5}$$

then

$$\overline{\boldsymbol{P}}'(\tau) e_v = \tfrac{1}{2}(D_{xx}^{\mathrm{T}} - D_x^{\mathrm{T}}) L^2_{\mathrm{LV}}(\tau) \overline{\boldsymbol{P}}(\tau) e_v + (r_d - r_f) D_x^{\mathrm{T}}\, \overline{\boldsymbol{P}}(\tau) e_v.$$

Hence, it follows that (7.5.4) holds whenever equation (7.5.2) has a unique solution.

Remark that in the definition (7.5.5) for the semidiscrete leverage function it is tacitly assumed that both vectors

$$\overline{\boldsymbol{P}}(\tau) \psi^2(\Lambda) e_v \quad \text{and} \quad \overline{\boldsymbol{P}}(\tau) e_v$$

only contain strictly positive values. By performing a spatial discretization with finite differences it is possible that some of the values $\overline{\boldsymbol{P}}_{j,k}$ become negative. In our experiments, both vectors often remained strictly positive for natural values of m_1, m_2.

Notice that *the derivation above is not restricted to the choice of finite difference formulas.* If the second order central formulas from Subsection 2.3.1 are replaced by alternative finite difference formulas for which (7.4.5) holds, and if these formulas are also applied for a similar semidiscretization in the LV model, then the SLV model is calibrated exactly to the LV model by employing (7.5.5).

We arrive at the following main result:

Theorem 7.5.1 *Assume semidiscretization of the backward Kolmogorov equation* (7.2.1) *is performed by consistent finite difference formulas on a Cartesian grid and that semidiscretization of the forward Kolmogorov equation* (7.2.4) *is performed by the adjoint spatial discretization. Then*

$$U_{\boldsymbol{l}_0}(T) = \overline{P}(\tau)^{\mathrm{T}} U(T-\tau) \quad \textit{for all } 0 \le \tau \le T,$$

where $\boldsymbol{l}_0$ *corresponds to the index* (j_0, k_0) *such that* $(x_{j_0}, v_{k_0}) = (X_0, V_0)$. *Hence, the two semidiscretizations define the same approximation for the fair value of non-path-dependent European options.*

Next, assume semidiscretization of the backward and forward equations (7.2.7) *and* (7.2.9) *under the LV model is performed in complete correspondence to that of* (7.2.1) *and* (7.2.4)*, respectively, under the SLV model by using the same grid and finite difference formulas in the* x*-direction. If* (7.4.5) *holds, if equation* (7.5.2) *has a unique solution and if the leverage function* σ_{SLV} *is defined on the grid in the* x*-direction by*

$$\sigma^2_{\mathrm{SLV}}(x_j, \tau) = \sigma^2_{\mathrm{LV}}(x_j, \tau) \frac{\sum_{k=1}^{m_2} \overline{\boldsymbol{P}}_{j,k}(\tau)}{\sum_{k=1}^{m_2} \psi^2(v_k) \overline{\boldsymbol{P}}_{j,k}(\tau)}, \tag{7.5.6}$$

then

$$\overline{\boldsymbol{P}}(\tau) e_v = \overline{P}_{\mathrm{LV}}(\tau) \quad \textit{for all } \tau \ge 0. \tag{7.5.7}$$

In particular, if the payoff depends only on the exchange rate S_T*, then the semidiscretizations of the LV model and the SLV model define the same approximation for the fair value of non-path-dependent European options:*

$$U_{\mathrm{LV},j_0}(T) = \overline{P}_{\mathrm{LV}}(\tau)^{\mathrm{T}} U_{\mathrm{LV}}(T-\tau) = \overline{P}(\tau)^{\mathrm{T}} U(T-\tau) = U_{\boldsymbol{l}_0}(T), \quad \text{for } 0 \le \tau \le T.$$

The second part of Theorem 7.5.1 can be regarded as the semidiscrete analogue of (7.1.4). Indeed, if σ_{SLV} is defined on the spatial grid in the x-direction by (7.5.6), then by the definition of $\overline{\boldsymbol{P}}$ it is directly seen that this is equivalent to applying (7.1.4), where the conditional expectation is approximated by

$$\mathbb{E}[\psi^2(V_\tau) | X_\tau = x_j] \approx \frac{\sum_{k=1}^{m_2} \psi^2(v_k) \boldsymbol{P}_{j,k}(\tau) w_{v,k}}{\sum_{k=1}^{m_2} \boldsymbol{P}_{j,k}(\tau) w_{v,k}} = \frac{\sum_{k=1}^{m_2} \psi^2(v_k) \overline{\boldsymbol{P}}_{j,k}(\tau)}{\sum_{k=1}^{m_2} \overline{\boldsymbol{P}}_{j,k}(\tau)}, \tag{7.5.8}$$

for $1 \le j \le m_1$.

7.6. Temporal Discretization

In this section we consider a suitable temporal discretization for the numerical solution of semidiscrete systems of the type (7.4.2) and (7.4.3) assuming that the matrix-valued function L is known. Since the pertinent systems of ODEs are stemming from spatial discretization of a multidimensional PDE with mixed spatial derivative term, the application of ADI time stepping schemes in the numerical solution can be very effective, see e.g. Chapter 3. Here, we opt to employ the MCS scheme (3.2.5) with parameter $\theta = 1/3$. A brief overview of the existing stability results for the MCS is presented in Section 3.3. In Chapter 4 it is shown that, under natural stability and smoothness assumptions, the MCS scheme is second order convergent with respect to the temporal step size whenever it is applied to semidiscrete two-dimensional convection-diffusion equations with mixed derivative term. In the present application, however, both vectors $U(0)$ and $\overline{P}(0)$ are stemming from initial functions that are non-smooth. Based on our analysis in Chapter 6, we replace the first two MCS time steps by four half-time steps of the implicit Euler scheme, i.e. we apply Rannacher time stepping with $N_0 = 2$.

As seen in Section 7.2, the fair values of non-path-dependent European options can be determined by solving either the backward equation (7.2.1) or the forward equation (7.2.4). In Section 7.4 spatial discretization of these two PDEs led to the semidiscrete systems (7.4.2) and (7.4.3), respectively. Denote by $N \geq 1$ again the total number of time steps and define uniform temporal grid points $t_n = n\Delta t$ with $\Delta t = T/N$. Application of the MCS scheme to (7.4.2) yields approximations U_n of $U(t_n)$ and the non-discounted fair option value at the spot is then approximated by $U_{\boldsymbol{l}_0,N} = \overline{P}_0^{\mathrm{T}} U_N$. Alternatively, take $\Delta\tau = \Delta t = T/N$ and let temporal grid points $\tau_n = n\Delta\tau = T - t_{N-n}$. Application of the MCS scheme to (7.4.3) yields approximations $\overline{P}_n$ of $\overline{P}(\tau_n)$ and the non-discounted fair option value at the spot is then approximated by $\overline{P}_N^{\mathrm{T}} U_0$.

It is possible, see Itkin [45, 46], to construct new ADI discretizations for (7.4.3) such that there is an exact match between the fully discretized backward equation and the fully discretized forward equation, that is,

$$\overline{P}_0^{\mathrm{T}} U_N = \overline{P}_N^{\mathrm{T}} U_0 = \overline{P}_{N-n}^{\mathrm{T}} U_n$$

for all $0 \leq n \leq N$. In this chapter we prefer to employ the MCS scheme for the numerical solution of both ODE systems (7.4.2) and (7.4.3). A main reason is that ample positive results are already available in the literature on the stability and convergence of the MCS scheme. In addition, practical experience shows that the temporal discretization error of the MCS scheme is often much smaller than the spatial discretization error.

7.7. Calibration of the SLV Model to the LV Model

In Section 7.5 we derived the expression (7.5.6) for the discrete leverage function σ_{SLV} that exactly calibrates the semidiscrete SLV model to the semidiscrete LV model. This expression involves the matrix function $\overline{\boldsymbol{P}}$. Combining (7.5.6) with

the semidiscrete forward equation (7.4.3) for $\overline{P} = \mathrm{vec}[\overline{\boldsymbol{P}}]$, one arrives at a large, non-linear system of ODEs. In the present section numerical time stepping is applied together with an inner iteration so as to numerically solve this system of ODEs and acquire the discrete leverage function that satisfies (7.5.6).

Suppose an approximation $\overline{P}_n$ to $\overline{P}(\tau_n)$ at time level τ_n is known. Let $\overline{\boldsymbol{P}}_n$ denote the $m_1 \times m_2$ matrix such that

$$\overline{P}_n = \mathrm{vec}[\overline{\boldsymbol{P}}_n].$$

Then the discrete leverage function is determined by

$$\sigma^2_{\mathrm{SLV}}(x_j, \tau_n)\mathbb{E}_{n,j} = \sigma^2_{\mathrm{LV}}(x_j, \tau_n), \tag{7.7.1}$$

where the quantity

$$\mathbb{E}_{n,j} = \frac{\sum_{k=1}^{m_2} \psi^2(v_k)\overline{\boldsymbol{P}}_{n,j,k}}{\sum_{k=1}^{m_2} \overline{\boldsymbol{P}}_{n,j,k}} \tag{7.7.2}$$

forms an approximation to the conditional expectation $\mathbb{E}[\psi^2(V_{\tau_n})|X_{\tau_n} = x_j]$, see (7.5.8). In order to arrive at the actual calibration procedure, we need to consider three practical issues concerning formula (7.7.2).

- Due to the spatial and temporal discretizations, it may happen that either the numerator or denominator of (7.7.2) becomes negative. In this case we assume that the conditional expectation is locally constant in time and set $\mathbb{E}_{n,j} = \mathbb{E}_{n-1,j}$. In most of our experiments, however, both parts of the quotient remained strictly positive for common values of $m_1, m_2, \Delta\tau$.
- Since $\overline{\boldsymbol{P}}_{n,j,k}$ can be viewed as an approximation of the probability of the event that

$$(X_{\tau_n}, V_{\tau_n}) \in [x_j - \tfrac{\Delta x_j}{2}, x_j + \tfrac{\Delta x_{j+1}}{2}] \times [v_k - \tfrac{\Delta v_k}{2}, v_k + \tfrac{\Delta v_{k+1}}{2}],$$

it can happen that both the numerator and denominator of (7.7.2) become very small, which can lead to unrealistic values of the leverage function. To resolve this, a regularized approximation of the conditional expectation is used (cf. [17]),

$$\mathbb{E}_{n,j} = \frac{\sum_{k=1}^{m_2} \psi^2(v_k)\overline{\boldsymbol{P}}_{n,j,k} + \psi^2(\eta)\epsilon}{\sum_{k=1}^{m_2} \overline{\boldsymbol{P}}_{n,j,k} + \epsilon} \tag{7.7.3}$$

for given small value ϵ. In this chapter $\epsilon = 10^{-8}$ is taken. By using the regularized version (7.7.3), the approximated conditional expectation is shifted towards $\psi^2(\eta)$ where η is the mean-reversion level of the process V_τ.

- At the spot $\tau = 0$, the matrix $\overline{\boldsymbol{P}}(0)$ has (j_0, k_0)-th entry equal to one and all its other entries are equal to zero. Consequently, if $n = 0$, then the expression (7.7.2) is only defined if $j = j_0$. To render the calibration procedure feasible, we extend this definition to all indices j and thus put

$$\mathbb{E}_{0,j} = \psi^2(V_0).$$

Notice that this agrees with (7.7.3) for $n = 0$ whenever $\eta = V_0$, which often holds in practice.

Let $Q \geq 1$ be a given integer. For calibrating the SLV model to the LV model, we employ the following numerical procedure. It consists of numerical

time stepping combined with an inner iteration, cf. [64].

for n is 1 to N do

  let $\overline{P}_n = \overline{P}_{n-1}$ be an initial approximation to $\overline{P}(\tau_n)$;

  for q is 1 to Q do

  (a) approximate $\mathbb{E}[\psi^2(V_{\tau_n})|X_{\tau_n} = x_j]$ by (7.7.3);

  (b) approximate $\sigma_{\mathrm{SLV}}(\cdot, \tau_n)$ on the grid in the x-direction by formula (7.7.1);

  (c) update $\overline{P}_n$ by performing a numerical time step for (7.4.3) from τ_{n-1} to τ_n;

  end

end

Whenever a time step from τ_{n-1} to τ_n with the MCS scheme is replaced by two half-time steps of the implicit Euler scheme, the inner iteration above is first performed for the substep from τ_{n-1} to $\tau_{n-1/2} = \tau_{n-1} + \Delta\tau/2$, yielding an approximation of $\overline{P}(\tau_{n-1/2})$ and $\sigma_{\mathrm{SLV}}(\cdot, \tau_{n-1/2})$. Next, the inner iteration is performed for the substep from $\tau_{n-1/2}$ to τ_n, yielding an approximation of $\overline{P}(\tau_n)$ and $\sigma_{\mathrm{SLV}}(\cdot, \tau_n)$.

Upon completion of the time stepping and iteration procedure above, the original approximation for $\sigma_{\mathrm{SLV}}(\cdot, 0)$ is replaced on the grid in the x-direction by

$$\sigma^2_{\mathrm{LV}}(x_j, 0) = \sigma^2_{\mathrm{SLV}}(x_j, 0)\frac{\sum_{k=1}^{m_2} \psi^2(v_k)\overline{\boldsymbol{P}}_{j,k,1} + \psi^2(\eta)\epsilon}{\sum_{k=1}^{m_2} \overline{\boldsymbol{P}}_{j,k,1} + \epsilon}$$

for $1 \leq j \leq m_1$. This appears more realistic as the original approximation was actually only valid for the index $j = j_0$.

7.8. Numerical Experiments

In this section, numerical experiments are presented to illustrate the effectiveness of the calibration procedure. Here, we opt to consider the popular and challenging Heston-based SLV model, i.e. SLV model (7.1.1) with $\psi(v) = \sqrt{v}$ and $\alpha = 1/2$, to describe the evolution of the EUR/USD exchange rate.

As stated in the introduction of the chapter, for the calibration of the SLV model it is customary to start from the parameters κ, η, ξ, ρ of the underlying (Heston) SV model such that this model reflects the market dynamics of the exchange rate. Denote $\mathfrak{f} = 2\kappa\eta/\xi^2 - 1$. It is readily seen that $\mathfrak{f} \geq -1$ and from Section 7.1 it follows that $V_\tau = 0$ is attainable if $\mathfrak{f} < 0$. In this chapter we consider the following four sets of parameters:

	κ	η	ξ	ρ	T	$\mathfrak{f}$
Set 1	3.02	0.015	0.31	-0.13	0.5	-0.0572
Set 2	0.30	0.04	0.90	-0.5	0.5	-0.9704
Set 3	0.75	0.015	0.15	-0.14	2	0
Set 4	0.30	0.04	0.90	-0.5	2	-0.9704

Table 7.1: Heston parameter sets for the numerical experiments with the adjoint calibration method.

The first and third parameter set are taken from [8, Table 6.5]. They correspond to the EUR/USD exchange rate for the pertinent maturities (market data as of 16 September 2008). The second and fourth parameter set are essentially the same and taken from [1, Case II]. The latter two parameter sets are challenging for the calibration procedure as the *Feller condition* [20] is strongly violated, i.e. $\mathfrak{f} \ll 0$, so that the probability mass is stacked up near $v = 0$, cf. [2]. Note that the spatial discretization of the backward Kolmogorov equation (7.2.1) from Subsection 7.4.1, and hence the total calibration procedure, is not dependent on the Feller condition.

The leverage function σ_{SLV} is determined in such a way that the SLV model is calibrated to the underlying LV model. The LV model is completely determined by the LV function σ_{LV} and the risk-free interest rates r_d, r_f. For the experiments we consider

$$r_d = 0.03, \quad r_f = 0.01,$$

and LV function displayed in Figure 7.2. This LV function originates from actual EUR/USD vanilla option data (market data as of 13 November 2015) and contains data up to two years. The corresponding spot rate is

$$S_0 = 1.0764.$$

For the spot value V_0 of the process V_τ in the SLV model we assume that it is equal to the long-term mean η of this process.

The aim is to construct the leverage function in such a way that the discretized LV model and the discretized SLV model yield identical approximate values for any given vanilla option whenever similar discretizations are employed. In our experiments, the semidiscretization of the backward Kolmogorov equation (7.2.1) is performed as described in Subsection 7.4.1 and the semidiscretization of the forward equation (7.2.4) is defined by the pertinent adjoint spatial discretization. The backward and forward PDEs (7.2.7) and (7.2.9) are semidiscretized analogously and by using the same finite difference schemes as described in Section 7.5. For the first numerical experiment we consider

$$m_1 = 100, \quad m_2 = 50.$$

The main Theorem 7.5.1 yields that if the leverage function is defined on the grid in the x-direction by (7.5.6), then the approximations obtained from the four semidiscrete systems (7.4.2), (7.4.3), (7.5.1), (7.5.2) of the fair value of any

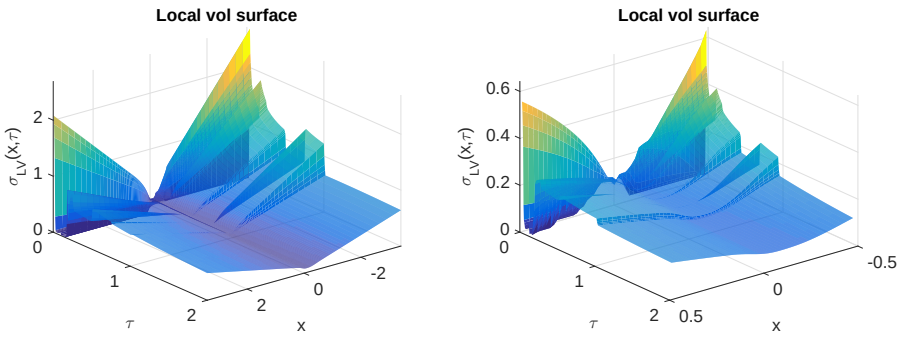


Figure 7.2: Local volatility function originating from actual EUR/USD vanilla option data (market data as of 13 November 2015) on the full domain in the x-direction (left) and on a subdomain around the spot rate (right). The spot rate $S_0 = 1.0764$.

given non-path-dependent European option are identical. The exact solution (7.5.6) is approximated by applying the calibration procedure described in Section 7.7. We choose to perform the temporal discretization in this procedure with values

$$\Delta\tau = 1/200, \quad \theta = 1/3, \quad Q = 2.$$

In Figure 7.3 the obtained discrete leverage function is shown for Set 4. If the SV model with parameters from Set 4 would fit the market prices for European call and put options exactly, then the leverage function would be identically equal to one and the SLV model reduces to the SV model. Clearly, Figure 7.3 indicates that the pure SV model with parameters from Set 4 does not match the market data very well. This outcome was to be expected, as the SV parameters stemming from [1] do not correspond to a EUR/USD exchange rate.

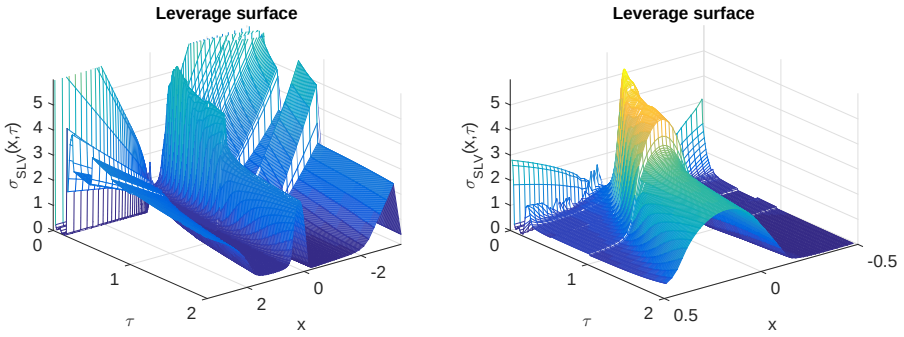


Figure 7.3: Leverage function on the full domain in the x-direction (left) and on a subdomain around the spot rate (right), stemming from the calibration procedure with local volatility function from Figure 7.2, SV parameters from Set 4 and with $m_1 = 100$, $m_2 = 50$, $\Delta\tau = 1/200$, $\theta = 1/3$, $Q = 2$.

With the obtained discrete leverage function, the performance of the calibration procedure is first tested by comparing the fully discrete approximations of the fair value of European call options which are acquired by numerically solving the systems of ODEs (7.4.2), (7.4.3), (7.5.1), (7.5.2). To this purpose we consider a range of strikes, given by

$$K = 0.7S_0,\ 0.8S_0,\ 0.9S_0,\ S_0,\ 1.1S_0,\ 1.2S_0,\ 1.3S_0.$$

Temporal discretization of (7.5.1) and (7.5.2) is performed by the classical *Crank–Nicolson scheme*. The systems (7.4.2) and (7.4.3), which are stemming from a two-dimensional PDE, are discretized in time by the MCS scheme (3.2.5) with parameter θ given above. For both methods we consider $\Delta\tau = 1/200$ and Rannacher time stepping is applied to handle the non-smoothness of the initial functions. In Table 7.2 the obtained fully discrete approximations FaV_i, $i \in \{\mathrm{LVB}, \mathrm{LVF}, \mathrm{SLVB}, \mathrm{SLVF}\}$, of the fair value (FaV) are presented for Set 1. Here $i = \mathrm{LVB}$, respectively $i = \mathrm{LVF}, \mathrm{SLVB}, \mathrm{SLVF}$, corresponds with the approximated fair value obtained via (7.5.1), respectively (7.5.2), (7.4.2), (7.4.3). In Table 7.2 one observes that the approximated option values are almost identical. To express in more detail the quality of the approximations, we present relative errors

$$\epsilon_{r,i} = (\mathrm{FaV}_i - \mathrm{FaV}_{\mathrm{LVB}})/\mathrm{FaV}_{\mathrm{LVB}}.$$

Here the option values given by solving (7.5.1), indicated by $\mathrm{FaV}_{\mathrm{LVB}}$, are considered as the reference values. This is motivated by the fact that in practice one starts from the underlying LV model and within the LV model it is common to solve the backward equation (7.2.7). Table 7.2 reveals the favourable result that all relative errors are smaller than 0.1%. Numerical experiments for the other SV parameter sets, i.e. for Sets $2, 3, 4$, yield the same observation. It can be concluded that the different approximations are almost identical in each of the four cases and the calibration procedure from Section 7.7 performs well.

