Technische Universität Chemnitz
Fakultät für Mathematik

Diploma Thesis

Pricing derivatives in stochastic volatility models using the finite difference method

submitted by:	Supervisors:
Tino Kluge	PD Dr. Thomas Apel (Technische Universität Chemnitz)
	Dr. Jürgen Hakala (Commerzbank, Quantitative Research)

Chemnitz, September 6, 2002

Contents

Preface

The topic of this thesis has its origin in a project I have worked on during an internship at Commerzbank in Frankfurt. Having been able to work in the dynamic and familiar Quantitative Research team has been a great experience for me. The objective of the project was to improve accuracy and to speed up the computation time of one option pricer for stochastic volatility markets based on the finite difference method. Mainly occupied with implementing the algorithm in a C++ environment the project still left some time for practical examinations. With help of many simulations the influence of different non uniform grids as well as different boundary conditions were studied.

This thesis aims to put some heuristics on a theoretical basis and gives further suggestions for an improvement of convergence. However, it has sometimes been quite difficult to find and acquire appropriate literature about finite differences which did not simplify the task.

I would like to take this opportunity to thank the Quantitative Research team at Commerzbank for giving me an insight in the banking business, in particular to Dr. Jürgen Hakala for providing me with important material, to Dr. Uwe Wystup for giving me the opportunity to attend conferences, to PD Dr. Thomas Apel, from Chemnitz University of Technology, for his valuable ideas and to Prof. Chris Rogers, from the University of Cambridge, for assisting in stochastic questions.

Chapter 1

Introduction

On financial markets many different products are traded. Focusing on the foreign exchange market we are basically faced with exchange rates of different currency pairs and derivative products which depend in a deterministic way on the underlying rates. The most common examples are swaps, forwards and different kind of options. At time of maturity T the value of a call option for instance is by contract equal to the value of an underlying minus a predefined so called strike value if the underlying exceeds the strike and otherwise zero. Even the trade of plain vanilla options is of high importance for the whole industry. Without going into too much detail the buyer of a call option of an exchange rate like EUR/USD protects himself from a strengthening Euro maybe because he will receive payments in US Dollar and thus hedges the currency risk. If the Euro looses value the option might become worthless but the incoming money will be worth more. The issuer which is mainly a bank on the other hand does not want to be exposed to the uncertainty and risk involved in the outcome of the deal. Therefore the bank hedges the value of the option using a portfolio consisting of the underlying and a money market account with the objective to replicate the option value at maturity T. The price of the option will be determined based on the value of the portfolio at the beginning of the contract. For an accurate assessment of the price it is essential to have a realistic stochastic model about the development of the underlying quantity. Black and Scholes [2] discussed in 1973 a simple model where the underlying obeys a geometric Brownian motion and derived option prices based on the no arbitrage principle. Their work was remarkable and the results are still widely used. Since then, several improvements and extensions were suggested in order to obtain a more realistic model and hence more accurate option prices. We concentrate on stochastic volatility models where as the name suggests the volatility is not constant like in the Black-Scholes model but a stochastic process itself. Of course there are other models which are equally important like Jump-Diffusion and Lévy processes. Further approaches also take into account which is called friction of the market meaning illiquidity effects, default risks and transaction costs.

In this thesis we examine the numerical part of option valuation and take stochastic volatility models for granted for which option prices in general are not available through an analytic formula. Besides Monte Carlo methods and evaluation using binomial or trinomial trees one can determine the price by solving a parabolic partial differential equation (p.d.e.) which is the approach we focus on. The actual price is then the solution of the p.d.e. evaluated in one particular point in space and time. In reality one needs an algorithm which computes this value as quickly as possible with an acceptable accuracy. Roughly, that means two accurate digits in less than one second. We employ the finite difference method (f.d.m.) to solve the p.d.e. which is of convection-diffusion type in two space dimensions. This numerical method has been chosen because it is simple to implement, flexible as far as modelling of different boundary conditions is concerned and offers ways to speed up the computation by using an alternating direction scheme. Still, practical implementations show that the relation between accuracy and time consumption of the numerical method is yet unsatisfactory. One of the main results of this thesis is a proposal for a non uniform structured grid which strongly improves the local error in that particular point in space we are interested in. Another contribution to reduce errors without increasing complexity is the choice of a discretisation at the zero variance boundary where no boundary condition is given. There, we discretise the p.d.e. using finite difference approximation from the right. These ideas which are coming from theoretical consideration are substantiated by numerical simulations.

Even though there exists a whole family of stochastic volatility models, the presented theory is mainly tailored to the p.d.e. arising from the Heston stochastic volatility model. Where possible, statements have been kept as general as possible in order to cover other p.d.e. as well. One difficulty of the Heston p.d.e. is its degenerating behaviour at the zero variance boundary. Existence and uniqueness properties have been shown in [8, Chapter 24] but only if the zero variance boundary is moved so that the p.d.e. is uniformly parabolic. In this work we present a proof which at least certifies a special finite difference method unconditionally stability even if the p.d.e. is not uniformly parabolic given a uniform grid and appropriate boundary conditions.

Chapter 2

Stochastic model

The aim of this chapter is to provide the reader with the basic idea on how to derive a differential equation for the value function of an option from the stochastic model of the underlying. It is not intended to introduce and exemplify the stochastic background like stochastic processes, stochastic differential equations (s.d.e.) and the no arbitrage principle. For an introductory text refer for example to [20].

For stochastic processes the subscript t denotes the random value which the process assumes at time t. In contrast a subscript t, v or s on a deterministic function means the derivative with respect to the variable. To distinct these two different meanings an additional prime $'$ is added to indicate derivatives. That will only be maintained within this chapter. In the following chapters a subscript t, x, y or v without a prime always denotes a partial derivatives.

2.1 Stochastic volatility models

Let $(\Omega, \mathcal{A}, P)$ a probability space. The value of the underlying financial product also called the spot value is in general modelled by a stochastic process denoted by S. At any time t the process is a random variable $S_t : \Omega \to \mathbb{R}_+$. In the Black-Scholes world the spot evolves according to the stochastic differential equation

$$\frac{\mathrm{d}S_t}{S_t} = \mu \,\mathrm{d}t + \sigma \,\mathrm{d}W_t$$

with a Wiener process W and constants $\mu \in \mathbb{R}$ and $\sigma > 0$ describing the drift and volatility, respectively. In stochastic volatility models the constant σ is replaced by a local times a stochastic volatility. The local volatility is a deterministic function of the spot and time. Following [14] the model can be characterised by

$$\begin{aligned} \frac{\mathrm{d}S_t}{S_t} &= \mu \,\mathrm{d}t + \sigma_L(S_t, t)\sqrt{v_t}\,\mathrm{d}W_t^{(1)}, \\ \mathrm{d}v_t &= \kappa(\theta - v_t)\,\mathrm{d}t + \xi\sqrt{v_t}\,\mathrm{d}W_t^{(2)}. \end{aligned}$$

In this approach the stochastic variance v is modelled by a mean reverting process with the mean variance $\theta > 0$, the strength of mean reversion $\kappa > 0$ and the so called vol of vol $\xi > 0$. The function $\sigma_L : \mathbb{R}_+ \times [0, T] \to \mathbb{R}_+$ is called local volatility. Both Wiener processes might be correlated. There exists several suggestion for the most appropriate choice of σ_L. For example Blacher suggests in [1] a local volatility which is quadratic in s

$$\sigma_L(s,t) = 1 + \alpha(s - S_0) + \beta(s - S_0)^2$$

with some constants α, β and the at time $t = 0$ observed spot S_0. The stochastic volatility model by Heston [10] is the particular case with $\sigma_L(s,t) = 1$. Figure 2.1 shows one sample path of the spot and volatility processes in Heston's model. The constants are taken from Table 6.1 and describe the exchange rate between US Dollar and Japanese Yen.

Remark 2.1.1 (Boundary behaviour of the variance process)
The variance process v never reaches the value zero if the inequality $\kappa\theta - \frac{1}{2}\xi^2 < 0$ is fulfilled. For details see [22, Chapter 5]. The idea can be summarised as follows: It is known that the so called

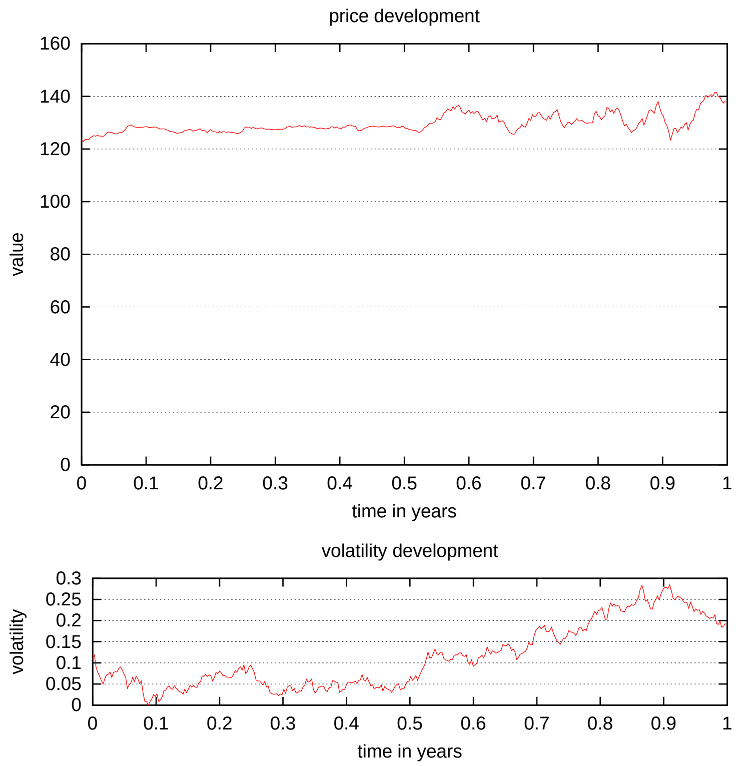

Figure 2.1: Sample path of the stochastic process S_t and $\sqrt{v_t}$

n-dimensional Bessel process $R_t : \Omega \to \mathbb{R}^1$ defined by the s.d.e.

$$\mathrm{d}R_t = \frac{n-1}{2R_t}\,\mathrm{d}t + \mathrm{d}W_t$$

never hits the origin if $n \geq 2$. The same obviously applies to the squared Bessel process $V_t := R_t^2$. By Itô's formula we see that

$$\mathrm{d}V_t = n\,\mathrm{d}t + 2\sqrt{V_t}\,\mathrm{d}W_t.$$

To obtain a similar s.d.e. as for the variance process we define $X_t := \frac{\xi^2}{4}V_t$ because then it is

$$\mathrm{d}X_t = \frac{\xi^2}{4}n\,\mathrm{d}t + \xi\sqrt{X_t}\,\mathrm{d}W_t.$$

The process X behaves qualitatively similar to v near the zero boundary if $\frac{\xi^2}{4}n = \kappa\theta$. It follows that the process v never reaches the zero value if

$$n = \frac{4\kappa\theta}{\xi^2} \geq 2.$$

2.2 Valuation of derivatives using the p.d.e. approach

We derive the p.d.e. for the value function u of an option only for the Heston model. The general stochastic volatility models can be treated in a similar way. First, it need to be remarked that the models above are incomplete in the sense that we are unable to perfectly replicate any given derivative with the underlying and the money market account. Different hedging strategies might be used. One example is the super hedging strategy where the value of the hedging portfolio is in almost any state $\omega \in \Omega$ at least as big as the payoff of the option at maturity T. The price of an option at issue time is therefore depending on the implemented hedging strategy and not unique as in the presence of complete markets.

In reality plain vanilla options are traded very liquidly in the market. One might tend to accept these prices and additionally use these products to hedge more exotic options. As it turns out we are then able to perfectly replicate derivatives and obtain unique prices.

We denote the value of a certain plain vanilla option by $c(s, v, t)$ where s is the current spot and v the current instantaneous volatility. Our hedging portfolio then consists of the following assets:

underlying

$$\begin{aligned} \mathrm{d}S_t &= S_t\mu\,\mathrm{d}t + S_t\sqrt{v_t}\,\mathrm{d}W_t^{(1)}, \\ \mathrm{d}v_t &= \kappa(\theta - v_t)\,\mathrm{d}t + \xi\sqrt{v_t}\,\mathrm{d}W_t^{(2)}, \\ \mathrm{d}W_t^{(1)}\,\mathrm{d}W_t^{(2)} &= \rho\,\mathrm{d}t, \end{aligned} \tag{2.1}$$

money market

$$\mathrm{d}M_t = M_t r\,\mathrm{d}t,$$

contingent claim

$$c(S_t, v_t, t).$$

At this stage we do not really need to know the value function $c : \mathbb{R}_+^2 \times [0, T] \to \mathbb{R}$. It is only important that c depends on no other values than s, v and t and that c is two times continously differentiable except for $t = T$.

Now, our objective is to hedge any given derivative of the underlying (with value function $u : \mathbb{R}_+^2 \times [0, T] \to \mathbb{R}$ to be determined) with a trading strategy $H_t : \Omega \to \mathbb{R}^3$, $H_t = (\alpha_t, \delta_t, \gamma_t)$, applied to the portfolio $\big(M_t, S_t, c(S_t, v_t, t)\big)$. The value of the trading strategy is then

$$h_t := \alpha_t M_t + \delta_t S_t + \gamma_t c(S_t, v_t, t).$$

We require the trading strategy to be self financing, i.e.

$$\mathrm{d}h_t := \alpha_t\,\mathrm{d}M_t + \delta_t\,\mathrm{d}S_t + \gamma_t\,\mathrm{d}c(S_t, v_t, t).$$

The value of the hedge portfolio must be equal to the value of the option

$$u(S_t, v_t, t) = h_t$$

and in particular, the instantaneous changes must be equal

$$\mathrm{d}u(S_t, v_t, t) = \mathrm{d}h_t.$$

Using Itô's formula we derive the partial differential equation (p.d.e.) which u must obey. In order to apply Itô's formula we have to assure that u and c are two times continuously differentiable. As it turns out this property is satisfied e.g. for any contingent claim or for barrier options. Itô's formula directly gives the expressions for $\mathrm{d}u(S_t, v_t, t)$ and $\mathrm{d}h_t$:

$$\begin{aligned} \mathrm{d}u(S_t, v_t, t) = &\left(u_t' + S_t\mu u_s' + \kappa(\theta - v)u_v' + \frac{1}{2}S_t^2 v_t u_{ss}'' + \frac{1}{2}\xi^2 v_t u_{vv}'' + S_t\xi v_t\rho u_{sv}''\right)\mathrm{d}t \\ &+ S_t\sqrt{v_t}u_s'\,\mathrm{d}W_t^{(1)} + \xi\sqrt{v_t}u_v'\,\mathrm{d}W_t^{(2)}, \end{aligned} \tag{2.2}$$

$$\begin{aligned} \mathrm{d}h_t = \gamma_t &\left(c_t' + S_t\mu c_s' + \kappa(\theta - v)c_v' + \frac{1}{2}S_t^2 v_t c_{ss}'' + \frac{1}{2}\xi^2 v_t c_{vv}'' + S_t\xi v_t\rho c_{sv}''\right)\mathrm{d}t \\ &+ (\alpha_t M_t r + \delta_t S_t\mu)\,\mathrm{d}t \\ &+ (\gamma_t S_t\sqrt{v_t}c_s' + \delta_t S_t\sqrt{v_t})\,\mathrm{d}W_t^{(1)} + \gamma_t\xi\sqrt{v_t}c_v'\,\mathrm{d}W_t^{(2)}. \end{aligned} \tag{2.3}$$

Given $|\rho| < 1$, the Itô processes $u(S_t, v_t, t)$ and h_t are identical if and only if the factors in front of $\mathrm{d}W^{(1)}$, $\mathrm{d}W^{(2)}$ and $\mathrm{d}t$ are equal (in a certain stochastic sense). Equality of the first two factors implies

$$\begin{aligned} S_t\sqrt{v_t}u_s' &= \gamma_t S_t\sqrt{v_t}c_s' + \delta_t S_t\sqrt{v_t}, \\ \xi\sqrt{v_t}u_v' &= \gamma_t\xi\sqrt{v_t}c_v' \end{aligned}$$

so that the γ- and δ-trading strategy are determined by

$$\begin{aligned}\gamma &= \frac{u'_v}{c'_v},\\ \delta &= u'_s - \gamma c'_s = u'_s - \frac{u'_v c'_s}{c'_v}.\end{aligned}$$

At this stage we already know how to perfectly hedge the option u with the money market account M_t, the underlying S_t and the option c. It remains the question how to hedge the option c but that will not be answered within this paper.

Having determined δ_t, γ_t and by replacing α_t using the relation $u(S_t, v_t, t) = h_t = \alpha_t M_t + \delta_t S_t + \gamma_t c(S_t, v_t, t)$ we now can compare the drift terms, yielding

$$\begin{aligned}&u'_t + S_t\mu u'_s + \kappa(\theta - v)u'_v + \frac{1}{2}S_t^2 v_t u''_{ss} + \frac{1}{2}\xi^2 v_t u''_{vv} + S_t\xi v_t \rho u''_{sv}\\ =&\gamma_t\left(c'_t + S_t\mu c'_s + \kappa(\theta - v)c'_v + \frac{1}{2}S_t^2 v_t c''_{ss} + \frac{1}{2}\xi^2 v_t c''_{vv} + S_t\xi v_t\rho c''_{sv}\right)\\ &+ (u - \delta_t S_t - \gamma_t c)\, r + \delta_t S_t \mu.\end{aligned}$$

By rearranging the terms and dividing the equation by u'_v we achieve that each side of the equation is either dependent on c or on u:

$$\begin{aligned}&\frac{1}{u'_v}\left(u'_t + S_t\mu u'_s + \kappa(\theta - v)u'_v + \frac{1}{2}S_t^2 v_t u''_{ss} + \frac{1}{2}\xi^2 v_t u''_{vv} + S_t\xi v_t\rho u''_{sv} - ru - (\mu - r)u'_s S_t\right)\\ =\ &\frac{1}{c'_v}\left(c'_t + S_t\mu c'_s + \kappa(\theta - v)c'_v + \frac{1}{2}S_t^2 v_t c''_{ss} + \frac{1}{2}\xi^2 v_t c''_{vv} + S_t\xi v_t\rho c''_{sv} - cr - (\mu - r)c'_s S_t\right).\end{aligned}$$

We can reproduce this result with any such option c. Given a sufficient richness of the set of these options we come to the conclusion that the left hand side of the equation does not depend on c but is a function of S_t, v_t and t, only. We denote this function by $\lambda : \mathbb{R}_+^2 \times [0, T] \to \mathbb{R}$ and obtain

$$\frac{1}{u'_v}\left(u'_t + S_t r u'_s + \kappa(\theta - v)u'_v + \frac{1}{2}S_t^2 v_t u''_{ss} + \frac{1}{2}\xi^2 v_t u''_{vv} + S_t\xi v_t\rho u''_{sv} - ru\right) = \lambda(S_t, v_t, t).$$

The partial differential equation the value function $u : \mathbb{R}_+^2 \times [0, T] \to \mathbb{R}$, $u(s, v, t)$ has to obey is therefore

$$u'_t + sru'_s + (\kappa(\theta - v) - \lambda(s, v, t))\, u'_v + \frac{1}{2}v\left(s^2 u''_{ss} + \xi^2 u''_{vv} + 2s\xi\rho u''_{sv}\right) - ru = 0.$$

The function λ is called market price of volatility risk and is not uniquely determined. That characterise an incomplete market where one cannot replicate derivatives with an portfolio only consisting of the money market account and the underlying. However, if one accepts the prices for plain vanilla call/put options (c) observed in the market and just wants to price other options (u), e.g. barrier options one gets a unique solution by finding an appropriate function λ which matches the prices for c. In practise one simply sets $\lambda(s, v, t) = 0$ and calibrates the parameters of the underlying to the observed prices for plain vanilla options (calibration to the smile).

If we have to deal with dividend paying assets S_t the situation slightly changes. We consider a continously dividend paying asset with the instantaneous rate r_f as it is the case in the foreign exchange market. Then, the instantaneous change in the hedge portfolio is:

$$\mathrm{d}h_t := \alpha_t\, \mathrm{d}M_t + \delta_t(\, \mathrm{d}S_t + r_\mathrm{f}\, \mathrm{d}t) + \gamma_t\, \mathrm{d}c(S_t, v_t, t).$$

Following the same steps as above we obtain the p.d.e.

$$u'_t + \frac{1}{2}v\left(s^2 u''_{ss} + \xi^2 u''_{vv} + 2s\xi\rho u''_{sv}\right) + s(r - r_\mathrm{f})u'_s + (\kappa(\theta - v) - \lambda(s, v, t))\, u'_v - ru = 0.$$

The final condition of the p.d.e. is given by the payoff of the option. A transformation from time t to time to maturity $T - t$ coordinates yields a p.d.e. with the payoff as initial condition. Additionally, we perform a transformation of the type $x = \log s$ in order to obtain a convection-diffusion type equation. Taking into account that Heston furthermore assumes that the market price of volatility risk is of the form $\lambda(s, v, t) = \lambda v$ with a constant λ so that $\kappa + \lambda > 0$ we obtain the equation which we call henceforth the Heston p.d.e.

$$u_t = \frac{1}{2}v\left(u_{xx} + \xi^2 u_{vv} + 2\rho\xi u_{xv}\right) + (r_\mathrm{d} - r_\mathrm{f} - \frac{1}{2}v)u_x + (\kappa(\theta - v) - \lambda v)u_v - r_\mathrm{d}u. \tag{2.4}$$

Chapter 3

Analysis of parabolic p.d.e.s

The field of parabolic partial differential equations has been extensively analysed in the literature. One of the most important examples is the convection-diffusion equation which has a variety of applications in physics and their analytical properties are well understood. Thanks to the concept of Fourier transformation the solution of constant coefficient p.d.e.s can be given in a fairly explicit way. Due to its simplicity and its powerful results which will also be used for consistency and stability estimates in 5.2 the basic ideas are discussed in the following sections. Additionally, fundamental solutions to simple equations are given which builds a basis for some rough estimates for the Heston p.d.e.

To begin with, we introduce the term parabolic p.d.e. and the notations we will use henceforth.

Definition 3.0.1 (Parabolic p.d.e.)
Let $\Omega \subset \mathbb{R}^d$ be a region (open and simply connected set) and $u(x,t)$, $u : \Omega \times [0,T] \to \mathbb{R}$ be a function which is two times continuously differentiable with respect to the space variable x and continuously differentiable with respect to the time variable t.

With the abbreviation $u(t) := u(\cdot,t)$ the partial differential equation

$$\frac{\partial}{\partial t}u(t) = Lu(t)$$

is then called parabolic if $L : \mathrm{C}^2(\Omega) \to \mathrm{C}(\Omega)$ is an elliptic differential operator or written more explicitly, the p.d.e.

$$\frac{\partial}{\partial t}u(x,t) = \sum_{i,j=1}^{d} a_{i,j}(x,t)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x,t)\frac{\partial u}{\partial x_i} + c(x,t)u + f(x,t)$$

is called parabolic if the Matrix $A(x,t) := (a_{i,j}(x,t))_{i,j=1}^d$ is positive definite for each $(x,t) \in \Omega \times (0,T)$. The p.d.e. is called uniformly parabolic if additionally $\langle A(x,t)y, y\rangle \geq C\,\|y\|^2$, $\forall y \in \mathbb{R}^d$ with a common constant $C > 0$ for all (x,t).

Introducing the multi-index $\alpha \in \mathbb{N}_0^d$ with the usual definitions $|\alpha| := \alpha_1 + \cdots + \alpha_d$, $x^\alpha := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ and $\mathbf{D}^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha 1} \cdots \partial x_d^{\alpha_d}}$ one can rewrite the p.d.e. in a more elegant way

$$\frac{\partial}{\partial t}u(x,t) = \sum_{|\alpha| \leq 2} p_\alpha(x,t)\mathbf{D}^\alpha u(x,t) + f(x,t).$$

We then define the characteristic polynomial of that p.d.e. by

$$P(x,t) := \sum_{|\alpha| \leq 2} p_\alpha(x,t)x^\alpha.$$

3.1 The convection-diffusion equation

The convection-diffusion equation is not just an example but is almost as general as the parabolic p.d.e. in the above definition. The main difference is the representation of the p.d.e. where the terms are no longer grouped in second-, first- and undifferentiated parts but as

$$\boldsymbol{\frac{\partial}{\partial t} u = \mathbf{div}(A \nabla u) - \mathbf{div}(ub) + f} \tag{3.1}$$

or in expanded form like

$$u_t = \sum_{i,j=1}^{d} a_{i,j} u_{x_i x_j} + \sum_{i=1}^{d} \left(\sum_{j=1}^{d} \frac{\partial a_{j,i}}{\partial x_j} - b_i \right) u_{x_i} - \operatorname{div}(b) u + f.$$

The matrix A is called diffusion matrix and b the convection vector of the convection-diffusion equation (3.1). Sinks and sources are represented by the function f. If one allows f to be dependent on u, i.e. if f is a functional of u then we can identify any parabolic p.d.e. with a convection-diffusion equation. In order to understand why one writes the p.d.e. as in (3.1) and why A and b are called diffusion matrix and convection vector, respectively, it is essential to discuss the physical background which will be the main objective of this section.

In physics these equations occur if we model the transport of a quantity for which a conservation law applies. This is the case, e.g. for the mass density of matter or the temperature (as a measure of energy density). Both, the diffusion of matter and the diffusion of heat are derived in the same principal way: One introduces a flow vector and states its dependence on the distribution of matter or temperature, respectively and then applies the conservation law. The following explanations are based on modelling the transport of a fluid (e.g. some pollution) inserted in an underlying flow of an other fluid (e.g. water in a river). It makes no difference if one imagines two gases (i.e. smoke in the air) instead of the fluids. First one needs to understand the basic physical law of mass conservation which will be the foundation and is derived in the next subsection. By defining what diffusion and convection quantitatively means we are able to deduce the convection-diffusion equation. Finally, the Heston p.d.e. as a non physical example will be stated.

3.1.1 Continuity equation

Let $\rho : \mathbb{R}^d \times [0,T] \to \mathbb{R}^+$, $\rho(x,t)$ be the mass density or concentration of the pollutant and $v : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, $v(x,t)$ be the velocity of a particle at the position $x \in \mathbb{R}^d$ and time t.

The conservation law now says that the change of mass of the matter in a certain region $G \subset \mathbb{R}^d$ is equal to the mass flow ρv over its borders ∂G

$$\frac{\partial}{\partial t} \int_G \rho(x,t)\, \mathrm{d}x = - \int_{\partial G} \rho(x,t) v(x,t)\, \mathrm{d}\sigma(x).$$

Using the Gauss integral theorem we obtain the equivalent formulation

$$\int_G \frac{\partial}{\partial t} \rho(x,t)\, \mathrm{d}x = - \int_G \operatorname{div}(\rho v)(x,t)\, \mathrm{d}x.$$

Since this has to remain true for any region $G \subset \Omega$ we obtain the differential formulation of the mass conservation which is

$$\boldsymbol{\frac{\partial}{\partial t} \rho + \mathbf{div}(\rho v) = 0}. \tag{3.2}$$

3.1.2 Modelling diffusion and convection

Diffusion is a process which forces the matter of the pollutant at a point with a higher concentration to flow towards a lower concentration. From molecular considerations (see Fick's law e.g. in [33, Subsection 5.4.5]) it follows that the mass flow of the pollutant defined by ρv is equal to

$$\rho v = -\kappa \nabla \rho$$

where κ is the conduction coefficient which describes the strength of diffusion and might depend on the position x and time t. It implies that the velocity of a particle imposed by the diffusion is $v = -\kappa \frac{\nabla \rho}{\rho}$. It might also be remarked that for heat conduction in non isotropic matter (i.e. crystal) the flow is not necessarily proportional to the density gradient but rather a linear expression of the gradient, i.e. $\rho v = -A \nabla \rho$ with a so called diffusion matrix A.

Coming to the modelling of convection it is natural to assume that the underlying fluid imposes its movement to the injected pollutant. Furthermore let this flow be known in advance and denoted by the vector field $b : \Omega \times [0, T] \to \mathbb{R}^d$.

Putting the effects of diffusion and convection additively together we conclude that the velocity of a particle of the pollutant is determined by

$$v(x,t) = b(x,t) - \kappa(x,t)\frac{\nabla \rho(x,t)}{\rho(x,t)}.$$

Having made an assumption about the flow of the injected fluid we would like to obtain an expression for the mass density ρ. Using the continuity equation leads to the desired result

$$\frac{\partial}{\partial t}\rho - \operatorname{div}(\kappa \nabla \rho) + \operatorname{div}(\rho b) = 0.$$

In the simplest case the parameter κ is a constant and the velocity field of the underlying fluid obeys $\operatorname{div} b = 0$ which is automatically satisfied if its mass density is constant, e.g. if the underlying fluid can be considered as incompressible. Under these assumptions the convection-diffusion equation simplifies to

$$\frac{\partial}{\partial t}\rho = \kappa \triangle \rho - \langle b, \nabla \rho \rangle .$$

Returning to the general form where the diffusion might also be non isotropic and additionally admitting sources and sinks represented by $f : \Omega \times [0, T] \to \mathbb{R}$ we obtain the most general form of the convection-diffusion equation

$$\frac{\partial}{\partial t}\rho = \operatorname{div}(A \nabla \rho) - \operatorname{div}(\rho b) + f.$$

3.1.3 Heston p.d.e. as convection-diffusion equation

The value function $u : \Omega \times [0, T] \to \mathbb{R}$, $u(x, v, t)$, with $\Omega \subset \mathbb{R} \times \mathbb{R}^+$ of an option obeys in the Heston model in its spot log transformed form ($x := \log s$) a convection-diffusion equation. That is basically the reason why one performs this transformation as convection-diffusion problems are well understood. Comparing the coefficients of the Heston p.d.e.

$$u_t \;=\; \frac{1}{2}v\left(u_{xx} + \xi^2 u_{vv} + 2\rho\xi u_{xv}\right) + (r_{\mathrm{d}} - r_{\mathrm{f}} - \frac{1}{2}v)u_x + (\kappa(\theta - v) - \lambda v)u_v - r_{\mathrm{d}}u$$

with the coefficients of the convection-diffusion p.d.e. we see that

$$\begin{aligned}
A &= \frac{1}{2}v\begin{pmatrix} 1 & \rho\xi \\ \rho\xi & \xi^2 \end{pmatrix}, \\
b &= \begin{pmatrix} \frac{1}{2}v - (r_{\mathrm{d}} - r_{\mathrm{f}}) + \frac{1}{2}\rho\xi \\ -(\kappa(\theta - v) - \lambda v) + \frac{1}{2}\xi^2 \end{pmatrix} = v\begin{pmatrix} \frac{1}{2} \\ \kappa + \lambda \end{pmatrix} + \begin{pmatrix} \frac{1}{2}\rho\xi + r_{\mathrm{f}} - r_{\mathrm{d}} \\ \frac{1}{2}\xi^2 - \kappa\theta \end{pmatrix}, \\
f &= (\kappa + \lambda - r_{\mathrm{d}})u.
\end{aligned}$$

We note that the second component of the flow vector b might be greater zero at the boundary $v = 0$ and in this case we are faced with an inflow boundary. Fortunately, with parameters usual in the markets the flow is quite small at $v = 0$.

3.2 Solution of parabolic p.d.e.s with constant coefficients

In some simple cases one is able to explicitly give solutions to parabolic p.d.e.s. The pure convection equation as well as the pure diffusion equation are such cases. Thanks to the Fourier transformation method a solution to any parabolic p.d.e. with constant coefficients can be found – at least in the Fourier transformed space.

3.2.1 The pure convection problem

The pure convection equation with a constant velocity vector field $b \in \mathbb{R}^d$

$$u_t + \langle b, \nabla u \rangle = 0$$

as derived in Section 3.1 should exactly exhibit the behaviour of the physical model, that is the transport of matter. Indeed, that is the case since with the initial condition $u_0 : \Omega \to \mathbb{R}$ the solution of the p.d.e. is given by

$$u(x,t) = u_0(x - bt), \qquad \forall (x,t) \in \Omega \times [0,T] : x - bt \in \Omega$$

simply because $u_t = -\langle \nabla u_0, b \rangle$. If the condition $x - bt \in \Omega$ is not satisfied the value $u(x,t)$ is determined by the boundary conditions. Assuming Dirichlet conditions only at the inflow boundary $\Gamma_{\text{in}} \subset \partial\Omega$ with $u(t)|_{\Gamma_{\text{in}}} = g(t)$ the solution at these points is

$$u(x + b\tau, t) = g(x, t - \tau), \qquad \forall (x,t) \in \partial\Omega \times [0,T] : x + bt \in \Omega.$$

It is clear that the values at the outflow boundary are determined by the inside values and can not be specified differently.

3.2.2 The pure diffusion problem – separation of variables

In the one dimensional case $x \in [a,b]$ and with help of the separation of variables approach we are able to solve the pure diffusion equation

$$u_t = \kappa u_{xx}.$$

The solution is assumed to have the form

$$u(x,t) = f(x)g(t).$$

It follows that

$$f(x)g'(t) = \kappa g(t) f''(x)$$

which is equivalent to

$$\frac{g'(t)}{\kappa g(t)} = \frac{f''(x)}{f(x)}.$$

Keeping $x \in [a,b]$ constant and varying $t \in [0,T]$ it follows that both terms have to be constant, e.g. $-c$:

$$\begin{aligned} g'(t) &= -c\kappa g(t), \\ f''(x) &= -cf(x). \end{aligned}$$

Considering the case $c > 0$ both ordinary differential equations have the general solution

$$\begin{aligned} g(t) &= c_1 \, \mathrm{e}^{-c\kappa t}, \\ f(x) &= c_2 \sin(\sqrt{c}x) + c_3 \cos(\sqrt{c}x) \end{aligned}$$

where the constants have to satisfy the given boundary conditions. For Dirichlet and Neumann boundaries only a countable number of constants $c > 0$ are admissible. Due to the physical interpretation as a frequency we define $\omega_i := \sqrt{c_i}$ and additionally rename the other constants. By superposition of solutions for each admissible ω_i we obtain the general solution of the pure diffusion equation

$$\begin{aligned} u(x,0) &= \sum_{i=0}^{\infty} a_i \sin(\omega_i x) + b_i \cos(\omega_i x), \\ u(x,t) &= \sum_{i=0}^{\infty} \big((a_i \sin(\omega_i x) + b_i \cos(\omega_i x) \big) \, \mathrm{e}^{-\kappa \omega_i^2 t}. \end{aligned} \tag{3.3}$$

The set of all admissible frequencies ω_i might also be uncountable if for example the infinite domain is considered in which case the sum has to be replaced by an integral expression. By this

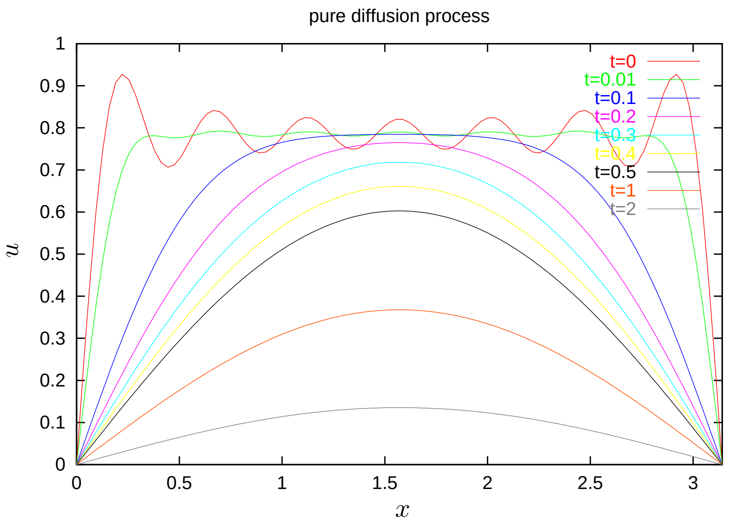

Figure 3.1: Solution of $u_t = u_{xx}$ with a step function as initial value

representation the solution of the pure diffusion equation becomes clear. Given the initial value function $u^{(0)} : \Omega \to \mathbb{R}$ one has to perform a Fourier analysis or transformation, respectively, to obtain the parameters a_i, b_i. High oscillations (ω_i big) in the initial condition are damped strongly with time by the factor $\mathrm{e}^{-\omega_i^2 t}$ so the solution becomes much smoother as time progresses. Figure 3.1 shows that behaviour where the initial condition

$$u^{(0)}(x) = \sum_{i=1}^{6} \frac{1}{2i+1} \sin(2i+1)x$$

has been chosen which is an approximation to the step function.

It also needs to be remarked that in the infinite domain exponentially growing functions can be solutions of the diffusion equation, too. This becomes obvious if we admit the constant c to be negative in the separation of variables approach. The solution then is

$$\begin{aligned} g(t) &= c_1 \,\mathrm{e}^{|c|\kappa t} \\ f(x) &= c_2 \,\mathrm{e}^{\sqrt{|c|}x} + c_3 \,\mathrm{e}^{-\sqrt{|c|}x} \end{aligned}$$

3.2.3 The general constant coefficient p.d.e. – Fourier transformation

As seen in the last subsection the structure of the solution looks much simpler in the frequency domain, i.e. in the Fourier transformed space. We will employ this method in order to examine solutions of parabolic p.d.e.s with constant coefficients in a more systematic way. For the definition and the properties of the Fourier transformation see appendix A.1. Particularly important is the differentiation rule (A.3) which makes it possible to solve the p.d.e. in the Fourier transformed space.

We consider the parabolic p.d.e. with constant coefficients in $\Omega = \mathbb{R}^d$

$$u_t = \sum_{|\alpha| \le 2} p_\alpha \mathbf{D}^\alpha u.$$

With the characteristic polynomial

$$P(x) := \sum_{|\alpha| \le 2} p_\alpha x^\alpha$$

we can formally write the p.d.e. as

$$u_t = P(\mathbf{D})u.$$

Since differentiation translates to multiplication in the Fourier transformed space the differential equation simplifies by the transformation. Transforming only the space variables $x \in \mathbb{R}^d$ the p.d.e.

rewrites to

$$\tilde{u}_t(\omega,t) = \sum_{|\alpha|\leq 2} (\mathrm{i}\omega)^\alpha p_\alpha \tilde{u}(\omega,t) = P(\mathrm{i}\omega)\tilde{u}(\omega,t).$$

This is an ordinary differential equation for each fixed $\omega \in \mathbb{R}^d$. Its unique solution is simply

$$\boldsymbol{\tilde{u}(\omega,t) = \tilde{u}^{(0)}(\omega)\exp\left(P(\mathrm{i}\omega)t\right).} \tag{3.4}$$

The importance of that result is worth mentioning. We now know the solution of p.d.e. in the Fourier transformed space in an explicit form so that we are able to interpret the solution. Furthermore we have shown existence and uniqueness of the parabolic p.d.e. in a constructive way. However, one needs to be careful with the statement of existence and uniqueness since that only holds true if the transformation and its inverse are applicable, e.g. $u \in \mathscr{S}(\mathbb{R}^d)$.

As an example we consider the pure diffusion equation

$$u_t = \kappa \triangle u.$$

The characteristic polynomial is therefore

$$P(x) = \kappa \sum_{i=1}^d x_i^2$$

and the solution of the p.d.e. in the Fourier transformed space is

$$\tilde{u}(\omega,t) = \tilde{u}^{(0)}(\omega)\exp\left(-\left\|\omega\right\|^2 t\right)$$

since $\omega \in \mathbb{R}^d$ is a real vector. This result coincides with the solution we have already obtained the one dimension using separation of variables, see equation (3.3).

3.3 Fundamental solution

The solution of a uniformly parabolic p.d.e. in $\mathbb{R}^d$ can be represented as an integral over a fundamental solution G

$$u(x,t) = \int_{\mathbb{R}^d} G(x,x',t)u(x',0)\,\mathrm{d}x', \tag{3.5}$$

see for example [32, Section 8].

The function G is sometimes also called Green's function. For constant coefficient p.d.e.s the fundamental solution can be given as an integral expression since with the Fourier transformed function $\tilde{u}$ and (3.4) we have

$$\begin{aligned}
u(x,t) &= (2\pi)^{-d/2}\int_{\mathbb{R}^d} \tilde{u}(\omega,t)\,\mathrm{e}^{\mathrm{i}\langle\omega,x\rangle}\ \mathrm{d}\omega \\
&= (2\pi)^{-d/2}\int_{\mathbb{R}^d} \tilde{u}^{(0)}(\omega)\exp\left(P(\mathrm{i}\omega)t\right)\mathrm{e}^{\mathrm{i}\langle\omega,x\rangle}\ \mathrm{d}\omega \\
&= (2\pi)^{-d/2}\int_{\mathbb{R}^d} (2\pi)^{-d/2}\int_{\mathbb{R}^d} u(x',0)\,\mathrm{e}^{-\mathrm{i}\left\langle\omega,x'\right\rangle}\ \mathrm{d}x'\exp\left(P(\mathrm{i}\omega)t\right)\mathrm{e}^{\mathrm{i}\langle\omega,x\rangle}\ \mathrm{d}\omega \\
&= \int_{\mathbb{R}^d} G(x,x',t)u(x',0)\,\mathrm{d}x'
\end{aligned}$$

with

$$G(x,x',t) := (2\pi)^{-d}\int_{\mathbb{R}^d} \exp\left(P(\mathrm{i}\omega)t\right)\mathrm{e}^{\mathrm{i}\left\langle\omega,x-x'\right\rangle}\ \mathrm{d}\omega.$$

For the pure diffusion equation that integral can be evaluated. Following [12, Chapter 4] the fundamental solution for $u_t = \kappa \triangle u$ is in d-dimensions given by

$$G(x,x',t) := (4\pi\kappa t)^{-d/2}\exp\left(-\frac{\left\|x-x'\right\|^2}{4\kappa t}\right)$$

which is illustrated in Figure 3.2.

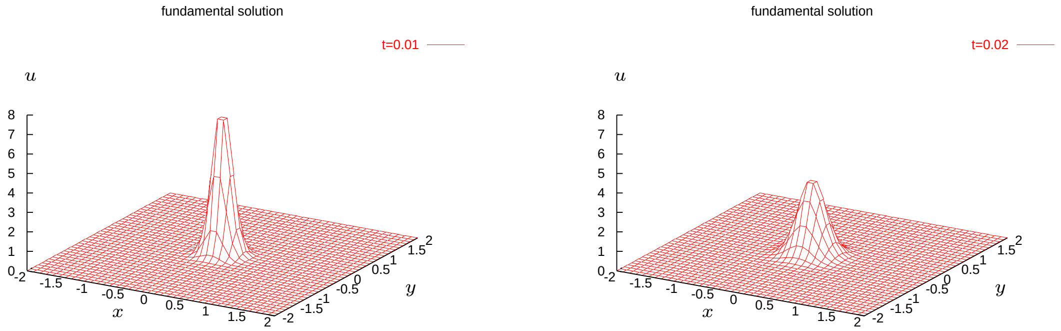

Figure 3.2: Fundamental solution of $u_t = u_{xx} + u_{yy}$

For analytical purposes it would be quite interesting to know the fundamental solution of the diffusion part of the Heston p.d.e. if existent. Only in one space dimension we will be able to find a fundamental solution. Hence consider the p.d.e. $u_t = \kappa u_{xx}$. By the variable transformation $z = (\frac{x}{2})^2$, i.e. $v : [0, \infty) \times [0, T] \to \mathbb{R}$, $v(z, t)$, is defined by $v\left((\frac{x}{2})^2, t\right) = u(x, t)$, we see that v fulfils the p.d.e.

$$v_t = \kappa \left(z v_{zz} + \frac{1}{2} v_z \right)$$

because $u_t = v_t$, $u_x = \frac{x}{2} v_z$ and $u_{xx} = \frac{1}{2} v_z + (\frac{x}{2})^2 v_{zz} = z v_{zz} + \frac{1}{2} v_z$. If the initial condition fulfils $u(x, 0) = 0$ for all $x \leq 0$ then we know from the fundamental solution for u denoted by G that

$$\begin{aligned} u(x,t) = v\left(\left(\frac{x}{2}\right)^2, t\right) &= \int_0^\infty G(x, x', t) v\left(\left(\frac{x'}{2}\right)^2, 0\right) \, \mathrm{d}x' \\ &= \int_0^\infty G(x, 2\sqrt{z'}, t) v(z', 0) \frac{1}{\sqrt{z'}} \, \mathrm{d}z' \end{aligned}$$

from which the following lemma directly follows.

