Pricing Equity Derivatives under Stochastic Volatility : A Partial Differential Equation Approach

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa, in fulfillment of the requirements of the degree of Master of Science.

Roelof Sheppard

August 30, 2007

Declaration

I declare that this is my own, unaided work. It is being submitted for the Degree of Master of Science to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.

(Signature)

(Date)

Acknowledgements

I would like to thank my supervisor Dr Graeme West for proposing the topic of this thesis and for his guidance and support.

I would also like to thank my family and friends for their support during my studies. A special thanks to my mother, Elsabé, without whom my academic achievements would not have been possible.

Roelof Sheppard
February 2007

"On the ostensible exactitude of certain branches of human knowledge, including mathematics: The exactness is a fake."

- ALFRED WHITEHEAD (1861 - 1947)

Contents

Chapter 1

Introduction

In the Black and Scholes model, Black and Scholes [1973], geometric Brownian motion with constant volatility is assumed for the underlying process, that is

$$dS_t/S_t = rdt + \sigma d\widetilde{W}_{S_t}$$

where r the constant risk-free rate, S_t the stock and σ the constant volatility of the stock. Under these assumptions closed form solutions for the values of European call and put options are derived. In practise the assumption of constant volatility is not reasonable, since we require different values for the volatility parameter for different strikes and different expiries to match market prices. The volatility parameter that is required in the Black-Scholes formula to reproduce market prices is called the implied volatility. This is a critical internal inconsistency, since the implied volatility of the underlying should not be dependent on the specifications of the contract. Thus to obtain market prices of options maturing at a certain date, volatility needs to be a function of the strike. This function is the so called volatility skew or smile. Furthermore for a fixed strike we also need different volatility parameters to match the market prices of options maturing on different dates written on the same underlying, hence volatility is a function of both the strike and the expiry date of the derivative security. This bivariate function is called the volatility surface. There are two prominent ways of working around this problem, namely, local volatility models and stochastic volatility models.

For local volatility models the assumption of constant volatility made in Black and Scholes [1973] is relaxed. The underlying risk-neutral stochastic process becomes

$$dS_t/S_t = r(t)dt + \sigma(S_t, t)d\widetilde{W}_{S_t}$$

where $r(t)$ is the instantaneous forward rate of maturity t implied by the yield curve and the function $\sigma(S_t, t)$ is chosen (calibrated) such that the model is consistent with market data, see Dupire [1994], Derman and Kani [1994] and [Wilmott, 2000, §25.6]. It is claimed in Hagan et al. [2002] that local volatility models predict that the smile shifts to higher prices (resp. lower prices) when the price of the underlying decreases (resp. increases). This is in contrast to the market behavior where the smile shifts to higher prices (resp. lower prices) when the price of the underlying increases (resp. decreases).

Another way of working around the inconsistency introduced by constant volatility is by introducing a stochastic process for the volatility itself; such models are called stochastic volatility models. The major advances in stochastic volatility models are Hull and White [1987], Heston [1993] and Hagan et al. [2002]. Such models have the following general form

$$\begin{aligned} dS_t &= p_S(S_t, \sigma_t, t)dt + q_S(S_t, \sigma_t, t)d\widetilde{W}_{S_t} \\ d\sigma_t &= p_\sigma(S_t, \sigma_t, t)dt + q_\sigma(S_t, \sigma_t, t)d\widetilde{W}_{\sigma_t} \end{aligned}$$

where the tradeable security S_t and its volatility σ_t are correlated, i.e. $d\widetilde{W}_{S_t} d\widetilde{W}_{\sigma_t} = \rho dt$ and the functional form of p_S, q_S, p_σ and q_σ are determined by the model used. The volatility process is no longer constant as in the Black-Scholes model nor deterministic as in the local volatility models, but is now subject to its own random process. In the Black-Scholes world the implied volatility is the only calibration parameter that needs to be determined by information in the market. In each of the above mentioned stochastic volatility models there is more than one unknown parameter in the processes involved. These parameters are solved by using the same backward reasoning as in the Black-Scholes model. First closed form solutions are derived for vanilla type options. These liquidly traded options are then used to determine the unknown parameters such that the total error between the theoretical and observed prices is minimized. Once the model is calibrated, i.e. the model returns market prices for vanilla options, we can price more exotic type options with the model.

In chapter 2 we derive the well known result that the price of a contingent claim in a stochastic volatility model can be represented as the solution of a two dimensional convection diffusion partial differential equation (PDE), with the initial condition given by the payoff function. In this chapter we will derive the PDEs relevant to the pricing of options in these major stochastic volatility models. In this thesis we will focus on a specific class of numerical procedures to obtain accurate approximations to the solution of the relevant PDE, called the finite difference method (FDM). Alternatively one can use Monte Carlo integration to obtain the values of derivatives in a stochastic volatility environment, see Fouque and Tullie [2002] and Kahl and Jäckel [2006].

We introduce the reader to the different concepts of the finite difference method by applying the numerical procedure to a one dimensional PDE in chapter 3. In this chapter we apply the θ-method to a one dimensional convection diffusion equation as well as a one dimensional convection equation. We define and prove the consistency and stability of these schemes. For each of these problems we investigate the stability of the θ-method by making use of von Neumann stability analysis as well as a matrix method of analysis (under the maximum norm). It is noted in chapter 3 that, strictly speaking, von Neumann stability analysis is not applicable to problems with variable coefficients or problems with non-smooth initial data. We also investigate a procedure called exponential fitting, introduced in Duffy [2006]. This procedure is used to improve the stability properties of the θ-method when it is applied to a one dimensional convection diffusion equation. We show how exponential fitting can be used to obtain schemes that are stable under the maximum norm.

It is well known that the second order accurate Crank-Nicolson scheme is only von Neumann stable and thus not able to handle non-smooth initial data, see Duffy [2003] and Giles and Carter [2006] for example[1]. In Gourlay and Morris [1980] extrapolation methods are applied to the one dimensional heat equation with homogenous Dirichlet boundary conditions, to obtain schemes that are second, third and fourth order accurate in time. A huge advantage of the extrapolation schemes derived in Gourlay and Morris [1980] is that these schemes are L_0-stable, meaning that these schemes are able to handle discontinuous initial data. In chapter 4 we extend these ideas to convection diffusion problems with non-zero Dirichlet boundary conditions. We also obtain a scheme that is L_0-stable and fifth order accurate in time.

In chapter 5 we extend the ideas of chapter 3 to two dimensions. We consider a generalization of the classical θ-method called the Implicit-Explicit method (IMEX-method) where the implicitness/explicitness of the convection, diffusion and mixed-derivative part of the FDM can differ. Conditions under which the IMEX-scheme is stable and the proof of unconditional consistency of the IMEX-scheme can be found in this chapter. We show how exponential fitting can be used to make special cases of the IMEX-method stable under the maximum norm. IMEX-schemes require the inversion of large non tri-diagonal matrices, which can be very time consuming. There are two main FDMs that are used to work around this problem: Alternating-Direction-Implicit schemes (ADI-schemes) and Locally-One-Dimensional schemes (LOD-schemes). LOD-schemes are also referred to as splitting schemes. With these schemes the original problem is rewritten as a sequence of simpler problems, each one of the simpler problems can be solved with a tri-diagonal solver. In chapter 5 we investigate the stability and consistency of a specific LOD-scheme called the Yanenko method, Yanenko [1971]. The Yanenko method resolves the problems ADI-methods show for parabolic PDE with mixed derivative terms, see Duffy [2006]. We motivate boundary conditions for the Yanenko scheme that retain the tri-diagonal property of the matrices and the stability of the scheme. For all methods in this chapter stability is investigated with von Neumann stability analysis and a matrix method of analysis (under the maximum norm).

In chapter 6 we extend the ideas of chapter 4 to two dimensions. In Khaliq and Twizell [1986] third and fourth order L_0-stable extrapolation schemes are developed for the simple two dimensional heat equation with homogenous boundary conditions. We show how the schemes developed in Khaliq and Twizell [1986] can be extended for two dimensional convection diffusion problems with a mixed derivative term.

In chapter 7 we discuss how non-uniform grids can be applied to stochastic volatility PDEs to improve the local order of convergence, as proposed in Kluge [2002]. We also show how exponential fitting can be used to improve the stability of FDMs on non-uniform grids. We derive transformations that removes the cross derivative term from the Heston PDE, this is in contrast to Zvan et al. [2003] where it is said that such transformations do not appear to be possible. Finally we give a précis of known remedies for the problems that arise when FDMs are applied to problems with non-smooth payoff functions.

In chapter 8 we show how the general two dimensional PDE solvers developed in chapters 5, 6 and 7 can be applied to these major stochastic volatility PDEs.

In chapter 9 of this thesis we combine extrapolation, exponential fitting and non-uniform grids to obtain a robust two dimensional PDE solver. In chapter 9 we compare the pricing formulae of vanilla European

[1] By non-smooth we mean that either the initial condition or the derivative of the initial condition is discontinuous. Almost all initial conditions, resulting from pricing problems, are non-smooth.

options in the SABR world, derived in Hagan et al. [2002], to the solutions obtained with the FDMs discussed in this thesis and confirm that the SABR formulae can give bad approximations of the true solution under certain parameter sets. Hence if calibration of the SABR model results in such a *bad* parameter set, it would be inconsistent to value more exotic options with some numerical procedure such as Monte Carlo or the FDM. We also compare the numerical solutions given by our FDMs to the semi-analytical pricing formulae of vanilla European options within the Heston model given in Vogt [2004].

Chapter 2

The PDE for Stochastic Volatility Models

In this chapter the PDE that needs to be solved to obtain the value surface of contingent claims in a general stochastic volatility environment is derived. In the first section we derive the PDE that the price of a derivative must solve, where the tradeable security as well as the volatility of the tradeable security follows general stochastic processes. In the second section we show that all the major stochastic volatility models are simplifications of the general model discussed in the first section.

2.1 PDE for general stochastic volatility processes

In this section the PDE that governs the prices of derivatives written on a tradeable security with stochastic volatility is derived. This derivation is based a derivation done in [Lewis, 2000, §1.4], a similar derivation can be found in Wilmott [2000]. The relevant processes are

$$dS_t = (p_S(S_t, \sigma_t, t) - D_t)dt + q_S(S_t, \sigma_t, t)d\widetilde{W}_{S_t} \tag{2.1}$$

$$d\sigma_t = p_\sigma(S_t, \sigma_t, t)dt + q_\sigma(S_t, \sigma_t, t)d\widetilde{W}_{\sigma_t} \tag{2.2}$$

where the tradeable security S_t and its volatility σ_t are correlated, i.e. $d\widetilde{W}_{S_t}d\widetilde{W}_{\sigma_t} = \rho dt$ [1]. From the fact that $S_t \geq 0$ for all t it follows that the dividend rate $D_t = D(S_t, t)$ must be smaller than $p_S(S_t, \sigma_t, t)$ at $S_t = 0$. This general stochastic volatility model reduces to the Black-Scholes model when $p_S = \mu S_t$, $q_S = \sigma_0 S_t$, $p_\sigma \equiv 0$ and $q_\sigma \equiv 0$ where σ_0 is a constant. Furthermore it is assumed that the price of a contingent claim is a function of S_t, σ_t and t, i.e. $V = V(S_t, \sigma_t, t)$. Thus in this setting we exclude path dependent options.

Using the processes above and a generalization of the hedging arguments given in Black and Scholes [1973] we will derive the PDE that the value of any contingent claim on a tradeable security must solve.

[1] Notation : For the rest for this chapter $\tilde{W}$ will denote the real world Wiener process while will W denote the process under the risk-neutral measure.

The first step is to construct a portfolio, containing the derivative, that is instantaneously riskless. The random volatility introduces an extra source of randomness which renders the traditional Black-Scholes world incomplete. To hedge this extra source of randomness we need another liquid derivative written on the same tradeable security with a different maturity. Let the replicating portfolio consist of the derivative that we want to hedge (V_t), $-\Delta$ shares of the stock (S_t) and $-\Delta_1$ shares of the other liquid contingent claim ($\widetilde{V}_t$), where Δ and Δ_1 are random variables. Denote our portfolio value by Π_t such that

$$\Pi_t = V_t - \Delta S_t - \Delta_1 \widetilde{V}_t \tag{2.3}$$

where $\Pi_t = \Pi(S_t, \sigma_t, t)$, $V_t = V(S_t, \sigma_t, t)$ and $\widetilde{V}_t = \widetilde{V}(S_t, \sigma_t, t)$. The replicating portfolio must be self-financing, which means that there is no additional cash inflow or outflow beyond the initial deposit Π_0. To derive the *self-financing condition* we are going to consider the dynamics of the replicating portfolio in discrete time $t, t + \Delta t, \ldots$, and then take the continuous time limit, $\Delta t \to dt$. It is assumed that events occur in the following order in this market:

- just prior to a possible dividend, the stock has value S_t and the replicating portfolio has a value Π_t,
- the stock pays a dividend D_t per share and the stock price drops to its ex-dividend value, $S_t^+ = S_t - D_t \Delta t$,
- A monetary value of Π_t gets invested as follows: one share of the derivative that needs to be valued, $-\Delta$ shares in the stock and $-\Delta_1$ shares of another liquidly traded contingent claim.

This results in the following equation for the first discrete time period

$$\begin{aligned}\Pi_t &= V_t - \Delta S_t^+ - \Delta_1 \widetilde{V}_t \\ &= V_t - \Delta S_t + \Delta D_t \Delta t - \Delta_1 \widetilde{V}_t \end{aligned} \tag{2.4}$$

and for the second discrete time period, $t + \Delta t$, it is easy to see that

$$\Pi_{t+\Delta t} = V_{t+\Delta t} - \Delta S_{t+\Delta t} - \Delta_1 \widetilde{V}_{t+\Delta t}.$$

Subtracting (2.4) from the equation above yields

$$\Pi_{t+\Delta t} - \Pi_t = V_{t+\Delta t} - V_t - \Delta(S_{t+\Delta t} - S_t) - \Delta D_t \Delta t - \Delta_1(\widetilde{V}_{t+\Delta t} - \widetilde{V}_t).$$

Taking the continues time limit, $\Delta t \to dt$, the instantaneous change in the value of the portfolio becomes

$$d\Pi_t = dV_t - \Delta(dS_t + D_t dt) - \Delta_1 d\widetilde{V}_t \tag{2.5}$$

where dS_t is known from (2.1). The value processes dV_t and $d\widetilde{V}_t$ can be obtained by making use of the multidimensional Itô's formula,

Proposition 2.1.1. *(Itô's formula, Björk [1998]) Take a vector Wiener process $W = (W_1, \ldots, W_n)$ with correlation matrix ρ, and assume that the vector process $X = (X_1, \ldots, X_k)^T$ has a stochastic differential. Then the following hold:*

- *For any $C^{1,\,2}$ function f, the stochastic differential process $f(t, X_t)$ is given by*

$$df(t, X_t) = \frac{\partial f}{\partial t}dt + \sum_{i\,=\,1}^{n} \frac{\partial f}{\partial x_i}dX_i + \tfrac{1}{2} \sum_{i,\,j\,=\,1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j} dX_i dX_j,$$

  with the formal multiplication table

$$\begin{aligned}(dt)^2 &= 0,\\ dt\, dW_i &= 0, \quad i = 1,\, 2,\, \ldots,\, n,\\ dW_i\, dW_j &= \rho_{i,\,j}dt.\end{aligned}$$

- *If, in particular, $k = n$ and dX has the structure*

$$dX_i = \mu_i dt + \sigma_i dW_i, \quad i = 1,\, 2,\, \ldots,\, n,$$

  where $\mu_1, \ldots, \mu_n$ and $\sigma_1, \ldots, \sigma_n$ are scalar processes, then the stochastic differential of the process $f(t, X_t)$ is given by

$$df = \left[\frac{\partial f}{\partial t} + \sum_{i\,=\,1}^{n} \mu_i \frac{\partial f}{\partial x_i} + \tfrac{1}{2} \sum_{i,\,j\,=\,1}^{n} \sigma_i \sigma_j \rho_{i,\,j} \frac{\partial^2 f}{\partial x_i \partial x_j}\right] dt + \sum_{i\,=\,1}^{n} \sigma_i \frac{\partial f}{\partial x_i} dW_i.$$

The derivation can be made compact by defining the following operator

$$\mathcal{A}V(S, \sigma, t) = (p_S - D)\frac{\partial V}{\partial S} + p_\sigma \frac{\partial V}{\partial \sigma} + \tfrac{1}{2}q_S^2 \frac{\partial^2 V}{\partial S^2} + \rho q_S q_\sigma \frac{\partial^2 V}{\partial S \partial \sigma} + \tfrac{1}{2}q_\sigma^2 \frac{\partial^2 V}{\partial \sigma^2}.$$

The application of Itô's formula to $V = V(S, \sigma, t)$ and $\widetilde{V} = \widetilde{V}(S, \sigma, t)$ results in

$$dV = \left(\frac{\partial V}{\partial t} + \mathcal{A}V\right) dt + q_S \frac{\partial V}{\partial S} d\widetilde{W}_S + q_\sigma \frac{\partial V}{\partial \sigma} d\widetilde{W}_\sigma$$

and

$$d\widetilde{V} = \left(\frac{\partial \widetilde{V}}{\partial t} + \mathcal{A}\widetilde{V}\right) dt + q_S \frac{\partial \widetilde{V}}{\partial S} d\widetilde{W}_S + q_\sigma \frac{\partial \widetilde{V}}{\partial \sigma} d\widetilde{W}_\sigma$$

respectively. Substituting the equations derived above together with (2.1) in the self financing condition, (2.5), results in

$$\begin{aligned} d\Pi &= dV - \Delta(dS + Ddt) - \Delta_1 d\widetilde{V}\\ &= \left[\frac{\partial V}{\partial t} + \mathcal{A}V - \Delta_1\left(\frac{\partial \widetilde{V}}{\partial t} + \mathcal{A}\widetilde{V}\right) - \Delta p_S\right] dt\\ &\quad + \left[q_S \frac{\partial V}{\partial S} - \Delta_1 q_S \frac{\partial \widetilde{V}}{\partial S} - \Delta q_S\right] d\widetilde{W}_S + \left[q_\sigma \frac{\partial V}{\partial \sigma} - \Delta_1 q_\sigma \frac{\partial \widetilde{V}}{\partial \sigma}\right] d\widetilde{W}_\sigma \end{aligned} \tag{2.6}$$

To eliminate all the noise in the portfolio process the coefficients of $d\widetilde{W}_S$ and $d\widetilde{W}_\sigma$ need to be set to zero. This can be achieved by solving a trivial system of two equations in two unknowns, Δ and Δ_1. The unique solution follows

$$\Delta_1 = \frac{\partial V}{\partial \sigma} \Big/ \frac{\partial \widetilde{V}}{\partial \sigma}, \tag{2.7}$$

$$\Delta = \frac{\partial V}{\partial S} - \Delta_1 \frac{\partial \widetilde{V}}{\partial S}. \tag{2.8}$$

Via this dynamic hedging procedure a risk-less portfolio has been obtained. From arbitrage arguments it follows that this portfolio must grow at the risk free rate

$$\begin{aligned} d\Pi &= r\Pi dt \\ &= r(V - \Delta S - \Delta_1 \widetilde{V})dt. \end{aligned} \tag{2.9}$$

Substituting (2.6), (2.7) and (2.8) in (2.9) results in

$$\begin{aligned} &\frac{\partial V}{\partial t} + \mathcal{A}V - \Delta_{1}\left(\frac{\partial \widetilde{V}}{\partial t} + \mathcal{A}\widetilde{V}\right) - \left(\frac{\partial V}{\partial S} - \Delta_1 \frac{\partial \widetilde{V}}{\partial S}\right) p_S \qquad (2.10) \\ &\quad = rV - rS\left(\frac{\partial V}{\partial S} - \Delta_1 \frac{\partial \widetilde{V}}{\partial S}\right) - r\Delta_1 \widetilde{V} \\ \Rightarrow &\frac{\partial V}{\partial t} + \mathcal{A}V - rV + rS\frac{\partial V}{\partial S} - p_S\frac{\partial V}{\partial S} \\ &\quad = \Delta_1\left[\frac{\partial \widetilde{V}}{\partial t} + \mathcal{A}\widetilde{V} - r\widetilde{V} + rS\frac{\partial \widetilde{V}}{\partial S} - p_S\frac{\partial \widetilde{V}}{\partial S}\right] \end{aligned}$$

which can be rearranged as follows

$$\frac{\left(\frac{\partial V}{\partial t} + \mathcal{A}V - rV + rS\frac{\partial V}{\partial S} - p_S\frac{\partial V}{\partial S}\right)}{\frac{\partial V}{\partial \sigma}} = \frac{\left(\frac{\partial \widetilde{V}}{\partial t} + \mathcal{A}\widetilde{V} - r\widetilde{V} + rS\frac{\partial \widetilde{V}}{\partial S} - p_S\frac{\partial \widetilde{V}}{\partial S}\right)}{\frac{\partial V_1}{\partial \sigma}}. \tag{2.11}$$

Since the left-hand side is independent of $\widetilde{V}$ and the right-hand side is independent of V, equation (2.11) must be equal to a function that is independent of the price of the derivative, hence

$$\frac{\left(\frac{\partial V}{\partial t} + \mathcal{A}V - rV + rS\frac{\partial V}{\partial S} - p_S\frac{\partial V}{\partial S}\right)}{\frac{\partial V}{\partial \sigma}} = q_\sigma \lambda_\sigma(F, \sigma, t)$$

which can be rearranged to give the familiar form

$$\begin{aligned} \frac{\partial V}{\partial t} &+ (rS - D)\frac{\partial V}{\partial S} + (p_\sigma - q_\sigma \lambda_\sigma(S, \sigma, t))\frac{\partial V}{\partial \sigma} \\ &+ \tfrac{1}{2}q_S^2\frac{\partial^2 V}{\partial S^2} + \rho q_S q_\sigma \frac{\partial^2 V}{\partial S \partial \sigma} + \tfrac{1}{2}q_\sigma^2\frac{\partial^2 V}{\partial \sigma^2} - rV = 0 \end{aligned} \tag{2.12}$$

The function $\lambda_\sigma(S, \sigma, t)$ is called the market price of volatility risk.

2.2 PDEs for the major stochastic volatility models

2.2.1 SABR model

Non-dynamic SABR model

In Hagan et al. [2002] closed form solutions for the values of European call and put options, written on a forward value, are derived. The underlying processes are given by,

$$dF_t = \sigma F_t^\beta dW_{F_t} \tag{2.13}$$

$$d\sigma_t = \nu \sigma_t dW_{\sigma_t} \tag{2.14}$$

where the forward value F_t and a volatility like parameter σ_t are correlated, $dW_{F_t}dW_{\sigma_t} = \rho dt$ [2]. The absence of drift in the forward process indicates that these processes are in the risk-neutral world. To use the results from section 2.1 we consider these processes under the real-world measure,

$$dF_t = p_F(F_t, \sigma_t, t)dt + \sigma F_t^\beta d\widetilde{W}_{F_t} \tag{2.15}$$

$$d\sigma_t = p_\sigma(F_t, \sigma_t, t)dt + \nu\sigma_t d\widetilde{W}_{\sigma_t} \tag{2.16}$$

where $d\widetilde{W}_{F_t}d\widetilde{W}_{\sigma_t} = \rho dt$. In order to use the PDE derived in the previous section we need to make a change of variable such that the PDE is applicable to a forward value. When we make the assumption that the underlying stock has a dividend yield of q, i.e. $D = qS$, then we can use of the fact that the arbitrage value of a forward is given by $F = Se^{(r-q)(T-t)}$. By making use of Itô's formula and the fact that $\frac{\partial F}{\partial t} = -(r-q)F$ we can obtain the dynamics of a forward,

$$\begin{aligned} dF_t &= -(r-q)F_t dt + e^{(r-q)(T-t)}dS_t \\ &= [e^{(r-q)(T-t)}p_S - rF_t]dt + e^{(r-q)(T-t)}q_S d\widetilde{W}_{S_t} \end{aligned} \tag{2.17}$$

Since a forward contract is a tradeable security the results of section 2.1 can be applied to obtain the relevant PDE,

$$\begin{aligned} \frac{\partial V}{\partial t} &+ (p_\sigma - q_\sigma\lambda_\sigma(Fe^{-(r-q)(T-t)}, \sigma, t))\frac{\partial V}{\partial \sigma} \\ &+ \tfrac{1}{2}e^{2(r-q)(T-t)}q_S^2\frac{\partial^2 V}{\partial F^2} + \rho e^{(r-q)(T-t)}q_S q_\sigma\frac{\partial^2 V}{\partial F\partial\sigma} + \tfrac{1}{2}q_\sigma^2\frac{\partial^2 V}{\partial\sigma^2} - rV = 0 \end{aligned} \tag{2.18}$$

By comparing (2.15) and (2.16) with (2.17) and (2.2) respectively, we see that the following substitutions,

$$e^{(r-q)(T-t)}q_S = \sigma F^\beta$$
$$q_\sigma(F, \sigma, t) = \nu\sigma$$

must be made in (2.18) to obtain,

$$\begin{aligned} \frac{\partial V}{\partial t} &+ (p_\sigma - \nu\sigma\lambda_\sigma(Fe^{-(r-q)(T-t)}, \sigma, t))\frac{\partial V}{\partial \sigma} \\ &+ \tfrac{1}{2}\sigma^2F^{2\beta}\frac{\partial^2 V}{\partial F^2} + \rho\nu\sigma^2F^\beta\frac{\partial^2 V}{\partial F\partial\sigma} + \tfrac{1}{2}\nu^2\sigma^2\frac{\partial^2 V}{\partial\sigma^2} - rV = 0. \end{aligned} \tag{2.19}$$

A useful fact of the SABR model is that λ_σ does not appear in the derivation of the model. The reason for this is that the authors made an implicit assumption about the market price of volatility risk by choosing (2.14) as the risk-neutral process for the volatility[3]. To see this, consider a change of measure in (2.16), under some technical conditions on λ_σ we may set

$$d\widetilde{W}_{\sigma_t} = dW_{\sigma_t} - \lambda_\sigma dt.$$

Substituting the equation above in (2.16) results in

$$d\sigma = (p_\sigma - \lambda_\sigma\nu\sigma)dt + \nu\sigma dW_{\sigma_t}.$$

[2] The volatility parameter is not a Black-Scholes volatility. It is function of the at-the-money volatility and other calibration parameters of the SABR model.

[3] S. Afshani pointed this out to me in one of our "academic sessions".

By comparing the final result above to (2.14) we obtain the implicit assumption on the market price of volatility risk,

$$\lambda_\sigma = \frac{p_\sigma}{\nu\sigma}. \tag{2.20}$$

Substitution into (2.19) results in the Black-Scholes type PDE that all derivatives, written on a forward, in the SABR world must solve [4]

$$\frac{\partial V}{\partial t} + \tfrac{1}{2}\sigma^2 F^{2\beta}\frac{\partial^2 V}{\partial F^2} + \rho\nu\sigma^2 F^\beta \frac{\partial^2 V}{\partial F \partial \sigma} + \tfrac{1}{2}\nu^2\sigma^2\frac{\partial^2 V}{\partial \sigma^2} - rV = 0. \tag{2.21}$$

This PDE together with the correct boundary conditions can to be solved to obtain the price of derivatives in the SABR model.

Closed form SABR formulae

In Hagan et al. [2002] they use perturbation expansions to obtain the following closed form approximation for a European call option with strike K and time to maturity $(T-t)$

$$V_{\text{SABR}}(K, F_0, \sigma_0, \beta, \nu, \rho) = e^{-r(T-t)}\left[F_0 N(d^+) - KN(d^-)\right] \tag{2.22}$$

where

$$d^\pm = \frac{\ln\left(\frac{F_0}{K}\right) \pm \frac{1}{2}\sigma^2_{\text{Imp}}(T-t)}{\sigma_{\text{Imp}}\sqrt{T-t}}, \tag{2.23}$$

F_0 and σ_0 are the spot underlying and spot volatility of the underlying respectively. The implied volatility σ_{Imp} is given by

$$\sigma_{\text{Imp}}(K) = \frac{\sigma_0\left\{1 + \left[\frac{(1-\beta)^2}{24}\frac{\sigma_0^2}{(e^{r(T-t)}K)^{1-\beta}} + \frac{1}{4}\frac{\rho\beta\nu\sigma_0}{(e^{r(T-t)}K)^{(1-\beta)/2}} + \frac{2-3\rho^2}{24}\nu^2\right](T-t) + \dots\right\}}{(e^{r(T-t)}K)^{(1-\beta)/2}\left\{1 + \frac{(1-\beta)^2}{24}\ln^2\left(\frac{e^{r(T-t)}}{K}\right) + \frac{(1-\beta)^4}{1920}\ln^4\left(\frac{e^{r(T-t)}}{K}\right) + \dots\right\}}\left(\frac{y}{\xi(y)}\right) \tag{2.24}$$

where

$$y = \frac{\nu}{\sigma_0}(e^{r(T-t)}K)^{(1-\beta)/2}\ln\left(\frac{e^{r(T-t)}}{K}\right) \tag{2.25}$$

$$\xi(y) = \ln\left(\frac{\sqrt{1-2\rho y + y^2} + y - \rho}{1-\rho}\right) \tag{2.26}$$

Dynamic SABR model

In Hagan et al. [2002] they propose the dynamic SABR model for derivatives which are path dependent and hence require a model not only calibrated to a single marginal distribution but to a whole range of marginal distributions. Examples of such options are Forward-Starting and American options. In the dynamic SABR model the forward value satisfies

$$\begin{aligned} dF_t &= \gamma(t)\sigma_t F_t^\beta dW_{F_t} \\ d\sigma_t &= \nu(t)\sigma_t dW_{\sigma_t} \\ dW_{F_t}dW_{\sigma_t} &= \rho(t)dt. \end{aligned}$$

[4] The same PDE can be obtained by making use of a result in Heath and Schweizer [2000], where the authors use a reverse Feyman Kač type theorem to obtain PDEs implied by risk-neutral stochastic processes.

By using the same arguments as in section (2.2.1) we obtain the following PDE

$$\frac{\partial V}{\partial t} + \tfrac{1}{2}\gamma^2(t)\sigma^2 F^{2\beta}\frac{\partial^2 V}{\partial F^2} + \rho(t)\nu(t)\sigma^2 F^{\beta}\frac{\partial^2 V}{\partial F \partial \sigma} + \tfrac{1}{2}\nu^2(t)\sigma^2\frac{\partial^2 V}{\partial \sigma^2} - rV = 0 \tag{2.27}$$

that the prices of derivatives on forward values must solve. In general time dependent coefficients makes the proofs of stability for finite difference schemes more complicated. For this model however, the time dependent functions will be step functions. By choosing the grid of the finite difference scheme appropriately, such that the discontinuities only occur on grid points, step functions can be handled as time-homogenous functions.