K/S_0	$\mathrm{FaV}_{\mathrm{LVB}}$	$\mathrm{FaV}_{\mathrm{LVF}}$	$\epsilon_{r,\mathrm{LVF}}$	$\mathrm{FaV}_{\mathrm{SLVB}}$	$\epsilon_{r,\mathrm{SLVB}}$	$\mathrm{FaV}_{\mathrm{SLVF}}$	$\epsilon_{r,\mathrm{SLVF}}$
0.7	0.3288	0.3288	0.0000%	0.3288	0.0000%	0.3288	0.0000%
0.8	0.2228	0.2228	0.0000%	0.2228	0.0000%	0.2228	0.0001%
0.9	0.1185	0.1185	0.0008%	0.1185	0.0004%	0.1185	0.0008%
1	0.0381	0.0381	0.0083%	0.0381	0.0004%	0.0381	0.0079%
1.1	0.0091	0.0091	0.0235%	0.0091	0.0017%	0.0091	0.0211%
1.2	0.0019	0.0019	0.0437%	0.0019	0.0073%	0.0019	0.0391%
1.3	0.0004	0.0004	0.0661%	0.0004	0.0223%	0.0004	0.0640%

Table 7.2: Comparison of the approximated option values $\mathrm{FaV}_{\mathrm{LVB}}$, $\mathrm{FaV}_{\mathrm{LVF}}$, $\mathrm{FaV}_{\mathrm{SLVB}}$, $\mathrm{FaV}_{\mathrm{SLVF}}$ *for Set 1 and for values* $m_1 = 100$, $m_2 = 50$, $\Delta\tau = 1/200$, $\theta = 1/3$, $Q = 2$.

When the strike increases relative to S_0, the fair value of European call options tends to zero and it is difficult to adequately compare approximations. In financial practice, European call and put options are often quoted in terms of *implied volatility*. Let $\sigma_{\mathrm{imp},i}$ denote the implied volatility (in %) corresponding to FaV_i. Figure 7.4 shows the implied volatilities obtained from the local

volatility function from Figure 7.2 and $\mathrm{FaV}_{\mathrm{LVF}}$ with $m_1 = 100$, $\Delta t = 1/200$. The implied volatilities are shown approximately every week.

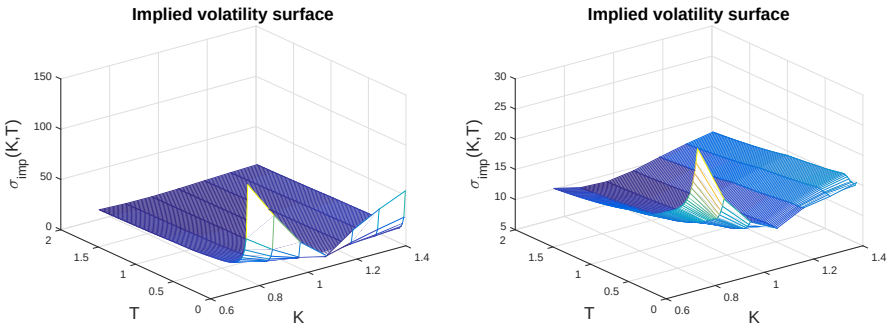


Figure 7.4: Implied volatilities obtained from the local volatility function from Figure 7.2 and $\mathrm{FaV}_{\mathrm{LVF}}$ *with* $m_1 = 100, \Delta t = 1/200$. *The implied volatilities are shown approximately every week from* $T = 1/200$ *(left) and from* $T = 1M$ *(right).*

In the following we test the performance of the calibration procedure by calculating the absolute implied volatility errors

$$\epsilon_i = |\sigma_{\mathrm{imp},i} - \sigma_{\mathrm{imp,LVB}}|.$$

In Table 7.3 these errors are presented for the different SV parameter sets, taking the same values of $m_1, m_2, \Delta\tau, \theta, Q$ as above. Since the adjoint spatial discretization is used for (7.2.9), the only source of error in ϵ_{LVF} is the temporal discretization error. Table 7.3 shows that these errors are small. The somewhat larger values ϵ_{LVF} for $T = 0.5$ compared to $T = 2$ can be explained from the fact that the implied volatility is more sensitive to changes in the fair value when the maturity is low. Table 7.3 subsequently reveals that for Set 1 and Set 3 the (small) errors ϵ_{SLVB} and ϵ_{SLVF} are of the same order of magnitude as ϵ_{LVF}. This indicates that the size of the error due to the calibration (with the iteration to handle the non-linearity) is not larger than the size of the temporal discretization error. As the calibration procedure includes numerical time stepping, this is the best result one can aim for. For Set 2 and Set 4 the magnitude of ϵ_{SLVB} is slightly higher than the one of ϵ_{LVF}. Additional experiments reveal that this is closely related to a strongly violated Feller condition ($\mathfrak{f} = -0.9704 \ll 0$). Probability mass is then stacked up near $v = 0$ and since our FD scheme is not positivity preserving, this sometimes leads to negative values for either the numerator or denominator of (7.7.2). In such situations, the calibration makes use of the conditional expectation of the previous time step, see Section 7.7, leading to a higher calibration error. Experiments with all parameter sets from [8, Table 6.5] indicate that the calibration error is of the same order of magnitude as the temporal discretization error whenever $\mathfrak{f} \geq -0.85$. The worst results that we encountered were for parameter set I from [1] where a strongly violated Feller condition ($\mathfrak{f} = -0.96$) is combined with a very high correlation ($\rho = -0.9$). Even for this extreme parameter set, however, the absolute implied volatility errors did not exceed 0.1.

$T = 0.5$			Set 1		Set 2	
K/S_0	$\sigma_{\text{imp,LVB}}$	ϵ_{LVF}	ϵ_{SLVB}	ϵ_{SLVF}	ϵ_{SLVB}	ϵ_{SLVF}
0.7	14.8646	0.0024	0.0024	0.0028	0.0008	0.0014
0.8	12.0458	0.0017	0.0015	0.0018	0.0004	0.0004
0.9	10.3061	0.0012	0.0006	0.0012	0.0057	0.0052
1	10.7970	0.0011	0.0000	0.0010	0.0006	0.0029
1.1	12.6393	0.0011	0.0001	0.0010	0.0176	0.0152
1.2	13.9656	0.0011	0.0002	0.0010	0.0260	0.0241
1.3	14.9511	0.0012	0.0004	0.0012	0.0094	0.0091

$T = 2$			Set 3		Set 4	
K/S_0	$\sigma_{\text{imp,LVB}}$	ϵ_{LVF}	ϵ_{SLVB}	ϵ_{SLVF}	ϵ_{SLVB}	ϵ_{SLVF}
0.7	10.2284	0.0008	0.0005	0.0008	0.0014	0.0012
0.8	9.1864	0.0005	0.0003	0.0005	0.0012	0.0011
0.9	8.9874	0.0004	0.0001	0.0004	0.0010	0.0011
1	9.6063	0.0004	0.0000	0.0004	0.0034	0.0036
1.1	10.6956	0.0004	0.0000	0.0004	0.0041	0.0044
1.2	11.6810	0.0004	0.0001	0.0004	0.0038	0.0040
1.3	12.4844	0.0004	0.0001	0.0004	0.0026	0.0026

Table 7.3: Comparison of the approximated implied volatilities $\sigma_{\text{imp,LVB}}$, $\sigma_{\text{imp,LVF}}$, $\sigma_{\text{imp,SLVB}}$, $\sigma_{\text{imp,SLVF}}$ *for values* $m_1 = 100, m_2 = 50, \Delta\tau = 1/200, \theta = 1/3, Q = 2$.

In order to verify the assertions above, we repeat the numerical experiments with a smaller step size. In Table 7.4 the absolute implied volatility errors are given for the same parameters as above but where the calibration and pricing is performed with

$$\Delta\tau = 1/400.$$

Comparing ϵ_{LVF} in Tables 7.3 and 7.4, it is clearly seen that the temporal error decreases if $\Delta\tau$ decreases. Moreover, for Set 1 and Set 3, the absolute implied volatility errors $\epsilon_{\text{SLVB}}, \epsilon_{\text{SLVF}}$ are again of the same size as ϵ_{LVF}. For Set 2 and Set 4 the errors do not decrease equally fast. This can be seen as a consequence of the fact that the spatial discretization has not been changed and, hence, negative numerators and denominators of (7.7.2) are still observed. From additional experiments it can be concluded that, if the Feller condition is not strongly violated ($\mathfrak{f} \geq -0.85$), then for realistic values of $\Delta\tau$ the error introduced by the calibration procedure is of the same order of magnitude as the temporal discretization error. When the Feller condition is strongly violated, the calibration error is slightly higher.

Observe that by decreasing the step size, the reference values $\sigma_{\text{imp,LVB}}$ have slightly changed. This is a consequence of the fact that more points from the LV surface are used and the fully discrete solution FaV_{LVB} converges to the semidiscrete solution $U_{\text{LV},\boldsymbol{\iota}_0}(T)$. We note that the approximated option values FaV_{LVB} acquired with $\Delta\tau = 1/400$ are identical to those in Table 7.2 up to the number of digits presented in that table. This confirms again that the temporal discretization error is small.

$T = 0.5$			Set 1		Set 2	
K/S_0	$\sigma_{\text{imp,LVB}}$	ϵ_{LVF}	ϵ_{SLVB}	ϵ_{SLVF}	ϵ_{SLVB}	ϵ_{SLVF}
0.7	14.8605	0.0001	0.0007	0.0006	0.0004	0.0004
0.8	12.0438	0.0000	0.0002	0.0001	0.0003	0.0005
0.9	10.3053	0.0000	0.0000	0.0000	0.0016	0.0015
1	10.7964	0.0000	0.0000	0.0000	0.0046	0.0050
1.1	12.6386	0.0000	0.0000	0.0000	0.0002	0.0002
1.2	13.9647	0.0000	0.0001	0.0000	0.0066	0.0059
1.3	14.9498	0.0000	0.0001	0.0000	0.0006	0.0001

$T = 2$			Set 3		Set 4	
K/S_0	$\sigma_{\text{imp,LVB}}$	ϵ_{LVF}	ϵ_{SLVB}	ϵ_{SLVF}	ϵ_{SLVB}	ϵ_{SLVF}
0.7	10.2278	0.0001	0.0001	0.0001	0.0003	0.0003
0.8	9.1859	0.0000	0.0000	0.0000	0.0000	0.0000
0.9	8.9870	0.0000	0.0000	0.0000	0.0010	0.0011
1	9.6060	0.0000	0.0000	0.0000	0.0022	0.0022
1.1	10.6953	0.0000	0.0000	0.0000	0.0027	0.0027
1.2	11.6806	0.0000	0.0000	0.0000	0.0029	0.0028
1.3	12.4840	0.0000	0.0000	0.0000	0.0026	0.0026

Table 7.4: Comparison of the approximated implied volatilities $\sigma_{\text{imp,LVB}}$, $\sigma_{\text{imp,LVF}}$, $\sigma_{\text{imp,SLVB}}$, $\sigma_{\text{imp,SLVF}}$ for values $m_1 = 100, m_2 = 50, \Delta\tau = 1/400, \theta = 1/3, Q = 2$.

As stated in Theorem 7.5.1, the condition (7.5.6) facilitates an exact match between the semidiscrete LV model and the semidiscrete SLV model whenever similar discretizations are used. This match is valid for any number of spatial grid points m_1, m_2. In order to test this property of the calibration procedure, we repeat the numerical experiments with the number of spatial grid points replaced by

$$m_1 = 200, \quad m_2 = 100,$$

and with step size $\Delta\tau = 1/200$. The obtained implied volatilities $\sigma_{\text{imp,LVB}}$ are presented in Table 7.5 as well as the absolute implied volatility errors $\epsilon_{\text{LVF}}, \epsilon_{\text{SLVB}}, \epsilon_{\text{SLVF}}$.

One observes that the size of the ϵ_{LVF} is similar in Tables 7.3 and 7.5. Hence, the experiments indicate that the performance of the calibration procedure is independent of the number of spatial grid points. The difference between the approximations of the fair value by discretizing either (7.5.1), (7.5.2), (7.4.2) or (7.4.3) is always small.

By increasing m_1, m_2 the values $\sigma_{\text{imp,LVB}}$ have noticeably changed, which is related to the convergence of $U_{\text{LV},\iota_0}(T)$ to the exact non-discounted fair value $u_{\text{LV}}(X_0, T)$. The differences in implied volatility in Tables 7.3, 7.4, 7.5 reveal that within the LV model and for the current, realistic values of m_1, m_2, τ the spatial discretization error is larger than the temporal discretization error (cf. Section 7.6). Since the calibration procedure from Section 7.7 matches the fully discrete LV and SLV models up to a difference with the size of the temporal discretization error, one can define an appropriate semidiscretization of the

$T = 0.5$			Set 1		Set 2	
K/S_0	$\sigma_{\text{imp,LVB}}$	ϵ_{LVF}	ϵ_{SLVB}	ϵ_{SLVF}	ϵ_{SLVB}	ϵ_{SLVF}
0.7	14.6017	0.0032	0.0039	0.0042	0.0028	0.0033
0.8	11.9199	0.0020	0.0019	0.0021	0.0006	0.0005
0.9	10.2664	0.0013	0.0006	0.0013	0.0052	0.0048
1	10.8100	0.0011	0.0001	0.0010	0.0022	0.0043
1.1	12.6442	0.0011	0.0001	0.0010	0.0134	0.0111
1.2	13.9412	0.0012	0.0002	0.0010	0.0294	0.0276
1.3	14.8890	0.0012	0.0004	0.0011	0.0102	0.0100

$T = 2$			Set 3		Set 4	
K/S_0	$\sigma_{\text{imp,LVB}}$	ϵ_{LVF}	ϵ_{SLVB}	ϵ_{SLVF}	ϵ_{SLVB}	ϵ_{SLVF}
0.7	10.1742	0.0009	0.0006	0.0010	0.0011	0.0009
0.8	9.1690	0.0006	0.0003	0.0006	0.0010	0.0009
0.9	8.9858	0.0004	0.0001	0.0004	0.0013	0.0014
1	9.6089	0.0004	0.0000	0.0004	0.0039	0.0041
1.1	10.6981	0.0004	0.0000	0.0004	0.0047	0.0050
1.2	11.6825	0.0004	0.0001	0.0004	0.0045	0.0047
1.3	12.4837	0.0004	0.0001	0.0004	0.0033	0.0034

Table 7.5: Comparison of the approximated implied volatilities $\sigma_{\text{imp,LVB}}$, $\sigma_{\text{imp,LVF}}$, $\sigma_{\text{imp,SLVB}}$, $\sigma_{\text{imp,SLVF}}$ for values $m_1 = 200$, $m_2 = 100$, $\Delta\tau = 1/200$, $\theta = 1/3$, $Q = 2$.

PDE (7.2.7), control the spatial discretization error within the LV model, and then calibrate the SLV model to the LV model such that the fully discrete SLV model matches the market data up to an error which is dominated by the controlled spatial error from semidiscretization within the LV model. If the Feller condition is strongly violated, one has to take into account a slightly higher calibration error.

We mention that all codes have been written in Matlab R2015a, where all matrices have been defined as sparse. The experiments have been performed on a computer with Intel Core i7-3540M 3.00GHz processor and 8GB memory. If the calibration is performed with $m_1 = 100$, $m_2 = 50$, $N = 100$ and $Q = 2$, then our implementation of the calibration procedure takes about 1 cpu-sec. Note that the computing time is (approximately) directly proportional to the number of spatial grid points $m = m_1 m_2$, the number of time steps N, and the number of iterations Q.

7.9. Conclusion

In financial practice, SLV models are calibrated to market data for European call and put options by calibrating them to their underlying LV models. Since there is often no closed-form analytical formula available for the fair value of vanilla options under an LV model, the best one can aim for is that the approximations of the fair value given by the two models are identical whenever similar numerical valuation methods are used. Here, we choose to perform the numer-

ical option valuation by semidiscretizing the respective backward Kolmogorov equations with finite differences. By making use of an adjoint semidiscretization of the corresponding forward Kolmogorov equations, we derived an expression for the leverage function such that the semidiscretized SLV model is calibrated exactly to the semidiscretized LV model. In order to employ this expression, one has to solve a large non-linear system of ODEs. For the actual numerical calibration, temporal discretization of this system by a suitable ADI method is combined with an inner iteration to deal with the non-linearity. Our numerical experiments reveal that the fully discrete approximations of the fair value of European call options under the LV and SLV models are the same up to the size of the temporal discretization error. Since the spatial discretization error is typically much larger than the temporal discretization error, one can control the former one by defining an appropriate semidiscretization of the LV model and then calibrate the fully discrete SLV model de facto exactly to the fully discrete LV model. Only if the Feller condition is strongly violated, one has to take into account a slightly higher calibration error.

CHAPTER 8

Finite Volume Calibration of SLV Models

8.1. Introduction

In financial practice, for the calibration of SLV models one will first determine the LV function $\sigma_{\mathrm{LV}}(x,\tau)$ such that LV model (7.1.2) yields the exact market prices for vanilla options, and the SV parameters such that the underlying SV model reflects the market dynamics of the underlying asset, see e.g. [8, 64]. Afterwards, given the SV parameters, the SLV model is calibrated to the LV model so that the former one also reproduces the known market prices for vanilla options. Let $S_\tau > 0$ denote the FX rate at time $\tau \geq 0$ and consider the standard transformed variable $X_\tau = \log(S_\tau/S_0)$. In this chapter we deal with SLV models of the type

$$\begin{cases} dX_\tau = (r_d - r_f - \frac{1}{2}\sigma^2_{\mathrm{SLV}}(X_\tau,\tau)\psi^2(V_\tau))d\tau + \sigma_{\mathrm{SLV}}(X_\tau,\tau)\psi(V_\tau)dW^{(1)}_\tau, \\ \\ dV_\tau = \kappa(\eta - V_\tau)d\tau + \xi V^\alpha_\tau dW^{(2)}_\tau, \end{cases} \tag{8.1.1}$$

with ψ a non-negative function on $\mathbb{R}^+$ such that $\psi(0) = 0$, $\alpha, \kappa, \eta, \xi$ strictly positive parameters, $dW^{(1)}_\tau \cdot dW^{(2)}_\tau = \rho d\tau$, $-1 \leq \rho \leq 1$ and given spot values $X_0 = 0, V_0$. These models are of the type (7.1.1), with the additional condition that α is strictly positive. Recall from Chapter 7 that the non-negative function $\sigma_{\mathrm{SLV}}(x,\tau)$ is called the leverage function and the constant r_d, respectively r_f, denotes the risk-free interest rate in the domestic currency, respectively in the foreign currency. In [2] it is shown that the process V_τ is always non-negative and that the boundary $V_\tau = 0$ is attainable for $0 < \alpha < 1/2$ and for $\alpha = 1/2$ if $2\kappa\eta < \xi^2$. For $\alpha > 1/2$ it holds that $V_\tau = 0$ is an unattainable boundary. Furthermore, $V_\tau = \infty$ is an unattainable boundary for all values of $\alpha > 0$. The choice $\psi(v) = \sqrt{v}, \alpha = 1/2$ corresponds to the Heston-based SLV model and the choice $\psi(v) = v, \alpha = 1$ corresponds to the SLV model described in [64].

Let $p(x, v, \tau; X_0, V_0)$ denote the joint density of (X_τ, V_τ) under the SLV model (8.1.1) and let $p_{\mathrm{LV}}(x,\tau; X_0)$ denote the density of $X_{\mathrm{LV},\tau}$ under (7.1.2). The SLV model and the underlying LV model define the same fair value for

This chapter is based on the article 'A finite volume - alternating direction implicit approach for the calibration of stochastic local volatility models', published in Int. J. Comput. Math., doi:10.1080/00207160.2017.1297805, 2017 [69].

vanilla options if they both yield the same marginal distribution for S_τ, i.e. if

$$p_{\mathrm{LV}}(x,\tau;X_0) = \int_0^\infty p(x,v,\tau;X_0,V_0)dv, \quad \text{for } \tau > 0. \tag{8.1.2}$$

From the previous chapter, see also [23,63], it follows that equality (8.1.2) holds if the leverage function is defined by

$$\sigma^2_{\mathrm{LV}}(x,\tau) = \mathbb{E}[\sigma^2_{\mathrm{SLV}}(X_\tau,\tau)\psi^2(V_\tau)|X_\tau = x] = \sigma^2_{\mathrm{SLV}}(x,\tau)\mathbb{E}[\psi^2(V_\tau)|X_\tau = x]. \tag{8.1.3}$$

Exact calibration of the SLV model can be performed by determining the conditional expectation above and defining the leverage function by (8.1.3). This is, however, highly non-trivial since the conditional expectation itself depends on the leverage function.

In the past years, a variety of numerical techniques, see e.g. [8,17,29,58,66] and Chapter 7, has been proposed in order to approximate the conditional expectation from (8.1.3) and to approximate the appropriate leverage function. The authors in [29,66] make use of Monte Carlo techniques to approximate the conditional expectation, whereas in [8,17,58] PDE methods are applied for the approximation of the density function $p(x,v,\tau;X_0,V_0)$. In Chapter 7 an adjoint PDE technique has been introduced for the calibration of SLV models. Note that, although the solution of e.g. (7.3.5) can be viewed as an approximation of the density function, there is no convergence result available for the adjoint spatial discretization when it is considered as a spatial discretization method for the forward Kolmogorov PDE (7.2.4).