Lemma 3.3.1
The fundamental solutions of the differential equations $u_t = \kappa u_{xx}$ denoted by G and of $u_t = \frac{\partial}{\partial x}(\kappa x u_x) - \frac{1}{2}\kappa u_x$ denoted by F are

$$\begin{aligned} G(x, x', t) &= \frac{1}{\sqrt{4\pi\kappa t}} \exp\left(-\frac{(x - x')^2}{4\kappa t} \right), \\ F(x, x', t) &= \frac{1}{\sqrt{x'}} G(2\sqrt{x}, 2\sqrt{x'}, t) \\ &= \frac{1}{\sqrt{4\pi\kappa x' t}} \exp\left(-\frac{(\sqrt{x} - \sqrt{x'})^2}{\kappa t} \right). \end{aligned} \tag{3.6}$$

Proof The first equation follows from [12, Chapter 4] and the second one is a direct result of the transformation $z = (\frac{x}{2})^2$ as shown above. □

Both functions are pictured in Figure 3.3 with the centre at $x' = 1$ and $x' = \frac{1}{2}$.

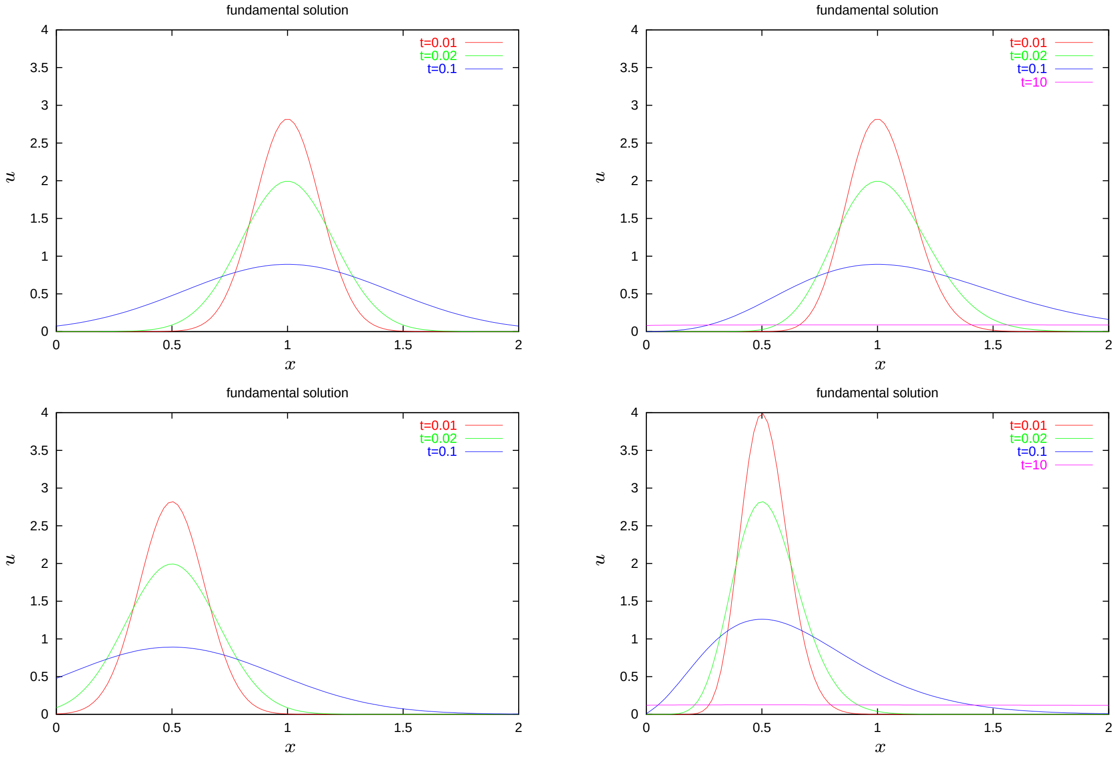

Figure 3.3: Fundamental solution of $u_t = u_{xx}$ on the left and of $u_t = xu_{xx} + \frac{1}{2}u_x$ on the right; in the upper pictures the centre is $x' = 1$ and in the lower pictures $x' = \frac{1}{2}$

Chapter 4

Description of the finite difference method

Solving parabolic p.d.e.s with the finite difference method (f.d.m.) is relatively straight forward. First the region $\bar{\Omega} \subset \mathbb{R}^d$ where the p.d.e. is defined needs to be approximated by a finite grid denoted by $\bar{\Omega}_h$. The variable $h > 0$ shall henceforth be used for the space discretisation parameter indicating the distance between two adjacent grid points. Section 4.1 answers the question how to generate non uniform grids. After that any continuous function defined on $\bar{\Omega}$, $u \in \mathrm{C}(\bar{\Omega})$, can be approximated by its values only at the grid-points. Functions only having values at grid points but which are related to the continuous function are denoted by a subscript h, i.e. $u_h : \bar{\Omega}_h \to \mathbb{R}$. All of those functions build the space of grid functions Φ_h which can be identified with the $\mathbb{R}^n$ given the grid points are numbered from one to n. Section 4.2 deals with the problem on how to approximate derivatives of a differentiable function if only its values in the grid points are known. After these introductory sections the numerical scheme is described in 4.3. Basically, the differentiable functions are replaced by grid functions and the partial derivatives in the p.d.e. are replaced by finite difference approximations so that one obtains a finite and linear equation system which can be solved by identifying Φ_h with the $\mathbb{R}^n$.

It needs to be remarked that the theory of finite differences was mainly developed in the years around 1960. The multitude of publications and the sometimes very time consuming process of obtaining literature made it impossible for me to get an absolute complete picture about the theory. I therefore can not rule out that some Theorems shown with much effort in the sections below were already known before. As a standard reference we use the articles [32] and [16] from the HANDBOOK OF NUMERICAL ANALYSIS.

4.1 Generating non uniform grids

The choice of an appropriate mesh is of very high importance since it directly influences the error we are making by approximating a continuous function with the function values in only grid points of the mesh. Basically, there are two different approaches of adapting grids. These are grid refinement where additional grid points are added in order to reduce approximation errors and grid redistribution. In the latter strategy the number of grid points remain constant and only the location of the grid points are changed. This has the advantage of no change in the complexity mainly depending on the number of grid points. That is why we focus on that method which will lead to the introduction of grid generating functions mapping from a uniform to a non uniform grid.

We restrict the discussion to so called structured grids which are the result of the direct product of d grids each only one-dimensional. For example in two dimensions these grids look like $\{(x_i, y_j) : i \in \{1, \dots, n\}, j \in \{1, \dots, m\}\}$. One example can bee seen in Figure 4.1. It is therefore sufficient to consider grids in only one dimension. In order to be more specific the following definition is provided.

Definition 4.1.1 (Structured grids, grid functions)
A grid of a subset $\Omega \subset \mathbb{R}^1$ is a finite set of points $\Omega_h := \left\{x^{(i)} : i \in \{1, \dots, m\}\right\} \subset \Omega$ with pints $x^{(i)} \in \mathbb{R}$ in strictly increasing order, i.e. $x^{(1)} < x^{(2)} < \dots x^{(m)}$.

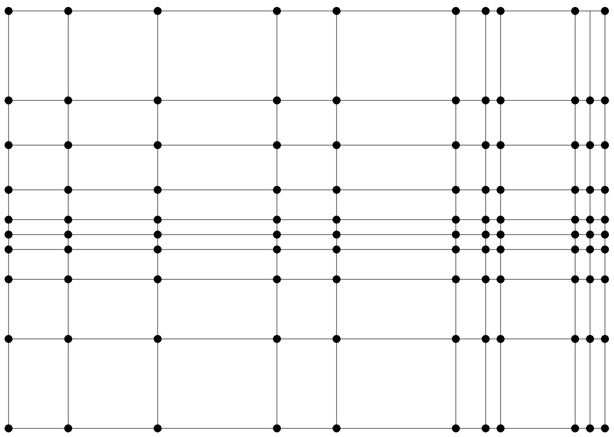

Figure 4.1: Example of a structured grid in two dimensions

A structured grid (sometimes also called tensor grid) of the closure of a region $\bar{\Omega} \subset \mathbb{R}^d$ is the set

$$\bar{\Omega}_h := \left\{ x^{(k)} = (x_1^{(k_1)}, \dots, x_d^{(k_d)}) : k \in I_h \right\} \subset \bar{\Omega}$$

based on d one dimensional meshes $\left\{ x_i^{(1)}, \dots, x_i^{(m_)i} \right\}$, $i = 1, \dots, d$. We call $I_h \subset \mathbb{N}_0^d$ the index set. To distinguish between boundary and inner points we define with Ω_h the set of all inner and with Γ_h the set of all boundary grid points.

The set of all functions $v : \bar{\Omega}_h \to \mathbb{R}$ denoted by Φ_h is called the space of grid functions. The multi index $k \in I_h$ as a subscript indicates the value at the k-th grid point: $v_k := v\big(x^{(k)}\big)$.

The discretisation parameter $h > 0$ can for example be defined as the maximum distance between two adjacent grid points or as the reciprocal of the number of grid points in one particular direction.

4.1.1 The grid generating function and the distance ratio function

One general approach to generate a non uniform mesh in one dimension is through a generating function. The idea is very simple. One has to specify an appropriate function $g : [0,1] \to [0,1]$ which is continuously differentiable, bijective and strictly monotonic increasing. Now, given a uniform grid on $[0,1]$ one applies the mapping g to these grid points and scales if necessary, i.e. the resulting non uniform mesh is then defined by $\{y_i\}_{i=0}^n$ with

$$y_i := cg(x_i) + d, \qquad x_i := \frac{i}{n}, \quad i = 1, \dots, n.$$

Figure 4.2 illustrates this process graphically. Unfortunately, the grid generating function is not very intuitive, that is to say if one wants the non uniform grid to be very dense at a certain point it is not immediately clear how to define the generating function. The example function $g(x) = x^2$ used in Figure 4.2 strongly increases the number of grid points at the lower boundary. A polynomial of third order for instance is able to concentrate grid points at any position with a pre-specified ratio.

With $\Delta x := \frac{1}{n}$ the distance between two adjacent grid point is obviously

$$y_{i+1} - y_i = cg(x_{i+1}) - cg(x_i) \approx cg'(x_i)\Delta x = cg'(g^{-1}(y_i))\Delta x.$$

Motivated by this result we introduce a distance ratio function $r : [0,1] \to \mathbb{R}^+$ by

$$r(y) := g'(g^{-1}(y)) \tag{4.1}$$

which characterises the ration between the distances of two adjacent points of the non uniform grid and the uniform grid. Having defined that distance ratio function one senses it more natural to first determine the distance ratio function and then deduce the mapping g from r. And indeed, that is possible since g satisfies the ordinary differential equation (o.d.e.)

$$g'(x) = r(g(x))$$

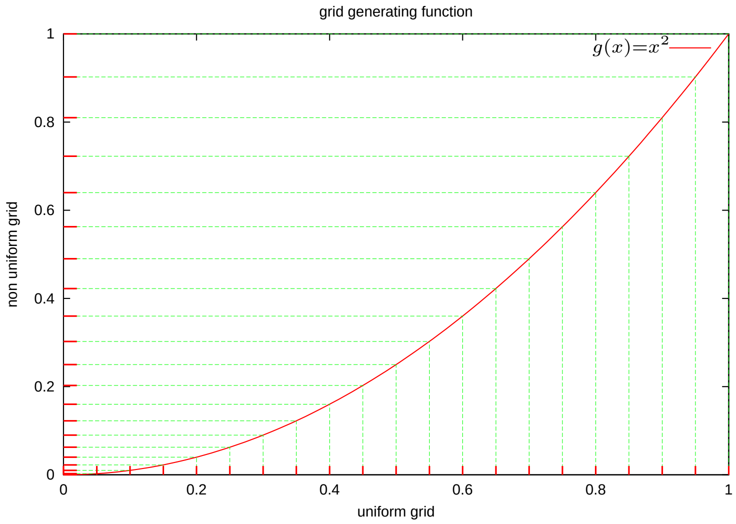

Figure 4.2: Grid generating function

which directly follows from the definition of r. This o.d.e. can be solved by the separation of variables approach

$$\int_0^x \frac{g'(z)}{r(g(z))}\,\mathrm{d}z = \int_0^x 1\,\mathrm{d}z$$

which yields

$$\int_0^{g(x)} \frac{1}{r(y)}\,\mathrm{d}y = x. \tag{4.2}$$

If it is possible to find a primitive to $\frac{1}{r}$ one has an implicit equation for the function g. In general where this might not be achievable the o.d.e. $g'(x) = r(g(x))$ can still be solved numerically. It is worth mentioning that the distance ratio function r can not be specified arbitrarily. Besides the obvious restrictions of $r(y) > 0$ and $r \in \mathrm{C}[0,1]$ it also has to satisfy the relation

$$\int_0^1 \frac{1}{r(y)}\,\mathrm{d}y = 1.$$

4.1.2 An example of the generating function

We choose a polynomial of degree three

$$g(x) = a_3(x-x^*)^3 + a_2(x-x^*)^2 + a_1(x-x^*) + a_0.$$

Let the position where the grid should be concentrated (concentration point) be denoted by $y^* \in [0,1]$. We then choose the constant $x^* \in [0,1]$ so that $g(x^*) = y^*$. The question is how to choose the parameters a_0 to a_3. With the additional requirement that the grid points at the concentration point have to be $\frac{1}{c}$ times as dense as in the uniform case we have to satisfy the following five equations with five unknowns (x^* and a_0 to a_3):

$$\begin{aligned} g(x^*) &= y^* \\ g(0) &= 0 \\ g(1) &= 1 \\ g'(x^*) &= c \\ g''(x^*) &= 0 \end{aligned}$$

The last equation has to hold since at y^* the concentration of grid points has to be the most dense. Inserting the function g into the equations we obtain

$$\begin{aligned} a_0 &= y^* \\ -a_3{x^*}^3 + a_2{x^*}^2 - a_1x^* + a_0 &= 0 \\ a_3(1-x^*)^3 + a_2(1-x^*)^2 + a_1(1-x^*) + a_0 &= 1 \\ a_1 &= c \\ 2a_2 &= 0. \end{aligned}$$

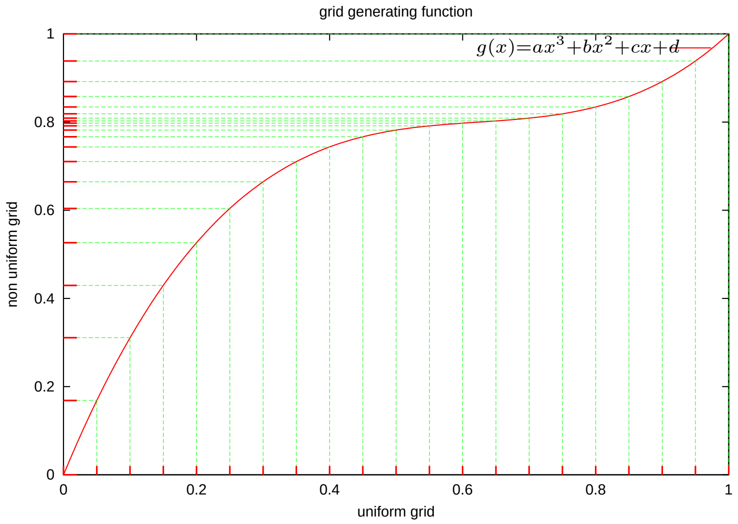

Figure 4.3: Polynomial of third degree as generating function with $y^* = 0.8$ and $c = 0.1$

We immediately see that $g(x) = a_3(x - x^*)^3 + c(x - x^*) + y^*$ and the remaining two parameters a_3 and x^* can be determined using a numerical method, e.g. Newton iteration method applied to the system of equations

$$\begin{aligned} -a_3{x^*}^3 - cx^* + y^* &= 0 \\ a_3(1 - x^*)^3 + c(1 - x^*) + y^* &= 1 \end{aligned}$$

which can be simplified to

$$\begin{aligned} \frac{-cx^* + y^*}{{x^*}^3} &= a_3 \\ (-cx^* + y^*)\frac{(1 - x^*)^3}{{x^*}^3} + c(1 - x^*) + y^* &= 1. \end{aligned}$$

With the parameter $y^* := 0.8$ and $c := 0.1$ the non uniform grid as shown in Figure 4.3 will be created. Using Newton iteration method it turns out that $x^* \approx 0.62353125$.

4.1.3 An example of the distance ratio function

Before choosing a distance ratio function one needs to be aware of the grid structure one would like to have, e.g. the following questions have to be answered: Is one concentration point sufficient or are there more points where the grid should be finer and how strong should the distance between two adjacent grid points increase as we go away from the concentration points? As a simple example we consider the distance ratio function

$$r(y) := \sqrt{c^2 + p^2(y - y^*)^2}.$$

The parameter y^* can be viewed as the centre of the grid point concentration with c as a measure of the intensity because r assumes its minimum at y^* with $r(y^*) = c$. For big values y the function is almost linear since $r(y) = \sqrt{c^2 + p^2(y - y^*)^2} \approx \sqrt{p^2y^2} = |py|$. The parameter p hast to be set appropriately so that the property of a density function is satisfied. A big advantage of the function r is that we are able to find an analytic solution for the grid generating function g by solving the o.d.e. $g' = r(g)$. From (4.2) it follows

$$\int_0^{g(x)} \frac{1}{\sqrt{c^2 + p^2(y - y^*)^2}}\,\mathrm{d}y = x.$$

Knowing the result of the integral

$$\int \frac{1}{\sqrt{ay^2 + by + c}}\,\mathrm{d}y = \frac{1}{\sqrt{a}}\operatorname{arsinh}\frac{2ay + b}{\sqrt{4ac - b^2}} \qquad \text{iff } a > 0, 4ac - b^2 > 0$$

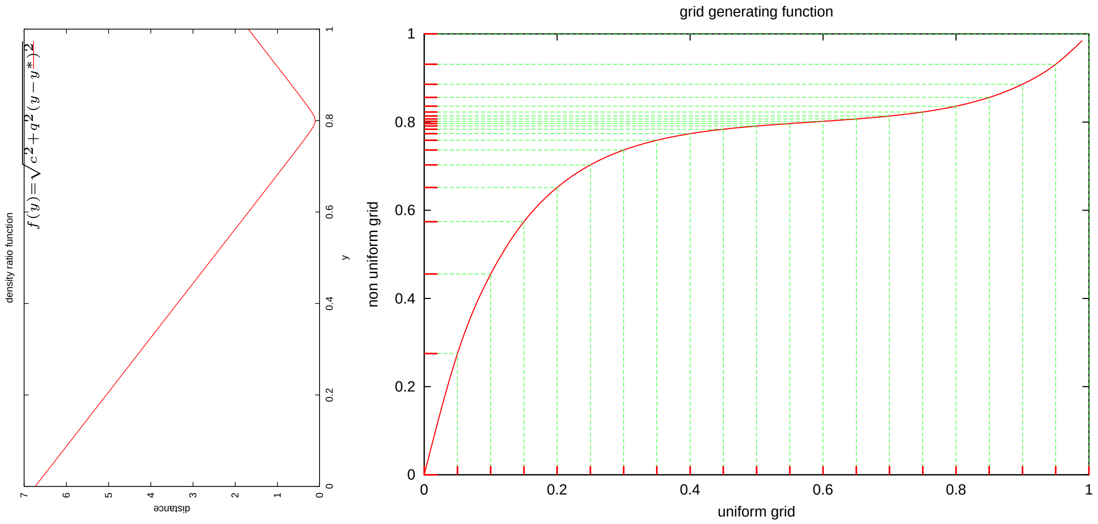

Figure 4.4: Concentration of grid points around $y^* = 0.8$ with 10 fold density $c = 0.1$

we conclude that

$$\frac{1}{p}\left(\operatorname{arsinh}\left(\frac{p}{c}(g(x)-y^*)\right)-\operatorname{arsinh}\left(-\frac{p}{c}y^*\right)\right)=x,$$

and thus

$$g(x)=y^*+\frac{c}{p}\sinh\left(px+\operatorname{arsinh}\left(-\frac{p}{c}y^*\right)\right). \tag{4.3}$$

The parameter p has to be chosen so that $g(1) = 1$. That can for example be performed using the Newton iteration method. With the particular parameter $y^* := 0.8$ and $c := 0.1$ it follows that $p \approx 8.42136$ which Figure 4.4 illustrates.

4.2 Approximation of derivatives

The general approach to approximate derivatives of a function $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, $f \in \mathrm{C}^3(\Omega)$, in a certain grid point $x^{(k)} \in \Omega_h$ is to use a weighted sum of the function values of adjacent grid points

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x^{(k)}) \approx \sum_{l \in I_0} a_l f(x^{(k+l)}). \tag{4.4}$$

The variables $k \in \mathbb{Z}^d$ and $l \in \mathbb{Z}^d$ are multi indices and $I_0 \subset \mathbb{Z}^d$ is the set of all indices considered to be adjacent to the origin.

In the following section we focus on cantered approximations using exactly 3^d grid points (compact stencil), i.e. the case $I_0 = \{-1, 0, 1\}^d$ is treated. However, right ($I_0 = \{0, 1, 2\}$) and left ($I_0 = \{-2, -1, 0\}$) hand side approximation are also mentioned.

4.2.1 Approximation of derivatives in one dimension

In one dimension we use subscripts instead of superscripts and abbreviate $\Delta x_k := x_{k+1} - x_k$. Before examining the general approach of a weighted sum of function values it is quite useful to describe how derivatives are approximated on uniform grids. We hence introduce the discrete backward and forward difference operators $\bar{\partial}$ and ∂, respectively, defined for any grid function $v \in \Phi_h$ by

$$(\partial v)_k := \frac{v_{k+1} - v_k}{\Delta x_k},$$
$$(\bar{\partial} v)_k := \frac{v_k - v_{k-1}}{\Delta x_{k-1}}.$$

For a good approximation it seems to be quite reasonable to use symmetric approximations. As it will be shown later in this subsection the approximation error with the abbreviation $f_k := f(x_k)$

of

$$\begin{aligned} f'(x_k) &\approx \frac{1}{2}(\partial + \bar{\partial}) f_k = \frac{f_{k+1} - f_{k-1}}{2\Delta x}, \\ f''(x_k) &\approx \partial\bar{\partial} f_k = \frac{f_{k+1} - 2f_k + f_{k-1}}{\Delta x^2} \end{aligned}$$

is of $\mathrm{O}(\Delta x^2)$ as Δx approaches zero. A generalisation to non uniform grids is non trivial. It will be shown that the best approximation using only three points is given by

$$\begin{aligned} f'(x_k) &\approx \big(\sigma\partial + (1-\sigma)\bar{\partial}\big) f_k, \qquad \sigma := \frac{\Delta x_{k-1}}{\Delta x_{k-1} + \Delta x_k}, \\ f''(x_k) &\approx \frac{\bar{\partial} f_{k+1} - \bar{\partial} f_k}{\frac{1}{2}(\Delta x_k + \Delta x_{k-1})}. \end{aligned}$$

Keeping the initial remarks in mind we now examine the very general approach

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x_k) \approx \sum_{i=l_1}^{l_2} a_i f(x_{k+i}).$$

As it turns out it suffices to set $l_1 = -1$ and $l_2 = 1$ to obtain an approximation to the first and second derivative which is second order accurate. The key to analyse the accuracy of this approximation is to apply Taylor series expansion around the grid point x_k.

With the abbreviations $f_k := f(x_k)$, $f_k' := f'(x_k)$, $f_k'' := f''(x_k)$, $\Delta x_1 := x_k - x_{k-1}$ and $\Delta x_2 := x_{k+1} - x_k$ we get:

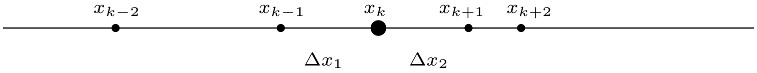

Figure 4.5: Δx_i

$$\begin{aligned} & a_{-1} f(x_{k-1}) + a_0 f(x_k) + a_1 f(x_{k+1}) \\ &= a_{-1}\left(f_k - \Delta x_1 f_k' + \frac{1}{2}\Delta x_1{}^2 f_k'' + R_3(-\Delta x_1)\right) + a_0 f_k \\ &\quad + a_1 \left(f_k + \Delta x_2 f_k' + \frac{1}{2}\Delta x_2{}^2 f_k'' + R_3(\Delta x_2)\right) \\ &= (a_{-1} + a_0 + a_1) f_k + (-a_{-1}\Delta x_1 + a_1 \Delta x_2) f_k' + (\frac{1}{2} a_{-1} \Delta x_1{}^2 + \frac{1}{2} a_1 \Delta x_2{}^2) f_k'' \\ &\quad + a_{-1} R_3(-\Delta x_1) + a_1 R_3(\Delta x_2) \\ &= \begin{pmatrix} f_k \\ f_k' \\ f_k'' \end{pmatrix}^{\tau} \begin{pmatrix} 1 & 1 & 1 \\ -\Delta x_1 & 0 & \Delta x_2 \\ \frac{1}{2}\Delta x_1{}^2 & 0 & \frac{1}{2}\Delta x_2{}^2 \end{pmatrix} \begin{pmatrix} a_{-1} \\ a_0 \\ a_1 \end{pmatrix} + a_{-1} R_3(-\Delta x_1) + a_1 R_3(\Delta x_2). \end{aligned}$$

Remark 4.2.1
Taylor series approximation gives an explicit expression for the remaining part R since it is

$$f(x + \Delta x) = f(x) + \sum_{k=1}^{n} \frac{1}{k!} \Delta x^k f^{(k)}(x) + R_{n+1}(\Delta x)$$

with

$$\begin{aligned} R_{n+1}(\Delta x) &= \frac{1}{n!} \int_0^{\Delta x} (\Delta x - t)^n f^{(n+1)}(x + t)\, \mathrm{d}t \\ &= \frac{1}{(n+1)!} f^{(n+1)}(\xi) \Delta x^{n+1}, \qquad \exists \xi \in [x, x + \Delta x] \end{aligned}$$

which can be estimated by

$$|R_{n+1}(\Delta x)| \le \frac{1}{(n+1)!} \left\| f^{(n+1)} \right\|_{\mathrm{C}[a,b]} |\Delta x|^{n+1}.$$

It follows that

$$\begin{aligned}\left|a_{-1}R_3(-\Delta x_1)+a_1R_3(\Delta x_2)\right| &\leq \frac{1}{6}\left\|f^{(n+1)}\right\|_{\mathrm{C}[a,b]}\left(\left|a_{-1}\right|\Delta x_1{}^3+\left|a_1\right|\Delta x_2{}^3\right)\\ &\leq \frac{1}{6}\left\|f^{(n+1)}\right\|_{\mathrm{C}[a,b]}\max\left\{\left|a_{-1}\right|+\left|a_1\right|\right\}h^3.\end{aligned}$$

In order to approximate the first derivative we have to choose the factors a_{-1}, a_0, a_1 so that factors before the function value f_k and its second derivative f_k'' are zero and the factor before the first derivative f_k' is one. In general, the following linear equations have to be solved, where exactly one of the δ's is one and the others are zero depending on which derivative has to be approximated.

$$\begin{pmatrix} 1 & 1 & 1 \\ -\Delta x_1 & 0 & \Delta x_2 \\ \frac{1}{2}\Delta x_1{}^2 & 0 & \frac{1}{2}\Delta x_2{}^2 \end{pmatrix}\begin{pmatrix} a_{-1} \\ a_0 \\ a_1 \end{pmatrix}=\begin{pmatrix} \delta \\ \delta_x \\ \delta_{xx} \end{pmatrix}$$

The solution is

$$\begin{pmatrix} a_{-1} \\ a_0 \\ a_1 \end{pmatrix}=\begin{pmatrix} 0 & \frac{-\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)} & \frac{2}{\Delta x_1(\Delta x_1+\Delta x_2)} \\ 1 & \frac{\Delta x_2-\Delta x_1}{\Delta x_1\Delta x_2} & \frac{-2}{\Delta x_1\Delta x_2} \\ 0 & \frac{\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)} & \frac{2}{\Delta x_2(\Delta x_1+\Delta x_2)} \end{pmatrix}\begin{pmatrix} \delta \\ \delta_x \\ \delta_{xx} \end{pmatrix}.$$

We summarise the result in Table 4.1 and additionally consider the uniform case where $\Delta x := \Delta x_1 = \Delta x_2$.

	non uniform case			uniform case		
	f	f'	f''	f	f'	f''
a_{-1}	0	$\frac{-\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)}$	$\frac{2}{\Delta x_1(\Delta x_1+\Delta x_2)}$	0	$\frac{-1}{2\Delta x}$	$\frac{1}{\Delta x^2}$
a_0	1	$\frac{\Delta x_2-\Delta x_1}{\Delta x_1\Delta x_2}$	$\frac{-2}{\Delta x_1\Delta x_2}$	1	0	$\frac{-2}{\Delta x^2}$
a_1	0	$\frac{\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)}$	$\frac{2}{\Delta x_2(\Delta x_1+\Delta x_2)}$	0	$\frac{1}{2\Delta x}$	$\frac{1}{\Delta x^2}$

Table 4.1: Central approximation of derivatives, inner points

The central difference scheme can not be applied on boundaries. For convection dominated parabolic p.d.e.s the Finite Difference Method with central difference approximation exhibits an oscillating behaviour. That is why we also need to discuss left and right hand side approximations. Beginning with right hand side approximation

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x_k)\approx\sum_{i=0}^{2}a_if(x_{k+i}),$$

redefining $\Delta x_1 := x_{k+1}-x_k$, $\Delta x_2 := x_{k+2}-x_{k+1}$ and applying the same method described above results in the system

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}=\begin{pmatrix} 1 & 1 & 1 \\ 0 & \Delta x_1 & \Delta x_1+\Delta x_2 \\ 0 & \frac{1}{2}\Delta x_1{}^2 & \frac{1}{2}(\Delta x_1+\Delta x_2)^2 \end{pmatrix}\begin{pmatrix} \delta \\ \delta_x \\ \delta_{xx} \end{pmatrix}.$$

The result of this equation system is given in Table 4.2.

The left hand side approximation of derivatives with

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x_k)\approx\sum_{i=-2}^{0}a_if(x_{k+i}).$$

	non uniform case			uniform case		
	f	f'	f''	f	f'	f''
a_0	1	$-\frac{2\Delta x_1+\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)}$	$\frac{2}{\Delta x_1(\Delta x_1+\Delta x_2)}$	1	$\frac{-3}{2\Delta x}$	$\frac{1}{\Delta x^2}$
a_1	0	$\frac{\Delta x_1+\Delta x_2}{\Delta x_1\Delta x_2}$	$\frac{-2}{\Delta x_1\Delta x_2}$	0	$\frac{2}{\Delta x}$	$\frac{-2}{\Delta x^2}$
a_2	0	$\frac{-\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)}$	$\frac{2}{\Delta x_2(\Delta x_1+\Delta x_2)}$	0	$\frac{-1}{2\Delta x}$	$\frac{1}{\Delta x^2}$

Table 4.2: Right hand side approximation of derivatives, left border

	non uniform case			uniform case		
	f	f'	f''	f	f'	f''
a_{-2}	0	$\frac{\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)}$	$\frac{2}{\Delta x_1(\Delta x_1+\Delta x_2)}$	0	$\frac{1}{2\Delta x}$	$\frac{1}{\Delta x^2}$
a_{-1}	0	$-\frac{\Delta x_1+\Delta x_2}{\Delta x_1\Delta x_2}$	$\frac{-2}{\Delta x_1\Delta x_2}$	0	$\frac{-2}{\Delta x}$	$\frac{-2}{\Delta x^2}$
a_0	1	$\frac{2\Delta x_2+\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)}$	$\frac{2}{\Delta x_2(\Delta x_1+\Delta x_2)}$	1	$\frac{3}{2\Delta x}$	$\frac{1}{\Delta x^2}$

Table 4.3: Left hand side approximation of derivatives, right border

is similar and we obtain the factors as shown in Table 4.3. There, the values of Δx_1 and Δx_2 have been redefined with $\Delta x_1 := x_{k-1} - x_{k-2}$ and $\Delta x_2 := x_k - x_{k-1}$.

The rate of convergence of the approximations to the real derivatives is given by the following proposition.

Proposition 4.2.2
Assume $f \in \mathrm{C}^3[a,b]$ and let $\{x_{-1}^{(h)}, x_0, x_1^{(h)}\} \subset (a,b)$, $h > 0$, be a series of points with $x_{-1}^{(h)} < x_0 < x_1^{(h)}$, $\Delta x_1 := x_0 - x_{-1}^{(h)}$, $\Delta x_2 := x_1^{(h)} - x_0$ and $h := \max\{\Delta x_1, \Delta x_2\}$ with the restriction that $\Delta x_1 \sim \Delta x_2 \quad (h \to 0)$, i.e. $\exists c_1, c_2 > 0 : c_1\Delta x_2 \leq \Delta x_1 \leq c_2\Delta x_2 \quad \forall h > 0$. It then applies

$$\begin{aligned} f'(x_0) &= f_h'(x_0) + \mathrm{O}(h^2) \\ f''(x_0) &= f_h''(x_0) + \mathrm{O}(h) \end{aligned}$$

where the $f_h'(x_0)$ and $f_h'(x_0)$ are defined as the approximation of the derivatives at the point x_0 using the method and the factors as derived above. On equidistant meshes ($\Delta_h x_1 = \Delta_h x_2$, $\forall h > 0$) the second derivative is even been approximated with second order accuracy if the function f is also four times continuous differentiable.

Proof Taking the factors for the approximation of the first derivative (see one of the Tables 4.1 to 4.3) and using the fact that $\Delta_h x_1 \sim \Delta_h x_2 \quad (h \to 0)$ one can find a constant $\alpha > 0$ so that $|a_{-1}| \leq \alpha\frac{1}{h}$ and $|a_1| \leq \alpha\frac{1}{h}$. Furthermore, the factors a_i were derived so that

$$f'(x_0) = f_h'(x_0) + \mathrm{O}(\max\{|a_{-1}h^3|, |a_1 h^3|\}).$$

With the estimate of a_{-1} and a_1 we obtain

$$f'(x_0) = f_h'(x_0) + \mathrm{O}(h^2).$$

Similarly, we get an estimate for the factors a_i in the case of the second derivative $|a_i| \leq \tilde{\alpha}\frac{1}{h^2} \quad i \in \{-1, 1\}$ which leads to the conclusion:

$$f''(x_0) = f_h''(x_0) + \mathrm{O}(h)$$

Second order accuracy on uniform meshes is obtained due to the fact that having chosen the factors a_{-1}, a_0 and a_1 so that the first derivative vanishes, the third derivative vanishes at the same time. More precisely,

$$\begin{aligned} f_h''(x_0) := \sum_{i=-1}^{1} a_i f(x_i^{(h)}) &= (a_{-1} + a_0 + a_1)f(x_0) + (-a_{-1}\Delta_h x_1 + a_1\Delta_h x_2)f'(x_0) \\ &+ \frac{1}{2}(a_{-1}\Delta_h {x_1}^2 + a_1\Delta_h {x_2}^2)f''(x_0) + \frac{1}{6}(-a_{-1}\Delta_h {x_1}^3 + a_1\Delta_h {x_2}^3)f'''(x_0) \\ &+ \mathrm{O}(\max\{|a_{-1}\Delta_h x_1^4|, |a_1\Delta_h x_2^4|\}). \end{aligned}$$

In the case that $\Delta_h x_1 = \Delta_h x_2$ the factor in front of the first derivative is zero if and only if the factor in front of the third derivative is zero. By construction of the second derivative the factor in front of the first derivative is zero. Therefore we obtain the estimate

$$f''(x_0) = f_h''(x_0) + \mathrm{O}(\max\{|a_{-1}h^4|, |a_1 h^4|\})$$

which results in

$$f''(x_0) = f_h''(x_0) + \mathrm{O}(h^2).$$

□

Let now the grid be created using a generating function $g : [0,1] \to [a,b]$ as defined in Subsection 4.1.1. We then define the discretisation parameter h as the distance between two adjacent grid points of the uniform grid in $[0,1]$. As we will see the accuracy of approximating the second derivative is similar to uniform grids, i.e. of order h^2. This comes not unexpected because as $h \to 0$ the two distances Δx_1 and Δx_2 become similar, more precisely $\Delta x_1 - \Delta x_2 \to 0$.

Lemma 4.2.3 (Accuracy of differential approximation)
Assume $f \in \mathrm{C}^4[a,b]$ and let $\{x_{-1}^{(h)}, x_0, x_1^{(h)}\} \subset (a,b)$, $h > 0$, be a series of points which are the result of a grid generating function $g \in \mathrm{C}^2[0,1]$, i.e. $x_i^{(h)} := g(z_0 + ih)$. Defining the distance ratio function as before $r(x) := g'(g^{-1}(x))$ the approximation errors of derivatives using the central three point scheme can then be estimated by

$$|f'(x_0) - f_h'(x_0)| \le \frac{1}{6} r(x_0)^2 \, \|f'''\|_{\mathrm{C}[x_{-1},x_1]} \, h^2 + \mathrm{O}(h^3)$$
$$|f''(x_0) - f_h''(x_0)| \le \left(\frac{1}{12} r(x_0)^2 \left\| f^{(4)} \right\|_{\mathrm{C}[x_{-1},x_1]} + \frac{1}{3} \|g''\|_{\mathrm{C}[z_0-h,z_0+h]} \, f'''(x_0) \right) h^2 + \mathrm{O}(h^3)$$

where the $f_h'(x_0)$ and $f_h''(x_0)$ are defined as the approximation of the derivatives at the point x_0 using the method and the factors as derived above and shown in Table 4.1. The higher order term can be estimated with a uniform constant, i.e. $\mathrm{O}(h^3) \le ch^3$ for any $x_0 \in [a,b]$ and c only depends on the functions f and g.

Proof The derivation of the factors a_i as shown above is already one important aspect of the proof since the factors have been chosen so that the lower parts of the Taylor series expansion of $\frac{\partial f}{\partial x}(x_k) - \sum_{i=-1}^{1} a_i f_{k+i}$ vanish. The same applies to the second derivative. We now estimate the higher order terms and first state that Δx_i is given by

$$\Delta x_1 = hg'(z_0) - \frac{1}{2} h^2 g''(\xi_1), \qquad \xi_1 \in [z_0 - h, z_0]$$
$$\Delta x_2 = hg'(z_0) + \frac{1}{2} h^2 g''(\xi_2), \qquad \xi_2 \in [z_0, z_0 + h]$$

Referring to Table 4.1 we estimate the error of the first derivative approximation. By obvious calculation we see that the factors a_{-1} and a_1 become similar to the factors of a uniform mesh as $h \to 0$:

$$a_{-1} = \frac{-1}{2hg'(z_0)} - \frac{g''(\xi)}{g'(z_0)^2} \, \mathrm{O}(1), \qquad h \to 0$$
$$a_1 = \frac{1}{2hg'(z_0)} - \frac{g''(\xi)}{g'(z_0)^2} \, \mathrm{O}(1), \qquad h \to 0$$

The factors in Table 4.1 are chosen so that the error is of third order in Δx_1 and Δx_2, i.e.

$$|f'(x_0) - f_h'(x_0)| = \frac{1}{6} \left(-a_{-1} {\Delta x_1}^3 f'''(\xi_1) + a_1 {\Delta x_2}^3 f'''(\xi_2) \right), \qquad \xi_1 \in [x_{-1}, x_0], \xi_2 \in [x_0, x_1]$$
$$\le \frac{1}{6} h^2 g'(z_0)^2 \, \|f'''\|_{\mathrm{C}[x_{-1},x_1]} + \|f'''\|_{\mathrm{C}[x_{-1},x_1]} \, \mathrm{O}(h^3).$$

For the factors a_i of the second derivative we similarly have

$$a_{-1} = \frac{1}{h^2 g'(z_0)^2} + \frac{g''(\xi)}{g'(z_0)^3} \, \mathrm{O}(h^{-1}), \qquad h \to 0$$
$$a_1 = \frac{1}{h^2 g'(z_0)^2} - \frac{g''(\xi)}{g'(z_0)^3} \, \mathrm{O}(h^{-1}), \qquad h \to 0.$$

The usage of one more term in the Taylor series expansion results in the estimate

$$\begin{aligned}
|f''(x_0) - f_h''(x_0)| &= \frac{1}{6}\left(-a_{-1}\Delta x_1{}^3 + a_1\Delta x_2{}^3\right) f'''(x_0) + \frac{1}{24}\left(a_{-1}\Delta x_1{}^4 f^{(4)}(\xi_1) + a_1\Delta x_2{}^4 f^{(4)}(\xi_2)\right) \\
&\le \frac{1}{6}\left(-a_{-1}\Delta x_1(h^2 g''(z_0)^2 - g'(z_0)g''(\xi_1)h^3) + a_1\Delta x_2(h^2 g''(z_0)^2 + g'(z_0)g''(\xi_2)h^3)\right) f'''(x_0) \\
&\quad + \frac{1}{12} g'(z_0)^2 h^2 \left\| f^{(4)} \right\|_{\mathrm{C}[x_{-1},x_1]} + \mathrm{O}(h^3) \\
&= \frac{1}{6}\left(a_{-1}\Delta x_1 g'(z_0)g''(\xi_1)h^3 + a_1\Delta x_2 g'(z_0)g''(\xi_1)h^3\right) f'''(x_0) \\
&\quad + \frac{1}{12} g'(z_0)^2 h^2 \left\| f^{(4)} \right\|_{\mathrm{C}[x_{-1},x_1]} + \mathrm{O}(h^3) \\
&= \left(\frac{1}{3}\left\| g'' \right\|_{\mathrm{C}[z_0-h,z_0+h]} f'''(x_0) + \frac{1}{12} g'(z_0)^2 \left\| f^{(4)} \right\|_{\mathrm{C}[x_{-1},x_1]}\right) h^2 + \mathrm{O}(h^3).
\end{aligned}$$

In the third step the relation $-a_{-1}\Delta x_1 + a_1\Delta x_2 = 0$ has been used. □

This Lemma roughly says that the approximation error for both derivatives reduces to one quarter if the distance ratio function halves. However, a second error component comes into action if the non uniformity is too strong, i.e. if $|g''(z_0)|$ is big.

4.2.2 Approximation of derivatives in two dimensions

Since other schemes can be treated in a similar way only the central approximation scheme will be discussed. All derivatives are then approximated using a compact nine point stencil, that is the point itself and all eight adjacent grid points as it is exemplarily shown for the mixed derivative

$$\frac{\partial^2 f}{\partial x \partial y}(x_k, y_l) \approx \sum_{i,j=-1}^{1} a_{i,j} f(x_{k+i}, y_{l+j}).$$

Before going into the analysis one already expects certain results of the one dimensional case to be applicable. So would it not be surprising if the approximation of all derivatives in the same direction is been done using only three grid points along this direction as in the one dimensional case. For the mixed derivative the relation

$$\frac{\partial^2 f}{\partial x \partial y}(x_k, y_l) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} f(x_k, y_l) \approx \frac{\partial}{\partial x}\left(\sum_{j=-1}^{1} c_j f(x_k, y_{l+j})\right) \approx \sum_{i,j=-1}^{1} b_i c_j f(x_{k+i}, y_{l+j})$$

suggests that $a_{i,j} = b_i c_j$ where b_i and c_j denote the factors approximating the first derivative in x and y direction, respectively. These first ideas will turn out to be one possible solution. However, there exist more possibilities to approximate derivatives with the same accuracy.