2.2.2 Heston model

One of the key differences between the SABR model and the model proposed in Heston [1993] are the assumptions made about the underlying. In the SABR model it is assumed that the underlying is a forward value whereas in the tradeable security is assumed to be the stock itself in Heston's model. For the model proposed in Heston [1993] the author used the following dynamics for the underlying

$$\begin{aligned} dS_t &= \mu S_t dt + \sqrt{\sigma_t} + S_t d\widetilde{W}_{S_t} \\ d\sigma_t &= \kappa(\theta - \sigma_t)dt + \nu\sqrt{\sigma_t} d\widetilde{W}_{\sigma_t} \end{aligned}$$

where the stock S_t and its volatility σ_t are correlated, $d\widetilde{W}_{S_t} d\widetilde{W}_{\sigma_t} = \rho dt$. By making the following substitutions

$$\begin{aligned} &p_S = \mu S, && q_S = \sqrt{\sigma}S, \\ &p_\sigma = \kappa(\theta - \sigma), && q_\sigma = \nu\sqrt{\sigma}, \\ &D = 0 \end{aligned}$$

and

$$\lambda_\sigma = \frac{\lambda\sqrt{\sigma}}{\nu}$$

in (2.12), to obtain the PDE that the value of derivatives, written on a tradeable security, must solve

$$\begin{aligned} \frac{\partial V}{\partial t} &+ rS\frac{\partial V}{\partial S} + (\kappa(\theta - \sigma) - \lambda\sigma)\frac{\partial V}{\partial \sigma} \\ &+ \tfrac{1}{2}\sigma S^2\frac{\partial^2 V}{\partial S^2} + \rho\nu\sigma S\frac{\partial^2 V}{\partial S \partial \sigma} + \tfrac{1}{2}\nu^2\sigma\frac{\partial^2 V}{\partial \sigma^2} - rV = 0 \end{aligned}$$

We can remove the market price of volatility risk from the pricing formulae by introducing the following new calibration parameters

$$\kappa^* = \kappa + \lambda \quad \text{and} \quad \theta^* = \frac{\kappa\theta}{\kappa + \lambda}.$$

The PDE above can now be written in its more familiar form

$$\begin{aligned} \frac{\partial V}{\partial t} &+ rS\frac{\partial V}{\partial S} + \kappa^*(\theta^* - \sigma)\frac{\partial V}{\partial \sigma} \\ &+ \tfrac{1}{2}\sigma S^2\frac{\partial^2 V}{\partial S^2} + \rho\nu\sigma S\frac{\partial^2 V}{\partial S \partial \sigma} + \tfrac{1}{2}\nu^2\sigma\frac{\partial^2 V}{\partial \sigma^2} - rV = 0. \end{aligned} \tag{2.28}$$

2.2.3 Hull & White model

Similarly as with Heston's model, the model proposed in Hull and White [1987] uses stock as the underlying tradeable security rather than a forward value. For this model the authors assumed the following dynamics for the underlying

$$\begin{aligned} dS_t &= \psi(S_t, \sigma_t, t)dt + \sqrt{\sigma_t} + S_t d\widetilde{W}_{S_t} \\ d\sigma_t &= \mu\sigma_t dt + \xi\sigma_t d\widetilde{W}_{\sigma_t} \end{aligned}$$

where the stock S_t and its volatility σ_t are correlated, $d\widetilde{W}_{S_t} d\widetilde{W}_{\sigma_t} = \rho dt$. By making the following substitutions

$$\begin{aligned} &p_S = \psi(S, \sigma, t), \quad && q_S = \sqrt{\sigma} S, \\ &p_\sigma = \mu\sigma, && q_\sigma = \xi\sigma, \\ &D = 0 && \end{aligned}$$

and $\lambda_\sigma = 0$, in (2.12), we obtain the PDE that the value of derivatives must solve

$$\begin{aligned} \frac{\partial V}{\partial t} &+ rS\frac{\partial V}{\partial S} + \mu\sigma\frac{\partial V}{\partial \sigma} \\ &+ \tfrac{1}{2}\sigma S^2 \frac{\partial^2 V}{\partial S^2} + \rho\xi\sigma^{3/2} S \frac{\partial^2 V}{\partial S \partial \sigma} + \tfrac{1}{2}\xi^2\sigma^2 \frac{\partial^2 V}{\partial \sigma^2} - rV = 0. \end{aligned} \tag{2.29}$$

Chapter 3

One Dimensional Finite Difference Methods

In this chapter we will introduce the finite difference method by making use of one dimensional parabolic PDEs. This chapter will mainly focus on numerical methods to find an approximation to the solution, $u : \Omega \to \mathbb{R}$, of the following parabolic equation

$$\frac{\partial u}{\partial \tau} = a(x)\frac{\partial^2 u}{\partial x^2} + d(x)\frac{\partial u}{\partial x} \tag{3.1}$$

on the domain $(x, \tau) \in \Omega$ where $a(x) > 0$ and $\Omega = [x_{\text{min}}, x_{\text{max}}] \times \mathbb{R}^+$. The consistency, stability and convergence of the numerical schemes considered will be discussed in detail. This chapter serves as an introduction to the finite difference method, hence we will for the most part only discuss the classical one dimensional solvers, namely: fully explicit, fully implicit and the Crank-Nicolson method. Some modifications on these schemes will also be discussed such as exponential fitting, see Duffy [2006]. For reasons that will become clear when we extend the ideas of this chapter to two dimensions we will also spend some time on the one dimensional convection equation

$$\frac{\partial u}{\partial \tau} = d(x)\frac{\partial u}{\partial x}.$$

We discretisize the domain, Ω, and derive the discrete approximations to the continuous derivatives as a first step.

3.1 Discrete approximations

In order to approximate the solutions of PDEs with the finite difference method we need to truncate our infinite domain Ω to the bounded domain $\overline{\Omega} = [x_{\text{min}}, x_{\text{max}}] \times [0, T]$. Define the following partitions for $[x_{\text{min}}, x_{\text{max}}]$ and $[0, T]$ respectively

$$0 = \tau_0 < \tau_1 < \ldots < \tau_l = T$$

$$x_{\text{min}} = x_0 < x_1 < \ldots < x_m = x_{\text{max}}.$$

Our aim is to find approximations of the exact solution of (3.1) on the mesh

$$\widehat{\Omega} = \{(x_i, \tau_k) | i = 0,\ 1,\ \ldots,\ m, \quad k = 0,\ 1,\ \ldots,\ l\}.$$

The approximation at each mesh point is denoted by

$$u_i^k \approx u(x_i, \tau_k).$$

Assuming that the mesh points are uniformly spaced we can write

$$\begin{aligned} x_i &= x_{\text{min}} + ih_x \quad \text{for } i = 0,\ 1,\ \ldots,\ m \\ \tau_k &= k\Delta t \quad \text{for } l = 0,\ 1,\ \ldots,\ l \end{aligned}$$

where $h_x = \frac{x_{\text{max}} - x_{\text{min}}}{m}$ and $\Delta t = \frac{T}{l}$. The finite difference approximations for the derivatives in (3.1) at a reference point (x_i, τ_k) can be obtained by considering the Taylor expansions at the surrounding nodes: (x_{i-1}, τ_k), (x_{i+1}, τ_k) and (x_i, τ_{k+1})

$$u(x_{i-1}, \tau_k) = u - h_x \frac{\partial u}{\partial x} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} - \frac{1}{6} h_x^3 \frac{\partial^3 u}{\partial x^3} + O(h_x^4) \tag{3.2}$$

$$u(x_{i+1}, \tau_k) = u + h_x \frac{\partial u}{\partial x} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} + \frac{1}{6} h_x^3 \frac{\partial^3 u}{\partial x^3} + O(h_x^4) \tag{3.3}$$

$$u(x_i, \tau_{k+1}) = u + \Delta t \frac{\partial u}{\partial t} + \tfrac{1}{2} \Delta t \frac{\partial^2 u}{\partial t^2} + O(\Delta t^3). \tag{3.4}$$

Rearranging (3.2) and (3.3) results in first order forward and backward approximations of the first order derivative

$$\frac{\partial u}{\partial x}(x_i, \tau_k) = \frac{u(x_i, \tau_k) - u(x_{i-1}, \tau_k)}{h_x} + \tfrac{1}{2} h_x \frac{\partial^2 u}{\partial x^2} - \frac{1}{6} h_x^2 \frac{\partial^3 u}{\partial x^3} + O(h_x^3) \tag{3.5}$$

$$\frac{\partial u}{\partial x}(x_i, \tau_k) = \frac{u(x_{i+1}, \tau_k) - u(x_i, \tau_k)}{h_x} - \tfrac{1}{2} h_x \frac{\partial^2 u}{\partial x^2} - \frac{1}{6} h_x^2 \frac{\partial^3 u}{\partial x^3} + O(h_x^3). \tag{3.6}$$

After rearranging (3.4) we obtain the first order approximation of the time derivative

$$\frac{\partial u}{\partial \tau}(x_i, \tau_k) = \frac{u(x_i, \tau_{k+1}) - u(x_i, \tau_k)}{\Delta t} - \tfrac{1}{2} \Delta t \frac{\partial^2 u}{\partial \tau^2} + O(\Delta t^3)$$

A second order central difference approximation to the first order spacial derivative can be obtained by subtracting (3.2) from (3.3)

$$\frac{\partial u}{\partial x}(x_i, \tau_k) = \frac{u(x_{i+1}, \tau_k) - u(x_{i-1}, \tau_k)}{2h_x} - \frac{1}{6} h_x^2 \frac{\partial^3 u}{\partial x^3} + O(h_x^4). \tag{3.7}$$

Finally by adding (3.2) and (3.3) we obtain a second order approximation to the second order spacial derivative

$$\frac{\partial^2 u}{\partial x^2}(x_i, \tau_k) = \frac{u(x_{i+1}, \tau_k) - 2u(x_i, \tau_k) + u(x_{i-1}, \tau_k)}{h_x^2} - \frac{1}{12} h_x^2 \frac{\partial^4 u}{\partial x^4} + O(h_x^4). \tag{3.8}$$

Discrete difference operators can be defined as follows

$$\begin{aligned}
\Delta_t^+ u_i^k &= \frac{u_i^{k+1} - u_i^k}{\Delta t} \\
\Delta_x^+ u_i^k &= \frac{u_{i+1}^k - u_i^k}{h_x} && (3.9) \\
\Delta_x^- u_i^k &= \frac{u_i^k - u_{i-1}^k}{h_x} && (3.10) \\
\Delta_x u_i^k &= \frac{u_{i+1}^k - u_{i-1}^k}{2h_x} \\
\Delta_x^2 u_i^k &= \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h_x^2}.
\end{aligned}$$

3.2 A model problem

Many valuation problems in the Black-Scholes world can be solved by finding the solution, $u(x,t)$, of a linear convection-diffusion PDE of the form

$$\frac{\partial u}{\partial \tau} = a(x)\frac{\partial^2 u}{\partial x^2} + d(x)\frac{\partial u}{\partial x}. \tag{3.11}$$

on the domain $(x,\tau) \in \overline{\Omega}$ where $a(x)$ is strictly positive. The solution of this PDE will need to satisfy an initial condition

$$u(x,0) = \Psi(x)$$

and boundary conditions at $x = x_{\text{min}}$ and $x = x_{\text{max}}$. The boundary condition at x_{min} will generally be of the form

$$\alpha_0(\tau)u + \alpha_1(\tau)\frac{\partial u}{\partial x} = \alpha_2(\tau) \tag{3.12}$$

where $\alpha_0(\tau) \geq 0$, $\alpha_1(\tau) \leq 0$ and $\alpha_0(\tau) - \alpha_1(\tau) > 0$. The boundary condition at x_{max} is given by

$$\beta_0(\tau)u + \beta_1(\tau)\frac{\partial u}{\partial x} = \beta_2(\tau) \tag{3.13}$$

where $\beta_0(\tau) \geq 0$, $\beta_1(\tau) \geq 0$ and $\beta_0(\tau) + \beta_1(\tau) > 0$. The reason for the constraints on $\alpha_0(\tau)$, $\alpha_1(\tau)$, $\beta_0(\tau)$ and $\beta_1(\tau)$ will become clear in later sections when we discuss stability.

3.3 θ-method

By substituting the discrete approximations to the continuous derivatives in equation (3.11) we obtain

$$\begin{aligned}
\frac{u_i^{k+1} - u_i^k}{\Delta t} =& \theta \left[a_i \frac{u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}}{h_x^2} + d_i \frac{u_{i+1}^{k+1} - u_{i-1}^{k+1}}{2h_x} \right] \\
&+ (1-\theta) \left[a_i \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h_x^2} + d_i \frac{u_{i+1}^k - u_{i-1}^k}{2h_x} \right] && (3.14)
\end{aligned}$$

for $i = 1,\ 2,\ \ldots,\ m-1$ and $k = 0,\ 1,\ \ldots,\ l-1$, where $a_i = a(x_i)$, $d_i = d(x_i)$ and $\theta \in [0,1]$. The classical fully implicit, fully explicit and Crank-Nicolson schemes are special cases of the θ-method and can be obtained by letting $\theta = 1$, $\theta = -1$, and $\theta = \frac{1}{2}$ respectively. After rearranging we obtain

$$\begin{aligned}(-\lambda_{xx}a_i + d_i\lambda_x)\theta u_{i-1}^{k+1} + (1 + 2\theta\lambda_{xx}a_i)\, u_i^{k+1} + (-\lambda_{xx}a_i - \lambda_x d_i)\theta u_{i+1}^{k+1} \\ = (\lambda_{xx}a_i - d_i\lambda_x)(1-\theta)u_{i-1}^{k} + (1 - 2(1-\theta)\lambda_{xx}a_i)\, u_i^{k} + (\lambda_{xx}a_i + \lambda_x d_i)(1-\theta)u_{i+1}^{k}\end{aligned} \tag{3.15}$$

for $i = 1,\ 2,\ \ldots,\ m-1$ and $k = 0,\ 1,\ \ldots,\ l-1$, where $\lambda_x = \frac{\Delta t}{2h_x}$, $\lambda_{xx} = \frac{\Delta t}{h_x^2}$. The boundary conditions at $x = x_{\text{min}}$ and $x = x_{\text{max}}$ can be rewritten in discrete form as

$$(\alpha_0^k h_x - \alpha_1^k)u_0^k + \alpha_1^k u_1^k = \alpha_2^k h_x \tag{3.16}$$

$$-\beta_1^k u_{m-1}^k + (\beta_0^k h_x + \beta_1^k)u_m^k = \beta_2^k h_x \tag{3.17}$$

respectively. These approximations remove the second order accuracy of the scheme on the boundaries and may have an adverse effect on the overall accuracy of the scheme. A well known solution to this accuracy problem is to choose a new uniform grid such that the boundary value at $x = x_{\text{min}}$ and $x = x_{\text{max}}$ occurs half way between the first two and last two grid points respectively, the discrete approximations of (3.12) and (3.13) then become

$$\begin{aligned}\tfrac{1}{2}\alpha_0^k(u_0^k + u_1^k) + \alpha_1^k\frac{u_1^k - u_0^k}{h_x} &= \alpha_2^k \\ \tfrac{1}{2}\beta_0^k(u_m^k + u_{m-1}^k) + \beta_1^k\frac{u_m^k - u_{m-1}^k}{h_x} &= \beta_2^k\end{aligned}$$

after rearranging we obtain

$$(\tfrac{1}{2}\alpha_0^k h_x - \alpha_1^k)u_0^k + (\tfrac{1}{2}\alpha_0^k h_x + \alpha_1^k)u_1^k = \alpha_2^k h_x \tag{3.18}$$

$$(\tfrac{1}{2}\beta_0^k h_x - \beta_1^k)u_{m-1}^k + (\tfrac{1}{2}\beta_0^k h_x + \beta_1^k)u_m^k = \beta_2^k h_x \tag{3.19}$$

It is well known that this approximation of the boundary conditions results in a second order scheme, see [Morton and Mayers, 1996, 2.13]. To ensure that the solution of (3.11) is unique, one of the boundary conditions must be a Dirichlet condition, see Morton and Mayers [1996] for example.

3.4 Consistency

This section is based on a discussion of consistency in [Smith, 1985, Chapter 2]. It is possible to construct a stable finite difference scheme to approximate a certain PDE that can converge to a different PDE as the step sizes tend to zero. Such a scheme is called inconsistent.

Consider the following initial value problem

$$\begin{aligned}\frac{\partial u}{\partial \tau} &= Lu, \quad x \in \mathbb{R}^n, \quad \tau > 0 \\ u(x,0) &= f(x), \quad x \in \mathbb{R}^n\end{aligned} \tag{3.20}$$

and consider some finite difference scheme to obtain an approximation to the solution $u(x,\tau) : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}$ of the continuous problem

$$\begin{aligned} L_{\mathbf{h}_x}^{\Delta t} u_i^k &= 0 \\ u_i^0 &= f(x_i) \end{aligned} \tag{3.21}$$

where i is a multi index, $L_{\mathbf{h}_x}^{\Delta t}$ is the discrete approximation of L with $\mathbf{h}_x$ a step size vector containing the step sizes in the respective directions and Δt the step size in the τ direction. The following definition is a well known definition of consistency, see Smith [1985] and Duffy [2006] for example.

Definition 3.4.1. *(Consistent, [Smith, 1985, Chapter 2]) The finite difference scheme* (3.21) *is consistent with the partial differential equation* (3.20) *if for any function $v = v(x,\tau)$, with a sufficient number of continues derivatives enabling Lv to a evaluated at (x_i, τ_k), the following relationship holds*

$$\mathcal{T}_{h_x}^{\Delta t} v = \left(\frac{\partial v}{\partial \tau} - Lv \right)_i^k - L_{h_x}^{\Delta t} v(x_i, \tau_k) \to 0 \quad \textit{as} \quad h_x, \Delta t \to 0 \quad \textit{and} \quad (x_i, \tau_k) \to (x, \tau).$$

$\mathcal{T}_{h_x}^{\Delta t} v$ is also known as the truncation error.

In other words, we say a finite difference scheme is consistent if the truncation error approaches zero when the step sizes in the respective directions tend to zero. In particular if we choose v in the definition above to be the solution of (3.11) we obtain the following very useful definition of consistency

Definition 3.4.2. *(Consistent, [Smith, 1985, Chapter 2]) The finite difference scheme* (3.21) *is consistent with the partial differential equation* (3.20) *if*

$$L_{h_x}^{\Delta t} v(x_i, \tau_k) \to 0 \quad \textit{as} \quad h_x, \Delta t \to 0 \quad \textit{and} \quad (x_i, \tau_k) \to (x, \tau)$$

where v is the analytical solution of (3.11).

The truncation error of the θ-method can then be written as

$$\begin{aligned} L_{h_x}^{\Delta t} v(x_i, \tau_k) = \Delta_t^+ v(x_i, \tau_k) &- \theta(a(x_i)\Delta_x^2 v(x_i, \tau_{k+1}) + d(x_i)\Delta_x v(x_i, \tau_{k+1})) \\ &- (1-\theta)(a(x_i)\Delta_x^2 v(x_i, \tau_k) + d(x_i)\Delta_x v(x_i, \tau_k)) \end{aligned} \tag{3.22}$$

To compute the truncation error of the θ-method we consider Taylor expansions about $(x_i, \tau_{k+\frac{1}{2}})$

$$\begin{aligned} v(x_i, \tau_k) &= v(x_i, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial v}{\partial \tau}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^2 v}{\partial \tau^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \\ v(x_i, \tau_{k+1}) &= v(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial v}{\partial t}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^2 v}{\partial \tau^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \end{aligned}$$

By subtracting and dividing by Δt we obtain

$$\Delta_t^+ v(x_i, \tau_k) = \frac{\partial v}{\partial t}(x_i, \tau_{k+\frac{1}{2}}) + \frac{1}{24}\Delta t^2 \frac{\partial^3 v}{\partial \tau^3}(x_i, \tau_{k+\frac{1}{2}}) + \dots$$

Rearranging (3.7) and applying the equation to both time levels τ_k and τ_{k+1} yields

$$\begin{aligned} \Delta_x v(x_i, \tau_k) &= \frac{\partial v}{\partial x}(x_i, \tau_k) + O(h_x^2) \\ \Delta_x v(x_i, \tau_{k+1}) &= \frac{\partial v}{\partial x}(x_i, \tau_{k+1}) + O(h_x^2) \end{aligned} \tag{3.23}$$

Expanding the terms on the right about $(x_i, \tau_{k+\frac{1}{2}})$ yields

$$\begin{aligned}
\Delta_x v(x_i, \tau_k) &= \left(\frac{\partial v}{\partial x}(x_i, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial x \partial t^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right) + O(h_x^2) \\
\Delta_x v(x_i, \tau_{k+1}) &= \left(\frac{\partial v}{\partial x}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial x \partial \tau^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right) + O(h_x^2).
\end{aligned}$$

Similarly by rearranging (3.8) and expanding about $(x_i, \tau_{k+\frac{1}{2}})$ we obtain

$$\begin{aligned}
\Delta_x^2 v(x_i, \tau_k) &= \left(\frac{\partial^2 v}{\partial x^2}(x_i, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial x^2}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right) + O(h_x^2) \\
\Delta_x^2 v(x_i, \tau_{k+1}) &= \left(\frac{\partial^2 v}{\partial x^2}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial x^2}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right) + O(h_x^2).
\end{aligned}$$

By substituting the equations above in (3.22) we obtain

$$\begin{aligned}
L_{h_x}^{\Delta t} v(x_i, \tau_k) = {} & \frac{\partial v}{\partial \tau}(x_i, \tau_{k+\frac{1}{2}}) + \frac{1}{24}\Delta t^2 \frac{\partial^3 v}{\partial \tau^3}(x_i, \tau_{k+\frac{1}{2}}) + \dots \\
& - a(x_i) \left[\frac{\partial^2 v}{\partial x^2}(x_i, \tau_{k+\frac{1}{2}}) + (2\theta - 1)\tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, \tau_{k+\frac{1}{2}}) \right. \\
& \qquad \left. + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right] \\
& - d(x_i) \left[\frac{\partial v}{\partial x}(x_i, \tau_{k+\frac{1}{2}}) + (2\theta - 1)\tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, \tau_{k+\frac{1}{2}}) \right. \\
& \qquad \left. + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right] \\
& + O(h_x^2) \\
= {} & \left[\frac{\partial v}{\partial \tau}(x_i, \tau_{k+\frac{1}{2}}) - a(x_i) \frac{\partial^2 v}{\partial x^2}(x_i, \tau_{k+\frac{1}{2}}) - d(x_i) \frac{\partial v}{\partial x}(x_i, \tau_{k+\frac{1}{2}}) \right] \\
& - \left[a(x_i) \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, \tau_{k+1}) + d(x_i)\tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, \tau_{k+\frac{1}{2}}) \right] (2\theta - 1)\tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, \tau_{k+1}) \\
& + O(\Delta t^2) + O(h_x^2) \\
= {} & - \left[a(x_i) \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, \tau_{k+1}) + d(x_i)\tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, \tau_{k+\frac{1}{2}}) \right] (2\theta - 1)\tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, \tau_{k+1}) \\
& + O(\Delta t^2) + O(h_x^2) \qquad (3.24)
\end{aligned}$$

where we used the fact that v is a solution of (3.11). From (3.24) it is clear that $L_{h_x}^{\Delta t} v(x_i, \tau_k) \to 0$ as $h_x, \Delta t \to 0$ for all $\theta \in [0, 1]$, hence we conclude that the θ-method is consistent. For general values of θ we see that the truncation error is of $O(\Delta t) + O(h_x^2)$, a scheme of $O(\Delta t^2) + O(h_x^2)$ can be obtained by letting $\theta = \frac{1}{2}$. This scheme is known as the Crank-Nicolson scheme.

3.5 Stability

Stability of the θ-method applied to (3.11) will be discussed using two different methods, namely von Neumann stability analysis and a matrix method of analysis. In Smith [1985] a nice heuristic description of stability is given:

> The essential idea defining stability is that the numerical process, applied exactly, should limit the amplification of all components of the initial conditions.

3.5.1 von Neumann stability analysis

This paragraph is based on a discussion of von Neumann stability analysis in [Smith, 1985, Chapter 2]. We are concerned with the θ-method as $h_x \to 0$ and $\Delta t \to 0$. The first step is to represent each initial data point as a Fourier series which is formulated in terms of complex exponential forms

$$u_i^0 = \sum_{s\,=\,0}^{m} A_s e^{I\beta_s i h_x}, \qquad \text{for } i = 0,\ 1,\ \ldots,\ m$$

where $I = \sqrt{-1}$ and $\beta_s = \frac{s\pi}{x_{\max} - x_{\min}}$. The unknown constants A_s, $s = 0,\ 1,\ \ldots,\ m$ are to be solved. This requires the solution of a system of $m+1$ equations in $m+1$ unknowns which has a unique solution. This shows that the initial data can be expressed in complex exponential form. The linearity of (3.11) enables us to add different solutions of (3.11) to obtain new solutions. This additive property allows us to investigate the propagation of one initial value only, $e^{I\beta i h_x}$ for example. Since A_s is constant, for all s, it can be neglected. We need to investigate the propagation of this term as time increases, for this purpose we let

$$u_i^k = e^{I\beta x} e^{\alpha t} = e^{I\beta i h_x} e^{\alpha k \Delta t} = \gamma^k e^{I\beta i h_x} \tag{3.25}$$

where $\gamma = e^{\alpha k}$ is called the amplification factor. A finite difference scheme is said to be stable, in the sense of Lax-Richtmyer, if the absolute value of the exact solution of the scheme remains bounded for all $k < l$ as $h_x \to 0$ and $\Delta t \to 0$. From (3.25) we see that a sufficient condition is

$$|\gamma| \leq 1.$$

Strictly speaking von Neumann stability analysis applies only to problems with constant coefficients. For problems with non-constant coefficients von Neumann stability analysis only gives a necessary condition, see Smith [1985] and Morton and Mayers [1996]. In Morton and Mayers [1996] it is stated that von Neumann stability analysis can still be applied to variable coefficient problems locally, since instability is a local phenomenon. In Smith [1985] it is stated that von Neumann stability analysis gives useful results even in cases where its application is not fully justified. In Craig and Sneyd [1988] and McKee et al. [1996] they use von Neumann stability analysis as an indicator of stability for their non-constant coefficient problems. We conclude the motivation for application of von Neumann stability analysis to variable coefficient problems with the following quote from Duffy [2005],

> Much of the literature uses the von Neumann theory to prove stability of finite difference schemes, Tavella and Randall [2000]. This theory was developed by John von Neumann, a Hungarian-American mathematician, the father of the modern computer and probably one of the greatest brains of the twentieth century.

To obtain the conditions under which the θ-method is stable we substitute (3.25) in (3.14) and assume that the coefficients of (3.11) are constant. The divided difference approximations becomes

$$\begin{aligned}
\Delta_t^+ u_i^k &= \frac{\gamma^{k+1} - \gamma^k}{\Delta t} e^{I\beta i h_x} \\
\Delta_x u_i^k &= \frac{1}{2h_x} \gamma^k e^{I\beta i h_x} (e^{I\beta h_x} - e^{-I\beta h_x}) \\
&= I \frac{1}{h_x} \gamma^k e^{I\beta i h_x} \sin(\alpha h_x) \\
\Delta_x^2 u_i^k &= \frac{1}{h_x^2} \gamma^k e^{I\beta i h_x} (e^{I\beta h_x} - 2 + e^{-I\beta h_x}) \\
&= -\frac{4}{h_x^2} \gamma^{k+1} \sin^2\left(\frac{\beta h_x}{2}\right) e^{I\beta i h_x}
\end{aligned}$$

By substituting these equations into (3.14) and rearranging we obtain

$$\gamma = \frac{1 - (1-\theta) a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right) + I(1-\theta) d \frac{\Delta t}{h_x} \sin(\alpha h_x)}{1 + \theta a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right) - I\theta d \frac{\Delta t}{h_x} \sin(\alpha h_x)}$$

which results in

$$|\gamma|^2 = \frac{\left[1 - (1-\theta) a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right)\right]^2 + \left[(1-\theta) d \frac{\Delta t}{h_x} \sin(\alpha h_x)\right]^2}{\left[1 + \theta a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right)\right]^2 + \left[\theta d \frac{\Delta t}{h_x} \sin(\alpha h_x)\right]^2}$$

A sufficient condition for $|\gamma| \leq 1$ is given by

$$\left[(1-\theta) d \frac{\Delta t}{h_x} \sin(\alpha h_x)\right]^2 \leq \left[\theta d \frac{\Delta t}{h_x} \sin(\alpha h_x)\right]^2 \tag{3.26}$$

$$\left[1 - (1-\theta) a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right)\right]^2 \leq \left[1 + \theta a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right)\right]^2 \tag{3.27}$$

By removing the common factors from equation (3.26) we see that the inequality will hold if $\theta \geq \frac{1}{2}$. Expanding (3.27) yields

$$\begin{aligned}
1 - 2(1-\theta) a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right) + (1-\theta)^2 a^2 \left(\frac{4\Delta t}{h_x^2}\right)^2 \sin^4\left(\frac{\beta h_x}{2}\right) \\
\leq 1 + 2\theta a \frac{4\Delta t}{h_x^2} \sin^2\left(\frac{\beta h_x}{2}\right) + \theta^2 a^2 \left(\frac{4\Delta t}{h_x^2}\right)^2 \sin^4\left(\frac{\beta h_x}{2}\right)
\end{aligned}$$

From the fact that $a \geq 0$ it is easy to see that this inequality will be satisfied when $(1-\theta)^2 \leq \theta^2$ which in turn is true if $\theta \geq \frac{1}{2}$. Thus we can conclude that the θ-method applied to (3.11) is von Neumann stable when $\theta \geq \frac{1}{2}$.

3.5.2 Matrix method of analysis

To state the Lax-Richtmyer definition of stability in terms of vector and matrix norms we will need to give a quick introduction of these norms. The following subsections of matrix and vector norms are taken from [Smith, 1985, Chapter 2].

Vector norms

The norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is a positive scalar valued function, $||\mathbf{x}|| : \mathbb{R}^n \rightarrow \mathbb{R}$, that satisfies the following axioms,

- $||\mathbf{x}|| > 0$ if $\mathbf{x} \neq \mathbf{0}$ and $||\mathbf{x}|| = 0$ if $\mathbf{x} = \mathbf{0}$.
- $||c\mathbf{x}|| = |c|||\mathbf{x}||$ for a complex scalar c.
- $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$.

The most commonly used norms are defined as follows,

- The 1-norm: $||\mathbf{x}||_1 = \sum_{i\,=\,1}^{n} |x_i|$.
- The infinity norm (maximum norm): $||\mathbf{x}||_\infty = \max_i |x_i|$.
- The 2-norm: $||\mathbf{x}||_2 = \left[\sum_{i\,=\,1}^{n} |x_i|^2\right]^{\frac{1}{2}}$.

The 2-norm gives the "length" of the vector.

Matrix norms

The norm of a matrix $\mathbf{A} \in \mathbb{R}^m \times \mathbb{R}^n$ is a positive scalar valued function, $||\mathbf{A}|| : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$, that satisfies the following axioms,

- $||\mathbf{A}|| > 0$ if $\mathbf{A} \neq \mathbf{0}$ and $||\mathbf{A}|| = 0$ if $\mathbf{A} = \mathbf{0}$.
- $||c\mathbf{A}|| = |c|||\mathbf{A}||$ for a complex scalar c.
- $||\mathbf{A} + \mathbf{B}|| \leq ||\mathbf{A}|| + ||\mathbf{B}||$.
- $||\mathbf{AB}|| \leq ||\mathbf{A}||||\mathbf{B}||$.