In this chapter, we consider a new PDE method for the approximation of the underlying density function $p(x,v,\tau;X_0,V_0)$. For the effective calibration, the conditional expectation is often rewritten as, cf. [8,17,58],

$$\mathbb{E}[\psi^2(V_\tau)|X_\tau = x] = \frac{\int_0^\infty \psi^2(v)p(x,v,\tau;X_0,V_0)dv}{\int_0^\infty p(x,v,\tau;X_0,V_0)dv}. \tag{8.1.4}$$

Recall from the previous chapter, cf. also [59], that the joint density function satisfies the forward Kolmogorov equation

$$\begin{aligned}\tfrac{\partial}{\partial\tau}p = \tfrac{\partial^2}{\partial x^2}\left(\tfrac{1}{2}\sigma^2_{\mathrm{SLV}}\psi^2(v)p\right) + \tfrac{\partial^2}{\partial x\partial v}\left(\rho\xi\sigma_{\mathrm{SLV}}\psi(v)v^\alpha p\right) + \tfrac{\partial^2}{\partial v^2}\left(\tfrac{1}{2}\xi^2 v^{2\alpha}p\right)\\ - \tfrac{\partial}{\partial x}\left((r_d - r_f - \tfrac{1}{2}\sigma^2_{\mathrm{SLV}}\psi^2(v))p\right) - \tfrac{\partial}{\partial v}\left(\kappa(\eta - v)p\right),\end{aligned} \tag{8.1.5}$$

for $x \in \mathbb{R}, v > 0, \tau > 0$ and with initial condition

$$p(x,v,0;X_0,V_0) = \delta(x)\delta(v - V_0),$$

where δ denotes the Dirac delta function. Once the joint density p is known, one can easily determine the leverage function by computing the integrals in (8.1.4) and the SLV model is calibrated exactly to the LV model. By combining (8.1.3)–(8.1.5) it is readily seen that one has to solve a highly non-linear PDE in order to perform the calibration.

In financial mathematics, convection-diffusion equations of the type (8.1.5) are often discretized by means of FD methods, see e.g. [8,58]. If the parameter α

is less or equal to 1/2, however, it holds that $V_\tau = 0$ is attainable and defining a proper boundary condition at $v = 0$ is a non-trivial task. Moreover, FD methods are often not mass-conservative whereas conservation of mass is a key property of forward Kolmogorov equations. The finite volume (FV) method proposed in [17] manages to deal with the issues above. The latter method, however, makes use of a transformation of the original PDE (8.1.5) which incorporates derivatives of the leverage function. As the leverage function is often non-smooth and only known at a finite number of points, this could lead to undesirable (erratic) behaviour.

In this chapter we will introduce a FV-ADI discretization for the numerical solution of general, non-transformed forward Kolmogorov equations of the type (8.1.5). The discretization makes use of the general MOL, cf. Chapter 1. The PDE is first discretized in the spatial variables by *finite volume methods* to keep the total numerical mass equal to one and to handle the boundary conditions in a natural way. This yields large systems of stiff ODEs. These semidiscrete systems are subsequently solved by applying the *Hundsdorfer–Verwer scheme* (3.2.6). Since we consider two-dimensional PDEs, this can yield a large computational advantage in comparison with standard (non-split) implicit time stepping methods. Finally, for the calibration of the SLV model to the LV model, an inner iteration similar to that in Section 7.7 is introduced in order to handle the non-linearity from inserting (8.1.4) into (8.1.5).

The chapter is organised as follows. In Section 8.2 a FV discretization is introduced for the spatial discretization of general one-dimensional and two-dimensional forward Kolmogorov equations. The performance of the FV spatial discretization is illustrated by ample numerical experiments. Semidiscretization results in a large system of ODEs. In Section 8.3 the HV scheme is applied to increase the computational efficiency in the numerical solution of this ODE system. In Section 8.4 the FV discretization is used for the calibration of SLV models, yielding a large non-linear system of ODEs. The HV scheme is applied for the temporal discretization of this system of ODEs and an iteration procedure is described for handling the non-linearity. In Section 8.5 numerical experiments are presented to illustrate the performance of the obtained calibration procedure and the final Section 8.7 concludes.

8.2. FV Discretization of Forward Kolmogorov Equations

In the general MOL approach the PDE is first discretized in the spatial variables by for example FD or FV methods. In this section a spatial discretization is proposed for a general two-dimensional forward Kolmogorov equation of the type

$$\frac{\partial}{\partial\tau}p + \frac{\partial}{\partial x}\left(\mu_1 p\right) + \frac{\partial}{\partial y}\left(\mu_2 p\right) = \frac{\partial^2}{\partial x^2}\left(\tfrac{1}{2}\sigma_1^2 p\right) + \frac{\partial^2}{\partial x\partial y}\left(\rho\sigma_1\sigma_2 p\right) + \frac{\partial^2}{\partial y^2}\left(\tfrac{1}{2}\sigma_2^2 p\right), \quad (8.2.1)$$

with $x, y \in \mathbb{R}$, $\tau > 0$, and where $\sigma_1, \sigma_2, \mu_1, \mu_2$ are real coefficient functions of x, y, τ. Moreover, the functions σ_1, σ_2 are required to be non-negative and it is assumed that there exist values X_0, Y_0 such that the initial function is given by $p(x, y, 0) = \delta(x - X_0)\delta(y - Y_0)$. Due to the form of the coefficients it is possible that the spatial domain is naturally restricted. For example, if

$\mu_2(x,y,\tau) = \kappa(\eta - y)$ and $\sigma_2(x,y,\tau) = \xi y^\alpha$ with $\kappa, \eta, \xi, \alpha$ strictly positive constants, then the domain in the y-direction is naturally restricted to $y \geq 0$, cf. PDE (8.1.5).

Since the solution of forward Kolmogorov equations represents the density of an underlying stochastic process, conservation of mass is a fundamental property and the use of FV schemes is appropriate. While FD methods are well-known in finance, FV methods are less common in financial applications and we briefly recall the basic idea.

8.2.1. Introduction to Finite Volume Discretizations

Finite volume methods were originally developed to solve conservation laws, or more generally to solve PDEs in conservative form. For example, consider the one-dimensional conservative PDE

$$\tfrac{\partial}{\partial \tau} p + \tfrac{\partial}{\partial x}\left(a(p,x,\tau)p\right) = \tfrac{\partial}{\partial x}\left(b(p,x,\tau)\tfrac{\partial}{\partial x}p\right), \tag{8.2.2}$$

for $x \in \Omega, \tau > 0$, where Ω is an interval in $\mathbb{R}$. Both sides of equation (8.2.2) can be integrated in x over an interval (more generally, a cell) $[x_l, x_u]$ in order to get

$$\tfrac{\partial}{\partial \tau}\int_{x_l}^{x_u} p dx = \boldsymbol{f}(p, x_l, \tau) - \boldsymbol{f}(p, x_u, \tau), \tag{8.2.3}$$

where $\boldsymbol{f}(p,x,\tau)$ is a function given by

$$\boldsymbol{f}(p,x,\tau) = a(p,x,\tau)p - b(p,x,\tau)\tfrac{\partial}{\partial x}p.$$

The function $\boldsymbol{f}$ is typically called the *flux* of p and $\boldsymbol{f}(p,x,\tau)|_{x=x_l}$, respectively $\boldsymbol{f}(p,x,\tau)|_{x=x_u}$, represents the flux at the left, respectively right, boundary of the cell $[x_l, x_u]$. Relationship (8.2.3) shows that the total integral of p, which typically represents a mass, momentum or some similar quantity, changes only as a result of the flux difference over the cell. If equation (8.2.2) is considered over a bounded domain $[x_{\min}, x_{\max}]$ and we assume that $\boldsymbol{f}(p, x_{\min}, \tau) = \boldsymbol{f}(p, x_{\max}, \tau)$ for all τ, i.e. the flux at the left boundary is exactly matched by the flux at the right boundary, then the space integral of p over $[x_{\min}, x_{\max}]$ is constant in time. This means that the total mass or momentum is conserved. If the spatial domain Ω of the PDE is unbounded, and if the interval $[x_{\min}, x_{\max}]$ is wide enough, then one will often have that $\boldsymbol{f}(p, x_{\min}, \tau) \approx \boldsymbol{f}(p, x_{\max}, \tau) \approx 0$ for all $\tau > 0$.

To construct a numerical FV scheme we start with a discretization of the spatial domain. If the spatial domain is unbounded, it needs to be truncated to a wide, finite interval $[x_{\min}, x_{\max}]$. Then, consider the discretization

$$x_{\min} = x_1 < x_2 < \cdots < x_m = x_{\max},$$

of the domain of interest and denote

$$\Delta x_j = x_j - x_{j-1}, \quad \text{for } 2 \leq j \leq m,$$

and $\Delta x_1 = \Delta x_{m+1} = 0$. Define mid-points

$$x_{j-0.5} = x_j - \tfrac{1}{2}\Delta x_j = \tfrac{x_{j-1}+x_j}{2}, \quad \text{for } 2 \leq j \leq m,$$

and let $\Omega_j = [x_{j-0.5}, x_{j+0.5}]$ be cells for $j = 1, 2, \ldots, m$ where we additionally define $x_{0.5} := x_1$ and $x_{m+0.5} := x_m$. This yields a *vertex centred grid* with cell vertices $x_{j-0.5}$. We can now consider the cell average $\overline{p}_j(\tau)$ which is defined by

$$\overline{p}_j(\tau) = \frac{1}{x_{j+0.5} - x_{j-0.5}} \int_{\Omega_j} p(x, \tau) dx,$$

and which is typically the quantity that FV schemes approximate. If we assume that the grid is smooth in the sense of (2.2.1), then the cell average $\overline{p}_j(\tau)$ is a second order approximation to $p(x_j, \tau)$. Differentiating $\overline{p}_j(\tau)$ in τ and using (8.2.3) gives

$$\overline{p}'_j(\tau) = \frac{\boldsymbol{f}(p, x_{j-0.5}, \tau) - \boldsymbol{f}(p, x_{j+0.5}, \tau)}{x_{j+0.5} - x_{j-0.5}} \tag{8.2.4}$$

which is just another way of stating the conservation property since if we sum over all cells and pull out the derivative in time we find that

$$\tfrac{\partial}{\partial \tau} \sum_{j=1}^{m} \overline{p}_j(\tau)(x_{j+0.5} - x_{j-0.5}) = 0,$$

provided $\boldsymbol{f}(p, x_{\min}, \tau) = \boldsymbol{f}(p, x_{\max}, \tau)$.

Equation (8.2.4) is typically taken as the starting point for the numerical discretization. Denote by

$$P_j(\tau) \approx \overline{p}_j(\tau), \quad \text{for } 1 \leq j \leq m,$$

the numerical approximations for the cell averages and let P be the vector that contains these approximations. The numerical discretization is then defined by

$$P'_j(\tau) = \tfrac{\boldsymbol{f}_{j-0.5}(P,\tau) - \boldsymbol{f}_{j+0.5}(P,\tau)}{x_{j+0.5} - x_{j-0.5}} = [\boldsymbol{f}_{j-0.5}(P, \tau) - \boldsymbol{f}_{j+0.5}(P, \tau)] \tfrac{2}{\Delta x_j + \Delta x_{j+1}}, \tag{8.2.5}$$

where the $\boldsymbol{f}_{j\pm0.5}$ are numerical fluxes that form approximations to the exact fluxes $\boldsymbol{f}(p, x_{j\pm0.5}, \tau)$. By defining a discretization of the type (8.2.5), it readily follows that the total numerical integral (mass)

$$\sum_{j=1}^{m_1} P_j(\tau)(x_{j+0.5} - x_{j-0.5}) \tag{8.2.6}$$

stays constant in time provided that $\boldsymbol{f}_{0.5}(P, \tau) = \boldsymbol{f}_{m+0.5}(P, \tau)$. It is clear that the exact fluxes from (8.2.4) involve the unknown function p at the cell boundaries $x_{i\pm0.5}$. Therefore, we define the numerical fluxes by

$$\boldsymbol{f}_{j\pm0.5}(P, \tau) = a(P_{j\pm0.5}, x_{j\pm0.5}, \tau) P_{j\pm0.5} - b(P_{j\pm0.5}, x_{j\pm0.5}, \tau) P_{x,j\pm0.5}, \tag{8.2.7}$$

where the $P_{j\pm0.5}(\tau)$ form approximations to the exact values $p(x_{j\pm0.5}, \tau)$ and the $P_{x,j\pm0.5}(\tau)$ form approximations to $\frac{\partial}{\partial x} p(x, \tau)|_{x=x_{j\pm0.5}}$. Since the cell average is a second order approximation to p, the approximations P_j can be used to define $P_{j\pm0.5}$ and $P_{x,j\pm0.5}$. The manner in which the latter values are computed at the cell boundaries from the surrounding P_j plays a large part in

defining the characteristics of the numerical scheme. Lastly, inserting (8.2.7) into (8.2.5) yields a system of (possibly non-linear) ODEs which is solved with a suitable time integration procedure.

Recall that conservation of mass is a fundamental property of forward Kolmogorov equations and the use of FV schemes is appropriate. Forward equations of the type (8.2.1) are, however, not in conservative form and hence straightforward application of standard FV schemes is not possible. Moreover, rewriting PDE (8.2.1) in conservative form would involve derivatives of the coefficient functions, which are not known in general practical applications. In the remainder of this section, a FV-based discretization of the spatial derivatives in the non-transformed PDE (8.2.1) is introduced such that conservation of total mass is guaranteed. We start by explaining the discretization for a general one-dimensional forward Kolmogorov equation, and then generalise it to the two-dimensional case.

8.2.2. One-Dimensional Forward Kolmogorov Equations

Standard one-dimensional forward Kolmogorov equations are also not written in conservative form and their solutions represent density functions of underlying stochastic processes. In this subsection a FV-based discretization is introduced for the general one-dimensional equation

$$\frac{\partial}{\partial \tau} p + \frac{\partial}{\partial x}(\mu p) = \frac{\partial^2}{\partial x^2}\left(\tfrac{1}{2}\sigma^2 p\right), \tag{8.2.8}$$

for $x \in \mathbb{R}$, $\tau > 0$, where σ, μ are real functions of x and τ, with σ non-negative and with initial function given by $p(x,0) = \delta(x - X_0)$ for some real X_0. Spatial discretization by FD or FV methods is often applied on a finite grid. By consequence, the spatial domain has to be truncated to $[x_{\min}, x_{\max}]$, where the boundaries are chosen sufficiently far away from X_0 such that the truncation error is negligible. Recall that the form of σ, μ can naturally restrict the spatial domain of the PDE to for example $x \geq 0$. In the latter case, the lower boundary is naturally defined as $x_{\min} = 0$.

As before, define a spatial mesh $x_{\min} = x_1 < x_2 < \ldots < x_m = x_{\max}$, let $\Delta x_j = x_j - x_{j-1}$ be the mesh widths, with $\Delta x_1 = \Delta x_{m+1} = 0$, and define

$$x_{j-0.5} = x_j - \tfrac{1}{2}\Delta x_j = \tfrac{x_{j-1}+x_j}{2} \quad \text{for } 2 \leq j \leq m,$$

with $x_{0.5} = x_1$ and $x_{m+0.5} = x_m$. This yields a vertex centred grid with cells $\Omega_j = [x_{j-0.5}, x_{j+0.5}]$. Let the $P_j(\tau)$ denote approximations to the exact cell averages

$$\overline{p}_j(\tau) = \frac{1}{x_{j+0.5} - x_{j-0.5}} \int_{\Omega_j} p(x,\tau)dx,$$

and let P be the vector containing these approximations. Analogously to the previous section (see equations (8.2.4) and (8.2.5), as well as [35]) we define discretizations of the form

$$P_j'(\tau) = [\boldsymbol{f}_{j-0.5}(P,\tau) - \boldsymbol{f}_{j+0.5}(P,\tau)]\tfrac{2}{\Delta x_j + \Delta x_{j+1}} \tag{8.2.9}$$

where the numerical fluxes are given by

$$\boldsymbol{f}_{j\pm0.5}(P,\tau) = \boldsymbol{f}_{a,j\pm0.5}(P,\tau) + \boldsymbol{f}_{d,j\pm0.5}(P,\tau),$$

with

$$\boldsymbol{f}_{a,j\pm0.5}(P,\tau) \approx \mu(x_{j\pm0.5},\tau)p(x_{j\pm0.5},\tau), \tag{8.2.10}$$

and

$$\boldsymbol{f}_{d,j\pm0.5}(P,\tau) \approx -\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma^2(x,\tau)p(x,\tau)\right)|_{x=x_{j\pm0.5}}. \tag{8.2.11}$$

For the ease of presentation, from now on we omit the dependence of the parameters on τ and set $\mu_{j\pm0.5} = \mu(x_{j\pm0.5},\tau)$ and $\sigma_j = \sigma(x_j,\tau)$. Note that $\boldsymbol{f}_{0,5}(P,\tau)$, respectively $\boldsymbol{f}_{m+0.5}(P,\tau)$, corresponds with the flux at the boundary $x_{\min} = x_1$, respectively $x_{\max} = x_m$.

The convection part of the PDE (8.2.8) is written in conservative form. For the inner cell boundaries, i.e. for $x_{j-0.5}$ with $2 \le j \le m$, we consider the second order central FV scheme, cf. [35], and define $\boldsymbol{f}_{a,j-0.5}(P,\tau)$ in (8.2.10) as

$$\boldsymbol{f}_{a,j-0.5}(P,\tau) = \mu_{j-0.5}\tfrac{P_{j-1}(\tau)+P_j(\tau)}{2}.$$

The diffusion part is not written in conservative form and hence it is not possible to apply standard FV schemes to this term directly. The idea of the second order FV scheme for (8.2.11), see e.g. [35], is generalised by defining

$$\boldsymbol{f}_{d,j-0.5}(P,\tau) = -\left(\tfrac{1}{2}\sigma_j^2 P_j(\tau) - \tfrac{1}{2}\sigma_{j-1}^2 P_{j-1}(\tau)\right)\tfrac{1}{\Delta x_j}$$

for $2 \le j \le m$. It is readily seen that

$$\left(\tfrac{1}{2}\sigma_j^2 p(x_j,\tau) - \tfrac{1}{2}\sigma_{j-1}^2 p(x_{j-1},\tau)\right)\tfrac{1}{\Delta x_j}$$

is a second order approximation of $\frac{\partial}{\partial x}(\frac{1}{2}\sigma^2 p)$ at the point $x_{j-0.5}$ which explains the choice for this discretization. Inserting these expressions back into (8.2.9) we get

$$\begin{aligned} P_j'(\tau) = {} & \tfrac{\sigma_{j-1}^2 P_{j-1}(\tau)}{\Delta x_j(\Delta x_j+\Delta x_{j+1})} - \tfrac{\sigma_j^2 P_j(\tau)}{\Delta x_j \Delta x_{j+1}} + \tfrac{\sigma_{j+1}^2 P_{j+1}(\tau)}{\Delta x_{j+1}(\Delta x_j+\Delta x_{j+1})} \\ & + \left[\mu_{j-0.5}\tfrac{P_{j-1}(\tau)+P_j(\tau)}{2} - \mu_{j+0.5}\tfrac{P_j(\tau)+P_{j+1}(\tau)}{2}\right]\tfrac{2}{\Delta x_j+\Delta x_{j+1}}, \end{aligned} \tag{8.2.12}$$

for $2 \le j \le m-1$. Note that by applying the second order central FD scheme for diffusion on non-uniform spatial grids, cf. Chapter 2, on the term $\frac{\partial^2}{\partial x^2}(\frac{1}{2}\sigma^2 p)$, one would end up with the same discretization for the diffusion term.

To complete this semidiscretization, it also has to be defined at the boundaries of the truncated domain such that conservation of the total mass is guaranteed. Given that p represents a density function, it follows that

$$\int_{-\infty}^{\infty} p(x,\tau)dx = 1, \quad \forall\tau > 0,$$

and hence

$$\int_{-\infty}^{\infty}\left[\tfrac{\partial}{\partial\tau}p\right]dx = \int_{-\infty}^{\infty}\left[\tfrac{1}{2}\tfrac{\partial^2}{\partial x^2}\left(\sigma^2 p\right) - \tfrac{\partial}{\partial x}\left(\mu p\right)\right]dx = 0.$$

Assuming that $[x_{\min}, x_{\max}]$ is chosen sufficiently wide, the condition above can be approximated by

$$\left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma^2 p\right) - (\mu p)\right]\Big|_{x=x_{\min}}^{x=x_{\max}} = 0,$$

reflecting the fact that the total flux over the interval $[x_{\min}, x_{\max}]$ is zero. As stated above, for some choices of coefficient functions the spatial domain is naturally restricted. For example, if the PDE (8.2.8) stems from a non-negative process, $x_{\min}$ can be set equal to zero. The no-flux boundary condition above then still holds on the naturally restricted domain.

In theory it is possible that there is a positive flux at one of the boundaries, and exactly the same negative flux at the other boundary. However, since the solution represents a density function, it is more realistic to impose that no mass is coming in or going out at each of the boundaries. In light of this, we assume that the following boundary conditions hold:

$$\begin{aligned}\left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma^2 p\right) - (\mu p)\right]\Big|_{x=x_{\min}} &= 0,\\ \left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma^2 p\right) - (\mu p)\right]\Big|_{x=x_{\max}} &= 0.\end{aligned} \tag{8.2.13}$$

The numerical equivalent of the first condition is to say that the flux at the left boundary $x_1 = x_{\min}$ is zero, i.e. $\boldsymbol{f}_1(P,\tau) \equiv \boldsymbol{f}_{0.5}(P,\tau) = 0$. This can be achieved by creating a ghost point $x_0 = x_1 - \Delta x_2$ and using (8.2.13) to define the value of $\frac{1}{2}\sigma_0^2 P_0(\tau)$ at the ghost point by

$$\frac{\frac{1}{2}\sigma_2^2 P_2(\tau) - \frac{1}{2}\sigma_0^2 P_0(\tau)}{2\Delta x_2} - \mu_1 P_1(\tau) = 0,$$

where we make use of the fact that the cell averages form second order approximations to the point values. Turning to (8.2.11) we define the diffusive flux $\boldsymbol{f}_{d,0.5}$ at x_1 as

$$\boldsymbol{f}_{d,0.5}(P,\tau) = -\frac{\frac{1}{2}\sigma_2^2 P_2(\tau) - \frac{1}{2}\sigma_0^2 P_0(\tau)}{2\Delta x_2} = -\mu_1 P_1(\tau).$$

Since x_1 is the left boundary of the first cell, the flux on the boundary $x_{\min}$ stemming from the convection part, see (8.2.10), can be approximated by the term $\boldsymbol{f}_{a,0.5} = \mu_1 P_1(\tau)$. Inserting these expressions into (8.2.9) we obtain

$$P_1'(\tau) = -\boldsymbol{f}_{1.5}(P,\tau)\tfrac{2}{\Delta x_1 + \Delta x_2} = -\boldsymbol{f}_{1.5}(P,\tau)\tfrac{2}{\Delta x_2}. \tag{8.2.14}$$

The boundary condition at $x_{\max}$ can be handled analogously in order to get

$$P_m'(\tau) = \boldsymbol{f}_{m-0.5}(P,\tau)\tfrac{2}{\Delta x_m}. \tag{8.2.15}$$

By performing the discretization of the boundary conditions in this way, it follows that $\boldsymbol{f}_{0.5}(P,\tau) = \boldsymbol{f}_{m+0.5}(P,\tau)$ and we ensure that mass is conserved in the numerical scheme.