Using Taylor approximation for $f(x_{k+i}, y_{l+j})$ around (x_k, y_l) and the abbreviation $f'_x := \frac{\partial}{\partial x} f(x_k, y_l)$ leads to

$$\begin{aligned}
\sum_{i,j=-1}^{1} a_{i,j} f(x_{k+i}, y_{l+j}) \quad = \quad & a_{0,0} f_k \\
+ \quad & a_{-1,0}\left(f - \Delta x_1 f'_x + \frac{1}{2}\Delta x_1{}^2 f''_{xx}\right) + a_{1,0}\left(f + \Delta x_2 f'_x + \frac{1}{2}\Delta x_2{}^2 f''_{xx}\right) \\
+ \quad & a_{0,-1}\left(f - \Delta y_1 f'_y + \frac{1}{2}\Delta y_1{}^2 f''_{yy}\right) + a_{0,1}\left(f + \Delta y_2 f'_y + \frac{1}{2}\Delta y_2{}^2 f''_{yy}\right) \\
+ \quad & a_{-1,-1}\left(f - \Delta x_1 f'_x - \Delta y_1 f'_y + \frac{1}{2}\left(\Delta x_1{}^2 f''_{xx} + \Delta y_1{}^2 f''_{yy}\right) + \Delta x_1 \Delta y_1 f''_{xy}\right) \\
+ \quad & a_{-1,1}\left(f - \Delta x_1 f'_x + \Delta y_2 f'_y + \frac{1}{2}\left(\Delta x_1{}^2 f''_{xx} + \Delta y_2{}^2 f''_{yy}\right) - \Delta x_1 \Delta y_2 f''_{xy}\right) \\
+ \quad & a_{1,-1}\left(f + \Delta x_2 f'_x - \Delta y_1 f'_y + \frac{1}{2}\left(\Delta x_2{}^2 f''_{xx} + \Delta y_1{}^2 f''_{yy}\right) - \Delta x_2 \Delta y_1 f''_{xy}\right) \\
+ \quad & a_{1,1}\left(f + \Delta x_2 f'_x + \Delta y_2 f'_y + \frac{1}{2}\left(\Delta x_2{}^2 f''_{xx} + \Delta y_2{}^2 f''_{yy}\right) + \Delta x_2 \Delta y_2 f''_{xy}\right) \\
+ \quad & \mathrm{O}\left(\max\{a_{i,j}\tilde{\Delta x}_i, \tilde{\Delta y}_j : i,j \in \{-1,0,1\}\}\right) \quad \left((\Delta x_1, \Delta x_2, \Delta y_1, \Delta y_2) \to 0\right).
\end{aligned}$$

The variables $\widetilde{\Delta x}_i := x_k - x_{k+i}$ have been introduced in order to write the order term O in a compressed way. Collecting all arguments for each derivative we see that this sum is equal (up to the term O(...)) to

$$\begin{pmatrix} f \\ f'_x \\ f'_y \\ f''_{xx} \\ f''_{yy} \\ f''_{xy} \end{pmatrix}^{\tau} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -\Delta x_1 & -\Delta x_1 & -\Delta x_1 & 0 & 0 & 0 & \Delta x_2 & \Delta x_2 & \Delta x_2 \\ \frac{1}{2}\Delta x_1{}^2 & \frac{1}{2}\Delta x_1{}^2 & \frac{1}{2}\Delta x_1{}^2 & 0 & 0 & 0 & \frac{1}{2}\Delta x_2{}^2 & \frac{1}{2}\Delta x_2{}^2 & \frac{1}{2}\Delta x_2{}^2 \\ -\Delta y_1 & 0 & \Delta y_2 & -\Delta y_1 & 0 & \Delta y_2 & -\Delta y_1 & 0 & \Delta y_2 \\ \frac{1}{2}\Delta y_1{}^2 & 0 & \frac{1}{2}\Delta y_2{}^2 & \frac{1}{2}\Delta y_1{}^2 & 0 & \frac{1}{2}\Delta y_2{}^2 & \frac{1}{2}\Delta y_1{}^2 & 0 & \frac{1}{2}\Delta y_2{}^2 \\ \Delta x_1\Delta y_1 & 0 & -\Delta x_1\Delta y_2 & 0 & 0 & 0 & -\Delta x_2\Delta y_1 & 0 & \Delta x_2\Delta y_2 \end{pmatrix} \begin{pmatrix} a_{-1,-1} \\ a_{-1,0} \\ a_{-1,1} \\ a_{0,-1} \\ a_{0,0} \\ a_{0,1} \\ a_{1,-1} \\ a_{1,0} \\ a_{1,1} \end{pmatrix}.$$

In order to obtain factors to approximate derivatives the equation system has to be solved, where again all δ's are zero except that δ which corresponds to the derivative to be approximated:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -\Delta x_1 & -\Delta x_1 & -\Delta x_1 & 0 & 0 & 0 & \Delta x_2 & \Delta x_2 & \Delta x_2 \\ \frac{1}{2}\Delta x_1{}^2 & \frac{1}{2}\Delta x_1{}^2 & \frac{1}{2}\Delta x_1{}^2 & 0 & 0 & 0 & \frac{1}{2}\Delta x_2{}^2 & \frac{1}{2}\Delta x_2{}^2 & \frac{1}{2}\Delta x_2{}^2 \\ -\Delta y_1 & 0 & \Delta y_2 & -\Delta y_1 & 0 & \Delta y_2 & -\Delta y_1 & 0 & \Delta y_2 \\ \frac{1}{2}\Delta y_1{}^2 & 0 & \frac{1}{2}\Delta y_2{}^2 & \frac{1}{2}\Delta y_1{}^2 & 0 & \frac{1}{2}\Delta y_2{}^2 & \frac{1}{2}\Delta y_1{}^2 & 0 & \frac{1}{2}\Delta y_2{}^2 \\ \Delta x_1\Delta y_1 & 0 & -\Delta x_1\Delta y_2 & 0 & 0 & 0 & -\Delta x_2\Delta y_1 & 0 & \Delta x_2\Delta y_2 \end{pmatrix} \begin{pmatrix} a_{-1,-1} \\ a_{-1,0} \\ a_{-1,1} \\ a_{0,-1} \\ a_{0,0} \\ a_{0,1} \\ a_{1,-1} \\ a_{1,0} \\ a_{1,1} \end{pmatrix} = \begin{pmatrix} \delta \\ \delta_x \\ \delta_{xx} \\ \delta_y \\ \delta_{yy} \\ \delta_{xy} \end{pmatrix}.$$

There may be many solutions to that equation system. However we are only interested in one solution so we try to solve it by a product approach $a_{i,j} = \hat{a}_i\tilde{a}_j$ which yields

$$\begin{pmatrix} \left(\sum\limits_{j=-1}^{1} \tilde{a}_j\right) \begin{pmatrix} 1 & 1 & 1 \\ -\Delta x_1 & 0 & \Delta x_2 \\ \frac{1}{2}\Delta x_1{}^2 & 0 & \frac{1}{2}\Delta x_2{}^2 \end{pmatrix} \begin{pmatrix} \hat{a}_{-1} \\ \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} \\ \left(\sum\limits_{i=-1}^{1} \hat{a}_i\right) \begin{pmatrix} 1 & 1 & 1 \\ -\Delta y_1 & 0 & \Delta y_2 \\ \frac{1}{2}\Delta y_1{}^2 & 0 & \frac{1}{2}\Delta y_2{}^2 \end{pmatrix} \begin{pmatrix} \tilde{a}_{-1} \\ \tilde{a}_0 \\ \tilde{a}_1 \end{pmatrix} \\ \sum\limits_{i,j\in\{-1,1\}} p_i q_j \hat{a}_i \tilde{a}_j \end{pmatrix} = \begin{pmatrix} \delta \\ \delta_x \\ \delta_{xx} \\ \delta \\ \delta_y \\ \delta_{yy} \\ \delta_{xy} \end{pmatrix},$$

where p and q are defined by $(p_{-1}, p_0, p_1) := (-\Delta x_1, 0, \Delta x_2)$ and $(q_{-1}, q_0, q_1) := (-\Delta y_1, 0, \Delta y_2)$. Now it becomes obvious that we can reduce the solution to the one dimensional case. If we, for instance, want to find the factors to approximate f_x, i.e. $\delta_x = 1$ and all other are equal to zero, we obtain a solution if we set $\tilde{a}_{-1} = 0$, $\tilde{a}_0 = 1$, $\tilde{a}_1 = 0$ and solve the equation system

$$\begin{pmatrix} 1 & 1 & 1 \\ -\Delta x_1 & 0 & \Delta x_2 \\ \frac{1}{2}\Delta x_1{}^2 & 0 & \frac{1}{2}\Delta x_2{}^2 \end{pmatrix} \begin{pmatrix} \hat{a}_{-1} \\ \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

which is exactly the same system we have already solved in the one dimensional case. The other lines in the big system are equal to zero since we set $\tilde{a}_{-1}$ and $\tilde{a}_1$ to zero. The same argument applies to derivatives with respect to y.

To find an approximation for the mixed derivative f_{xy}, let $\hat{a}_i$ and $\tilde{a}_j$ are the factors which approximate f_x and f_y, respectively. As it can easily be seen, $a_{i,j} := \hat{a}_i\tilde{a}_j$ are the sought factors to approximate f_{xy}. That is because $\sum_{i=-1}^{1} \hat{a}_i = 0$ and $\sum_{i=-1}^{1} \tilde{a}_i = 0$. Finally, we have $\sum_{i\in\{-1,1\}} p_i\hat{a}_i = 1$ and $\sum_{j\in\{-1,1\}} q_i\tilde{a}_i = 1$. Consequently $\sum_{i,j\in\{-1,1\}} p_i\tilde{a}_i\, q_j\tilde{a}_j = 1$. Table 4.4 sums up the result.

Remark 4.2.4
By similar estimation as shown in Lemma 4.2.3 one can show that the error by approximating the mixed derivative is also of quadratic order in h.

4.3 Finite difference method (f.d.m.)

Besides the finite element and finite volume method the finite difference method is one method to solve partial differential equations numerically. In its simplest form it requires a structured

$a_{i,j}$	non uniform case		
	-1	0	1
-1	$\frac{\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)}\frac{\Delta y_2}{\Delta y_1(\Delta y_1+\Delta y_2)}$	$\frac{-\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)}\frac{\Delta y_2-\Delta y_1}{\Delta y_1\Delta y_2}$	$\frac{-\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)}\frac{\Delta y_1}{\Delta y_2(\Delta y_1+\Delta y_2)}$
0	$\frac{\Delta x_2-\Delta x_1}{\Delta x_1\Delta x_2}\frac{-\Delta y_2}{\Delta y_1(\Delta y_1+\Delta y_2)}$	$\frac{\Delta x_2-\Delta x_1}{\Delta x_1\Delta x_2}\frac{\Delta y_2-\Delta y_1}{\Delta y_1\Delta y_2}$	$\frac{\Delta x_2-\Delta x_1}{\Delta x_1\Delta x_2}\frac{\Delta y_1}{\Delta y_2(\Delta y_1+\Delta y_2)}$
1	$\frac{\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)}\frac{-\Delta y_2}{\Delta y_1(\Delta y_1+\Delta y_2)}$	$\frac{\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)}\frac{\Delta y_2-\Delta y_1}{\Delta y_1\Delta y_2}$	$\frac{\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)}\frac{\Delta y_1}{\Delta y_2(\Delta y_1+\Delta y_2)}$

$a_{i,j}$	uniform case		
	-1	0	1
-1	$\frac{1}{4\Delta x\Delta y}$	0	$\frac{-1}{4\Delta x\Delta y}$
0	0	0	0
1	$\frac{-1}{4\Delta x\Delta y}$	0	$\frac{1}{4\Delta x\Delta y}$

Table 4.4: Central approximation of the mixed derivative

grid that is why the method is not suitable for regions $\Omega \subset \mathbb{R}^d$ with a smooth boundary. Since the Heston p.d.e. can be approximated by a rectangular domain that causes no problems. The big advantage of f.d.m. is its simplicity and straight forward implementation where in short the derivatives are replaced by differential quotients. Additionally, under certain conditions it achieves second order convergence.

In the following we consider the parabolic p.d.e. as introduced in Chapter 3 with $u(x,t)$, $u : \Omega \times [0,T] \to \mathbb{R}$ and the abbreviation $u(t) := u(\cdot, t)$

$$\frac{\partial u(t)}{\partial t} = Lu(t) + f(t) \qquad \text{in } \Omega, \quad \forall t \in [0,T] \tag{4.5}$$

where Ω is a rectangular domain and

$$L = \sum_{|\alpha|\leq 2} p_\alpha(x)\mathbf{D}^\alpha.$$

In general p_α might also depend on time. However, as the time dependency of the coefficients is not of particular interest within the topic of this thesis we do not consider it. It is no problem, though, to extend the following statements to the time dependent case.

4.3.1 The method

Given any structured grid $\bar{\Omega}_h = \{x^{(i)} : i \in I_h\}$ of the space variables with the index set $I_h := \{0,\dots,m_1\}\times\dots\times\{0,\dots,m_d\}$ and a one dimensional grid of the time $\{t_k\}_{k=0}^{m_0} \subset [0,T]$ the function $u : \Omega \times [0,T] \to \mathbb{R}$ will be approximated at time point t_k by the grid function $\bar{u}^{(k)} \in \Phi_h$, defined as

$$\bar{u}^{(k)}(x) := u(x, t_k), \qquad \forall x \in \bar{\Omega}_h.$$

As indicated before we sometimes write in short $\bar{u}_i^{(k)} := u^{(k)}(x^{(i)})$ for any multi index $i \in I_h$. The space of grid functions will often be identified with the $\mathbb{R}^n$ where $n := |I_h| := (m_1+1)\cdot\ldots\cdot(m_d+1)$.

We first discuss the discretisation of the differential operator in space, L, which is straight forward since we know how to approximate derivatives of the form $\mathbf{D}^\alpha u(x,t)$ for $|\alpha| \leq 2$ according to Section 4.2. In general the approximation takes the form

$$\mathbf{D}^\alpha u(x^{(i)}, t_k) \approx \sum_{j\in I_0} a_{i,j}\bar{u}_{i+j}^{(k)}.$$

with the constants $a_{i,j}$, e.g. as shown in Table 4.1 and 4.4. The constants $a_{i,j}$ are of course differing for different derivatives α, but in order not to over-stretch the index it has been left out. It is natural to define the discrete differential operator by

$$\left(\mathbf{D}_h^\alpha \bar{u}^{(k)}\right)(x^{(i)}) := \sum_{j\in I} a_{i,j}\,\bar{u}_{i+j}^{(k)}, \qquad \forall x^{(i)} \in \Omega_h.$$

By identifying the space of grid functions Φ_h with the $\mathbb{R}^n$ the differential operator can be represented by a sparse $n \times n$ matrix. The discrete operator of L is now defined by

$$\left(L_h \bar{u}^{(k)}\right)(x^{(i)}) := \sum_{|\alpha|\leq 2} p_\alpha(x^{(i)}) \left(\mathbf{D}_h^\alpha \bar{u}^{(k)}\right)(x^{(i)}), \qquad \forall x^{(i)} \in \Omega. \tag{4.6}$$

And again, L_h can be represented by a sparse $n \times n$ matrix. Its sparsity results from the fact that derivatives at a grid point are approximated using function values of only the immediate adjacent grid points and not all points of the grid. In the one dimensional case where $\Omega = (a, b)$, L_h is represented by a band matrix, e.g. tridiagonal for central differences.

The first step to discretise a parabolic p.d.e. is to discretise the space variables only. In order to accomplish that and not to confuse with the existing vector $\bar{u}^{(k)}$ we introduce the vector function $\varphi : [0, T] \to \Phi_h$ depending on time as an approximation to the solution u on all grid points

$$\varphi_i(t) \approx u(x^{(i)}, t).$$

Replacing the continuous space operator L by its discrete version L_h in the parabolic p.d.e. (4.5) and thus only considering the function values of u in the spacial grid points the p.d.e. is approximated by a system of ordinary differential equations (o.d.e.). Since the vector function φ is the representation of u in its spacial grid points we require φ to comply with

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi(t) = L_h \varphi(t). \tag{4.7}$$

This system of o.d.e.s is called the semi-discrete system of the parabolic p.d.e. (4.5). It has constant coefficients $(L_h)_{i,j}$. In the case of a time dependent spacial operator L the coefficients in the system of o.d.e.s are time dependent, too. One now could employ a sophisticated numerical o.d.e. solver to obtain a highly accurate solution for φ. However, the numerical algorithm has to be able to deal with large sparse systems L_h. We take a different approach and use the simplest solver available – the explicit or implicit Euler method. The advantages are its simplicity, the ability to deal with sparse systems and an easy to establish error analysis which finally reveals an acceptable convergence order. In order to explain the method we discretise the variable t and define the series of vectors $\hat{u}^{(0)}, \ldots, \hat{u}^{(m_0)}$ as an approximation to φ at the time points $t_0 = 0, \ldots, t_{m_0} = T$

$$\hat{u}^{(k)} \approx \varphi(t_k).$$

There are many different ways to approximate the time derivative. Taking more than two points for the derivative results in a multi-step scheme. Also, the operator L_h can be applied at slightly different times and an average can be taken. That all leaves a lot of freedom to construct a stable and high order convergent scheme. However, as said before, we only consider the Euler- or single step method which approximates the time derivative with two points in time and the right hand side at a time point in between. With a parameter $\theta \in [0, 1]$ and the abbreviation $\Delta t_k := t_{k+1} - t_k$ the method is defined by

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k)}}{\Delta t_k} = L_h \left(\theta \hat{u}^{(k+1)} + (1 - \theta) \hat{u}^{(k)} \right) \tag{4.8}$$

and is called the θ-method or weighted method.

Changing from the space Φ_h to $\mathbb{R}^n$ all grid function become vectors and operators like L_h become matrices. Given the vector $\hat{u}^{(k)}$ one can determine $\hat{u}^{(k+1)}$. The initial condition is given and therefore $\hat{u}^{(0)}$ known. Step by step $\hat{u}^{(1)}$, $\hat{u}^{(2)}$, ..., $\hat{u}^{(m_0)}$ are being determined always solving the linear equation system

$$\boldsymbol{(I - \theta \Delta t_k L_h) \hat{u}^{(k+1)} = (I + (1 - \theta) \Delta t_k L_h) \hat{u}^{(k)}} \tag{4.9}$$

or in short

$$A_h \hat{u}^{(k+1)} = B_h \hat{u}^{(k)}. \tag{4.10}$$

The involved matrices $A_h(k) := I - \theta \Delta t_k L_h$ and $B_h(k) := I + (1 - \theta) \Delta t_k L_h$ play an important role in the stability analysis. The three particular schemes where $\theta = 0$, $\theta = \frac{1}{2}$ and $\theta = 1$ are called forward Euler (fully explicit), Crank-Nicholson and backward Euler (fully implicit) scheme, respectively. The Crank-Nicholson scheme is of particular importance since it has second order consistency in time (see Section 5.1). The second order consistency of (4.8) is almost expected because the difference quotient in time is second order accurate at the time point $\frac{t_k + t_{k+1}}{2}$ and the right hand side of the equation is only second order accurate at this time point if and only if $\theta \hat{u}^{(k+1)} + (1 - \theta) \hat{u}^{(k)}$ is a second order approximation to $\varphi\big(\frac{t_k + t_{k+1}}{2}\big)$ which is only fulfilled if $\theta = \frac{1}{2}$.

Many numerical methods exists for the solution of the linear equation system (4.10) which is of order n. A simple Gauss elimination algorithm which does not exploit the structure of the matrix A_h would need memory storage of order $\mathrm{O}(n^2)$ and $\mathrm{O}(n^3)$ operations to solve the equation.

Under practical consideration that would limit its application to $n \leq 10000$. Fortunately, there are more sophisticated algorithms available. One basically distinguishes between iterative and direct methods which all take advantage of the sparsity of the matrix A_h. Iterative methods do not alter the matrix A_h but find successively a better approximation to the solution. Examples are the SOR, Bi-CG and GmRes methods. In contrast, direct methods change the structure of the matrix. The LU method with column permutation for example, decomposes the matrix multiplicatively into $A_h = LU$ with a left lower matrix L and a right upper matrix U using Gauss elimination. The previous permutation of columns is needed in order to give A_h the sparsest band structure as possible because L and U have the same band width as A_h but in general all values within that band are not zero. The advantage of direct methods is, once the decomposition has been established the solution of $A_h\hat{u}^{(k)} = b$ can be determined very efficiently, i.e. the number of calculations is of order $\mathrm{O}(l^2 n)$. If the time grid is uniform then A_h does not change with time and the decomposition process needs to be done only once. The disadvantage is that it needs $\mathrm{O}(ln)$ bytes of memory where l denotes the number of non zero diagonals after the permutation. In practice that is much more memory space than the sparse matrix A_h occupies but also much less than $\mathrm{O}(n^2)$.

4.3.2 Componentwise splitting

Even though numerical methods solving the equation system (4.10) are quite efficient they do not reach the efficiency of solving a tridiagonal system of the same order n. That is why one introduces splitting methods where a series of tridiagonal systems are solved instead of the original general sparse system A_h. As soon as a more efficient algorithm for solving general sparse matrices is developed splitting methods might not be superior, any more. The following descriptions are all based on [16].

If the operator L does not contain any mixed derivative one can decompose the matrix L_h additively into

$$L_h = \sum_{i=1}^{d} D_i$$

where D_i is that part of L_h which discretises derivatives in direction i. After permutation all matrices D_i are tridiagonal or have at most five non zero diagonals. A method which in each step solves an equation system only containing one of the matrices D_i is called a splitting method.

The objective of this section is to introduce splitting methods which are at least consistent of order $\mathrm{O}(\Delta t)$ and preferably of order $\mathrm{O}(\Delta t^2)$ like the Crank-Nicholson method. For the sake of simplicity no strict proof for consistency and stability is given. More details can be found in the [16]. To start with, we describe the simplest case where the matrix L_h is decomposed into two parts

$$L_h = D_1 + D_2.$$

The general case can be discussed in a similar way. The only difficulty there is to construct a second order consistent scheme in time. The basic idea is to introduce half time steps, i.e. vectors $\hat{u}^{(k+1/2)}$. We now add a zero to the equation of the θ-method (4.8)

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k+1/2)} + \hat{u}^{(k+1/2)} - \hat{u}^{(k)}}{\Delta t_k} = (D_1 + D_2)\left(\theta\hat{u}^{(k+1)} + (1-\theta)\hat{u}^{(k)}\right).$$

As an approximation we instead solve

$$\begin{aligned}\frac{\hat{u}^{(k+1/2)} - \hat{u}^{(k)}}{\Delta t_k} &= D_1\left(\theta\hat{u}^{(k+1/2)} + (1-\theta)\hat{u}^{(k)}\right),\\ \frac{\hat{u}^{(k+1)} - \hat{u}^{(k+1/2)}}{\Delta t_k} &= D_2\left(\theta\hat{u}^{(k+1)} + (1-\theta)\hat{u}^{(k+1/2)}\right).\end{aligned} \tag{4.11}$$

To see whether this splitting method is an approximation to the original θ-method (4.8) we eliminate $\hat{u}^{(k+1/2)}$ from

$$\begin{aligned}\big(I - \theta\Delta t_k D_1\big)\hat{u}^{(k+1/2)} &= \big(I + (1-\theta)\Delta t_k D_1\big)\hat{u}^{(k)},\\ \big(I - \theta\Delta t_k D_2\big)\hat{u}^{(k+1)} &= \big(I + (1-\theta)\Delta t_k D_2\big)\hat{u}^{(k+1/2)}\end{aligned}$$

to obtain

$$\hat{u}^{(k+1)} = \big(I - \theta\Delta t_k D_2\big)^{-1}\big(I + (1-\theta)\Delta t_k D_2\big)\big(I - \theta\Delta t_k D_1\big)^{-1}\big(I + (1-\theta)\Delta t_k D_1\big)\hat{u}^{(k)}.$$

Given the norm of $\Delta t_k D_1$ and $\Delta t_k D_2$ is sufficiently small we can apply Neumann series $(I-T)^{-1} = \sum_{i=0}^{\infty} T^i$:

$$\begin{aligned}\hat{u}^{(k+1)} &= \big(I + \theta\Delta t_k D_2 + \theta^2 \Delta t_k^2 D_2^2 + \dots\big)\big(I + (1-\theta)\Delta t_k D_2\big)\\ &\cdot \big(I + \theta\Delta t_k D_1 + \theta^2 \Delta t_k^2 D_1^2 + \dots\big)\big(I + (1-\theta)\Delta t_k D_1\big)\hat{u}^{(k)}\\ &= \big(I + \Delta t_k D_2 + \Delta {t_k}^2 \theta D_2^2 + \dots\big)\big(I + \Delta t_k D_1 + \Delta {t_k}^2 \theta D_1^2 + \dots\big)\hat{u}^{(k)}\\ &= \big(I + \Delta t_k (D_1 + D_2) + \Delta {t_k}^2 (D_2 D_1 + \theta(D_1^2 + D_2^2)) + \dots\big)\hat{u}^{(k)}.\end{aligned}$$

Hence, the system can be written in the form

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k)}}{\Delta t_k} = \theta L_h \hat{u}^{(k)} + (1-\theta)L_h\hat{u}^{(k)} + \Delta t_k (D_2 D_1 + (1-\theta)(D_1^2 + D_2^2)) + \dots\big)\hat{u}^{(k)}.$$

Replacing $\hat{u}^{(k)}$ with $\hat{u}^{(k+1)} - \Delta t_k L_h \hat{u}^{(k)} - \Delta {t_k}^2 (\dots)\hat{u}^{(k)} + \dots$ we obtain a scheme which is similar to the θ-method

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k)}}{\Delta t_k} = \theta L_h \hat{u}^{(k+1)} + (1-\theta)L_h\hat{u}^{(k)} + \Delta t_k (D_2 D_1 + \theta(D_1^2 + D_2^2 - L_h^2)) + \dots\big)\hat{u}^{(k)}.$$

Since the difference to the θ-method is only of order $\mathrm{O}(\Delta t_k)$ one expects it to be first order accurate in time. Even for $\theta = \frac{1}{2}$ the consistency in general is also only of first order. In special cases it can reach second order, though, but only if D_1 and D_2 commute, i.e. $D_1 D_2 = D_2 D_1$. Then the term $D_2 D_1 + \frac{1}{2}(D_1^2 + D_2^2 - L_h^2)$ becomes zero as $L_h^2 = D_1^2 + D_2^2 + 2D_1 D_2$. Unfortunately, the restriction $D_1 D_2 = D_2 D_1$ is too strict and in practice mainly does not hold. In order to compete with the Crank-Nicholson method as far as the order of convergence is concerned one needs to construct a splitting method which approximates the Crank-Nicholson scheme up to a term of order $\mathrm{O}(\Delta {t_k}^2)$. As it turns out that will be achieved if one swaps D_1 and D_2 on the right hand side of the splitting method just discussed, namely by solving

$$\begin{aligned}\frac{\hat{u}^{(k+1/2)} - \hat{u}^{(k)}}{\Delta t_k} &= \Big(\theta D_1 \hat{u}^{(k+1/2)} + (1-\theta) D_2 \hat{u}^{(k)}\Big),\\ \frac{\hat{u}^{(k+1)} - \hat{u}^{(k+1/2)}}{\Delta t_k} &= \Big(\theta D_2 \hat{u}^{(k+1)} + (1-\theta) D_1 \hat{u}^{(k+1/2)}\Big).\end{aligned} \tag{4.12}$$

This method is called Alternating Direction scheme and plays a very important role for $\theta = \frac{1}{2}$. Eliminating $\hat{u}^{(k+1/2)}$ from the equation and applying the same estimation as above we see that the Alternating Direction method (4.12) approximates the θ-method (4.8) as follows.

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k)}}{\Delta t_k} = \theta L_h \hat{u}^{(k+1)} + (1-\theta)L_h\hat{u}^{(k)} + \Delta t_k\big((1-\theta)(D_1 D_2) + \theta(D_2 D_1 + D_1^2 + D_2^2 - L_h^2)\big)\hat{u}^{(k)} + \dots$$

For $\theta = (1-\theta)$, i.e. $\theta = \frac{1}{2}$ the term after Δt_k becomes zero because $L_h^2 = D_1^2 + D_2^2 + D_1 D_2 + D_2 D_1$. That suggests that this method in general is second order accurate in time like the Crank-Nicholson method. A proof is given in [16, Section 26]. The restriction of the above splitting method that D_1 and D_2 have to commute is no longer necessary. Unfortunately, it is not possible to generalise this result for splitting methods with more than two components.

Finally, we turn to the practically relevant situation where mixed derivative occur. Focusing on the two dimensional case we decompose L_h into two (up to permutation) tridiagonal matrices D_i, each representing derivatives with respect to one dimension, and a matrix $D_{1,2}$ approximating the mixed derivative. The part of the p.d.e. with no derivatives, i.e. $a_0 u$, is represented by a diagonal matrix and can therefore be put into D_1 or D_2

$$L_h = D_1 + D_2 + D_{1,2}.$$

The difficulty with $D_{1,2}$ is that it is not of tridiagonal but rather of block tridiagonal form and it is inefficient to solve an equation system like $D_{1,2} z = b$. Hence, the part $D_{1,2}$ is considered to be explicit in the splitting method to avoid solving such equation systems. Again, the basic idea is to introduce half time steps. Focusing on approximating the Crank-Nicholson method ($\theta = \frac{1}{2}$) we derive from the θ-method (4.8)

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k+1/2)} + \hat{u}^{(k+1/2)} - \hat{u}^{(k)}}{\Delta t_k} = (D_1 + D_2)\frac{1}{2}\big(\hat{u}^{(k+1)} + \hat{u}^{(k)}\big) + D_{1,2}\hat{u}^{(k+1/2)}.$$

In order not to loose one order of consistency in time the operator $D_{1,2}$ ideally would have to be applied to the average of $\hat{u}^{(k)}$ and $\hat{u}^{(k+1)}$. As a fairly good approximation we have chosen $\hat{u}^{(k+1/2)}$. The splitting method is then defined by

$$\begin{aligned}\frac{\hat{u}^{(k+1/2)}-\hat{u}^{(k)}}{\Delta t_k} &= \frac{1}{2}D_1\left(\hat{u}^{(k+1/2)}+\hat{u}^{(k)}\right),\\ \frac{\hat{u}^{(k+1)}-\hat{u}^{(k+1/2)}}{\Delta t_k} &= \frac{1}{2}D_2\left(\hat{u}^{(k+1)}+\hat{u}^{(k+1/2)}\right)+D_{1,2}\hat{u}^{(k+1/2)}.\end{aligned}$$

An analysis reveals, however, that it does not reach the desired second order consistency:

$$\frac{\hat{u}^{(k+1)}-\hat{u}^{(k)}}{\Delta t_k} = \frac{1}{2}L_h\big(\hat{u}^{(k+1)}+\hat{u}^{(k)}\big)+\frac{\Delta t_k}{2}\big((D_1+D_2)^2+(D_1+2D_2)D_{1,2}-L_h^2\big)\hat{u}^{(k)}+\dots.$$

In order to improve accuracy one has somehow to find a better approximation to $\varphi(t_{k+1/2})$ to what the operator $D_{1,2}$ has to be applied. We call this the predictor step. High accuracy is not needed but first order accuracy in time is necessary. That is not difficult to achieve. One can for example use a splitting method for the half step which is implicit for D_1, D_2 and explicit in $D_{1,2}$. Having found a good approximation to φ at time point $t_{k+1/2}$ the entire operator L_h can be applied to this approximation. This strategy is motivated by the fact that the time derivative is second order consistent to the real value only at the time $(t_k+t_{k+1})/2$ and so is the entire scheme if the space discretisation L_h is applied to $\hat{u}$ at the same point in time. As a whole this method is called predictor-corrector scheme. The predictor step in the example described is

$$\begin{aligned}\frac{\hat{u}^{(k+1/4)}-\hat{u}^{(k)}}{2\Delta t_k} &= D_1\hat{u}^{(k+1/4)},\\ \frac{\hat{u}^{(k+1/2)}-\hat{u}^{(k+1/4)}}{2\Delta t_k} &= D_2\hat{u}^{(k+1/2)}+D_{1,2}\hat{u}^{(k+1/4)}\end{aligned} \tag{4.13}$$

and the corrector step can be written as

$$\frac{\hat{u}^{(k+1)}-\hat{u}^{(k)}}{\Delta t_k} = L_h\hat{u}^{(k+1/2)}. \tag{4.14}$$

In [16, Section 23] the method is shown to be stable and consistent for the simple case of $D_{1,2}=0$.

There exists a variety of different splitting methods. A method which is second order consistent and unconditionally stable for equations with mixed derivatives is for example shown in [17].

Chapter 5

Analysis of the finite difference method

If for a numerical method it can not be shown convergence to the analytical solution in the limit case $h \to 0$ and $\tau \to 0$, their results have to be used with caution. The main objective of this chapter is to show convergence of the finite difference method for general parabolic p.d.e.s without the restriction to be uniformly parabolic. First the general theory of error analysis is described and the important concepts of consistency and stability are introduced. For the initial value problem and p.d.e.s with constant coefficients simple criteria can be found to determine whether a scheme is stable. That is illustrated in Section 5.2. The description of the theory is mainly based on [32]. In [32, Chapter 3] the mixed initial boundary value problem is treated, mainly requiring uniform parabolicity. Hence we can not apply that to the p.d.e.s derived from stochastic volatility models. In Section 5.3 we generalise these results and show unconditionally stability for the Crank-Nicholson method under zero Dirichlet boundary conditions.

In order to conclude convergence of a scheme it is essential that the original problem is well posed which in particular means that the p.d.e. has a unique solution. Under the assumption of the uniform parabolicity of the Heston p.d.e. (can be achieved by displacing the zero volatility boundary) it is shown in [8, Chapter 24] that there exists a unique solution in the weak sense. Semi analytic formulas (represented by integrals) are known for plain vanilla options (see [10] or [8, Chapter 23]) and under some restrictions also for barrier options (see [14] or [6]). Any following statement about convergence implicitly assumes the well posedness of the problem.

5.1 Basic error analysis

The concepts of consistency[1], stability and convergence are introduced and applied to numerical schemes for parabolic equations.

5.1.1 Error analysis of general operator equations

In order to illustrate the basic ideas in error analysis techniques we begin with a general linear operator equation where a parabolic p.d.e. is always the example we have in mind. Let $G \subset \mathbb{R}^{d+1}$ be a region of $\mathbb{R}^{d+1}$ (open and simple connected set) and $A : \mathrm{C}^2(G) \to \mathrm{C}(G)$ a linear operator. Let $f \in \mathrm{C}(G)$ then we look for solutions $u \in \mathrm{C}^2(G)$ of

$$Au = f.$$

To solve that numerically we need to approximate u, f and the operator A in a finite dimensional space. By some approximation method, for instance the finite difference method, we then obtain a linear equation system

$$A_h \hat{u} = f_h \tag{5.1}$$

[1] Based on [32, Section 3] we mainly call a scheme accurate instead of consistent to indicate that the approximation order is better than o(1).

defined over the space of grid functions Φ_h. If we identify the space of grid functions with the $\mathbb{R}^n$, n equal to the number of grid points, the grid functions $\hat{u}$ and f_h can be represented by n-dimensional vectors and A_h by an $n \times n$ matrix. The parameter $h > 0$ describes the fineness of the discretisation, for example the distance between two adjacent grid points of the grid $\bar{\Omega}_h$. It also suggests that we are provided with a series of approximations and that we are interested in the limit case where h approaches zero. The dimension n of the approximation space is somehow indirectly related to the discretisation parameter h, i.e. as h approaches zero n goes to infinity. We call the series of approximations convergent if the solution $\hat{u}$ converges in a certain sense to u for $h \to 0$. To put this idea on a more precise basis we need to introduce the projection operator $P_h : \mathrm{C}(G) \to \Phi_h$ and the interpolation operator $Q_h : \Phi_h \to \mathrm{C}^2(G)$. One particular example is the operator which restricts a function to the grid points, i.e. $P_h f : \bar{\Omega}_h \to \mathbb{R}$, $P_h f(x) = f(x)$ for all $x \in \bar{\Omega}_h$. Now, the objective is to find an expression for the error which can be represented with the introduced projection operator by $P_h u - \hat{u}$. If this error converges in a certain norm to zero as $h \to 0$ the approximation method is called convergent. Examples of discrete norms defined over the space of grid points are the maximum norm $\|\hat{u}\| := \max_{x \in \bar{\Omega}_h} |\hat{u}(x)|$ or the L_2 norm which is on uniform grids defined by $\|\hat{u}\| := h^{d/2} \sqrt{\sum_{x \in \bar{\Omega}_h} \hat{u}(x)^2}$.

In order to analyse the approximation error we multiply with the discrete operator A_h and obtain

$$A_h(P_h u - \hat{u}) = A_h P_h u - f_h.$$

The expression on the right hand side is called the consistency or truncation error and describes how well the exact solution u projected to Φ_h fulfils the discrete equation system $A_h \hat{u} = f_h$. Multiplying both sides with A_h^{-1} results in the fundamental relation

$$\boldsymbol{P_h u - \hat{u} = A_h^{-1}(A_h P_h u - f_h)} \tag{5.2}$$

which says that the approximation error is the result of the multiplication of the inverse discrete operator A_h^{-1} with the consistency error. If A_h^{-1} is bounded by a constant independent of h the scheme is said to be stable, i.e. if $\left\|A_h^{-1} v\right\| \leq c \|v\| \; \forall v \in \Phi_h, h > 0$. The immediate conclusion is that if an approximation method is consistent and stable it is also convergent with respect to the same norm. Under certain assumptions one can show even more. By the famous Lax equivalence theorem a consistent scheme is convergent if and only if it is stable. The necessity of stability is more difficult to show and uses some fundamental theorems of functional analysis.

Proving consistency is quite often the simplest part of an error analysis. It might sometimes be advantageous to use a second linear projection operator $\tilde{P}_h$ different from P_h. If we furthermore define $f_h := \tilde{P}_h f$ in the discretisation we then see that using the relation $\tilde{P}_h(Au - f) = 0$ we can estimate the consistency error denoted by γ_h with

$$\gamma_h := A_h P_h u - f_h = A_h P_h u - \tilde{P}_h A u + \tilde{P}_h f - f_h = A_h P_h u - \tilde{P}_h A u \tag{5.3}$$

which can be examined componentwise using Taylor series expansion for instance.

Example 5.1.1 (Consistency of $u_{xx} + u_{yy} = 0$)
To clarify the idea we consider for a moment the two dimensional Laplace equation, i.e. $Au := \triangle u$ and $f = 0$. Abbreviating $\hat{u}_{i,j} := \hat{u}(x_i, y_j)$ the finite difference method leads on a uniform grid to

$$(A_h \hat{u})_{i,j} = \frac{1}{h^2} \left(-4\hat{u}_{i,j} + \hat{u}_{i+1,j} + \hat{u}_{i-1,j} + \hat{u}_{i,j+1} + \hat{u}_{i,j-1}\right)$$

and Taylor series expansion of $u(x_{i\pm1}, y_{j\pm1})$ around (x_i, y_j) gives the following approximation (see also Section 4.2.1)

$$\begin{aligned}
(P_h A u)_{i,j} &= \triangle u(x_i, y_j) \\
&= \frac{1}{h^2} \left(-4u(x_i, y_j) + u(x_{i+1}, y_j) + u(x_{i-1}, y_j) + u(x_i, y_{j+1}) + u(x_i, y_{j-1})\right) + R_4(h) \\
&= (A_h P_h u)_{i,j} + R_4(h)
\end{aligned}$$

with

$$|R_4(h)| \leq \frac{2}{4!} \left(\left\| \frac{\partial^4 u}{\partial x^4} \right\|_{\mathrm{C}[a,b]^2} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{\mathrm{C}[a,b]^2} \right) h^2.$$

It follows immediately for the consistency error

$$(P_h A u - A_h P_h u)_{i,j} = R_4(h).$$

Returning to the general case, investigating stability might be more difficult since it involves properties of the inverse operator A_h^{-1} which one most likely does not know a priori but can be calculated numerically. If it is possible to estimate eigenvalues of A_h then one can conclude whether A_h^{-1} is bounded. As we will see later on this is possible for the finite difference method in an unbounded domain if the coefficients of the p.d.e. are constant. Otherwise one has to use other estimation techniques.

5.1.2 Error analysis of parabolic p.d.e.s

The results of the general case can directly be applied to parabolic p.d.e.s. In order to do that we first have to recall some basic notations already introduced in Section 4.3 where the finite difference method is described. The first d components of the $\mathbb{R}^{d+1}$ are conceived as the representation of space and the last component is understood to be the representation of time. Let $\Omega \subset \mathbb{R}^d$ be a region of the space and $\bar{\Omega}_h = \left\{x^{(i)} : i \in I_h\right\} \subset \bar{\Omega}$ a structured grid consisting of n grid points. The spatial discretisation parameter $h > 0$ might be defined as the greatest distance between two adjacent grid points. For simplicity of notation let the time interval $[0, T]$ be represented by the uniform grid $\{0, t_1, \ldots, t_{m_0-1}, T\}$ and we set $\tau := t_{i+1} - t_i$. The projection operator mapping continuous function to grid functions over the space time grid can now be defined by $P_{h,\tau} u := \bar{u}$ with

$$\bar{u}^{(k)}(x) := u(x, t_k), \qquad \forall x \in \bar{\Omega}_h.$$

With the time independent linear elliptic differential operator $L : \mathrm{C}^2(\Omega) \to \mathrm{C}(\Omega)$ a parabolic p.d.e. can be written in the form

$$\begin{aligned} \frac{\partial}{\partial t} u &= Lu + f, \\ u(x, 0) &= v(x). \end{aligned}$$

General two step finite difference methods can be represented by two $n \times n$ matrices forming the equation

$$\frac{1}{\tau} A_{h,\tau} \hat{u}^{(k+1)} = \frac{1}{\tau} B_{h,\tau} \hat{u}^{(k)} + f_h(k). \tag{5.4}$$

Note that by definition $\bar{u}$ is the projection of the function u which is the exact solution of the p.d.e. However, the vector $\hat{u}$ in the first place is not at all related to u but is the solution of the discrete system. Under the assumptions of consistency and stability $\hat{u}$ finally converges to $\bar{u}$.

In the θ-method (4.9) for example with the discrete space operator L_h we know that $A_{h,\tau} = (I - \theta\tau L_h)$ and $B_{h,\tau} = (I + (1-\theta)\tau L_h)$. With the solution matrix $E_{h,\tau} := A_{h,\tau}^{-1} B_{h,\tau}$ which applied to the solution in a certain time step k gives the solution of the time step $k+1$ we can write the method in an explicit way:

$$\begin{aligned} \hat{u}^{(1)} &= E_{h,\tau}\bar{v} + \tau A_{h,\tau}^{-1} f_h^{(0)} \\ \hat{u}^{(2)} &= E_{h,\tau}\hat{u}^{(1)} + \tau A_{h,\tau}^{-1} f_h^{(1)} \\ &= E_{h,\tau}^2 \bar{v} + \tau E_{h,\tau} A_{h,\tau}^{-1} f_h^{(0)} + \tau A_{h,\tau}^{-1} f_h^{(1)} \\ &\vdots \\ \hat{u}^{(k)} &= E_{h,\tau}^k \bar{v} + \tau \sum_{k=1}^{m_0} E_{h,\tau}^{m_0-k} A_{h,\tau}^{-1} f_h^{(k)}. \end{aligned} \tag{5.5}$$

For homogeneous equations ($f = 0$) everything simplifies to

$$\hat{u}^{(k)} = E_{h,\tau}^k \bar{v}. \tag{5.6}$$

It is thus not surprising that the stability of the method is directly connected to the boundedness of the m_0-th power of $E_{h,\tau}$. A more detailed analysis as shown below reveals that the boundedness of $A_{h,\tau}^{-1}$ is important as well.