Since matrix and vector norms occur together it is essential that they satisfy a condition similar to the last inequality above. A matrix norm is said to be compatible with a vector norm if

$$||\mathbf{Ax}|| \leq ||\mathbf{A}||||\mathbf{x}||, \quad ||\mathbf{x}|| \neq \mathbf{0}.$$

Subordinate matrix norms

Let $\mathbf{A}$ be a real $n \times n$ matrix and $\mathbf{x} \in S \subset \mathbb{R}^n$ where

$$S = \{\mathbf{x} \in \mathbb{R}^n | \quad ||\mathbf{x}|| = 1\}$$

An example of a matrix norm that automatically satisfies that compatibility condition is

$$||\mathbf{A}|| = \max_{\mathbf{x}\,\in\,S} ||\mathbf{Ax}||.$$

To see this, let $\mathbf{x}_1 \in S$ then $||\mathbf{A}\mathbf{x}_1|| \leq \max_{\mathbf{x}\,\in\,S} ||\mathbf{A}\mathbf{x}|| = ||\mathbf{A}|| = ||\mathbf{A}||||\mathbf{x}_1||$. This matrix norm is said to be subordinate to the vector norm and always has the following property

$$||\mathbf{I}|| = \max_{\mathbf{x}\,\in\,S} ||\mathbf{I}\mathbf{x}|| = \max_{\mathbf{x}\,\in\,S} ||\mathbf{x}|| = 1$$

where $\mathbf{I}$ is the identity matrix. The definitions of the 1, 2 and ∞ norms with $||\mathbf{x}|| = 1$ leads to the following well known matrix norms, let $\mathbf{A} = (a_{i,\,j})$

- The 1-norm: $||\mathbf{A}||_1 = \max_j \sum_{i\,=\,1}^n |a_{i,\,j}|$.
- The infinity norm: $||\mathbf{A}||_\infty = \max_i \sum_{j\,=\,1}^n |a_{i,\,j}|$.
- The 2-norm: $||\mathbf{A}||_2 = \rho(\mathbf{A}^T\mathbf{A})$, where $\rho : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the spectral radius function.

Useful definitions

The following definitions will aid our proof for the stability of the θ-method,

Definition 3.5.1. *(M-matrix, Johanson et al. [2002]) A matrix $M \in \mathbb{R}^n \times \mathbb{R}^n$ of the form*

$$M = \begin{bmatrix} a_{1,\,1} & -a_{1,\,2} & -a_{1,\,3} & \cdots \\ -a_{2,\,1} & a_{2,\,2} & -a_{2,\,3} & \cdots \\ -a_{3,\,1} & -a_{3,\,2} & a_{3,\,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where $a_{i,\,j}$, $i \neq j$ non-negative and $a_{i,\,i}$ positive, is called a non-singular M-matrix if there exists a positive vector $\mathbf{x} \in \mathbb{R}^n$ such that the vector $M\mathbf{x}$ is positive[1].

Theorem 3.5.1. *(M-matrix properties, Horváth [2004]) If $\mathbf{A} \in \mathbb{R}^n \times \mathbb{R}^n$ is an M-matrix, then*

- *$\mathbf{A}$ is nonsingular,*
- *$\mathbf{A}^{-1}$ is non-negative ($\mathbf{A}^{-1} \geq 0$, so-called monotone matrix),*
- *the estimation*

$$||\mathbf{A}^{-1}||_\infty \leq \frac{||\mathbf{x}||_\infty}{\min_i(\mathbf{A}\mathbf{x})_i}$$

where $(\mathbf{A}\mathbf{x})_i$ is the ith element of the vector $\mathbf{A}\mathbf{x}$ and, $\mathbf{x}$ is a positive vector such that $\mathbf{A}\mathbf{x} > 0$, holds.

Lax-Richtmyer stability

This subsection is taken from [Smith, 1985, Chapter 2]. Consider a finite difference scheme where the numerical approximations along the k-th and $(k+1)$-th time-rows are related by the following matrix equation

$$\mathbf{A}\mathbf{u}^{k+1} = \mathbf{C}\mathbf{u}^k + \mathbf{b}^k$$

[1] We call a vector (resp. matrix) positive if all the elements of the vector (resp. matrix) are positive.

where $\mathbf{A}$ and $\mathbf{C}$ are time-homogenous matrices and $\mathbf{b}^k$ is a residual vector containing boundary information. Assuming that $\mathbf{A}$ is non-singular we can rewrite this relationship as

$$\mathbf{u}^{k+1} = \mathbf{A}^{-1}\mathbf{C}\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}^k$$

which can be expressed more conveniently as

$$\mathbf{u}^{k+1} = \mathbf{B}\mathbf{u}^k + \mathbf{f}^k$$

where $\mathbf{B} = \mathbf{A}^{-1}\mathbf{C}$ and $\mathbf{f}^k = \mathbf{A}^{-1}\mathbf{b}^k$. Applied recursively yields

$$\begin{aligned}\mathbf{u}^k &= \mathbf{B}\mathbf{u}^{k-1} + \mathbf{f}^{k-1}\\ &= \mathbf{B}(\mathbf{B}\mathbf{u}^{k-2} + \mathbf{f}^{k-2}) + \mathbf{f}^{k-1}\\ &= \mathbf{B}^2\mathbf{u}^{k-2} + \mathbf{B}\mathbf{f}^{k-2} + \mathbf{f}^{k-1}\\ &= \dots\\ &= \mathbf{B}^k\mathbf{u}^0 + \mathbf{B}^{k-1}\mathbf{f}^0 + \mathbf{B}^{k-2}\mathbf{f}^1 + \dots + \mathbf{f}^{k-1}\end{aligned}$$

where $\mathbf{u}^0$ is the vector of initial values. To investigate the stability of the numerical scheme consider the effect of a perturbation on the initial vector from $\mathbf{u}^0$ to $\widehat{\mathbf{u}}^0$. The exact solution at the k-th time step will then be given by

$$\widehat{\mathbf{u}}^k = \mathbf{B}^k\widehat{\mathbf{u}}^0 + \mathbf{B}^{k-1}\mathbf{f}^0 + \mathbf{B}^{k-2}\mathbf{f}^1 + \dots + \mathbf{f}^{k-1}$$

The perturbation error at the k-th time level is defined by

$$\mathbf{e}^k = \widehat{\mathbf{u}}^k - \mathbf{u}^k$$

It follows that

$$\mathbf{e}^k = \widehat{\mathbf{u}}^k - \mathbf{u}^k = \mathbf{B}^k(\widehat{\mathbf{u}}^0 - \mathbf{u}^0) = \mathbf{B}^k\mathbf{e}^0$$

This equation shows that the matrix $\mathbf{B}$ plays an important part in the propagation of the error. For a scheme to be considered stable we require the initial error not to explode but to dampen, for compatible matrix and vector norms we have

$$||\mathbf{e}^k|| \leq ||\mathbf{B}^k||||\mathbf{e}^0||$$

Lax and Richtmyer define a scheme to be stable when there exists a number $M \in \mathbb{R}^+$, independent of k, h_x and Δt such that

$$||\mathbf{B}^k|| \leq M, \qquad \text{for} \quad k = 1,\ 2,\ \dots,\ l$$

This limits the amplification of any initial perturbation as well as rounding errors since

$$||\mathbf{e}^k|| \leq M||\mathbf{e}^0||$$

From the fact that

$$||\mathbf{B}^k|| \leq ||\mathbf{B}||^k$$

it follows that a finite difference scheme will be stable, in the sense of Lax-Richtmyer, if it can be shown that

$$||\mathbf{B}|| \leq 1$$

If this condition holds, the scheme is said to be stable under the appropriate matrix norm. For the rest of this thesis the maximum norm (∞-norm) will be used when we prove stability with the matrix method of analysis.

Stability analysis under the maximum norm

An advantage of the following analysis over von Neumann stability analysis is that no assumptions have to be made about the coefficients of the PDE. A key idea in the proof of stability using the matrix method is that the matrices that need to be solved during the time marching procedure are required to be M-matrices. We start the discussion by investigating the effect of the boundary conditions on the structure of the matrices. When we use (3.18) and (3.19) as our boundary conditions we need to make the additional assumption that

$$\begin{aligned} \tfrac{1}{2}\alpha_0^k h_x + \alpha_1^k &\leq 0 \\ \tfrac{1}{2}\beta_0^k h_x - \beta_1^k &\leq 0 \end{aligned}$$

to ensure that we invert only M-matrices. Although these boundary conditions will give higher order of convergence near the boundary, using (3.16) and (3.17) does not require extra assumptions about the relevant parameters. For simplicity we assume that the boundary conditions are time-homogenous and write them as

$$\overline{\alpha}_0 u_0^k - \overline{\alpha}_1 u_1^k = \overline{\alpha}_2 \tag{3.28}$$

$$-\overline{\beta}_1 u_{m-1}^k + \overline{\beta}_0 u_m^k = \overline{\beta}_2 \tag{3.29}$$

where $\overline{\alpha}_0$ and $\overline{\beta}_0$ are strictly positive parameters and $\overline{\alpha}_1$ and $\overline{\beta}_1$ are non-negative real values.

Equations (3.15), (3.28) and (3.29) can be written in matrix form as

$$(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})\mathbf{u}^{k+1} = (\mathbf{I}_{\text{ex}} + (1-\theta)\Delta t\mathbf{A})\mathbf{u}^{k+1} + \mathbf{b}$$

for $k = 1,\ 2,\ \ldots,\ l-1$, where $\mathbf{u}^k$ is the solution vector given by

$$\mathbf{u}^k = (u_0, u_1, \ldots, u_m)^T$$

the matrices $\mathbf{I}_{\text{ex}}$, $\mathbf{I}_{\text{im}}$ and $\mathbf{A}$ are respectively given by,

$$\mathbf{I}_{\text{im}} = \begin{bmatrix} \overline{\alpha}_0 & -\overline{\alpha}_1 & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & -\overline{\beta}_1 & \overline{\beta}_0 \end{bmatrix}$$

$$\mathbf{I}_{\text{ex}} = \begin{bmatrix} 0 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 0 \end{bmatrix}$$

and

$$\mathbf{A} = \begin{bmatrix} 0 & & & & & \\ \frac{a_1}{h_x^2} - \frac{d_1}{2h_x} & -2\frac{a_1}{h_x^2} & \frac{a_1}{h_x^2} + \frac{d_1}{2h_x} & & & \\ & & \ddots & & & \\ & & & \frac{a_{m-1}}{h_x^2} - \frac{d_{m-1}}{2h_x} & -2\frac{a_{m-1}}{h_x^2} & \frac{a_{m-1}}{h_x^2} + \frac{d_{m-1}}{2h_x} \\ & & & & & 0 \end{bmatrix}.$$

For all matrices of this thesis blank entries are zeros. The next step is to prove that $\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A}$ is a M-matrix. Since $a(x) > 0$ it is trivial to see that the diagonal of $\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A}$ will always be strictly positive. From the fact that non-positive off-diagonal elements are a necessary condition for the M-matrix property we deduce the following inequality as a necessary condition for M-matrix property of $\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A}$:

$$h_x \leq \text{sgn}(d_i)\frac{2a_i}{d_i} \tag{3.30}$$

where

$$\text{sgn}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0. \end{cases}$$

Assuming Dirichlet boundary conditions at both boundaries and that inequality (3.30) is satisfied we can utilize theorem 3.5.1, with $\mathbf{x} = (1, \ldots, 1)^T$, to obtain

$$||(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})^{-1}||_\infty \leq 1.$$

Thus

$$\begin{aligned} ||(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})^{-1}(\mathbf{I}_{\text{im}} + (1-\theta)\Delta t\mathbf{A})||_\infty &\leq ||(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})^{-1}||_\infty ||(\mathbf{I}_{\text{im}} + (1-\theta)\Delta t\mathbf{A})||_\infty \\ &\leq ||(\mathbf{I}_{\text{im}} + (1-\theta)\Delta t\mathbf{A})||_\infty \\ &= \max_i \left\{ \left|1 - \frac{2a_i}{h_x^2}(1-\theta)\Delta t\right| + \left|\frac{a_i}{h_x^2} - \frac{d_i}{2h_x}\right| + \left|\frac{a_i}{h_x^2} + \frac{d_i}{2h_x}\right|, 1 \right\} \\ &= \max_i \left\{ \left|1 - \frac{2a_i}{h_x^2}(1-\theta)\Delta t\right| + \frac{2a_i}{h_x^2}, 1 \right\} \\ &= 1 \end{aligned}$$

where the second last equality follows from the assumption that (3.30) holds and the last equality is true if

$$\frac{a_i\Delta t}{h_x^2} < \frac{1}{2(1-\theta)} \tag{3.31}$$

From this it is possible to deduce that the θ-method is only conditionally stable under the maximum norm for all $\theta \neq 1$. This is a well known constraint, see Morton and Mayers [1996] and Duffy [2006] for example. The key step in this proof of stability was the restrictive assumption that (3.30) holds. For convection dominated problems, where $\frac{a_i}{d_i}$ is very small, this restriction will be too severe. In Duffy [2006] an exponential fitting method is introduced that can be used to overcome this restriction.

Exponential fitting

In Duffy [2006] it is shown that the restrictive necessary condition for the M-matrix property in equation (3.30) can be removed by substituting the coefficient of the diffusion term with the following function

$$f(a_i, d_i, h_x) = \frac{d_i h_x}{2} \coth \frac{d_i h_x}{2a_i}. \tag{3.32}$$

The off-diagonal elements of the matrix $\mathbf{A}$ becomes

$$\frac{f(a_i, d_i, h_x)}{h_x^2} + \frac{d_i}{2h_x} \qquad \text{and} \qquad \frac{f(a_i, d_i, h_x)}{h_x^2} - \frac{d_i}{2h_x} \tag{3.33}$$

Furthermore the proposed function has the property that

$$f(a_i, d_i, h_x) \geq \text{sgn}(d_i)\frac{d_i h_x}{2} \tag{3.34}$$

for all h_x, implying that the off diagonal elements will always be non-positive. Equation (3.34) follows from the fact that $\coth\theta - 1 \geq 0$ if $\theta \geq 0$

$$f(x, y, \epsilon) - \frac{y\epsilon}{2} = \frac{y\epsilon}{2}\Big(\coth\frac{y\epsilon}{2x} - 1\Big) \geq 0 \quad \text{if} \quad y \geq 0$$

similarly from the fact that $\coth\theta + 1 < 0$ if $\theta < 0$

$$f(x, y, \epsilon) + \frac{y\epsilon}{2} = \frac{y\epsilon}{2}\Big(\coth\frac{y\epsilon}{2x} + 1\Big) > 0 \quad \text{if} \quad y < 0.$$

Using the analysis in the previous section it is easy to see that the fitted scheme will be stable if the following inequality holds

$$\frac{f(a_i, d_i, h_x)\Delta t}{h_x^2} < \frac{1}{2(1-\theta)} \tag{3.35}$$

To prove consistency of $f(x, y, \epsilon)$ with x we use L'Hopital's rule

$$\begin{aligned} \lim_{\epsilon \to 0} f(x, y, \epsilon) &= \lim_{\epsilon \to 0} \frac{y\epsilon}{2\tanh\frac{y\epsilon}{2x}} \\ &= \lim_{\epsilon \to 0} x\cosh^2\frac{y\epsilon}{2x} \\ &= x \end{aligned}$$

Consistency (Revisited)

We have shown that by making use of exponential fitting we can obtain a scheme that has favorable stability properties. Exponential fitting should not affect the order of convergence of our scheme. Second order convergence of $f(x, y, \epsilon)$ to x can be shown via a Taylor series expansion of $\coth(\theta)$

$$\coth\theta = \frac{1}{\theta} + \frac{1}{3}\theta - \frac{1}{45}\theta^3 + \frac{2}{945}\theta^5 - + \dots$$

which implies that

$$\begin{aligned} f(x, y, \epsilon) &= x + \frac{1}{3}\Big(\frac{y\epsilon}{2}\Big)^2\frac{1}{x} - \frac{1}{45}\Big(\frac{y\epsilon}{2}\Big)^4\frac{1}{x^3} + \frac{1}{945}\Big(\frac{y\epsilon}{2}\Big)^6\frac{1}{x^5} - + \dots \\ &= x + O(\epsilon^2) \end{aligned} \tag{3.36}$$

To ensure convergence to the true solution we require the modification of the coefficients to maintain the consistency of the original scheme. The truncation error for the modified scheme is given by

$$\begin{aligned}\widetilde{L}_{h_x}^{\Delta t} v(x_i,\tau_k) = \Delta_t^+ v(x_i,\tau_k) &- \theta(f(a(x_i),d(x_i),h_x)\Delta_x^2 v(x_i,\tau_{k+1}) + d(x_i)\Delta_x v(x_i,\tau_{k+1})) \\ &- (1-\theta)(f(a(x_i),d(x_i),h_x)\Delta_x^2 v(x_i,\tau_k) + d(x_i)\Delta_x v(x_i,\tau_k))\end{aligned}$$

With similar analysis to section 3.4 we obtain

$$\begin{aligned}\widetilde{L}_{h_x}^{\Delta t} v(x_i,\tau_k) &= \left(\frac{\partial v}{\partial \tau}(x_i,\tau_{k+\frac{1}{2}}) - f(a(x_i),d(x_i),h_x)\frac{\partial^2 v}{\partial x^2}(x_i,\tau_{k+\frac{1}{2}}) - d(x_i)\frac{\partial v}{\partial x}(x_i,\tau_{k+\frac{1}{2}})\right) \\ &\quad - \left(f(a(x_i),d(x_i),h_x)\frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i,\tau_{k+1}) + d(x_i)\frac{\partial^2 v}{\partial \tau \partial x}(x_i,\tau_{k+1})\right)(2\theta-1)\tfrac{1}{2}\Delta t \\ &\quad + O(\Delta t^2) + O(h_x^2) \\ &= \left(\frac{\partial v}{\partial \tau}(x_i,\tau_{k+\frac{1}{2}}) - a(x_i)\frac{\partial^2 v}{\partial x^2}(x_i,\tau_{k+\frac{1}{2}}) - d(x_i)\frac{\partial v}{\partial x}(x_i,\tau_{k+\frac{1}{2}})\right) \\ &\quad - \left(a(x_i)\frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i,\tau_{k+1}) + d(x_i)\frac{\partial^2 v}{\partial \tau \partial x}(x_i,\tau_{k+1})\right)(2\theta-1)\tfrac{1}{2}\Delta t \\ &\quad + O(\Delta t^2) + O(h_x^2) \\ &= -\left(a(x_i)\frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i,\tau_{k+1}) + d(x_i)\frac{\partial^2 v}{\partial \tau \partial x}(x_i,\tau_{k+1})\right)(2\theta-1)\tfrac{1}{2}\Delta t + O(\Delta t^2) + O(h_x^2)\end{aligned}$$

where we obtained the second equality by substituting (3.36). Again it is clear that $\widetilde{L}_{h_x}^{\Delta t} v(x_i,\tau_k) \to 0$ as $h_x, \Delta t \to 0$ for all $\theta \in [0,1]$, we can conclude that the modified θ-method is consistent. Second order convergence can be obtained by letting $\theta = \frac{1}{2}$.

3.6 Convergence

A finite difference scheme is said to be convergent under the relevant norm if

$$||\mathbf{u^k} - \mathbf{v}^k|| \to 0 \quad \text{as} \quad \Delta t \to 0$$

for every $\mathbf{u}^0$ for which the problem is well posed in the norm. The vector $\mathbf{v}^k$ contains the exact solutions of the original problem evaluated at time steps $t = k\Delta t$, for $k = 0,\ 1,\ \ldots,\ l$. The following theorem relates convergence, consistency and stability

Theorem 3.6.1. *(The Lax equivalence theorem) A consistent two-level scheme of the form*

$$\mathbf{u}^{k+1} = \mathbf{B}\mathbf{u}^k + \mathbf{b} \tag{3.37}$$

where $\mathbf{b}$ *is a vector containing the boundary information, for a well-posed linear initial value problem is convergent if and only if it is stable.*

Proof. Let $\mathbf{v}^k = (v(x_1,\tau_1),\ldots,v(x_{m-1},\tau_{m-1}))^T$ be the exact solution to the continuous operator. From definition 3.4.2 it follows that

$$\mathbf{v}^{k+1} = \mathbf{B}\mathbf{v}^k + \mathbf{b} + \mathbf{w}^k \tag{3.38}$$

where $\mathbf{w}^k$ is the truncation error at time k. By subtracting (3.37) from (3.38) we obtain

$$\begin{aligned}\mathbf{v}^{k+1} - \mathbf{u}^{k+1} &= \mathbf{B}(\mathbf{v}^k - \mathbf{u}^k) + \mathbf{w}^k \\ &= \mathbf{B}^2(\mathbf{v}^{k-1} - \mathbf{u}^{k-1}) + \mathbf{B}\mathbf{w}^{k-1} + \mathbf{w}^k \\ &= \dots \\ &= \mathbf{B}^{k+1}(\mathbf{v}^0 - \mathbf{u}^0) + \sum_{i=0}^{k} \mathbf{B}^i \mathbf{w}^{k-i}\end{aligned}$$

Assuming that $\mathbf{v}^0 - \mathbf{u}^0 = \mathbf{0}$ we obtain

$$||\mathbf{v}^{k+1} - \mathbf{u}^{k+1}|| \leq \sum_{i=0}^{k} ||\mathbf{B}||^i ||\mathbf{w}^{k-i}||$$

If the scheme is stable it follows that $||\mathbf{B}|| \leq 1$ from which it follows that

$$||\mathbf{v}^{k+1} - \mathbf{u}^{k+1}|| \leq \sum_{i=0}^{k} ||\mathbf{w}^{k-i}||$$

thus if the truncation error tends to zero, i.e. the scheme is consistent, then the scheme will be convergent. The proof of the necessary condition uses the principle of uniform boundedness in functional analysis and is outside of the scope of this project, see Morton and Mayers [1996] and references therein for a more complete proof. □

We conclude that the exponentially fitted θ-method is convergent under the maximum norm if

$$\frac{f(a_i, d_i, h_x)\Delta t}{h_x^2} < \frac{1}{2(1-\theta)}$$

holds for all $i = 1,\ 2,\ \dots,\ m-1$.

3.7 A convection equation

In this section we will consider finite difference methods to solve the one dimensional convection equation

$$\frac{\partial u}{\partial \tau} = d(x)\frac{\partial u}{\partial x} \tag{3.39}$$

on the domain $(x, \tau) \in \Omega = [x_{\text{min}}, x_{\text{max}}] \times \mathbb{R}^+$. The initial condition is given by

$$u(x, 0) = \Psi(x). \tag{3.40}$$

Since this is a first order equation we only need to pose one boundary condition, either at $x = x_{\text{min}}$ or $x = x_{\text{max}}$. The sign of $d(x)$ determines the direction of information flow which in turn determines the position of the boundary condition. If $d(x) < 0$ then information goes from $x = x_{\text{min}}$ to $x = x_{\text{max}}$ and if $d(x) > 0$ information flows in the opposite direction. To see this, note that the analytical solution of

$$\begin{aligned}\frac{\partial u}{\partial \tau} &= d\frac{\partial u}{\partial x} \\ u(x, 0) &= \Psi(x)\end{aligned}$$

on the domain $(x, \tau) \in \mathbb{R} \times \mathbb{R}^+$ is given by

$$u(x, \tau) = \Psi(x + d\tau)$$

where we assumed that d is constant. If $d(x) < 0$ the well defined problem is given by (3.39), (3.40) and the boundary condition by

$$u(x_{\min}, \tau) = \alpha(\tau)$$

if $d(x) > 0$ the boundary condition becomes

$$u(x_{\max}, \tau) = \beta(\tau).$$

3.7.1 θ-method

The θ-method for the one dimensional convection equation is given by,

$$\Delta_t^+ u_i^k = \theta d_i \Delta_x u_i^{k+1} + (1-\theta) d_i \Delta_x u_i^k \tag{3.41}$$

for $i = 1, 2, \ldots, m-1$ and $k = 0, 1, \ldots, l-1$, where $a_i = a(x_i)$, $d_i = d(x_i)$ and $\theta \in [0, 1]$. If we insist on central second order approximation for the spacial derivative we obtain

$$\begin{aligned}\lambda_x a_i \theta u_{i-1}^{k+1} + u_i^{k+1} - \lambda_x d_i \theta u_{i+1}^{k+1}& \\ = -\lambda_x d_i (1-\theta) u_{i-1}^k + u_i^k + \lambda_x d_i (1-\theta) u_{i+1}^k&\end{aligned} \tag{3.42}$$

for $i = 1, 2, \ldots, m-1$ and $k = 0, 1, \ldots, l-1$, where $\lambda_x = \frac{\Delta t}{2h_x}$. Depending on the sign of $d(x)$ the discrete form of the boundary condition is given by

$$u_0^k = \alpha^k \quad \text{or} \quad u_m^k = \beta^k$$

for $k = 0, 1, \ldots, l$.

3.7.2 Consistency

The truncation error of the θ-method can then be written as

$$\begin{aligned}L_{h_x}^{\Delta t} v(x_i, \tau_k) = \Delta_t^+ v(x_i, \tau_k) - \theta d(x_i) \Delta_x v(x_i, \tau_{k+1})& \\ - (1-\theta) d(x_i) \Delta_x v(x_i, \tau_k)&\end{aligned} \tag{3.43}$$

By simply letting $a(x) \equiv 0$ in section 3.4 it easy to see that

$$L_{h_x}^{\Delta t} v(x_i, \tau_k) = -d(x_i)(2\theta - 1)\tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, \tau_{k+1}) + O(\Delta t^2) + O(h_x^2)$$

Thus the θ-method is consistent in general and second order when $\theta = \frac{1}{2}$.

3.7.3 Stability

von Neumann stability analysis

Assuming constant coefficient and periodic initial data we can make use of von Neumann stability analysis by substituting (3.25) in (3.41) to obtain

$$\gamma = \frac{1 + I(1-\theta)d\frac{\Delta t}{h_x}\sin(\alpha h_x)}{1 - I\theta d\frac{\Delta t}{h_x}\sin(\alpha h_x)}.$$

From which it follows that

$$|\gamma|^2 = \frac{1 + \left[(1-\theta)d\frac{\Delta t}{h_x}\sin(\alpha h_x)\right]^2}{1 + \left[\theta d\frac{\Delta t}{h_x}\sin(\alpha h_x)\right]^2}$$

which will be less or equal than one if $\theta \geq \frac{1}{2}$.

Maximum norm analysis

Similarly as in section 3.5.2 the θ-method for the one dimensional convection equation can be written in matrix form as

$$(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})\mathbf{u}^{k\,+\,1} = (\mathbf{I}_{\text{ex}} + (1-\theta)\Delta t\mathbf{A})\mathbf{u}^{k\,+\,1} + \mathbf{b}$$

for $k = 1,\ 2,\ \ldots,\ l-1$, where the matrices $\mathbf{I}_{\text{ex}}$, $\mathbf{I}_{\text{im}}$ and $\mathbf{A}$ are respectively given by,

$$\mathbf{I}_{\text{im}} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

$$\mathbf{I}_{\text{ex}} = \begin{bmatrix} 0 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 0 \end{bmatrix}$$

and

$$\mathbf{A} = \begin{bmatrix} 0 & & & & \\ -\frac{d_0}{2h_x} & 0 & \frac{d_0}{2h_x} & & \\ & & \ddots & & \\ & & -\frac{d_{m\,-\,1}}{2h_x} & 0 & \frac{d_{m\,-\,1}}{2h_x} \\ & & & & 0 \end{bmatrix}$$

From this it is clear that $\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A}$ will only be an M-matrix when $\theta = 0$, which implies that a simple stability proof under the maximum norm, as done in section 3.5.2, is not possible. To obtain stability under the maximum norm we must make use of a well known first order finite difference method called upwinding.

3.7.4 θ-method for an upwind scheme

An upwind scheme uses backward difference approximations for the spacial derivative if $d(x)$ is negative and forward difference approximations if $d(x)$ is positive. For the case when $d(x)$ is positive the θ-method becomes

$$\Delta_t^+ u_i^k = \theta d_i \Delta_x^+ u_i^{k+1} + (1-\theta) d_i \Delta_x^+ u_i^k \tag{3.44}$$

which can be rewritten as

$$(2\theta d_i \lambda_x + 1) u_i^{k+1} - 2\theta d_i \lambda_x u_{i+1}^{k+1} = (-2(1-\theta) d_i \lambda_x + 1) u_i^{k+1} + 2(1-\theta) d_i \lambda_x u_{i+1}^{k+1}$$

for $i = 1,\, 2,\, \dots,\, m-1$ and $k = 0,\, 1,\, \dots,\, l-1$. Similarly for the case when $d(x)$ is negative the θ-method is given by

$$\Delta_t^+ u_i^k = \theta d_i \Delta_x^- u_i^{k+1} + (1-\theta) d_i \Delta_x^- u_i^k \tag{3.45}$$

which can be rewritten as

$$2\theta d_i \lambda_x u_{i-1}^{k+1} + (1 - 2\theta d_i \lambda_x) u_i^{k+1} = -2(1-\theta) d_i \lambda_x u_{i-1}^{k+1} + (2(1-\theta) d_i \lambda_x + 1) u_i^{k+1}$$

for $i = 1,\, 2,\, \dots,\, m-1$ and $k = 0,\, 1,\, \dots,\, l-1$.

3.7.5 Consistency

The truncation error of the θ-method for the case when $d(x)$ is positive can be written as

$$\begin{aligned} L_{h_x}^{\Delta t} v(x_i, \tau_k) = \Delta_t^+ v(x_i, \tau_k) &- \theta d(x_i) \Delta_x^+ v(x_i, \tau_{k+1}) \\ &- (1-\theta) d(x_i) \Delta_x^+ v(x_i, \tau_k) \end{aligned} \tag{3.46}$$

To compute the truncation error of the θ-method applied to the one dimensional convection equation we consider Taylor expansions about $(x_i, \tau_{k+\frac{1}{2}})$. Rearranging (3.5) and applying the equation to both time levels τ_k and τ_{k+1} yields

$$\begin{aligned} \Delta_x^+ v(x_i, \tau_k) &= \frac{\partial v}{\partial x}(x_i, \tau_k) + O(h_x) \\ \Delta_x^+ v(x_i, \tau_{k+1}) &= \frac{\partial v}{\partial x}(x_i, \tau_{k+1}) + O(h_x). \end{aligned}$$

Expanding the terms on the right about $(x_i, \tau_{k+\frac{1}{2}})$ yields

$$\begin{aligned} \Delta_x^+ v(x_i, \tau_k) &= \left(\frac{\partial v}{\partial x}(x_i, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial t \partial x}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial x \partial \tau^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right) + O(h_x) \\ \Delta_x^+ v(x_i, \tau_{k+1}) &= \left(\frac{\partial v}{\partial x}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial x \partial \tau^2}(x_i, \tau_{k+\frac{1}{2}}) + \dots \right) + O(h_x). \end{aligned}$$

By substituting the equations above in (3.46) we obtain

$$\begin{aligned}
L_{h_x}^{\Delta t}v(x_i,\tau_k) &= \frac{\partial v}{\partial \tau}(x_i,\tau_{k+\frac{1}{2}}) + \frac{1}{24}\Delta t^2\frac{\partial^3 v}{\partial \tau^3}(x_i,\tau_{k+\frac{1}{2}}) + \dots \\
&\quad - d(x_i)\left[\frac{\partial v}{\partial x}(x_i,\tau_{k+\frac{1}{2}}) + (2\theta-1)\tfrac{1}{2}\Delta t\frac{\partial^2 v}{\partial \tau\partial x}(x_i,\tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2\frac{\partial^3 v}{\partial \tau^2\partial x}(x_i,\tau_{k+\frac{1}{2}}) + \dots\right] \\
&\quad + O(h_x) \\
&= \left(\frac{\partial v}{\partial \tau}(x_i,\tau_{k+\frac{1}{2}}) - d(x_i)\frac{\partial v}{\partial x}(x_i,\tau_{k+\frac{1}{2}})\right) \\
&\quad - d(x_i)(2\theta-1)\tfrac{1}{2}\Delta t\frac{\partial^3 v}{\partial \tau^2\partial x}(x_i,\tau_{k+1}) + O(\Delta t^2) + O(h_x) \\
&= -d(x_i)(2\theta-1)\tfrac{1}{2}\Delta t\frac{\partial^3 v}{\partial \tau^2\partial x}(x_i,\tau_{k+1}) + O(\Delta t^2) + O(h_x) \qquad (3.47)
\end{aligned}$$

where we used the fact that v is a solution of (3.11). From (3.47) it is clear that $L_{h_x}^{\Delta t}v(x_i,\tau_k) \to 0$ as $h_x, \Delta t \to 0$ for all $\theta \in [0,1]$, hence we conclude that the θ-method is consistent. Note that this scheme is first order accurate independent of the choice of θ. A similar proof for consistency can by given for the case when $d(x) < 0$.