Combining (8.2.12), (8.2.14) and (8.2.15) we see that the total discretization can be written as a system of ODEs

$$P'(\tau) = A(\tau)P(\tau) \tag{8.2.16}$$

for $\tau > 0$, with given matrix $A(\tau)$. Since the $P_j(\tau)$ represent cell averages it is natural to define the initial vector as

$$P_j(0) = \begin{cases} \frac{2}{\Delta x_j + \Delta x_{j+1}} & \text{if } X_0 \in [x_{j-0.5},\ x_{j+0.5}], \\ 0 & \text{otherwise.} \end{cases}$$

In general, the exact solution of the system of ODEs (8.2.16) can not be computed analytically and one relies on numerical methods in order to approximate it. Since the discretization above often leads to stiff semidiscrete systems, suitable implicit time stepping schemes such as the Crank–Nicolson scheme are widely considered, see e.g. [65].

8.2.3. Numerical Experiments for One-Dimensional Forward Kolmogorov Equations

In this subsection the performance of the FV discretization is tested by considering two practical examples. As a first example, consider the SDE

$$dS_\tau = (r_d - r_f)S_\tau d\tau + \sigma_{BS} S_\tau dW_\tau,$$

with $\sigma_{BS} > 0$, corresponding to the classical Black–Scholes model [6]. Then, the underlying density is known exactly and given by

$$p(s,\tau) = \frac{1}{\sigma_{BS}\sqrt{\tau}} \phi\left(\frac{\log(s/S_0) - (r_d - r_f - \frac{1}{2}\sigma_{BS}^2)\tau}{\sigma_{BS}\sqrt{\tau}} \right) \frac{1}{s}, \quad \text{for } s > 0, \tau > 0, \tag{8.2.17}$$

where $\phi(x)$ is the density function of a standard normally distributed random variable. The density function $p(s,\tau)$ from (8.2.17) satisfies the PDE

$$\tfrac{\partial}{\partial \tau} p = \tfrac{\partial^2}{\partial s^2}\left(\tfrac{1}{2}\sigma_{BS}^2 s^2 p\right) - \tfrac{\partial}{\partial s}\left((r_d - r_f) s p\right),$$

for $s, \tau > 0$, with $p(s,0) = \delta(s - S_0)$. This PDE is of the form (8.2.8) with a natural restriction of the spatial domain. Note that for the numerical experiment we don't apply the log-transformation from Section 8.1 in order to have non-constant coefficients which makes the problem more challenging.

Firstly, the spatial domain is truncated to $[s_{\min}, s_{\max}] = [0, 30S_0]$ and we construct a non-uniform grid $s_{\min} = s_1 < s_2 < \cdots < s_m = s_{\max}$ as described in Subsection 2.2.1. In Figure 8.1 the spatial grid is shown for the sample values $S_0 = 100, m = 50$ and from $s = 0$ to $s = 5S_0$ to illustrate the smaller mesh widths around the point $s = S_0$. Applying the FV discretization from Subsection 8.2.2 then yields approximations $P_j(\tau)$ of the exact values $p(s_j, \tau)$.

When trying to determine the performance of a numerical method with respect to a reference solution, it is important to take note of the computational environment in which values are calculated, and to understand the impact that has on the comparison. Our calculations take place in 64 bit IEEE floating point arithmetic. Since the solution of forward Kolmogorov equations represents a probability density function, the magnitude of the solution varies dramatically over the computational domain. This is especially true of the initial condition (a Dirac delta), and also more generally with naturally bounded

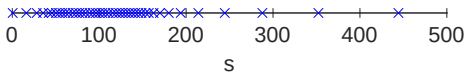


Figure 8.1: Illustration of the non-uniform grid around $s = S_0$ for the Black–Scholes example and the actual values $S_0 = 100, m = 50$.

stochastic processes with an attainable boundary, cf. the SLV model (8.1.1) with $0 < \alpha \leq 1/2$. Since IEEE floating point has a fixed-length mantissa, the density function cannot be represented to a high *absolute* accuracy uniformly over the domain. In areas where the density function is large, only high *relative* accuracy (correct number of digits) can be achieved. In addition, since the numerical solution is obtained by using implicit time stepping, we should not expect high relative accuracy of the numerical density in regions where the exact solution is small. This is because when solving the linear systems we combine many terms with very different magnitudes and then sum them up, which in IEEE arithmetic will lead to a loss of relative accuracy. Therefore when comparing the numerical solution and the reference solution we adopt a mixed absolute-relative error metric: we use relative error when the reference solution is larger than 1, absolute error if the reference solution is less than 1, and we take the maximal error value over the whole domain. More precisely, let

$$\epsilon_j(m) = \begin{cases} \left| \frac{p(s_j,T) - P_j(T)}{p(s_j,T)} \right| & \text{if } p(s_j,T) > 1, \\ |p(s_j,T) - P_j(T)| & \text{else.} \end{cases}$$

The total mixed spatial error is then defined by

$$\epsilon(m) = \max_{1 \leq j \leq m} \epsilon_j(m).$$

The value of 1 is somewhat arbitrary. The results, however, are not that sensitive to the crossover value as long as it is not too small. For the actual experiments, the values $P_j(T)$ are approximated by applying the Crank–Nicolson time stepping scheme with a large number of steps such that the temporal discretization error is negligible. In the left plot of Figure 8.2 the total mixed spatial error is shown for the relevant situation where $r_d = 0.03$, $r_f = 0.01$, $\sigma_{BS} = 0.2, T = 1$ and for the number of spatial grid points $m = \{50, 100, \ldots, 1000\}$. The corresponding numerical solution for $m = 200$ is shown in the right plot of Figure 8.2. The convergence plot clearly indicates that the FV discretization is second order convergent with respect to the current initial-boundary value problem.

As a second example we consider the Cox–Ingersoll–Ross (CIR) process, cf. [9],

$$dV_\tau = \kappa(\eta - V_\tau)d\tau + \xi\sqrt{V_\tau}dW_\tau,$$

where κ, η, ξ are strictly positive parameters. The corresponding density function is given by, see e.g. [9],

$$p(v,\tau) = ce^{-u_0-u_1}\left(\frac{u_1}{u_0}\right)^{\mathrm{f}/2}\mathrm{I}_{\mathrm{f}}(2\sqrt{u_0u_1}), \tag{8.2.18}$$

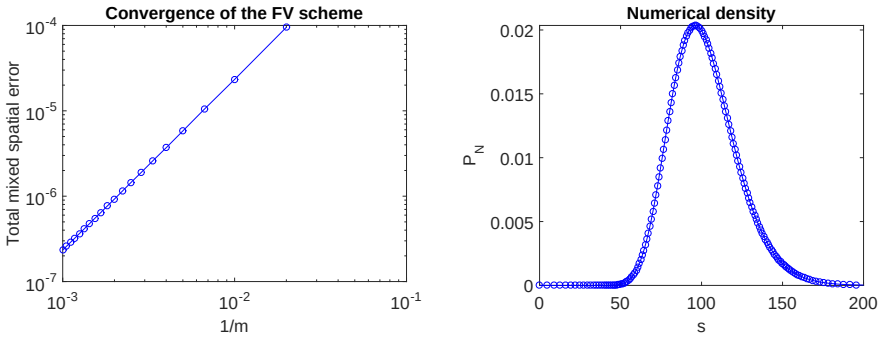


Figure 8.2: Convergence results within the 1D Black–Scholes model. The parameter values are $r_d = 0.03, r_f = 0.01, \sigma_{BS} = 0.2, T = 1$.

where

$$c = \tfrac{2\kappa}{\xi^2(1-e^{-\kappa\tau})}, \quad u_0 = cV_0 e^{-\kappa\tau}, \quad u_1 = cv, \quad \mathfrak{f} = \tfrac{2\kappa\eta}{\xi^2} - 1,$$

and $\mathrm{I}_{\mathfrak{f}}(\cdot)$ is the modified Bessel function of the first kind of order $\mathfrak{f}$. Note that the value of $\mathfrak{f}$ is directly related with the so-called Feller condition, i.e. with the possibility that $V_\tau = 0$ is attainable, cf. Chapter 7. The density function (8.2.18) satisfies the forward Kolmogorov equation

$$\tfrac{\partial}{\partial\tau}p = \tfrac{\partial^2}{\partial v^2}\left(\tfrac{1}{2}\xi^2 vp\right) - \tfrac{\partial}{\partial v}\left(\kappa(\eta - v)p\right), \tag{8.2.19}$$

for $v, \tau > 0$, with $p(v,0) = \delta(v - V_0)$. It is readily seen that if the Feller condition is violated, i.e. if $\mathfrak{f} < 0$, then the density from (8.2.18) is not defined at $v = 0$ and the density function tends to infinity as v tends to zero. In addition, around $v = 0$ the PDE (8.2.19) is strongly convection dominated which is very challenging for numerical discretization methods.

The domain is truncated to $[v_{\min}, v_{\max}] = [0, 15]$ and we construct a non-uniform grid $0 = v_1 < v_2 < \cdots < v_m = v_{\max}$ of the type (2.2.3). The spatial grid is shown in Figure 8.3 for the values $V_0 = 0.0625, m = 50$ and from $v = 0$ to $v = 0.2$ to illustrate the smaller mesh widths around $v = 0$ and $v = V_0$. Afterwards the FV discretization is applied which leads to approximations $P_k(T)$, $(1 \le k \le m)$. Recall that if $\mathfrak{f} < 0$, then the density function tends to infinity as v tends to zero and at $v = 0$ the exact density function is not defined. By increasing the number of spatial grid points m, the value of the second grid point v_2 tends to zero and adequately comparing the difference between $p(v_2, T)$ and $P_2(T)$ becomes difficult. In view of this, we opt to compute the error on similar spatial domains. Let v_{low} be the smallest non-zero grid point, i.e. the point v_2, if the total number of spatial grid points m is 50. We then define the total mixed spatial error by

$$\max_{k_1 \le k \le m} \epsilon_k(m),$$

where, for given m, k_1 is the lowest index such that $v_{k_1} \geq v_{low}$ and

$$\epsilon_k(m) = \begin{cases} \left| \frac{p(v_k,T)-P_k(T)}{p(v_k,T)} \right| & \text{if } p(v_k,T) > 1, \\ |p(v_k,T) - P_k(T)| & \text{else.} \end{cases}$$

The approximations $P_k(T)$ are determined by considering a large number of time steps with the Crank–Nicolson scheme such that the temporal discretization error is negligible. Please note that the choice $m = 50$ for defining v_{low} is not crucial. The conclusions of the numerical experiments are essentially unchanged as long as v_{low} is defined via one of the coarsest grids considered in the experiment.

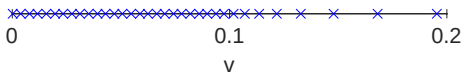


Figure 8.3: Illustration of the non-uniform grid around $v = 0$ and $v = V_0$ for the CIR example and the actual values $V_0 = 0.0625, m = 50$.

For the actual experiment we consider two sets of parameters:

	κ	η	ξ	V_0	T	$\mathfrak{f}$
Set A	5	0.16	0.9	0.0625	0.25	0.98
Set B	1.15	0.0348	0.39	0.0348	0.25	−0.47

Table 8.1: Parameter sets for the CIR example.

These sets are taken from [19] and were also used in [60]. For Set A we have $\mathfrak{f} = 0.98$ and the variance process remains strictly positive. For Set B we have $\mathfrak{f} = -0.47$ and $V_\tau = 0$ is attainable. In the left plot of Figure 8.4, respectively Figure 8.5, the total mixed spatial error is shown for the parameters of Set A, respectively Set B, and for the number of spatial grid points $m = \{50, 100, \ldots, 1000\}$. In the right plots, the corresponding numerical solutions are shown for $m = 200$. The convergence plots indicate that the FV discretization is convergent with respect to the current initial-boundary value problems. Additional experiments suggest that FV discretization is second order convergent if the Feller condition is satisfied. If $\mathfrak{f} < 0$ the order of convergence can drop to one. In addition, all the experiments confirm that the total numerical mass (8.2.6) stays constantly equal to one, even if the Feller condition is strongly violated.

8.2.4. Two-Dimensional Forward Kolmogorov Equations

In this subsection, the FV discretization from the one-dimensional case is used to define a spatial discretization for the general two-dimensional forward Kolmogorov equation (8.2.1). Suppose the spatial domain is truncated

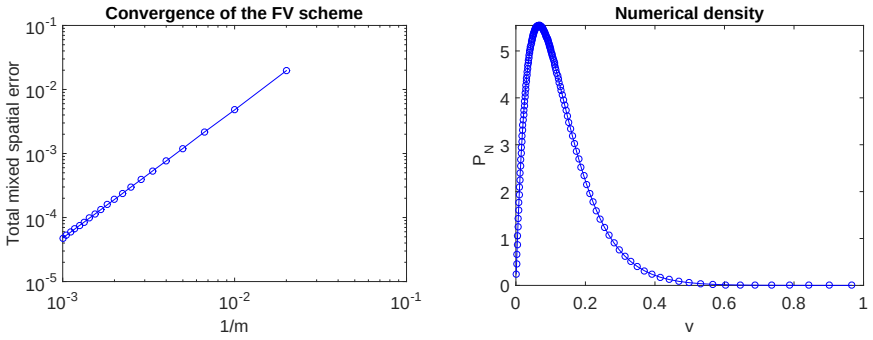


Figure 8.4: Convergence results within the CIR model. The parameters are given by Set A.

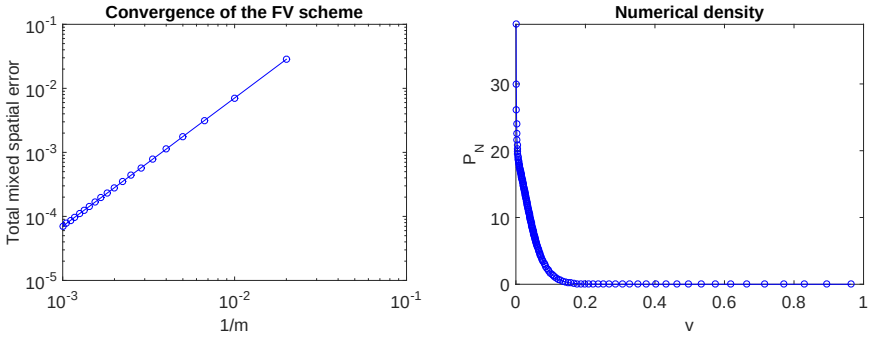


Figure 8.5: Convergence results within the CIR model. The parameters are given by Set B.

to $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$, and the spatial grid points in the x-direction, respectively y-direction, are given by

$$x_{\min} = x_1 < x_2 < \ldots < x_{m_1} = x_{\max},$$

respectively

$$y_{\min} = y_1 < y_2 < \ldots < y_{m_2} = y_{\max}.$$

Let $\Delta x_j = x_j - x_{j-1}$ and $\Delta y_k = y_k - y_{k-1}$ be the spatial mesh widths, where $\Delta x_1 = \Delta x_{m_1+1} = \Delta y_1 = \Delta y_{m_2+1} = 0$, and define volumes

$$\Omega_{j,k} := [x_{j-0.5}, x_{j+0.5}] \times [y_{k-0.5}, y_{k+0.5}],$$

where $x_{i\pm0.5}$, respectively $y_{k\pm0.5}$, are defined analogously as in the one-dimensional case.

We can now consider the two-dimensional equivalent of the cell averages, volume averages, $\overline{p}_{j,k}(\tau)$ which are defined by

$$\overline{p}_{j,k}(\tau) = \frac{1}{|\Omega_{j,k}|} \int_{\Omega_{j,k}} p(x, y, \tau) dx dy,$$

with $|\Omega_{j,k}| = (x_{j+0.5} - x_{j-0.5})(y_{k+0.5} - y_{k-0.5})$ the area of the corresponding volume. The volume average $\overline{p}_{j,k}(\tau)$ is again a second order approximation to $p(x_j, y_k, \tau)$, provided that the underlying meshes are smooth, and is the quantity that is approximated by the FV discretization. It is readily verified that

$$|\Omega_{j,k}|\frac{\partial}{\partial\tau}\overline{p}_{j,k}(\tau) = \int_{y_{k-0.5}}^{y_{k+0.5}} \left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma_1^2 p\right) - \mu_1 p\right]\Big|_{x=x_{j-0.5}}^{x=x_{j+0.5}} dy \tag{8.2.20a}$$

$$+ \int_{x_{j-0.5}}^{x_{j+0.5}} \left[\tfrac{\partial}{\partial y}\left(\tfrac{1}{2}\sigma_2^2 p\right) - \mu_2 p\right]\Big|_{y=y_{k-0.5}}^{y=y_{k+0.5}} dx \tag{8.2.20b}$$

$$+ \left[\left[\rho\sigma_1\sigma_2 p\right]\Big|_{x=x_{j-0.5}}^{x=x_{j+0.5}}\right]\Big|_{y=y_{k-0.5}}^{y=y_{k+0.5}} \tag{8.2.20c}$$

and the two-dimensional discretization is based on the approximation

$$|\Omega_{j,k}|\frac{\partial}{\partial\tau}\overline{p}_{j,k}(\tau) \approx \left[\left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma_1^2 p\right) - \mu_1 p\right]\Big|_{x=x_{j-0.5}}^{x=x_{j+0.5}}\right]\Big|_{y=y_k} \frac{\Delta y_k + \Delta y_{k+1}}{2} \tag{8.2.21a}$$

$$+ \left[\left[\tfrac{\partial}{\partial y}\left(\tfrac{1}{2}\sigma_2^2 p\right) - \mu_2 p\right]\Big|_{y=y_{k-0.5}}^{y=y_{k+0.5}}\right]\Big|_{x=x_j} \frac{\Delta x_j + \Delta x_{j+1}}{2} \tag{8.2.21b}$$

$$+ \left[\left[\rho\sigma_1\sigma_2 p\right]\Big|_{x=x_{j-0.5}}^{x=x_{j+0.5}}\right]\Big|_{y=y_{k-0.5}}^{y=y_{k+0.5}}. \tag{8.2.21c}$$

Equation (8.2.20) includes several flux terms that are similar to the flux terms in Subsections 8.2.1 and 8.2.2. It reflects the fact that the total integral of p over a volume changes only as a result of the flux difference over the volume boundary. This is completely analogous with the one-dimensional interpretation of (8.2.3). If the total flux over the boundary of the spatial domain is zero, i.e. if

$$\begin{aligned} 0 = &\int_{y_{\min}}^{y_{\max}} \left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma_1^2 p\right) - \mu_1 p\right]\Big|_{x=x_{\min}}^{x=x_{\max}} dy \\ &+ \int_{x_{\min}}^{x_{\max}} \left[\tfrac{\partial}{\partial y}\left(\tfrac{1}{2}\sigma_2^2 p\right) - \mu_2 p\right]\Big|_{y=y_{\min}}^{y=y_{\max}} dx \\ &+ \left[\left[\rho\sigma_1\sigma_2 p\right]\Big|_{x=x_{\min}}^{x=x_{\max}}\right]\Big|_{y=y_{\min}}^{y=y_{\max}}, \end{aligned}$$

then the total integral of p over the entire domain is constant in time.