After these introductory remarks we are now able to apply the general theory described in the above subsection. We approximate the operator equation

$$\begin{aligned} \left(\frac{\partial}{\partial t} - L\right) u &= f, \\ u(x, 0) &= v(x) \end{aligned}$$

with the discretised version $A_h\hat{u} = f_h$ where A_h represented as a matrix is of order $n(m_0+1)$ and consists of the sub-matrices $A_{h,\tau}$ and $B_{h,\tau}$ as follows:

$$\frac{1}{\tau}\begin{pmatrix} \tau I & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} \\ -B_{h,\tau} & A_{h,\tau} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & -B_{h,\tau} & A_{h,\tau} & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbb{O} & \mathbb{O} & \mathbb{O} & \cdots & A_{h,\tau} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} & \cdots & -B_{h,\tau} & A_{h,\tau} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{O} & \cdots & \mathbb{O} & -B_{h,\tau} & A_{h,\tau} \end{pmatrix}\begin{pmatrix} \hat{u}^{(0)} \\ \hat{u}^{(1)} \\ \hat{u}^{(2)} \\ \vdots \\ \hat{u}^{(m-2)} \\ \hat{u}^{(m-1)} \\ \hat{u}^{(m)} \end{pmatrix} = \begin{pmatrix} \bar{v} \\ f_h^{(0)} \\ f_h^{(1)} \\ \vdots \\ f_h^{(m-3)} \\ f_h^{(m-2)} \\ f_h^{(m-1)} \end{pmatrix}.$$

Hereby the vector $\bar{v}$ is the projection of the initial value function $v : \Omega \to \mathbb{R}$. Consequently, the inverse of the operator A_h is

$$\begin{pmatrix} \hat{u}^{(0)} \\ \hat{u}^{(1)} \\ \hat{u}^{(2)} \\ \vdots \\ \hat{u}^{(m-2)} \\ \hat{u}^{(m-1)} \\ \hat{u}^{(m)} \end{pmatrix} = \tau \begin{pmatrix} \frac{1}{\tau}I & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \frac{1}{\tau}E_{h,\tau} & A_{h,\tau}^{-1} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \frac{1}{\tau}E_{h,\tau}^2 & E_{h,\tau}A_{h,\tau}^{-1} & A_{h,\tau}^{-1} & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{1}{\tau}E_{h,\tau}^{m_0-2} & E_{h,\tau}^{m_0-3}A_{h,\tau}^{-1} & E_{h,\tau}^{m_0-4}A_{h,\tau}^{-1} & \cdots & A_{h,\tau}^{-1} & \mathbb{O} & \mathbb{O} \\ \frac{1}{\tau}E_{h,\tau}^{m_0-1} & E_{h,\tau}^{m_0-2}A_{h,\tau}^{-1} & E_{h,\tau}^{m_0-3}A_{h,\tau}^{-1} & \cdots & E_{h,\tau}A_{h,\tau}^{-1} & A_{h,\tau}^{-1} & \mathbb{O} \\ \frac{1}{\tau}E_{h,\tau}^{m_0} & E_{h,\tau}^{m_0-1}A_{h,\tau}^{-1} & E_{h,\tau}^{m_0-2}A_{h,\tau}^{-1} & \cdots & E_{h,\tau}^2A_{h,\tau}^{-1} & E_{h,\tau}A_{h,\tau}^{-1} & A_{h,\tau}^{-1} \end{pmatrix}\begin{pmatrix} \bar{v} \\ f_h^{(0)} \\ f_h^{(1)} \\ \vdots \\ f_h^{(m-3)} \\ f_h^{(m-2)} \\ f_h^{(m-1)} \end{pmatrix}.$$

The consistency or truncation error defined in equation (5.3) is

$$\begin{pmatrix} \mathbb{O} \\ \gamma_{h,\tau}^{(1)} \\ \gamma_{h,\tau}^{(2)} \\ \vdots \\ \gamma_{h,\tau}^{(m_0)} \end{pmatrix} = \frac{1}{\tau}\begin{pmatrix} \tau\bar{u}^{(0)} \\ A_{h,\tau}\bar{u}^{(1)} - B_{h,\tau}\bar{u}^{(0)} \\ A_{h,\tau}\bar{u}^{(2)} - B_{h,\tau}\bar{u}^{(1)} \\ \vdots \\ A_{h,\tau}\bar{u}^{(m_0)} - B_{h,\tau}\bar{u}^{(m_0-1)} \end{pmatrix} - \begin{pmatrix} \bar{v} \\ f^{(0)} \\ f^{(1)} \\ \vdots \\ f^{(m-1)} \end{pmatrix}.$$

We call the component $\gamma_{h,\tau}^{(k)}$ the truncation error at the time step k and according to (5.3) it is equal to

$$\gamma_{h,\tau}^{(k)} = \frac{1}{\tau}\left(A_{h,\tau}\bar{u}^{(1)} - B_{h,\tau}\bar{u}^{(0)}\right) - \tilde{P}_h\left(\frac{\partial u}{\partial t} - Lu\right) \tag{5.7}$$

if the right hand side is defined as $f_h = \tilde{P}_h f$. The relation between convergence, consistence and stability as seen in (5.2) is given by $P_h u - \hat{u} = A_h^{-1}(A_h P_h u - f_h) = A_h^{-1}\gamma_h$. With the knowledge of the inverse of the big matrix A_h and the definition of the truncation error we conclude at once that the error at the time step k is given by

$$\boldsymbol{(P_h u - \hat{u})^{(k)} = \tau\sum_{i=1}^{k} E_{h,\tau}^{k-i}A_{h,\tau}^{-1}\gamma_{h,\tau}^{(i)}}. \tag{5.8}$$

This result is of fundamental importance in local error analysis. Since the truncation error $\gamma_{h,\tau}^{(i)}$ can be estimated using Taylor series expansion we can directly deduce the approximation error at any time step k. Furthermore, with the knowledge of the spatial distribution of the truncation error, which depends on the elliptic operator L and the choice of the non uniform grid, one can estimate the distribution of the approximation error. Formula (5.8) tells that this error is a sum of the terms $\hat{v}^{(1)}$, $\hat{v}^{(2)}$, ..., $\hat{v}^{(k)}$ where $\hat{v}^{(i)} := E_{h,\tau}^{k-i}A_{h,\tau}^{-1}\gamma_{h,\tau}^{(i)}$. However, it is mostly impracticable to obtain sufficient information about the operator $E_{h,\tau}$ which would allow us to perform a local error analysis. Comparing with the finite difference method (5.6) it becomes clear that the single error terms are approximations to solution of the p.d.e. with the truncation error as part of the initial condition $\bar{v} = A_{h,\tau}^{-1}\gamma_{h,\tau}^{(i)}$. This point of view opens the way for a local error analysis without examining the matrix $E_{h,\tau}$. Assuming that the operator is convergent and given we know a fundamental solution G of the p.d.e. we can state that for one $x \in \Omega_h$

$$\hat{v}^{(i)}(x) \approx \int_\Omega G(x, x', t_i)Q_h(A_{h,\tau}^{-1}\gamma_{h,\tau}^{(i)})(x')\,\mathrm{d}x'$$

and hence

$$
\begin{aligned}
(P_h u - \hat{u})^{(k)}(x) &\approx \tau \sum_{i=1}^{k} \int_\Omega G(x, x', t_i) Q_h (A_{h,\tau}^{-1} \gamma_{h,\tau}^{(i)})(x')\,\mathrm{d}x' \\
&\approx \int_\Omega \int_0^{t_k} G(x, x', t) Q_h (A_{h,\tau}^{-1} \gamma_{h,\tau})(x', t)\,\mathrm{d}t\,\mathrm{d}x' \\
&\approx \int_\Omega \int_0^{t_k} G(x, x', t)\,\mathrm{d}t Q_h (A_{h,\tau}^{-1} \gamma_{h,\tau})(x')\,\mathrm{d}x'
\end{aligned}
\tag{5.9}
$$

where the last approximation only holds if the truncation error is similarly distributed over time. If one can show that the operator $A_{h,\tau}^{-1}$ does not have a big effect the influence of the truncation error at a point x' on the numerical solution at x is then expressed by the the fundamental solution averaged over the time period $[0, t_k]$. Although these estimates are quite remote one can at least get an insight in local errors.

Writing the relation of the truncation and approximation error in a different way, i.e. like $A_h(P_h u - \hat{u}) = \gamma_h$ we see that the approximation error $z := P_h u - \hat{u}$ is the solution of the finite difference method applied to the same p.d.e. but with the right hand side given by the truncation error and the initial condition $z^{(0)} = 0$:

$$
\boldsymbol{A_{h,\tau} z^{(k+1)} = B_{h,\tau} z^{(k)} + \tau \gamma_{h,\tau}^{(k+1)}}. \tag{5.10}
$$

For non uniform time steps one has to replace τ by $\Delta t_k := t_{k+1} - t_k$. If for example it turns out that the truncation error remains not constant in each time step then it is advantageous to choose different time steps. Smaller time steps reduce the influence of the truncation error. One example where one should use non uniform time steps is if the initial condition contains discontinuities in a derivative or even in the values. The truncation error is then very big at the beginning until these discontinuities are no longer visible due to the effect of diffusion. Especially for reverse barrier options the discontinuity is huge directly next to the knockout boundary.

Since the ideas illustrated above are crucial in further error analysis we give the definitions of consistency and stability according to [32, Section 3] and mention the Lax equivalence theorem for the special case of parabolic p.d.e.s.

Definition 5.1.2 (Consistency)
A numerical scheme is called consistent with the parabolic differential equation if the truncation error for a sufficiently smooth solution u tends uniformly to zero as the discretisation parameters h and τ approach zero, i.e.

$$
\gamma_{h,\tau}^{(k)}(x) := \frac{1}{\tau} A_{h,\tau} \bar{u}^{(k)}(x) - B_{h,\tau} \bar{u}^{(k-1)}(x) - \tau f_h^{(k)}(x) \to 0, \quad \forall x \in \bar{\Omega}_h, \quad (h, \tau \to 0).
$$

The scheme is said to be accurate of order μ in x and ν in t if additionally the order of convergence is uniformly for all grid and time points

$$
\gamma_{h,\tau,i}^{(k)} := \gamma_{h,\tau}^{(k)}(x^{(i)}) = \mathrm{O}(h^\mu + \tau^\nu), \quad (h, \tau \to 0).
$$

The requirement that the convergence is uniform, i.e. the ϵ-estimates have to be independent of i, is equivalent to the formulation that the maximum norm of $\gamma_{h,\tau}$ converges. One sometimes says that a method is consistent or accurate in the maximum norm. Since in literature the above definition of consistency is the most common we do not emphasise the maximum norm convergence and only say it is consistent.

Definition 5.1.3 (Stability)
Let $(\Phi_h, \|\cdot\|_h)$ be a series of normed spaces of the spatial variables. The difference scheme (5.5) is said to be stable with respect to the normed space if the discrete operator $E_{h,\tau}^{m_0}$ is bounded with a common constant for any discretisation parameters $h > 0$ and $\tau > 0$, i.e.

$$
\left\| E_{h,\tau}^k v \right\|_h \le c \left\| v \right\|_h, \quad \forall v \in \Phi_h, k \in \{1, \ldots, m_0\}, h > 0, \tau > 0.
$$

Note that the variable $m_0 \in \mathbb{N}$ is directly linked to the discretisation parameters τ, respectively, i.e. if $\tau \to 0$ then $m_0 \to \infty$.

The Lax Richtmyer equivalence theorem now says that for a consistent method stability is necessary and sufficient for convergence if the original problem is well posed. For more details see the original work [13] or [32, Section 3]. The theorem is a remarkable achievement in numerical analysis. It is of simple structure, does not require much information about the schemes except consistency and stability and links these properties in an equivalent way together with the much more difficult to prove and most interesting property of convergence.

Unfortunately, it does not establish a connection between the order of accuracy and the order of convergence. Relation (5.8) suggests that order of consistency and order of convergence are the same given the scheme is stable and the solution u is sufficiently smooth. Additionally it requires that $A_{h,\tau}^{-1}$ is bounded in the considered normed space. The relation between accuracy and convergence follows directly from (5.8) since

$$\begin{aligned}\left\|(P_h u - \hat{u})^{(m_0)}\right\| &\leq \tau \sum_{i=1}^{m_0} \left\|E_{h,\tau}^{m_0-k}\right\| \left\|A_{h,\tau}^{-1}\right\| \left\|\gamma_{h,\tau}^{(k)}\right\| \leq \tau m_0 c \left\|A_{h,\tau}^{-1}\right\| \max_{k=1\dots m_0} \left\|\gamma_{h,\tau}^{(k)}\right\| \\ &\leq Tc \left\|A_{h,\tau}^{-1}\right\| C(u,T)(h^\mu + \tau^\nu).\end{aligned}$$

As it turns the requirement that $A_{h,\tau}^{-1}$ is bounded does not need to be checked for convergence analysis in the methods demonstrated in the following sections.

Finally, we prove a statement about the θ method which reveals its accuracy in time.

Lemma 5.1.4 (Consistency of the θ-method)
Let the discrete operator L_h be accurate of order μ to the differential operator in space L and let additionally the truncation error of L_h applied to $v(x,t) := \theta u(x, t + \Delta t_k) + (1-\theta)u(x,t)$ denoted by $\lambda_{h,i}^{(k)} := \left(L_h \bar{v}^{(k)}\right)_i - Lv(x_i, t_k)$ be of order h^μ uniformly for all $k \in \{0, \dots, m_0\}$ if the solution u is sufficiently smooth. Let additionally $f_h^{(k)}$ be defined so that it consists of the function values of f at the time $t_{k+1/2}$, i.e. $f_h^{(k)} := \left(f(x_i, t_k + \frac{1}{2}\tau)\right)_{i_1,\dots,i_d=0}^{m_1,\dots,m_d}$ then the θ-method (4.8)

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k)}}{\Delta t_k} - L_h \left(\theta \hat{u}^{(k+1)} + (1-\theta)\hat{u}^{(k)}\right) = f_h^{(k)}$$

is accurate of order μ in x and of order one in t to the p.d.e.

$$\left(\frac{\partial}{\partial t} - L\right) u = f$$

and the truncation error of the method is bounded by

$$\begin{aligned}\left|\gamma_{h,\tau,i}^{(k+1)}\right| &\leq \left(\frac{1}{6}\left\|u_{ttt}\right\|_{C(\bar{\Omega}\times[0,T])} + \frac{1}{2}\left\|Lu_{tt}\right\|_{C(\bar{\Omega}\times[0,T])}\right)(\Delta t_k/2)^2 \\ &\quad + \left|(2\theta - 1)Lu_t(x_i, t_{k+1/2})(\Delta t_k/2) + \lambda_{h,i}^{(k)}\right| \\ &= O(h^\mu + \tau).\end{aligned} \tag{5.11}$$

For $\theta = \frac{1}{2}$ the scheme is obviously accurate of order two in t.

Proof For the θ-method we have $A_{h,\tau} = (I - \tau\theta L_h)$ and $B_{h,\tau} = (I + \tau(1-\theta)L_h)$ and so the truncation error is given by

$$\begin{aligned}\gamma_{h,\tau}^{(k+1)} &= \frac{1}{\tau}\left(A_{h,\tau}\bar{u}^{(k+1)} - B_{h,\tau}\bar{u}^{(k)}\right) - \tilde{P}_h\left(\frac{\partial u}{\partial t} - Lu\right) \\ &= \frac{\bar{u}^{(k+1)} - \bar{u}^{(k)}}{\Delta t_k} - L_h\left(\theta\bar{u}^{(k+1)} + (1-\theta)\bar{u}^{(k)}\right) - \tilde{P}_h\left(\frac{\partial u}{\partial t} - Lu\right)\end{aligned}$$

if $f_h = \tilde{P}_h f$. As it turns out it is appropriate to choose $\tilde{P}_h$ so that it projects to the values in the spatial grid points and in time to the values of the middle of the time interval, more precisely $(\tilde{P}_h u)_i^{(k)} := u(x_i, t_k + \frac{1}{2}\Delta t_k)$. For a sufficiently smooth solution u we estimate the truncation error componentwise using the abbreviation $t_{k+1/2} := t_k + \frac{1}{2}\Delta t_k$:

$$\gamma_{h,\tau,i}^{(k+1)} = \frac{u(x_i, t_{k+1}) - u(x_i, t_k)}{\Delta t_k} - \frac{\partial u}{\partial t}(x_i, t_{k+\frac{1}{2}}) + Lu(x_i, t_{k+\frac{1}{2}}) - L_h\left(\theta\bar{u}^{(k+1)} + (1-\theta)\bar{u}^{(k)}\right)_i \tag{5.12}$$

For the first part we find by Taylor series expansion around the time point $t_{k+1/2}$

$$\left|\frac{u(x_i,t_{k+1})-u(x_i,t_k)}{\Delta t_k}-\frac{\partial u}{\partial t}(x_i,t_{k+1/2})\right| \le \frac{(\Delta t_k/2)^2}{6}\,\|u_{ttt}\|_{\mathrm{C}(\bar\Omega\times[0,T])}$$

and the second part can be estimated using the truncation error for the spatial operator $\lambda_{h,i}^{(k)} := \left(L_h\bar v^{(k)}\right)_i - Lv(x_i,t_k)$ with $v(x,t) := \theta u(x,t+\Delta t_k) + (1-\theta)u(x,t)$:

$$\begin{aligned} L_h\left(\theta\bar u^{(k+1)} + (1-\theta)\bar u^{(k)}\right)_i &= \theta Lu(x_i,t_{k+1}) + (1-\theta)Lu(x_i,t_k) + \lambda_{h,i}^{(k)} \\ &= Lu(x_i,t_{k+1/2}) + (2\theta-1)Lu_t(x_i,t_{k+1/2})(\Delta t_k/2) \\ &\quad + \frac{1}{2}\big(\theta Lu_{tt}(x_i,\xi_1) + (1-\theta)Lu_{tt}(x_i,\xi_2)\big)(\Delta t_k/2)^2 + \lambda_{h,i}^{(k)} \end{aligned}$$

with $\xi_1 \in [t_{k+1/2},t_{k+1}]$ and $\xi_2 \in [t_k,t_{k+1/2}]$. Since L is assumed to be a differential operator and u sufficiently smooth the operators $\frac{\partial}{\partial t}$ and L commute which has been used in the last Taylor series expansion. Putting both estimates together into (5.12) we conclude that the truncation error is uniformly for $i \in \mathbb{N}^d$ and $k \in \mathbb{N}$ bounded by

$$\begin{aligned} \left|\gamma_{h,\tau,i}^{(k+1)}\right| &\le \left(\frac{1}{6}\|u_{ttt}\|_{\mathrm{C}(\bar\Omega\times[0,T])} + \frac{1}{2}\|Lu_{tt}\|_{\mathrm{C}(\bar\Omega\times[0,T])}\right)(\Delta t_k/2)^2 \\ &\quad + \left|(2\theta-1)Lu_t(x_i,t_{k+1/2})(\Delta t_k/2) + \lambda_{h,i}^{(k)}\right| \\ &= \mathrm{O}(h^\mu + \tau). \end{aligned}$$

In the Crank-Nicholson scheme ($\theta = \frac{1}{2}$) the first order term in Δt_k vanishes and the truncation error then is

$$\begin{aligned} \left|\gamma_{h,\tau,i}^{(k+1)}\right| &\le \left(\frac{1}{6}\|u_{ttt}\|_{\mathrm{C}(\bar\Omega\times[0,T])} + \frac{1}{2}\|Lu_{tt}\|_{\mathrm{C}(\bar\Omega\times[0,T])}\right)(\Delta t_k/2)^2 + \left|\lambda_{h,i}^{(k)}\right| \\ &= \mathrm{O}(h^\mu + \tau^2). \end{aligned}$$

□

Remark 5.1.5
Together with Lemma 4.2.3 and Remark 4.2.4 this lemma concludes the second order accuracy in space as well as in time for the Crank-Nicholson method with the three point central approximation of derivatives and the space operator L_h as defined in (4.6) for any parabolic p.d.e. with sufficiently smooth and bounded coefficients p_α. That also applies for any non uniform structured grid created by a sufficiently smooth grid generating function for which $g'(z) > 0$, $\forall z \in [0,1]$.

5.1.3 Super-convergent methods for parabolic p.d.e.s

The estimate (5.11) in the above lemma can sometimes be used to construct schemes which are super-convergent, that is if the convergence rate is greater than accuracy suggests. This is for example achievable if one order term in $\lambda_{h,i}^{(k)}$ cancels with $(2\theta-1)Lu_t(x_i,t_{k+1/2})(\Delta t_k/2)$. At first sight that seems to be impossible since the spatial truncation error $\lambda_{h,i}^{(k)}$ will never depend on a time derivative of u. However, for the homogeneous equation one might use the relation that u solves $u_t = Lu$ to rewrite the truncation error

$$\begin{aligned} \left|\gamma_{h,\tau,i}^{(k+1)}\right| &\le \left(\frac{1}{6}\|u_{ttt}\|_{\mathrm{C}(\bar\Omega\times[0,T])} + \frac{1}{2}\|Lu_{tt}\|_{\mathrm{C}(\bar\Omega\times[0,T])}\right)(\Delta t_k/2)^2 \\ &\quad + \left|(2\theta-1)LLu(x_i,t_{k+1/2})(\Delta t_k/2) + \lambda_{h,i}^{(k)}\right|. \end{aligned}$$

For the simple example $u_t = \kappa u_{xx}$ on a uniform grid it can easily be seen how the higher consistency can be achieved. The spatial truncation error is by Taylor series expansion (see also Lemma 4.2.3)

$$\begin{aligned} \lambda_{h,i}^{(k)} &:= \left(L_h\bar v^{(k)}\right)_i - Lv(x_i,t_k) = \kappa\left((D_h^2\bar v^{(k)})_i - v_{xx}(x_i,t_k)\right) \\ &= \frac{\kappa}{12}v_{xxxx}(x_i,t_k)h^2 + \mathrm{O}(h^4) = \frac{\kappa}{12}u_{xxxx}(x_i,t_{k+1/2})h^2 + \mathrm{O}(h^4) + \mathrm{O}(\tau^2) \end{aligned}$$

and we obtain for the truncation error

$$\begin{aligned}\left|\gamma_{h,\tau,i}^{(k+1)}\right| \leq & \left(\frac{1}{6}\left\|u_{ttt}\right\|_{\mathrm{C}(\bar{\Omega}\times[0,T])} + \frac{1}{2}\left\|Lu_{tt}\right\|_{\mathrm{C}(\bar{\Omega}\times[0,T])}\right)(\Delta t_k/2)^2 \\ & + \left|(2\theta-1)\kappa^2 u_{xxxx}(x_i, t_{k+1/2})(\Delta t_k/2) + \frac{\kappa}{12} u_{xxxx}(x_i, t_{k+1/2})h^2 + \mathrm{O}(\tau^2) + \mathrm{O}(h^4)\right|.\end{aligned}$$

If we now choose $\theta < \frac{1}{2}$ and set the time step size according to

$$\tau = \frac{1}{6\kappa(1-2\theta)} h^2$$

the first order term in τ and the second order term in h cancel each other so that the truncation error is of order $\mathrm{O}(h^4 + \tau^2)$. As will be demonstrated in Subsection 5.2.3 this method is stable in the L_2 norm and one can show that it is then convergent in L_2 with the order of accuracy. Comparing to equation (5.24) the time step size is exactly one third of what the stability criterion requires. It is important to note that the super-convergence only applies to the specified p.d.e. $u_t = \kappa u_{xx}$. For other parabolic p.d.e.s the estimates have to be repeated with possibly different results for the parameters. In general, though, the above method will not be able to eliminate the order $\mathrm{O}(h^2)$ in $\lambda_{h,i}^{(k)}$ with the term LLu and hence super-convergence will not be achieved. Fortunately, the idea can be saved if one allows certain extensions of the Crank-Nicholson method for instance the parameter θ might differ for some terms in the p.d.e.

In quite resent research Ronald Smith ([25] to [29]) extensively studied an extension of the Crank-Nicholson method with some more parameters. With this additional degree of freedom he was able to match all the first four order terms in the spatial direction so that the method proposed by him is accurate of the astonishing order $\mathrm{O}(h^5 + \tau^2)$. In [25] the method is stated for the non-homogeneous parabolic p.d.e. with constant coefficients in one dimension. For a simplified discussion we consider the parabolic p.d.e.

$$u_t = \kappa u_{xx} - b u_x - \lambda u.$$

With the three point discrete differential operator $\mathbf{D}_h^2$ and $\mathbf{D}_h^1$ approximating second and first derivative with respect to x the numerical scheme is defined by

$$\begin{aligned}&\frac{\vartheta}{4}\frac{\hat{u}_{i-1}^{(k+1)} - u_{i-1}^{(k)}}{\Delta t} + \frac{4-2\vartheta}{4}\frac{\hat{u}_{i}^{(k+1)} - u_{i}^{(k)}}{\Delta t} + \frac{\vartheta}{4}\frac{\hat{u}_{i+1}^{(k+1)} - u_{i+1}^{(k)}}{\Delta t} = \\ &\qquad = \bar{\kappa}\mathbf{D}_h^2(\theta_2\hat{u}^{(k+1)} + \theta_2\hat{u}^{(k)})_i - \bar{b}\mathbf{D}_h^1(\theta_1\hat{u}^{(k+1)} + \theta_1\hat{u}^{(k)})_i - \lambda(\theta_0\hat{u}^{(k+1)} + \theta_0\hat{u}^{(k)})_i.\end{aligned}$$

The optimal choice of the parameters ϑ, θ_2, θ_1, θ_0, $\bar{\kappa}$ and $\bar{b}$ to achieve the mentioned high accuracy is given in [25]. They depend on the coefficients of the p.d.e. as well as the discretisation parameters Δt and Δx. A generalisation to non uniform grids and non constant coefficients is given in [26] and [28]. The case of a two dimensional p.d.e.s with constant coefficients is covered in [29] where an alternating direction scheme is proposed. As far as I can judge, this method is a very promising approach to achieve more accurate numerical results within less computing time and should the centre of attention in further investigations.

5.2 Analysis of f.d.m. for constant coefficient p.d.e.s

In this chapter a general overview over the main ideas of analysing the error for constant coefficients p.d.e.s is given. The theory is well established and mainly relies on the Fourier transformation for L_2 norm estimates. Even though it is not directly applicable to variable coefficient p.d.e.s and is restricted to uniform grids and the pure initial value problem the theory is nevertheless of high importance. First of all, it gives clear statements under which circumstances a certain scheme is stable and secondly there exists a generalisation to variable coefficient p.d.e.s. Furthermore, convergence information not only for function values but also for derivative values are obtained. For an exact presentation of the L_2 theory which uses the Fourier transformation a very detailed description would be necessary which is out of scope of this thesis. The reader might refer to [32, Section 4] or for a more detailed treatment to [31, chapter 10]. Consequently, only the ideas will be shown and the main theorems be stated.

5.2.1 Introduction and Definitions

We quickly recall some facts of Subsection 3.2.3 which are important in the following treatment. The parabolic p.d.e. with constant coefficients

$$u_t = P(\mathbf{D})u$$

is called parabolic in the sense of Petrovskii if

$$\Re P(\mathrm{i}\omega) \leq -c_1 |\omega|^2 + c_2, \qquad \forall \omega \in \mathbb{R}^d$$

with a positive constant $c_1 > 0$. As derived in Subsection 3.2.3 the solution of the p.d.e. in the Fourier transformed space is

$$\tilde{u}(\omega, t) = \tilde{u}^{(0)}(\omega) \exp\left(P(\mathrm{i}\omega)t\right).$$

The definition of the parabolicity as above makes sense because by Parseval's relation we have

$$\begin{aligned} \|u(\cdot,t)\| = \|\tilde{u}(\cdot,t)\| = \left\|\tilde{u}^{(0)} \exp\left(P(\mathrm{i}\omega)t\right)\right\| &\leq \sup_{\omega\in\mathbb{R}^d} \exp\left((-c_1|\omega|^2 + c_2)t\right) \left\|\tilde{u}^{(0)}\right\| \\ &\leq \exp(c_2 t)\left\|u^{(0)}\right\|, \end{aligned}$$

i.e. the problem is well posed in L_2.

We exclusively consider one step finite difference schemes for the homogeneous pure initial value problem, i.e. $\Omega = \mathbb{R}^d$. We thus bypass the difficulty of boundary conditions but have to deal with infinite domains. Also, the grid in the space dimensions has to be uniform, for simplicity we say $\bar{\Omega}_h := \left\{x^{(i)} : i \in \mathbb{Z}^d\right\}$ and $x^{(i)} = hi$. In order to define stability we have to introduce a normed space and since we want to use Fourier transformation the discrete version of the L_2 space is appropriate, i.e. $(\Phi_h, \|\cdot\|)$ with

$$\|v\| := h^{d/2}\sqrt{\sum_{i\in\mathbb{Z}^d} v_i^2}, \qquad v \in \Phi_h.$$

One step finite difference schemes are represented by $A_{h,\tau}\hat{u}^{(k+1)} = B_{h,\tau}\hat{u}^{(k)}$ or equivalently with $E_{h,\tau} := A_{h,\tau}^{-1}B_{h,\tau}$ by $\hat{u}^{(k+1)} = E_{h,\tau}\hat{u}^{(k)}$. Due to the structure of the approximation of derivatives and the spatial homogeneity of the differential operator the discrete operator are as follows

$$\begin{aligned} (A_{h,\tau}v)_i &= \sum_{k\in I_0} a_k v_{i+k}, \\ (B_{h,\tau}v)_i &= \sum_{k\in I_0} b_k v_{i+k} \end{aligned}$$

where $I_0 \subset \mathbb{Z}^d$ is the finite set of grid point indices which are used to approximate derivatives of a function at the origin. We have to keep in mind that the real numbers a_k etc. are clearly also dependent on the discretisation parameters τ and h and are known explicitly. The simplicity of the operators $A_{h,\tau}$ and $B_{h,\tau}$ will be very useful and is due to the absence of boundary conditions.

The basic idea to examine stability of the difference scheme, which is by definition equivalent to the boundedness of the m_0-th power of the operator $E_{h,\tau}$, is to perform an eigenvalue analysis. Due to the simple structure of $A_{h,\tau}$ and $B_{h,\tau}$ we are even able to explicitly provide the eigenvectors. We define the series of grid functions $\left\{w^{(\omega)} : \omega \in \mathbb{R}^d\right\}$ with

$$w_i^{(\omega)} := \mathrm{e}^{\mathrm{i}\left\langle\omega, x^{(i)}\right\rangle} \tag{5.13}$$

and as one can easily see these are all eigenvectors of $A_{h,\tau}$ and $B_{h,\tau}$:

$$\left(A_{h,\tau}w^{(\omega)}\right)_i = \sum_{k\in I_0} a_k\,\mathrm{e}^{\mathrm{i}\left\langle\omega, x^{(i+k)}\right\rangle} = \mathrm{e}^{\mathrm{i}\left\langle\omega, x^{(i)}\right\rangle}\sum_{k\in I_0} a_k\,\mathrm{e}^{\mathrm{i}\langle\omega, x_k\rangle}.$$

Motivated by this relation one defines the symbol of a discrete operator.

Definition 5.2.1 (Symbol of discrete operators)
The function $\lambda_A : \mathbb{R}^d \to \mathbb{R}$ which assigns each $\omega \in \mathbb{R}^d$ the eigenvalue of the operator $A_{h,\tau}$ to the eigenvector $w^{(\omega)}$ is called the symbol of $A_{h,\tau}$, i.e.

$$A_{h,\tau}w^{(\omega)} = \lambda_A(\omega)w^{(\omega)}.$$

The same applies to the operators $B_{h,\tau}$ and $E_{h,\tau}$.

The eigenvalues are known explicitly and so are the symbols. As seen above the symbol of $A_{h,\tau}$ for example is given by

$$\lambda_A(\omega) = \sum_{k\in I_0} a_k \, \mathrm{e}^{\mathrm{i}\left\langle \omega, x^{(k)}\right\rangle} = \sum_{k\in I_0} a_k \, \mathrm{e}^{\mathrm{i}h\langle \omega, k\rangle} \,. \tag{5.14}$$

Since an eigenvector of $A_{h,\tau}$ is an eigenvector of $B_{h,\tau}$ and vise versa we immediately conclude that the eigenvalues of $A_{h,\tau}^{-1}B_{h,\tau}$ are

$$\lambda_E(\omega) = \frac{\lambda_B(\omega)}{\lambda_A(\omega)}.$$

One might suspect that the boundedness of $E_{h,\tau}^{m_0}$ is strongly related to the boundedness of $\lambda_E(\omega)^{m_0}$. As it turns out both properties are even equivalent. That will be justified by the fact that any grid function $v \in \mathrm{L}_2$ can be represented as an integral over all eigenvectors $w^{(\omega)}$ and will be proved with help of the Fourier transformation. It is remarkable how simple the analysis of stability turns out to be if one uses the symbol of $E_{h,\tau}$.

5.2.2 Analysis using Fourier transformation

As said in the introductory part of this section only the basic ideas and some explanatory remarks on why the results of the theorems make sense are given. To begin, we establish a connection between the Fourier transformation and the principles of the subsection above. Let the projection operator $Q_h : \Phi_h \to \mathrm{L}_2$ be defined as the constant interpolation, i.e. $(Q_h v)(x) := v_i \;\forall x_1 \in [x_1^{(i)}, x_1^{(i)} + h), \ldots, x_d \in [x_d^{(i)}, x_d^{(i)} + h)$. We are now able to introduce the Fourier transformation for the grid function $v \in \Phi_h$ by

$$\tilde{v}(\omega) := \mathscr{F}(Q_h v)(\omega) = (2\pi)^{-d/2} h^d \sum_{i\in\mathbb{Z}^d} v_i \, \mathrm{e}^{-\mathrm{i}\left\langle \omega, x^{(i)}\right\rangle} = (2\pi)^{-d/2} \left\langle v, w^{(-\omega)} \right\rangle$$

and we see that it is simply the scalar product of v with the the vector $w^{(-\omega)}$ defined in (5.13). The further objective is to fully represent the finite difference scheme in the Fourier transformed space. We note that the multiplication $A_{h,\tau}v$ simplifies after transforming due to the simple structure of $A_{h,\tau}$:

$$\begin{aligned}
\mathscr{F}(A_{h,\tau}v)(\omega) &= (2\pi)^{-d/2} h^d \sum_{i\in\mathbb{Z}^d} (A_{h,\tau}v)_i \, \mathrm{e}^{-\mathrm{i}\left\langle \omega, x^{(i)}\right\rangle} = (2\pi)^{-d/2} h^d \sum_{i\in\mathbb{Z}^d} \sum_{k\in I_0} a_k v_{i+k} \, \mathrm{e}^{-\mathrm{i}\left\langle \omega, x^{(i)}\right\rangle} \\
&= (2\pi)^{-d/2} h^d \sum_{k\in I_0} a_k \sum_{i\in\mathbb{Z}^d} v_i \, \mathrm{e}^{-\mathrm{i}\left\langle \omega, x^{(i-k)}\right\rangle} = \sum_{k\in I_0} a_k \, \mathrm{e}^{\mathrm{i}\left\langle \omega, x^{(k)}\right\rangle} \, \tilde{v}(\omega) \\
&= \lambda_A(\omega)\tilde{v}(\omega)
\end{aligned}$$

From these remarks it follows that the finite difference scheme $A_{h,\tau}\hat{u}^{(k+1)} = B_{h,\tau}\hat{u}^{(k)}$ simplifies to a multiplication with the symbols of $A_{h,\tau}$ and $B_{h,\tau}$ in the Fourier transformed space

$$\tilde{\hat{u}}^{(k+1)}(\omega) = \frac{\lambda_B(\omega)}{\lambda_A(\omega)} \tilde{\hat{u}}^{(k)}(\omega) = \lambda_E(\omega)\tilde{\hat{u}}^{(k)}(\omega). \tag{5.15}$$

Comparing the discrete scheme with the analytical solution

$$\tilde{u}(\omega, t) = \tilde{u}^{(0)}(\omega) \exp\left(P(\mathrm{i}\omega)t\right)$$

one suspects a very strong relationship between the two functions $\exp\big(P(\mathrm{i}\omega)\tau\big)$ and $\lambda_E(\omega)$. As it turns out it is consistency which requires both functions to be similar. In order to see that, we apply the definition of the truncation error (5.7) to one particular solution of the p.d.e.

$$u(x,t) = e^{\mathrm{i}\langle \omega, x\rangle} \exp(P(\mathrm{i}\omega)t)$$

where $\omega \in \mathbb{R}^d$ is a fixed frequency. The truncation error then is

$$\begin{aligned}
\gamma_{h,\tau}^{(k)} &= \frac{1}{\tau} \Big(\exp(P(\mathrm{i}\omega)t_{k+1}) A_{h,\tau} w^{(\omega)} - \exp(P(\mathrm{i}\omega)t_k) B_{h,\tau} w^{(\omega)} \Big) \\
&= \frac{1}{\tau} \exp(P(\mathrm{i}\omega)t_k) \left(\exp(P(\mathrm{i}\omega)\tau)\lambda_A(\omega) - \lambda_B(\omega) \right) w^{(\omega)} \\
&= \frac{1}{\tau} \exp(P(\mathrm{i}\omega)t_k) \lambda_A(\omega) \big(\exp(P(\mathrm{i}\omega)\tau) - \lambda_E(\omega) \big) w^{(\omega)} = \mathrm{O}(h^\mu + \tau^\nu)
\end{aligned}$$

which concludes that $\frac{1}{\tau}\lambda_A(\omega)\big(\exp(P(\mathrm{i}\omega)\tau) - \lambda_E(\omega)\big) = \mathrm{O}(h^\mu + \tau^\nu)$ for all $\omega \in \mathbb{R}^d$ but not necessarily uniformly. The dependence of the truncation error on ω can be examined using Taylor expansions of λ_E and the exp function. Motivated by these remarks we give a further definition of consistency.

Definition 5.2.2 (Consistency in the Fourier transformed space)
A one step finite difference scheme (5.4) for a constant coefficient p.d.e. on a uniform grid with a given relation between space and time discretisation $\tau = \sigma(h)$ is called accurate of order μ in space and ρ in ω if there exists a constant $c > 0$ so that

$$\frac{1}{\tau}\left|\exp(P(i\omega)\tau) - \lambda_E(\omega)\right| \le ch^\mu(1 + |\omega|)^\rho, \qquad \forall\, |h\omega| \le \pi \tag{5.16}$$

This definition makes sense as shown by the following theorem.

Theorem 5.2.3 (Consistency)
Let the p.d.e. be parabolic in the sense of Petrovskii. The one step finite difference scheme (5.4) with a predefined relation between space and time discretisation $\tau = \sigma(h)$ is accurate of order μ in space according to definition 5.1.2 if and only if it is accurate of order μ in space and $2 + \mu$ in ω according to definition 5.2.2.

The theorem is a slight modification of Theorem 4.1 in [32, page 45]. A similar theorem with proof can be found in [31, chapter 10].

The criterion for stability defined by the boundedness of the m_0-th power of $E_{h,\tau} := A_{h,\tau}^{-1}B_{h,\tau}$ obviously translates to the boundedness of $\lambda_E(\omega)^{m_0}$ uniformly for all ω. Since $m_0 = \frac{T}{\tau}$ it follows that a constant $c > 0$ exists so that $\lambda_E(\omega) \le 1 + c\tau$. Following [32, Theorem 4.2] we state:

Theorem 5.2.4 (Stability)
A one step finite difference scheme (5.4) for a constant coefficient p.d.e. on a uniform grid is stable in L_2 if and only if there exists a constant $c > 0$ so that

$$|\lambda_E(\omega)| \le 1 + c\tau, \qquad \forall \omega \in \mathbb{R}^d, \tau > 0. \tag{5.17}$$

The above condition is referred to as the von Neumann stability criterion. Following the line of [32, Section 4] and [31, chapter 10] we state the theorem about convergence where the usual $\mathbf{H}^\rho$ space will be used. The $\mathbf{H}^\rho$-norm is defined as

$$\|v\|^2_{\mathbf{H}^\rho} := \sum_{|\alpha|\le\rho} \|\mathbf{D}^\alpha v\|^2_{\mathrm{L}_2}, \qquad \forall t_k \le T.$$

Theorem 5.2.5 (Convergence)
Let the p.d.e. be parabolic in the sense of Petrovskii and the initial condition sufficiently smooth $u^{(0)} \in \mathbf{H}^\rho$. If the one step finite difference scheme (5.4) with a predefined relation between space and time discretisation $\tau = \sigma(h)$ is stable and accurate of order μ in space and ρ in ω with $\rho > \max\left\{\mu, \frac{1}{2}\right\}$ then it is convergent in L_2 of order μ and the error can be estimated by

$$\left\|P_h u(\cdot, t_k) - \hat{u}^{(k)}\right\| \le C_T h^\mu \left\|u^{(0)}\right\|_{\mathbf{H}^\rho}. \tag{5.18}$$

Proof See [31, Theorem 10.1.4] or [32, Theorem 4.6]. □

The idea on why one obtains the same order of convergence as the order of accuracy can be made plausible by considering the error in the L_2-norm and switching to the Fourier transformed space:

$$\begin{aligned}\left\|u(\cdot,t_k) - Q_h\hat{u}^{(k)}\right\|_{\mathrm{L}_2} &= \left\|\tilde{u}(\cdot,t_k) - \tilde{\hat{u}}^{(k)}\right\|_{\mathrm{L}_2} = \left\|\exp(P(\mathrm{i}\cdot)t_k)\tilde{u}^{(0)} - \lambda_E^k\tilde{\hat{u}}^{(0)}\right\|_{\mathrm{L}_2} \\ &\le \sup_{\omega\in\mathbb{R}^d}\left|\exp(P(\mathrm{i}\omega)t_k) - \lambda_E(\omega)^k\right| \left\|\tilde{u}^{(0)}\right\|_{\mathrm{L}_2} + \sup_{\omega\in\mathbb{R}^d}\left|\lambda_E(\omega)^k\right| \left\|\tilde{u}^{(0)} - \tilde{\hat{u}}^{(0)}\right\|_{\mathrm{L}_2}.\end{aligned}$$

The last term describes the difference between the Fourier transformed of the initial condition and the approximated initial condition and is equal to $\left\|\mathscr{F}(u^{(0)} - Q_hP_hu^{(0)})\right\|$. Neglecting this term just for the moment the first term can be estimated using the easy to prove relation $a^k - b^k = (a-b)\sum_{i=0}^{n-1} a^{n-1-i}b^i$ and as $a - b = \exp(P(\mathrm{i}\omega)\tau) - \lambda_E(\omega)$ which is the definition of accuracy of the scheme one can imagine that accuracy and convergence are strongly related. Stability is necessary to estimate the other terms. However a much more thorough analysis is necessary to prove the theorem.

A further interesting question is how accurate are the approximated derivatives of the numerical solution compared to the derivatives of the analytical solution. As it turns out stability is no longer a sufficient condition and we have to introduce a stronger criteria comparing to (5.17).