3.7.6 Stability

In this subsection we will find conditions under which the upwind θ-method is stable.

von Neumann stability analysis

Again assuming constant coefficients we can apply von Neumann analysis to find necessary conditions for stability. Substitute (3.25) in (3.9) and (3.10) to obtain

$$\begin{aligned}
\Delta_x^+ u_i^k &= \frac{1}{h_x}\gamma^k e^{I\beta i h_x}(e^{I\beta h_x}-1) \\
\Delta_x^- u_i^k &= \frac{1}{h_x}\gamma^k e^{I\beta i h_x}(1-e^{I\beta h_x})
\end{aligned}$$

Substituting into (3.44) and (3.45) respectively yields

$$\begin{aligned}
\gamma &= \frac{1+2(1-\theta)d_i\lambda_x(e^{I\beta h_x}-1)}{1-2\theta d_i\lambda_x(e^{I\beta h_x}-1)} \\
\gamma &= \frac{1-2(1-\theta)d_i\lambda_x(e^{I\beta h_x}-1)}{1+2\theta d_i\lambda_x(e^{I\beta h_x}-1)}
\end{aligned}$$

For the case when $d(x) > 0$ we have

$$|\gamma|^2 = \frac{[1+2(1-\theta)d_i\lambda_x(\cos(\beta h_x)-1)]^2 + [2(1-\theta)d_i\lambda_x\sin(\beta h_x)]^2}{[1-2\theta d_i\lambda_x(\cos(\beta h_x)-1)]^2 + [2\theta d_i\lambda_x\sin(\beta h_x)]^2}$$

From the fact that $\cos(\theta)-1 \le 0$ it is easy to see that $|\gamma|^2 \le 1$ if $\theta \ge \frac{1}{2}$. Hence we can deduce that the scheme is von Neumann stable for $\theta \ge \frac{1}{2}$. Similarly when $d(x) < 0$ we have

$$|\gamma|^2 = \frac{[1-2(1-\theta)d_i\lambda_x(1-cos(\beta h_x))]^2 + [2(1-\theta)d_i\lambda_x\sin(\beta h_x)]^2}{[1+2\theta d_i\lambda_x(1-\cos(\beta h_x))]^2 + [2\theta d_i\lambda_x\sin(\beta h_x)]^2}.$$

Following the same arguments we will be able to deduce that the θ-method is von Neumann stable when $\theta \ge \frac{1}{2}$.

Maximum norm analysis

The following proof for stability under the maximum norm will have the same general structure as the previous proofs given for stability under the maximum norm. The θ-method for the upwind scheme can be written in matrix form as

$$(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})\mathbf{u}^{k+1} = (\mathbf{I}_{\text{ex}} + (1-\theta)\Delta t\mathbf{A})\mathbf{u}^{k+1} + \mathbf{b}$$

for $k = 1,\ 2,\ \ldots,\ l-1$. The matrix $\mathbf{A}$ is given by,

$$\mathbf{A} = \begin{bmatrix} 0 & & & & \\ & -\frac{d_1}{h_x} & \frac{d_1}{h_x} & & \\ & & \ddots & & \\ & & & -\frac{d_{m-1}}{h_x} & \frac{d_{m-1}}{h_x} \\ & & & & 0 \end{bmatrix}$$

when $d(x) > 0$ and by

$$\mathbf{A} = \begin{bmatrix} 0 & & & & \\ -\frac{d_1}{h_x} & \frac{d_1}{h_x} & & & \\ & & \ddots & & \\ & & -\frac{d_{m-1}}{h_x} & \frac{d_{m-1}}{h_x} & \\ & & & & 0 \end{bmatrix}$$

for the case when $d(x) < 0$. From this it is clear that $\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A}$ is an M-matrix, hence we can apply definition 3.5.1 with $\mathbf{x} = (1,\ldots,1)^T$ to deduce that

$$||(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})^{-1}||_\infty \leq 1$$

Thus

$$\begin{aligned} ||(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})^{-1}(\mathbf{I}_{\text{im}} + (1-\theta)\Delta t\mathbf{A})||_\infty &\leq ||(\mathbf{I}_{\text{im}} - \theta\Delta t\mathbf{A})^{-1}||_\infty||(\mathbf{I}_{\text{im}} + (1-\theta)\Delta t\mathbf{A})||_\infty \\ &\leq ||(\mathbf{I}_{\text{im}} + (1-\theta)\Delta t\mathbf{A})||_\infty \\ &= \max_i\left\{\left|-(1-\theta)\text{sgn}(d_i)d_i\frac{\Delta t}{h_x} + 1\right| + (1-\theta)\text{sgn}(d_i)d_i\frac{\Delta t}{h_x}, 1\right\} \\ &= 1 \end{aligned}$$

where the last equality is true if and only if

$$\frac{\text{sgn}(d_i)d_i\Delta t}{h_x} \leq \frac{1}{1-\theta} \tag{3.48}$$

From this it is clear that the upwind θ-method is stable under the maximum norm when (3.48) holds. From (3.48) we see that unconditional stability only occurs when we set $\theta = 1$. Also note that when $\theta = 1$ the scheme will only be of first order, the following sub section gives some fundamental results on this.

Convergence results

Definition 3.7.1. *(FDM of positive type, Duffy [2006]) A finite difference scheme of the form*

$$\mathbf{u}^{k+1} = \mathbf{B}\mathbf{u}^{k} + \mathbf{b} \tag{3.49}$$

is said to be of positive type if the matrix $\mathbf{B}$ is non-negative.

The following theorem states that finite difference approximations of positive type that approximates (3.39) have a maximum order of convergence of one, see Duffy [2006],

Theorem 3.7.1. *(Order of convergence for convection approximations, Duffy [2006]) If the scheme* (3.49) *is consistent with* (3.39) *and is of positive type, then it is of order 1 or ∞.*

This theorem shows us that we will not be able to find schemes that are both stable under the maximum norm and with an higher order of convergence than the one we have obtained.

3.7.7 Convergence

Similarly as in section 3.6 we simply use the Lax equivalence theorem to deduce that the θ-method is convergent when it is stable[2].

[2] Since it is unconditionally consistent we only need stability to ensure convergence.

Chapter 4

Alternative Approaches for the One Dimensional FDM

In this chapter we introduce an alternative approach for the finite difference method by making use of parabolic equations in one space variable. Most of the results of this chapter, excluding section 4.4, are based on [Smith, 1985, Chapter 3]. In Smith [1985] only numerical methods that approximate the solution of the simple one dimensional heat equation are considered. We will attempt to obtain accurate approximations for the solution, $u : \Omega \to \mathbb{R}$, of the following convection diffusion equation

$$\frac{\partial u}{\partial \tau} = a\frac{\partial^2 u}{\partial x^2} + d\frac{\partial u}{\partial x} \tag{4.1}$$

on the domain $(x, \tau) \in \Omega$ where $a > 0$ and d are a constants and $\Omega = [x_{\text{min}}, x_{\text{max}}] \times \mathbb{R}^+$. For the problem to be well posed it must satisfy an initial condition $u(x, 0) = \Psi(x)$ and Dirichlet boundary conditions

$$u(x_{\text{min}}, \tau) = \alpha \qquad \text{and} \qquad u(x_{\text{max}}, \tau) = \beta.$$

4.1 Reduction to a system of ordinary differential equations

To obtain a system of ODEs we only discretisize Ω in the spatial direction and keep the time axis continuous. Consider the following partition of $[x_{\text{min}}, x_{\text{max}}]$

$$x_{\text{min}} = x_0 < x_1 < \ldots < x_m = x_{\text{max}}.$$

Truncate the domain Ω such that $(x, \tau) \in \overline{\Omega} = [x_{\text{min}}, x_{\text{max}}] \times [0, T]$. The semi-discrete mesh is then given by

$$\widehat{\Omega} = \{(x_i, \tau) | i = 0,\ 1,\ \ldots,\ m, \tau \in [0, T]\}$$

By making use of the second order approximations derived in section 3.1 we can rewrite (4.1) in semi-discrete form as

$$\frac{du(x_i, \tau)}{d\tau} = \left(\frac{a}{h^2} - \frac{d}{2h}\right) u(x_{i\,-\,1}, \tau) - \frac{2a}{h^2} u(x_i, \tau) + \left(\frac{a}{h^2} + \frac{d}{2h}\right) u(x_{i\,+\,1}, \tau) + O(h^2) \tag{4.2}$$

Let the approximations on the *lines* of $\widehat{\Omega}$ be denoted by $u_i(\tau) \approx u(x_i, \tau)$. Incorporating the boundary conditions allows us to approximate (4.2) by

$$\begin{aligned} u_0(\tau) &= \alpha \\ \frac{du_i(\tau)}{d\tau} &= \left(\frac{a}{h^2} - \frac{d}{2h}\right) u_{i-1}(\tau) - \frac{2a}{h^2} u_i(\tau) + \left(\frac{a}{h^2} + \frac{d}{2h}\right) u_{i+1}(\tau), \quad \text{for} \quad i = 1,\ 2,\ \ldots,\ m-1 \\ u_m(\tau) &= \beta. \end{aligned}$$

Which can be written in matrix form as

$$\frac{d\mathbf{u}(\tau)}{d\tau} = \mathbf{A}\mathbf{u}(\tau) + \mathbf{b} \tag{4.3}$$

where $\mathbf{u} = (u_1(\tau), \ldots, u_{m-1}(\tau))^T$ and the boundary vector, $\mathbf{b}$, is given by

$$\mathbf{b} = \left(\left(\frac{a}{h^2} - \frac{d}{2h}\right)\alpha, 0, \ldots, 0, \left(\frac{a}{h^2} + \frac{d}{2h}\right)\beta\right)^T.$$

The time homogenous matrix, $\mathbf{A}$, is defined as follows

$$\mathbf{A} = \begin{bmatrix} -2\frac{a}{h_x^2} & \frac{a}{h_x^2} + \frac{d}{2h_x} & & & \\ \frac{a}{h_x^2} - \frac{d}{2h_x} & -2\frac{a}{h_x^2} & \frac{a}{h_x^2} + \frac{d}{2h_x} & & \\ & & \ddots & & \\ & & \frac{a}{h_x^2} - \frac{d}{2h_x} & -2\frac{a}{h_x^2} & \frac{a}{h_x^2} + \frac{d}{2h_x} \\ & & & \frac{a}{h_x^2} - \frac{d}{2h_x} & -2\frac{a}{h_x^2} \end{bmatrix}.$$

4.1.1 The solution of $\frac{d\mathbf{u}(\tau)}{d\tau} = \mathbf{A}\mathbf{u}(\tau) + \mathbf{b}$

The exponential of a real valued square matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is defined as

$$\exp(\mathbf{Q}) := \sum_{k=0}^{\infty} \frac{\mathbf{Q}^k}{k!} \tag{4.4}$$

where $\mathbf{Q}^0 = \mathbf{I}$ and $\mathbf{I}$ the identity matrix[1]. By making use of (4.4) it is easy to see that

$$e^{\mathbf{Q}} e^{-\mathbf{Q}} = \mathbf{I} = e^{-\mathbf{Q}} e^{\mathbf{Q}}. \tag{4.5}$$

From equation (4.5) it follows that

$$e^{-\mathbf{Q}} = (e^{\mathbf{Q}})^{-1}.$$

It is well known that the solution of

$$\frac{du(\tau)}{d\tau} = Au(\tau) + b$$

where A and b are constants, together with the initial condition $u(0) = \Psi$, is given by

$$u(\tau) = -\frac{b}{A} + e^{\tau A}\left(\Psi + \frac{b}{A}\right).$$

[1] Let M be a real number such that $|\mathbf{Q}_{i,j}| < M$ for all i and j. Note that $|(\mathbf{Q}^k)_{i,j}| < n^k M_{k+1}$ where $(\mathbf{Q}^k)_{i,j}$ denotes the ith row and jth column of $\mathbf{Q}^k$. From that fact that $\sum_{k=0}^{\infty} \frac{n^k M^{k+1}}{k!}$ converges it follows that $\exp(\mathbf{Q})$ converges to a real valued $n \times n$ matrix, see McLeman [2006].

This can be used as a motivation for the proposition of

$$\mathbf{u}(\tau) = -\mathbf{A}^{-1}\mathbf{b} + e^{\tau\mathbf{A}}\left(\mathbf{\Psi} + \mathbf{A}^{-1}\mathbf{b}\right) \tag{4.6}$$

as a solution of (4.3). This equation clearly satisfies the initial condition $\mathbf{u}(0) = \mathbf{\Psi}$. By differentiation we obtain

$$\begin{aligned}\frac{d\mathbf{u}(\tau)}{d\tau} &= \mathbf{A}e^{\tau\mathbf{A}}\left(\mathbf{\Psi} + \mathbf{A}^{-1}\mathbf{b}\right)\\ &= \mathbf{A}\mathbf{u}(\tau) + \mathbf{b}\end{aligned}$$

which shows that equation (4.6) is the solution of (4.3). The last equation follows directly from (4.6).

4.2 Alternative derivation of classical one dimensional finite difference schemes

From the previous section it follows that an approximate solution of (4.1) together with the initial and boundary conditions is given by (4.6). Consider the following partition of $[0, T]$

$$0 = \tau_0 < \tau_1 < \ldots < \tau_l = T$$

where $\tau_k = k\Delta t$ and $\Delta t = \frac{T}{l}$. If the solution at time τ is known then the solution at time $\tau + \Delta t$ can be obtained with

$$\begin{aligned}\mathbf{u}(\tau + \Delta t) &= -\mathbf{A}^{-1}\mathbf{b} + e^{\Delta t\mathbf{A}}e^{\tau\mathbf{A}}\left(\mathbf{\Psi} + \mathbf{A}^{-1}\mathbf{b}\right)\\ &= -\mathbf{A}^{-1}\mathbf{b} + e^{\Delta t\mathbf{A}}\left(\mathbf{u}(\tau) + \mathbf{A}^{-1}\mathbf{b}\right) \qquad (4.7)\end{aligned}$$

where we made use of the fact that $e^{\Delta t\mathbf{A}}e^{t\mathbf{A}} = e^{(t+\Delta t)\mathbf{A}}$. From the definition of the exponential of a matrix it follows that

$$e^{\Delta t\mathbf{A}} = \mathbf{I} + \Delta t\mathbf{A} + \tfrac{1}{2}\Delta t^2\mathbf{A}^2 + \tfrac{1}{3!}\Delta t^3\mathbf{A}^3 + \ldots$$

Since we cannot afford to add an infinite amount of matrices we need to find an approximation of $e^{\Delta t\mathbf{A}}$ in order to implement this scheme. The higher the order of approximation the more accurate, in time, the scheme will be. A trivial approximation is given by $e^{\Delta t\mathbf{A}} \approx \mathbf{I} + \Delta t\mathbf{A}$, with an error term of $O(\Delta t^2)$. The finite difference scheme becomes

$$\begin{aligned}\mathbf{u}^{k+1} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + \Delta t\mathbf{A}\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right)\\ &= (\mathbf{I} + \Delta t\mathbf{A})\mathbf{u}^k + \Delta t\mathbf{b}\end{aligned}$$

which is the fully explicit scheme discussed in chapter 3. Different approximations of $e^{\Delta t\mathbf{A}}$ will give rise to different finite difference schemes, in particular fully implicit and the Crank-Nicolson scheme. To derive these schemes we make use of Padé's rational functions that approximate $e^{\Delta t\mathbf{A}}$.

4.2.1 The Padé approximations of e^θ, $\theta \in \mathbb{R}$

It is possible to approximate e^θ by

$$e^\theta = \frac{\sum_{i\,=\,0}^{T} p_i\theta^i}{\sum_{i\,=\,0}^{S} q_i\theta^i} + c_{S\,+\,T\,+\,1}\theta^{S+T+1} + O(\theta^{S+T+2})$$

where $c_{S\,+\,T\,+\,1}$, p_i and q_i are constants for all i, see [Smith, 1985, Chapter 3]. The rational function

$$R_{S,\,T}(\theta) = \frac{\sum_{i\,=\,0}^{T} p_i\theta^i}{\sum_{i\,=\,0}^{S} q_i\theta^i} = \frac{P_T(\theta)}{Q_S(\theta)}$$

is called the (S, T) Padé approximation of e^θ and is of $O(S + T)$. As an example consider $S = T = 1$

$$e^\theta = \frac{1 + p_1\theta}{1 + q_1\theta} + c_3\theta^3 + O(\theta^4). \tag{4.8}$$

The unknown constants p_1, q_1 and c_3 can be obtained by making use of the Taylor expansion of e^θ

$$e^\theta = 1 + \theta + \tfrac{1}{2}\theta^2 + \tfrac{1}{3!}\theta^3 + O(\theta^4). \tag{4.9}$$

Substituting (4.8) in (4.9) and rearranging results in

$$(1 + q_1\theta)(1 + \theta + \tfrac{1}{2}\theta^2 + \tfrac{1}{3!}\theta^3 + O(\theta^4)) = 1 + p_1\theta + (1 + q_1\theta)(c_3\theta^3 + O(\theta^4)).$$

Group the terms to obtain

$$(1 + q_1 - p_1)\theta + (\tfrac{1}{2} + q_1)\theta^2 + (\tfrac{1}{3!} + \tfrac{1}{2}q_1 - c_3)\theta^3 + O(\theta^4) = 0.$$

It is easy to see that this is uniquely satisfied, with an error term of $O(\theta^4)$, by

$$p_1 = \tfrac{1}{2}, \quad q_1 = -\tfrac{1}{2} \quad \text{and} \quad c_3 = -\tfrac{1}{12}.$$

Thus the $(1, 1)$ Padé approximation of e^θ is given by

$$e^\theta = \frac{1 + \frac{1}{2}\theta}{1 - \frac{1}{2}\theta} - \tfrac{1}{12}\theta^3 + O(\theta^4).$$

The major Padé approximations are given by,

(S, T)	$R_{S,\,T}(\theta)$	Principal error term
$(0, 1)$	$1 + \theta$	$\frac{1}{2}\theta^2$
$(1, 0)$	$\frac{1}{1-\theta}$	$-\frac{1}{2}\theta^2$
$(1, 1)$	$\frac{1+\frac{1}{2}\theta}{1-\frac{1}{2}\theta}$	$-\frac{1}{12}\theta^3$

4.2.2 Classical FDMs via Padé approximations

We have already shown how the $(0, 1)$ Padé approximation leads to the fully explicit scheme in the introduction of this section. By approximating $e^{\Delta t\mathbf{A}}$ in equation (4.7) with the $(1, 0)$ Padé approximation we obtain

$$\begin{aligned} \mathbf{u}^{k\,+\,1} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - \Delta t\mathbf{A})^{-1}(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}) \\ \Rightarrow (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k\,+\,1} &= \mathbf{u}^k + \Delta t\mathbf{b} \end{aligned}$$

which is the fully implicit method discussed in chapter 3.
We can obtain the Crank-Nicolson scheme by making use of the $(1,1)$ Padé approximation

$$\begin{aligned} \mathbf{u}^{k+1} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - \tfrac{1}{2}\Delta t\mathbf{A})^{-1}(\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A})\,(\mathbf{u}^{k} + \mathbf{A}^{-1}\mathbf{b}) \\ \Rightarrow (\mathbf{I} - \tfrac{1}{2}\Delta t\mathbf{A})\mathbf{u}^{k+1} &= (\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A})\mathbf{u}^{k} + \Delta t\mathbf{b}. \end{aligned}$$

4.3 A_0-stability, L_0-stability and the symbol of the method

To determine the stability of the finite difference scheme we investigate the propagation of a perturbation made on the initial data from Ψ to $\widehat{\Psi}$ at time τ_0. Let the perturbed vector be given by $\widehat{\mathbf{u}}(\tau)$. Define the error vector by $\mathbf{e}(\tau)$. At time $\tau + \Delta t$ we have

$$\begin{aligned} \mathbf{e}(\tau+\Delta t) &:= \widehat{\mathbf{u}}(\tau+\Delta t) - \mathbf{u}(\tau+\Delta t) \\ &= e^{\Delta t\mathbf{A}}\mathbf{e}(\tau). \end{aligned}$$

where the last equation follows from (4.7). Let $e^{\Delta t\mathbf{A}}$ be approximated with the (S,T) Padé approximation. If the perturbation vector is known at time τ_k, we can obtain the approximation of the perturbation vector at time τ_{k+1} as follows

$$\mathbf{e}^{k+1} = R_{S,T}(\Delta t\mathbf{A})\mathbf{e}^{k}$$

which can be written in terms of the initial data as

$$\mathbf{e}^{k+1} = [R_{S,T}(\Delta t\mathbf{A})]^{k}\mathbf{e}^{0}. \tag{4.10}$$

We state the following useful lemma without proof,

Lemma 1. *(Hulley and Lotter [2004]) The eigenvalues of an tri-diagonal $m \times m$ matrix*

$$\begin{pmatrix} a & b & & & & \\ c & a & b & & & \\ & c & a & b & & \\ & & & \cdots & & \\ & & & c & a & b \\ & & & & c & a \end{pmatrix}$$

where a, b and c may be real or complex, are given by

$$\lambda_s = a + 2\sqrt{bc}\cos\left(\frac{s\pi}{m+1}\right), \quad \textit{for} \quad s = 1,\ 2,\ \ldots,\ m$$

The eigenvector, $\mathbf{v}_s$, is given by

$$\mathbf{v}_s = \left[\left(\frac{c}{b}\right)^{\frac{1}{2}}\sin\left(\frac{s\pi}{m+1}\right), \frac{c}{b}\sin\left(\frac{2s\pi}{m+1}\right), \ldots, \left(\frac{c}{b}\right)^{\frac{m}{2}}\sin\left(\frac{ms\pi}{m+1}\right)\right]^{T}$$

for $s = 1,\ 2,\ \ldots,\ m$.

Using lemma 1 it follows that the eigenvalues of the $(m-1) \times (m-1)$ matrix $\mathbf{A}$ are given by

$$\lambda_s = -\frac{2a}{h_x^2} + 2\sqrt{\left(\frac{a}{h_x^2} - \frac{d}{2h_x}\right)\left(\frac{a}{h_x^2} + \frac{d}{2h_x}\right)} \cos\left(\frac{s\pi}{m}\right) \tag{4.11}$$

for $s = 1,\ 2,\ \ldots,\ m-1$, and are distinct. Proofs of the following theorems can be found in Kreyszig [1999]

Theorem 4.3.1. *(Kreyszig [1999]) Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be distinct eigenvalues of a $m \times m$ matrix. Then the corresponding eigenvectors are linearly independent.*

Theorem 4.3.2. *(Kreyszig [1999]) If an $m \times m$ matrix $\mathbf{A}$ has m distinct eigenvalues, then $\mathbf{A}$ has a basis of eigenvectors for $\mathbb{C}^m$ (or $\mathbb{R}^m$).*

From these theorems it follows that the initial error vector, $\mathbf{e}^0 = \widehat{\mathbf{\Psi}} - \mathbf{\Psi}$, can be written as an linear combination of eigenvectors

$$\mathbf{e}^0 = \sum_{s=1}^{m-1} c_s \mathbf{v}_s$$

where c_s are constants and $\mathbf{v}_s$ are the eigenvectors of $\mathbf{A}$ for $s = 1,\ 2,\ \ldots,\ m-1$. Substituting this into (4.10) results in

$$\begin{aligned}
\mathbf{e}^{k+1} &= [R_{S,\,T}(\Delta t\mathbf{A})]^k \sum_{s=1}^{m-1} c_s \mathbf{v}_s \\
&= \sum_{s=1}^{m-1} c_s [R_{S,\,T}(\Delta t\mathbf{A})]^k \mathbf{v}_s \\
&= \sum_{s=1}^{m-1} c_s [R_{S,\,T}(\Delta t\lambda_s)]^k \mathbf{v}_s
\end{aligned}$$

where the last equation follows from the fact that $\mathbf{A}\mathbf{v}_s = \lambda_s \mathbf{v}_s$ and that $f(\mathbf{A})\mathbf{v}_s = f(\lambda_s)\mathbf{v}_s$, where the function f is a rational function. From this equation it is clear that the perturbation will diminish if and only if $|R_{S,\,T}(\Delta t\lambda_s)| < 1$ when $k \to \infty$. In Gourlay and Morris [1980] and Smith [1985] $R_{S,\,T}(-z)$, where $z = -\Delta t\lambda_s$, is defined to be the symbol of the method.

Definition 4.3.1. *(A_α-stable, [Smith, 1985, Chapter 3]) If the eigenvalues of $\mathbf{A}$ are within the wedge defined by $\pi + \alpha < \arg \lambda_s < \pi - \alpha$, $\alpha \in (0, \frac{\pi}{2})$, then the method is A_α-stable when, within this wedge, $|R_{S,\,T}(\Delta t\lambda_s)| < 1$ for all $s = 1,\ 2,\ \ldots,\ m-1$, $\Delta t > 0$ and $h_x > 0$.*

When all the eigenvalues of the matrix are real and $|R_{S,\,T}(\Delta t\lambda_s)| < 1$ for all $s = 1,\ 2,\ \ldots,\ m-1$, $\Delta t > 0$ and $h_x > 0$, then the scheme is A_0-stable. For a scheme that is A_0-stable it is possible that $R_{S,\,T}(\Delta t\lambda_s)$ is close to -1 for particular s, say $\widehat{s}$. If $c_{\widehat{s}}$ is large then the numerical solution may oscillate finitely as k increases. These oscillations can be avoided by ensuring that the scheme is L_0-stable, see Gourlay and Morris [1980], Smith [1985] and Khaliq and Twizell [1986].

Definition 4.3.2. *(L_0-stable, [Smith, 1985, Chapter 3]) A finite difference scheme is L_0-stable when*

$$\max_{z \geq 0} |R_{S,\,T}(-z)| < 1 \quad \textit{and} \quad \lim_{z \to \infty} R_{S,\,T}(-z) = 0$$

where $z = -\Delta t\lambda_s \in \mathbb{R}^+ \backslash \{0\}$.

A similar definition of L_0-stability is given in Cash [1984],

Definition 4.3.3. *(L_0-stable, Cash [1984]) Suppose a numerical integration method is applied to* (4.3) *gives a numerical approximation of the form*

$$\mathbf{u}^{k+1} = -\mathbf{A}^{-1}\mathbf{b} + R(\Delta t\mathbf{A})\,(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b})$$

where R is a rational function of $\Delta t\mathbf{A}$. Suppose further

$$\max_{z\,\geq\,0}|R_{S,\,T}(-z)| < 1 \quad \textit{and} \quad \lim_{z\,\to\,\infty} R_{S,\,T}(-z) = 0$$

where $z = -\Delta t\lambda_s \in \mathbb{R}^+\backslash\{0\}$. Then the numerical method is said to be L_0-stable.

Unwanted oscillations are damped much faster by a L_0-stable scheme than an A_0-stable scheme damps oscillations, hence L_0-stability is preferred to A_0-stability. These definitions are based on a definition of A-stability posed in Dahlquist [1963]

Definition 4.3.4. *(L_0-stable, Bui [1979]) A method to solve the following system of ODEs*

$$\frac{\partial\mathbf{u}}{\partial\tau} = f(\mathbf{u}),$$
$$\mathbf{u}(0) = \mathbf{\Psi}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$, is called A-stable in the sense of Dahlquist *if and only if $|\mathbf{u}^{k+1}| = c|\mathbf{u}^k|$ when the method is applied with any positive step size Δt to the test function*

$$\frac{d\mathbf{u}}{d\tau} = \lambda\mathbf{u},$$

where λ is a complex constant with negative real part and $c \leq 1$ is a real constant.

Definition 4.3.5. *(L-stable, Bui [1979]) A method is called L-stable if it is A-stable and $c \to 0$ as $\lambda\Delta t \to -\infty$.*

By applying the numerical method, described in the previous sections, to (4.3) we can use the definitions above to deduce that the scheme will be A-stable if $|R_{S,\,T}(\Delta t\lambda)| < 1$, where $\mathrm{Re}(\lambda) < 0$ and, L-stable if it is A-stable and $\lambda\Delta t \to -\infty$

4.3.1 Stability analysis of classical schemes

In this section we are going to consider the stability of the fully implicit and the Crank-Nicolson scheme applied to the convection diffusion equation (4.1) and the simple Heat equation.

Heat equation

The heat equation can be obtained by letting $d = 0$ in (4.1). The semi-discrete form of the heat equation is given by (4.7), where

$$\mathbf{A} = \begin{bmatrix} -2\frac{a}{h_x^2} & \frac{a}{h_x^2} & & & \\ \frac{a}{h_x^2} & -2\frac{a}{h_x^2} & \frac{a}{h_x^2} & & \\ & & \ddots & & \\ & & \frac{a}{h_x^2} & -2\frac{a}{h_x^2} & \frac{a}{h_x^2} \\ & & & \frac{a}{h_x^2} & -2\frac{a}{h_x^2} \end{bmatrix}$$

Thus when $d = 0$ the eigenvalues of $\mathbf{A}$ are given by

$$\lambda_s = -\frac{4}{h_x^2}\sin^2\left(\frac{s\pi}{2m}\right) \quad \text{for} \quad s = 1,\ 2,\ \ldots,\ m-1$$

and are real, negative and non-zero. Alternatively the sign of the eigenvalues of $\mathbf{A}$ can also by obtained by making use of the following theorem,

Theorem 4.3.3. *(Gerschgorin's circle theorem, Kreyszig [1999]) Let λ be an eigenvalue of an arbitrary $n \times n$ matrix $\mathbf{A} = (a_{i,\,j})$. Then for some integer i, $i = 1,\ 2,\ \ldots,\ n$, we have*

$$|a_{i,\,i} - \lambda| \le \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{i,\,j}|$$

Using Gerschgorin's theorem and the fact that $a > 0$ it follows that

$$-\frac{4a}{h_x^2} \le \lambda \le 0$$

for any eigenvalue, λ, of $\mathbf{A}$. From the Padé approximations we see that the symbols of the fully implicit and the Crank-Nicolson schemes are respectively given by

$$R_{1,\,0}(-z) = \frac{1}{1+z} \quad \text{and} \quad R_{1,\,1}(-z) = \frac{1-\frac{1}{2}z}{1+\frac{1}{2}z}.$$

It is easy to see that $|R_{1,\,0}(-z)| < 1$ which implies that the fully implicit scheme is unconditionally A_0-stable. From the fact that

$$\lim_{z\,\to\,\infty} R_{1,\,0}(-z) = \lim_{z\,\to\,\infty} \frac{1}{1+z} = 0$$

it follows that the fully implicit scheme is L_0-stable. The A_0-stability of the Crank-Nicolson scheme follows trivially from the fact that $|R_{1,\,1}(-z)| < 1$. Note

$$\lim_{z\,\to\,\infty} R_{1,\,1}(-z) = \lim_{z\,\to\,\infty} \frac{1-\frac{1}{2}z}{1+\frac{1}{2}z} = -1$$

hence the Crank-Nicolson scheme is not L_0-stable and oscillations might occur. At points of discontinuity in the initial data these oscillations will be pronounced, see Smith [1985].