Let $\boldsymbol{P}_{j,k}(\tau)$ denote approximations to the exact value $\overline{p}_{j,k}(\tau)$, denote by $\boldsymbol{P}(\tau)$ the $m_1 \times m_2$ matrix with entries $\boldsymbol{P}_{j,k}(\tau)$ and let

$$P(\tau) = \mathrm{vec}[\boldsymbol{P}(\tau)],$$

where vec$[\cdot]$ again denotes the operator that turns any given matrix into a vector by putting its successive columns below each other. The bold notation is only introduced to indicate the subtle difference between the matrix form and the vectorised form of the approximations. Similarly to the one-dimensional

discretization in Subsections 8.2.1 and 8.2.2, we discretize (8.2.21) in the following way by introducing numerical fluxes

$$\boldsymbol{P}'_{j,k}(\tau) = \left[\boldsymbol{f}_{j-0.5,k}(P,\tau) - \boldsymbol{f}_{j+0.5,k}(P,\tau)\right] \tfrac{2}{\Delta x_j+\Delta x_{j+1}} \tag{8.2.22a}$$

$$+ \left[\boldsymbol{f}_{j,k-0.5}(P,\tau) - \boldsymbol{f}_{j,k+0.5}(P,\tau)\right] \tfrac{2}{\Delta y_k+\Delta y_{k+1}} \tag{8.2.22b}$$

$$+ \sum_{j_1,k_1=0}^{1} (-1)^{j_1+k_1} \boldsymbol{f}_{m,j-0.5+j_1,k-0.5+k_1} \tfrac{2}{\Delta x_j+\Delta x_{j+1}} \tfrac{2}{\Delta y_k+\Delta y_{k+1}}, \tag{8.2.22c}$$

for $1 \le j \le m_1, 1 \le k \le m_2$. For the ease of presentation, denote

$$\begin{aligned}
&\mu_{1,j\pm0.5,k} = \mu_1(x_{j\pm0.5}, y_k, \tau), && \sigma_{1,j,k} = \sigma_1(x_j, y_k, \tau),\\
&\mu_{2,j,k\pm0.5} = \mu_2(x_j, y_{k\pm0.5}, \tau), && \sigma_{2,j,k} = \sigma_2(x_j, y_k, \tau),\\
&\sigma_{1,j\pm0.5,k\pm0.5} = \sigma_1(x_{j\pm0.5}, y_{k\pm0.5}, \tau), && \sigma_{2,j\pm0.5,k\pm0.5} = \sigma_2(x_{j\pm0.5}, y_{k\pm0.5}, \tau).
\end{aligned}$$

Since the actual form of the fluxes in (8.2.21) is completely similar to the form of the fluxes in Subsection 8.2.4, we define the numerical fluxes by

$$\boldsymbol{f}_{j\pm0.5,k}(P,\tau) = \boldsymbol{f}_{a,j\pm0.5,k}(P,\tau) + \boldsymbol{f}_{d,j\pm0.5,k}(P,\tau),$$

with

$$\boldsymbol{f}_{a,j-0.5,k}(P,\tau) = \mu_{1,j-0.5,k} \tfrac{\boldsymbol{P}_{j-1,k}(\tau)+\boldsymbol{P}_{j,k}(\tau)}{2} \approx \mu_{1,j-0.5,k} p(x_{j-0.5}, y_k, \tau),$$

and

$$\begin{aligned}
\boldsymbol{f}_{d,j-0.5,k}(P,\tau) &= \tfrac{\frac{1}{2}\sigma^2_{1,j-1,k}\boldsymbol{P}_{j-1,k}(\tau) - \frac{1}{2}\sigma^2_{1,j,k}\boldsymbol{P}_{j,k}(\tau)}{\Delta x_j}\\
&\approx -\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma_1^2(x, y_k, \tau)p(x, y_k, \tau)\right)|_{x=x_{j-0.5}},
\end{aligned}$$

for $2 \le j \le m_1, 1 \le k \le m_2$. Moreover

$$\boldsymbol{f}_{j,k\pm0.5}(P,\tau) = \boldsymbol{f}_{a,j,k\pm0.5}(P,\tau) + \boldsymbol{f}_{d,j,k\pm0.5}(P,\tau),$$

where

$$\boldsymbol{f}_{a,j,k-0.5}(P,\tau) = \mu_{2,j,k-0.5} \tfrac{\boldsymbol{P}_{j,k-1}(\tau)+\boldsymbol{P}_{j,k}(\tau)}{2} \approx \mu_{2,j,k-0.5} p(x_j, y_{k-0.5}, \tau),$$

and

$$\begin{aligned}
\boldsymbol{f}_{d,j,k-0.5}(P,\tau) &= \tfrac{\frac{1}{2}\sigma^2_{2,j,k-1}\boldsymbol{P}_{j,k-1}(\tau) - \frac{1}{2}\sigma^2_{2,j,k}\boldsymbol{P}_{j,k}(\tau)}{\Delta y_k}\\
&\approx -\tfrac{\partial}{\partial y}\left(\tfrac{1}{2}\sigma_2^2(x_j, y, \tau)p(x_j, y, \tau)\right)|_{y=y_{k-0.5}},
\end{aligned}$$

for $1 \le j \le m_1, 2 \le k \le m_2$ Finally, for the mixed spatial derivative we define

$$\begin{aligned}
\boldsymbol{f}_{m,j-0.5,k-0.5}(P,\tau) &= \rho\sigma_{1,j-0.5,k-0.5}\sigma_{2,j-0.5,k-0.5}\\
&\quad \times \tfrac{\boldsymbol{P}_{j-1,k-1}(\tau)+\boldsymbol{P}_{j-1,k}(\tau)+\boldsymbol{P}_{j,k-1}(\tau)+\boldsymbol{P}_{j,k}(\tau)}{4}\\
&\approx \rho\sigma_{1,j-0.5,k-0.5}\sigma_{2,j-0.5,k-0.5} p(x_{j-0.5}, y_{k-0.5}, \tau),
\end{aligned} \tag{8.2.23}$$

for $1 \leq j \leq m_1 + 1, 1 \leq k \leq m_2 + 1$, where it is assumed that

$$\begin{aligned} \boldsymbol{P}_{0,k}(\tau) &:= \boldsymbol{P}_{1,k}(\tau), \quad \boldsymbol{P}_{m_1+1,k}(\tau) := \boldsymbol{P}_{m_1,k}(\tau), \\ \boldsymbol{P}_{j,0}(\tau) &:= \boldsymbol{P}_{j,1}(\tau), \quad \boldsymbol{P}_{j,m_2+1}(\tau) := \boldsymbol{P}_{j,m_2}(\tau), \end{aligned} \tag{8.2.24}$$

such that the general formula is naturally extended at the boundaries of the spatial domain. The numerical flux term (8.2.23) can be viewed as the result of applying the discretization for the convection part first in the x-direction and then in the y-direction.

The semidiscretization is completed by defining boundary conditions and discretizations at the boundaries of the truncated domain. Since p is again a density function, it follows that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\tfrac{\partial}{\partial \tau} p\right] dxdy = 0.$$

Inserting the right-hand side of the PDE (8.2.1), this can be rewritten as

$$\begin{aligned} 0 = &\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \left[\tfrac{\partial^2}{\partial x^2} \left(\tfrac{1}{2}\sigma_1^2 p \right) - \tfrac{\partial}{\partial x} \left(\mu_1 p \right) \right] dx \right) dy \\ &+ \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \left[\tfrac{\partial^2}{\partial y^2} \left(\tfrac{1}{2}\sigma_2^2 p \right) - \tfrac{\partial}{\partial y} \left(\mu_2 p \right) \right] dy \right) dx \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tfrac{\partial^2}{\partial x \partial y} \left(\rho \sigma_1 \sigma_2 p \right) dxdy. \end{aligned}$$

Analogously to the one-dimensional case it is assumed that the boundaries are chosen sufficiently far away from the spot value (X_0, Y_0) or that they are defined by a natural truncation of the spatial domain. The condition above is then approximated by

$$\begin{aligned} 0 = &\int_{y_{\min}}^{y_{\max}} \left[\tfrac{\partial}{\partial x} \left(\tfrac{1}{2}\sigma_1^2 p \right) - \mu_1 p \right] \Big|_{x=x_{\min}}^{x=x_{\max}} dy \\ &+ \int_{x_{\min}}^{x_{\max}} \left[\tfrac{\partial}{\partial y} \left(\tfrac{1}{2}\sigma_2^2 p \right) - \mu_2 p \right] \Big|_{y=y_{\min}}^{y=y_{\max}} dx \\ &+ \int_{y_{\min}}^{y_{\max}} \int_{x_{\min}}^{x_{\max}} \tfrac{\partial^2}{\partial x \partial y} \left(\rho \sigma_1 \sigma_2 p \right) dxdy. \end{aligned} \tag{8.2.25}$$

Note that by assuming that

$$\begin{aligned} \rho\sigma_1\sigma_2 p|_{x=x_{\min},y=y_{\min}} &= \rho\sigma_1\sigma_2 p|_{x=x_{\min},y=y_{\max}} = 0, \\ \rho\sigma_1\sigma_2 p|_{x=x_{\max},y=y_{\min}} &= \rho\sigma_1\sigma_2 p|_{x=x_{\max},y=y_{\max}} = 0, \end{aligned} \tag{8.2.26}$$

the last integral, corresponding to the mixed derivative term, is always equal to zero. Next, we generalise the idea that there are no fluxes at the boundaries, i.e. that no mass is coming in or going out at the boundaries. In light of this

it is assumed that the following boundary conditions hold:

$$\begin{aligned}
\left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma_1^2 p\right) - (\mu_1 p)\right]\Big|_{x=x_{\min}} &= 0, \quad \text{for } y_{\min} \leq y \leq y_{\max},\\
\left[\tfrac{\partial}{\partial x}\left(\tfrac{1}{2}\sigma_1^2 p\right) - (\mu_1 p)\right]\Big|_{x=x_{\max}} &= 0, \quad \text{for } y_{\min} \leq y \leq y_{\max},\\
\left[\tfrac{\partial}{\partial y}\left(\tfrac{1}{2}\sigma_2^2 p\right) - (\mu_2 p)\right]\Big|_{y=y_{\min}} &= 0, \quad \text{for } x_{\min} \leq x \leq x_{\max},\\
\left[\tfrac{\partial}{\partial y}\left(\tfrac{1}{2}\sigma_2^2 p\right) - (\mu_2 p)\right]\Big|_{y=y_{\max}} &= 0, \quad \text{for } x_{\min} \leq x \leq x_{\max}.
\end{aligned}$$

By combining these boundary conditions with the assumption (8.2.26) it follows that condition (8.2.25) is satisfied.

For the discretization of the one-dimensional fluxes in (8.2.21a), respectively (8.2.21b), at the boundaries of the truncated domain, the approach from the one-dimensional case is generalised. By using a similar discretization of the boundary conditions, one ends up with numerical fluxes which are zero at the boundaries, i.e. with

$$\boldsymbol{f}_{0.5,k}(P,\tau) = \boldsymbol{f}_{m_1+0.5,k}(P,\tau) = \boldsymbol{f}_{j,0.5}(P,\tau) = \boldsymbol{f}_{j,m_2+0.5}(P,\tau) = 0,$$

for $1 \leq j \leq m_1, 1 \leq k \leq m_2$. The fluxes stemming from the mixed derivative term, see (8.2.21c), are discretized at the boundaries by using (8.2.23) in combination with (8.2.24). It is readily verified that if

$$\begin{aligned}
0 &= \rho\sigma_{1,1,1}\sigma_{2,1,1}\boldsymbol{P}_{1,1}(\tau) = \rho\sigma_{1,1,m_2}\sigma_{2,1,m_2}\boldsymbol{P}_{1,m_2}(\tau)\\
&= \rho\sigma_{1,m_1,1}\sigma_{2,m_1,1}\boldsymbol{P}_{m_1,1}(\tau) = \rho\sigma_{1,m_1,m_2}\sigma_{2,m_1,m_2}\boldsymbol{P}_{m_1,m_2}(\tau),
\end{aligned}$$

which is the semidiscrete version of (8.2.26), then the total numerical flux over the boundary of the spatial domain is equal to zero and the total numerical mass is kept constant in time, i.e.

$$\sum_{j=1}^{m_1}\sum_{k=1}^{m_2}\boldsymbol{P}_{j,k}(\tau)|\Omega_{j,k}| = \text{constant}, \quad \text{for } \tau \geq 0.$$

As stated above, some processes are naturally bounded, e.g. the general variance process from Section 8.1 can never become negative. Suppose for example that the process corresponding to the y-variable in the PDE (8.2.1) is bounded from below. Then, $y_{\min}$ is naturally taken equal to this lower boundary. Moreover, it can happen that this lower boundary is attainable (cf. the variance process from Section 8.1 with $\alpha < 1/2$) and probability mass can stack up at this boundary. This can cause instabilities in the approximation of the mixed derivative term near this boundary. In order to deal with this, if for example the boundary $y_{\min}$ is attainable, the central FV scheme in the y-direction for the "mixed derivative fluxes" (8.2.21c) at $y_{\min} + \frac{1}{2}\Delta y_2$ are replaced by a first-order forward scheme. More precisely, the $\boldsymbol{f}_{m,j\pm 0.5,1.5}(P,\tau)$ from above are then replaced by

$$\boldsymbol{f}_{m,j-0.5,1.5}(P,\tau) = \rho\sigma_{1,j-0.5,1.5}\sigma_{2,j-0.5,1.5}\frac{\boldsymbol{P}_{j-1,2}(\tau)+\boldsymbol{P}_{j,2}(\tau)}{2},$$

for $1 \leq j \leq m_1 + 1$, where

$$\boldsymbol{P}_{0,2}(\tau) := \boldsymbol{P}_{1,2}(\tau), \quad \boldsymbol{P}_{m_1+1,2}(\tau) := \boldsymbol{P}_{m_1,2}(\tau).$$

The total spatial discretization (8.2.22) yields a large system of differential equations. By making use of a well-known property of the Kronecker product, this system of differential equations can be written as a system of ODEs

$$P'(\tau) = A(\tau)P(\tau), \tag{8.2.27}$$

for $\tau > 0$ and with given matrix $A(\tau)$. Analogously to the one-dimensional case, since the values $\boldsymbol{P}_{j,k}(\tau)$ can be seen as approximations of the cell averages $\overline{p}(x_j, y_k, \tau)$, it is natural to define the initial vector as $P(0) = \mathrm{vec}[\boldsymbol{P}(0)]$ where

$$\boldsymbol{P}_{j,k}(0) = \begin{cases} \frac{1}{|\Omega_{j,k}|} & \text{if } (X_0, V_0) \in \Omega_{j,k}, \\ 0 & \text{else.} \end{cases}$$

Note that the matrix A can be split as

$$A = A_0 + A_1 + A_2, \tag{8.2.28}$$

where A_1, respectively A_2, represents the discretization of the spatial derivatives in the first, respectively second, spatial dimension. The matrix A_0 represents the discretization of the mixed spatial derivative term in (8.2.1). Due to this decomposition, the four ADI schemes from Section 3.2 can easily be applied for the temporal discretization of the semidiscrete system (8.2.27).

8.2.5. Numerical Experiments for Two-Dimensional Forward Kolmogorov Equations

In this section the performance of the two-dimensional FV discretization is tested for two practical examples. For the first experiment we consider the two-dimensional Black–Scholes model which can be described by the system of SDEs

$$\begin{cases} dS_{1,\tau} = rS_{1,\tau}d\tau + \sigma_{1,BS}S_{1,\tau}dW_\tau^{(1)}, \\ \\ dS_{2,\tau} = rS_{2,\tau}d\tau + \sigma_{2,BS}S_{2,\tau}dW_\tau^{(2)}, \end{cases}$$

with $dW_\tau^{(1)} \cdot dW_\tau^{(2)} = \rho d\tau$, $-1 \leq \rho \leq 1$ and $r, \sigma_{1,BS}, \sigma_{2,BS}$ strictly positive constants. The corresponding forward Kolmogorov equation is given by

$$\begin{aligned} \tfrac{\partial}{\partial\tau}p = {} & \tfrac{\partial^2}{\partial s_1^2}\left(\tfrac{1}{2}\sigma_{1,BS}^2 s_1^2 p\right) + \tfrac{\partial^2}{\partial s_1 \partial s_2}\left(\rho\sigma_{1,BS}\sigma_{2,BS}s_1 s_2 p\right) + \tfrac{\partial^2}{\partial s_1^2}\left(\tfrac{1}{2}\sigma_{2,BS}^2 s_2^2 p\right) \\ & - \tfrac{\partial}{\partial s_1}\left(r s_1 p\right) - \tfrac{\partial}{\partial s_2}\left(r s_2 p\right), \end{aligned}$$

for $s_1, s_2, \tau > 0$ and with $p(s_1, s_2, 0) = \delta(s_1 - S_{1,0})\delta(s_2 - S_{2,0})$ for given values $S_{1,0}, S_{2,0}$. The exact solution is known analytically and can be written as

$$p(s_1, s_2, \tau) = n_2(\log(s_1/S_{1,0}), \log(s_2/S_{2,0}), \tau)\tfrac{1}{s_1}\tfrac{1}{s_2}, \quad \text{for } s_1 > 0, s_2 > 0, \tau > 0,$$

where this time $n_2(x, y, \tau)$ is the density function of a two-dimensional normally distributed random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\mu} = \begin{bmatrix} (r - \frac{1}{2}\sigma_{1,BS}^2)\tau \\ \\ (r - \frac{1}{2}\sigma_{2,BS}^2)\tau \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,BS}^2\tau & \rho\sigma_{1,BS}\sigma_{2,BS}\tau \\ \\ \rho\sigma_{1,BS}\sigma_{2,BS}\tau & \sigma_{2,BS}^2\tau \end{bmatrix}.$$

8

Similarly to the domain truncation in the one-dimensional numerical experiment from Subsection 8.2.3, the spatial domain is truncated to $[s_{1,\min}, s_{1,\max}] \times [s_{2,\min}, s_{2,\max}] = [0, 30S_{1,0}] \times [0, 30S_{2,0}]$. The Cartesian spatial grid,

$$(s_{1,j}, s_{2,k}) \quad \text{for } 1 \leq j \leq m_1, 1 \leq k \leq m_2,$$

is constructed by considering the spatial grid from Subsection 8.2.3 in both spatial dimensions. In Figure 8.6 this spatial grid is illustrated within the region $[0, 5S_{1,0}] \times [0, 5S_{2,0}]$ for $S_{1,0} = S_{1,0} = 100$ and $m_1 = m_2 = 50$.

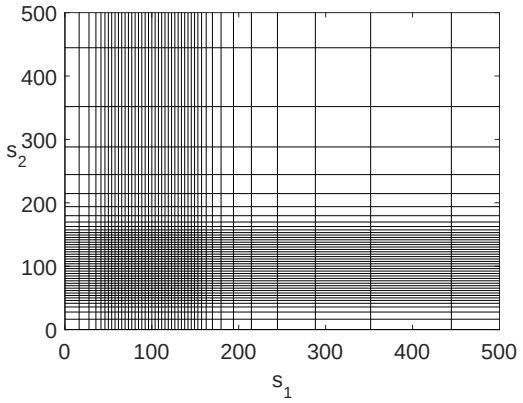


Figure 8.6: Illustration of the non-uniform grid around $(S_{1,0}, S_{2,0})$ for the 2D Black–Scholes example and the actual values $S_{1,0} = S_{2,0} = 100$, $m_1 = m_2 = 50$.

The FV discretization from Subsection 8.2.4 then defines approximations $\boldsymbol{P}_{j,k}(\tau)$ to $p(s_{1,j}, s_{2,k}, \tau)$ and we compute the total mixed spatial error

$$\max_{1 \leq j \leq m_1, 1 \leq k \leq m_2} \epsilon_{j,k}(m),$$

where

$$\epsilon_{j,k}(m) = \begin{cases} \left| \frac{p(s_{1,j}, s_{2,k}, T) - \boldsymbol{P}_{j,k}(T)}{p(s_{1,j}, s_{2,k}, T)} \right| & \text{if } p(s_{1,j}, s_{2,k}, T) > 1, \\ |p(s_{1,j}, s_{2,k}, T) - \boldsymbol{P}_{j,k}(T)| & \text{else.} \end{cases}$$

The values $\boldsymbol{P}_{j,k}(T)$ are calculated by using the HV scheme with $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ and a large number of time steps such that the temporal discretization error is negligible. In the left plot of Figure 8.7 the total mixed spatial error is shown for the parameter values $r = 0.03$, $\sigma_{1,BS} = 0.2$, $\sigma_{2,BS} = 0.25$, $\rho = -0.7$, $T = 1$ and for the number of spatial grid points $m_1 = m_2 = \{50, 100, \ldots, 500\}$. The right plot shows the numerical solution for $m_1 = m_2 = 100$. The convergence plot indicates that the FV discretization is second order convergent with respect to the current initial-boundary value problem.

For the second example we consider the popular Heston model [30], i.e. the SV model (7.1.3) with $\psi(v) = \sqrt{v}$ and $\alpha = 1/2$. The underlying density

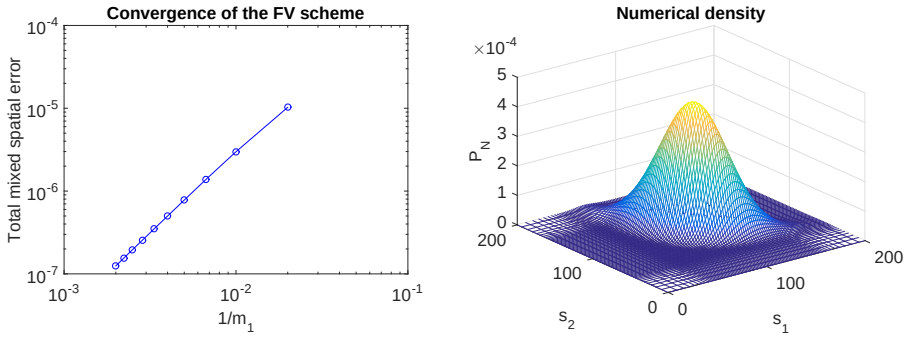


Figure 8.7: Convergence results within the 2D Black–Scholes model. The parameter values are $r = 0.03, \sigma_{1,BS} = 0.2, \sigma_{2,BS} = 0.25, \rho = -0.7, T = 1$.

function satisfies the forward Kolmogorov equation

$$\begin{aligned}\frac{\partial}{\partial\tau}p = &\frac{\partial^2}{\partial x^2}\left(\tfrac{1}{2}vp\right) + \frac{\partial^2}{\partial x\partial v}\left(\rho\xi vp\right) + \frac{\partial^2}{\partial v^2}\left(\tfrac{1}{2}\xi^2 vp\right)\\ &- \frac{\partial}{\partial x}\left((r_d - r_f - \tfrac{1}{2}v)p\right) - \frac{\partial}{\partial v}\left(\kappa(\eta - v)p\right),\end{aligned} \tag{8.2.29}$$

for $x \in \mathbb{R}, v > 0, \tau > 0$ and with initial condition $p(x, v, 0) = \delta(x)\delta(v - V_0)$ where $X_0 = 0$ and V_0 is given. To the best of our knowledge, no analytical solution is available in the literature for the density function p that satisfies (8.2.29). In order to test the performance of our FV discretization, we compute a reference solution with an alternative discretization method described in [19]. The latter method is based on rewriting

$$p(x, v, \tau) = p_1(x, \tau|V_{\mathrm{SV},\tau} = v)p_2(v, \tau), \tag{8.2.30}$$

where $p_2(v, \tau)$ denotes the one-dimensional density of the volatility process, and $p_1(x, \tau|V_{\mathrm{SV},\tau} = v)$ denotes the conditional density of X_τ given the variance value $V_{\mathrm{SV},\tau} = v$. In [19] the characteristic function

$$\psi(\omega|V_{\mathrm{SV},\tau} = v) = \mathbb{E}[\exp(\mathbf{i}\omega X_{\mathrm{SV},\tau})|V_{\mathrm{SV},\tau} = v]$$

corresponding to $p_1(x, \tau|V_{\mathrm{SV},\tau} = v)$ is given in semi-analytical form and it is stated that $p_2(v, \tau)$ is given by (8.2.18). By using the COS-method [18] we approximate the conditional density function $p_1(x, \tau|V_{\mathrm{SV},\tau} = v)$ and our reference solution p_{ref} is then defined via (8.2.30).