Definition 5.2.6
A one step finite difference scheme (5.4) for a constant coefficient p.d.e. on a uniform grid is called parabolic in the sense of John if for the symbol of $E_{h,\tau} := A_{h,\tau}^{-1} B_{h,\tau}$ exist constants $\delta > 0$ and $c > 0$ so that for sufficiently small $\tau > 0$

$$|\lambda_E(\omega)| \leq 1 - \delta h^2 \|\omega\|^2 + c\tau, \qquad \forall \omega \in \mathbb{R}^d. \tag{5.19}$$

The following theorem is a special case of [32, Theorem 4.7].

Theorem 5.2.7
Let the p.d.e. be parabolic in the sense of Petrovskii and the initial condition sufficiently smooth $u^{(0)} \in \mathbf{H}^\rho$. We consider the difference operator $\mathbf{D}_h^\alpha$ which is assumed to be accurate of order q to $\mathbf{D}^\alpha$. If the one step finite difference scheme (5.4) with a predefined relation between space and time discretisation $\tau = \sigma(h)$ is accurate of order μ in space and parabolic in John's sense then

$$\left\| P_h \mathbf{D}^\alpha u(\cdot, t_k) - D_h^\alpha \hat{u}^{(k)} \right\| \leq C h^\mu t_k^{-q/2} \left\| u^{(0)} \right\|_{H^\mu}. \tag{5.20}$$

5.2.3 Some examples of schemes

This subsection is devoted to the examination of stability of some particular numerical schemes. The coverage is limited to the θ-method for some simple p.d.e.s and different approximations of derivatives in space direction. First, some general statements are given.

We recall that for the θ-method (4.9) the discrete operators are defined as $A_{h,\tau} := I - \theta\tau L_h$ and $B_{h,\tau} := I + (1-\theta)\tau L_h$. Hence the symbol of $E_{h,\tau}$ is simply

$$\lambda_E(\omega) = \frac{1 + (1-\theta)\tau\lambda_L(\omega)}{1 - \theta\tau\lambda_L(\omega)}, \qquad \omega \in \mathbb{R}^d. \tag{5.21}$$

In some simple cases, for instance if $\lambda_L(\omega)$ is real, the fraction can be simplified. The discrete Laplace operator is one example where the symbol is real. By polynomial division we see that under the assumption of real values equality (5.21) becomes

$$\lambda_E(\omega) = -\frac{1-\theta}{\theta} + \frac{\theta^{-1}}{1 - \theta\tau L_h(\omega)}.$$

If additionally the inequality $\lambda_L(\omega) \leq 0$ holds which hints that the operator L_h is negative semidefinite we can immediately give stability estimates since we conclude from the above equation that

$$-\frac{1-\theta}{\theta} + \frac{\theta^{-1}}{1 + \theta\tau \sup\limits_{\omega\in\mathbb{R}^d} |\lambda_L(\omega)|} \leq \lambda_E \leq 1.$$

Corollary 5.2.8
Let the symbol of the discrete space operator L_h be real and not positive, i.e. $\lambda_L(\omega) \leq 0$, then the θ-method for the pure initial value problem is unconditionally stable in L$_2$ *for all $\frac{1}{2} \leq \theta \leq 1$ and otherwise stable if the condition*

$$\tau \leq \frac{2}{1-2\theta} \frac{1}{\sup\limits_{\omega\in\mathbb{R}^d} |\lambda_L(\omega)|}, \qquad 0 \leq \theta < \frac{1}{2} \tag{5.22}$$

is satisfied.

Proof For $\frac{1}{2} \leq \theta \leq 1$ we know from the above estimate that $-1 \leq \lambda_E \leq 1$ holds automatically. For all other values of $\theta \in [0,1]$ the relation $-1 \leq \lambda_E \leq 1$ is also fulfilled if we additionally restrict the step length by (5.22). It follows by Theorem 5.2.4 that the method is then stable. □

In general the symbol λ_L is complex and one has to calculate in the complex plane in order to evaluate λ_E. However, as only the absolute value of λ_E is relevant for stability the calculation of $|\lambda_E|$ is simpler:

$$\begin{aligned} |\lambda_E|^2 &= \frac{(1+(1-\theta)\tau\Re\lambda_L)^2 + ((1-\theta)\tau\Im\lambda_L)^2}{(1-\theta\tau\Re\lambda_L)^2 + (\theta\tau\Im\lambda_L)^2} \\ &= \frac{1+((1-\theta)\tau)^2 |\lambda_L|^2 + 2(1-\theta)\tau\Re\lambda_L}{1+(\theta\tau)^2 |\lambda_L|^2 - 2\theta\tau\Re\lambda_L}. \end{aligned} \tag{5.23}$$

Corollary 5.2.9
Let the real part of the symbol of the discrete space operator L_h be negative or zero, i.e. $\Re\lambda_L(\omega) \leq 0$, then the θ-method for the pure initial value problem is unconditionally stable in L_2 for all $\frac{1}{2} \leq \theta \leq 1$.

Proof For $\frac{1}{2} \leq \theta \leq 1$ the inequality $(1-\theta) \leq \theta$ holds. As additionally $\Re\lambda_L \leq 0$ it follows that

$$1+((1-\theta)\tau)^2 |\lambda_L|^2 + 2(1-\theta)\tau\Re\lambda_L \leq 1+(\theta\tau)^2 |\lambda_L|^2 - 2\theta\tau\Re\lambda_L$$

and hence by (5.23) it is

$$|\lambda_L|^2 \leq 1.$$

The stability follows immediately by Theorem 5.2.4. □

Example 5.2.10 ($u_t = \kappa u_{xx}$ with central differences)
With the approximation of the second derivative in space on a uniform grid as shown Subsection 4.2.1 the discrete space operator is

$$(L_h v)_i = \frac{\kappa}{h^2}(v_{i-1} - 2v_i + v_{i+1})$$

and by definition (5.14) its symbol is

$$\lambda_L(\omega) = \frac{\kappa}{h^2}(\mathrm{e}^{\mathrm{i}h\omega} - 2 + \mathrm{e}^{-\mathrm{i}h\omega}) = \frac{2\kappa}{h^2}(\cos(h\omega) - 1)$$

which is obviously real and non positive so that Corollary 5.2.8 is applicable. The θ-method is thus unconditionally stable for $\theta \geq \frac{1}{2}$ and otherwise conditionally stable if

$$\tau \leq \frac{1}{2\kappa(1-2\theta)} h^2, \qquad 0 \leq \theta < \frac{1}{2}. \tag{5.24}$$

Example 5.2.11 (Pure diffusion equation with central differences)
The pure diffusion equation in d dimensions is given by

$$u_t = \operatorname{div}(G \nabla u) = \sum_{i,j=1}^{d} g_{i,j} \frac{\partial^2 u}{\partial x_i \partial x_j}$$

with a symmetric positive semidefinite matrix G of order $d \times d$. We allow a uniform grid with different discretisation parameters $(h_1, \ldots, h_d)$. Referring to Subsection 4.2.1 where among others the approximation of mixed derivatives is explained the discrete operator in space is with $k \in \mathbb{Z}^d$

$$(L_h v)_k = \sum_{i=1}^{d} \frac{g_{ii}}{h_i^2}(v_{k-e_i} - 2v_k + v_{k+e_i}) + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \frac{g_{ij}}{h_i h_j}(v_{k+e_i+e_j} + v_{k-e_i-e_j} - v_{k+e_i-e_j} - v_{k-e_i+e_j})$$

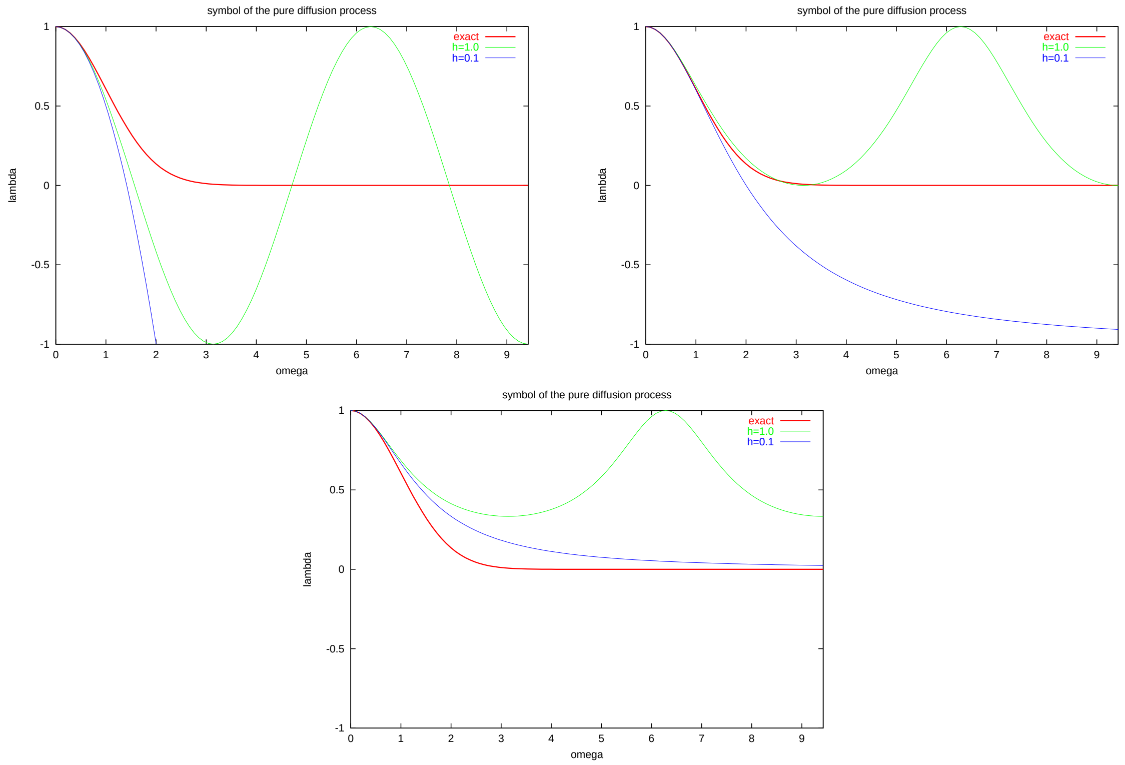

Figure 5.1: Symbol λ_E of $u_t = u_{xx}$ with $\tau = \frac{1}{2}$ and $\theta = 0$, $\theta = \frac{1}{2}$ and $\theta = 1$, respectively

where $e_i \in \mathbb{Z}^d$ represents the i-th unit vector, i.e. $e_1 = (1, 0, \ldots, 0)$, $\ldots$, $e_d = (0, \ldots, 0, 1)$. It follows for the symbol (5.14) of L_h which depends on $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$

$$\begin{aligned}
\lambda_L(\omega) &= \sum_{i=1}^{d} \frac{g_{ii}}{h_i^2} \left(\mathrm{e}^{\mathrm{i} h_i \omega_i} - 2 + \mathrm{e}^{-\mathrm{i} h_i \omega_i} \right) \\
&\quad + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \frac{g_{ij}}{h_i h_j} \left(\mathrm{e}^{-\mathrm{i}(h_i\omega_i + h_j\omega_j)} + \mathrm{e}^{\mathrm{i}(h_i\omega_i + h_j\omega_j)} - \mathrm{e}^{-\mathrm{i}(h_i\omega_i - h_j\omega_j)} - \mathrm{e}^{\mathrm{i}(h_i\omega_i - h_j\omega_j)} \right) \\
&= \sum_{i=1}^{d} \frac{2g_{ii}}{h_i^2} \left(\cos(h_i\omega_i) - 1 \right) + \sum_{\substack{i,j=1 \\ i \neq j}}^{d} \frac{2g_{ij}}{h_i h_j} \left(\cos(h_i\omega_i + h_j\omega_j) - \cos(h_i\omega_i - h_j\omega_j) \right) \\
&= -\sum_{i,j=1}^{d} \frac{4g_{ij}}{h_i h_j} \sin(h_i\omega_i) \sin(h_j\omega_j).
\end{aligned}$$

The relation $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$ has been used to simplify the sum. It immediately follows that the symbol λ_L is real and not positive since G is positive semidefinite and with the definition of the vector $s(\omega) := \left(\frac{1}{h_1} \sin(h_1\omega_1), \ldots, \frac{1}{h_d} \sin(h_d\omega_d) \right)^\tau$ we have

$$\lambda_L(\omega) = -4 s(\omega)^\tau G s(\omega) \leq 0.$$

Hence Corollary 5.2.8 is applicable and we conclude that the θ-method is unconditionally stable for all $\theta \geq \frac{1}{2}$. We first must find the supremum of $|\lambda_L|$ in order to say whether the scheme is stable for all other values θ. If G is a diagonal matrix the maximum is then

$$|\lambda_L(\omega)| = 4 \sum_{i=1}^{d} \frac{g_{i,i}}{h_i^2} \sin^2(h_i\omega_i) \leq 4 \sum_{i=1}^{d} \frac{g_{i,i}}{h_i^2}.$$

Let the discretisation parameters be tight together by the relation $(h_1, \ldots, h_d) = \bar{h}(c_1, \ldots, c_d)$ with the scalar $\bar{h} > 0$. The stability criterion then is

$$\tau \leq \frac{1}{2(1-2\theta) \sum_{i=1}^{d} \frac{g_{i,i}}{c_i^2}} \bar{h}^2, \qquad 0 \leq \theta < \frac{1}{2}. \tag{5.25}$$

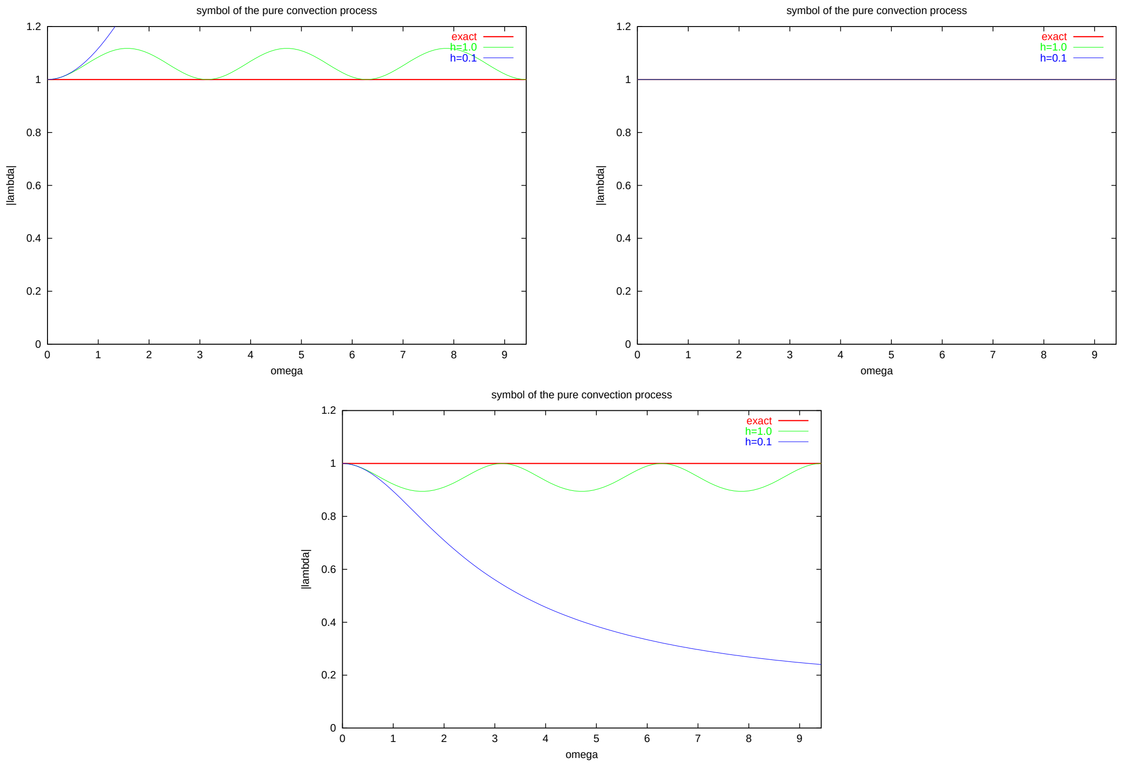

Figure 5.2: Symbol $|\lambda_E|$ of $u_t = -u_x$ with $\tau = \frac{1}{2}$ and $\theta = 0$, $\theta = \frac{1}{2}$ and $\theta = 1$, respectively; central differences

Example 5.2.12 ($u_t = -bu_x$ **with central differences)**
The equation $u_t = -bu_x$ characterises pure advection and is a degenerated parabolic differential equation (→ hyperbolic equation). With the discretised space operator

$$(L_h v)_i = -\frac{b}{2h}(v_{i+1} - v_{i-1})$$

we obtain for the symbol

$$\lambda_L(\omega) = -\frac{b}{2h}(\mathrm{e}^{-\mathrm{i}h\omega} - \mathrm{e}^{\mathrm{i}h\omega}) = \frac{b}{h}\mathrm{i}\sin(h\omega).$$

As the symbol λ_L is pure imaginary it follows from Corollary 5.2.9 that the θ-method is unconditionally stable for $\frac{1}{2} \leq \theta \leq 1$. For the Crank-Nicholson scheme the absolute value of the symbol is equal to one for any frequency. We remark that the scheme then is not parabolic in the sense of John and thus approximation to derivatives might be unreliable. In practice one observes visible oscillations. That gives rise to an alternative scheme where space derivatives are approximated in that direction where the flow comes from which will be shown in the next example.

For $\theta < \frac{1}{2}$ we see from (5.23) that

$$|\lambda_E|^2 = \frac{1 + ((1-\theta)\tau\Im\lambda_L)^2}{1 + (\theta\tau\Im\lambda_L)^2}.$$

Stability is obtained if there exists a constant $c > 0$ so that for all sufficiently small τ the inequality $|\lambda_E|^2 \leq 1 + c\tau$ holds. That is the case if the time step size obeys

$$\tau \leq \inf_{\omega\in\mathbb{R}} \frac{c}{(1-2\theta)\Im\lambda_L{}^2} = \frac{c}{b^2(1-2\theta)}h^2, \qquad 0 \leq \theta < \frac{1}{2}. \tag{5.26}$$

Therefore it is sufficient for stability to choose $\tau = ch^2$ with any constant $c > 0$. Figure 5.2 shows the symbol $|\lambda_E|$ graphically.

Example 5.2.13 ($u_t = -bu_x$ **with a simple upwind scheme)**
Let $b > 0$. We then approximate the first derivative with backward differences, i.e.

$$(L_h v)_i = -\frac{b}{h}(v_i - v_{i-1}).$$

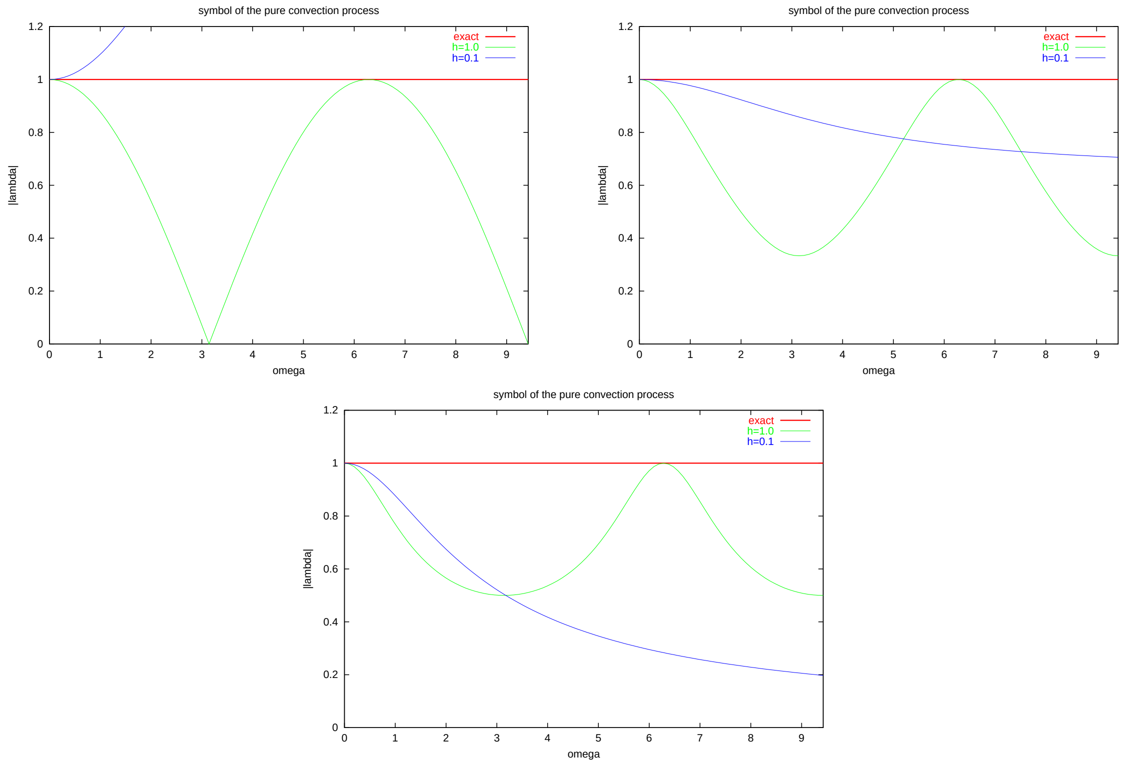

Figure 5.3: Symbol $|\lambda_E|$ of $u_t = -u_x$ with $\tau = \frac{1}{2}$ and $\theta = 0$, $\theta = \frac{1}{2}$ and $\theta = 1$, respectively; upwind

This choice is understandable if one imagines the solution of the equation (see also Subsection 3.2.1). Because $b > 0$ the matter flows from the left to the right so that one might prefer to approximate the derivatives from the side where the information comes from. The symbol of L_h is obviously

$$\lambda_L(\omega) = -\frac{b}{h}(1 - \mathrm{e}^{\mathrm{i}h\omega}) = -\frac{b}{h}\big(1 - \cos(h\omega) - \mathrm{i}\sin(h\omega)\big).$$

The real part of λ_L is negative so that from Corollary 5.2.9 we deduce the unconditionally stability of the θ-scheme for $\theta \geq \frac{1}{2}$. To be more precise we calculate the symbol of λ_E by referring to (5.23) and noting that $|\lambda_L|^2 = 2\left(\frac{b}{h}\right)^2(1-\cos(h\omega))$, $\Re\lambda_L = -\frac{b}{h}(1-\cos(h\omega))$:

$$|\lambda_E| = \frac{1 + 2(1-\cos(h\omega))(1-\theta)\tau\frac{b}{h}\left((1-\theta)\tau\frac{b}{h} - 1\right)}{1 + 2(1-\cos(h\omega))\theta\tau\frac{b^2}{h^2}\frac{b}{h}\left(\theta\tau\frac{b}{h} + 1\right)}$$

The sufficient criteria for stability $|\lambda_E| \leq 1$ is achieved if $(1-\theta)\left((1-\theta)\tau\frac{b}{h} - 1\right) \leq \theta\left(\theta\tau\frac{b}{h} + 1\right)$ which is equivalent to

$$\tau \leq \frac{h}{(1-2\theta)b}, \qquad 0 \leq \theta < \frac{1}{2}. \tag{5.27}$$

It indicates that by using an upwind scheme for a pure convection equation the stability strongly increases. First of all the scheme is parabolic in John's sense as illustrated in Figure 5.3. Secondly, the stability criterion for more explicit schemes ($\theta < \frac{1}{2}$) is less strict as it only requires that the time step size approaches zero with the same speed as the space step size goes to zero. For the central difference approximation the criterion for stability is more restrictive as seen in (5.26). The disadvantage though is that we lose one order of consistency in space direction.

There exists a very illustrative explanation for the stability criterion in the explicit case ($\theta = 0$) which is shown in figure 5.4. The value of a grid point $\hat{u}_i^{k+1}$ only depends on $\hat{u}_i^k$ and $\hat{u}_{i-1}^k$ because the derivatives are approximated using exactly these two points. By iteration we see that $\hat{u}_i^k$ only depends on $\hat{u}_{i-k}^{(0)}$ to $\hat{u}_i^{(0)}$. It follows that only the initial condition in the interval $[x_i - kh, x_i]$ influences the value $\hat{u}_i^k$. Since the analytical solution shows that the graph of the initial condition moves with speed b (from the left to the right if $b > 0$). It follows that a value $u(x, 0)$ influences $u(x + bt, t)$. Hence we only expect convergence if at least $bk\tau \leq kh$ and thus $\tau \leq \frac{h}{b}$ is fulfilled.

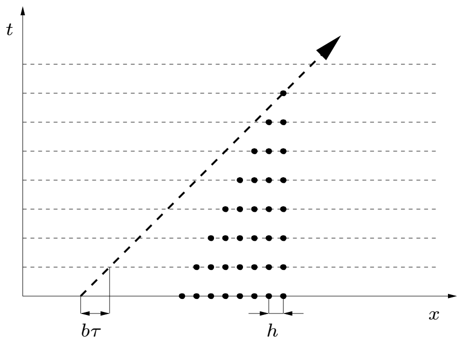

Figure 5.4: Obvious necessary stability criterion $b\tau \le h$

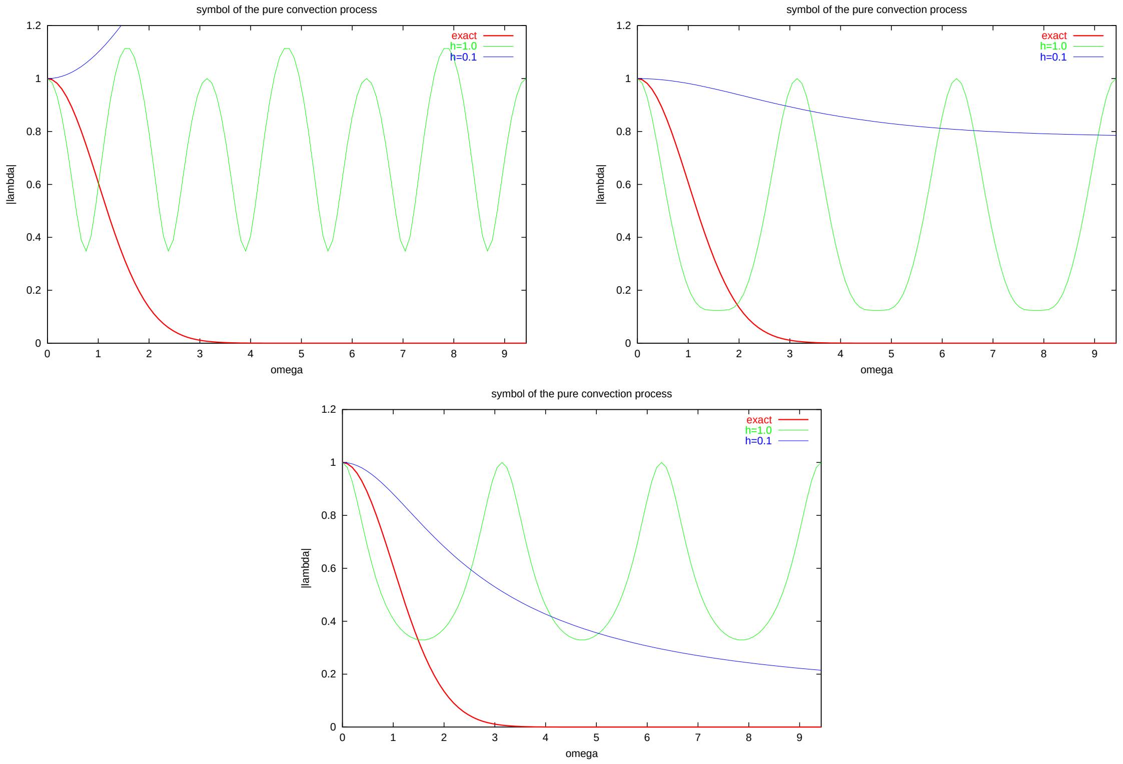

Figure 5.5: Symbol $|\lambda_E|$ of $u_t = u_{xx} - u_x$ with $\tau = \frac{1}{2}$ and $\theta = 0$, $\theta = \frac{1}{2}$ and $\theta = 1$, respectively

Example 5.2.14 (General diffusion convection with central differences)
Using the results already obtained in the previous examples the symbol of the general diffusion convection equation

$$u_t = \operatorname{div}(G\,\nabla\, u) - \langle b, \nabla\, u\rangle$$

is given by

$$\lambda_L(\omega) = -\sum_{i,j=1}^{d} \frac{4g_{ij}}{h_i h_j} \sin(h_i\omega_i)\sin(h_j\omega_j) + \mathrm{i}\left(\sum_{i=1}^{d} \frac{b_i}{h_i}\sin(h_i\omega_i)\right)$$

where h_i denotes the discretisation parameter in direction i. With the definition of the vector function $s(\omega) := \left(\frac{1}{h_1}\sin(h_1\omega_1), \ldots, \frac{1}{h_d}\sin(h_d\omega_d)\right)^\tau$ the symbol simplifies to

$$\lambda_L(\omega) = -4s(\omega)^\tau G s(\omega) + \mathrm{i}\,\langle b, s(\omega)\rangle\,. \tag{5.28}$$

As the real part of the symbol is not positive it follows by Corollary 5.2.9 that the θ-method is unconditionally stable for $\frac{1}{2} \le \theta \le 1$. Both Figures 5.5 and 5.6 show the symbol of $E_{h,\tau}$ where the latter is the result of a convection dominated equation.

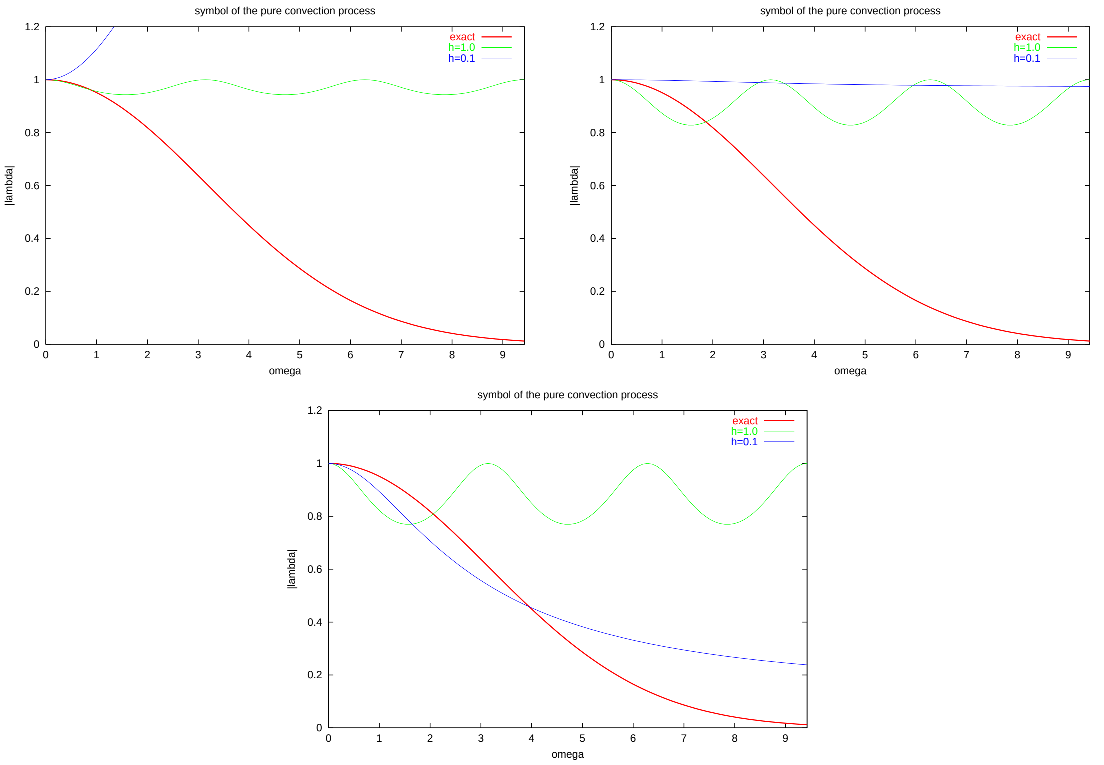

Figure 5.6: Symbol $|\lambda_E|$ of $u_t = \frac{1}{10}u_{xx} - u_x$ with $\tau = \frac{1}{2}$ and $\theta = 0$, $\theta = \frac{1}{2}$ and $\theta = 1$, respectively

5.3 Stability of f.d.m. for variable coefficient p.d.e.s

There exists a broad range of different approaches to determine stability for the variable coefficient case. In [32, Section 8] it is described how the theory of the constant coefficients can be generalised. Essentially, one fixes one $x \in \Omega$, determines the coefficients in that point and examines stability of the corresponding constant coefficient p.d.e. Stability in all points $x \in \Omega$ is related to the stability of the scheme, if the p.d.e. is uniformly parabolic. Samarskii gives in [24, chapter 6] general estimates and stability criteria in the energetic norm, e.g. $\|v\|^2_{-L_h} := -\langle L_h v, v\rangle$. However, we concentrate on estimates for the weaker L_2-norm in which we are finally able to infer stability of the Crank-Nicholson scheme on uniform grids applied to the Heston p.d.e. under appropriate boundary conditions even if the left variance boundary is at $x_2 = 0$.

5.3.1 General L_2 stability analysis in the $\mathbb{R}^d$

The main ideas presented in this subsection are taken from [32, Section 10]. Basically, we estimate $\hat{u}^{(k+1)}$ by scalar multiplying the finite difference scheme with a combination of the two vectors $\hat{u}^{(k)}$ and $\hat{u}^{(k+1)}$. Restricting the analysis to the θ-method with emphasis on the Crank-Nicholson method we will see that stability depends on the properties of the space operator L_h. Hence in the second part of this subsection the discrete operator L_h will be examined and a result proved which certifies the Crank-Nicholson method with a particular space discretisation unconditionally stability if boundary conditions are chosen appropriately.

In order to state the first important result we recall the finite dimensional Hilbert space of grid functions $(\Phi_h, \langle\cdot,\cdot\rangle)$ over a uniform grid $\bar{\Omega}_h$ where the scalar product is defined by

$$\langle u, v\rangle := h^d \sum_{x\in\bar{\Omega}_h} u(x)v(x) = h^d \sum_{i_1,\ldots,i_d=0}^{m_1,\ldots,m_d} u_i v_i.$$

For convenience reasons let $\bar{\Omega}_h := \left\{x^{(k)} := hk | k \in \{0,\ldots,m_1\} \times \ldots \times \{0,\ldots,m_d\}\right\}$ for any $h > 0$. The values $m_1,\ldots,m_d$ are obviously indirect proportional to h.

Lemma 5.3.1
For the θ-method (4.8) at time step $k \in \mathbb{N}$

$$\frac{\hat{u}^{(k+1)} - \hat{u}^{(k)}}{\Delta t_k} = L_h \hat{u}^{(k+\theta)} + f_h^{(k)}$$

with the abbreviation $\hat{u}^{(k+\theta)} := \theta \hat{u}^{(k+1)} + (1-\theta)\hat{u}^{(k)}$, the following L_2*-norm equality holds:*

$$\left\| \hat{u}^{(k+1)} \right\|^2 + (2\theta - 1) \left\| \hat{u}^{(k+1)} - \hat{u}^{(k)} \right\|^2 = \left\| \hat{u}^{(k)} \right\|^2 + 2\Delta t_k \left(\left\langle L_h \hat{u}^{(k+\theta)} + f_h^{(k)}, \hat{u}^{(k+\theta)} \right\rangle \right). \tag{5.29}$$

Proof Multiplying both sides of the θ-method with the grid function $\hat{u}^{(k+\theta)}$ results in

$$\left\langle \hat{u}^{(k+1)} - \hat{u}^{(k)}, \hat{u}^{(k+\theta)} \right\rangle = \Delta t_k \left(\left\langle L \hat{u}^{(k+\theta)}, \hat{u}^{(k+\theta)} \right\rangle + \left\langle f_h^{(k)}, u^{(k+\theta)} \right\rangle \right).$$

We would like to convert the term on the left hand side which is essentially $(a-b)(\theta a + (1-\theta)b)$ to a form of $c_1(a^2 - b^2) + c_2(a-b)^2$. Determining the constants we see that

$$(a-b)(\theta a + (1-\theta)b) = \frac{1}{2}(a^2 - b^2) + \frac{1}{2}(2\theta - 1)(a-b)^2$$

and hence

$$2 \left\langle \hat{u}^{(k+1)} - \hat{u}^{(k)}, \hat{u}^{(k+\theta)} \right\rangle = \left\| \hat{u}^{(k+1)} \right\|^2 - \left\| \hat{u}^{(k)} \right\|^2 + (2\theta - 1) \left\| \hat{u}^{(k+1)} - \hat{u}^{(k)} \right\|^2$$

from which the desired result immediately follows. □

Corollary 5.3.2
For the θ-method (4.8), $\frac{1}{2} \le \theta \le 1$, at time step $k \in \mathbb{N}$ and the same abbreviation as used above the following L_2*-norm inequality holds:*

$$\boldsymbol{\left\| \hat{u}^{(k+1)} \right\|^2 \le \left\| \hat{u}^{(k)} \right\|^2 + 2\Delta t_k \left(\left\langle L_h \hat{u}^{(k+\theta)}, \hat{u}^{(k+\theta)} \right\rangle + \left\| \hat{u}^{(k+\theta)} \right\|^2 + \left\| f_h^{(k)} \right\|^2 \right)}. \tag{5.30}$$

Proof It is a direct result of (5.29) with the estimates $2\theta - 1 \ge 0$ and $\left\langle f_h^{(k)}, \hat{u}^{(k+\theta)} \right\rangle \le \left\| \hat{u}^{(k+\theta)} \right\|^2 + \left\| f_h^{(k)} \right\|^2$. □

It is now quite simple to find criteria for stability in the sense that $\left\| \hat{u}^{(m_0)} \right\| \le c \left\| \hat{u}^{(0)} \right\|$ for all $h, \tau \to 0$. The next corollary gives the stability criterion if the space operator L_h is negative semidefinite which is a generalisation of Corollary 5.2.8 and 5.2.9.

Corollary 5.3.3
Let the discrete space operator L_h be negative semidefinite, i.e. $\langle L_h v, v \rangle \le 0$, $\forall v \in \Phi_h$ satisfying current boundary conditions. The θ-method for the homogeneous equation is then unconditionally stable in L_2 *for $\theta \ge \frac{1}{2}$. If L_h is additionally symmetric the stability criterion for $0 \le \theta < \frac{1}{2}$ can be written as*

$$\Delta t_k \le \frac{2}{1 - 2\theta} \frac{1}{\|L_h\|}. \tag{5.31}$$

Proof If $\theta \ge \frac{1}{2}$ then $2\theta - 1 \ge 0$ and with the above lemma and the negative semidefinite operator L_h it follows

$$\left\| \hat{u}^{(k+1)} \right\|^2 \le \left\| \hat{u}^{(k)} \right\|^2 + 2\Delta t_k \left\langle L_h \hat{u}^{(k+\theta)}, \hat{u}^{(k+\theta)} \right\rangle \le \left\| \hat{u}^{(k)} \right\|^2 \le \ldots \le \left\| \hat{u}^{(0)} \right\|^2$$

which is stability by definition.

Otherwise $2\theta - 1 < 0$ and the neglected term $\left\| \hat{u}^{(k+1)} - \hat{u}^{(k+1)} \right\|^2$ has to be taken into account. Using the relation $\hat{u}^{(k+1)} - \hat{u}^{(k)} = \Delta t_k L_h u^{k+\theta}$ and the above lemma the following equality holds:

$$\left\| \hat{u}^{(k+1)} \right\|^2 = \left\| \hat{u}^{(k)} \right\|^2 + (1 - 2\theta)\Delta t_k^2 \left\| L_h \hat{u}^{(k+\theta)} \right\|^2 + 2\Delta t_k \left\langle L_h \hat{u}^{(k+\theta)}, \hat{u}^{(k+\theta)} \right\rangle.$$

For stability it is sufficient if

$$(1-2\theta)\Delta t_k^2 \left\| L_h \hat{u}^{(k+\theta)} \right\|^2 + 2\Delta t_k \left\langle L_h \hat{u}^{(k+\theta)}, \hat{u}^{(k+\theta)} \right\rangle \leq 0$$

which is satisfied if

$$\Delta t_k \leq \frac{2}{1-2\theta} \inf_{v\in\Phi_h} \frac{|\langle L_h v, v\rangle|}{\|L_h v\|^2}.$$

For a symmetric matrix there exists an eigenvector basis. Writing the vector as a linear combination of the eigenvectors and finding the infimum it turns out that it is reached for the eigenvector v^* to the greatest absolute eigenvalue

$$\inf_{v\in\mathbb{R}^n} \frac{|\langle L_h v, v\rangle|}{\|L_h v\|^2} = \frac{\lambda_{\max}\langle v^*, v^*\rangle}{\lambda_{\max}^2 \|v^*\|^2} = \frac{1}{\lambda_{\max}} = \frac{1}{\|L_h\|}$$

which completes the proof. □

The requirement $\langle L_h v, v\rangle \leq 0$ is strict and quite often not satisfied. An stronger stability result gives the following corollary where only $\langle L_h v, v\rangle \leq C\|v\|^2$ is demanded.

Corollary 5.3.4
Let the space grid $\bar{\Omega}_h$ be uniform and the time steps be proportional to the parameter τ, $\Delta t_k \sim \tau$, i.e. there exist constants $c_1 > 0, c_2 > 0$ so that $\tau \leq c_1 \min_k \Delta t_k$ and $\max_k \Delta t_k \leq c_2\tau$. The θ-method with $\frac{1}{2} \leq \theta \leq 1$ for the non-homogeneous equation is unconditionally stable in L_2 if there exists a constant independent of τ and h so that

$$\langle L_h v, v\rangle \leq C\|v\|^2$$

for all $v \in \Phi_h$ satisfying the actual boundary conditions.

The norm of the approximated solution at the final time T can then be estimated with a constant $\tilde{c} > 0$ by

$$\left\|\hat{u}^{(m_0)}\right\|^2 \leq \mathrm{e}^{\tilde{c}T}\left(\left\|\hat{u}^{(0)}\right\|^2 + \sup_{k\in\{0,\dots,m_0\}} \left\|f_h^{(k)}\right\|^2\right), \qquad h,\tau \to 0. \tag{5.32}$$

Proof By Corollary 5.3.2 the solution of the numerical scheme for $\theta \geq \frac{1}{2}$ at time step $k+1 \in \{1,\dots,m_0\}$ denoted by $\hat{u}^{(k+1)} \in \Phi_h$ can be estimated by

$$\begin{aligned}\left\|\hat{u}^{(k+1)}\right\|^2 &\leq \left\|\hat{u}^{(k)}\right\|^2 + 2\Delta t_k\left(\left\langle L_h\hat{u}^{(k+\theta)}, \hat{u}^{(k+\theta)}\right\rangle + \left\|\hat{u}^{(k+\theta)}\right\|^2 + \left\|f_h^{(k)}\right\|^2\right)\\ &\leq \left\|\hat{u}^{(k)}\right\|^2 + 2(2C+1)\Delta t_k\left(\theta^2\left\|\hat{u}^{(k+1)}\right\|^2 + (1-\theta)^2\left\|\hat{u}^{(k)}\right\|^2\right) + 2\Delta t_k\left\|f_h^{(k)}\right\|^2\end{aligned}$$

from which

$$\begin{aligned}\left\|\hat{u}^{(k+1)}\right\|^2 &\leq \frac{1+(1-\theta)^2\tilde{C}\Delta t_k}{1-\theta^2\tilde{C}\Delta t_k}\left\|\hat{u}^{(k)}\right\|^2 + 2\Delta t_k\left\|f_h^{(k)}\right\|^2\\ &\leq (1+c\Delta t_k)\left\|\hat{u}^{(k)}\right\|^2 + 2\Delta t_k\left\|f_h^{(k)}\right\|^2, \qquad (\Delta t_k \to 0)\end{aligned}$$

follows. By iterative application of the estimate and keeping in mind that $\max_k \Delta t_k \leq c_2\tau$ we conclude

$$\begin{aligned}\left\|\hat{u}^{(m_0)}\right\|^2 &\leq (1+c\tau)^{m_0}\left\|\hat{u}^{(0)}\right\|^2 + 2c_2\tau\sum_{k=0}^{m_0}(1+c\tau)^k\left\|f_h^{(k)}\right\|^2\\ &\leq \mathrm{e}^{c^*T}\left(\left\|\hat{u}^{(0)}\right\|^2 + 2c_2\tau\sum_{k=0}^{m_0}\left\|f_h^{(k)}\right\|^2\right)\\ &\leq \mathrm{e}^{\tilde{c}T}\left(\left\|\hat{u}^{(0)}\right\|^2 + \sup_{k\in\{0,\dots,m_0\}}\left\|f_h^{(k)}\right\|^2\right), \qquad \tau \to 0.\end{aligned}$$

In the last step the relation $m_0\tau \leq c_1 T$ has been used which follows from $m_0 \min_k \Delta t_k \leq T$ and $\tau \leq c_1 \min_k \Delta t_k$. □

As seen in (5.10) the error of the numerical solution of a homogeneous p.d.e. satisfies the same difference equation but with the truncation error as right hand side f and zero as initial condition. Stability of the θ-method for non-homogeneous p.d.e.s therefore implies the convergence of a homogeneous equation.