Convection diffusion equation

If we can show that the eigenvalues of $\mathbf{A}$ ($d \ne 0$) are real, negative and non-zero then we can apply the analysis of the previous section to deduce the following table,

	A_0-stable	L_0-stable
Fully implicit	Unconditionally	Unconditionally
Crank-Nicolson	Unconditionally	Unstable

The eigenvalues of $\mathbf{A}$ are given by equation (4.11) and may be complex[2]. To keep the terms under the square root positive, and the eigenvalues real, we can implement exponential fitting. The eigenvalues become

$$\begin{aligned}\lambda_s &= -\frac{2f(a,d,h_x)}{h_x^2} + 2\sqrt{\left(\frac{f(a,d,h_x)}{h_x^2} - \frac{d}{2h_x}\right)\left(\frac{f(a,d,h_x)}{h_x^2} + \frac{d}{2h_x}\right)}\cos\left(\frac{s\pi}{m}\right)\\ &= -\frac{2f(a,d,h_x)}{h_x^2} + 2\sqrt{\left(\frac{f(a,d,h_x)}{h_x^2} - \frac{d}{2h_x}\right)\left(\frac{f(a,d,h_x)}{h_x^2} + \frac{d}{2h_x}\right)}\left(1 - 2\sin^2\left(\frac{s\pi}{m}\right)\right)\\ &\leq -\frac{2f(a,d,h_x)}{h_x^2} + 2\sqrt{\left(\frac{f(a,d,h_x)}{h_x^2} - \frac{d}{2h_x}\right)\left(\frac{f(a,d,h_x)}{h_x^2} + \frac{d}{2h_x}\right)}\\ &= -\frac{2f(a,d,h_x)}{h_x^2} + 2\sqrt{\left(\frac{f(a,d,h_x)}{h_x^2}\right)^2 - \left(\frac{d}{2h_x}\right)^2}\\ &< 0\end{aligned}$$

for $s = 1,\ 2,\ \ldots,\ m-1$, where the function f is given by (3.32). The last inequality follows from the fact that $d \neq 0$ and $f(a,d,h_x) > 0$. The negativity of the eigenvalues of $\mathbf{A}$ can also be shown by making use of Gerschgorin's theorem 4.3.3. For any eigenvalue of $\mathbf{A}$ it follows that

$$\begin{aligned}\lambda &\leq -\frac{2f(a,d,h_x)}{h_x^2} + \left|\frac{f(a,d,h_x)}{h_x^2} - \frac{d}{2h_x}\right| + \left|\frac{f(a,d,h_x)}{h_x^2} + \frac{d}{2h_x}\right|\\ \lambda &\geq -\frac{2f(a,d,h_x)}{h_x^2} - \left|\frac{f(a,d,h_x)}{h_x^2} - \frac{d}{2h_x}\right| - \left|\frac{f(a,d,h_x)}{h_x^2} + \frac{d}{2h_x}\right|.\end{aligned}$$

From a property of the fitting function (3.34) it follows that

$$-\frac{4f(a,d,h_x)}{h_x^2} \leq \lambda \leq 0.$$

4.4 Extrapolation methods

In this section we will consider extrapolation methods to increase the accuracy in time of the classical finite difference methods. Consider the L_0-stable $(1,0)$ Padé approximation (exponentially fitted fully implicit method) of (4.7)

$$\mathbf{u}^{k+1} = -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - \Delta t\mathbf{A})^{-1}(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b})$$

which can be used to obtain the unknown approximation at time τ_{k+1} if the solution at time τ_k is known[3]. In Gourlay and Morris [1980] second, third and fourth order accurate methods for the simple heat equation with homogenous Dirichlet boundary conditions are derived. We will discuss these methods for the more general case when a convection term is present and non-zero Dirichlet boundary conditions are posed.

[2] Since the terms under the square root may be negative.

[3] We make use of exponential fitting, as done in section 4.3.1, to ensure the the eigenvalues of $\mathbf{A}$ are real and non-positive.

4.4.1 Second order extrapolation scheme

Say the solution at time τ_k is known, then there are two different ways of obtaining the solution at time τ_{k+2}. We can simply set the time increment to $2\Delta t$

$$\mathbf{u}_{(1)}^{k+2} = -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 2\Delta t\mathbf{A})^{-1}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right)$$

or make two time steps, with the time increment equal to Δt

$$\begin{aligned}\mathbf{u}_{(2)}^{k+2} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - \Delta t\mathbf{A})^{-1}\left(\mathbf{u}^{k+1} + \mathbf{A}^{-1}\mathbf{b}\right)\\ &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - \Delta t\mathbf{A})^{-2}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right).\end{aligned}$$

By making use of the following Binomial expansions

$$(1+x)^{-1} = 1 - x + x^2 - x^3 + x^4 - x^5 + - \dots \tag{4.12}$$

$$(1+x)^{-2} = 1 - 2x + 3x^2 - 4x^3 + 5x^4 - 6x^5 + - \dots \tag{4.13}$$

we obtain

$$\mathbf{u}_{(1)}^{k+2} = -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} + 2\Delta t\mathbf{A} + 4\Delta t^2\mathbf{A}^2)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^3) \tag{4.14}$$

$$\mathbf{u}_{(2)}^{k+2} = -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} + 2\Delta t\mathbf{A} + 3\Delta t^2\mathbf{A}^2)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^3). \tag{4.15}$$

From the definition of the exponential of an matrix we see that

$$e^{2\Delta t\mathbf{A}} = \mathbf{I} + 2\Delta t\mathbf{A} + 2\Delta t^2\mathbf{A}^2 + O(\Delta t^3).$$

Hence the approximation to the true solution, with the correct higher order terms, at time τ_{k+2} is given by

$$\begin{aligned}\mathbf{u}^{k+2} &= -\mathbf{A}^{-1}\mathbf{b} + e^{2\Delta t\mathbf{A}}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right)\\ &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 2\Delta t\mathbf{A} + 2\Delta t^2\mathbf{A}^2 + \tfrac{4}{3}\Delta t^3\mathbf{A}^3 + \dots\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right).\end{aligned}$$

From this it is clear that neither (4.14) nor (4.15) approximates the second order term correctly. With the following linear combination

$$\mathbf{u}^{k+1} = 2\mathbf{u}_{(2)}^{k+2} - \mathbf{u}_{(1)}^{k+2}$$

we obtain a method that is second order accurate in time. The algorithm for this second order accurate extrapolated scheme is given by

$$(\mathbf{I} - 2\Delta t\mathbf{A})\mathbf{u}_{(1)}^{k+2} = \mathbf{u}^k + 2\Delta t\mathbf{b}$$

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k+1} &= \mathbf{u}^k + \Delta t\mathbf{b}\\ (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}_{(2)}^{k+2} &= \mathbf{u}^{k+1} + \Delta t\mathbf{b}\end{aligned}$$

and

$$\mathbf{u}^{k+1} = 2\mathbf{u}_{(2)}^{k+2} - \mathbf{u}_{(1)}^{k+2} \tag{4.16}$$

This scheme is also called the Lawson-Morris scheme, see Lawson and Morris [1978] and Gourlay and Morris [1980].

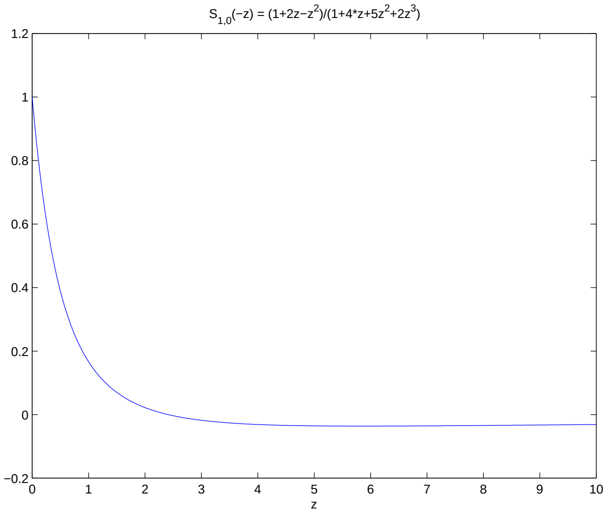


Figure 4.1: The symbol of the Lawson-Morris scheme, $S^{\text{LM}}_{1,\,0}(-z)$.

Stability of the Lawson-Morris scheme

The extrapolation step, (4.16), can be written as

$$\mathbf{u}^{k\,+\,2} = -\mathbf{A}^{-1}\mathbf{b} + S^{\text{LM}}_{1,\,0}(\Delta t\mathbf{A})\,(\mathbf{u}^{k} + \mathbf{A}^{-1}\mathbf{b})$$

where

$$S^{\text{LM}}_{1,\,0}(\Delta t\mathbf{A}) = 2(\mathbf{I} - \Delta t\mathbf{A})^{-2} - (\mathbf{I} - 2\Delta t\mathbf{A}).$$

This shows that the symbol of the Lawson-Morris scheme is given by

$$S^{\text{LM}}_{1,\,0}(-z) = \frac{2}{(1+z)^2} - \frac{1}{1+2z}$$

It is easy to see that

$$\lim_{z\,\to\,\infty} S^{\text{LM}}_{1,\,0}(-z) = 0$$

Figure 4.1 shows that $\max_{z\,\geq\,0} |S^{\text{LM}}_{1,\,0}(-z)| < 1$ and that the symbol attains small negative values for $z > 1 + \sqrt{2}$. We conclude that the Lawson-Morris scheme is L_0-stable. The negative numbers might introduce some oscillations but the effect will not be too severe since these numbers will be very small in absolute value.

4.4.2 Third order extrapolation scheme

In the previous section we showed that a second order L_0-stable scheme can be obtained by extrapolating over two time steps. Similarly we can obtain third and fourth order schemes by extrapolating over three and four time steps respectively, see Gourlay and Morris [1980]. If the solution at time τ_k is known, then there are three different ways of obtaining the solution at time τ_{k+3}. We can simply take the time increment to be $3\Delta t$

$$\mathbf{u}_{(1)}^{k+3} = -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 3\Delta t\mathbf{A})^{-1}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right),$$

march one time step with a time increment of Δt followed by a time step with the step size equal to $2\Delta t$

$$\begin{aligned}\mathbf{u}_{(2)}^{k+3} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 2\Delta t\mathbf{A})^{-1}\left(\mathbf{u}^{k+1} + \mathbf{A}^{-1}\mathbf{b}\right)\\ &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 2\Delta t\mathbf{A})^{-1}(\mathbf{I} - \Delta t\mathbf{A})^{-1}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) \qquad (4.17)\end{aligned}$$

or by making three time steps with an increment size of Δt

$$\mathbf{u}_{(3)}^{k+3} = -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - \Delta t\mathbf{A})^{-3}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right).$$

By making use of the binomial expansions in (4.12), (4.13) and the following

$$(1+x)^{-3} = 1 - 3x + 6x^2 - 10x^3 + 15x^4 - 21x^5 + - \dots \qquad (4.18)$$

we obtain

$$\begin{aligned}\mathbf{u}_{(1)}^{k+3} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 3\Delta t\mathbf{A} + 9\Delta t^2\mathbf{A}^2 + 27\Delta t^3\mathbf{A}^3\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^4) \qquad (4.19)\\ \mathbf{u}_{(2)}^{k+3} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 2\Delta t\mathbf{A} + 4\Delta t^2\mathbf{A}^2 + 8\Delta t^3\mathbf{A}^3\right)\\ &\quad\cdot\left(\mathbf{I} + \Delta t\mathbf{A} + \Delta t^2\mathbf{A}^2 + \Delta t^3\mathbf{A}^3\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^4)\\ &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 3\Delta t\mathbf{A} + 7\Delta t^2\mathbf{A}^2 + 15\Delta t^3\mathbf{A}^3\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^4) \qquad (4.20)\end{aligned}$$

and

$$\mathbf{u}_{(3)}^{k+3} = -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 3\Delta t\mathbf{A} + 6\Delta t^2\mathbf{A}^2 + 10\Delta t^3\mathbf{A}^3\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^4). \qquad (4.21)$$

From the definition of the exponential of an matrix we see that

$$e^{3\Delta t\mathbf{A}} = \mathbf{I} + 3\Delta t\mathbf{A} + \tfrac{9}{2}\Delta t^2\mathbf{A}^2 + \tfrac{9}{2}\Delta t^3\mathbf{A}^3 + O(\Delta t^4).$$

Hence the approximation of the solution, with the correct higher order terms, at time τ_{k+3} is given by

$$\begin{aligned}\mathbf{u}^{k+3} &= -\mathbf{A}^{-1}\mathbf{b} + e^{3\Delta t\mathbf{A}}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right)\\ &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 3\Delta t\mathbf{A} + \tfrac{9}{2}\Delta t^2\mathbf{A}^2 + \tfrac{9}{2}\Delta t^3\mathbf{A}^3 + \dots\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right). \qquad (4.22)\end{aligned}$$

From this it is clear that neither (4.19), (4.20) nor (4.21) approximates the second or third order terms correctly. A linear combination of (4.19), (4.20) and (4.21) will match the terms of (4.22) up to $O(\Delta t^3)$

$$\mathbf{u}^{k+3} = \eta_1\mathbf{u}_{(1)}^{k+3} + \eta_2\mathbf{u}_{(2)}^{k+3} + \eta_3\mathbf{u}_{(3)}^{k+3}$$

where η_1, η_2 and η_3 satisfies the following equations

$$\begin{aligned}\eta_1 + \eta_2 + \eta_3 &= 1\\ 9\eta_1 + 7\eta_2 + 6\eta_3 &= \tfrac{9}{2}\\ 27\eta_1 + 15\eta_2 + 10\eta_3 &= \tfrac{9}{2}.\end{aligned}$$

This system of equations has the unique solution

$$\eta_1 = 1, \quad \eta_2 = -\tfrac{9}{2} \quad \text{and} \quad \eta_3 = \tfrac{9}{2}.$$

The algorithm for this third order accurate extrapolated scheme is given by

$$(\mathbf{I} - 3\Delta t\mathbf{A})\mathbf{u}^{k\,+\,3}_{(1)} = \mathbf{u}^{k} + 3\Delta t\mathbf{b}$$

$$\begin{aligned} (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k\,+\,1} &= \mathbf{u}^{k} + \Delta t\mathbf{b} \\ (\mathbf{I} - 2\Delta t\mathbf{A})\mathbf{u}^{k\,+\,3}_{(2)} &= \mathbf{u}^{k\,+\,1} + 2\Delta t\mathbf{b} \end{aligned}$$

$$\begin{aligned} (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k\,+\,1} &= \mathbf{u}^{k} + \Delta t\mathbf{b} \\ (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k\,+\,2} &= \mathbf{u}^{k\,+\,1} + \Delta t\mathbf{b} \\ (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k\,+\,3}_{(3)} &= \mathbf{u}^{k\,+\,2} + \Delta t\mathbf{b} \end{aligned}$$

and

$$\mathbf{u}^{k\,+\,3} = \mathbf{u}^{k\,+\,3}_{(1)} - \tfrac{9}{2}\mathbf{u}^{k\,+\,3}_{(2)} + \tfrac{9}{2}\mathbf{u}^{k\,+\,3}_{(3)}. \tag{4.23}$$

In Gourlay and Morris [1980] exactly the same results are obtained. In the section that follows we will investigate stability by making use of the symbol of the method as done in Gourlay and Morris [1980].

Stability of the third order Gourlay-Morris scheme

We can rewrite the extrapolation step, (4.23), as

$$\mathbf{u}^{k\,+\,3} = -\mathbf{A}^{-1}\mathbf{b} + S^{\text{GM3}}_{1,\,0}(\Delta t\mathbf{A})\left(\mathbf{u}^{k} + \mathbf{A}^{-1}\mathbf{b}\right)$$

where

$$S^{\text{GM3}}_{1,\,0}(\Delta t\mathbf{A}) = (\mathbf{I} - 3\Delta t\mathbf{A})^{-1} - \tfrac{9}{2}(\mathbf{I} - 2\Delta t\mathbf{A})^{-1}(\mathbf{I} - \Delta t\mathbf{A})^{-1} + \tfrac{9}{2}(\mathbf{I} - \Delta t\mathbf{A})^{-3}.$$

This shows that the symbol of the third order Gourlay-Morris scheme is given by

$$S^{\text{GM3}}_{1,\,0}(-z) = \frac{1}{1+3z} - \frac{9}{2(1+2z)(1+z)} + \frac{9}{2t(1+z)^3}$$

It is easy to see that

$$\lim_{z\,\rightarrow\,\infty} S^{\text{GM3}}_{1,\,0}(-z) = 0.$$

Figure 4.2 shows that $\max_{z\,\geq\,0} |S^{\text{GM3}}_{1,\,0}(-z)| < 1$ and that the symbol is positive. We can conclude that the third order Gourlay-Morris scheme is L_0-stable.

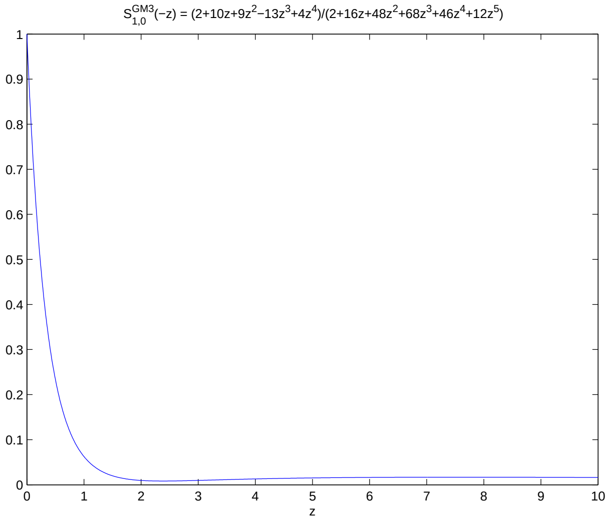


Figure 4.2: The symbol of the third order Gourlay-Morris scheme, $S^{\mathrm{GM3}}_{1,\,0}(-z)$.

4.4.3 Fourth order extrapolation scheme

If the solution at time τ_k is known, then there are five different ways of obtaining the solution at time $\tau_{k\,+\,4}$

$$\begin{aligned}
\mathbf{u}^{k\,+\,4}_{(1)} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 4\Delta t\mathbf{A})^{-1}\,(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b})\\
\mathbf{u}^{k\,+\,4}_{(2)} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 3\Delta t\mathbf{A})^{-1}(\mathbf{I} - \Delta t\mathbf{A})^{-1}\,(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b})\\
\mathbf{u}^{k\,+\,4}_{(3)} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 2\Delta t\mathbf{A})^{-2}\,(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b})\\
\mathbf{u}^{k\,+\,4}_{(4)} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - 2\Delta t\mathbf{A})^{-1}(\mathbf{I} - \Delta t\mathbf{A})^{-2}\,(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b})\\
\mathbf{u}^{k\,+\,4}_{(5)} &= -\mathbf{A}^{-1}\mathbf{b} + (\mathbf{I} - \Delta t\mathbf{A})^{-4}\,(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b})\,.
\end{aligned}$$

By making use of the binomial expansions in (4.12), (4.13) and the following expansion

$$(1+x)^{-4} = 1 - 4x + 10x^2 - 20x^3 + 35x^4 - 56x^5 + - \ldots \tag{4.24}$$

we obtain

$$\begin{aligned}
\mathbf{u}_{(1)}^{k+4} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 4\Delta t\mathbf{A} + 16\Delta t^2\mathbf{A}^2 + 64\Delta t^3\mathbf{A}^3 + 256\Delta t^4\mathbf{A}^4 + 1024\Delta t^5\mathbf{A}^5\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) \\
&\quad + O(\Delta t^6) \qquad (4.25) \\
\mathbf{u}_{(2)}^{k+4} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 3\Delta t\mathbf{A} + 9\Delta t^2\mathbf{A}^2 + 27\Delta t^3\mathbf{A}^3 + 81\Delta t^4\mathbf{A}^4 + 243\Delta t^5\mathbf{A}^5\right) \\
&\qquad \cdot \left(\mathbf{I} + \Delta t\mathbf{A} + \Delta t^2\mathbf{A}^2 + \Delta t^3\mathbf{A}^3 + \Delta t^4\mathbf{A}^4 + \Delta t^5\mathbf{A}^5\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^6) \\
&= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 4\Delta t\mathbf{A} + 13\Delta t^2\mathbf{A}^2 + 40\Delta t^3\mathbf{A}^3 + 121\Delta t^4\mathbf{A}^4 + 364\Delta t^5\mathbf{A}^5\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) \\
&\quad + O(\Delta t^6) \qquad (4.26) \\
\mathbf{u}_{(3)}^{k+4} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 4\Delta t\mathbf{A} + 12\Delta t^2\mathbf{A}^2 + 32\Delta t^3\mathbf{A}^3 + 80\Delta t^4\mathbf{A}^4 + 192\Delta t^5\mathbf{A}^5\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) \\
&\quad + O(\Delta t^6) \qquad (4.27) \\
\mathbf{u}_{(4)}^{k+4} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 2\Delta t\mathbf{A} + 4\Delta t^2\mathbf{A}^2 + 8\Delta t^3\mathbf{A}^3 + 16\Delta t^4\mathbf{A}^4 + 32\Delta t^5\mathbf{A}^5\right) \\
&\qquad \cdot \left(\mathbf{I} + 2\Delta t\mathbf{A} + 3\Delta t^2\mathbf{A}^2 + 4\Delta t^3\mathbf{A}^3 + 5\Delta t^4\mathbf{A}^4 + 6\Delta t^5\mathbf{A}^5\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) + O(\Delta t^6) \\
&= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 4\Delta t\mathbf{A} + 11\Delta t^2\mathbf{A}^2 + 26\Delta t^3\mathbf{A}^3 + 57\Delta t^4\mathbf{A}^4 + 120\Delta t^5\mathbf{A}^5\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) \\
&\quad + O(\Delta t^6) \qquad (4.28)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{u}_{(5)}^{k+4} &= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 4\Delta t\mathbf{A} + 10\Delta t^2\mathbf{A}^2 + 20\Delta t^3\mathbf{A}^3 + 35\Delta t^4\mathbf{A}^4 + 56\Delta t^5\mathbf{A}^5\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) \\
&\quad + O(\Delta t^6). \qquad (4.29)
\end{aligned}$$

By making use of the definition, of the exponential of a matrix we obtain

$$e^{4\Delta t\mathbf{A}} = \mathbf{I} + 4\Delta t\mathbf{A} + 8\Delta t^2\mathbf{A}^2 + \tfrac{32}{3}\Delta t^3\mathbf{A}^3 + \tfrac{32}{3}\Delta t^4\mathbf{A}^4 + \tfrac{128}{15}\Delta t^5\mathbf{A}^5 + O(\Delta t^6).$$

Hence the approximation of the solution, with the correct higher order terms, at time τ_{k+4} is given by

$$\begin{aligned}
\mathbf{u}^{k+4} &= -\mathbf{A}^{-1}\mathbf{b} + e^{4\Delta t\mathbf{A}}\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right) \\
&= -\mathbf{A}^{-1}\mathbf{b} + \left(\mathbf{I} + 4\Delta t\mathbf{A} + 8\Delta t^2\mathbf{A}^2 + \tfrac{32}{3}\Delta t^3\mathbf{A}^3 + \tfrac{32}{3}\Delta t^4\mathbf{A}^4 + \tfrac{128}{15}\Delta t^5\mathbf{A}^5 + \ldots\right)\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right). \qquad (4.30)
\end{aligned}$$

We want to find a linear combination of (4.25), (4.26), (4.27), (4.28) and (4.29) which matches the terms of equation (4.30) up to $O(\Delta t^4)$

$$\mathbf{u}^{k+4} = \eta_1\mathbf{u}_{(1)}^{k+4} + \eta_2\mathbf{u}_{(2)}^{k+4} + \eta_3\mathbf{u}_{(3)}^{k+4} + \eta_4\mathbf{u}_{(4)}^{k+4} + \eta_5\mathbf{u}_{(5)}^{k+4}$$

where η_1, η_2, η_3, η_4 and η_5 satisfies the following equations

$$\begin{aligned}
\eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 &= 1 \\
16\eta_1 + 13\eta_2 + 12\eta_3 + 11\eta_4 + 10\eta_5 &= 8 \\
64\eta_1 + 40\eta_2 + 32\eta_3 + 26\eta_4 + 20\eta_5 &= \tfrac{32}{3} \\
256\eta_1 + 121\eta_2 + 80\eta_3 + 57\eta_4 + 35\eta_5 &= \tfrac{32}{3}
\end{aligned}$$

Since this is a system of four equations in five unknowns it has an infinite amount of solutions. In Gourlay and Morris [1980] they propose the following solutions

	η_1	η_2	η_3	η_4	η_5
GM4.1	$-\frac{7}{9}$	$\frac{40}{9}$	0	$-\frac{32}{3}$	8
GM4.2	$-\frac{1}{9}$	$\frac{16}{9}$	-6	$\frac{16}{3}$	0
GM4.3	$\frac{1}{3}$	0	-10	16	$-\frac{16}{3}$
GM4.4	$-\frac{1}{3}$	$\frac{8}{3}$	-4	0	$\frac{8}{3}$

The algorithm for these fourth order accurate extrapolated schemes are given by

$$(\mathbf{I} - 4\Delta t\mathbf{A})\mathbf{u}_{(1)}^{k+4} = \mathbf{u}^k + 4\Delta t\mathbf{b}$$

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k+1} &= \mathbf{u}^k + \Delta t\mathbf{b}\\ (\mathbf{I} - 3\Delta t\mathbf{A})\mathbf{u}_{(2)}^{k+4} &= \mathbf{u}^{k+1} + 3\Delta t\mathbf{b}\end{aligned}$$

$$\begin{aligned}(\mathbf{I} - 2\Delta t\mathbf{A})\mathbf{u}^{k+2} &= \mathbf{u}^k + 2\Delta t\mathbf{b}\\ (\mathbf{I} - 2\Delta t\mathbf{A})\mathbf{u}_{(3)}^{k+4} &= \mathbf{u}^{k+2} + 2\Delta t\mathbf{b}\end{aligned}$$

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k+1} &= \mathbf{u}^k + \Delta t\mathbf{b}\\ (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k+2} &= \mathbf{u}^{k+1} + \Delta t\mathbf{b}\\ (\mathbf{I} - 2\Delta t\mathbf{A})\mathbf{u}_{(4)}^{k+4} &= \mathbf{u}^{k+2} + 2\Delta t\mathbf{b}\end{aligned}$$

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k+1} &= \mathbf{u}^k + \Delta t\mathbf{b}\\ (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k+2} &= \mathbf{u}^{k+1} + \Delta t\mathbf{b}\\ (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}^{k+3} &= \mathbf{u}^{k+2} + \Delta t\mathbf{b}\\ (\mathbf{I} - \Delta t\mathbf{A})\mathbf{u}_{(5)}^{k+4} &= \mathbf{u}^{k+2} + \Delta t\mathbf{b}\end{aligned}$$

and

$$\mathbf{u}^{k+4} = \eta_1\mathbf{u}_{(1)}^{k+4} + \eta_2\mathbf{u}_{(2)}^{k+4} + \eta_3\mathbf{u}_{(3)}^{k+4} + \eta_4\mathbf{u}_{(4)}^{k+4} + \eta_5\mathbf{u}_{(5)}^{k+4}. \tag{4.31}$$

In the section that follows we will investigate stability of the different types of fourth order schemes by making use of the symbol of the method as done in Gourlay and Morris [1980].

Stability of the fourth order Gourlay-Morris scheme

We can rewrite the extrapolation step, (4.31), as

$$\mathbf{u}^{k+4} = -\mathbf{A}^{-1}\mathbf{b} + S_{1,0}^{\text{GM4.x}}(\Delta t\mathbf{A})\left(\mathbf{u}^k + \mathbf{A}^{-1}\mathbf{b}\right)$$

where

$$\begin{aligned}S_{1,0}^{\text{GM4.x}}(\Delta t\mathbf{A}) = {} & \eta_1(\mathbf{I} - 4\Delta t\mathbf{A})^{-1} + \eta_2(\mathbf{I} - 3\Delta t\mathbf{A})^{-1}(\mathbf{I} - \Delta t\mathbf{A})^{-1}\\ & + \eta_3(\mathbf{I} - 2\Delta t\mathbf{A})^{-2} + \eta_4(\mathbf{I} - 2\Delta t\mathbf{A})^{-1}(\mathbf{I} - \Delta t\mathbf{A})^{-2} + \eta_5(\mathbf{I} - \Delta t\mathbf{A})^{-4}\end{aligned}$$

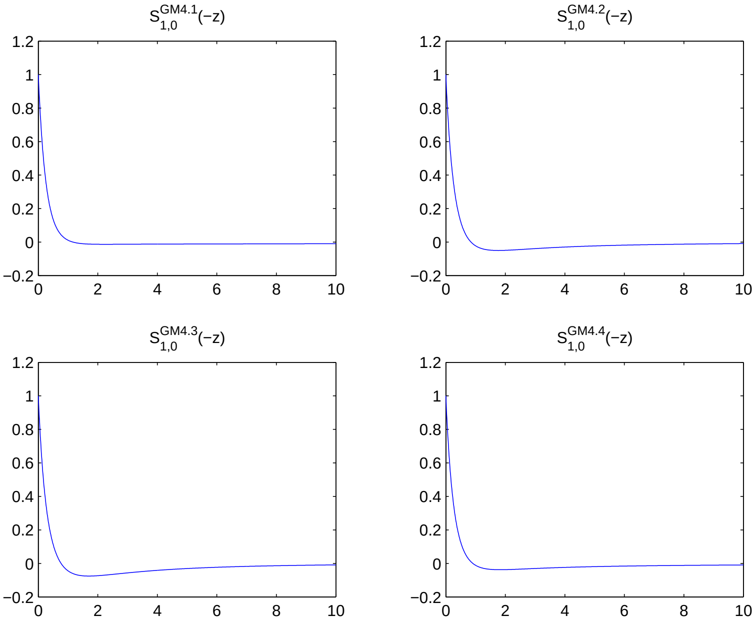


Figure 4.3: The symbols of the fourth order Gourlay-Morris schemes.

for $x = 1, 2, 3, 4$. This shows that the symbol of the fourth order Gourlay-Morris scheme is given by

$$S^{\text{GM4.x}}_{1,\,0}(-z) = \frac{\eta_1}{1+4z} + \frac{\eta_2}{(1+3z)(1+z)} + \frac{\eta_3}{(1+2z)^2} + \frac{\eta_4}{(1+2z)(1+z)^2} + \frac{\eta_5}{(1+z)^4}.$$

Clearly

$$\lim_{z\,\to\,\infty} S^{\text{GM4.x}}_{1,\,0}(-z) = 0.$$

Figure 4.3 shows that $\max_{z\,\geq\,0} |S^{\text{GM4.x}}_{1,\,0}(-z)| < 1$, $x = 1, 2, 3, 4$. We can conclude that the proposed fourth order Gourlay-Morris schemes are L_0-stable.