In the numerical experiment, we opt to truncate the domain in the x-direction to $[X_0 - \log(30), X_0 + \log(30)]$. The spatial domain in the v-direction is truncated to $[0, 15]$, analogously to the CIR example. We consider spatial grids

$$-\log(30) = x_1 < x_2 < \cdots < x_{m_1} = \log(30),$$

$$0 = v_1 < v_2 < \cdots < v_{m_2} = 15,$$

with $m_1 = 2m_2$, which are similar to the ones described in Subsection 2.2.1 and such that there exist indices j_0, k_0 such that $(x_{j_0}, v_{k_0}) = (X_0, V_0)$. In

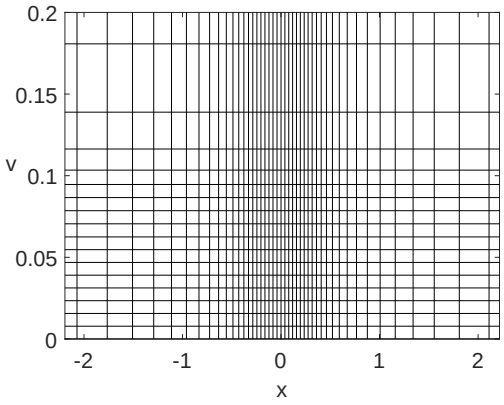


Figure 8.8: Illustration of the non-uniform grid around (X_0, V_0) for the Heston example and the actual values $X_0 = 0, V_0 = 0.0625$, $m_1 = 2m_2 = 50$.

Figure 8.8 the spatial grid is shown for $X_0 = 0, V_0 = 0.0625$ and the sample value $m_1 = 2m_2 = 50$.

Applying the FV discretization from Subsection 8.2.4 yields approximations $\boldsymbol{P}_{j,k}(\tau)$ to $p(x_j, v_k, \tau)$. From equation (8.2.30) and the CIR example it follows that if $\mathfrak{f} = \frac{2\kappa\eta}{\xi^2} - 1 < 0$, then the density function can tend to infinity as v tends to zero and at $v = 0$ the exact density is not defined. Analogously to the remark in Subsection 8.2.3, it is readily seen that adequately comparing the difference between $p(x_j, v_2, T)$ and $P_{j,2}(T)$ then becomes difficult for increasing values of m_2. The error is therefore computed on similar spatial domains. Let v_{low} again be the smallest non-zero grid point in the v-direction if the total number of grid points in that direction is $m_2 = 50$. For given m_2, let k_1 be the lowest index such that $v_{k_1} \geq v_{low}$ and define the total mixed spatial error by

$$\max_{1 \leq j \leq m_1, k_1 \leq k \leq m_2} \epsilon_{j,k}(m),$$

where

$$\epsilon_{j,k}(m) = \begin{cases} \left| \frac{p_{ref}(x_j, v_k, T) - \boldsymbol{P}_{j,k}(T)}{p_{ref}(x_j, v_k, T)} \right| & \text{if } p_{ref}(x_j, v_k, T) > 1, \\ |p_{ref}(x_j, v_k, T) - \boldsymbol{P}_{j,k}(T)| & \text{else.} \end{cases}$$

The values $P_{j,k}(T)$ are once more approximated by applying the HV scheme with $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ and a small temporal step size such that the temporal discretization error is negligible.

For the actual numerical experiments we consider an extension of the two sets of parameters used in Subsection 8.2.3. The extensions are also taken from [19], used in [60], and given by

	κ	η	ξ	ρ	r_d	V_0	T	$\mathfrak{f}$
Set C	5	0.16	0.9	0.1	0.1	0.0625	0.25	0.98
Set D	1.15	0.0348	0.39	−0.64	0.04	0.0348	0.25	−0.47

Table 8.2: Parameter sets for the Heston example.

with $X_0 = 0$ and $r_f = 0$. Recall that for Set C we have that $\mathfrak{f} = 0.98$ and for Set D it holds that $\mathfrak{f} = -0.47$. In the left plot of Figure 8.9, respectively Figure 8.10, the total mixed spatial error is shown for the parameters from Set C, respectively from Set D. The number of spatial grid points is given by $m_1 = 2m_2 = \{50, 100, \ldots, 500\}$. In the right plots, the corresponding numerical solutions are shown for $m_1 = 2m_2 = 100$. The convergence plots indicate that the FV discretization is convergent with respect to the current initial-boundary value problems. Additional experiments again suggest that the FV discretization is second order convergent if the Feller condition is satisfied. If $\mathfrak{f} < 0$ the order of convergence can drop to one. Please note that the conclusions of the numerical experiments are essentially unchanged for different values of v_{low} as long as it is defined via one of the coarsest grids considered in the experiment. The two-dimensional tests also confirm that the total numerical mass is, indeed, kept constantly equal to one, even if the Feller condition is strongly violated.

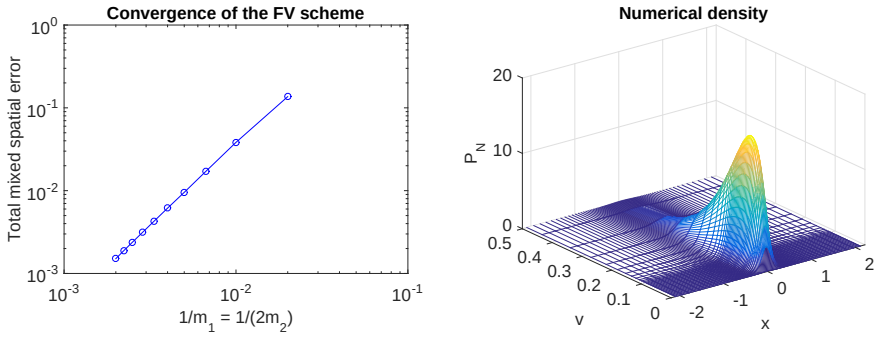


Figure 8.9: Convergence results within the Heston model. The parameters are given by Set C.

8

8.3. Temporal Discretization

Spatial discretization of forward Kolmogorov equations with the FV method leads to large systems of stiff ODEs. In the one-dimensional case, the matrix corresponding to the semidiscrete system (8.2.16) is tridiagonal and temporal discretization can be performed very efficiently with standard implicit time stepping methods such as the Crank–Nicolson scheme, see e.g. Subsection 2.3.4. It is readily seen that in the two-dimensional case, the semidiscretization matrix stemming from (8.2.27) has in general nine non-zero elements per row and

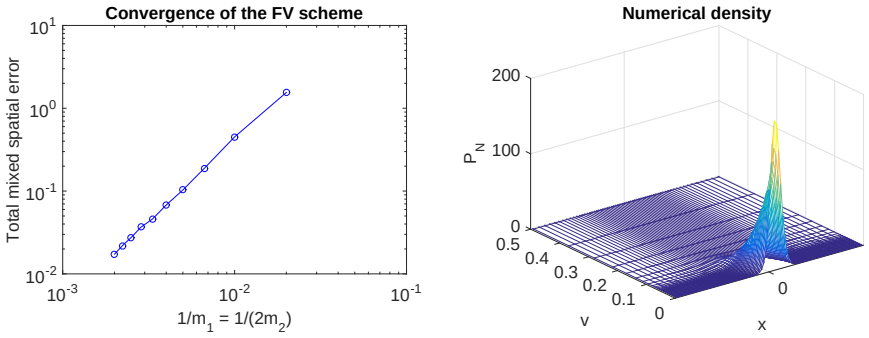


Figure 8.10: Convergence results within the Heston model. The parameters are given by Set D.

column. Application of standard implicit time stepping methods often requires solving linear systems of equations involving a matrix

$$B = I - \theta \Delta t A,$$

where I denotes again the identity matrix of the same size as A, Δt is the temporal step size and θ denotes a parameter of the method. As illustrated in Figure 8.11, FV discretization of two-dimensional forward Kolmogorov equations gives rise to a matrix B with non-zero subdiagonals that lie at a distance $m_1 + 1$ from the main diagonal. The number of operations required in every time step can grow faster than the total number of spatial grid points, cf. Subsection 2.3.4, which is not favourable.

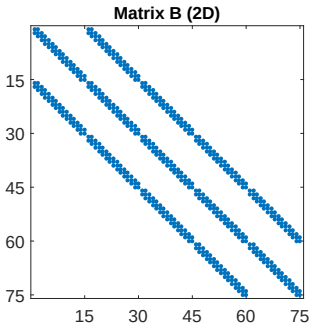


Figure 8.11: Sparsity structure of the matrix B corresponding to the semidiscrete system (8.2.27) *where $m_1 = 15$, $m_2 = 5$.*

At the end of Subsection 8.2.4 it is shown that these semidiscrete systems, stemming from spatial discretization with the FV method of two-dimensional forward Kolmogorov equations, allow for a splitting (8.2.28). This splitting is of the type (3.2.2) and, hence, the four ADI schemes introduced in Chapter 3 can be applied very naturally. It is readily verified that A_1 from (8.2.28) is

tridiagonal, A_2 is essentially tridiagonal and A_0 has in general nine non-zero elements per row and column. Application of the ADI schemes requires solving linear systems of equations involving matrices

$$B_1 = I - \theta \Delta t A_1 \quad \text{and} \quad B_2 = I - \theta \Delta t A_2,$$

which are both essentially tridiagonal. The matrix A_0 is only used in the explicit steps. Since solving linear systems of equations involving essentially tridiagonal matrices can be performed very efficiently, the application of ADI schemes can lead to a major computational advantage in comparison with classical implicit time stepping methods, cf. Section 3.1.

In this chapter, we opt to employ the HV scheme (3.2.6) with parameter $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$. An overview of the existing stability results for the HV scheme, relevant to semidiscretized two-dimensional convection-diffusion equations with mixed derivative term, is presented in Section 3.3. In Chapter 5 it is shown that, under natural stability and smoothness assumptions, the HV scheme is second order convergent with respect to the temporal step size. The temporal convergence result has the key property that it holds uniformly in the spatial mesh width. In this chapter, the vector $P(0)$ is stemming from an initial function that is non-smooth. The analysis in Chapter 6 shows that convergence can then be seriously impaired. Based on our analysis in the pertinent chapter, we apply Rannacher time stepping with $N_0 = 2$, i.e. we replace the first two HV time steps by four half-time steps of the implicit Euler scheme.

8.4. Calibration of the SLV Model to the LV Model

As stated in the introduction of the chapter, the goal is to calibrate state-of-the-art SLV models to its underlying LV model in order to reproduce the known market prices for European call and put options. This is done by defining the leverage function σ_{SLV} such that the relationship (8.1.3) is satisfied. By combining equations (8.1.3)–(8.1.5), it is readily seen that a highly non-linear PDE needs to be solved. In this section the FV-ADI method is used in combination with an inner iteration to approximate the corresponding leverage function and density function.

In order to use the FV-ADI discretization, one first has to define spatial and temporal grids. Since the initial function of forward Kolmogorov equations is highly non-smooth around the spot value (X_0, V_0), and the region of interest is also situated around this value, it is natural to consider non-uniform Cartesian grids that are concentrated around the value (X_0, V_0). If the parameter α from the SLV model is chosen smaller than or equal to $1/2$, then the natural boundary $V_\tau = 0$ can be reached and probability mass stacks up at the boundary $v = 0$. It is then natural to additionally require smaller mesh widths in the v-direction at this boundary. The non-uniform grids define volumes of which the volume average is approximated by the FV scheme. Denote by m_1, respectively m_2, the number of spatial grid points in the x-direction, respectively v-direction. We consider spatial grids

$$x_{\min} = x_1 < x_2 < \cdots < x_{m_1} = x_{\max},$$

$$0 = v_1 < v_2 < \cdots < v_{m_2} = v_{\max},$$

which are similar to the ones described in Subsection 2.2.1 and such that there exist indices j_0, k_0 such that $(x_{j_0}, v_{k_0}) = (X_0, V_0)$. Recall that the pertinent meshes are smooth. Denote the corresponding mesh widths by $\Delta x_j, \Delta v_k$ and define volumes

$$\Omega_{j,k} = [x_{j-0.5}, x_{j+0.5}] \times [v_{k-0.5}, v_{k+0.5}],$$

where the values $x_{j-0.5}, v_{k-0.5}$ are defined similarly as in Section 8.2. The values $x_{\min}, x_{\max}, v_{\max}$ are chosen sufficiently far away from the spot value such that the boundary conditions from Section 8.2 can be applied. An example of the spatial grid for the small sample values $m_1 = 2m_2 = 50$ was already shown in Figure 8.8 for the case where $(X_0, V_0) = (0, 0.0625)$. For the discretization in time we always consider uniform grids $\tau_n = n\Delta\tau$ where the temporal step size is given by $\Delta\tau = T/N$ and N denotes the total number of time steps.

Once the spatial grid and corresponding volumes are defined, the FV discretization from Section 8.2 can be applied. This yields a large system of ordinary differential equations

$$P'(\tau) = A(\tau)P(\tau) = (A_0(\tau) + A_1(\tau) + A_2(\tau))P(\tau) \quad (0 \leq \tau \leq T), \tag{8.4.1}$$

with given matrices $A_0(\tau), A_1(\tau), A_2(\tau)$ and initial function defined via

$$\boldsymbol{P}_{j,k}(0) = \begin{cases} \frac{2}{\Delta x_{j-1} + \Delta x_j} \frac{2}{\Delta v_{k-1} + \Delta v_k} & \text{if } (j,k) = (j_0, k_0), \\ 0 & \text{else.} \end{cases}$$

The matrices A_0, A_1 contain, however, the unknown function σ_{SLV} at the spatial grid points. $P(\tau)$ can be used in combination with a numerical integration technique to approximate the conditional expectations (8.1.4) and hence the pertinent leverage function. We opt to perform numerical integration with the trapezoidal rule and define approximations

$$\mathbb{E}_j(\tau) = \frac{\sum_{k=1}^{m_2} \psi^2(v_k) \boldsymbol{P}_{j,k}(\tau) \frac{\Delta v_k + \Delta v_{k+1}}{2}}{\sum_{k=1}^{m_2} \boldsymbol{P}_{j,k}(\tau) \frac{\Delta v_k + \Delta v_{k+1}}{2}} \approx \mathbb{E}[\psi^2(V_\tau) | X_\tau = x_j], \tag{8.4.2}$$

where we recall that $\Delta v_1 = \Delta v_{m_2+1} = 0$. Inserting the approximations (8.4.2) into (8.4.1) leads to a non-linear system of ODEs.

As a final step, the system of ODEs (8.4.1) is discretized in time with the HV scheme and an inner iteration to handle the non-linearity, cf. e.g. Section 7.7 and [64]. By applying the HV scheme, the conditional expectations (8.4.2) are naturally replaced by their fully discrete versions

$$\mathbb{E}_{n,j} = \frac{\sum_{k=1}^{m_2} \psi^2(v_k) \boldsymbol{P}_{n,j,k} \frac{\Delta v_k + \Delta v_{k+1}}{2}}{\sum_{k=1}^{m_2} \boldsymbol{P}_{n,j,k} \frac{\Delta v_k + \Delta v_{k+1}}{2}}, \tag{8.4.3}$$

and we define the leverage function σ_{SLV} at the spatial and temporal grid by

$$\sigma_{\mathrm{SLV}}(x_j, \tau_n) = \frac{\sigma_{\mathrm{LV}}(x_j, \tau_n)}{\sqrt{\mathbb{E}_{n,j}}}. \tag{8.4.4}$$

It is readily seen that at the initial time level, i.e. at $n = 0$, the expression (8.4.3) is only defined if $j = j_0$. To render the calibration procedure feasible we put

$$\mathbb{E}_{0,j} = \psi^2(V_0) \quad \text{for } 1 \le j \le m_1.$$

For strictly positive time levels, i.e. for $n > 0$, the exact density $p(x_j, v_k, \tau_n)$ is always non-negative. By performing the spatial and temporal discretization, however, it is possible that some of the values $\boldsymbol{P}_{n,j,k}$ become (slightly) negative. In order to prevent the numerical solution from undesirable behaviour, the expression (8.4.3) is replaced in the calibration procedure by

$$\mathbb{E}_{n,j} = \frac{\sum_{k=1}^{m_2} \psi^2(v_k) |\boldsymbol{P}_{n,j,k}| \frac{\Delta v_k + \Delta v_{k+1}}{2}}{\sum_{k=1}^{m_2} |\boldsymbol{P}_{n,j,k}| \frac{\Delta v_k + \Delta v_{k+1}}{2}} \quad \text{for } 1 \le j \le m_1,\ n > 0. \tag{8.4.5}$$

Note that this approach, involving absolute values, is slightly different than our approach in the previous chapter. If some of the $\overline{\boldsymbol{P}}_{j,k}(\tau)$ in (7.5.6) are negative, the numerator and denominator in the pertinent expression can still be positive and Theorem 7.5.1 remains valid. Applying absolute values in such a situation would lead to the unfavourable result that equality (7.5.7) no longer holds.

Theoretically it is possible that the denominator (and hence also the numerator) of (8.4.5) equals zero and the fully discrete conditional expectation is undefined. In this case we assume that the conditional expectation is locally constant in time and set $\mathbb{E}_{n,j} = \mathbb{E}_{n-1,j}$.

Let $Q \ge 1$ be a given integer. For the actual calibration of the SLV model to the LV model, we employ the following numerical procedure.

```
for n is 1 to N do

    let P_n = P_{n-1} be an initial approximation to P(τ_n);

    for q is 1 to Q do

    (a) approximate E[ψ²(V_{τ_n}) | X_{τ_n} = x_j] by (8.4.5);
    (b) Define σ_SLV(·, τ_n) on the grid in the x-direction by
        formula (8.4.4);
    (c) update P_n by performing a numerical time step for (8.4.1)
        from τ_{n-1} to τ_n;

    end

end
```

Whenever a time step from τ_{n-1} to τ_n with the HV scheme is replaced by two half-time steps of the implicit Euler scheme, the inner iteration above is first performed for the substep from τ_{n-1} to $\tau_{n-1/2} = \tau_{n-1} + \Delta\tau/2$, yielding an approximation of $P(\tau_n - \Delta\tau/2)$ and $\sigma_{\mathrm{SLV}}(\cdot, \tau_n - \Delta\tau/2)$. Next, the inner iteration is performed for the substep from $\tau_n - \Delta\tau/2$ to τ_n, yielding an approximation of $P(\tau_n)$ and $\sigma_{\mathrm{SLV}}(\cdot, \tau_n)$. Upon completion of the time stepping and

iteration procedure above, the original approximation for $\sigma_{\mathrm{SLV}}(\cdot, 0)$ is replaced on the grid in the x-direction by $\sigma_{\mathrm{SLV}}(\cdot, \tau_1)$. This appears more realistic as the original approximation was actually only valid for the index $j = j_0$.

8.5. Numerical Experiments

In this section, the effectiveness of the calibration procedure is illustrated by applying it to a practical example. Here, we opt to consider the popular and challenging Heston-based SLV model, i.e. SLV model (8.1.1) with $\psi(v) = \sqrt{v}$ and $\alpha = 1/2$, to describe the evolution of the EUR/USD exchange rate.

As stated in the introduction of the chapter, in financial practice it is common to first determine the SV parameters of the underlying SV model and to define the LV function such that the LV model reproduces the known market prices for European call and put options. Afterwards, the calibration procedure aims at matching the SLV model with its underlying LV model, i.e. at obtaining equality (8.1.2). We assume that the SV parameters and the LV function are known and we then apply the calibration procedure from Section 8.4. In this chapter, the performance is illustrated by comparing the numerical densities stemming from the LV model and from the SLV model, and by comparing the corresponding fair values of European call options.

For the actual experiments we consider the following sets of SV parameters:

	κ	η	ξ	ρ	T	V_0	$\mathfrak{f}$
Set E	5	0.16	0.9	0.1	0.25	0.0625	0.98
Set F	1.15	0.0348	0.39	−0.64	0.25	0.0348	−0.47
Set G	1.50	0.0154	0.24	−0.11	1	0.0154	−0.20

Table 8.3: SV parameter sets for the SLV calibration experiments.

The Sets E and F correspond with the SV parameters from Sets C and D, and are taken from [19]. Set G is taken from [8] and corresponds to the EUR/USD exchange rate for a maturity of $T = 1$ (market data as of 16 September 2008). For Set E it holds that $\mathfrak{f} = \frac{2\kappa\eta}{\xi^2} - 1 = 0.98$ and the process V_τ is strictly positive. For Set F, respectively Set G, it holds that $\mathfrak{f} = -0.47$, respectively $\mathfrak{f} = -0.20$, such that $V_\tau = 0$ is attainable. The LV model is completely determined by the LV function, the risk-free interest rates and the spot value S_0. We assume that the risk-free interest rates are given by

$$r_d = 0.02, \quad r_f = 0.01,$$

and that the LV function is as displayed in Figure 8.12. The pertinent LV function originates from actual EUR/USD vanilla option data (market data as of 2 March 2016) and is constructed by using an SSVI-type method for implied volatility interpolation, [21]. The corresponding spot rate is given by

$$S_0 = 1.08815.$$

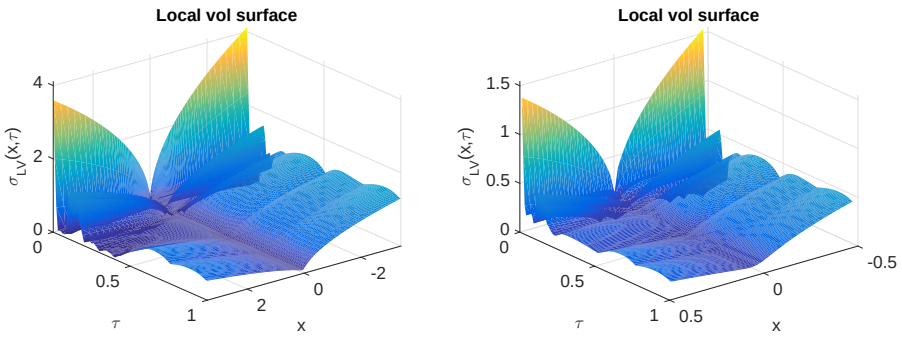


Figure 8.12: Local volatility function originating from actual EUR/USD vanilla option data (market data as of 2 March 2016) on the full domain in the x-direction (left) and on a subdomain around the spot rate (right). The spot rate $S_0 = 1.08815$.