Theorem 5.3.5 (Convergence)
Let a finite difference scheme be accurate of order $\mathrm{O}(h^\mu+\tau^\nu)$ *and stable in the sense that constants* $c_1>0$ *and* $c_2>0$ *exists so that independently of* h, τ *and* f *the estimate*

$$\left\|\hat{u}^{(m_0)}\right\|^2 \leq \mathrm{e}^{\tilde{c}T}\left(\left\|\hat{u}^{(0)}\right\|^2+\sup_{k\in\{0,\ldots,m_0\}}\left\|f_h^{(k)}\right\|^2\right), \qquad h,\tau\to 0$$

for any right hand side f *holds. The method is then convergent of the order* $\mathrm{O}(h^\mu+\tau^\nu)$, *i.e. the difference between the analytic solution* $u(\cdot,T)$ *and the numerical solution* $\hat{u}^{(m_0)}$ *is*

$$\left\|P_h u(\cdot,T)-\hat{u}^{(m_0)}\right\| = \mathrm{O}(h^\mu+\tau^\nu).$$

Proof We set $z^{(k)}:=P_h u(\cdot,t_k)-\hat{u}^{(k)}$. Since $z^{(0)}=\mathbb{O}$ it follows from (5.10) and the stability of the numerical scheme that for a $\tilde{c}>0$

$$\left\|z^{(m_0)}\right\| \leq \mathrm{e}^{\tilde{c}T}\sup_{k\in\{0,\ldots,m_0\}}\left\|\gamma_{h,\tau}^{(k)}\right\|, \qquad h,\tau\to 0$$

with the truncation error $\gamma_{h,\tau}^{(k)}$. By definition of accuracy 5.1.2 the truncation error is

$$\left\|\gamma_{h,\tau}^{(k)}\right\|_\infty = \mathrm{O}(h^\mu+\tau^\nu), \quad (h,\tau\to 0)$$

uniformly for all $k\in\{0,\ldots,m_0\}$. It follows for the error

$$\left\|z^{(m_0)}\right\| \leq \mathrm{e}^{\tilde{c}T}\left|\bar{\Omega}\right|\mathrm{O}(h^\mu+\tau^\nu), \qquad h,\tau\to 0$$

where $\left|\bar{\Omega}\right|$ denotes area of $\bar{\Omega}\subset\mathbb{R}^d$. □

With Theorem 5.3.5 and Corollary 5.3.4 one only needs to check the boundedness of the space operator L_h, i.e. $\langle L_h v, v\rangle \leq C\|v\|^2$, in order to determine whether the θ-method is stable and thus convergent if consistent. Hence from now on we concentrate on estimating $\langle L_h v, v\rangle$. It might already be said that this is not a straight forward process and requires some effort especially if different boundary conditions are taken into account. As it turns out for Dirichlet-type boundary conditions with zero values on Γ the analysis mainly based on the summation by parts rule – the discrete analogue of integration by parts – simplifies a lot. Hence only these conditions are discussed below. It is remarkable that the uniform parabolicity of the p.d.e. will not be required. In order to apply the summation by parts rule and thus simplifying the proof of the theorems a space operator L_h slightly differing from the space discretisation described in Subsection 4.3.1 will be introduced.

We now consider the general d-dimensional convection-diffusion equation in a rectangular domain $\Omega\subset\mathbb{R}^d$,

$$u_t = \operatorname{div}(G\,\nabla u)-\langle b,\nabla u\rangle - cu + f, \qquad , u|_{\partial\Omega}=0 \tag{5.33}$$

with in x continuous functions $G:\bar{\Omega}\to\mathbb{R}^{d\times d}$, $b:\bar{\Omega}\to\mathbb{R}^d$ and $c:\bar{\Omega}\to\mathbb{R}$. We require the matrix $G(x)$ to be positive semidefinite for all $x\in\bar{\Omega}$. No further restrictions are imposed on the p.d.e. which in particular means that $G(x)=\mathbb{O}$ is allowed for any $x\in\bar{\Omega}$. Written more explicitly the p.d.e. under consideration is

$$\frac{\partial u}{\partial t}(x,t) = \sum_{i,j=1}^{d}\frac{\partial}{\partial x_i}\left(g_{i,j}(x)\frac{\partial u}{\partial x_j}\right)(x,t)+\sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i}(x,t)+c(x)u(x,t).$$

To specify the discrete space operator L_h we use the backward, forward and centred difference approximation introduced in Subsection 4.2.1. Let $v\in\Phi_h$ be a grid function then these operators are defined for $i\in\{1,\ldots,d\}$ and $k\in\{0,\ldots,m_1\}\times\ldots\times\{0,\ldots,m_d\}$ by

$$\begin{aligned}
(\partial_i v)_k &:= \frac{v_{k+e_i}-v_k}{h}, && k_i\neq m_i,\\
(\bar{\partial}_i v)_k &:= \frac{v_k-v_{k-e_i}}{h}, && k_i\neq 0,\\
(\hat{\partial}_i v)_k &:= \frac{1}{2}(\partial_i v+\bar{\partial}_i v)_k, && k_i\notin\{0,m_i\}.
\end{aligned}$$

The vectors $e_i \in \mathbb{Z}^d$ are unit vectors in direction $i \in \{1,\ldots,d\}$, i.e. $e_1 = (1,0,\ldots,0)$ to $e_d = (0,\ldots,0,1)$. If the operators are undefined in the above sense, e.g. if $k + e_i$ is not a grid point we assign the value zero. This artificial declaration causes no problems when it comes to the examination of $\langle L_h v, v\rangle$ as we always assume that $v|_{\Gamma_h} = 0$. All difference operators are then mappings on the space of grid functions $\partial_i : \Phi_h \to \Phi_h$ which simplifies the notation below.

The following lemma lies the foundation to estimate $\langle L_h v, v\rangle$.

Lemma 5.3.6 (Discrete version of integration and differentiation rules)
Let $u \in \Phi_h$ and $v \in \Phi_h$ be grid functions with $u|_{\Gamma_h} = v|_{\Gamma_h} = 0$. The following equation then holds for any $i \in \{1,\ldots,d\}$

$$\langle \partial_i u, v\rangle = -\left\langle u, \bar{\partial}_i v\right\rangle \tag{5.34}$$

which is called the summation by parts rule.

For the difference operator $\hat{\partial}$ the multiplication rule with $k \in \{0,\ldots,m_1\} \times \ldots \times \{0,\ldots,m_d\}$

$$\hat{\partial}_i(u\cdot v)_k = u_k(\hat{\partial} v)_k + \frac{1}{2}\left(v_{k+e_i}(\partial_i u)_k + v_{k-e_i}(\bar{\partial}_i u)_k\right), \qquad k_i \notin \{0, m_i\} \tag{5.35}$$

applies for all $u, v \in \Phi_h$ where $(u \cdot v)_k := u_k v_k$.

Proof For simplicity of notation the summation by parts rule is only shown for the derivative with respect to the first component. Applying the definition, rearranging the terms of the sum and taking into account that u, v are zero on the boundary Γ_h yields

$$\begin{aligned}
\langle \partial_1 u, v\rangle &= \frac{1}{h}\sum_{k_1,\ldots,k_d=1}^{m_1-1,\ldots,m_d-1} u_{k+e_1}v_k - u_k v_k \\
&= \frac{1}{h}\sum_{k_2,\ldots,k_d=1}^{m_2-1,\ldots,m_d-1}\left(\sum_{k_1=2}^{m_1} u_k v_{k-e_1} - \sum_{k_1=1}^{m_1-1} u_k v_k\right) \\
&= \frac{1}{h}\sum_{k_2,\ldots,k_d=1}^{m_2-1,\ldots,m_d-1}\sum_{k_1=1}^{m_1-1} u_k(v_{k-e_1} - v_k) \\
&= -\left\langle u, \bar{\partial}_1 v\right\rangle.
\end{aligned}$$

By applying the definition we see that both terms of the multiplication rule are equal:

$$\hat{\partial}_i(u\cdot v)_k = \frac{1}{2h}(u_{k+e_i}v_{k+e_i} - u_{k-e_i}v_{k-e_i}),$$

$$\begin{aligned}
u_k(\hat{\partial} v)_k + \frac{1}{2}\left(v_{k+e_i}(\partial u)_k + v_{k-e_i}(\bar{\partial} u)_k\right) = &\ \frac{u_k v_{k+e_i} - u_k v_{k-e_i}}{2h} \\
&+ \frac{v_{k+e_i}u_{k+e_i} - v_{k+e_i}u_k + v_{k-e_i}u_k - v_{k-e_i}u_{k-e_i}}{2h}.
\end{aligned}$$

□

Remark 5.3.7
It is easy to prove that the above summation by parts rule remains true even if $u_k \neq 0$ for $k_i = m_i$ as it is exemplary the case for $u := \bar{\partial} v$. Lemma 5.3.12 shows for the $\mathbb{R}^1$ how the rule generalises if no restrictions on the boundary values of u and v are imposed.

We now introduce the discretisation L_h for which we will show stability of the θ-method. The three different parts of the p.d.e., that is to say the diffusion, the convection and the undifferentiated term are treated separately. We thus set $L_h := M_h + N_h + O_h$ and define them as follows

$$\begin{aligned}
(M_h v)_k &:= \sum_{i,j=1}^{d} \partial_i\left(g_{i,j}(x^{(k)})\bar{\partial}_j v\right)_k, \\
(N_h v)_k &:= \sum_{i=1}^{d} b_i(x^{(k)})(\hat{\partial}_i v)_k, \\
(O_h v)_k &:= c(x^{(k)})v_k.
\end{aligned} \tag{5.36}$$

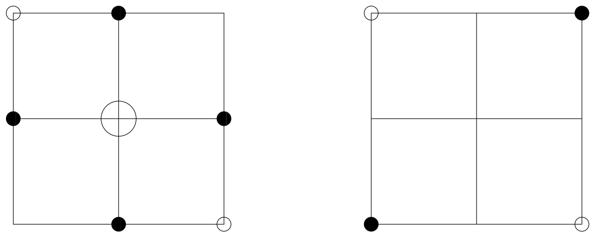

Figure 5.7: Stencil of mixed derivatives for L_h and $\tilde{L}_h$

Remark 5.3.8
Comparing the spatial operator L_h with the general discretisation defined in Subsection 4.3.1 and derived in Section 4.2 the convection and undifferentiated term N_h and O_h, respectively, coincide in both definitions. The diffusion term M_h differs, mainly due to the approximation of the mixed derivatives. If in (5.36) all entries of the matrix $G(x)$ are constant 1 then we see that mixed derivatives are as follows compared to the four point approximation shown in Table 4.4:

$$\begin{aligned}
\frac{\partial^2 v}{\partial x \partial y}(x_i, y_j) &\approx \frac{1}{2}\left(\partial_1 \bar{\partial}_2 \bar{v} + \bar{\partial}_1 \partial_2 \bar{v}\right)_{i,j} \\
&= \frac{h^2}{2}\left(\bar{v}_{i-1,j} - \bar{v}_{i-1,j+1} + \bar{v}_{i,j-1} - 2\bar{v}_{i,j} + \bar{v}_{i,j+1} + \bar{v}_{i+1,j} - \bar{v}_{i+1,j-1}\right), \\
\frac{\partial^2 v}{\partial x \partial y}(x_i, y_j) &\approx \left(\hat{\partial}_1 \hat{\partial}_2 \bar{v}\right)_{i,j} \\
&= \frac{h^2}{4}\left(\bar{v}_{i-1,j-1} - \bar{v}_{i-1,j+1} - \bar{v}_{i+1,j-1} + \bar{v}_{i+1,j+1}\right).
\end{aligned}$$

The stencil of these two difference approximations is illustrated in Figure 5.7. Filled circles represent the value one and empty circles represent minus one or minus two depending on the diameter.

Only if the p.d.e. does not contain any mixed derivatives we are able to conclude stability of $\tilde{L}_h$ if L_h is stable. That will below exemplary be shown in only one space dimension. The reason why it has been defined that way is to simplify the proof of stability. One disadvantage of that definition is its first and not second order accuracy if G is not constant. Second order accuracy can be achieved with the alternative definition

$$(M_h v)_k := \sum_{i,j=1}^{d} \partial_i \left(g_{i,j}(x^{(k)} - 1/2he_i)\bar{\partial}_j v\right)_k . \tag{5.37}$$

As by Lemma 5.3.1 it is decisive for stability of the θ-method which values the scalar product $\langle L_h v, v\rangle = \langle M_h v, v\rangle + \langle N_h v, v\rangle + \langle O_h v, v\rangle$ can attain:

Lemma 5.3.9
Assume $G(x)$ is symmetric, positive semidefinite for all $x \in \bar{\Omega}$ and let $b \in \mathrm{C}^1(\bar{\Omega}, \mathbb{R}^d)$. For the operators M_h, N_h and O_h as defined in (5.36) the following estimates hold for any $v \in \Phi_h$ with $v|_{\Gamma_h} = 0$:

$$\begin{aligned}
\langle M_h v, v\rangle &\leq 0, \\
\langle N_h v, v\rangle &\leq C_2(b)\left\|v\right\|^2, \\
\langle O_h v, v\rangle &\leq C_3(c)\left\|v\right\|^2.
\end{aligned} \tag{5.38}$$

If the diffusion part is of the form $a(x)G$ with a constant positive semidefinite matrix $G \in \mathbb{R}^{d\times d}$ and a scalar $a(x) \geq 0\ \forall x \in \bar{\Omega}$ then the same estimate applies to the second order accurate operator M_h as defined in (5.37).

Proof Let $\bar{g}_{i,j} := P_h g_{i,j}$ and $\bar{b}_i := P_h b_i$ be the restrictions of the corresponding continuous function to the grid $\bar{\Omega}_h$. With the definition of M_h and the summation by parts rule we establish the estimate

$$
\begin{aligned}
\langle M_h v, v\rangle &= \sum_{i,j=1}^{d} \left\langle \partial_i(\bar{g}_{i,j}\cdot\bar{\partial}_j v), v\right\rangle = -\sum_{i,j=1}^{d}\left\langle \bar{g}_{i,j}\cdot\bar{\partial}_j v, \bar{\partial}_i v\right\rangle \\
&= -\sum_{k_1,\dots,k_d=1}^{m_1,\dots,m_d}\sum_{i,j=1}^{d} g_{i,j}(x^{(k)})(\bar{\partial}_i v)_k(\bar{\partial}_j v)_k \le 0.
\end{aligned}
$$

The last expression is smaller or equal zero because $G(x)$ is positive semidefinite for all $x\in\bar{\Omega}$, i.e. $\sum_{i,j=1}^{d} g_{i,j}(x)y_i y_j \ge 0$, $\forall y\in\mathbb{R}^d$. If the diffusion coefficient is of the form $a(x)G$ with a constant matrix G and if M_h is defined as in (5.37) the equation modifies

$$
\begin{aligned}
\langle M_h v, v\rangle &= -\sum_{k_1,\dots,k_d=1}^{m_1,\dots,m_d}\sum_{i,j=1}^{d} a(x^{(k)}-h/2e_i)g_{i,j}(\bar{\partial}_i v)_k(\bar{\partial}_j v)_k \\
&= -\sum_{k_1,\dots,k_d=1}^{m_1,\dots,m_d}\sum_{i,j=1}^{d} \tilde{g}_{i,j}(x^{(k)})(\bar{\partial}_i v)_k(\bar{\partial}_j v)_k
\end{aligned}
$$

with the matrix $\tilde{G}(x)$ defined by

$$
\tilde{G}(x) := \begin{pmatrix} a(x-h/2e_1) & 0 & \cdots & 0 \\ 0 & a(x-h/2e_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a(x-h/2e_d)\end{pmatrix} G
$$

which is still positive semidefinite since G is positive semidefinite and $a(x)\ge 0$. It can be shown by the determinant criteria that a positive semidefinite matrix remains so if it is multiplied line by line with a non negative value. That is why we conclude as above that $\langle M_h v, v\rangle \le 0$.

Assisted by the multiplication rule and the abbreviation $v_{\pm e_i}(x^{(k)}) := v(x^{(k)}\pm he_i)$ or $v_{\pm e_i}(x^{(k)}) := 0$ if $x^{(k)}\pm he_i \notin \bar{\Omega}_h$, we see that

$$
\begin{aligned}
\langle N_h v, v\rangle &= \sum_{i=1}^{d}\left\langle \bar{b}_i\cdot\hat{\partial}_i v, v\right\rangle = \sum_{i=1}^{d}\left\langle \hat{\partial}_i v, \bar{b}_i\cdot v\right\rangle = -\sum_{i=1}^{d}\left\langle v, \hat{\partial}_i(\bar{b}_i\cdot v)\right\rangle \\
&= -\sum_{i=1}^{d}\left(\left\langle v, \bar{b}_i\cdot\hat{\partial}_i v)\right\rangle + \frac{1}{2}\left(\langle v, v_{+e_i}\partial_i\bar{b}_i\rangle + \langle v, v_{-e_i}\bar{\partial}_i\bar{b}_i\rangle\right)\right) \\
&= -\langle N_h v, v\rangle - \sum_{i=1}^{d}\left(\frac{1}{2}\left(\langle v, v_{+e_i}\partial_i\bar{b}_i\rangle + \langle v, v_{-e_i}\bar{\partial}_i\bar{b}_i\rangle\right)\right).
\end{aligned}
$$

It follows

$$
\begin{aligned}
\langle N_h v, v\rangle &= -\frac{1}{4}\sum_{i=1}^{d}\left(\langle v, v_{+e_i}\partial_i\bar{b}_i\rangle + \langle v, v_{-e_i}\bar{\partial}_i\bar{b}_i\rangle\right) \\
&\le \frac{1}{4}\sum_{i=1}^{d}\left\|\frac{\partial b_i}{\partial x_i}\right\|_{\mathrm{C}(\bar{\Omega})}\left(\langle v, v_{-e_i}\rangle + \langle v, v_{+e_i}\rangle\right) \\
&\le \frac{1}{2}\sum_{i=1}^{d}\left\|\frac{\partial b_i}{\partial x_i}\right\|_{\mathrm{C}(\bar{\Omega})}\|v\|^2.
\end{aligned}
$$

By the Cauchy-Schwarz inequality we have $\langle v, v_{+e_i}\rangle \le \|v\|\,\|v_{+e_i}\| \le \|v\|^2$ which justifies the last step in the estimate.

Finally, we remark that obviously $\langle O_h v, v\rangle \le \|c\|_{\mathrm{C}(\bar{\Omega})}\|v\|^2$. □

The estimates of this lemma together with Corollary 5.3.4 paves the way for a general stability result.

Theorem 5.3.10 (Stability of the θ-method)
Let the space grid $\bar{\Omega}_h$ be uniform and $\Delta t_k \sim \tau$. The θ-method with $\frac{1}{2}\le\theta\le 1$ and in particular the Crank-Nicholson scheme is then unconditionally stable in L_2 for the non-homogeneous p.d.e.

(5.33) with zero boundary conditions and the space discretisation $L_h := M_h + N_h + O_h$ as defined in (5.36) if the matrix $G(x)$ is positive semidefinite for all $x \in \bar{\Omega}$ and $b \in \mathrm{C}^1(\bar{\Omega}, \mathbb{R}^d)$. Is the diffusion matrix of the form $a(x)G$ with a constant matrix then we also obtain stability if M_h is defined as in (5.37).

Proof The prerequisite of Corollary 5.3.4 which is $\langle L_h v, v\rangle \leq C\,\|v\|^2$ is ensured by (5.38). □

Remark 5.3.11
Any homogeneous p.d.e. with Dirichlet conditions can be transformed to an inhomogeneous p.d.e. with a zero boundary: Let u be the solution of $u_t = Lu$ with $u|_\Gamma = g$ and a smooth function $g : \Gamma \times [0,T] \to \mathbb{R}$ defined on the boundary. If one can find a smooth function $w : \Omega \times [0,T] \to \mathbb{R}$ fulfilling $w|_\Gamma = g$ we define $v := u - w$. Since $Lv = Lu - Lw = u_t - Lw = v_t + w_t - Lw$ and $v|_\Gamma = g - g = 0$ the function v obeys an inhomogeneous p.d.e. with a zero boundary: $v_t = Lv + f$ with $f := w_t - Lw$ known a priori.

5.3.2 Further analysis in the $\mathbb{R}^1$

The stability result of Theorem 5.3.10 is general but on the other hand quite restrictive in the sense that it requires zero boundary conditions. Estimating the scalar product $\langle L_h v, v\rangle$ is more difficult in $\mathbb{R}^d$ if v is not zero on Γ_h because the summation by parts rule complicates as many boundary terms come into the equation. Also, the inevitable usage of multi indices in the $\mathbb{R}^d$ can make equations look unnecessarily complex. Hence the remaining questions, which are stability of the discrete space operator as defined in 4.3.1 and under other boundary conditions, are examined in the $\mathbb{R}^1$. There the homogeneous p.d.e. can be written in the two equivalent ways

$$\begin{aligned} u_t &= \frac{\partial}{\partial x}(g(x)u_x) + b(x)u_x + c(x)u, \\ u_t &= \big(g(x)u_{xx} + g_x(x)u_x\big) + b(x)u_x + c(x)u. \end{aligned} \tag{5.39}$$

Let the grid for simplicity be $\bar{\Omega}_h = \{x_k = hk : k = 1 \dots m\}$.

First we restate the summation by parts rule valid for any boundary condition. To write the following equations in an elegant way it is useful to define the scalar product with subscript:

$$\langle u, v\rangle_{(k,l)} := \sum_{i=k}^{l} u_i v_i.$$

Lemma 5.3.12
Let $u \in \Phi_h$ and $v \in \Phi_h$ be grid functions then the following equations are valid:

$$\begin{aligned} \langle \partial u, v\rangle_{(1,m-1)} &= -\left\langle u, \bar{\partial} v\right\rangle_{(1,m)} + u_m v_m - u_1 v_0, \\ \left\langle \hat{\partial} u, v\right\rangle_{(1,m-1)} &= -\left\langle u, \hat{\partial} v\right\rangle_{(1,m-1)} + \frac{1}{2}\left(u_m v_{m-1} + u_{m-1} v_m - u_0 v_1 - u_1 v_0\right). \end{aligned}$$

Proof All equations follow directly from definition. □

We now look into the differences between the discretisation defined in (5.36) with M_h as in (5.37) and Subsection 4.3.1. Being in line with the approach above and discretising the three parts of the convection-diffusion equation separately the method described in 4.3.1 for which we will show stability is given on a uniform grid by $\tilde{L}_h = \tilde{M}_h + N_h + O_h$ with

$$\begin{aligned} (\tilde{M}_h v)_k &:= \frac{1}{h^2} g(x_k)(v_{k+1} - 2v_k + v_{k-1}) + \frac{1}{2h} g_x(x_k)(v_{k+1} - v_{k-1}), \\ (N_h v)_k &:= \frac{1}{2h} b(x_k)(v_{k+1} - v_{k-1}), \\ (O_h v)_k &:= c(x_k) v_k. \end{aligned} \tag{5.40}$$

In one space dimension (5.36) and M_h defined as in (5.37) simplifies to $L_h = M_h + N_h + O_h$ with N_h, O_h as above and

$$\begin{aligned}(M_h v)_k &:= \frac{1}{h^2}\left(g(x_k + 1/2h)(v_{k+1} - v_k) - g(x_k - 1/2h)(v_k - v_{k-1})\right)\\ &= \frac{1}{h^2} g(x_k)(v_{k+1} - 2v_k + v_{k-1}) + \frac{1}{2h} g_x(x_k)(v_{k+1} - v_{k-1})\\ &\quad + \frac{1}{4}\left(g_{xx}(\xi_1)(v_{k+1} - v_k) - g_{xx}(\xi_2)(v_k - v_{k-1})\right).\end{aligned} \tag{5.41}$$

From Lemma 5.3.9 follows that $\langle M_h v, v\rangle_{(1,m-1)} \leq 0$ for all $v \in \Phi_h$ with $v|_{\Gamma_h} = 0$. Using the relation

$$(\tilde{M}_h v)_k = (M_h v)_k - \frac{1}{4}\left(g_{xx}(\xi_1)(v_{k+1} - v_k) + g_{xx}(\xi_2)(v_{k-1} - v_k)\right)$$

we conclude that

$$\left\langle \tilde{M}_h v, v \right\rangle \leq \|g_{xx}\|_{\mathrm{C}(\bar{\Omega})} \|v\|^2 ,$$

which is sufficient for unconditionally stability of the θ-method if $\theta \geq \frac{1}{2}$ as shown in 5.3.4. We summarise the result in a lemma.

Lemma 5.3.13
Let the coefficients of (5.39) be sufficiently smooth and $g(x) \geq 0$, $\forall x \in \bar{\Omega}_h \subset \mathbb{R}^1$. The Crank-Nicholson method is then stable for the space operator L_h as defined in Subsection 4.3.1 and zero boundary conditions if the grid $\bar{\Omega}_h$ is uniform and $\Delta t_k \sim \tau$. For $f = 0$ the method is convergent of order $\mathrm{O}(h^2 + \tau^2)$.

Proof By the above considerations the space operator is bounded by $\left\langle \tilde{L}_h v, v \right\rangle \leq C \|v\|^2$ for all $v \in \Phi_h$ with $v|_{\Gamma_h} = 0$. Stability then follows from Corollary 5.3.4. By Remark 5.1.5 the Crank-Nicholson method with that space operator is accurate of order $\mathrm{O}(h^2 + \tau^2)$ and by Theorem 5.3.5 convergent of the same order if the p.d.e. is homogeneous. □

We now turn to the question of stability under different boundary conditions. In [32, Section 10] Thomée shows the unconditional stability of the θ-method ($\frac{1}{2} \leq \theta \leq 1$) for Dirichlet, Neumann and third-kind boundary continuous. However, he always requires the boundedness of a from above and below, i.e. constants $c_1, c_2 > 0$ exist so that $c_1 \leq a(x) \leq c_2$, $\forall x \in \bar{\Omega}_h$. This assumption is been used to obtain a stronger estimate for M_h: $\langle M_h v, v\rangle \leq -C \|v\|^2$ with a positive constant $C > 0$.

Finally, we discuss stability of the numerical boundary condition where the p.d.e. is discretised using finite difference approximations from the left and the right, respectively. For example on the outflow boundary of the pure convection equation no boundary condition can be imposed so that it is sound to approximate the p.d.e. even in the border points. In 5.4.2 we will also suggest this numerical condition for the Heston p.d.e. at the zero variance boundary $x_2 = 0$. Even though the stability analysis is inconclusive we give the estimates which might be useful in further examinations of the problem.

We use the discretisation similar to (5.36) with (5.37) in all inner grid points Ω_h and define on the boundary

$$\begin{aligned}(M_h v)_0 &:= \frac{a(1/2h)}{h^2}(v_0 - 2v_1 + v_2) \qquad & (M_h v)_m &:= \frac{a((m-1/2)h)}{h^2}(v_{m-2} - 2v_{m-1} + v_m)\\ (N_h v)_0 &:= \frac{b(0)}{h}(v_1 - v_0) & (N_h v)_m &:= \frac{b(mh)}{h}(v_m - v_{m-1}).\end{aligned}$$

Employing the summation by parts rule we obtain the following estimates:

Lemma 5.3.14
For the above operators the following estimates are valid for any $v \in \Phi_h$:

$$\begin{aligned}\langle M_h v, v\rangle &= \left\langle \bar{a}_{-1/2}\bar{\partial} v, \bar{\partial} v\right\rangle_{(1,m)} + \frac{1}{h}\left(\bar{a}_{m-1/2}(2v_m - 3v_{m-1} + v_{m-2})v_m + \bar{a}_{1/2}(2v_0 - 3v_1 + v_2)\right),\\ \langle N_h v, v\rangle &\leq (1/2 + 1/4)\|b_x\|_{\mathrm{C}[a,b]} \|v\|^2 + \frac{1}{2}\left(2\bar{b}_m v_m^2 - 2\bar{b}_0 v_0^2 - \bar{b}_m v_m v_{m-1} + \bar{b}_0 v_0 v_1\right).\end{aligned}$$

Proof The first equation follows directly from applying summation by parts and adding the boundary terms. Following the same steps and additionally performing a Taylor expansion for $b(x_1)$ and $b(x_{m-1})$ leads to

$$\begin{aligned}\left\langle \hat{\partial} v, \bar{b}\cdot v\right\rangle_{(1,m-1)} &= -\left\langle v, \hat{\partial}(\bar{b}\cdot v)\right\rangle_{(1,m-1)} + \frac{1}{2}\left(v_m v_{m-1}(\bar{b}_m + \bar{b}_{m-1}) - v_0 v_1(\bar{b}_0 + \bar{b}_1)\right) \\ &= -\left\langle v, \hat{\partial}(\bar{b}\cdot v)\right\rangle_{(1,m-1)} + v_m v_{m-1}\bar{b}_m - v_0 v_1 \bar{b}_0 + \frac{h}{2}\left(b_x(\xi_1) v_m v_{m-1} + b_x(\xi_2) v_0 v_1\right).\end{aligned}$$

Estimating the last term $\frac{h}{2}\left(b_x(\xi_1)v_m v_{m-1} + b_x(\xi_2)v_0 v_1\right) \leq \frac{1}{2}\left\|b_x\right\|_{\mathrm{C}(\bar{\Omega}_h)}\left\|v\right\|^2$ and using the multiplication rule similar as in proof of Lemma 5.3.9 yields

$$\left\langle \hat{\partial} v, \bar{b}\cdot v\right\rangle_{(1,m-1)} \leq (1/2 + 1/4)\left\|b_x\right\|_{\mathrm{C}[a,b]}\left\|v\right\|^2 + \frac{1}{2}\left(v_m v_{m-1}\bar{b}_m + u_0 u_1 \bar{b}_0\right).$$

We finally infer

$$\begin{aligned}\left\langle N_h v, v\right\rangle_{(0,m)} &= \left\langle N_h v, v\right\rangle_{(1,m-1)} + \left(\bar{b}_m(v_m - v_{m-1})v_m + \bar{b}_0(v_1 - v_0)v_0\right) \\ &\leq (1 + 1/4)\left\|b_x\right\|_{\mathrm{C}[a,b]}\left\|v\right\|^2 + \frac{1}{2}\left(2\bar{b}_m v_m^2 - 2\bar{b}_0 v_0^2 - \bar{b}_m v_m v_{m-1} + \bar{b}_0 v_0 v_1\right).\end{aligned}$$

□

For outflow boundaries we have $\bar{b}_m v_m^2 - \bar{b}_0 v_0^2 \leq 0$. However, the estimation of $\bar{b}_m v_m v_{m-1} + \bar{b}_0 v_0 v_1$ still remains.

5.4 Conclusions for the Heston p.d.e.

Based on the general theory presented in this chapter we are now able to derive properties and make suggestions for the particular p.d.e. arising from Heston's stochastic volatility model. Due to the meaning of the space variables we switch from the general notation $(x_1, x_2) \in \Omega$ to $(x, v) \in \Omega$. The variable x represents the logarithm of the spot which is the value of the underlying financial product and v the square of its instantaneous volatility. The price of an option is denoted by $u(x, v, t)$. We write the p.d.e. u has to obey in the convection-diffusion form

$$u_t = \operatorname{div}(G\,\nabla\, u) - \operatorname{div}(ub) + cu$$

with

$$\begin{aligned}G(v) &= \frac{1}{2}v\begin{pmatrix}1 & \rho\xi \\ \rho\xi & \xi^2\end{pmatrix}, \\ b(v) &= v\begin{pmatrix}\frac{1}{2} \\ \kappa + \lambda\end{pmatrix} + \begin{pmatrix}\frac{1}{2}\rho\xi + r_{\mathrm{f}} - r_{\mathrm{d}} \\ \frac{1}{2}\xi^2 - \kappa\theta\end{pmatrix}, \\ c &= \kappa + \lambda - r_{\mathrm{d}}.\end{aligned}$$

The region Ω where the p.d.e. is defined depends on the type of option but is always a rectangular domain with $\Omega \subset \mathbb{R} \times \mathbb{R}^+$.

5.4.1 Consistency, stability and convergence

We note that the diffusion matrix $G(v)$ is obviously positive semidefinite for all $v \geq 0$ which follows directly from the determinant criteria since $g_{1,1}(v) \geq 0$, $g_{2,2}(v) \geq 0$ and $\det G(v) = \frac{1}{4}v^2\xi^2(1-\rho^2) \geq 0$ because ρ is the correlation: $|\rho| \leq 1$. Given Dirichlet conditions on the entire boundary it follows by Theorem 5.3.10 and Remark 5.3.11 that the Crank-Nicholson scheme is unconditionally stable in L_2 on a uniform grid with the space discretisation defined in (5.36) and (5.37) even if the left variance boundary is at $v = 0$. Given smooth initial data, convergence then follows immediately from Theorem 5.3.5.

Unfortunately, we felt short of the initial expectation to show stability on non uniform structured grids with the approximation of derivatives introduced in Section 4.2. By Taylor series expansion the second order accuracy in space as well as in time of the Crank-Nicholson method on these non uniform grids has been shown in Lemma 5.1.4 as long as the grids are created by a generating function. Without a conclusive stability result convergence can not be guaranteed. Numerical calculations show that second order convergence in space and time is reached in those cases where an analytic formula is known. However, we can not rule out that there exist parameter constellations so that the numerical scheme becomes unstable.

From the theory of constant coefficients described in Section 5.2 we know by Theorem 5.2.7 and Examples 5.2.10 to 5.2.14 that for diffusion dominated equations all central finite differences converge to the corresponding analytical solution. If the diffusion term vanishes and the convection dominates this is no longer the case and oscillations can be observed in the numerical solution. Upwind schemes as shown in Example 5.2.13 are suggested to cure that behaviour. In the Heston p.d.e. the diffusion term $G(v)$ is linear in v and so is the convection term $b(v)$ up to an additional constant. In my opinion the constant vector $(\frac{1}{2}\rho\xi + r_\mathrm{f} - r_\mathrm{d}, \frac{1}{2}\xi^2 - \kappa\theta)$ with parameters usual in finance can be neglected as far as the necessity of an upwind scheme is concerned. Only for huge and unrealistic asymmetries in the foreign and domestic currency an upwind scheme has to be considered. A demonstration is given in chapter 6 which substantiates these statements.

5.4.2 Boundaries and boundary conditions

Initial and boundary conditions are determined by the kind of option. The initial condition is always equal to the payoff of the derivative at time to maturity. Is the payoff not dependent on the path of the underlying value $s = \mathrm{e}^x$ but only on the value at maturity then no boundaries in spot direction exist, i.e. $\Omega = \mathbb{R} \times \mathbb{R}^+$. Up-and-out or down-and-out barrier options lose the value if at any time the value s hits a predefined level. That is modelled by setting the spot boundary to the upper and lower barrier level, respectively, and imposing Dirichlet conditions with value equal zero or equal to the rebate value if that has been specified. All options are independent of the instantaneous variance v and hence no additional boundary in v-direction is introduced. Even on the $v = 0$ boundary no conditions are required by the option or the stochastic model. That seems to be reasonable if the convection at $v = 0$ in v-direction given by $\frac{1}{2}\xi^2 - \kappa\theta$ is flowing outside. We know from Subsection 3.2.1 that for the pure convection equation no boundary condition can be imposed if the flow points outside. Since diffusion slowly disappears as v approaches the zero boundary it is possible that the same applies to the Heston p.d.e.

Motivated by the stability result of the numerical scheme we suggest to leave the left variance boundary at $v = 0$ and not to impose any artificial boundary condition. Instead we recommend to discretise the p.d.e. at these boundary grid points by finite differences from the right as it is for example done in upwind schemes. The accurate numerical implementation of the left variance boundary is decisive for errors of the numerical solution at the point (x^*, v^*) we are interested in since the boundary $v = 0$ is not far away from (x^*, v^*) so that errors at $v = 0$ influence the value at (x^*, v^*). However, I was neither able to show stability of this numerical boundary condition nor could I present a proof that no boundary condition can be imposed at $v = 0$. It only remains a well tested heuristic with some theoretical motivations.

For path independent options we have to introduce three artificial boundaries at $x = x_\mathrm{min}$, $x = x_\mathrm{max}$ and $v = v_\mathrm{max}$ because it is impossible to numerically calculate with infinite many grid points. Barrier options only require the artificial boundary at $v = v_\mathrm{max}$ since the other two boundaries are defined in a natural way. The question on how to choose x_min, x_max and v_max are addressed below. For the choice of the most appropriate condition we refer to the literature. We note that the current implementation of these conditions are uncritical since errors made there have a negligible effect at (x^*, v^*) if only the boundary is sufficiently far away. That is a nice property of parabolic p.d.e.s in general. One sometimes uses Dirichlet conditions with values one would expect from simpler models like the model by Black and Scholes. Dirichlet conditions have the advantage that they do not jeopardise stability of a scheme. In chapter 6 the boundary condition suggested for $v = 0$ is also used for all the artificial boundaries. It seems that the error made at the boundary is then the smallest compared to other conditions. However, as soon as the flow vector b is big enough and points inside, the numerical scheme becomes unstable.

In order to decide where the artificial boundaries should be placed we examine which influence a small disturbance at the boundary would have on the solution at (x^*, v^*). Since we do not know an analytic solution for arbitrary initial conditions we have to use our comprehension of convection-diffusion equations and consider both effects, convection and diffusion, separately. A typical vector field of b is shown in Figure 5.8. In a pure convection equation the disturbance moves with a velocity b and the way it passes is Tb assuming a constant b. For the $v = v_\mathrm{max}$ boundary

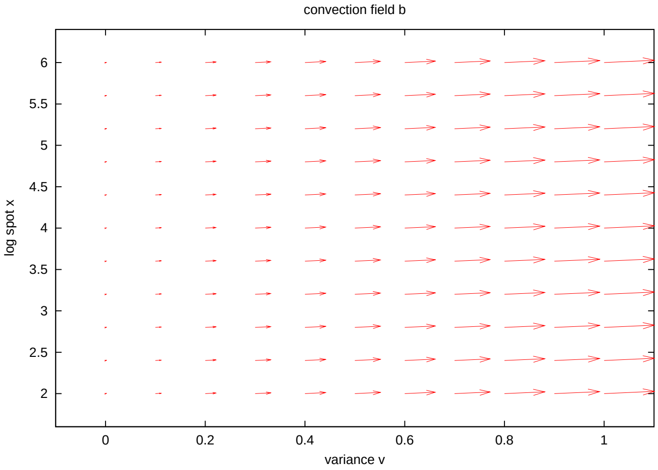

Figure 5.8: A typical convection field of foreign exchange markets

we only need to satisfy that $b_2(v_{\max}) = (\kappa+\lambda)v_{\max}\frac{1}{2}\xi^2 - \kappa\theta \geq 0$ since $\kappa+\lambda > 0$. The boundaries in spot direction need to satisfy $x_{\min} + b_1(v^*)T < x^*$ and $x_{\max} + b_1(v^*)T > x^*$. Given a diffusion only p.d.e. the disturbance has an immediate effect on any point of the solution. However, the magnitude strongly reduces as the distance to the disturbance increases. Quantitatively, this effect can be described using a fundamental solution. As we do not know the fundamental solution for the Heston p.d.e. we consider both dimensions separately and take as a fairly good approximation the fundamental solution to $u_t = \gamma u_{xx}$ and $u_t = \frac{\partial}{\partial v}(\gamma v u_v) - \frac{1}{2}\gamma u_v$ for the x- and v- direction, respectively. It follows from Lemma 3.3.1 that the fundamental solution of each respective p.d.e. is given by

$$\begin{aligned} G(x,x',t) &= \frac{1}{\sqrt{4\pi\gamma t}} \exp\left(-\frac{(x-x')^2}{4\gamma t}\right), \\ F(v,v',t) &= \frac{1}{\sqrt{4\pi\gamma v' t}} \exp\left(-\frac{(\sqrt{v}-\sqrt{v'})^2}{\gamma t}\right) \\ &\leq \frac{10}{\sqrt{4\pi\gamma t}} \exp\left(-\frac{(\sqrt{v}-\sqrt{v'})^2}{\gamma t}\right), \qquad \forall v' \geq \frac{1}{100}. \end{aligned}$$

A disturbance at $v_{\max}$ has a negligible influence on the solution at v^* if $F(v^*, v_{\max}, T) \leq \epsilon$ which is satisfied if $v_{\max} \geq \frac{1}{100}$ and

$$v_{\max} \geq \left(\sqrt{v^*} + \sqrt{-\gamma T \log\left(\frac{\epsilon}{10}\sqrt{4\pi\gamma T}\right)}\right)^2.$$

Since we want to approximate the diffusion part of the Heston p.d.e. in v-direction with $u_t = \frac{\partial}{\partial v}(\gamma v u_v) - \frac{1}{2}\gamma u_v$ we choose $\gamma := \frac{1}{2}\xi^2$. Similarly, we demand for the boundaries $x_{\min}$ and $x_{\max}$ that $G(x^*, x_{\max}, T) \leq \epsilon$ and $G(x^*, x_{\min}, T) \leq \epsilon$ which is equivalent to

$$\begin{aligned} x_{\min} &\leq x^* - \sqrt{-4\gamma T \log(\epsilon\sqrt{4\pi\gamma T})}, \\ x_{\max} &\geq x^* + \sqrt{-4\gamma T \log(\epsilon\sqrt{4\pi\gamma T})}. \end{aligned}$$

It is not obvious in this situation which value to assign to γ. A conservative approach would be to take $\gamma = \frac{1}{2}v_{\max}$, a more realistic choice maybe $\gamma = \frac{1}{2}\frac{v^*+v_{\max}}{2}$.

We have chosen the boundary sufficiently far away so that errors at the boundary have no significant impact on the quality of the solution at (x^*, v^*). We therefore do not discuss optimal conditions for the artificial boundaries and only refer to literature. In [19] optimal conditions for the Laplace equations are derived. Analytical properties of the condition $u_t = -bu_{xx}$ have been examined in [15]. Finally, absorbing boundary conditions are constructed and analysed in [9].

5.4.3 Space discretisation

The value function u obeys a p.d.e. which we solve numerically. With the finite difference method the numerical solution is calculated in all grid points and all time steps but at the end we are only interested in one value, namely $u(x^*, v^*, T)$ and sometimes also in some derivatives. Therefore the objective is to choose a grid which is optimal in the sense that the error at exactly that point (x^*, v^*, T) is minimal. Local error analysis is quite complicated since the operator $E_{h,\tau} = \big(I - \theta\tau L_h\big)^{-1}\big(I + (1-\theta)\tau L_h\big)$ mainly influences the error but is in general only numerically known. We present an approach which is not fully justified by theory but still yields very useful results.