4.4.4 Higher order schemes

We can do even better than fourth order accuracy by simply imposing the following extra constraint on η_1, η_2, η_3, η_4 and η_5

$$1024\eta_1 + 364\eta_2 + 192\eta_3 + 120\eta_4 + 56\eta_5 = \tfrac{128}{15}$$

which allows us to match $e^{4\Delta t\mathbf{A}}$ up to $O(\Delta t^5)$. The unique solution of the system of equations is given by

$$\eta_1 = -\tfrac{17}{45}, \quad \eta_2 = \tfrac{128}{45}, \quad \eta_3 = -\tfrac{18}{5}, \quad \eta_4 = -\tfrac{16}{15} \quad \text{and} \quad \eta_5 = \tfrac{16}{5}.$$

Figure 4.4 shows that $\max_{z\,\geq\,0} |S^{\text{HO}}_{1,\,0}(-z)| < 1$. We conclude that the proposed fifth order scheme is L_0-stable. It is clear that the methodology can be extended to higher order schemes. For example if we chose to extrapolate over five time intervals we will be able to obtain a scheme of $O(\Delta t^7)$. This comes from the fact that there are seven different ways of proceeding to $\tau_{k\,+\,5}$ if the solution at τ_k is known.

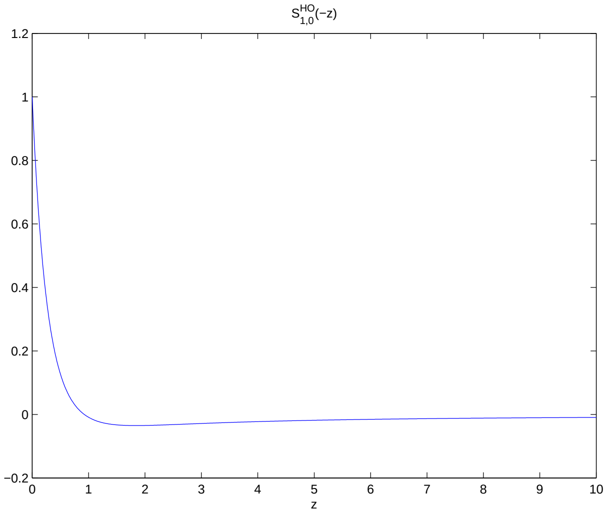


Figure 4.4: The symbol of the fifth order scheme, $S^{\mathrm{HO}}_{1,\,0}(-z)$.

Chapter 5

Two Dimensional Finite Difference Methods

In this chapter we will extend the finite difference schemes discussed in chapter 3 to two dimensions. The focus of this chapter will be on PDEs of the form

$$\frac{\partial u}{\partial \tau} = a(x,y)\frac{\partial^2 u}{\partial x^2} + 2b(x,y)\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\frac{\partial^2 u}{\partial y^2} + d(x,y)\frac{\partial u}{\partial x} + e(x,y)\frac{\partial u}{\partial y} \tag{5.1}$$

on the domain $(x, y, \tau) \in \Omega = [l_x, r_x] \times [l_y, r_y] \times \mathbb{R}^+$ where the coefficients satisfy the following inequalities

$$ac - b^2 > 0, \quad a > 0, \quad c > 0, \tag{5.2}$$

$$d \geq 0, \quad e(x, l_y) \geq 0 \quad \text{and} \quad e(x, r_y) \leq 0 \tag{5.3}$$

and the following equalities

$$a(x, l_y) = b(x, l_y) = c(x, l_y) = 0. \tag{5.4}$$

The inequalities in (5.2) must hold for the PDE to be parabolic. The inequalities in (5.3) and the equalities in (5.4) are general enough to allow for all PDEs that arise in stochastic volatility models. For the problem to be well posed an initial condition, $u(x, y, 0) = \Psi(x, y)$, and four boundary conditions need to be specified. A Dirichlet condition will be used for the boundary at $x = l_x$ and a von Neumann condition will be used at $x = r_x$

$$\begin{aligned} u(l_x, y, \tau) &= c_l \\ \frac{\partial u}{\partial x}(r_x, y, \tau) &= c_r. \end{aligned}$$

Neither a Dirichlet nor a von Neumann condition is used as a boundary condition in the y-direction, the only requirement is that the PDE itself must be solved on the boundaries, see Zvan et al. [2003] and

Duffy [2006]. Using equation (5.4) and the assumption that $\frac{\partial u}{\partial y}(x, r_y) = 0$[1] we obtain

$$\begin{aligned}\frac{\partial u}{\partial \tau} &= d(x,y)\frac{\partial u}{\partial x} + e(x,y)\frac{\partial u}{\partial y}, \qquad y = l_y\\ \frac{\partial u}{\partial \tau} &= a(x,y)\frac{\partial^2 u}{\partial x^2} + d(x,y)\frac{\partial u}{\partial x}, \qquad y = r_y.\end{aligned}$$

The fully implicit and Crank-Nicolson schemes require the inversion of non tri-diagonal matrices. Such schemes turn out to be very slow, we will discuss how splitting methods can be employed to overcome this problem. Stability, consistency and convergence of these methods will be shown on a uniform grid. In the first section the general Taylor approximations used to obtain the discrete approximations to our continuous derivatives will be derived. Then we will investigate different types of finite difference schemes that can be used to approximate the solutions of two dimensional parabolic PDEs.

5.1 Discrete approximations (Continued)

This section can be seen as an extension of section 3.1 to two dimensions. Again we truncate the unbounded domain Ω to the bounded domain $\overline{\Omega} = [l_x, r_x] \times [l_y, r_y] \times [0, T]$. Keeping the notation the same as in section 3.1 we see that our aim is to obtain approximations to the true solution on the three dimensional mesh

$$\widehat{\Omega} = \{(x_i, y_j, \tau_k) | i = 0,\ 1,\ \ldots,\ m, \quad j = 0,\ 1,\ \ldots,\ n, \quad k = 0,\ 1,\ \ldots,\ l\}.$$

with the approximation at each mesh point given by

$$u^k_{i,j} \approx u(x_i, y_j, \tau_k).$$

Assuming, for the moment, that the mesh points are uniformly spaced we can write

$$\begin{aligned}x_i &= l_x + ih_x \quad \text{for } i = 0,\ 1,\ \ldots,\ m\\ y_i &= l_y + jh_y \quad \text{for } j = 0,\ 1,\ \ldots,\ n\\ \tau_k &= k\Delta t \quad \text{for } l = 0,\ 1,\ \ldots,\ l\end{aligned}$$

where $h_x = \frac{r_x - l_x}{m}$, $h_y = \frac{r_y - l_y}{n}$ and $\Delta t = \frac{T}{l}$. The finite difference approximations of the derivatives in (5.1), excluding the approximation of the mixed derivative, can be obtained in exactly the same manner

[1] In later sections it will become clear that x denotes the underlying and y denotes the volatility of the underlying. It is not unreasonable to assume that $\frac{\partial u}{\partial \sigma}(x, \sigma_{max}) = 0$ for very large σ_{max}.

as was done in section 3.1. The difference operators are given by

$$\Delta_x^+ u_{i,j}^k = \frac{u_{i+1,j}^k - u_{i,j}^k}{h_x} \tag{5.5}$$

$$\Delta_y^+ u_{i,j}^k = \frac{u_{i,j+1}^k - u_{i,j}^k}{h_y} \tag{5.6}$$

$$\Delta_x^- u_{i,j}^k = \frac{u_{i,j}^k - u_{i-1,j}^k}{h_x} \tag{5.7}$$

$$\Delta_y^- u_{i,j}^k = \frac{u_{i,j}^k - u_{i,j-1}^k}{h_y} \tag{5.8}$$

$$\Delta_x u_{i,j}^k = \frac{u_{i+1,j}^k - u_{i-1,j}^k}{2h_x} \tag{5.9}$$

$$\Delta_y u_{i,j}^k = \frac{u_{i,j+1}^k - u_{i,j-1}^k}{2h_y} \tag{5.10}$$

$$\Delta_x^2 u_{i,j}^k = \frac{u_{i+1,j}^k - 2u_{i,j}^k + u_{i-1,j}^k}{h_x^2} \tag{5.11}$$

$$\Delta_y^2 u_{i,j}^k = \frac{u_{i,j+1}^k - 2u_{i,j}^k + u_{i,j-1}^k}{h_y^2} \tag{5.12}$$

We still need to derive the divided difference approximation of the mixed derivative. We give a derivation of the second order approximation to the mixed derivative proposed in Hout and Welfert [2006]. Second order approximations of the cross derivative can be derived with the aid of the following Taylor expansions about the reference point (x_i, y_j, τ_k)

$$\begin{aligned} u(x_{i+1}, y_{j+1}, \tau_k) = {} & u + h_x \frac{\partial u}{\partial x} + h_y \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} + h_x h_y \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^2 \frac{\partial^2 u}{\partial y^2} \\ & + \frac{1}{3!} h_x^3 \frac{\partial^3 u}{\partial x^3} + \frac{3}{3!} h_x^2 h_y \frac{\partial^3 u}{\partial x^2 \partial y} + \frac{3}{3!} h_x h_y^2 \frac{\partial^3 u}{\partial x \partial y^2} + \frac{1}{3!} h_y^3 \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^4, h_x^3 h_y, h_x^2 h_y^2, h_x h_y^3, h_y^4) \end{aligned} \tag{5.13}$$

$$\begin{aligned} u(x_{i-1}, y_{j+1}, \tau_k) = {} & u - h_x \frac{\partial u}{\partial x} + h_y \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} - h_x h_y \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^2 \frac{\partial^2 u}{\partial y^2} \\ & - \frac{1}{3!} h_x^3 \frac{\partial^3 u}{\partial x^3} + \frac{3}{3!} h_x^2 h_y \frac{\partial^3 u}{\partial x^2 \partial y} - \frac{3}{3!} h_x h_y^2 \frac{\partial^3 u}{\partial x \partial y^2} + \frac{1}{3!} h_y^3 \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^4, h_x^3 h_y, h_x^2 h_y^2, h_x h_y^3, h_y^4) \end{aligned} \tag{5.14}$$

$$\begin{aligned} u(x_{i+1}, y_{j-1}, \tau_k) = {} & u + h_x \frac{\partial u}{\partial x} - h_y \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} - h_x h_y \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^2 \frac{\partial^2 u}{\partial y^2} \\ & + \frac{1}{3!} h_x^3 \frac{\partial^3 u}{\partial x^3} - \frac{3}{3!} h_x^2 h_y \frac{\partial^3 u}{\partial x^2 \partial y} + \frac{3}{3!} h_x h_y^2 \frac{\partial^3 u}{\partial x \partial y^2} - \frac{1}{3!} h_y^3 \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^4, h_x^3 h_y, h_x^2 h_y^2, h_x h_y^3, h_y^4) \end{aligned} \tag{5.15}$$

$$\begin{aligned} u(x_{i-1}, y_{j-1}, \tau_k) = {} & u - h_x \frac{\partial u}{\partial x} - h_y \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} + h_x h_y \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^2 \frac{\partial^2 u}{\partial y^2} \\ & - \frac{1}{3!} h_x^3 \frac{\partial^3 u}{\partial x^3} - \frac{3}{3!} h_x^2 h_y \frac{\partial^3 u}{\partial x^2 \partial y} - \frac{3}{3!} h_x h_y^2 \frac{\partial^3 u}{\partial x \partial y^2} - \frac{1}{3!} h_y^3 \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^4, h_x^3 h_y, h_x^2 h_y^2, h_x h_y^3, h_y^4) \end{aligned} \tag{5.16}$$

$$u(x_{i-1}, y_j, \tau_k) = u - h_x \frac{\partial u}{\partial x} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} - \frac{1}{3!} h_x^3 \frac{\partial^3 u}{\partial x^3} + O(h_x^4) \tag{5.17}$$

$$u(x_{i+1}, y_j, \tau_k) = u + h_x \frac{\partial u}{\partial x} + \tfrac{1}{2} h_x^2 \frac{\partial^2 u}{\partial x^2} + \frac{1}{3!} h_x^3 \frac{\partial^3 u}{\partial x^3} + O(h_x^4) \tag{5.18}$$

$$u(x_i, y_{j-1}, \tau_k) = u - h_y \frac{\partial u}{\partial y} + \tfrac{1}{2} h_y^2 \frac{\partial^2 u}{\partial y^2} - \frac{1}{3!} h_y^3 \frac{\partial^3 u}{\partial y^3} + O(h_y^4) \tag{5.19}$$

$$u(x_i, y_{j+1}, \tau_k) = u + h_y \frac{\partial u}{\partial y} + \tfrac{1}{2} h_y^2 \frac{\partial^2 u}{\partial y^2} + \frac{1}{3!} h_y^3 \frac{\partial^3 u}{\partial y^3} + O(h_y^4). \tag{5.20}$$

To construct a general second order approximation the following linear combinations are of critical importance

$$\begin{aligned} u(x_{i+1}, y_{j+1}, \tau_k) + u(x_{i-1}, y_{j-1}, \tau_k) &= 2u + h_x^2 \frac{\partial^2 u}{\partial x^2} + 2 h_x h_y \frac{\partial^2 u}{\partial x \partial y} + h_y^2 \frac{\partial^2 u}{\partial y^2} + O(h_x^4, h_x^3 h_y, h_x^2 h_y^2, h_x h_y^3, h_y^4) \\ u(x_{i-1}, y_{j+1}, \tau_k) + u(x_{i+1}, y_{j-1}, \tau_k) &= 2u + h_x^2 \frac{\partial^2 u}{\partial x^2} - 2 h_x h_y \frac{\partial^2 u}{\partial x \partial y} + h_y^2 \frac{\partial^2 u}{\partial y^2} + O(h_x^4, h_x^3 h_y, h_x^2 h_y^2, h_x h_y^3, h_y^4) \end{aligned}$$

$$\begin{aligned} u(x_{i-1}, y_j, \tau_k) + u(x_{i+1}, y_j, \tau_k) + u(x_i, y_{j-1}, \tau_k) + u(x_i, y_{j+1}, \tau_k) = {} & 4u + h_x^2 \frac{\partial^2 u}{\partial x^2} + h_y^2 \frac{\partial^2 u}{\partial y^2} \\ & + O(h_x^4, h_y^4). \end{aligned} \tag{5.21}$$

By making an appropriate linear combination of the first two equations we obtain

$$\begin{aligned} (1 + \omega)[u(x_{i+1}, y_{j+1}, \tau_k) + u(x_{i-1}, y_{j-1}, \tau_k)] - (1 - \omega)[u(x_{i-1}, y_{j+1}, \tau_k) + u(x_{i+1}, y_{j-1}, \tau_k)] \\ = 4\omega u + 2\omega h_x^2 \frac{\partial^2 u}{\partial x^2} + 4 h_x h_y \frac{\partial^2 u}{\partial x \partial y} + 2\omega h_x^2 \frac{\partial^2 u}{\partial x^2} + O(h_x^4, h_x^3 h_y, h_x^2 h_y^2, h_x h_y^3, h_y^4) \end{aligned}$$

where $\omega \in [-1,1]$. The $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ terms can now be eliminated by adding -2ω of (5.21), after rearranging we obtain

$$\begin{aligned}\frac{\partial^2 u}{\partial x \partial y} &= \frac{(1+\omega)[u(x_{i+1}, y_{j+1}, \tau_k) + u(x_{i-1}, y_{j-1}, \tau_k)] - (1-\omega)[u(x_{i-1}, y_{j+1}, \tau_k) + u(x_{i+1}, y_{j-1}, \tau_k)]}{4h_x h_y} \\ &+ \frac{4\omega u - 2\omega[u(x_{i-1}, y_j, \tau_k) + u(x_{i+1}, y_j, \tau_k) + u(x_i, y_{j-1}, \tau_k) + u(x_i, y_{j+1}, \tau_k)]}{4h_y h_y} + O(h_x^2, h_x h_y, h_y^2)\end{aligned} \tag{5.22}$$

The difference operator for the mixed derivative can be written as

$$\begin{aligned}\Delta_{xy}^{\omega} u_{i,j}^k = &\frac{(1+\omega)[u_{i+1,j+1}^k + u_{i-1,j-1}^k] - (1-\omega)[u_{i-1,j+1}^k + u_{i+1,j-1}^k]}{4h_x h_y} \\ &+ \frac{4\omega u_{i,j}^k - 2\omega[u_{i-1,j}^k + u_{i+1,j}^k + u_{i,j-1}^k + u_{i,j+1}^k]}{4h_x h_y}\end{aligned} \tag{5.23}$$

5.2 Implicit Explicit schemes (IMEX-schemes)

IMEX-schemes can be seen as a generalization of the θ-method in the sense that these schemes allow the implicitness (or explicitness) of the convection and diffusion parts of the PDE to differ. The discretized PDE for the two dimensional IMEX-method can be written as

$$\begin{aligned}\Delta_t^+ u_{i,j}^k = &\,\theta_{\text{diff}}[a_{i,j}\Delta_x^2 u_{i,j}^{k+1} + c_{i,j}\Delta_y^2 u_{i,j}^{k+1}] + \theta_{\text{cov}}[d_{i,j}\Delta_x u_{i,j}^{k+1} + e_{i,j}\Delta_y u_{i,j}^{k+1}] + 2\theta_{\text{cross}} b_{i,j}\Delta_{xy}^{\omega} u_{i,j}^{k+1} \\ &+ (1-\theta_{\text{diff}})[a_{i,j}\Delta_x^2 u_{i,j}^k + c_{i,j}\Delta_y^2 u_{i,j}^k] + (1-\theta_{\text{cov}})[d_{i,j}\Delta_x u_{i,j}^k + e_{i,j}\Delta_y u_{i,j}^k] + 2(1-\theta_{\text{cross}}) b_{i,j}\Delta_{xy}^{\omega} u_{i,j}^k\end{aligned} \tag{5.24}$$

for $i = 1, 2, \ldots, m-1$, $j = 1, 2, \ldots, n-1$ and $k = 0, 1, \ldots, l-1$, where $a_{i,j} = a(x_i, y_j)$, $b_{i,j} = b(x_i, y_j)$ $c_{i,j} = c(x_i, y_j)$, $d_{i,j} = d(x_i, y_j)$, $e_{i,j} = e(x_i, y_j)$ and $\theta_{\text{diff}}, \theta_{\text{cov}}, \theta_{\text{cross}} \in [0,1]$. The parameters θ_{diff}, θ_{cov} and θ_{cross} respectively determines the implicitness of the diffusion, convection and cross derivative part of the parabolic PDE. After rearranging we obtain

$$\begin{aligned}&(-\lambda_{xx}\theta_{\text{diff}} a_{i,j} + \lambda_x \theta_{\text{cov}} d_{i,j})\theta u_{i-1,j}^{k+1} + (-\lambda_{yy}\theta_{\text{diff}} c_{i,j} + \lambda_y \theta_{\text{cov}} e_{i,j}) u_{i,j-1}^{k+1} + (1 + 2\theta_{\text{diff}} a_{i,j}\lambda_{xx} + 2\theta_{\text{cov}} c_{i,j}\lambda_{yy}) u_{i,j}^{k+1} \\ &\quad + (-\lambda_{yy}\theta_{\text{diff}} c_{i,j} - \lambda_y \theta_{\text{cov}} e_{i,j}) u_{i,j+1}^{k+1} + (-\lambda_{xx}\theta_{\text{diff}} a_{i,j} - \lambda_x \theta_{\text{cov}} d_{i,j}) u_{i+1,j}^{k+1} \\ &\quad + 2\lambda_{xy} b_{i,j}(1+\omega)\theta_{\text{cross}}[u_{i+1,j+1}^{k+1} + u_{i-1,j-1}^{k+1}] - 2\lambda_{xy} b_{i,j}(1-\omega)\theta_{\text{cross}}[u_{i-1,j+1}^{k+1} + u_{i+1,j-1}^{k+1}] \\ &\quad + 8\lambda_{xy} b_{i,j}\omega\theta_{\text{cross}} u_{i,j}^{k+1} - 4\lambda_{xy} b_{i,j}\omega\theta_{\text{cross}}[u_{i-1,j}^{k+1} + u_{i+1,j}^{k+1} + u_{i,j-1}^{k+1} + u_{i,j+1}^{k+1}] \\ &= (\lambda_{xx}(1-\theta_{\text{diff}}) a_{i,j} - \lambda_x(1-\theta_{\text{cov}}) d_{i,j}) u_{i-1,j}^k + (\lambda_{yy}(1-\theta_{\text{diff}}) c_{i,j} - \lambda_y(1-\theta_{\text{cov}}) e_{i,j}) u_{i,j-1}^k \\ &\quad + (1 - 2(1-\theta_{\text{diff}}) a_{i,j}\lambda_{xx} - 2(1-\theta_{\text{diff}}) c_{i,j}\lambda_{yy}) u_{i,j}^k \\ &\quad + (\lambda_{yy}(1-\theta_{\text{diff}}) c_{i,j} + \lambda_y(1-\theta_{\text{cov}}) e_{i,j}) u_{i,j+1}^k + (\lambda_{xx}(1-\theta_{\text{diff}}) a_{i,j} + \lambda_x(1-\theta_{\text{cov}}) d_{i,j})(1-\theta) u_{i+1,j}^k \\ &\quad + 2\lambda_{xy} b_{i,j}(1+\omega)(1-\theta_{\text{cross}})[u_{i+1,j+1}^k + u_{i-1,j-1}^k] - 2\lambda_{xy} b_{i,j}(1-\omega)(1-\theta_{\text{cross}})[u_{i-1,j+1}^k + u_{i+1,j-1}^k] \\ &\quad + 8\lambda_{xy} b_{i,j}\omega(1-\theta_{\text{cross}}) u_{i,j}^k - 4\lambda_{xy} b_{i,j}\omega(1-\theta_{\text{cross}})[u_{i-1,j}^k + u_{i+1,j}^k + u_{i,j-1}^k + u_{i,j+1}^k]\end{aligned} \tag{5.25}$$

for $i = 1,\ 2,\ \ldots,\ m-1$, $j = 1,\ 2,\ \ldots,\ n-1$ and $k = 0,\ 1,\ \ldots,\ l-1$, where $\lambda_{xy} = \frac{\Delta t}{4h_x h_y}$. In discrete form the boundary conditions can be written as

$$\begin{aligned}
u^{k+1}_{0,\,j} &= c_l \quad \text{for} \quad j = 0,\ 1,\ \ldots,\ n && (5.26)\\
\frac{u^{k+1}_{m,\,j} - u^{k+1}_{m-1,\,j}}{h_x} &= c_r \quad \text{for } j = 0,\ 1,\ \ldots,\ n && (5.27)\\
\frac{u^{k+1}_{i,\,0} - u^{k}_{i,\,0}}{\Delta t} &= \theta_{\text{cov}} d_{i,\,0}\Delta^+_x u^{k+1}_{i,\,0} + \theta_{\text{cov}} e_{i,\,0}\Delta^+_y u^{k+1}_{i,\,0} && (5.28)\\
&\quad + (1-\theta_{\text{cov}}) d_{i,\,0}\Delta^+_x u^{k}_{i,\,0} + (1-\theta_{\text{cov}}) e_{i,\,0}\Delta^+_y u^{k}_{i,\,0} \quad \text{for} \quad i = 1,\ 2,\ \ldots,\ m-1\\
\frac{u^{k+1}_{i,\,n} - u^{k}_{i,\,n}}{\Delta t} &= \theta_{\text{diff}} a_{i,\,n}\Delta^2_x u^{k+1}_{i,\,n} + \theta_{\text{cov}} d_{i,\,n}\Delta_x u^{k+1}_{i,\,n} && (5.29)\\
&\quad + (1-\theta_{\text{diff}}) a_{i,\,n}\Delta^2_x u^{k}_{i,\,n} + (1-\theta_{\text{cov}}) d_{i,\,n}\Delta_x u^{k}_{i,\,n} \quad \text{for} \quad i = 1,\ 2,\ \ldots,\ m-1
\end{aligned}$$

for $k = 0,\ 1,\ \ldots,\ l-1$. To ensure the stability of the scheme at the boundary, upwinding is used in equation (5.28). In particular from section 3.7.6 we see that upwinding will ensure that the scheme is stable under the maximum norm at the boundary for the case when $\theta_{\text{cov}} = 1$.

5.3 Consistency

In this section we will prove the consistency of the IMEX-method, the proof will be of the same form as the proof of consistency given in section 3.4. The truncation error of the IMEX-method in two dimensions is given by

$$\begin{aligned}
L^{\Delta t}_{h_x,\,h_y} v(x_i, y_j, \tau_k) &= \Delta^+_t v(x_i, y_j, \tau_k) && (5.30)\\
&\quad - \theta_{\text{diff}}[a_{i,\,j}\Delta^2_x v(x_i, y_j, \tau_{k+1}) + c_{i,\,j}\Delta^2_y v(x_i, y_j, \tau_{k+1})] - \theta_{\text{cov}}[d_{i,\,j}\Delta_x v(x_i, y_j, \tau_{k+1})\\
&\quad + e_{i,\,j}\Delta_y v(x_i, y_j, \tau_{k+1})] - 2\theta_{\text{cross}} b_{i,\,j}\Delta^\omega_{xy} v(x_i, y_j, \tau_{k+1})\\
&\quad - (1-\theta_{\text{diff}})[a_{i,\,j}\Delta^2_x v(x_i, y_j, \tau_k) + c_{i,\,j}\Delta^2_y v(x_i, y_j, \tau_k)] - (1-\theta_{\text{cov}})[d_{i,\,j}\Delta_x v(x_i, y_j, \tau_k)\\
&\quad + e_{i,\,j}\Delta_y v(x_i, y_j, \tau_k)] - 2(1-\theta_{\text{cross}}) b_{i,\,j}\Delta^\omega_{xy} v(x_i, y_j, \tau_k)
\end{aligned}$$

As done in section 3.4 we need to compute the Taylor expansions about $(x_i, y_j, \tau_{k+\frac{1}{2}})$ in order to obtain the truncation error. Expanding $v(x_i, y_j, \tau_k)$ and $v(x_i, y_j, \tau_{k+1})$ about $(x_i, y_j, \tau_{k+\frac{1}{2}})$ and subtracting results in

$$\Delta^+_t v(x_i, y_j, \tau_k) = \frac{\partial v}{\partial \tau}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \frac{1}{24}\Delta t^2 \frac{\partial^3 v}{\partial \tau^3}(x_i, y_j, \tau_{k+\frac{1}{2}}) + O(\Delta t^4)$$

Similarly as in section 3.4 we can compute the Taylor expansions of the difference operators at time τ_k about the point $(x_i, y_j, \tau_{k+\frac{1}{2}})$ to obtain

$$\begin{aligned}
\Delta_x v(x_i, y_j, \tau_k) &= \Big(\frac{\partial v}{\partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial x \partial \tau^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) \\
&\quad + O(h_x^2) \\
\Delta_x^2 v(x_i, y_j, \tau_k) &= \Big(\frac{\partial^2 v}{\partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) \\
&\quad + O(h_x^2) \\
\Delta_y v(x_i, y_j, \tau_k) &= \Big(\frac{\partial v}{\partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial y \partial \tau^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) \\
&\quad + O(h_y^2) \\
\Delta_y^2 v(x_i, y_j, \tau_k) &= \Big(\frac{\partial^2 v}{\partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) \\
&\quad + O(h_y^2)
\end{aligned}$$

whereas difference operators applied at time τ_{k+1} and expanded about the point $(x_i, y_j, \tau_{k+\frac{1}{2}})$ yield

$$\begin{aligned}
\Delta_x v(x_i, y_j, \tau_{k+1}) &= \Big(\frac{\partial v}{\partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
&\quad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial x \partial \tau^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) + O(h_x^2) \\
\Delta_x^2 v(x_i, y_j, \tau_{k+1}) &= \Big(\frac{\partial^2 v}{\partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
&\quad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) + O(h_x^2) \\
\Delta_y v(x_i, y_j, \tau_{k+1}) &= \Big(\frac{\partial v}{\partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
&\quad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial y \partial \tau^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) + O(h_y^2) \\
\Delta_y^2 v(x_i, y_j, \tau_{k+1}) &= \Big(\frac{\partial^2 v}{\partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
&\quad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \dots\Big) + O(h_y^2)
\end{aligned}$$

To obtain these results for the approximations of the mixed derivative we rearrange (5.22) to obtain

$$\begin{aligned}
\Delta_{xy}^{\omega} v(x_i, y_j, \tau_k) &= \frac{\partial^2 v}{\partial x \partial y}(x_i, y_j, \tau_k) + O(h_x^2, h_x h_y, h_y^2) \\
\Delta_{xy}^{\omega} v(x_i, y_j, \tau_{k+1}) &= \frac{\partial^2 v}{\partial x \partial y}(x_i, y_j, \tau_{k+1}) + O(h_x^2, h_x h_y, h_y^2)
\end{aligned}$$

Expanding the terms on the right about $(x_i, y_j, \tau_{k+\frac{1}{2}})$ yields

$$\begin{aligned}
\Delta^{\omega}_{xy} v(x_i, y_j, \tau_k) = \Bigg(& \frac{\partial^2 v}{\partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) - \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
& + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \Bigg) + O(h_x^2, h_x h_y, h_y^2) \\
\Delta^{\omega}_{xy} v(x_i, y_j, \tau_{k+1}) = \Bigg(& \frac{\partial^2 v}{\partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
& + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \Bigg) + O(h_x^2, h_x h_y, h_y^2)
\end{aligned}$$

Substitution of these equations into equation (5.30) results in

$$\begin{aligned}
& L^{\Delta t}_{h_x, h_y} v(x_i, y_j, \tau_k) = \\
& \quad \frac{\partial v}{\partial \tau}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \frac{1}{24}\Delta t^2 \frac{\partial^3 v}{\partial \tau^3}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \qquad (5.31) \\
& \quad - a(x_i, y_j) \Bigg[\frac{\partial^2 v}{\partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + (1 - 2\theta_{\text{diff}})\tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
& \qquad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \Bigg] \\
& \quad - d(x_i, y_j) \Bigg[\frac{\partial v}{\partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) + (1 - 2\theta_{\text{cov}})\tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
& \qquad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \Bigg] \\
& \quad - c(x_i, y_j) \Bigg[\frac{\partial^2 v}{\partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + (1 - 2\theta_{\text{diff}})\tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau^2 \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
& \qquad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial y^2}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \Bigg] \\
& \quad - e(x_i, y_j) \Bigg[\frac{\partial v}{\partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + (1 - 2\theta_{\text{cov}})\tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
& \qquad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial \tau^2 \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \Bigg] \\
& \quad - 2b(x_i, y_j) \Bigg[\frac{\partial^2 v}{\partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + (1 - 2\theta_{\text{cross}})\tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau \partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) \\
& \qquad + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x \partial y}(x_i, y_j, \tau_{k+\frac{1}{2}}) + \ldots \Bigg] \\
& \quad + O(h_x^2, h_x h_y, h_y^2).
\end{aligned}$$

Grouping the terms and using the fact that v is the solution of (5.1) we obtain

$$
\begin{aligned}
&L^{\Delta t}_{h_x,\,h_y} v(x_i,y_j,\tau_k)\\
&= \Bigg[\frac{\partial v}{\partial \tau}(x_i,y_j,\tau_{k+\frac{1}{2}}) - a(x_i,y_j)\frac{\partial^2 v}{\partial x^2}(x_i,y_j,\tau_{k+\frac{1}{2}}) - d(x_i,y_j)\frac{\partial v}{\partial x}(x_i,y_j,\tau_{k+\frac{1}{2}}) \qquad (5.32)\\
&\quad -c(x_i,y_j)\frac{\partial^2 v}{\partial y^2}(x_i,y_j,\tau_{k+\frac{1}{2}}) - e(x_i,y_j)\frac{\partial v}{\partial y}(x_i,y_j,\tau_{k+\frac{1}{2}}) - 2b(x_i,y_j)\frac{\partial^2 v}{\partial x \partial y}(x_i,y_j,\tau_{k+\frac{1}{2}})\Bigg]\\
&\quad - \left[a(x_i,y_j)\frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i,y_j,\tau_{k+\frac{1}{2}}) + c(x_i,y_j)\frac{\partial^3 v}{\partial \tau^2 \partial y}(x_i,y_j,\tau_{k+\frac{1}{2}})\right](1-2\theta_{\text{diff}})\tfrac{1}{2}\Delta t \qquad (5.33)\\
&\quad - \left[d(x_i,y_j)\frac{\partial^2 v}{\partial \tau \partial x}(x_i,y_j,\tau_{k+\frac{1}{2}}) + e(x_i,y_j)\frac{\partial^2 v}{\partial \tau \partial y}(x_i,y_j,\tau_{k+\frac{1}{2}})\right](1-2\theta_{\text{cov}})\tfrac{1}{2}\Delta t\\
&\quad - 2(1-2\theta_{\text{cross}})b(x_i,y_j)\tfrac{1}{2}\Delta t\frac{\partial^3 v}{\partial \tau \partial x \partial y}(x_i,y_j,\tau_{k+\frac{1}{2}})\\
&\quad + O(\Delta t^2) + O(h_x^2, h_x h_y, h_y^2)\\
&= -\left[a(x_i,y_j)\frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i,y_j,\tau_{k+\frac{1}{2}}) + c(x_i,y_j)\frac{\partial^3 v}{\partial \tau^2 \partial y}(x_i,y_j,\tau_{k+\frac{1}{2}})\right](1-2\theta_{\text{diff}})\tfrac{1}{2}\Delta t\\
&\quad - \left[d(x_i,y_j)\frac{\partial^2 v}{\partial \tau \partial x}(x_i,y_j,\tau_{k+\frac{1}{2}}) + e(x_i,y_j)\frac{\partial^2 v}{\partial \tau \partial y}(x_i,y_j,\tau_{k+\frac{1}{2}})\right](1-2\theta_{\text{cov}})\tfrac{1}{2}\Delta t\\
&\quad - 2(1-2\theta_{\text{cross}})b(x_i,y_j)\tfrac{1}{2}\Delta t\frac{\partial^3 v}{\partial \tau \partial x \partial y}(x_i,y_j,\tau_{k+\frac{1}{2}}) + O(\Delta t^2) + O(h_x^2, h_x h_y, h_y^2).
\end{aligned}
$$

From the last equation it is clear that $L^{\Delta t}_{h_x,\,h_y} v(x_i,y_j,\tau_k) \to 0$ as $\Delta t, h_x, h_y \to 0$, hence we can deduce that the IMEX-method is consistent. It is also clear that this method will be of second order only if $\theta_{\text{diff}} = \theta_{\text{cov}} = \theta_{\text{cross}} = \frac{1}{2}$ [2].