Note that, even if the LV function is given, there is often no analytical expression available for the density function p_{LV} or the option values. It is well-known, see e.g. Chapter 7, that the density function satisfies the one-dimensional forward Kolmogorov equation

$$\frac{\partial}{\partial \tau} p_{\mathrm{LV}} = \frac{\partial}{\partial x^2}\left(\tfrac{1}{2}\sigma_{\mathrm{LV}}^2 p_{\mathrm{LV}}\right) - \frac{\partial}{\partial x}\left((r_d - r_f - \tfrac{1}{2}\sigma_{\mathrm{LV}}^2) p_{\mathrm{LV}}\right), \tag{8.5.1}$$

for $x \in \mathbb{R}, \tau > 0$. By applying the FV discretization described in Subsection 8.2.2 one defines approximations $P_{\mathrm{LV},j}(\tau)$ of the exact density values $p_{\mathrm{LV}}(x_j, \tau)$. Fully discrete approximations $P_{\mathrm{LV},N,j}$ of $p_{\mathrm{LV}}(x_j, T)$ are then obtained by applying a suitable time stepping method. In Figure 8.13 the latter approximations are shown for $\tau = 0.25$, respectively $\tau = 1$, and the practical value $m = 400$.

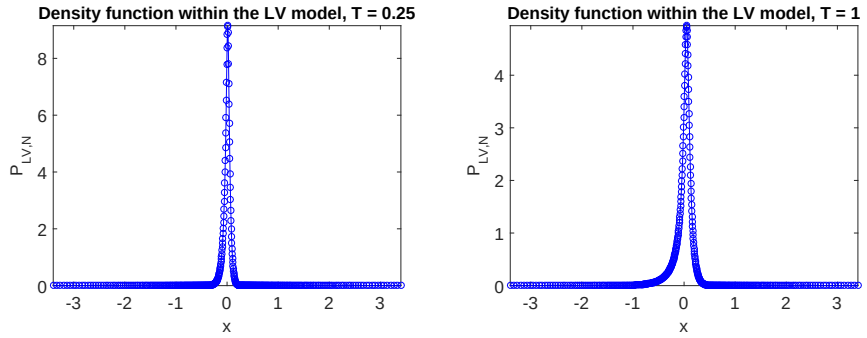


Figure 8.13: Approximation of the density function $p_{\mathrm{LV}}(x, 0.25)$ (left) and $p_{\mathrm{LV}}(x, 1)$ (right) by applying the FV discretization from Subsection 8.2.2 with $m = 400$.

Once the underlying SV model and LV model are specified, the calibration procedure from Section 8.4 can be applied. For the actual experiments we

consider the discretization parameters

$$m_1 = 400, \quad m_2 = 200, \quad \Delta\tau = 1/200, \quad \theta = \tfrac{1}{2} + \tfrac{1}{6}\sqrt{3}, \quad Q = 2.$$

In Figure 8.14 the resulting discrete leverage function is shown for set G. In order to illustrate the performance of the calibration, we first consider a discrete version of (8.1.2). More precisely, the discrete numerical densities $P_{\mathrm{LV},N,j}$ from the LV model are compared with

$$P_{\mathrm{SLV},N,j} := \sum_{k=1}^{m_2} \boldsymbol{P}_{N,j,k} \tfrac{\Delta v_k + \Delta v_{k+1}}{2},$$

which can be seen as the fully discrete approximations of

$$\int_0^\infty p(x_j, v, T)dv,$$

within the SLV model after applying the trapezoidal rule for numerical integration.

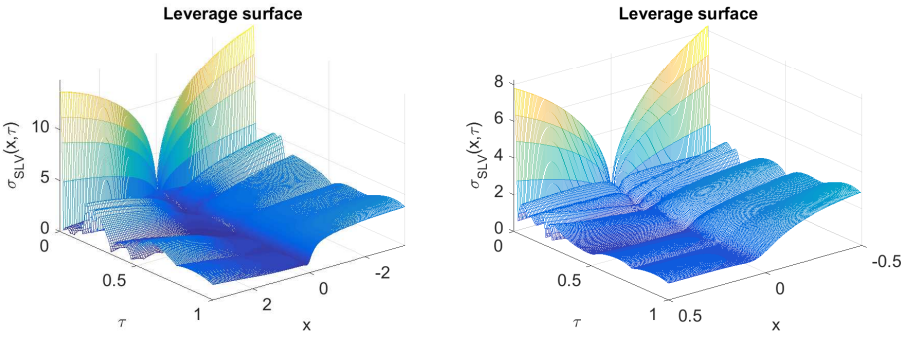


Figure 8.14: Leverage function on the full domain in the x-direction (left) and on a subdomain around the spot rate (right), stemming from the calibration procedure with local volatility function from Figure 8.12, SV parameters from Set G and values $m_1 = 400$, $m_2 = 200$, $\Delta\tau = 1/200$, $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$, $Q = 2$.

In the left plots of Figure 8.15 the approximations $P_{\mathrm{SLV},N}$ are shown for each of the three sets of parameters. In the right plots, the corresponding differences $P_{\mathrm{LV},N} - P_{\mathrm{SLV},N}$ are plotted. Note that the final time $T = 0.25$ for Set E and Set F, and $T = 1$ for Set G. From Figure 8.15 it is readily seen that the difference between the fully discrete numerical densities is very small and hence that the calibration procedure performs well.

The main goal of the calibration procedure is to define the leverage function in such a way that the LV model and the SLV model define the same fair values for non-path-dependent European options. If the leverage function is defined by (8.1.3), then it follows for the fair value (FaV) of such an option with payoff $u_0(x), x \in \mathbb{R}$, at maturity T that

$$\mathrm{FaV} = e^{-r_d T} \int_{-\infty}^{\infty} p_{\mathrm{LV}}(x, T) u_0(x) dx = e^{-r_d T} \int_0^\infty \int_{-\infty}^{\infty} p(x, v, T) u_0(x) dx dv.$$

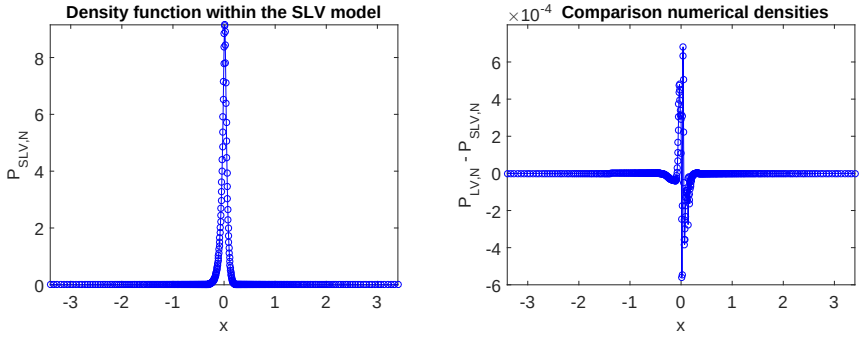


Illustration of the density in case of Set E.

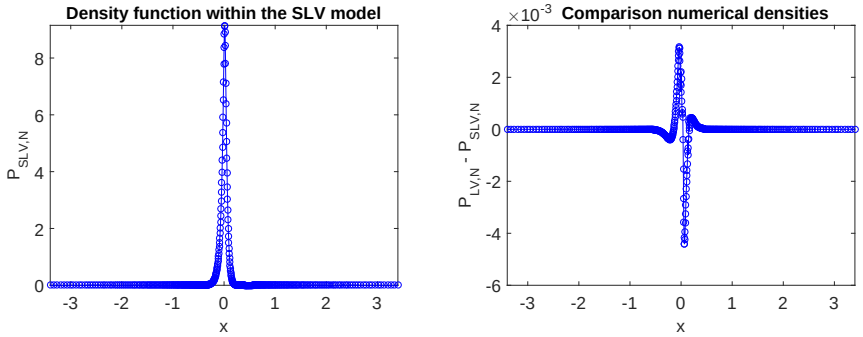


Illustration of the density in case of Set F.

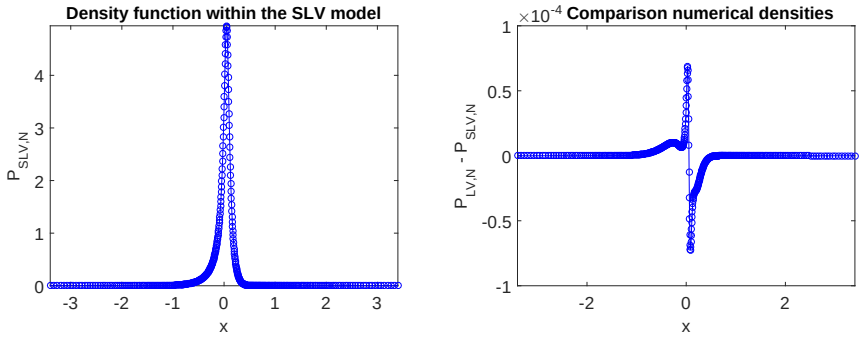


Illustration of the density in case of Set G.

Figure 8.15: Comparison of the fully discrete density functions $P_{\mathrm{LV},N}$ and $P_{\mathrm{SLV},N}$ for each of the parameters sets and for values $m_1 = 400$, $m_2 = 200$, $\Delta\tau = 1/200$, $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$, $Q = 2$.

Given the approximations $P_{\mathrm{LV},N}$ and $\boldsymbol{P}_N$, the pertinent fair value can easily be approximated by applying numerical integration with the trapezoidal rule. In case of the SLV model, it is readily seen that defining the approximated fair value via $\boldsymbol{P}_N$ and the trapezoidal rule is equivalent with defining the fair value via $P_{\mathrm{SLV},N}$ and the trapezoidal rule. Denote by $\mathrm{FaV}_{\mathrm{LV}}$, respectively FaV_{SLV}, the approximated fair values obtained via $P_{\mathrm{LV},N}$, respectively $P_{\mathrm{SLV},N}$. We now compare these approximations for a set of European call options with a range of strikes given by

$$K = 0.75S_0,\ 0.8S_0,\ 0.9S_0,\ S_0,\ 1.1S_0,\ 1.2S_0,\ 1.25S_0.$$

When the strike increases relatively to S_0, the fair value of European call options tends to zero and it is difficult to adequately compare approximations.

	$T = 0.25$	Set E	Set F	$T = 1$	Set G
K/S_0	$\sigma_{\mathrm{imp,LV}}$	ϵ_{imp}	ϵ_{imp}	$\sigma_{\mathrm{imp,LV}}$	ϵ_{imp}
0.75	19.18	0.1005	0.1208	21.94	0.0021
0.80	18.40	0.0212	0.0454	20.20	0.0015
0.90	15.01	0.0033	0.0154	16.65	0.0008
1.0	11.26	0.0011	0.0030	13.14	0.0004
1.10	11.59	0.0011	0.0153	11.38	0.0003
1.20	13.20	0.0009	0.0937	11.77	0.0003
1.25	14.03	0.0006	0.1888	12.12	0.0003

Table 8.4: Comparison of the approximated implied volatilities $\sigma_{\mathrm{imp,LV}}$ and $\sigma_{\mathrm{imp,SLV}}$ for values $m_1 = 400$, $m_2 = 200$, $\Delta\tau = 1/200$, $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$, $Q = 2$.

In financial practice, European call and put options are often quoted in terms of implied volatility. Let $\sigma_{\mathrm{imp,LV}}$, respectively $\sigma_{\mathrm{imp,SLV}}$, denote the implied volatility (in %) corresponding to $\mathrm{FaV}_{\mathrm{LV}}$, respectively $\mathrm{FaV}_{\mathrm{SLV}}$. In the following we test the performance of the calibration procedure by calculating the absolute implied volatility errors

$$\epsilon_{\mathrm{imp}} = |\sigma_{\mathrm{imp,LV}} - \sigma_{\mathrm{imp,SLV}}|.$$

In Table 8.4 these errors are presented for the different SV parameter sets, taking the same values of $m, m_1, m_2, \Delta\tau, \theta, Q$ as above. The somewhat larger values ϵ_{imp} for $T = 0.25$ compared to $T = 1$ can be explained from the fact that the implied volatility is more sensitive to changes in the fair value when the maturity is low. The results in Table 8.4 confirm that the calibration procedure performs well. They indicate that the fully discrete leverage surface is, indeed, defined such that the SLV model reproduces accurately the known market prices for European call options.

8.6. Comparison of the Calibration Methods

In Chapter 7, the adjoint calibration procedure is based on Theorem 7.5.1 which creates an exact calibration at the semidiscrete level. The implied volatility

errors $\epsilon_{\text{SLVB}}, \epsilon_{\text{SLVF}}$ in the pertinent chapter are mainly dependent on the error introduced by numerically solving the non-linear system of ODEs (7.4.3) with coefficients defined by (7.5.6). Calibration of the implied volatilities $\sigma_{\text{imp,SLVB}}$, $\sigma_{\text{imp,SLVF}}$ to the market implied volatility can be decoupled into two separate steps. First, the number of spatial and temporal grid points can be chosen such that the error introduced by discretization of the one-dimensional backward Kolmogorov equation (7.2.7), i.e. the error in the fully discrete LV model, is sufficiently small. In a second step, one can choose a temporal step size $\Delta\tau$ and a number of iterations Q such that the implied volatility errors $\epsilon_{\text{SLVB}}, \epsilon_{\text{SLVF}}$ are sufficiently small.

In the current chapter, there is no direct relationship between the implied volatilities $\sigma_{\text{imp,LV}}$ and $\sigma_{\text{imp,SLV}}$. They both form approximations to the market implied volatility, but the FV discretization of the one-dimensional PDE (8.5.1) is not used explicitly for the FV calibration of the SLV model. To the best of our knowledge, neither the implied volatility error ϵ_{imp}, nor the difference between $\sigma_{\text{imp,SLV}}$ and the market implied volatility, can be decoupled. They are dependent on the accuracy of the discretization of the one-dimensional forward Kolmogorov equation, on the accuracy of the discretization of the two-dimensional forward Kolmogorov equation, and on the error introduced by handling the non-linearity in the system of ODEs (8.4.1) with coefficients given by (8.4.2). This constitutes a disadvantage in comparison with the adjoint calibration method.

The decoupling in the calibration method from Chapter 7 has the additional advantage that, for non-path-dependent options, the approximations of the fair value under the LV model and under the SLV model are the same up to a small temporal discretization error. As such, one can use the one-dimensional LV model for the (very) fast valuation of vanilla options consistently with the two-dimensional SLV model. The latter model can be used for the correct valuation of path-dependent options.

By using the adjoint spatial discretization, the difference between the approximated fair value obtained by numerically solving the forward equation and the one obtained by discretization of the backward equation, is of the (very small) size of the temporal discretization error. This is a useful property since the solution of the forward Kolmogorov equation can be used very efficiently for the valuation of a non-path-dependent option for a range of strikes, whilst the backward equation is often used to calculate the Greeks, i.e. the sensitivity of the option value to its underlying variables. By matching the approximations, the numerical solutions obtained via the forward and backward Kolmogorov equation can be interchanged. In the current Chapter 8, the backward Kolmogorov equation is never used and such a property is not applicable.

The numerical experiments in Section 8.2 reveal that the FV spatial discretization method is convergent with respect to the pertinent initial-boundary value problems. If the boundary conditions are smooth, then the FV method shows second order convergence behaviour, see e.g. the experiments for the one-dimensional and two-dimensional Black–Scholes model and the experiments for the CIR and Heston model with $\mathfrak{f} \geq 0$. Next, consider the adjoint spatial discretization method for the numerical solution of forward Kolmogorov

equations. Recall that this adjoint discretization is completely defined by the spatial discretization of the corresponding backward equation. To the best of our knowledge, there is no clear relationship between the convergence properties of the original spatial discretization method and the convergence properties of the corresponding adjoint spatial discretization.

In our opinion, the adjoint method from Chapter 7 is preferable for the calibration of SLV models and the valuation of financial options. However, if one is interested in the properties of the numerical solution of the forward Kolmogorov equation, then the FV spatial discretization method is more appropriate.

8.7. Conclusion

Stochastic local volatility models constitute state-of-the-art models to describe asset price processes. Their calibration to the underlying local volatility model is, however, highly non-trivial. It incorporates the solution of non-linear forward Kolmogorov equations. In general, no analytical solution is available and one relies on numerical methods in order to approximate the exact solution. Here, we introduce a FV-ADI method for the numerical solution of general one-dimensional and two-dimensional forward Kolmogorov equations. The FV spatial discretization does not require a transformation of the PDE, which constitutes a main advantage in the calibration of SLV models, and handles the boundary conditions in a natural way. Moreover, the FV scheme preserves the crucial property that the total mass of a density function is always equal to one. Our numerical experiments for relevant practical applications confirm that the pertinent spatial discretization is convergent. Temporal discretization is performed by using the HV scheme and the non-linearity in the calibration procedure of stochastic local volatility models is handled by introducing an inner iteration. Our numerical experiments reveal that the proposed calibration procedure performs well. The calibrated stochastic local volatility model matches the underlying local volatility model, both in terms of the density function and of the implied volatilities of European call options.

Comparing the FV-ADI method with the adjoint method from Chapter 7, we prefer the former one for the numerical solution of forward Kolmogorov equations. For the calibration of SLV models and the valuation of financial options, we prefer the adjoint method.

CHAPTER 9

Conclusions & Outlook

9.1. Conclusions

In this thesis a convergence analysis has been presented for four ADI time stepping methods adapted to mixed spatial derivative terms that are widely used for the numerical solution of partial differential equations (PDEs) from financial mathematics. More precisely, we considered convergence of the Douglas (Do) scheme, the Craig–Sneyd (CS) scheme, the Modified Craig–Sneyd (MCS) scheme and the Hundsdorfer–Verwer (HV) scheme, in application to semidiscretized two-dimensional time-dependent convection-diffusion equations with mixed derivative term. Subsequently, it was demonstrated that the ADI schemes can be very useful for the fast, stable and accurate calibration of stochastic local volatility (SLV) models. We proposed two techniques for the calibration of SLV models that are new in the literature. Ample numerical experiments indicate that both techniques perform well.

The preliminary Chapter 2 dealt with the first step of the method-of-lines, i.e. with spatial discretization. We introduced the smooth, non-uniform Cartesian grids and second order finite difference (FD) schemes that are used throughout the thesis. Application of the FD schemes for semidiscretization of initial-boundary value problems for general one-dimensional and two-dimensional convection-diffusion equations leads to large systems of ordinary differential equations (ODEs). An analysis of the sparsity structure of the semidiscretization matrices revealed that, if the semidiscrete system is stemming from a two-dimensional PDE, then application of standard implicit time stepping methods can lead to a computational effort that increases faster than the total number of spatial grid points.

In the preliminary Chapter 3 the four ADI time stepping schemes were formally introduced. They employ a splitting of the semidiscrete operator in the different spatial dimensions, which can lead to a major computational advantage. We presented an overview of the existing stability and consistency results relevant to semidiscretized two-dimensional convection-diffusion equations. It is shown that the analysis of the local discretization errors in [34] leads to a convergence result for the Do scheme.

In Chapter 4 we proved a first convergence result for the CS scheme and the MCS scheme. Considering a perturbed version of the MCS scheme led to

a recursion formula for the total error. By Taylor expansion, expressions for the local errors in the perturbed scheme have been obtained. We split the local discretization error so that each component allowed application of a key lemma from [33]. Under natural stability and smoothness assumptions, this resulted in a second order convergence theorem. The convergence result has the important property that it holds uniformly in the arbitrarily small spatial mesh width. Positive results on the stability assumptions have been obtained in the von Neumann framework.

A convergence result for the HV scheme has been obtained in Chapter 5. As in the previous chapter, it was shown that the total error satisfies a recursion formula and expressions for the local errors were derived by Taylor expansion. A subtle splitting of the local discretization error was used to prove a second order convergence theorem under natural stability and smoothness assumptions. As before, we obtained positive results on the stability assumptions in the von Neumann framework.

A motivating example next showed that convergence of ADI schemes can be seriously impaired if the initial data are non-smooth. In Chapter 6 we analysed the positive effect of Rannacher time stepping, i.e. replacing the first N_0 ADI time steps by $2N_0$ half-time steps of the implicit Euler scheme, on the order of convergence of the ADI schemes when they are applied for the numerical solution of a model two-dimensional convection-diffusion equation with constant coefficients and provided with Dirac delta initial data. A discrete/continuous Fourier transformation led to the important insight that, for every ADI scheme, the total discretization error can be written as the sum of a low-wavenumber error and a high-wavenumber error. An asymptotic analysis for the MCS scheme (and CS scheme) revealed that its low-wavenumber error decreases in a second order fashion as a function of the temporal step size. The order of the high-wavenumber error is $2N_0 - 2$. Based on ample numerical experiments we conjectured that for the Do scheme, respectively the HV scheme, the order of the low-wavenumber error is equal to the classical order of consistency of the ADI scheme, and that the order of the high-wavenumber error is $2N_0 - 2$. The value $N_0 = 2$ is then a lower bound on N_0 for the Rannacher time stepping in order to ensure convergence of the numerical solution to the exact solution. A brief consideration of alternative initial data revealed that the order of the high-wavenumber error is mainly dependent on N_0 and the degree of smoothness of the initial data. Finally, our analysis showed that for each of the ADI schemes it seems favourable to consider smaller values of the ADI parameter θ.

In Chapter 7 an adjoint method has been introduced for the exact calibration of SLV models to the underlying local volatility (LV) model. Given a spatial discretization for the backward Kolmogorov equation, we defined an adjoint spatial discretization for the corresponding forward Kolmogorov equation such that both discretizations yield the same approximation for the fair value of non-path-dependent options. It was shown that, if similar spatial discretizations are considered for the backward equation stemming from the LV model and for the backward equation stemming from the SLV model, then the adjoint spatial discretization can be used to create an exact match between the semidiscretized LV model and the semidiscretized SLV model. Since there is

often no analytical solution available for the fair value of vanilla options under (S)LV models, this is the best result one can aim for. The adjoint spatial discretization that is used for the exact calibration leads to a large system of non-linear ODEs. The MCS scheme proved to be very useful for the efficient numerical solution of this system of ODEs. An inner iteration was described to handle the non-linearity.