From error analysis in Subsection 5.1.2 and particularly (5.10) it follows that the numerical error, i.e. the difference between analytical and numerical solution, is given by the same finite difference equation with the truncation error at the right hand side. In an equivalent formulation (5.8) the error can be expressed as a sum over solutions with the truncation error as part of the initial condition. If the numerical scheme is convergent, in a rough estimate (5.9) the local error can be expressed by the fundamental solution G_{Heston} of the Heston p.d.e. and the truncation error $\gamma_{h,\tau}$ which shall be be similarly distributed over time. Ignoring $A_{h,\tau}^{-1}$ in the estimate the error then is

$$(P_h u - \hat{u})(x^*, v^*, T) \approx C \int_\Omega \int_0^T G_{\text{Heston}}(x^*, v^*, x', v', t)\,\mathrm{d}t\; Q_h \gamma_{h,\tau}(x', v')\,\mathrm{d}x'\,\mathrm{d}v'.$$

The finite difference scheme requires a structured grid so that we construct one dimensional grids in x- and in v- direction. In both directions we approximate the Heston p.d.e. so that the coefficients best fit the diffusion term at the point (x^*, v^*). As above we choose the equation $u_t = \gamma u_{xx}$ and $u_t = \frac{\partial}{\partial v}(\gamma v u_v) - \frac{1}{2}\gamma u_v$ for x- and v- dimension, respectively, and try to find an optimal grid for each one dimensional problem. Beginning with the choice of a grid in the x-direction and denoting the integral of the fundamental solution G over t by Gi the error at the point x^* of the p.d.e. $u_t = \gamma u_{xx}$ is

$$(P_h u - \hat{u})(x^*, T) \approx C \int_\Omega Gi(x^*, x', T) Q_h \gamma_{h,\tau}(x')\,\mathrm{d}x'. \tag{5.42}$$

The truncation error $\gamma_{h,\tau}$ of the Crank-Nicholson method can by Lemma 4.2.3 and 5.1.4 estimated as follows

$$|Q_h \gamma_{h,\tau}(x_i)| = |\gamma_{h,\tau}(x_i)| \le C_1 \tau^2 + \gamma C_2 h^2 r(x_i)^2 + \gamma C_3 \|g''\|_{C[0,1]} h^2,$$

where r denotes the distance ratio function and g the grid generating function as introduced in Subsection 4.1.1. The value $r(x_i)$ describes the ratio between the distance to the next grid point and the distances in the corresponding uniform grid. Hence the truncation error consists of a time step error, an error proportional to the square of the distances of two grid points and an error due to the non uniformity of the grid. We only try to minimise the influence of the second part, still keeping in mind that if the resulting grid becomes strongly non uniform it will not be optimal since the error due to non uniformity can not be ignored. The constants C_1, C_2, C_3 depend on the norm of some derivatives of u. In the following we assume that the solution u is equally smooth over the entire domain $\Omega \subset \mathbb{R}$ so that the constants are similar in each grid point. This requirement excludes for example the initial and boundary conditions of reverse barrier options. Inserting the estimate for the truncation error into (5.42) and minimising the error at x^* leads to the optimisation problem

$$\begin{cases} \int_{x_{\min}}^{x_{\max}} Gi(x^*, x', T) r(x')^2\,\mathrm{d}x' \to \min \\ \int_{x_{\min}}^{x_{\max}} \frac{1}{r(x')}\,\mathrm{d}x' = x_{\max} - x_{\min} \end{cases}$$

which we do not solve but instead apply the equidistribution principle in the hope it gives a result not far away from the optimal solution for r. The equidistribution principle requires

$$Gi(x^*, x', T) r(x')^2 = C, \qquad \forall x' \in [x_{\min}, x_{\max}].$$

We therefore choose

$$\boldsymbol{r(x') = \frac{C}{\sqrt{Gi(x^*, x', T)}}}, \tag{5.43}$$

with a constant C defined by

$$C := \int_{x_{\min}}^{x_{\max}} \frac{1}{\sqrt{Gi(x^*, x', T)}}\,\mathrm{d}x'.$$

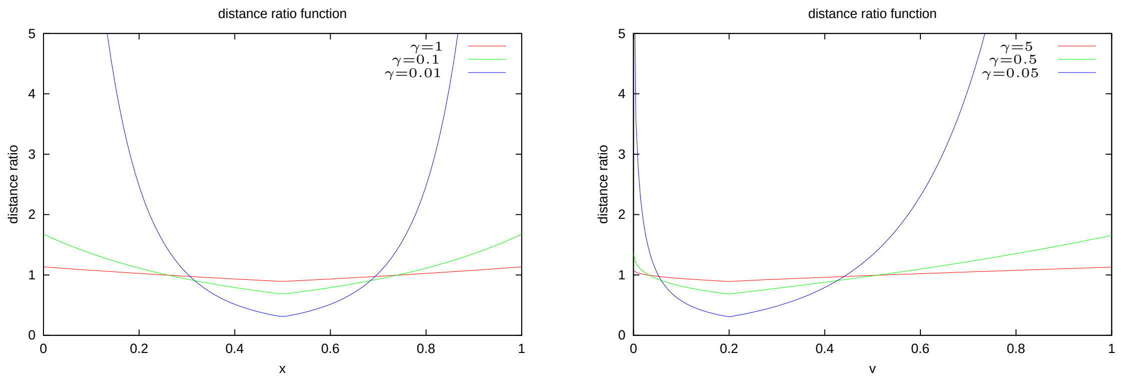

Figure 5.9: Distance ratio function for $u_t = \gamma u_{xx}$ and $u_t = \frac{\partial}{\partial v}(\gamma v u_v) - \frac{1}{2}\gamma u_v$ at $T = 1$

For the v-axis we propose by similar deliberations

$$\boldsymbol{r(v') = \frac{C\sqrt[4]{v'}}{\sqrt{Fi(v^*, v', T)}}}. \tag{5.44}$$

It remains the question whether there is an explicit representation of the functions Gi and Fi defined as the integral of the respective fundamental solution over t. Since both functions G and F are of the same principal form we answer the question at once by giving the result of the integral

$$\begin{aligned}\int_0^T \frac{c(x,x')}{\sqrt{t}} \exp\left(-\frac{d(x,x')}{t}\right) \mathrm{d}t \; &= 2c(x,x')\left(\sqrt{T}\exp\left(-\frac{d(x,x')}{T}\right) + \sqrt{d(x,x')\pi}\,\operatorname{erf}\sqrt{\frac{d(x,x')}{T}}\right) \\ &- 2c(x,x')\sqrt{d(x,x')\pi}\end{aligned}$$

where the so called error function erf is defined as the commutative normal distribution

$$\operatorname{erf}(x) = \int_{-\infty}^{x} \frac{1}{2\pi}\,\mathrm{e}^{-\frac{x'^2}{2}}\,\mathrm{d}x'.$$

As an illustration Figure 5.9 shows the distance ratio functions for some diffusion parameters γ. The central points (x^*, v^*) in these examples are $x^* = \frac{1}{2}$ and $v^* = \frac{1}{5}$. It can clearly be seen that if diffusion is quite big the distance ratio function has values near to one and the corresponding grid will be almost uniform. This is not unexpected because strong diffusion means that a disturbance at some remote point is blurred quickly over the entire domain.

5.4.4 Time discretisation

In Subsection 5.1.2 it has already been remarked that the usage of non uniform time steps might be useful. It is in particular important if the truncation error is not of the same order of magnitude over time. From (5.10) it follows that time steps should be made smaller if the truncation error increases. The truncation error mainly depends on the smoothness of the solution. Since uniformly parabolic equations exhibit a smoothing property, i.e. the solution becomes smoother with time, the truncation error is expected to be bigger near the initial time. If the initial condition is not at least two times continuously differentiable the truncation error might be severely big in the first time steps and a refinement of the grid in time is highly recommended. Since the payoff functions of options are often not differentiable in some points and might even be discontinuous in the case of reverse barrier options we always use a finer grid near the time zero.

In [3] some particular choices of time grids were analysed, for example

$$\Delta t_k = (\alpha + \beta t_k)h^2$$

with some positive parameters $\alpha, \beta > 0$. We use a similar grid which is generated by a distance ratio function r of the form discussed in Subsection 4.1.3:

$$r(t) = \sqrt{\alpha^2 + \beta t^2}.$$

Numerical calculation have shown that in general a value of $\alpha = \frac{1}{10}$ or even $\alpha = \frac{1}{100}$ is advantageous which means that time grid points are ten or one hundred times as dense at $t = 0$ as in the uniform case. If one uses a direct matrix solver in each time step the LU matrix decomposition has to be performed in each step because if Δt changes the matrix $A_{h,\tau} = I - \theta \tau L_h$ changes, too. In this case one should consider to choose the time steps piecewise uniform.

Chapter 6

Numerical results

This chapter is dedicated to the practical aspects of the finite difference method as described in Chapter 4 applied to the Heston p.d.e. We show that the stability of the method is not in question even though the in Chapter 5 derived stability result only applies for uniform grids and a different approximation of mixed derivative as proposed in Chapter 4. The suggestion given in Section 5.4 to improve accuracy of the numerical solution like the particular boundary condition at $v = 0$ and the choice of the non uniform grids in space and time direction are also substantiated below.

6.1 Details of the algorithm

The finite difference method (Crank-Nicholson) with a 3^d-point compact stencil capable of dealing with non uniform grids (see Section 4.2) has been implemented in C++. Also a simple ADI scheme where the mixed derivative part is treated as explicit (see Subsection 4.3.2) has been integrated. The method is only second order accurate and stable if the coefficient in front of the mixed derivative is negligible. The most computing intensive part of the θ-method and in particular of the Crank-Nicholson method is solving the linear equation system. Having tested some different algorithms like SOR, BiCG with pre-conditioning and also a few direct methods based on the LU decomposition I recommend to use direct methods which strongly outperform iterative methods if the same equation system has to be solved many times with a different right hand side. That is usual for time independent parabolic p.d.e.s if the time grid is chosen to be uniform. The implementation of a direct solver which takes advantage of the sparsity of matrices is expensive. That is why I use the freely available SuperLU package.

The programme is compiled with the GNU C++ compiler and maximum optimisation (gcc -O3). Computing times are given for a Pentium II with 300 MHz equipped with 512 MB RAM under Linux. The time for generating the non uniform grid is excluded in the figures because as yet, this is only poorly and inefficiently implemented and might take up to 0.5 seconds.

For convergence analysis we use the semi analytic formulas for plain vanilla options [8, Chapter 23] implemented by Gunter Winkler and for barrier options with zero correlation and foreign equal domestic interest rate [6].

6.2 Selected examples

We present three examples, two of them have been calibrated to the foreign exchange market of USD/JPY (US Dollar in Japanese Yen) at some time and the parameters of the third example has been taken from [8, Chapter 24].

6.2.1 Call option on USD/JPY

We consider the exchange rate of USD/JPY that is the value of one US Dollar in Japanese Yen. The Heston parameters were calibrated to the market and shown in Table 6.1. We want to by a call option with strike $K = S^* = 123.4$ and a time of $T = 0.50137$ years. In order to get an imagination of the market some sample paths of the underlying Heston process can be viewed in

S^*	123.4	x^*	4.8154	v^*	0.014328
r_d	log(1.0005)	r_f	log(1.0375)		
θ	0.011876	κ	1.98937		
ξ	0.33147	ρ	0.0258519	λ	0
K	123.4	T	0.50137		
exact price	2.7207095899				

Table 6.1: Call option on USD/JPY

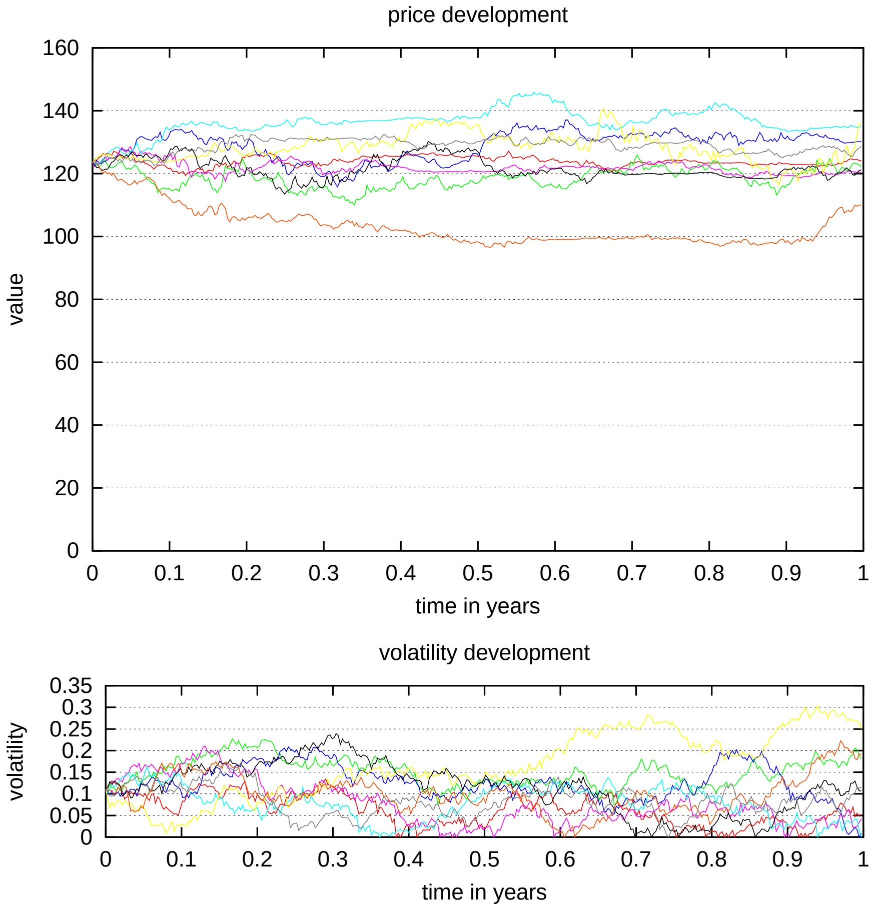

Figure 6.1: Sample path of the stochastic process S_t and $\sqrt{v_t}$

Figure 6.1. The convection field b of the corresponding p.d.e. can be seen in Figure 6.2. According to Subsection 5.4.2 the boundaries are set to $x \in [2.990790, 6.640072]$ and $v \in [0.0, 0.559951]$ if $\epsilon = 10^{-5}$. If we follow the method in Subsection 5.4.3 the non uniform grid as in Figure 6.3 will be generated.

The solution of the finite difference scheme is shown in Figure 6.4 where the x-axis is already back transformed to spot values. First the initial condition is portrayed which is equal to the payoff of the option and then the solution at maturity is shown. A numerical error near the edge $(s_{\text{min}}, v_{\text{max}})$ can be seen but which has no impact on the solution at (s^*, v^*).

6.2.2 Up and out call option on USD/JPY

In the second example we use almost the same market as in the first. Since we want to compare the numerical results with the analytical solution which is only available for $\rho = 0$ and $r_{\text{d}} = r_{\text{f}}$ we choose the market with parameters as in Table 6.2. We specify the barrier option as follows: $T = 0.50137$, $K = 120$ and the up and out barrier is at 127. The boundaries in this example are then set to $x \in [2.990790, 4.844187]$ and $v \in [0.0, 0.559951]$ and the grid looks like in Figure 6.5. The initial condition and the final solution are shown in Figure 6.6 where the x-axis has been transformed back to spot values. For a more detailed view the third picture again shows the final

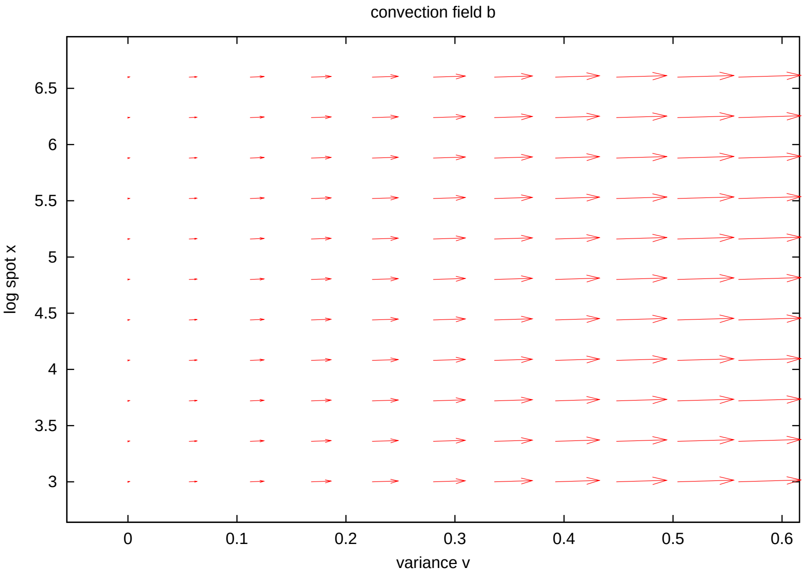

Figure 6.2: Convection field b of the p.d.e.

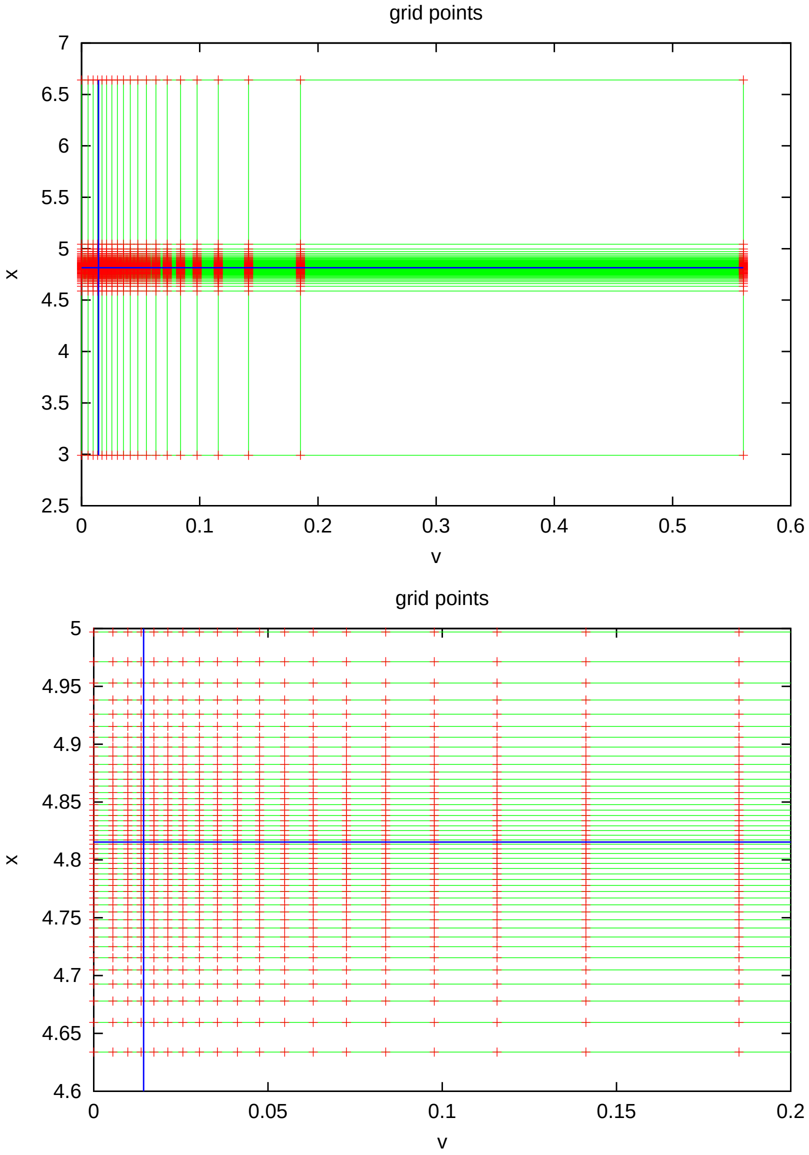

Figure 6.3: Non uniform structured grid

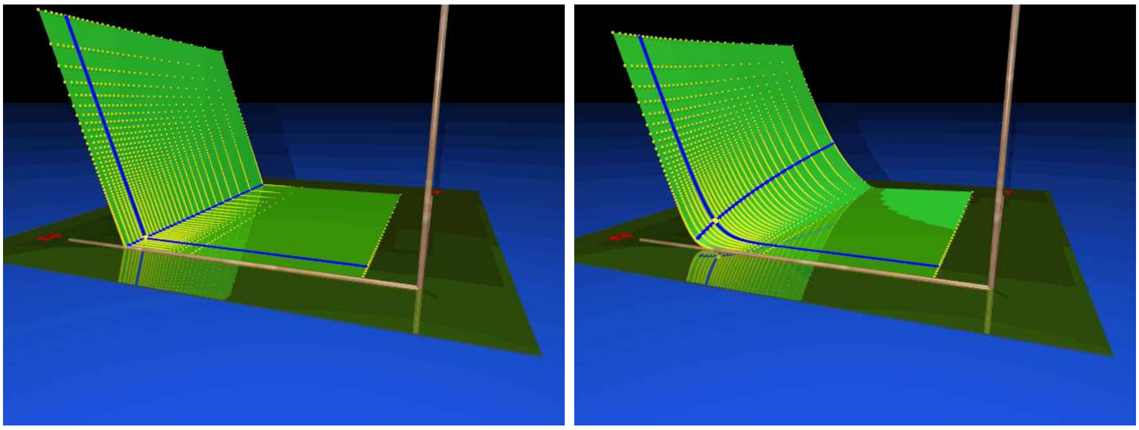

Figure 6.4: Initial condition and final solution of the call option (z-scale: 50)

S^*	123.4	x^*	4.8154	v^*	0.014328
r_d	log(1.0005)	r_f	log(1.0005)		
θ	0.011876	κ	1.98937		
ξ	0.33147	ρ	0	λ	0
K	120	B	127	T	0.50137
exact price	0.3312511368				

Table 6.2: Up and out call on USD/JPY

solution but ten times zoomed in.

6.2.3 Call option on an underlying with higher volatility

The parameters of Table 6.3 are identical to the example treated in [8, Chapter 24]. We consider the call option with strike $K = 1$ and time to maturity $T = 0.25$. The boundaries are set to $x \in [-1.583349, 1.583349]$ and $v \in [0.0, 0.804015]$. The convection field and the non uniform grid can be seen in Figure 6.7 and 6.8.

6.3 Influence of diffusion and convection

The critical area of the Heston p.d.e. is where the diffusion vanishes and therefore the flow dominates. By theory we know that the Crank-Nicholson scheme is still stable in L_2 but the derivatives at these points are unreliable since the solution begins to oscillate. This problem can be bypassed using an upwind scheme where derivative are approximated in that direction from where the flow is coming. We show, however, that the flow at $v = 0$ can be neglected for realistic market parameters and hence the implementation of an upwind scheme is not essential. We demonstrate that by choosing some peaks as initial condition and solve the p.d.e. In Figure 6.9 the solution u is shown at four different time points with parameters as in Table 6.1. Only if we strongly increase the difference between the domestic and foreign interest rate the influence of the convection can be seen. For the same parameters but with $r_f = \log(10)$, i.e. the foreign interest rate is at 1000%, the solution looks as in Figure 6.10. The same is shown in Figure 6.12 but with an even higher interest rate of $r_f = \log(100)$.

S^*	1.0	x^*	0.0	v^*	0.05225
r_d	log(1.052)	r_f	log(1.048)		
θ	0.06	κ	2.5		
ξ	0.5	ρ	-0.1	λ	0
K	1	T	0.25		
exact price	0.0449439664				

Table 6.3: Call option on an underlying with higher volatility

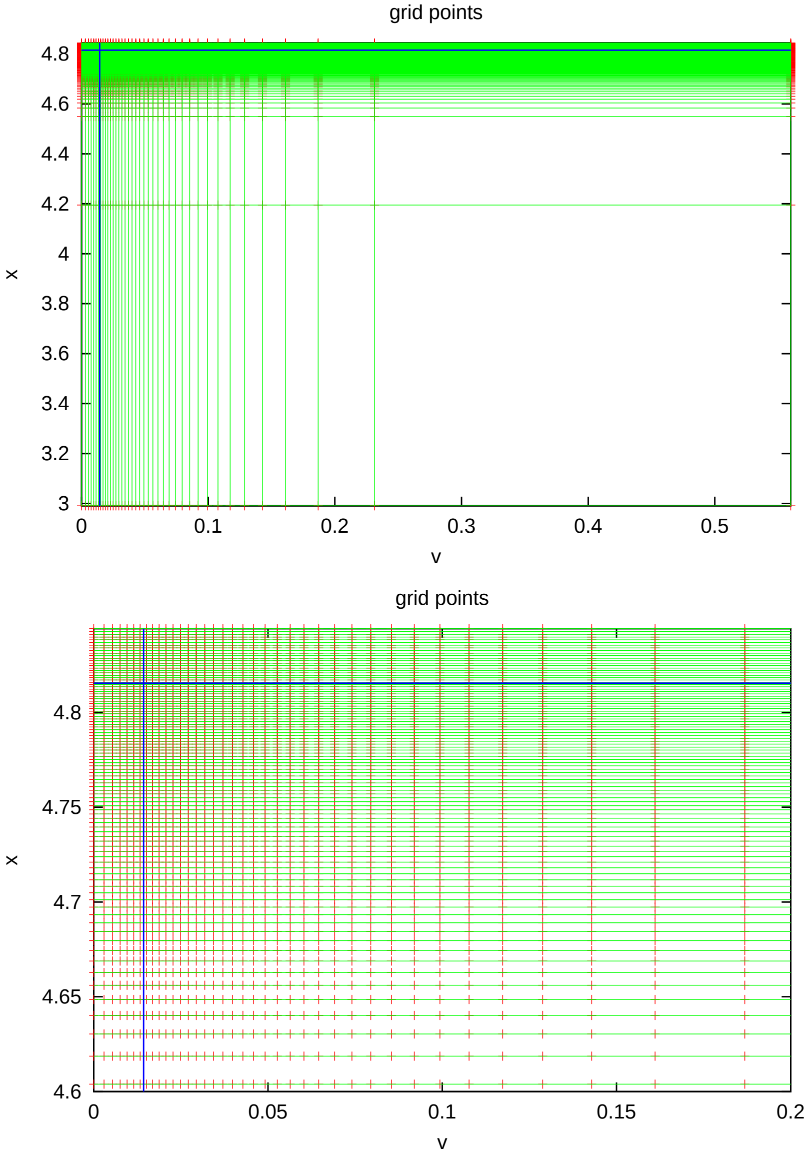

Figure 6.5: Non uniform structured grid for an up and out barrier

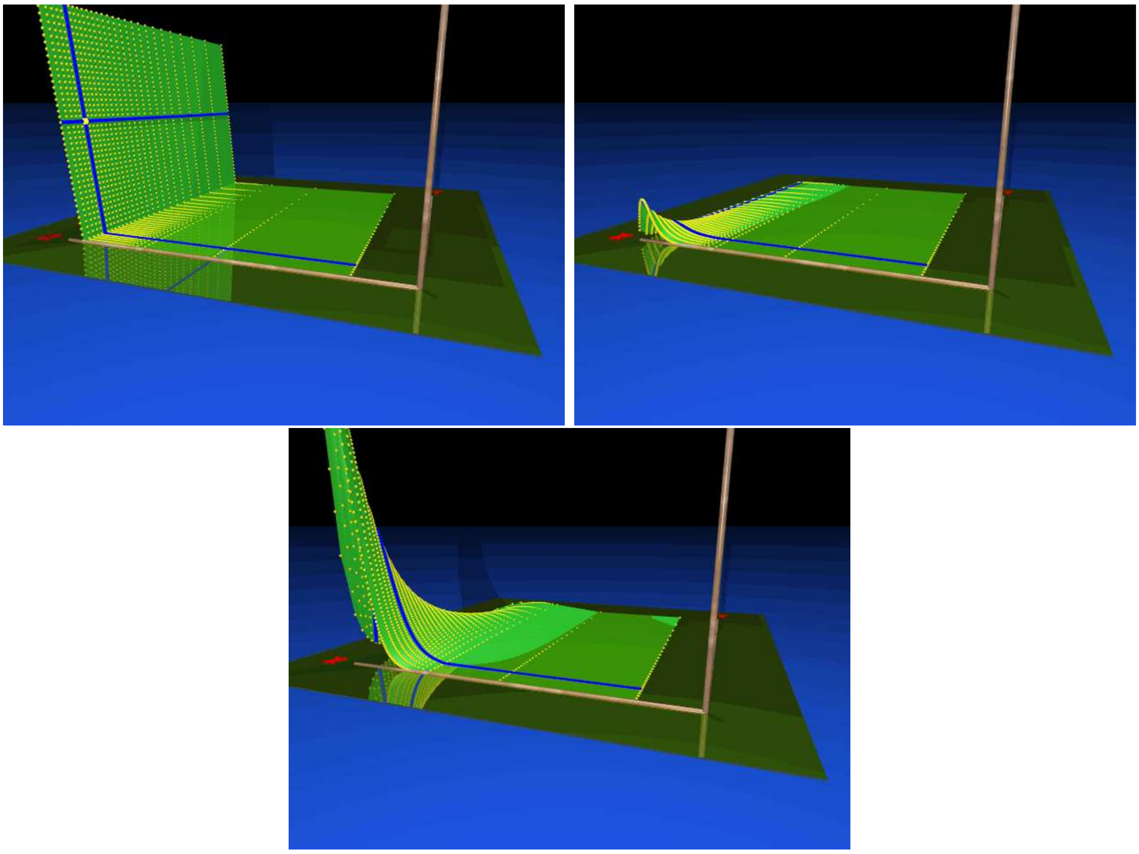

Figure 6.6: Initial condition and final solution of the barrier option (z-scale: 10) and the last picture shows the final solution zoomed in (z-scale: 1)

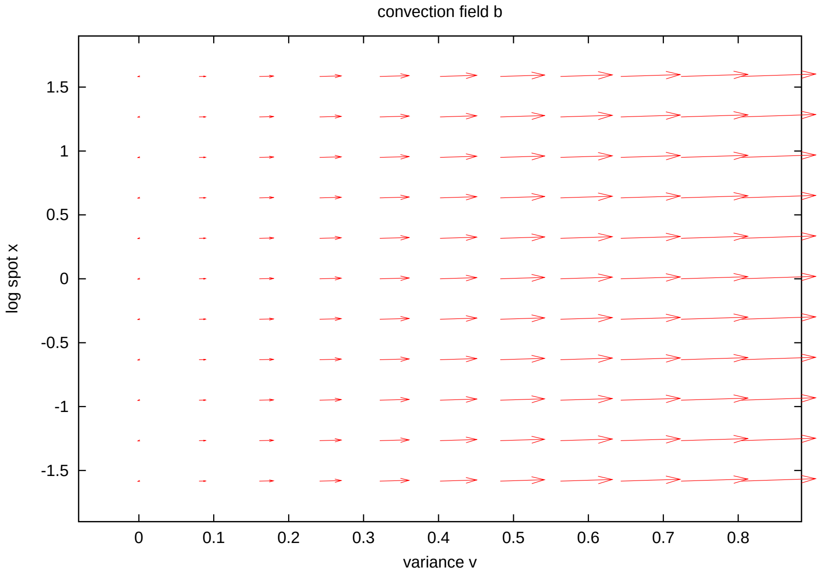

Figure 6.7: Convection field b of the p.d.e.

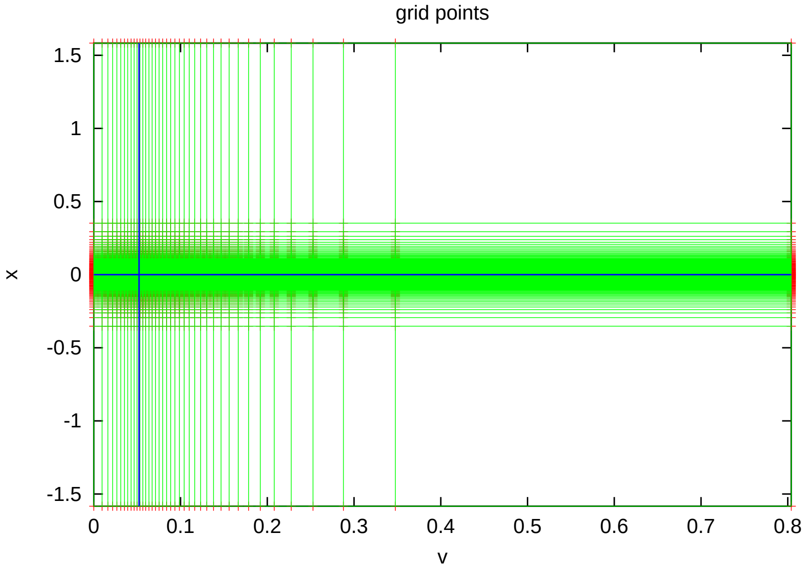

Figure 6.8: Non uniform structured grid

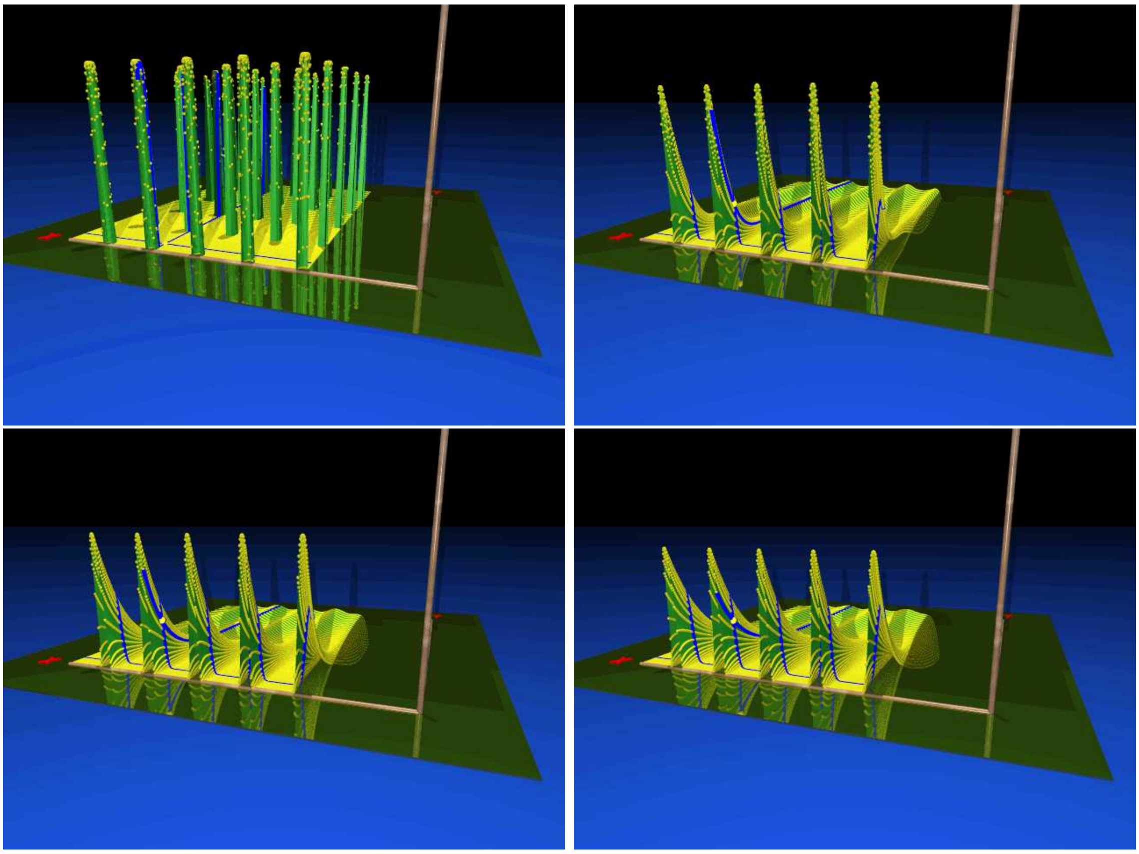

Figure 6.9: Convection and diffusion of the Heston p.d.e.

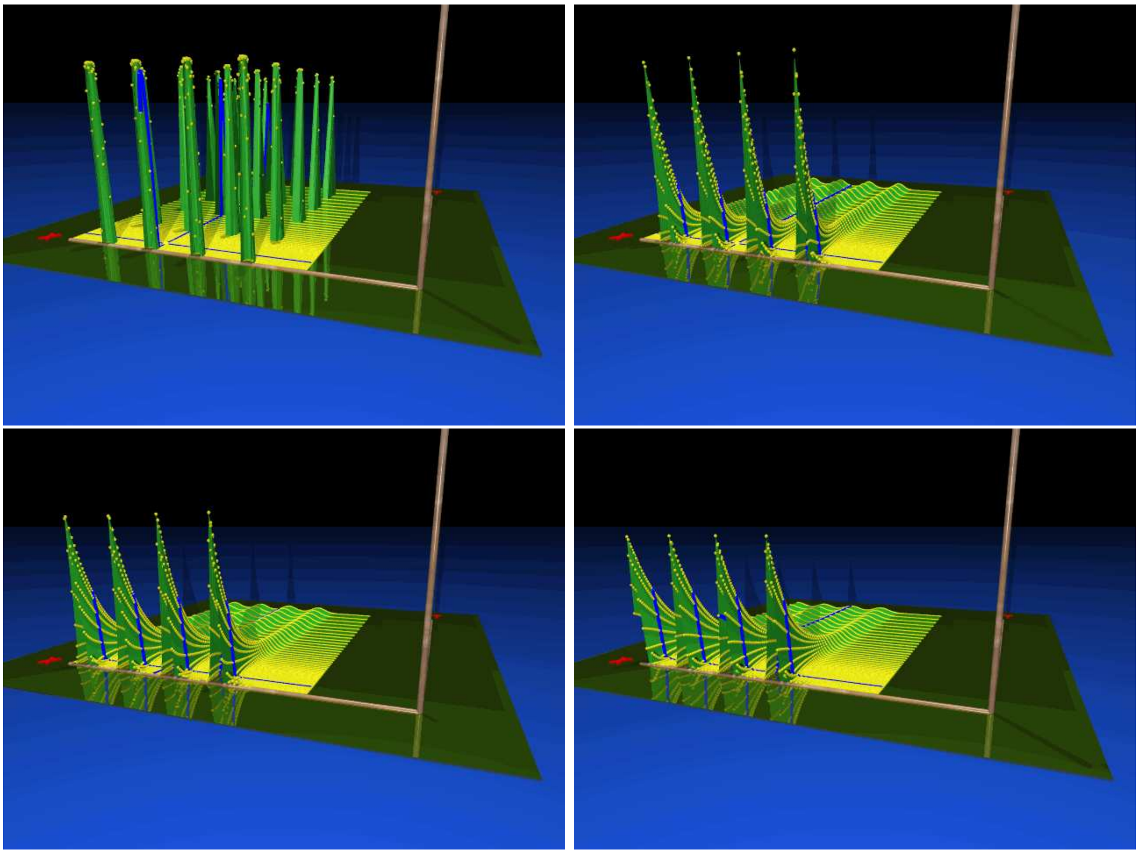

Figure 6.10: Convection and diffusion of the Heston p.d.e. with $r_{\mathrm{f}} = \log(10)$

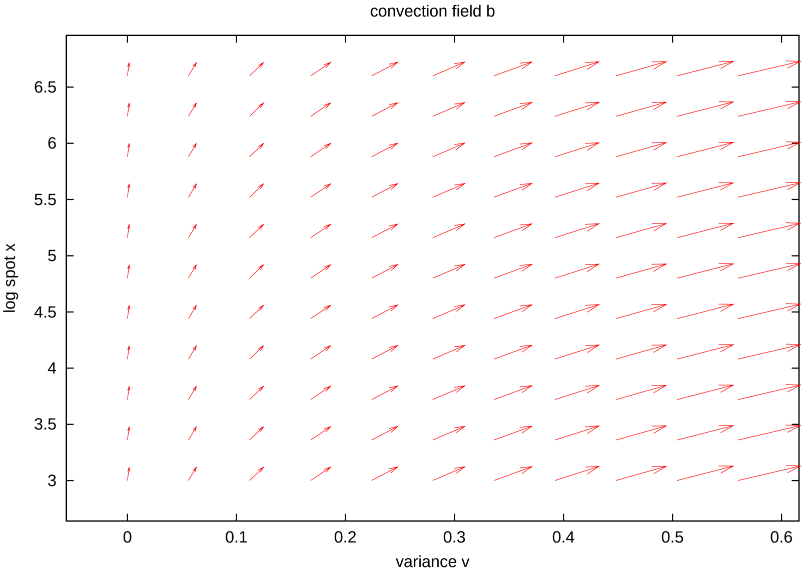

Figure 6.11: Convection field b for $r_{\mathrm{f}} = \log(10)$

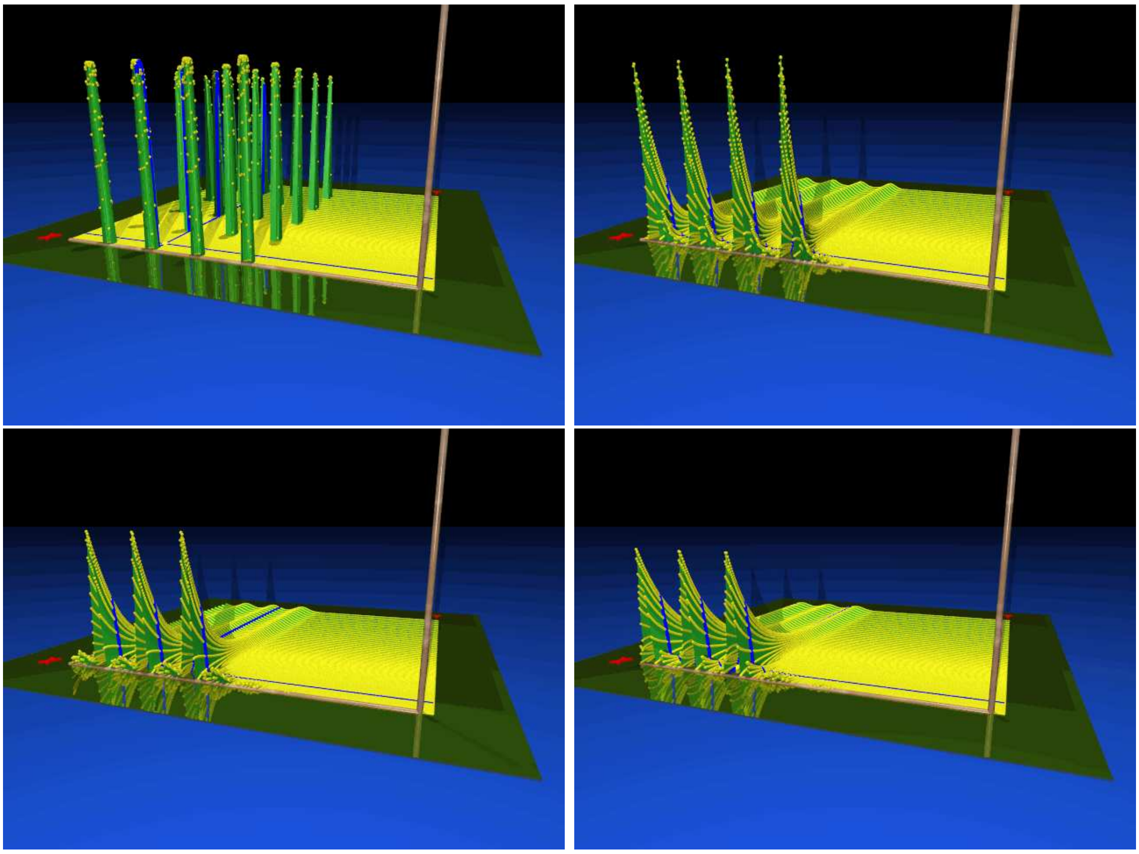

Figure 6.12: Convection and diffusion of the Heston p.d.e. with $r_\mathrm{f} = \log(100)$

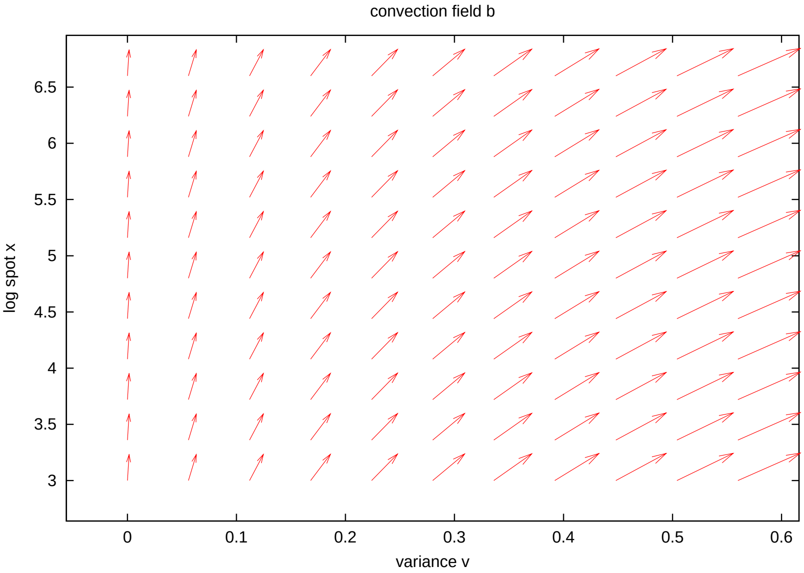

Figure 6.13: Convection field b for $r_\mathrm{f} = \log(100)$

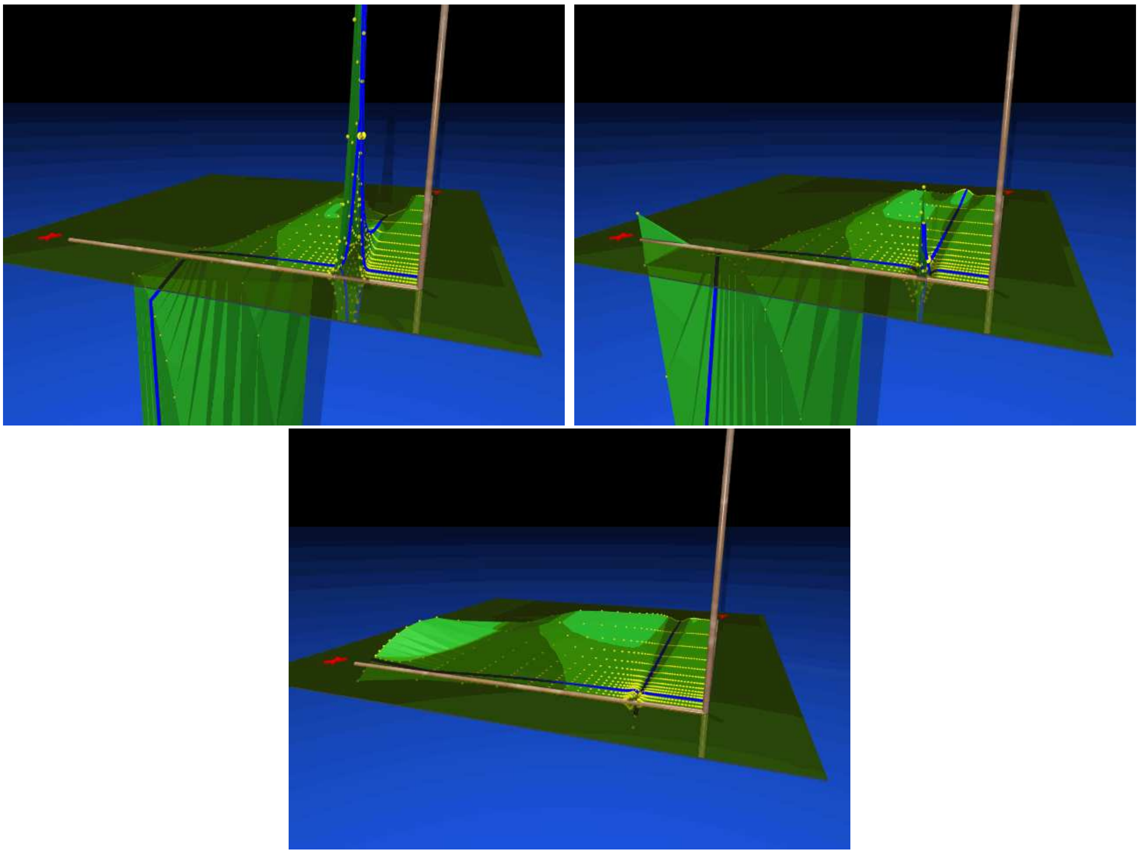

Figure 6.14: The error $\hat{u} - P_h u$ at T for different numerical boundary conditions

6.4 Improvements of the accuracy of the numerical solution

The effect of the particular boundary condition suggested in Subsection 5.4.2 and the effect of non uniform grids are demonstrated. We compare the numerical solution of the first example (call option on USD/JPY) with its analytical solution and plot the error over the entire region Ω_h.