5.4 Stability

Similarly as in section 3.5 we will discuss the stability of the IMEX-method in two dimensions using both von Neumann stability analysis and the matrix method of analysis.

5.4.1 von Neumann stability analysis

As done in the previous sections we will assume constant coefficients when we apply von Neumann stability analysis. Similarly as for the one dimensional case we can investigate the propagation of initial data points by substituting

$$u^k_{i,\,j} = \gamma^k e^{I\alpha i h_x} e^{I\beta j h_y} \qquad (5.34)$$

[2] Note that this is the θ-method in two dimensions with $\theta = \frac{1}{2}$, also known as the Crank-Nicolson scheme.

where $I = \sqrt{-1}$ in (5.24), see Thomas [1995] and Duffy [2006]. We start the proof by computing $\Delta_x u^k_{i,\,j}$, $\Delta_y u^k_{i,\,j}$, $\Delta^2_x u^k_{i,\,j}$, $\Delta^2_y u^k_{i,\,j}$ and $\Delta^0_{xy} u^k_{i,\,j}$ for a general time step τ_k

$$\begin{aligned}
\Delta_x u^k_{i,\,j} &= \frac{u^k_{i\,+\,1,\,j} - u^k_{i\,-\,1,\,j}}{2h_x} \\
&= \frac{1}{2h_x}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\left(e^{I\alpha h_x} - e^{-I\alpha h_x}\right) \\
&= I\frac{1}{h_x}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\sin(\alpha h_x)
\end{aligned}$$

similarly

$$\Delta_y u^k_{i,\,j} = I\frac{1}{h_y}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\sin(\beta h_y).$$

For the approximations of the second order terms we have

$$\begin{aligned}
\Delta^2_x u^k_{i,\,j} &= \frac{u^k_{i\,+\,1,\,j} - 2u^k_{i,\,j} + u^k_{i\,-\,1,\,j}}{h^2_x} \\
&= \frac{1}{h^2_x}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\left(e^{-I\alpha h_x} - 2 + e^{I\alpha h_x}\right) \\
&= \frac{2}{h^2_x}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}(\cos(\alpha h_x) - 1) \\
&= -\frac{4}{h^2_x}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\sin^2\left(\frac{\alpha h_x}{2}\right)
\end{aligned}$$

and similarly

$$\Delta^2_y u^k_{i,\,j} = -\frac{4}{h^2_y}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\sin^2\left(\frac{\beta h_y}{2}\right).$$

The cross derivative term becomes

$$\begin{aligned}
\Delta^0_{xy} u^k_{i,\,j} &= \frac{u^k_{i\,+\,1,\,j\,+\,1} - u^k_{i\,-\,1,\,j\,+\,1} - u^k_{i\,+\,1,\,j\,-\,1} + u^k_{i\,-\,1,\,j\,-\,1}}{4h_x h_y} \\
&= \frac{1}{4h_x h_y}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\left(e^{I(\alpha h_x+\beta h_y)} + e^{-I(\alpha h_x+\beta h_y)} - e^{I(\alpha h_x-\beta h_y)} - e^{-I(\alpha h_x-\beta h_y)}\right) \\
&= \frac{1}{2h_x h_y}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\left(\cos(\alpha h_x + \beta h_y) - \cos(\alpha h_x - \beta h_y)\right) \\
&= -\frac{1}{h_x h_y}\gamma^k e^{I\alpha i h_x} e^{I\beta j h_y}\sin(\alpha h_x)\sin(\beta h_y).
\end{aligned}$$

By substituting these results in (5.24) we obtain

$$\gamma = \frac{1 - (1-\theta_{\text{diff}})\Delta t\varepsilon_1 - (1-\theta_{\text{cross}})\Delta t\omega_1 + I(1-\theta_{\text{cov}})\varepsilon_2}{1 + \theta_{\text{diff}}\Delta t\varepsilon_1 + \theta_{\text{cross}}\Delta t\omega_1 - I\theta_{\text{cov}}\Delta t\varepsilon_2}$$

where

$$\begin{aligned}
\varepsilon_1 &= a\frac{4}{h^2_x}\sin^2\left(\frac{\alpha h_x}{2}\right) + c\frac{4}{h^2_y}\sin^2\left(\frac{\beta h_y}{2}\right) \\
\varepsilon_2 &= d\frac{1}{h_x}\sin(\alpha h_x) + e\frac{1}{h_y}\sin(\beta h_y) \\
\omega_1 &= \frac{2b}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y).
\end{aligned}$$

It follows that

$$|\gamma|^2 = \frac{(1-(1-\theta_{\text{diff}})\Delta t\varepsilon_1 - (1-\theta_{\text{cross}})\Delta t\omega_1)^2 + ((1-\theta_{\text{cov}})\varepsilon_2)^2}{(1+\theta_{\text{diff}}\Delta t\varepsilon_1 + \theta_{\text{cross}}\Delta t\omega_1)^2 + (\theta_{\text{cov}}\Delta t\varepsilon_2)^2}.$$

From section 3.7.6 we see that the θ-method for convection equation is stable when $\theta \geq \frac{1}{2}$. We will use this as our main motivation for assuming that $\theta_{\text{cov}} \geq \frac{1}{2}$, which results in

$$|\gamma|^2 \leq \frac{(1-(1-\theta_{\text{diff}})\Delta t\varepsilon_1 - (1-\theta_{\text{cross}})\Delta t\omega_1)^2}{(1+\theta_{\text{diff}}\Delta t\varepsilon_1 + \theta_{\text{cross}}\Delta t\omega_1)^2}. \tag{5.35}$$

Finding general ranges of θ_{diff}, θ_{cross} and ω for which $|\gamma| \leq 1$ is a non-trivial task. Since second order accuracy is only obtained when $\theta_{\text{diff}} = \theta_{\text{cross}} = \frac{1}{2}$ we feel that it is unnecessary to prove stability for complete ranges of these values. We will focus on specific choices of these parameters.

$\theta_{\text{diff}} = \theta_{\text{cross}} = \Theta$ **and** $\omega = 0$

Substituting in (5.35) results in

$$|\gamma|^2 \leq \frac{(1-(1-\Theta)\Delta t(\varepsilon_1+\omega_1))^2}{(1+\Theta\Delta t(\varepsilon_1+\omega_1))^2}.$$

From the assumption that $\Theta \geq \frac{1}{2}$ it follows that $|\gamma|^2 \leq 1$ if we can show that $\varepsilon_1 + \omega_1 \geq 0$. To show this we will make use of the following lemma

Lemma 5.4.1. *(McKee et al. [1996])*

$$q\sin^2\frac{\theta}{2} + r\sin^2\frac{\phi}{2} + \frac{s}{4}\sin\theta\sin\phi \geq 0 \tag{5.36}$$

if $q \geq 0$, $r \geq 0$ *and* $4qr > s^2$.

Proof. The inequality trivially holds if $\frac{s}{4}\sin\theta\sin\phi \geq 0$. If $\frac{s}{4}\sin\theta\sin\phi < 0$ then

$$\begin{aligned}
&q\sin^2\frac{\theta}{2} + r\sin^2\frac{\phi}{2} + \frac{s}{4}\sin\theta\sin\phi \\
&= q\sin^2\frac{\theta}{2} + r\sin^2\frac{\phi}{2} - \frac{|s|}{4}|\sin\theta\sin\phi| \\
&> q\sin^2\frac{\theta}{2} + r\sin^2\frac{\phi}{2} - \frac{\sqrt{qr}}{2}|\sin\theta\sin\phi| \\
&= \left(\sqrt{q}\left|\sin\frac{\theta}{2}\right| - \sqrt{r}\left|\sin\frac{\phi}{2}\right|\right)^2 + 2\sqrt{qr}\left|\sin\frac{\theta}{2}\sin\frac{\phi}{2}\right| - \frac{\sqrt{qr}}{2}|\sin\theta\sin\phi| \\
&= \left(\sqrt{q}\left|\sin\frac{\theta}{2}\right| - \sqrt{r}\left|\sin\frac{\phi}{2}\right|\right)^2 + 2\sqrt{qr}\left|\sin\frac{\theta}{2}\sin\frac{\phi}{2}\right| - \frac{\sqrt{qr}}{2}\left|4\sin\frac{\theta}{2}\cos\frac{\theta}{2}\sin\frac{\phi}{2}\cos\frac{\phi}{2}\right| \\
&= \left(\sqrt{q}\left|\sin\frac{\theta}{2}\right| - \sqrt{r}\left|\sin\frac{\phi}{2}\right|\right)^2 + 2\sqrt{qr}\left|\sin\frac{\theta}{2}\sin\frac{\phi}{2}\right|\left(1-\left|\cos\frac{\theta}{2}\cos\frac{\phi}{2}\right|\right) \\
&\geq 0
\end{aligned}$$

where the first inequality follows from the assumption that $4qr > s^2$ and the final inequality from the fact that $\left|\cos\frac{\theta}{2}\cos\frac{\phi}{2}\right| \leq 1$. □

Corollary 5.4.1.

$$q_c \sin^2 \frac{\theta}{2} + r_c \sin^2 \frac{\phi}{2} - \frac{s_c}{4} \sin\theta \sin\phi \geq 0$$

if $4q_c r_c \geq s_c^2$.

Proof. Simply use lemma 5.4.1 with $q = q_c$, $r = r_c$ and $s = -s_c$

$$4qr = 4q_c r_c \geq s_c^2 = (-s_c)^2 = s^2$$

□

By letting

$$q = \frac{a}{h_x^2}, \qquad r = \frac{c}{h_y^2} \quad \text{and} \quad s = \frac{2b}{h_x h_y}$$

in lemma 5.4.1 we see that $\varepsilon_1 + \omega_1 \geq 0$ if we can show that $4qr \geq s^2$. Consider

$$4qr = \frac{4ac}{h_x^2 h_y^2} \geq \frac{4b^2}{h_x^2 h_y^2} = s^2$$

where the inequality follows from the assumption that the PDE is parabolic. From the corollary it follows that $\varepsilon_1 - \omega_1 \geq 0$. In particular we have shown that the second order Crank-Nicolson scheme is von Neumann stable. The explicitness of a finite difference method scheme is directly linked to the stability and the computational effort of the scheme. Explicit schemes require less computational effort but tend to be more unstable than the slower implicit schemes. In the following subsection we discuss a combination of these schemes where we keep the cross derivative term explicit and the diffusion part implicit,

$\theta_{\text{diff}} = 1$**,** $\theta_{\text{cross}} = 0$ **and** $\omega = 0$

Substituting in (5.35) results in

$$\begin{aligned} |\gamma|^2 &\leq \frac{(1 - \Delta t \omega_1)^2}{(1 + \Delta t \varepsilon_1)^2} \\ &= \frac{1 - 2\Delta t \omega_1 + \Delta t^2 \omega_1^2}{1 + 2\Delta t \varepsilon_1 + \Delta t^2 \varepsilon_1^2} \end{aligned}$$

From the fact that $\varepsilon_1 + \omega_1 \geq 0$ it follows that $|\gamma|^2 \leq 1$ if we can show that $\varepsilon_1^2 \geq \omega_1^2$. Consider

$$\varepsilon_1^2 - \omega_1^2 = (\varepsilon_1 + \omega_1)(\varepsilon_1 - \omega_1) \geq 0.$$

where the inequality follows from the fact that $\varepsilon_1 + \omega_1 \geq 0$ and $\varepsilon_1 - \omega_1 \geq 0$. This case will play an important role when we consider more modern FDMs in the later sections.

5.4.2 Matrix formulation

Before we attempt to prove stability under the maximum norm we will rewrite the IMEX-method in matrix form. Equation (5.25) together with the boundary conditions can be written in matrix form as

$$\begin{aligned}(\mathbf{I}_{\text{im}} - \theta_{\text{diff}}\Delta t(\mathbf{A}_{\text{diff}} + \mathbf{C}_{\text{diff}}) - \theta_{\text{cov}}\Delta t(\mathbf{A}_{\text{cov}} + \mathbf{C}_{\text{cov}}) - 2\theta_{\text{cross}}\Delta t\mathbf{B})\mathbf{u}^{k\,+\,1} \\ = (\mathbf{I}_{\text{ex}} + (1-\theta_{\text{diff}})\Delta t(\mathbf{A}_{\text{diff}} + \mathbf{C}_{\text{diff}}) + (1-\theta_{\text{cov}})\Delta t(\mathbf{A}_{\text{cov}} + \mathbf{C}_{\text{cov}}) + 2(1-\theta_{\text{cross}})\Delta t\mathbf{B})\mathbf{u}^{k\,+\,1} + \mathbf{b}\end{aligned} \tag{5.37}$$

for $k = 0,\ 1,\ \ldots,\ l-1$. To proceed from one time level to the next, this matrix equation (5.37) must be solved. The $(m+1)(n+1) \times (m+1)(n+1)$-matrices $\mathbf{A}_{\text{diff}}$ and $\mathbf{C}_{\text{diff}}$ contain the divided difference approximations for the diffusion operator in the x and y directions respectively, whereas the approximation of the convection operators are contained in $\mathbf{A}_{\text{cov}}$ and $\mathbf{C}_{\text{cov}}$. The $(m+1)(n+1) \times (m+1)(n+1)$-matrix $\mathbf{B}$ contains the approximations of the cross derivative at the respective grid points. The $(m+1)(n+1)$ vector $\mathbf{b}$ contains the information given by the boundary conditions. The rows of $\mathbf{I}_{\text{im}}$ and $\mathbf{I}_{\text{ex}}$ corresponding to interior grid points have one as a diagonal element and zeros as off-diagonal elements. The rows of these matrices corresponding to boundary points, at $x = l_x$ and $x = r_x$, incorporate the relevant boundary conditions.

The ordering of the elements in $\mathbf{u}$ will determine the structure of $\mathbf{A}_{\text{diff}}$, $\mathbf{C}_{\text{diff}}$, $\mathbf{A}_{\text{cov}}$, $\mathbf{C}_{\text{cov}}$ and $\mathbf{B}$. Two different orderings will be considered in this project, the first will ensure that $\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{cov}}$ is a tri-diagonal matrix and the second that $\mathbf{C}_{\text{diff}} + \mathbf{C}_{\text{cov}}$ is a tri-diagonal matrix. Tri-diagonal matrices can be viewed as banded matrices with upper and lower bandwidths of one[3]. We will denote the first ordering by the subscript A and the second by the subscript C

$$\begin{aligned}\mathbf{u}_A &= (u_{0,\,0}, u_{1,\,0}, \ldots, u_{m,\,0}; \ldots\ldots; u_{0,\,j}, u_{1,\,j}, \ldots, u_{m,\,j}; \ldots\ldots; u_{0,\,n}, u_{1,\,n}, \ldots, u_{m,\,n})^T \\ \mathbf{u}_C &= (u_{0,\,0}, u_{0,\,1}, \ldots, u_{0,\,n}; \ldots\ldots; u_{i,\,0}, u_{i,\,1}, \ldots, u_{i,\,n}; \ldots\ldots; u_{m,\,0}, u_{m,\,1}, \ldots, u_{m,\,n})^T\end{aligned}$$

A banded matrix with an upper band of b and an lower band of a can be represented in a-1-b bandwidth form as suggested in Press et al. [1992] and Hagan and West [2006][4]. For an arbitrary matrix, $\mathbf{G}$, $|\mathbf{G}||\mathbf{u}|$ will denote $\mathbf{Gu}$ in bandwidth form.

To illustrate the structure of the matrices in (5.37) it is convenient to give a simple example. The case when $m = 3$ and $n = 2$ will be used to illustrate the structure of the matrices $\mathbf{A}_{\text{diff}} + \mathbf{A}_{\text{cov}}$, $\mathbf{C}_{\text{diff}} + \mathbf{C}_{\text{cov}}$, $\mathbf{B}$, $\mathbf{I}_{\text{im}}$ and $\mathbf{I}_{\text{ex}}$ for both of the orderings. Figure (5.1) shows the finite difference grid for our example. The 1-1-1 bandwidth form of $\mathbf{A}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}$ is given by,

[3] When $\mathbf{A}$ is banded : $A_{i,\,j} = 0$ if $i - j > a$ or $i - j < -b$ where $0 \leq a, b < n$. The numbers a and b are called the lower bandwidth and upper bandwidth respectively.

[4] A matrix in a-1-b bandwidth form means that the entry in the i^{th} row and j^{th} column in this representation actually lies in the i^{th} row and $i + j - a - 1^{th}$ column in the original matrix representation. The entries denoted by × are irrelevant.

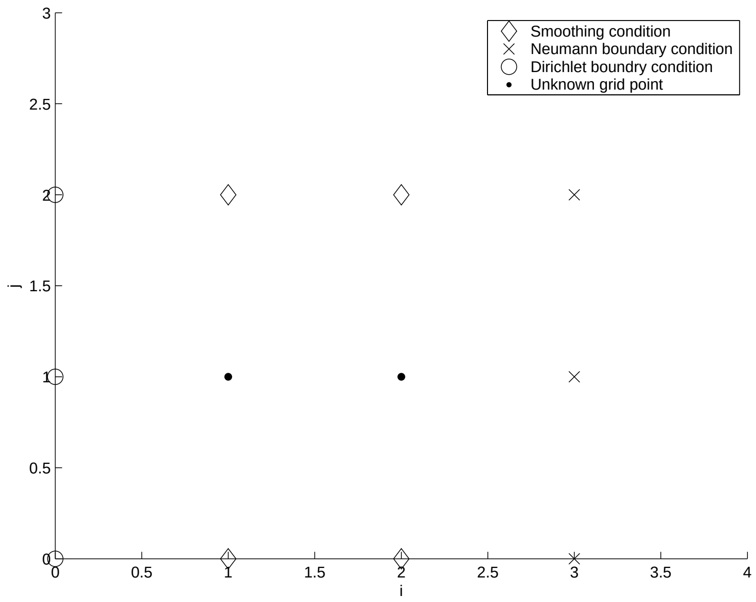


Figure 5.1: Finite difference grid when $m = 3$ and $n = 2$.

$$|\mathbf{A}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}||\mathbf{u}_A^k| = \left|\begin{array}{c|c|c||c|}
\times & 0 & & u^k_{0,\,0} \\
 & -\frac{d_{1,\,0}}{h_x} & \frac{d_{1,\,0}}{h_x} & u^k_{1,\,0} \\
 & -\frac{d_{2,\,0}}{h_x} & \frac{d_{2,\,0}}{h_x} & u^k_{2,\,0} \\
 & 0 & & u^k_{3,\,0} \\
\hline
 & 0 & & u^k_{0,\,1} \\
\frac{a_{1,\,1}}{h_x^2} - \frac{d_{1,\,1}}{2h_x} & -\frac{2a_{1,\,1}}{h_x^2} & \frac{a_{1,\,1}}{h_x^2} + \frac{d_{1,\,1}}{2h_x} & u^k_{1,\,1} \\
\frac{a_{2,\,1}}{h_x^2} - \frac{d_{2,\,1}}{2h_x} & -\frac{2a_{2,\,1}}{h_x^2} & \frac{a_{2,\,1}}{h_x^2} + \frac{d_{2,\,1}}{2h_x} & u^k_{2,\,1} \\
 & 0 & & u^k_{3,\,1} \\
\hline
 & 0 & & u^k_{0,\,2} \\
\frac{a_{1,\,2}}{h_x^2} - \frac{d_{1,\,2}}{2h_x} & -\frac{2a_{1,\,2}}{h_x^2} & \frac{a_{1,\,2}}{h_x^2} + \frac{d_{1,\,2}}{2h_x} & u^k_{1,\,2} \\
\frac{a_{2,\,2}}{h_x^2} - \frac{d_{2,\,2}}{2h_x} & -\frac{2a_{2,\,2}}{h_x^2} & \frac{a_{2,\,2}}{h_x^2} + \frac{d_{2,\,2}}{2h_x} & u^k_{2,\,2} \\
 & 0 & \times & u^k_{3,\,2}
\end{array}\right|$$

$\mathbf{C}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}$ can be written in $(m+1)$-1-$(m+1)$ bandwidth form as

$$
|\mathbf{C}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}||\mathbf{u}_A^k| =
\left|\begin{array}{c|c|c||c|}
\times \qquad \overbrace{\cdots}^{m \text{ columns}} & 0 & & u_{0,\,0}^k \\
\times & -\frac{e_{1,\,0}}{h_y} & \frac{e_{1,\,0}}{h_y} & u_{1,\,0}^k \\
\times & -\frac{e_{2,\,0}}{h_y} & \frac{e_{2,\,0}}{h_y} & u_{2,\,0}^k \\
\times & 0 & & u_{3,\,0}^k \\
\hline
 & 0 & & u_{0,\,1}^k \\
\frac{c_{1,\,1}}{h_y^2} - \frac{e_{1,\,1}}{2h_y} & -\frac{2c_{1,\,1}}{h_y^2} & \frac{c_{1,\,1}}{h_y^2} + \frac{e_{1,\,1}}{2h_y} & u_{1,\,1}^k \\
\frac{c_{2,\,1}}{h_y^2} - \frac{e_{2,\,1}}{2h_y} & -\frac{2c_{2,\,1}}{h_y^2} & \frac{c_{2,\,1}}{h_y^2} + \frac{e_{2,\,1}}{2h_y} & u_{2,\,1}^k \\
 & 0 & & u_{3,\,1}^k \\
\hline
 & 0 & \times & u_{0,\,2}^k \\
 & 0 & \times & u_{1,\,2}^k \\
 & 0 & \times & u_{2,\,2}^k \\
 & 0 & \overset{m \text{ columns}}{\underbrace{\cdots}} \qquad \times & u_{3,\,2}^k
\end{array}\right.
$$

and $\mathbf{B}_A$ can be written in $(m+2)$-1-$(m+2)$ bandwidth form as

$$
|\mathbf{B}_A||\mathbf{u}_A^k| =
\left|\begin{array}{ccccc|c|ccccc||c}
\times & \times & \times & \overbrace{\cdots}^{m-2\,\mathrm{col}} & \times & 0 & & & & & & u_{0,\,0}^k \\
\times & \times & \times & & & 0 & & & & & & u_{1,\,0}^k \\
\times & \times & \times & & & 0 & & & & & & u_{2,\,0}^k \\
\times & \times & & & & 0 & & & & & & u_{3,\,0}^k \\
\hline
\times & & & & & 0 & & & & & & u_{0,\,1}^k \\
\frac{(1+\omega)b_{1,\,1}}{4h_xh_y} & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & -\frac{(1-\omega)b_{1,\,1}}{4h_xh_y} & & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & \frac{4\omega b_{1,\,1}}{4h_xh_y} & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & & -\frac{(1-\omega)b_{1,\,1}}{4h_xh_y} & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & \frac{(1+\omega)b_{1,\,1}}{4h_xh_y} & u_{1,\,1}^k \\
\frac{(1+\omega)b_{2,\,1}}{4h_xh_y} & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & -\frac{(1-\omega)b_{2,\,1}}{4h_xh_y} & & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & \frac{4\omega b_{2,\,1}}{4h_xh_y} & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & & -\frac{(1-\omega)b_{2,\,1}}{4h_xh_y} & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & \frac{(1+\omega)b_{2,\,1}}{4h_xh_y} & u_{2,\,1}^k \\
& & & & & 0 & & & & & \times & u_{3,\,1}^k \\
\hline
& & & & & 0 & & & & \times & \times & u_{0,\,2}^k \\
& & & & & 0 & & & \times & \times & \times & u_{1,\,2}^k \\
& & & & & 0 & & & \times & \times & \times & u_{2,\,2}^k \\
& & & & & 0 & \times & \underbrace{\cdots}_{m-2\,\mathrm{col}} & \times & \times & \times & u_{3,\,2}^k
\end{array}\right|
$$

The *modified* identity matrices are given by

$$
|\mathbf{I}_{\mathrm{im},A}||\mathbf{u}_A^k| =
\left|\begin{array}{c|c||c}
\times & 1 & u_{0,\,0}^k \\
& 1 & u_{1,\,0}^k \\
& 1 & u_{2,\,0}^k \\
-1 & 1 & u_{3,\,0}^k \\
\hline
& 1 & u_{0,\,1}^k \\
& 1 & u_{1,\,1}^k \\
& 1 & u_{2,\,1}^k \\
-1 & 1 & u_{3,\,1}^k \\
\hline
& 1 & u_{0,\,2}^k \\
& 1 & u_{1,\,2}^k \\
& 1 & u_{2,\,2}^k \\
-1 & 1 & u_{3,\,2}^k
\end{array}\right|
\qquad \text{and} \qquad
|\mathbf{I}_{\mathrm{ex},A}||\mathbf{u}_A^k| =
\left|\begin{array}{c||c}
0 & u_{0,\,0}^k \\
1 & u_{1,\,0}^k \\
1 & u_{2,\,0}^k \\
0 & u_{3,\,0}^k \\
\hline
0 & u_{0,\,1}^k \\
1 & u_{1,\,1}^k \\
1 & u_{2,\,1}^k \\
0 & u_{3,\,1}^k \\
\hline
0 & u_{0,\,2}^k \\
1 & u_{1,\,2}^k \\
1 & u_{2,\,2}^k \\
0 & u_{3,\,2}^k
\end{array}\right|
$$

The boundary vector is given by

$$\mathbf{b}_A = (c_l, 0, 0, c_r h_x^-; c_l, 0, 0, c_r h_x^-; c_l, 0, 0, c_r h_x^-)^T$$

For the second ordering of the elements in $\mathbf{u}$ it is easy to see that $\mathbf{A}_{\text{diff,C}} + \mathbf{A}_{\text{cov,C}}$ can be written in $(n+1)$-1-$(n+1)$ bandwidth form as

$$|\mathbf{A}_{\text{diff,C}} + \mathbf{A}_{\text{cov,C}}||\mathbf{u}_C^k| = \begin{array}{|c|c|c||c|}
\times \qquad \overbrace{\cdots}^{n \text{ col}} & 0 & & u_{0,\,0}^k \\
\times & 0 & & u_{0,\,1}^k \\
\times & 0 & & u_{0,\,2}^k \\
\hline
 & -\frac{d_{1,\,0}}{h_x} & \frac{d_{1,\,0}}{h_x} & u_{1,\,0}^k \\
\frac{a_{1,\,1}}{h_x^2} - \frac{d_{1,\,1}}{2h_x} & -\frac{2a_{1,\,1}}{h_x^2} & \frac{a_{1,\,1}}{h_x^2} + \frac{d_{1,\,1}}{2h_x} & u_{1,\,1}^k \\
\frac{a_{1,\,2}}{h_x^2} - \frac{d_{1,\,2}}{2h_x} & -\frac{2a_{1,\,2}}{h_x^2} & \frac{a_{1,\,2}}{h_x^2} + \frac{d_{1,\,2}}{2h_x} & u_{1,\,2}^k \\
\hline
 & -\frac{d_{2,\,0}}{h_x} & \frac{d_{2,\,0}}{h_x} & u_{2,\,0}^k \\
\frac{a_{2,\,1}}{h_x^2} - \frac{d_{2,\,1}}{2h_x} & -\frac{2a_{2,\,1}}{h_x^2} & \frac{a_{2,\,1}}{h_x^2} + \frac{d_{2,\,1}}{2h_x} & u_{2,\,1}^k \\
\frac{a_{2,\,2}}{h_x^2} - \frac{d_{2,\,2}}{2h_x} & -\frac{2a_{2,\,2}}{h_x^2} & \frac{a_{2,\,2}}{h_x^2} + \frac{d_{2,\,2}}{2h_x} & u_{2,\,2}^k \\
\hline
 & 0 & \times & u_{3,\,0}^k \\
 & 0 & \times & u_{3,\,1}^k \\
 & 0 & \underbrace{\cdots}_{n \text{ col}} \qquad \times & u_{3,\,2}^k
\end{array}$$