A common approach for the calibration of SLV models is numerically solving the associated two-dimensional non-linear forward Kolmogorov equation. In Chapter 8 we presented a new finite volume (FV) method for the spatial discretization of general one-dimensional and two-dimensional forward Kolmogorov equations. The FV scheme is mass-conservative. This is a key property since the solution of these convection-diffusion equations represents a density function. Moreover, the FV method does not require a transformation of the PDE. This constitutes a major advantage in comparison with standard FV methods as the PDE coefficients are often non-smooth. Application of the FV scheme for the calibration of SLV models to the underlying LV model led to a large system of non-linear ODEs. In Chapter 8 we demonstrated that the HV scheme is very useful for the efficient numerical solution of such systems of ODEs. As before, the non-linearity was handled by an iteration procedure.

Ample numerical experiments exemplified that both calibration techniques are promising. For the adjoint technique we showed that the calibration error can be decoupled and that each part can be controlled. A similar property does not hold for the FV method. In our opinion, the adjoint calibration technique is more appropriate for the calibration of SLV models to the underlying LV model. If the objective is numerically solving the forward Kolmogorov equation, then we prefer the FV spatial discretization.

9.2. Outlook

In this thesis we have derived a variety of results that provide important new insight in the convergence properties and applicability of ADI schemes. It is inherent to research that every answer raises new questions. Completing the research in this PhD thesis we arrived at a number of interesting problems and ideas for future research.

The use of Cartesian grids for the numerical solution of higher-dimensional PDEs can lead to an enormous amount of spatial grid points, even if the number of points in each spatial direction is small. This is known as *the curse of dimensionality* and forms an important topic in computational finance. Sparse grid methods reduce the total number of spatial grid points and can be promising.

Three- and four-dimensional time-dependent convection-diffusion equations are becoming more common in financial mathematics and their approximate solution is often obtained via the numerical techniques described in this thesis. Applying the ADI schemes leads to a computational advantage in comparison with classical implicit time stepping schemes. The advantage is often more prominent than in the two-dimensional case. To the best of our knowledge, there are no second order convergence results available for the CS scheme, the

MCS scheme and the HV scheme if the number of spatial dimensions $l \geq 3$. A convergence analysis is, however, imperative. It provides a basis for the schemes being used in practice.

The analysis in [34], Chapter 4 and Chapter 5 can be used as a starting point for a convergence analysis relevant to higher-dimensional PDEs provided with smooth initial and boundary-data. The expressions for the total error and the local discretization error derived in the pertinent literature are valid for a general number of spatial directions. In order to arrive at a useful second order convergence result for the (M)CS scheme and the HV scheme with $l \geq 3$, a new ingenious splitting of the local discretization error can be performed. Based on the second order convergence result in [34] for the Do scheme with $l = 3$ and $F_0 = 0$, we foresee a mild restriction on the temporal step size.

The adverse effect of non-smooth initial data on the convergence of ADI schemes is also present in higher-dimensional applications. Consider for example a l-dimensional model convection-diffusion equation with constant coefficients, provided with an initial function that is the product of Dirac delta functions in each of the l spatial dimensions. Comparing the results in [22] for the Crank-Nicolson scheme in one spatial dimension and our results in Chapter 6 for the ADI schemes in two spatial dimensions, we expect that for larger values of l a larger value N_0 for the Rannacher time stepping is needed to ensure convergence of the numerical solution the exact solution. A theoretical convergence analysis can be performed by extending the discrete/continuous Fourier transformation to a higher number of spatial dimensions.

In this thesis it has been assumed that all risk-free interest rates are constant. Considering SLV models with stochastic interest rates for the modelling of foreign exchange rates naturally leads to four-dimensional forward and backward Kolmogorov equations. Under some assumptions, an expression for the leverage function that calibrates a four-factor SLV model exactly to the underlying LV model has been obtained in [10]. Determining this leverage function for a four-factor SLV model is, however, inherently more difficult than for the corresponding SLV model with deterministic interest rates. It is interesting to examine whether the adjoint spatial discretization can be used to create an exact match between the semidiscretized SLV model with stochastic interest rates and the semidiscretized LV model. Alternatively, our FV spatial discretization can be extended for the numerical solution of a four-dimensional forward Kolmogorov equation. Its numerical solution is useful for the approximation of the leverage function that calibrates the SLV model exactly.

The existing stability results for ADI schemes relevant to higher-dimensional pure diffusion equations with mixed derivative terms reveal that a larger number of spatial directions l often leads to a stronger restriction on the ADI parameter θ. Moreover, the number of linear systems that has to be solved in each time step of the ADI scheme is directly proportional with l. For higher-dimensional PDEs it might be interesting to split the semidiscrete operator into suboperators that represent spatial derivatives in multiple spatial dimensions. For example, the semidiscrete operator stemming from a four-dimensional convection-diffusion equation could be split into an operator F_0, representing all the mixed spatial derivative terms, and suboperators F_1, F_2 that each represent unidirectional spatial derivatives in two spatial dimensions.

The total number of spatial dimensions then no longer corresponds with the number of suboperators that are handled implicitly. A stability and convergence analysis relevant to this situation is lacking in the literature. It can be investigated whether the results from Chapter 4 and Chapter 5 are applicable. In the case of non-smooth (Dirac delta) initial data, we conjecture that the lower bound on N_0 for the Rannacher time stepping in order to ensure convergence of the numerical solution to the exact solution, is mainly dependent on the number of spatial directions and not on the number of suboperators.

Bibliography

[1] L. B. G. Andersen. Simple and efficient simulation of the Heston stochastic volatility model. *Journal of Computational Finance*, 11:1–42, 2008.

[2] L. B. G. Andersen and V. V. Piterbarg. Moment explosions in stochastic volatility models. *Finance and Stochastics*, 11:29–50, 2007.

[3] L. B. G. Andersen and V. V. Piterbarg. *Interest Rate Modeling, Volume I: Foundations and Vanilla Models*. London: Atlantic Financial Press, 2010.

[4] J. Andreasen and B. Huge. Volatility interpolation. *Risk*, March:76–79, 2011.

[5] T. Björk. *Arbitrage Theory in Continuous Time*. Oxford: Oxford University Press, 1998.

[6] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–654, 1973.

[7] P. L. T. Brian. A finite-difference method of high-order accuracy for the solution of three-dimensional transient heat conduction problems. *AIChE Journal*, 7:367–370, 1961.

[8] I. J. Clark. *Foreign Exchange Option Pricing: A Practitioner's Guide*. Chichester: John Wiley & Sons, 2011.

[9] J. C. Cox, J. E. Ingersoll, and S. A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53:385–407, 1985.

[10] A. Cozma, M. Mariapragassam, and C. Reisinger. Calibration of a four-factor hybrid local-stochastic-volatility model with a new control variate particle method, 2017. Available at `https://arxiv.org/abs/1701.06001`.

[11] I. J. D. Craig and A. D. Sneyd. An alternating-direction implicit scheme for parabolic equations with mixed derivatives. *Computers and Mathematics with Applications*, 16:341–350, 1988.

[12] J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, 43:50–67, 1947.

[13] J. Douglas. Alternating direction methods for three space variables. *Numerische Mathematik*, 4:41–63, 1962.

[14] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82:421–439, 1956.

[15] B. Dupire. Pricing with a smile. *Risk*, January:18–20, 1994.

[16] E. Ekström and J. Tysk. Boundary conditions for the single-factor term structure equation. *The Annals of Applied Probability*, 21:332–350, 2011.

[17] B. Engelmann, F. Koster, and D. Oeltz. Calibration of the Heston stochastic local volatility model: a finite volume scheme, 2012. Available at SSRN 1823769.

[18] F. Fang and C. W. Oosterlee. A novel pricing method for European options based on Fourier-cosine series expansions. *SIAM Journal on Scientific Computing*, 31:826–848, 2008.

[19] F. Fang and C. W. Oosterlee. A Fourier-based valuation method for Bermudan and barrier options under Heston's model. *SIAM Journal on Financial Mathematics*, 2:439–463, 2011.

[20] W. Feller. Two singular diffusion problems. *Annals of Mathematics*, 54:173–182, 1951.

[21] J. Gatheral and A. Jacquier. Arbitrage-free SVI volatility surfaces. *Quantitative Finance*, 14:59–71, 2014.

[22] M. B. Giles and R. Carter. Convergence analysis of Crank–Nicolson and Rannacher time-marching. *Journal of Computational Finance*, 9:89–112, 2006.

[23] I. Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an Ito differential. *Probability Theory and Related Fields*, 71:501–516, 1986.

[24] T. Haentjens. Efficient and stable numerical solution of the Heston–Cox–Ingersoll–Ross partial differential equation by alternating direction implicit finite difference schemes. *International Journal of Computer Mathematics*, 90:2409–2430, 2013.

[25] T. Haentjens and K. J. in 't Hout. Alternating direction implicit finite difference schemes for the Heston–Hull–White partial differential equation. *Journal of Computational Finance*, 16:83–110, 2012.

[26] T. Haentjens and K. J. in 't Hout. ADI schemes for pricing American options under the Heston model. *Applied Mathematical Finance*, 22:207–237, 2015.

[27] P. S. Hagan, D. Kumar, A. S. Lesniewski, and D. E. Woodward. Managing smile risk. *Wilmott Magazine*, January:84–108, 2002.

[28] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II.* Berlin: Springer, 1996.

[29] P. Henry-Labordère. Calibration of local stochastic volatility models to market smiles. *Risk*, September:112–117, 2009.

[30] S. L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343, 1993.

[31] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge: Cambridge University Press, 1991.

[32] J. Hull. *Options, Futures, & Other Derivatives.* New Jersey: Prentice Hall, 5th edition, 2002.

[33] W. Hundsdorfer. Unconditional convergence of some Crank–Nicolson LOD methods for initial-boundary value problems. *Mathematics of Computation*, 58:35–53, 1992.

[34] W. Hundsdorfer. Accuracy and stability of splitting with stabilizing corrections. *Applied Numerical Mathematics*, 42:213–233, 2002.

[35] W. Hundsdorfer and J. G. Verwer. *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations.* Berlin: Springer, 2003.

[36] S. Ikonen and J. Toivanen. Componentwise splittingmethods for pricing American options under stochastic volatility. *International Journal of Theoretical and Applied Finance*, 10:331–361, 2007.

[37] K. J. in 't Hout and S. Foulon. ADI finite difference schemes for option pricing in the Heston model with correlation. *International Journal of Numerical Analysis and Modeling*, 7:303–320, 2010.

[38] K. J. in 't Hout and C. Mishra. A stability result for the modified Craig–Sneyd scheme applied to 2D and 3D pure diffusion equations. *AIP Conference Proceedings*, 1281:2029–2032, 2010.

[39] K. J. in 't Hout and C. Mishra. Stability of the modified Craig–Sneyd scheme for two-dimensional convection-diffusion equations with mixed derivative term. *Mathematics and Computers in Simulation*, 81:2540–2548, 2011.

[40] K. J. in 't Hout and C. Mishra. Stability of ADI schemes for multidimensional diffusion equations with mixed derivative terms. *Applied Numerical Mathematics*, 74:83–94, 2013.

[41] K. J. in 't Hout and B. D. Welfert. Stability of ADI schemes applied to convection-diffusion equations with mixed derivative terms. *Applied Numerical Mathematics*, 57:19–35, 2007.

[42] K. J. in 't Hout and B. D. Welfert. Unconditional stability of second-order ADI schemes applied to multi-dimensional diffusion equations with mixed derivative terms. *Applied Numerical Mathematics*, 59:677–692, 2009.

[43] K. J. in 't Hout and M. Wyns. Convergence of the Hundsdorfer–Verwer scheme for two-dimensional convection-diffusion equations with mixed derivative term. *AIP Conference Proceedings*, 1648:850054–1–850054–5, 2015.

[44] K. J. in 't Hout and M. Wyns. Convergence of the Modified Craig–Sneyd scheme for two-dimensional convection-diffusion equations with mixed derivative term. *Journal of Computational and Applied Mathematics*, 296:170–180, 2016.

[45] A. Itkin. High-order splitting methods for forward PDEs and PIDEs. *International Journal of Theoretical and Applied Finance*, 18:1550031–1 – 1550031–24, 2015.

[46] A. Itkin. Modelling stochastic skew of FX options using SLV models with stochastic spot/vol correlation and correlated jumps, 2017. Available at `https://arxiv.org/abs/1701.02821`.

[47] A. Itkin and P. Carr. Jumps without tears: a new splitting technology for barrier options. *International Journal of Numerical Analysis and Modeling*, 8:667–704, 2011.

[48] D. Lanser, J. G. Blom, and J. G. Verwer. Time integration of the shallow water equations in spherical geometry. *Journal of Computational Physics*, 171:373–393, 2001.

[49] A. Lipton. *Mathematical Methods for Foreign Exchange*. Singapore: World Scientific, 2001.

[50] A. Lipton. The vol smile problem. *Risk*, February:61–65, 2002.

[51] S. McKee and A. R. Mitchell. Alternating direction methods for parabolic equations in two space dimensions with a mixed derivative. *The Computer Journal*, 13:81–86, 1970.

[52] S. McKee, D. P. Wall, and S. K. Wilson. An alternating direction implicit scheme for parabolic equations with mixed derivative and convective terms. *Journal of Computational Physics*, 126:64–76, 1996.

[53] C. Mishra. *Stability of Alternating Direction Implicit Schemes with Application to Financial Option Pricing Equations*. PhD thesis, University of Antwerp, 2014.

[54] C. Mishra. A new stability result for the modified Craig–Sneyd scheme applied to two-dimensional convection-diffusion equations with mixed derivatives. *Applied Mathematics and Computation*, 285:41–50, 2016.

[55] D. W. Peaceman and H. H. Rachford. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics*, 3:28–41, 1955.

[56] D. M. Pooley, K. R. Vetzal, and P. A. Forsyth. Convergence remedies for non-smooth payoffs in option pricing. *Journal of Computational Finance*, 6:25–40, 2003.

[57] R. Rannacher. Finite element solution of diffusion problems with irregular data. *Numerische Mathematik*, 43:309–327, 1984.

[58] Y. Ren, D. Madan, and M. Q. Qian. Calibrating and pricing with embedded local volatility models. *Risk*, September:138–143, 2007.

[59] H. Risken. *The Fokker-Planck Equation: Methods of Solution and Applications*. Berlin: Springer, 2nd edition, 1989.

[60] M. J. Ruijter and C. W. Oosterlee. Two-dimensional Fourier cosine series expansion method for pricing financial options. *SIAM Journal on Scientific Computing*, 34:B642–B671, 2012.

[61] L. O. Scott. Option pricing when the variance changes randomly: theory, estimation, and an application. *The Journal of Financial and Quantitative Analysis*, 22:419–438, 1987.

[62] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Belmont: Wadsworth Publishing Company, 1989.

[63] R. Tachet. *Non-Parametric Model Calibration in Finance*. PhD thesis, Ecole Centrale Paris, 2011.

[64] G. Tataru and T. Fisher. Stochastic local volatility. *Technical report, Quantitative Development Group, Bloomberg*, 2010.

[65] D. Tavella and C. Randall. *Pricing Financial Instruments: The Finite Difference Method*. New York: John Wiley & Sons, 2000.

[66] A. W. van der Stoep, L. A. Grzelak, and C. W. Oosterlee. The Heston stochastic-local volatility model: Efficient Monte Carlo simulation. *International Journal of Theoretical and Applied Finance*, 17:1450045–1–1450045–30, 2014.

[67] J. G. Verwer, E. J. Spee, J. G. Blom, and W. Hundsdorfer. A second-order Rosenbrock method applied to photochemical dispersion problems. *SIAM Journal on Scientific Computing*, 20:1456–1480, 1999.

[68] M. Wyns. Convergence analysis of the Modified Craig–Sneyd scheme for two-dimensional convection-diffusion equations with nonsmooth initial data. *IMA Journal of Numerical Analysis*, 37:798–831, 2017.

[69] M. Wyns and J. Du Toit. A finite volume - alternating direction implicit approach for the calibration of stochastic local volatility models. *International Journal of Computer Mathematics*, 2017. http://dx.doi.org/10.1080/00207160.2017.1297805.

[70] M. Wyns and K. J. in 't Hout. An adjoint method for the exact calibration of stochastic local volatility models. *Journal of Computational Science*, 2017. http://dx.doi.org/10.1016/j.jocs.2017.02.004.

Scientific Résumé

Education

- **2013-2017**
  PhD candidate - Mathematics, University of Antwerp, Belgium
  PhD thesis: *Convergence Analysis and Application of ADI Schemes for Partial Differential Equations from Financial Mathematics*
  Advisor: Prof. dr. Karel in 't Hout
- **2011-2013**
  Master of Science - Financial Mathematics, Vrije Universiteit Brussel, Belgium
  Master dissertation (in dutch): *Lokale stochastische volatiliteitsmodellen*
  Advisor: Prof. dr. Karel in 't Hout
- **2008-2011**
  Bachelor of Science - Mathematics, Vrije Universiteit Brussel, Belgium
- **2002-2008**
  Secondary education, Sint-Theresiacollege Kapelle-op-den-Bos, Belgium

Assistant teaching

University of Antwerp, Master of Science - Mathematics, 2013-2017:

- Finite difference methods and financial mathematics
  Academic year 2013-2014, 2016-2017
- Applications of differential equations
  Academic year 2013-2014, 2014-2015, 2015-2016
- Actuarial models
  Academic year 2016-2017

Awards and grants

- NAG Student Prize 'Direct Award' 2017 for the work on a new finite volume approach for the calibration of stochastic local volatility models.
  `https://www.nag.co.uk/content/student-awards`

Conference talks and poster presentations

- **SPRING 2017:** Spring Meeting of the Dutch-Flemish Numerical Analysis community, *An adjoint method for the calibration of SLV models*, Antwerp, Belgium, May 19, 2017.
- **MF 2017:** MathFinance Conference 2017, *ADI finite difference schemes for the calibration of stochastic local volatility models*, Frankfurt am Main, Germany, April 20-21, 2017.
- **SIAMFM 2016:** SIAM Conference on Financial Mathematics and Engineering 2016, *ADI finite difference schemes for the calibration of stochastic local volatility models*, Austin, Texas, United States, November 17-19, 2016.
- **SCFD 2016:** Student Computational Finance Day 2016, *ADI finite difference schemes for the calibration of stochastic local volatility models*, Delft, The Netherlands, May 23, 2016.
- **SEM 2015:** Seminar of the Mathematical Finance research group at the University of Manchester, *ADI schemes for pricing European options under the Heston model*, Manchester, United Kingdom, December 10, 2015.
- **WSC 2015:** 40th Woudschoten Conference of the Dutch-Flemish Numerical Analysis community: Poster presentation, *Convergence analysis of the Modified Craig–Sneyd scheme in 2 dimensions for nonsmooth initial data*, Zeist, The Netherlands, October 7-9, 2015.
- **NUMDIFF 2015:** 14th Conference on the Numerical Solution of Differential and Differential-Algebraic Equations, *Convergence analysis of the Modified Craig–Sneyd scheme for two–dimensional convection–diffusion equations with nonsmooth initial data*, Halle, Germany, September 7-11, 2015.
- **WMNFM 2015:** Workshop Models and Numerics in Financial Mathematics 2015: Poster presentation, *Convergence analysis of the Modified Craig–Sneyd scheme in 2 dimensions for nonsmooth initial data*, Leiden, The Netherlands, May 26-29, 2015.
- **ICNAAM 2014:** 12th International Conference of Numerical Analysis and Applied Mathematics, *Convergence of the Hundsdorfer–Verwer scheme for two–dimensional convection–diffusion equations with mixed derivative term*, Rhodes, Greece, September 22-28, 2014.
- **WCBFS 2014:** 8th World Congress of the Bachelier Finance Society, *Convergence of the Modified Craig–Sneyd scheme for multi–dimensional convection–diffusion equations with application to the Heston PDE*, Brussels, Belgium, June 2-6, 2014.

Further conferences and workshops attended

- **SWFI 2017:** Study Week with Financial Industry, Madrid, Spain, May 8-12, 2017.
- **QTC 2017:** 2nd Annual QuanTech Conference, London, United Kingdom, April 27-28, 2017.
- **AFMATH 2017:** Actuarial and Financial Mathematics Conference 2017, Brussels, Belgium, February 9-10, 2017.
- **WSC 2016:** 41st Woudschoten Conference of the Dutch-Flemish Numerical Analysis community, Zeist, The Netherlands, October 5-7, 2016.
- **AFMATH 2016:** Actuarial and Financial Mathematics Conference 2016, Brussels, Belgium, February 1-2, 2016.
- **SPRING 2015:** Spring Meeting of the Dutch-Flemish Numerical Analysis community, Antwerp, Belgium, May 8, 2015.
- **MF 2015:** MathFinance Conference 2015, Frankfurt am Main, Germany, March 23-24, 2015.
- **PAR 2014:** Introduction to OpenFoam, MPI and GPU programming, Delft, The Netherlands, February 24-28, 2014.
- **WSC 2013:** 38th Woudschoten Conference of the Dutch-Flemish Numerical Analysis community, Zeist, The Netherlands, October 2-4, 2013.

Publications

- [**A1**] M. Wyns and J. Du Toit. A finite volume - alternating direction implicit approach for the calibration of stochastic local volatility models. *International Journal of Computer Mathematics*, 2017. doi:10.1080/00207160.2017.1297805
- [**A1**] M. Wyns and K. J. in 't Hout. An adjoint method for the exact calibration of stochastic local volatility models. *Journal of Computational Science*, 2017. doi:10.1016/j.jocs.2017.02.004
- [**A1**] M. Wyns. Convergence analysis of the Modified Craig–Sneyd scheme for two-dimensional convection-diffusion equations with nonsmooth initial data. *IMA Journal of Numerical Analysis*, 37:798–831, 2017. doi:10.1093/imanum/drw028
- [**A1**] K. J. in 't Hout and M. Wyns. Convergence of the Modified Craig–Sneyd scheme for two-dimensional convection-diffusion equations with mixed derivative term. *Journal of Computational and Applied Mathematics*, 296:170–180, 2016. doi:10.1016/j.cam.2015.09.023

- **[P1]** K. J. IN 'T HOUT AND M. WYNS. Convergence of the Hundsdorfer–Verwer scheme for two-dimensional convection-diffusion equations with mixed derivative term.
  AIP Conference Proceedings, 1648:850054-1–850054-5, 2015.
  doi:10.1063/1.4913109