First we demonstrate the impact of different boundary conditions, see Figure 6.14. In the upper left picture homogeneous Neumann conditions have been imposed. We observe a quite big error at the $v = 0$ boundary near the strike. The error reduces if we instead require that second derivatives have to be zero and almost vanishes if we discretise the p.d.e. in the boundary points as shown in the lower picture. It also becomes clear that an accurate modelling of the boundary $v = 0$ is crucial for good numerical values at (s^*, v^*).

Using appropriate non uniform grids can greatly improve the quality of the solution at the particular point (s^*, v^*). In the upper left picture of Figure 6.15 a uniform grid has been chosen and in the right picture the grid has been created using a generating function as described in Subsection 4.1.3. The grid on the lower picture is based on the grid which has been derived in 5.4.3 and aims to be locally optimal for the point (s^*, v^*). And indeed, the local error is impressively small (as shown in Table 6.5) but obviously only in a small neighbourhood of (s^*, v^*).

Finally we show the result of using non uniform grids in time. The effect is particularly impressive for reverse barrier options. Hence Figure 6.16 and 6.17 show the error of the barrier option problem, the first on a uniform and the latter on a non uniform time grid (100 times as dense at $t = 0$). Both figures show a series of errors at four different times. It can be seen that if time steps are uniform a huge error is made in the first step. The error then continously reduces with time but still has an undesirable effect on the final solution.

6.5 Results of the ADI- and Crank-Nicholson scheme

The following calculations are based on grids in space direction as defined in Table 6.4. For plain vanilla call options we choose the a non uniform time grid so that at $t = 0$ the time step size is ten

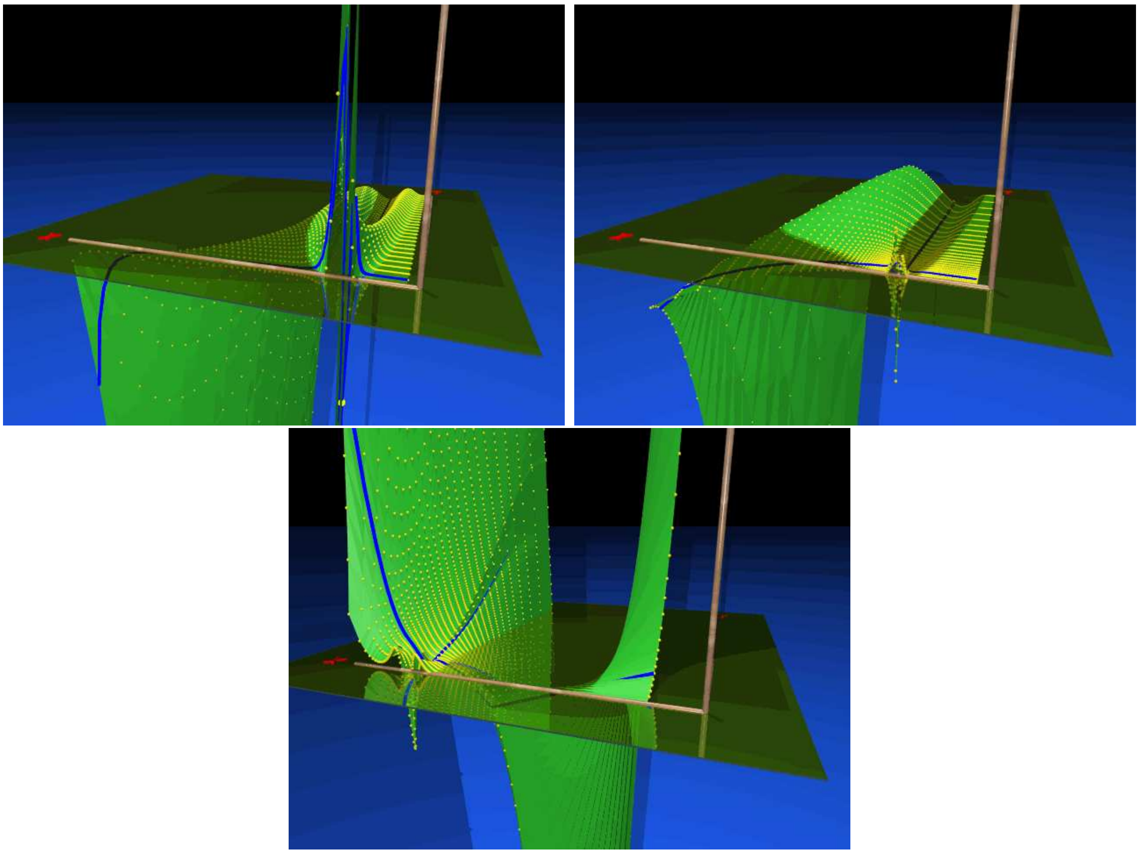

Figure 6.15: The error $\hat{u} - P_h u$ at T with different space discretisation (z-scale: 0.1)

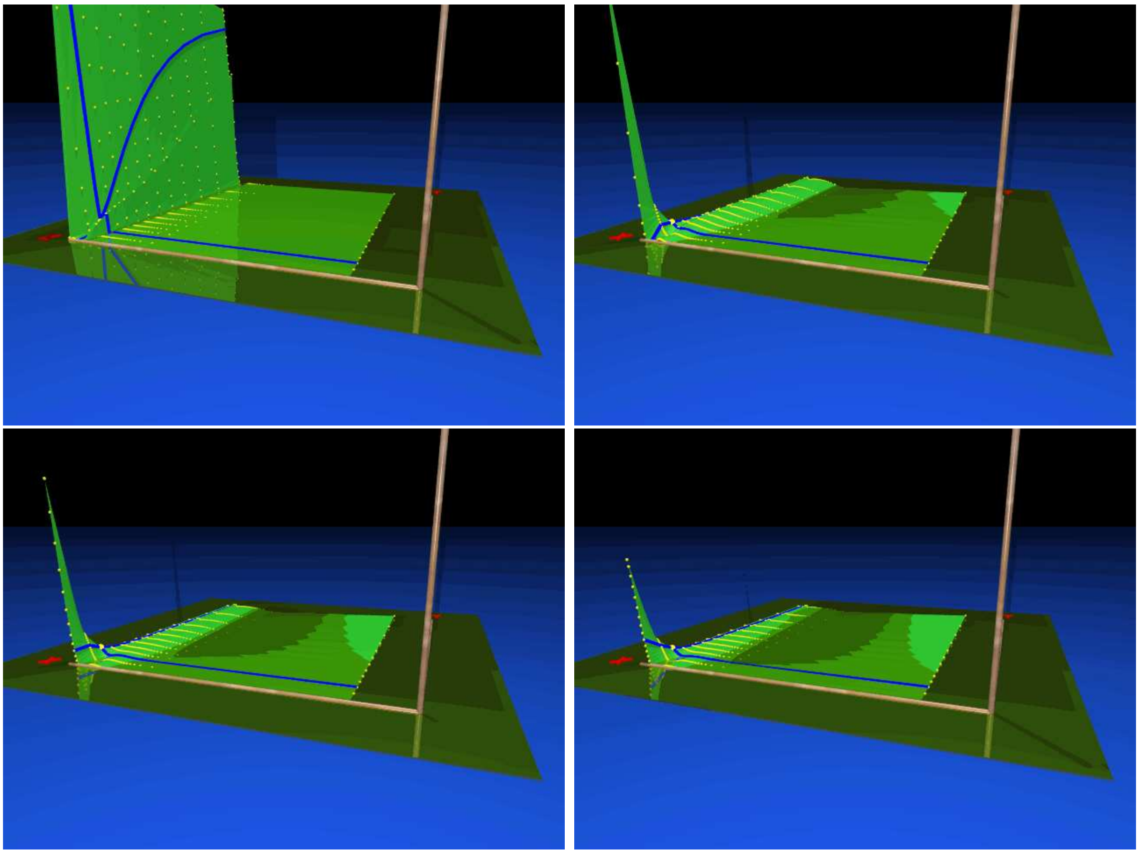

Figure 6.16: The error $\hat{u} - P_h u$ at times to maturity $t = 4.6$, 64.1, 119.0 and 173.9 days, respectively (z-scale: 2)

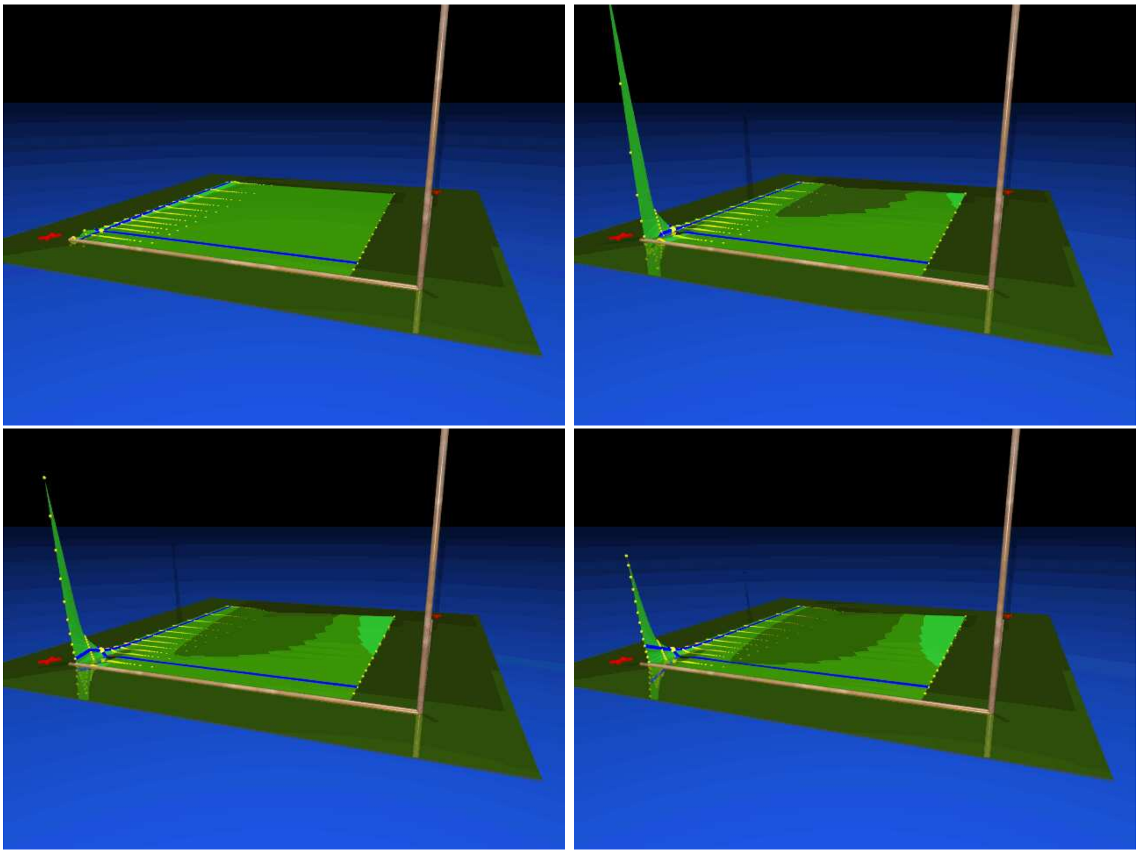

Figure 6.17: The error $\hat{u} - P_h u$ at times to maturity $t = 4.2$, 60.2, 112.8 and 165.5, respectively; time grid non uniform (z-scale: 2)

grid	points in x	points in y	matrix size	time points (call)	time points (barrier)
1	25	8	200	10	20
2	50	15	750	20	40
3	100	30	3000	40	80
4	200	60	12000	80	160
5	400	120	48000	160	320

Table 6.4: Grids used in the numerical simulations

grid	result of scenario			absolute error of scenario		
	1	2	3	1	2	3
1	6.5208303	2.5240578	2.6975496	3.8000348	-0.1967376	-0.0232458
2	5.4382319	2.7044851	2.7174149	2.7174363	-0.0163104	-0.0033806
3	4.0524806	2.7174085	2.7201532	1.3316851	-0.0033870	-0.0006422
4	3.3863823	2.7201582	2.7206148	0.6655867	-0.0006372	-0.0001806
5	3.0786712	2.7206342	2.7207095	0.3578757	-0.0001613	-0.0000859

Table 6.5: Results of different improvements

grid	time	LU-time	s-time	a-time	a-b-time	memory	value	error
1	0.3						2.6975496	-0.0232458
2	0.9					3 MB	2.7174149	-0.0033806
3	5.8					6 MB	2.7201532	-0.0006422
4	41.0	2.6	0.1	0.4	0.1	20 MB	2.7206148	-0.0001806
5	439.8	18.7	0.8	1.7	0.4	80 MB	2.7207095	-0.0000859

Table 6.6: Efficiency of the Crank-Nicholson method

times as large as for uniform grids. For reverse barrier option we choose 100 as the factor. In both cases the time grid is modified to obtain piecewise uniform time steps so that the LU-decomposition only has to be performed at five time points.

6.5.1 Improvement of the solution

Based on the example of a call option on USD/JPY we demonstrate how the suggestion made within this thesis strongly improve the accuracy of the numerical solution. In Scenario one we use a uniform grid in space and in time. The boundary condition is set to homogeneous Neumann conditions. As can be seen in Table 6.5 the results are unacceptable. Scenario two uses a non uniform grid in space according to Subsection 4.1.3 with the parameter $c = 0.2$ and the concentration points K and v^*, respectively. It uses the boundary condition where the p.d.e. is discretised. The last scenario is based on the non uniform grid as proposed in Subsection 5.4.3 and shown in Figure 6.3. The accuracy with that grid is remarkable.

6.5.2 Comparison between Crank-Nicholson and ADI method

From now on we only discuss the results of the best discretisation, i.e. only Scenario three is considered. We compare the efficiency and accuracy of the ADI and Crank-Nicholson scheme with the parameters of the first example (call option on USD/JPY). In Table 6.6 and 6.7 some information about the computing time is given. The titles have the following meanings:

time	total computing time
LU-time	time for solving the equation system $Ax = b$ including the LU decomposition
s-time	time for solving the equation system $Ax = b$ if LU decomposition is known
a-time	time to assemble the matrix A and the vector b
a-b-time	time to assemble the vector b

Three observations can be made. Both methods are quite efficient and accurate. Even with Grid two the error is about 0.1% and the total computing time less than one second which is sufficient for most of the practical applications. In both methods the time to solve the equation system once the LU-decomposition has been performed is less than the time to assemble the vector b which

grid	time	LU-time	s-time	a-time	a-b-time	memory	value	error
1	0.1						2.6975496	-0.0232458
2	0.4						2.7188348	-0.0019606
3	2.4						2.7205944	-0.0002010
4	17.8	0.1	0.0	0.3	0.2	4 MB	2.7207878	-0.0000076
5	142.9	0.4	0.1	1.1	0.6	8 MB	2.7211212	0.0003257

Table 6.7: Efficiency of the ADI method

grid	time	value	relative error	time	value	relative error
1	0.1	0.0448298	-0.25397%	0.1	0.3902796	17.81988%
2	0.3	0.0449514	0.01656%	0.6	0.3526465	6.45897%
3	2.3	0.0449475	0.00798%	4.5	0.3378257	1.98476%
4	17.6	0.0448605	-0.18566%	33.5	0.3327874	0.46380%

Table 6.8: Comparison between accuracy of plain vanilla and reverse barrier option

might be due to some inefficiencies in the assembly routines. The order of convergence seems to be partially better than second order. In the last grid of the ADI scheme the error increases significantly. That is because the method I use is only stable if the correlation ρ is equal zero.

6.5.3 Solution of the second and third example

We finally compare the accuracy of the results for a reverse boundary option with a plain vanilla call option. The ADI method with improved boundary conditions and the fairly optimal grid is used. With the parameters of the second (Table 6.2) and third example (Table 6.3) we obtain the results which are summarised in Table 6.8.

The numerical errors for reverse barrier options are very big compared to plain vanilla options. I believe that the error is mainly caused by inappropriate numerical boundary conditions at $v = 0$. According to Subsection 5.4.2 the suggested numerical boundary condition shall only be used if $\frac{1}{2}\xi^2 - \kappa\theta < 0$ which is not fulfilled in that example.

Again, the particular ADI method which I use shows instabilities if the correlation is not zero.

Appendix A

Mathematical methods

A.1 Fourier transformation

The Fourier transformation defined over $\mathbb{R}^d$ can be considered to be a generalisation of the Fourier series expansion. In short, the Fourier coefficients are the coordinates of a function $v : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ with respect to a complete orthogonal basis (given appropriate Ω) consisting of sine and cosine functions $\{s_k\}_{k\in\mathbb{N}_0^d} \cup \{c_k\}_{k\in\mathbb{N}_0^d}$, $s_k(x) := \sin\langle k, x\rangle$, $c_k(x) := \cos\langle k, x\rangle$. In that basis the coordinates of the function v are

$$v = \sum_{k\in\mathbb{N}_0^+} \alpha_k c_k + \beta_k s_k, \quad \text{with } \alpha_k := \|c_k\|^{-1} \langle v, c_k\rangle\,, \text{ and } \beta_k := \|s_k\|^{-1} \langle v, s_k\rangle\,.$$

There exist similar results if we admit real values for k. Due to its interpretation as a frequency we denote it by ω. Let now $v : \mathbb{R}^d \to \mathbb{R}$ be a function which satisfies certain conditions, especially the absolute integrability condition, i.e. $\int_{R^d} |v(x)| \,\mathrm{d}x$ exists and is finite. One then can show that

$$v(x) = \int_{\mathbb{R}^d} a(\omega)\cos\langle\omega, x\rangle + b(\omega)\sin\langle\omega, x\rangle \;\mathrm{d}\omega$$

with

$$\begin{aligned} a(\omega) &:= (2\pi)^{-d}\int_{\mathbb{R}^d} v(x)\cos(\omega x)\,\mathrm{d}x, \\ b(\omega) &:= (2\pi)^{-d}\int_{\mathbb{R}^d} v(x)\sin(\omega x)\,\mathrm{d}x. \end{aligned}$$

That can be expressed more elegantly if we go to the complex plane, more precisely we consider functions $v : \mathbb{R}^d \to \mathbb{C}$. So far, only the basic idea has been shown. In order to introduce the Fourier transformation in an accurate way we need to define the space of all test functions which consists of functions where the function values as well as all derivatives tend strongly to zero as we go to infinity. Base on [7, Subsection 10.4.6]

Definition A.1.1 (Space of test functions)
With $u : \mathbb{R}^d \to \mathbb{C}$, $u \in \mathrm{C}^\infty(\mathbb{R}^d)$ we define a system of half norms

$$\rho_{k,r}(u) := \sup_{x\in\mathbb{R}^d}\left((\|x\|^r + 1) \sum_{\|\alpha\|\le k} \|\mathbf{D}^\alpha u(x)\| \right).$$

The space of test functions is the subset of infinite often differentiable functions where $\rho_{k,r}$ is finite, i.e.

$$\mathscr{S}(\mathbb{R}^d) := \left\{u \in \mathrm{C}^\infty(\mathbb{R}^d) : \rho_{k,r}(u) < \infty\; \forall k, r > 0\right\}$$

The strict requirement to be a test function guarantees the existence of the integral in the Fourier transform.

Definition A.1.2 (Fourier transformation)
For $v \in \mathscr{S}(\mathbb{R}^d)$ the Fourier transformation $\mathscr{F} : \mathscr{S}(\mathbb{R}^d) \to \mathscr{S}(\mathbb{R}^d)$ is defined as $\mathscr{F}(v) := \tilde{v}$ with

$$\tilde{v}(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} v(x)\, \mathrm{e}^{-i\langle \omega, x\rangle}\ dx. \tag{A.1}$$

The function $\tilde{v}$ is called the Fourier transformed of v. Formally, we define the inverse by $\mathscr{I}(\tilde{v}) := v$ with

$$v(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \tilde{v}(\omega)\, \mathrm{e}^{i\langle \omega, x\rangle}\ d\omega. \tag{A.2}$$

As it turns out the formally defined inverse transformation $\mathscr{I}$ is the inverse of $\mathscr{F}$.

Theorem A.1.3 (Fourier transformation)
The Fourier transformation is well defined, i.e. the integral exists and the operator $\mathscr{F} : \mathscr{S}(\mathbb{R}^d) \to \mathscr{S}(\mathbb{R}^d)$ is bijective and it is $\mathscr{F}^{-1} = \mathscr{I}$. Furthermore, $\mathscr{F}$ as well as $\mathscr{F}^{-1}$ are continuous. The operator $\mathscr{F}$ is uniquely extendable to the Hilbert space $\mathrm{L}_2(\mathbb{R}^d, \mathbb{C})$.

One of its most valuable properties is that differentiation in the real space translates to multiplication with iω in the Fourier transformed space.

Lemma A.1.4
The following properties apply to the Fourier transformation $\mathscr{F}$ for all $u, v \in \mathscr{S}(\mathbb{R}^d)$.

Parseval's relation *With the $\mathrm{L}_2(\mathbb{R}^d, \mathbb{C})$ scalar product it is*

$$\langle \mathscr{F}(u), \mathscr{F}(v)\rangle = \langle u, v\rangle\,.$$

Differentiation *With the multi-index $\alpha \in \mathbb{N}_0^d$ it applies*

$$\begin{aligned} \mathscr{F}(\mathbf{D}^\alpha v)(\omega) &= (i\omega)^\alpha \mathscr{F}(v)(\omega), \\ \mathscr{F}(x^\alpha v)(\omega) &= (-i)^{|\alpha|} \mathbf{D}^\alpha\big(\mathscr{F}(v)\big)(\omega). \end{aligned} \tag{A.3}$$

Convolution *The convolution of u and v is defined by $(u * v)(x) := \int_{\mathbb{R}^d} u(y)v(x-y)\ dy$. Then it is*

$$\mathscr{F}(u * v) = \mathscr{F}(u)\mathscr{F}(v).$$

A.2 Approximation of derivatives using polynomial interpolation

An alternative approach to obtain an approximation to derivatives is to use polynomial interpolation of three adjacent points and to determine the derivative of that polynomial. As it turns out we obtain the same result as with the approach of using Taylor expansion shown in subsection 4.2.1.

A.2.1 One dimensional functions

We use the trial function $g(x) := c_2(x - x_k)^2 + c_1(x - x_k) + c_0$ to approximate the function f around x_k. It has to be satisfied that $g(x_{k+i}) = f(x_{k+i})$, $\forall i \in \{-1, 0, 1\}$ which restricts the choice of c_0, c_1 and c_2 by the system of equations

$$\begin{pmatrix} 1 & -\Delta x_1 & {\Delta x_1}^2 \\ 1 & 0 & 0 \\ 1 & \Delta x_2 & {\Delta x_2}^2 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} f_{k-1} \\ f_k \\ f_{k+1} \end{pmatrix}.$$

By inverting this matrix we obtain the solution for the coefficients

$$\begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{-\Delta x_2}{\Delta x_1(\Delta x_1 + \Delta x_2)} & \frac{\Delta x_2 - \Delta x_1}{\Delta x_1 \Delta x_2} & \frac{\Delta x_1}{\Delta x_2(\Delta x_1 + \Delta x_2)} \\ \frac{1}{\Delta x_1(\Delta x_1 + \Delta x_2)} & \frac{-1}{\Delta x_1 \Delta x_2} & \frac{1}{\Delta x_2(\Delta x_1 + \Delta x_2)} \end{pmatrix} \begin{pmatrix} f_{k-1} \\ f_k \\ f_{k+1} \end{pmatrix}.$$

Furthermore we have

$$\begin{aligned} g(x_k) &= c_0 \\ g'(x_k) &= c_1 \\ g''(x_k) &= 2c_2 \end{aligned}$$

and thus

$$\begin{pmatrix} g(x_k) \\ g'(x_k) \\ g''(x_k) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{-\Delta x_2}{\Delta x_1(\Delta x_1+\Delta x_2)} & \frac{\Delta x_2-\Delta x_1}{\Delta x_1 \Delta x_2} & \frac{\Delta x_1}{\Delta x_2(\Delta x_1+\Delta x_2)} \\ \frac{2}{\Delta x_1(\Delta x_1+\Delta x_2)} & \frac{-2}{\Delta x_1 \Delta x_2} & \frac{2}{\Delta x_2(\Delta x_1+\Delta x_2)} \end{pmatrix} \begin{pmatrix} f_{k-1} \\ f_k \\ f_{k+1} \end{pmatrix}.$$

This is exactly the same result already obtained using Taylor series expansion.

A.2.2 Two dimensional functions

Interpolation over nine points in two dimensions requires multi-quadratic trial functions

$$g(x,y) := \sum_{i,j=0}^{2} a_{i,j}(x-x_k)^i(x-x_l)^j$$

With the abbreviation $f_{i,j} := f(x_{k+i,l+j})$ the interpolation requirements $g(x_{k+i}, y_{l+j}) = f_{i,j}$ is equivalent to the linear equation system

$$\begin{pmatrix} 1 & -\Delta y_1 & {\Delta y_1}^2 & -\Delta x_1 & \Delta x_1 \Delta y_1 & -\Delta x_1 {\Delta y_1}^2 & {\Delta x_1}^2 & -{\Delta x_1}^2 \Delta y_1 & {\Delta x_1}^2 {\Delta y_1}^2 \\ 1 & 0 & 0 & -\Delta x_1 & 0 & 0 & {\Delta x_1}^2 & 0 & 0 \\ 1 & \Delta y_2 & {\Delta y_2}^2 & -\Delta x_1 & -\Delta x_1 \Delta y_2 & -\Delta x_1 {\Delta y_2}^2 & {\Delta x_1}^2 & {\Delta x_1}^2 \Delta y_2 & {\Delta x_1}^2 {\Delta y_2}^2 \\ 1 & -\Delta y_1 & {\Delta y_1}^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & \Delta y_2 & {\Delta y_2}^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -\Delta y_1 & {\Delta y_1}^2 & \Delta x_2 & -\Delta x_2 \Delta y_1 & \Delta x_2 {\Delta y_1}^2 & {\Delta x_2}^2 & -{\Delta x_2}^2 \Delta y_1 & {\Delta x_2}^2 {\Delta y_1}^2 \\ 1 & 0 & 0 & \Delta x_2 & 0 & 0 & {\Delta x_2}^2 & 0 & 0 \\ 1 & \Delta y_2 & {\Delta y_2}^2 & \Delta x_2 & \Delta x_2 \Delta y_2 & \Delta x_2 {\Delta y_2}^2 & {\Delta x_2}^2 & {\Delta x_2}^2 \Delta y_2 & {\Delta x_2}^2 {\Delta y_2}^2 \end{pmatrix} \begin{pmatrix} a_{0,0} \\ a_{0,1} \\ a_{0,2} \\ a_{1,0} \\ a_{1,1} \\ a_{1,2} \\ a_{2,0} \\ a_{2,1} \\ a_{2,2} \end{pmatrix} = \begin{pmatrix} f_{-1,-1} \\ f_{-1,0} \\ f_{-1,1} \\ f_{0,-1} \\ f_{0,0} \\ f_{0,1} \\ f_{1,-1} \\ f_{1,0} \\ f_{1,1} \end{pmatrix}.$$

We are only interested in the coefficients $a_{1,0}$, $a_{0,1}$, $a_{2,0}$, $a_{0,2}$ and $a_{1,1}$ since

$$\frac{\partial g}{\partial x}(x_k, y_l) = a_{1,0} \qquad \frac{\partial g}{\partial y}(x_k, y_l) = a_{0,1},$$

$$\frac{\partial^2 g}{\partial x^2}(x_k, y_l) = 2a_{2,0} \qquad \frac{\partial^2 g}{\partial y^2}(x_k, y_l) = 2a_{0,2},$$

$$\frac{\partial^2 g}{\partial x \partial y}(x_k, y_l) = a_{1,1}.$$

After inverting the Matrix and determining the coefficients we see that the result is identical to that we obtained by the Taylor approximation approach. This is not very astonishing since, up to some permutation of columns and a transposition of the matrix, both matrices are the same.

Appendix B

List of symbols

In general for $x \in \mathbb{R}$ we denote with x_i the i-th point. For $x \in \mathbb{R}^d$, however, subscripts are reserved to indicate the i-th component of the vector x. Therefore $x^{(k)}$ with a multi-index k denotes the k-th point. If there is no bracket superscripts always indicate the i-th power.

Throughout the text, the time variable t is always considered as a distinct variable which means that all operators like Laplace, the gradient or the Fourier transformation operate only in the space variables.

o.d.e.	abbreviation for ordinary differential equation
p.d.e.	abbreviation for partial differential equation
f.d.m.	abbreviation for finite difference method
$\mathbb{R}_+$	positive real numbers $(0, \infty)$
$\mathbb{N}_0$	non negative natural numbers $\{0, 1, 2, \ldots\}$
$\langle \cdot, \cdot \rangle$	scalar product
$\mathscr{F}(u)$	Fourier transformation
$\triangle u$	Laplace operator
Δx	delta x, $\Delta x := x_{k+1} - x_k$
I	identity matrix
$\Omega \subset \mathbb{R}^d$	region of the $\mathbb{R}^d$ where the p.d.e. is defined
$\bar{\Omega}$	closure of the set Ω
$\Gamma = \partial\Omega$	boundary of Ω
$\mathbf{D}^\alpha$	derivative operator
$\mathrm{C}^p(\bar{\Omega})$	space of functions defined over Ω which are p-times continuously differentiable
u	mainly used in the context of the solution of a p.d.e.
h	space discretisation parameter, $h > 0$
τ	time discretisation parameter, $\tau > 0$
I_h	index set of grid points
$\bar{\Omega}_h$	finite grid approximating $\bar{\Omega}$
Γ_h	boundary points of the grid $\bar{\Omega}_h$
Ω_h	inner points of the grid $\bar{\Omega}_h$
f_h	grid function defined over $\bar{\Omega}_h$ approximating a continuous function f
Φ_h	space of all grid function defined on $\bar{\Omega}_h$, can be identified with the $\mathbb{R}^n$ where $n \in \mathbb{N}$ is the number of grid points in $\bar{\Omega}_h$
$v_k = v(x^{(k)})$	for grid functions $v \in \Phi_h$ the multi index k indicates the value of v at the k-th grid point
$\bar{u} = P_h u$	projection of a continuous function to the space of grid functions
$Q_h v$	interpolation of a grid function $v \in \Phi_h$ to the continuous space.
$\hat{u}^{(k)}$	solution of the finite difference scheme at time step k
∂	forward finite difference operator defined on Φ_h
$\bar{\partial}$	backward finite difference operator defined on Φ_h
$\hat{\partial}$	centred finite difference operator defined on Φ_h

Bibliography

[1] G. Blacher. Applying stochastic volatility models for pricing and hedging derivatives. Risk Training presentation, 2 2002.

[2] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–654, 1973.

[3] J. Douglas and T.M. Gallie. Variable time steps in the solution of the heat flow equation by a difference equation. *Proc. Amer. Math. Soc.*, 6:787–793, 1955.

[4] J. Douglas and J.E. Gunn. A general formulation of alternating direction methods. *Numer. Math.*, 6:428–453, 1964.

[5] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations.* Cambridge University Press, 1996.

[6] O. Faulhaber. Analytic methods for pricing double barrier options in the presence od stochastic volatility. Master's thesis, University of Kaiserslautern, Germany, 8 2002.

[7] G. Grosche, V. Ziegler, D. Ziegler, and E. Zeidler. *Teubner-Taschenbuch der Mathematik, Teil II.* Teubner, Stuttgart, Leipzig, Germany, 1995.

[8] J. Hakala and U. Wystup. *Foreign Exchange Risk: Models, Insturments and Strategies.* Risk Books, 2002.

[9] L. Halpern and J. Rauch. Absorbing boundary conditions for diffusion equations. *Numer. Math.*, 71:185–224, 1995.

[10] S. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2), 1973.

[11] J.D. Hoffman. Relationship between the truncation errors of centered finite difference approximations on uniform and nonuniform meshes. *J. Comput. Phys.*, 46:469–474, 1982.

[12] J. Jost. *Partielle Differentialgleichungen.* Springer, 1998.

[13] P.D. Lax and R.D. Richtmyer. Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, 9:267–293, 1956.

[14] A. Lipton. *Mathematical methods for foreign exchange.* World Scientific, 2001.

[15] J.P. Lohéac. An artificial boundary condition for an advection-diffusion equation. *Math. Meth. Appl. Sci.*, 14:155–175, 1991.

[16] G. Marchuk. Splitting and alternating direction methods. In J. Lions P. Ciarlet, editor, *Handbook of Numerical Analysis, Vol. I: Finite Difference Methods (Part 1)*, pages 197–462. Elsevier Science Publishers, 1989.

[17] S. McKee, D.P. Wall, and S.K. Wilson. An alternating direction implicit scheme for parabolic equations with mixed derivative and convective terms. *J. Comput. Phys.*, 126:64–76, 1996.

[18] K. W. Morton. *Numerical Solution of Convection-Diffusion Problems.* Chapman & Hall, 1996.

[19] S.A. Nazarov and M. Specovius-Neugebauer. Approximation of exterior problems. optimal conditions for the Laplacian. *Analysis*, 16:305–324, 1996.

[20] L.T. Nielsen. *Pricing and Hedging of Derivative Securities.* Oxford University Press, 1999.

[21] L.C.G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales – Volume one: Foundations.* Wiley, 1994.

[22] L.C.G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales – Volume two: Itô Calculus.* Cambridge University Press, Cambridge, 2000.

[23] H. Roos, M. Stynes, and L. Tobiska. *Numerical Methods for Singularly Perturbed Differential Equations, Convection-Diffusion and Flow Problems.* Springer, 1996.

[24] A. Samarskij. *Theorie der Differenzenverfahren.* Nauka, 1982.

[25] R. Smith. Optimal and near-optimal advection-diffusion finite-difference schemes. i. constant coefficient in one dimension. *Proc. Roy. Soc. Lond.*, 455(1987):2371–2387, 1999.

[26] R. Smith. Optimal and near-optimal advection-diffusion finite-difference schemes. ii. unsteadiness and non-uniform grid. *Proc. Roy. Soc. Lond.*, 456(1995):489–502, 2000.

[27] R. Smith. Optimal and near-optimal advection-diffusion finite-difference schemes. iii. black-scholes equation. *Proc. Roy. Soc. Lond.*, 456(1997):1019–1028, 2000.

[28] R. Smith. Optimal and near-optimal advection-diffusion finite-difference schemes. iv. spatial non-uniformity. *Proc. Roy. Soc. Lond.*, 457(2005):45–65, 2001.

[29] R. Smith and Y. Tang. Optimal and near-optimal advection-diffusion finite-difference schemes. vi. two-dimensional alternating directions. *Proc. Roy. Soc. Lond.*, 456(2014):2379–2396, 2001.

[30] W. F. Spotz and G. F. Carey. Extension of high-order compact schemes to time-dependent problems. *Numer. Methods Partial Differential Equations*, 17(6):657–672, 2001.

[31] J. C. Strikwerda. *Finite difference schemes and partial differential equations.* Wadsworth & Brooks, 1989.

[32] V. Thomée. Finite difference methods for linear parabolic equations. In J. Lions P. Ciarlet, editor, *Handbook of Numerical Analysis, Vol. I: Finite Difference Methods (Part 1)*, pages 5–196. Elsevier Science Publishers, 1989.

[33] H. Vogel. *Gerthsen Physik.* Springer, 1995.

[34] N.K. Yamaleev. Minimization of the truncation error by grid adaption. *J. Comput. Phys.*, 170:459–497, 2001.

[35] N.K. Yamaleev. Optimal two-dimensional finite difference grids providing superconvergence. *SIAM J. Sci. Comput.*, 23(5):1707–1730, 2002.

Erklärung

Ich erkläre an Eides Statt, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Chemnitz, August 4, 2009 Tino Kluge

Thesen

- Stochastische Volatilitätsmodelle ermöglichen marktnahe Optionspreisbewertung. Aus dem stochastischen Modellen können partielle Differentialgleichungen für die Optionspreisfunktion abgeleitet werden. Die Differentialgleichung ist von parabolischer Form und hat zwei Ortsrichtungen, die den momentanen Spot und die Varianz beschreiben (s, v). Die Lösung der Gleichung ist nur in einem bestimmten Punkt von Interesse.

- Das Finite Differenzenverfahren ist zur Lösung dieser parabolischen Differentialgleichung geeignet, da das Gebiet durch ein Rechtecksnetz approximiert werden kann.

- Die untersuchten Verfahren – Crank-Nicholson und Alternating Direction Implicit (ADI) – sind auf uniformen und nicht uniformen Gittern konsistent von der Ordnung zwei in Ort und in Zeitrichtung.

- Die Crank-Nicholson Methode ist stabil bei Dirichlet-Randbedingungen.

- Es kann ein nicht uniformes Netz angegeben werden, dass unter der Voraussetzung von glatten Anfangsbedingungen die Genauigkeit der Lösung in einem lokalen Punkt erheblich verbessert. Bei *Reverse Barrier* Optionen ist das nicht erfüllt. Dort sind weitere Untersuchungen nötig, um ein passendes Gitternetz zu ermitteln.

- Die Wahl eines nicht uniformen Netzes in Zeitrichtung ist empfehlenswert und besonders wichtig für *Reverse Barrier* Optionen. Am Beginn $(t = 0)$ sollte mit feinen Zeitschritten gerechnet werden.

- Die Modellierung der Randbedingung an dem linken Rand in Varianzrichtung $(v = 0)$ hat entscheidende Auswirkung auf die Genauigkeit der Lösung in dem Punkt, an dem wir interessiert sind. Die Diskretisierung der Differentialgleichung erwieß sich als bestmögliche Variante.

- Die anderen Ränder genügen entweder Dirichlet Bedingungen oder müssen künstlich gesetzt werden um den unendlichen Bereich zu approximieren. Es kann angegeben werden, wie weit die künstlichen Ränder entfernt sein müssen damit der Einfluss auf die Lösung im interessanten Punkt vernachlässigbar ist.

- Mit allen genannten Vorschlägen ist das Finite Differenzenverfahren praxistauglich für einfache pfadunabhängige Optionen. Für Call Optionen erreicht man Genauigkeiten von 0.1% innerhalb einer Sekunde.

- Für das Crank-Nicholson Verfahren wird ein direkter Gleichungssystem-Löser empfohlen.

- ADI ist schneller als das Crank-Nicholson Verfahren bei gleicher Genauigkeit.