The tri-diagonal matrix $\mathbf{C}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}$ can be written in 1-1-1 bandwidth form as

$$|\mathbf{C}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}||\mathbf{u}_C^k| = \begin{array}{|c|c|c||c|}
\times & 0 & & u_{0,\,0}^k \\
 & 0 & & u_{0,\,1}^k \\
 & 0 & & u_{0,\,2}^k \\
\hline
 & -\frac{e_{1,\,0}}{h_y} & \frac{e_{1,\,0}}{h_y} & u_{1,\,0}^k \\
\frac{c_{1,\,1}}{h_y^2} - \frac{e_{1,\,1}}{2h_y} & -\frac{2c_{1,\,1}}{h_y^2} & \frac{c_{1,\,1}}{h_y^2} + \frac{e_{1,\,1}}{2h_y} & u_{1,\,1}^k \\
 & 0 & & u_{1,\,2}^k \\
\hline
 & -\frac{e_{2,\,0}}{h_y} & \frac{e_{2,\,0}}{h_y} & u_{2,\,0}^k \\
\frac{c_{2,\,1}}{h_y^2} - \frac{e_{2,\,1}}{2h_y} & -\frac{2c_{2,\,1}}{h_y^2} & \frac{c_{2,\,1}}{h_y^2} + \frac{e_{2,\,1}}{2h_y} & u_{2,\,1}^k \\
 & 0 & & u_{2,\,2}^k \\
\hline
 & 0 & & u_{3,\,0}^k \\
 & 0 & & u_{3,\,1}^k \\
 & 0 & \times & u_{3,\,2}^k
\end{array}$$

and $\mathbf{B}_C$ can be written in $(n+2)$-1-$(n+2)$ bandwidth form as

$$|\mathbf{B}_C||\mathbf{u}_C^k| =$$

$$\left|\begin{array}{ccccc|c|ccccc||c}
\times & \times & \times & \overbrace{\cdots}^{n-2\,\text{col}} & \times & 0 & & & & & & u^k_{0,\,0} \\
\times & \times & \times & & & 0 & & & & & & u^k_{0,\,1} \\
\times & \times & & & & 0 & & & & & & u^k_{0,\,2} \\
\hline
\times & & & & & 0 & & & & & & u^k_{1,\,0} \\
\frac{(1+\omega)b_{1,\,1}}{4h_xh_y} & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & -\frac{(1-\omega)b_{1,\,1}}{4h_xh_y} & & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & \frac{4\omega b_{1,\,1}}{4h_xh_y} & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & & -\frac{(1-\omega)b_{1,\,1}}{4h_xh_y} & -\frac{2\omega b_{1,\,1}}{4h_xh_y} & \frac{(1+\omega)b_{1,\,1}}{4h_xh_y} & u^k_{1,\,1} \\
 & & & & & 0 & & & & & & u^k_{1,\,2} \\
\hline
 & & & & & 0 & & & & & & u^k_{2,\,0} \\
\frac{(1+\omega)b_{2,\,1}}{4h_xh_y} & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & -\frac{(1-\omega)b_{2,\,1}}{4h_xh_y} & & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & \frac{4\omega b_{2,\,1}}{4h_xh_y} & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & & -\frac{(1-\omega)b_{2,\,1}}{4h_xh_y} & -\frac{2\omega b_{2,\,1}}{4h_xh_y} & \frac{(1+\omega)b_{2,\,1}}{4h_xh_y} & u^k_{2,\,1} \\
 & & & & & 0 & & & & & \times & u^k_{2,\,2} \\
\hline
 & & & & & 0 & & & & \times & \times & u^k_{3,\,0} \\
 & & & & & 0 & & & \times & \times & \times & u^k_{3,\,1} \\
 & & & & & 0 & \times & \underbrace{\cdots}_{n-2\,\text{col}} & \times & \times & \times & u^k_{3,\,2}
\end{array}\right|$$

The *modified* identity matrices are given by,

$$|\mathbf{I}_{\text{im},C}||\mathbf{u}_C^k| = \left|\begin{array}{cc|c||c}
\times & \overbrace{\cdots}^{n\,\text{col}} & 1 & u^k_{0,\,0} \\
\times & & 1 & u^k_{0,\,1} \\
\times & & 1 & u^k_{0,\,2} \\
\hline
 & & 1 & u^k_{1,\,0} \\
 & & 1 & u^k_{1,\,1} \\
 & & 1 & u^k_{1,\,2} \\
\hline
 & & 1 & u^k_{2,\,0} \\
 & & 1 & u^k_{2,\,1} \\
 & & 1 & u^k_{2,\,2} \\
\hline
-1 & & 1 & u^k_{3,\,0} \\
-1 & & 1 & u^k_{3,\,1} \\
-1 & & 1 & u^k_{3,\,2}
\end{array}\right| \quad \text{and} \quad |\mathbf{I}_{\text{ex},C}||\mathbf{u}_C^k| = \left|\begin{array}{c||c}
0 & u^k_{0,\,0} \\
0 & u^k_{0,\,1} \\
0 & u^k_{0,\,2} \\
\hline
1 & u^k_{1,\,0} \\
1 & u^k_{1,\,1} \\
1 & u^k_{1,\,2} \\
\hline
1 & u^k_{2,\,0} \\
1 & u^k_{2,\,1} \\
1 & u^k_{2,\,2} \\
\hline
0 & u^k_{3,\,0} \\
0 & u^k_{3,\,1} \\
0 & u^k_{3,\,2}
\end{array}\right|$$

The boundary vector is given by

$$\mathbf{b}_C = (c_l, c_l, c_l; 0, 0, 0; 0, 0, 0; c_r h_x^-, c_r h_x^-, c_r h_x^-)^T.$$

5.4.3 Stability under the maximum norm

For this subsection we replace the von Neumann condition in (5.27) with a Dirichlet condition. It is easy to see that if

$$\mathbf{I}_{\text{im}} - \theta_{\text{diff}}\Delta t(\mathbf{A}_{\text{diff}} + \mathbf{C}_{\text{diff}}) - \theta_{\text{cov}}\Delta t(\mathbf{A}_{\text{cov}} + \mathbf{C}_{\text{cov}}) - 2\theta_{\text{cross}}\Delta t\mathbf{B} \tag{5.38}$$

is a M-matrix, definition 3.5.1 can be applied with $\mathbf{x} = (1, \ldots, 1)^T$ to deduce that

$$||(\mathbf{I}_{\text{im}} - \theta_{\text{diff}}\Delta t(\mathbf{A}_{\text{diff}} + \mathbf{C}_{\text{diff}}) - \theta_{\text{cov}}\Delta t(\mathbf{A}_{\text{cov}} + \mathbf{C}_{\text{cov}}) - 2\theta_{\text{cross}}\Delta t\mathbf{B})^{-1}|| \leq 1.$$

Hence, as done in section 3.5, the IMEX-method will be stable if we can show that (5.38) is a M-matrix and that

$$||\mathbf{I}_{\text{ex}} + (1-\theta_{\text{diff}})\Delta t(\mathbf{A}_{\text{diff}} + \mathbf{C}_{\text{diff}}) + (1-\theta_{\text{cov}})\Delta t(\mathbf{A}_{\text{cov}} + \mathbf{C}_{\text{cov}}) + 2(1-\theta_{\text{cross}})\Delta t\mathbf{B}||_\infty \leq 1.$$

We will make use of the following theorems to find conditions under which the IMEX-scheme is stable under the maximum norm.

Lemma 2. *For any sequence of real numbers $(x_i)_{i=1}^n \in \mathbb{R}$ we have*

$$\sum_{i=1}^n (|x_i| - x_i) \geq 0 \quad \textit{and} \quad \sum_{i=1}^n (|x_i| - x_i) = 0 \iff x_i \geq 0 \quad \forall i.$$

Proof. The first statement follows directly from the fact that $|x| \geq x \quad \forall x \in \mathbb{R}$. It is easy to see that $\sum_{i=1}^n (|x_i| - x_i) = 0$ if $x_i \geq 0$ for all i, to prove the converse we will make use of induction. It is clear that the statement must be true if $n = 1$, assume the statement holds for $n = k$. Then for $n = k + 1$ we have

$$0 = \sum_{i=1}^{k+1} (|x_i| - x_i) = \sum_{i=1}^{k} (|x_i| - x_i) + |x_{k+1}| - x_{k+1}.$$

From the fact that $|x| - x \geq 0 \quad \forall x \in \mathbb{R}$ it follows that the equation above can only be true if $\sum_{i=1}^k (|x_i| - x_i) = 0$ and $|x_{k+1}| - x_{k+1} = 0$. By making use of the induction assumptions we see that it must be true that $x_i \geq 0 \quad \forall i$. □

Lemma 3. *Consider the parabolic PDE*

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

on the domain $\Omega \in \mathbb{R}^2$. Let $\overline{\Omega}$ denote the truncated computational domain and $\overline{L}$ the discrete difference operator. Let the discrete operator $\overline{L}$ be a linear combination of the Taylor expansions about the reference grid point and immediate surrounding grid points

$$\overline{L}u_{i,j}^k = \sum_{q=i-1}^{i+1} \sum_{r=j-1}^{j+1} \alpha_{q,r} u_{q,r}.$$

If $\overline{L}$ is consistent with L then it will always be the case that

$$\alpha_{i,\,j} = -\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\alpha_{q,\,r}.$$

Proof. The linear parabolic operator L can be approximated by taking linear combinations of equations (5.13) to (5.20) that removes the lower order terms

$$\begin{aligned}\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\alpha_{q,\,r}u(x_q,y_r,\tau_k) &= \left(\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\alpha_{q,\,r}\right)u(x_i,y_j,\tau_k)+Lu(x_i,y_j,\tau_k)+O(h_x^3,h_x^3h_y,h_xh_y^3,h_y^3)\\ &\approx \left(\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\alpha_{q,\,r}\right)u_{i,\,j}^k+\overline{L}u_{i,\,j}^k\end{aligned}$$

which can be rewritten as

$$\overline{L}u_{i,\,j}^k = \sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\alpha_{q,\,r}u_{q,\,r}^k - \left(\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\alpha_{q,\,r}\right)u_{i,\,j}^k. \tag{5.39}$$

□

Theorem 5.4.1. *Consider a consistent finite difference scheme*

$$\Delta_t^+u_{i,\,j}^k = \overline{L}u_{i,\,j}^{k\,+\,1}+\widetilde{L}u_{i,\,j}^k \tag{5.40}$$

where $\overline{L}$ and $\widetilde{L}$ are of the form (5.39). *If the coefficients of the difference operators $\overline{L}$ and $\widetilde{L}$, excluding the coefficient of $u_{i,\,j}$, that approximates the continues parabolic operator L is non-negative and*

$$1+\Delta t\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\overline{\alpha}_{q,\,r}\geq 0$$

$$1-\Delta t\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\widetilde{\alpha}_{q,\,r}\geq 0$$

then the θ-method is stable.

Proof. By making use of lemma 3 we see that equation (5.40) can be rewritten as

$$\begin{aligned}&-\Delta t\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\overline{\alpha}_{q,\,r}u_{q,\,r}^{k\,+\,1}+\left(1+\Delta t\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\overline{\alpha}_{q,\,r}\right)u_{i,\,j}^{k\,+\,1}\\ &=\Delta t\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\widetilde{\alpha}_{q,\,r}u_{q,\,r}^k+\left(1-\Delta t\sum_{\substack{q=i-1\\ q\neq i}}^{i\,+\,1}\sum_{\substack{r=j-1\\ r\neq j}}^{j\,+\,1}\widetilde{\alpha}_{q,\,r}\right)u_{i,\,j}^k\end{aligned}$$

Ignoring the boundary conditions for the moment, note that these equations can be written in matrix form as

$$(\mathbf{I}-\Delta t\overline{\mathbf{X}})\mathbf{u}^{k\,+\,1} = (\mathbf{I}+\Delta t\widetilde{\mathbf{X}})\mathbf{u}^k.$$

As noted in section 3.5.2 this scheme will be stable under the maximum norm if we can show that

- the implicit matrix, $(\mathbf{I}-\Delta t\overline{\mathbf{X}})$, is an M-matrix,
- $(\mathbf{I}-\Delta t\overline{\mathbf{X}})\mathbf{e}=\mathbf{e}$ where $\mathbf{e}=(1,\ldots,1)^T$,
- $||\mathbf{I}+\Delta t\widetilde{\mathbf{X}}||_\infty \leq 1$.

The second property follows directly from lemma 3. By making use of the the definition of an M-matrix, definition 3.5.1, we see that the implicit matrix implied by the equation above will only be an M-matrix when $\overline{\alpha}_{q,\,r}\geq 0 \quad \forall (q,r)\in\{(i+1,j+1),(i-1,j-1),(i+1,j-1),(i-1,j+1),(i+1,j),(i-1,j),(i,j+1),(i,j-1)\}$ and if

$$1+\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}\overline{\alpha}_{q,\,r}\geq 0.$$

By definition of the maximum norm we have

$$||\mathbf{I}+\Delta t\widetilde{\mathbf{X}}||_\infty=\max_{i,\,j}\left\{\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}|\widetilde{\alpha}_{q,\,r}|+\left|1-\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}\widetilde{\alpha}_{q,\,r}\right|\right\}.$$

Which will be less equal to one if we can show that

$$\Upsilon:=\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}|\widetilde{\alpha}_{q,\,r}|+\left|1-\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}\widetilde{\alpha}_{q,\,r}\right|\leq 1 \tag{5.41}$$

for all (i,j). Assume that there exists a node point (i,j) such that

$$1-\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}\widetilde{\alpha}_{q,\,r}<0. \tag{5.42}$$

Using lemma 2 we see that

$$\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}|\widetilde{\alpha}_{q,\,r}|\geq\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}\widetilde{\alpha}_{q,\,r}>1$$

where the last inequality follows from the assumption made in equation (5.42). This implies that

$$\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}(|\widetilde{\alpha}_{q,\,r}|+\widetilde{\alpha}_{q,\,r})>2.$$

If we substitute (5.42) in (5.41) we see that for the scheme to be stable under the maximum norm it must be true that

$$\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}(|\widetilde{\alpha}_{q,\,r}|+\widetilde{\alpha}_{q,\,r})\leq 2.$$

This contradiction implies that for the scheme to be stable we must have that

$$1-\Delta t\sum_{\substack{q=i-1\\q\neq i}}^{i+1}\sum_{\substack{r=j-1\\r\neq j}}^{j+1}\widetilde{\alpha}_{q,\,r}\geq 0. \tag{5.43}$$

Substituting in (5.41) results in

$$\Upsilon = \Delta t \sum_{\substack{q=i-1\\ q\neq i}}^{i+1} \sum_{\substack{r=j-1\\ r\neq j}}^{j+1} (|\widetilde{\alpha}_{q,\,r}| - \widetilde{\alpha}_{q,\,r}) \leq 0.$$

Using lemma 2 we see that Υ can only be equal to zero and that this will only be true when all the coefficients are non-negative. □

In order to prove stability we are going to use theorem 5.4.1, hence we need to obtain the relevant coefficients. These coefficients can obtained by expanding and rearranging equation (5.24)

$$\Delta_t^+ u_{i,\,j}^k = \overline{L}_{\theta_{\text{diff}},\,\theta_{\text{cov}},\,\theta_{\text{cross}},\,\omega} u_{i,\,j}^{k\,+\,1} + \overline{L}_{1\,-\,\theta_{\text{diff}},\,1\,-\,\theta_{\text{cov}},\,1\,-\,\theta_{\text{cross}},\,\omega} u_{i,\,j}^k$$

where

$$\begin{aligned}
\overline{L}_{\theta_{\text{diff}},\,\theta_{\text{cov}},\,\theta_{\text{cross}},\,\omega} u_{i,\,j}^k = {} & \left(\theta_{\text{diff}}\frac{a_{i,\,j}}{h_x^2} - \theta_{\text{cov}}\frac{d_{i,\,j}}{2h_x} - \theta_{\text{cross}}\omega\frac{b_{i,\,j}}{h_x h_y}\right) u_{i\,-\,1,\,j}^k + \left(\theta_{\text{diff}}\frac{a_{i,\,j}}{h_x^2} + \theta_{\text{cov}}\frac{d_{i,\,j}}{2h_x} - \theta_{\text{cross}}\omega\frac{b_{i,\,j}}{h_x h_y}\right) u_{i\,+\,1,\,j}^k \\
& + \left(\theta_{\text{diff}}\frac{c_{i,\,j}}{h_x^2} - \theta_{\text{cov}}\frac{e_{i,\,j}}{2h_x} - \theta_{\text{cross}}\omega\frac{b_{i,\,j}}{h_x h_y}\right) u_{i,\,j\,-\,1}^k + \left(\theta_{\text{diff}}\frac{c_{i,\,j}}{h_x^2} + \theta_{\text{cov}}\frac{e_{i,\,j}}{2h_x} - \theta_{\text{cross}}\omega\frac{b_{i,\,j}}{h_x h_y}\right) u_{i,\,j\,+\,1}^k \\
& + \left(\theta_{\text{cross}}(1+\omega)\frac{b_{i,\,j}}{2h_x h_y}\right) u_{i\,+\,1,\,j\,+\,1}^k + \left(\theta_{\text{cross}}(1+\omega)\frac{b_{i,\,j}}{2h_x h_y}\right) u_{i\,-\,1,\,j\,-\,1}^k \\
& + \left(\theta_{\text{cross}}(1-\omega)\frac{-b_{i,\,j}}{2h_x h_y}\right) u_{i\,-\,1,\,j\,+\,1}^k + \left(\theta_{\text{cross}}(1-\omega)\frac{-b_{i,\,j}}{2h_x h_y}\right) u_{i\,+\,1,\,j\,-\,1}^k \\
& + \left(-\theta_{\text{diff}}\frac{2a_{i,\,j}}{h_x^2} - \theta_{\text{diff}}\frac{2c_{i,\,j}}{h_y^2} + 2\theta_{\text{cross}}\omega\frac{b_{i,\,j}}{h_x h_y}\right) .u_{i,\,j}^k
\end{aligned} \tag{5.44}$$

From theorem 5.4.1 it follows that this scheme will be stable if we can prove that all coefficients, excluding the coefficients of $u_{i,\,j}^k$, are non-negative. First we consider the case when $b_{i,\,j} \neq 0$. From the fact that $\theta_{\text{diff}}, \theta_{\text{cov}}, \theta_{\text{cross}} \in [0,1]$ and $\omega \in [-1,1]$ it follows that the coefficients of $u_{i\,+\,1,\,j\,+\,1}^{k\,+\,1}$, $u_{i\,-\,1,\,j\,-\,1}^{k\,+\,1}$, $u_{i\,+\,1,\,j\,-\,1}^{k\,+\,1}$, $u_{i\,-\,1,\,j\,+\,1}^{k\,+\,1}$, $u_{i\,+\,1,\,j\,+\,1}^k$, $u_{i\,-\,1,\,j\,-\,1}^k$, $u_{i\,+\,1,\,j\,-\,1}^k$ and $u_{i\,-\,1,\,j\,+\,1}^k$ will be non-negative if and only if $\omega = \text{sgn}(b_{i,\,j})$.

For purposes of computational efficiency the implicit matrices must not be functions of the corner points i.e. the coefficients of $u_{i\,+\,1,\,j\,+\,1}^{k\,+\,1}$, $u_{i\,-\,1,\,j\,-\,1}^{k\,+\,1}$, $u_{i\,+\,1,\,j\,-\,1}^{k\,+\,1}$ and $u_{i\,-\,1,\,j\,+\,1}^{k\,+\,1}$ must be zero. The motivation for this point will become clearer when we introduce splitting. For this reason we propose to keep the cross-derivative explicit and the convection terms implicit. The difference operator becomes

$$\Delta_t^+ u_{i,\,j}^k = \overline{L}_{\theta_{\text{diff}},\,1,\,0,\,\text{sgn}(b_{i,\,j})} u_{i,\,j}^{k\,+\,1} + \overline{L}_{1\,-\,\theta_{\text{diff}},\,0,\,1,\,\text{sgn}(b_{i,\,j})} u_{i,\,j}^k$$

where $\theta_{\text{diff}} \in (0,1)$, we do not allow $\theta_{\text{diff}} = 1$ or $\theta_{\text{diff}} = 0$ since it is needed to control the positivity of the off-diagonal elements for both the implicit and explicit side of the matrix equation. Control of the positivity will be done via exponential fitting.

Exponential fitting

Using the fact that

$$f(x,y,\epsilon) = \frac{y\epsilon}{2}\coth\frac{y\epsilon}{2x} \geq \frac{|y|\epsilon}{2} \tag{5.45}$$

we deduce that the fitted scheme defined by

$$\begin{aligned}\Delta_t^+ u_{i,\,j}^k &= \left[\theta_{\text{diff}}\overline{a}_{i,\,j}\Delta_x^2 u_{i,\,j}^{k\,+\,1} + \theta_{\text{diff}}\overline{c}_{i,\,j}\Delta_y^2 u_{i,\,j}^{k\,+\,1} + d_{i,\,j}\Delta_x u_{i,\,j}^{k\,+\,1} + e_{i,\,j}\Delta_y u_{i,\,j}^{k\,+\,1}\right] \\ &\quad + \left[(1-\theta_{\text{diff}})\widetilde{a}_{i,\,j}\Delta_x^2 u_{i,\,j}^k + (1-\theta_{\text{diff}})\widetilde{c}_{i,\,j}\Delta_y^2 u_{i,\,j}^k\right] + 2b_{i,\,j}\Delta_{xy}^{\text{sgn}(b_{i,\,j})} u_{i,\,j}^k,\end{aligned}$$

where

$$\begin{aligned}\overline{a}_{i,\,j} &= f\left(a_{i,\,j}, \frac{d_{i,\,j}}{\theta_{\text{diff}}}, h_x\right) \\ \overline{c}_{i,\,j} &= f\left(c_{i,\,j}, \frac{e_{i,\,j}}{\theta_{\text{diff}}}, h_y\right) \\ \widetilde{a}_{i,\,j} &= f\left(a_{i,\,j}, \frac{2|b_{i,\,j}|}{(1-\theta_{\text{diff}})}h_y, h_x\right) \\ \widetilde{c}_{i,\,j} &= f\left(c_{i,\,j}, \frac{2|b_{i,\,j}|}{(1-\theta_{\text{diff}})}h_x, h_y\right),\end{aligned}$$

will be stable under the maximum norm if the coefficients of $u_{i,\,j}^{k\,+\,1}$ and $u_{i,\,j}^k$ are positive

$$1 + 2\theta_{\text{diff}}\frac{\overline{a}_{i,\,j}\Delta t}{h_x^2} + 2\theta_{\text{diff}}\frac{\overline{c}_{i,\,j}\Delta t}{h_y^2} > 0 \tag{5.46}$$

$$1 - 2(1-\theta_{\text{diff}})\frac{\widetilde{a}_{i,\,j}\Delta t}{h_x^2} - 2(1-\theta_{\text{diff}})\frac{\widetilde{c}_{i,\,j}\Delta t}{h_y^2} + 2\frac{|b_{i,\,j}|\Delta t}{h_x h_y} > 0. \tag{5.47}$$

The parabolic nature of the PDE ensures that the first inequality will always be true. Furthermore θ_{diff} can be chosen such that the constraint in the second equation is not too severe. Another possibility is to define separate parameters that determine implicitness of the diffusion operator for the x and y direction respectively, we will not pursue this idea further. Considering both orderings, we can rewrite these equations in matrix form as follows

$$\begin{aligned}&(\mathbf{I}_{\text{im,A}} - \theta_{\text{diff}}\Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \overline{\mathbf{C}}_{\text{diff,A}}) - \Delta t(\mathbf{A}_{\text{cov,A}} + \mathbf{C}_{\text{cov,A}}))\mathbf{u}_A^{k\,+\,1} \\ &\quad = (\mathbf{I}_{\text{ex,A}} + (1-\theta_{\text{diff}})\Delta t(\widetilde{\mathbf{A}}_{\text{diff,A}} + \widetilde{\mathbf{C}}_{\text{diff,A}}) + 2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k\,+\,1} + \mathbf{b}_A\end{aligned}$$

for the first ordering, and

$$\begin{aligned}&(\mathbf{I}_{\text{im,C}} - \theta_{\text{diff}}\Delta t(\overline{\mathbf{A}}_{\text{diff,C}} + \overline{\mathbf{C}}_{\text{diff,C}}) - \Delta t(\mathbf{A}_{\text{cov,C}} + \mathbf{C}_{\text{cov,C}}))\mathbf{u}_C^{k\,+\,1} \\ &\quad = (\mathbf{I}_{\text{ex,C}} + (1-\theta_{\text{diff}})\Delta t(\widetilde{\mathbf{A}}_{\text{diff,C}} + \widetilde{\mathbf{C}}_{\text{diff,C}}) + 2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k\,+\,1} + \mathbf{b}_C\end{aligned}$$

for the second, where the implicit matrices are M-matrices under the stability condition (5.46). As a special case we will consider the stability of the scheme when $b_{i,\,j} = 0$ on the computational domain $\overline{\Omega}$.

Stability when $b(x, y) \equiv 0$

The operator defined in (5.44) becomes

$$\begin{aligned}\overline{L}_{\theta_{\text{diff}},\,\theta_{\text{cov}},\,\theta_{\text{cross}},\,\omega} u_{i,\,j}^k &= \left(\theta_{\text{diff}}\frac{a_{i,\,j}}{h_x^2} - \theta_{\text{cov}}\frac{d_{i,\,j}}{2h_x}\right)u_{i\,-\,1,\,j}^k + \left(\theta_{\text{diff}}\frac{a_{i,\,j}}{h_x^2} + \theta_{\text{cov}}\frac{d_{i,\,j}}{2h_x}\right)u_{i\,+\,1,\,j}^k \\ &\quad + \left(\theta_{\text{diff}}\frac{c_{i,\,j}}{h_x^2} - \theta_{\text{cov}}\frac{e_{i,\,j}}{2h_x}\right)u_{i,\,j\,-\,1}^k + \left(\theta_{\text{diff}}\frac{c_{i,\,j}}{h_x^2} + \theta_{\text{cov}}\frac{e_{i,\,j}}{2h_x}\right)u_{i,\,j\,+\,1}^k \\ &\quad + \left(-\theta_{\text{diff}}\frac{2a_{i,\,j}}{h_x^2} - \theta_{\text{diff}}\frac{2c_{i,\,j}}{h_y^2}\right)u_{i,\,j}^k.\end{aligned}$$

Since the cross derivative term is zero we do not have to keep the convection part fully implicit, we only need to make sure that the convection is never fully explicit (resp. fully implicit) whenever the diffusion part is fully implicit (resp. explicit). The reason for this is that we need the diffusion part to control the positivity of the off-diagonal elements. Using (5.45) we see that the following IMEX-scheme will have positive off-diagonal terms

$$\begin{aligned}\Delta_t^+ u_{i,j}^k = &\left[\theta_{\text{diff}}\overline{a}_{i,j}\Delta_x^2 u_{i,j}^{k+1} + \theta_{\text{diff}}\overline{c}_{i,j}\Delta_y^2 u_{i,j}^{k+1} + \theta_{\text{cov}} d_{i,j}\Delta_x u_{i,j}^{k+1} + \theta_{\text{cov}} e_{i,j}\Delta_y u_{i,j}^{k+1}\right] \\ &+ \left[(1-\theta_{\text{diff}})\widetilde{a}_{i,j}\Delta_x^2 u_{i,j}^k + (1-\theta_{\text{diff}})\widetilde{c}_{i,j}\Delta_y^2 u_{i,j}^k + (1-\theta_{\text{cov}})d_{i,j}\Delta_x u_{i,j}^k + (1-\theta_{\text{cov}})e_{i,j}\Delta_y u_{i,j}^k\right],\end{aligned}$$

where

$$\begin{aligned}\overline{a}_{i,j} &= f\left(a_{i,j}, \frac{\theta_{\text{cov}} d_{i,j}}{\theta_{\text{diff}}}, h_x\right) \\ \overline{c}_{i,j} &= f\left(c_{i,j}, \frac{\theta_{\text{cov}} e_{i,j}}{\theta_{\text{diff}}}, h_y\right) \\ \widetilde{a}_{i,j} &= f\left(a_{i,j}, \frac{(1-\theta_{\text{cov}})d_{i,j}}{(1-\theta_{\text{diff}})}, h_x\right) \\ \widetilde{c}_{i,j} &= f\left(c_{i,j}, \frac{(1-\theta_{\text{cov}})e_{i,j}}{(1-\theta_{\text{diff}})}, h_y\right).\end{aligned}$$

Hence this scheme will be stable if the diagonal terms of the implicit and explicit matrices are positive. This is trivially true for the implicit matrix, for the explicit matrix we need

$$1 - (1-\theta_{\text{diff}})\frac{2a_{i,j}\Delta t}{h_x^2} - (1-\theta_{\text{diff}})\frac{2c_{i,j}\Delta t}{h_y^2} > 0$$

for the scheme to be stable under the maximum norm. Note for this *very* special case we can obtain second order convergence and a scheme that is stable under the maximum norm if we choose $\theta_{\text{diff}} = \theta_{\text{cov}} = \frac{1}{2}$ and ensure that

$$1 - \frac{a_{i,j}\Delta t}{h_x^2} - \frac{c_{i,j}\Delta t}{h_y^2} > 0$$

We can obtain a scheme that is unconditionally stable under the maximum norm if we let $\theta_{\text{diff}} = \theta_{\text{cov}} = 1$.

5.5 Convergence

Similarly as in section 3.6 we can use the Lax equivalence theorem to show convergence. Since IMEX-schemes discussed are unconditionally consistent, we deduce that they are convergent whenever they are stable.

5.6 The Yanenko scheme

For equations of the form (5.1) Yanenko, Yanenko [1971], proposed the following stable and convergent scheme

$$\frac{\widetilde{u}_{i,\,j} - u^k_{i,\,j}}{\Delta t} = a_{i,\,j}\Delta^2_x\widetilde{u}_{i,\,j} + d_{i,\,j}\Delta_x\widetilde{u}_{i,\,j} + b_{i,\,j}\Delta^\omega_{xy}u^k_{i,\,j} \tag{5.48}$$

$$\frac{u^{k\,+\,1}_{i,\,j} - \widetilde{u}_{i,\,j}}{\Delta t} = c_{i,\,j}\Delta^2_y u^{k\,+\,1}_{i,\,j} + e_{i,\,j}\Delta_y u^{k\,+\,1}_{i,\,j} + b_{i,\,j}\Delta^\omega_{xy}\widetilde{u}_{i,\,j} \tag{5.49}$$

where $a_{i,\,j} = a(x_i, y_j)$, $b_{i,\,j} = b(x_i, y_j)$ $c_{i,\,j} = c(x_i, y_j)$, $d_{i,\,j} = d(x_i, y_j)$ and $e_{i,\,j} = e(x_i, y_j)$ for $k = 0,\ 1,\ \ldots,\ l$. From these equations it is clear why we call this method a fractional step method, to proceed from one time level to the next we first compute the solution at some fictitious intermediate time level. It is clear that we only need to invert tri-diagonal matrices in the time marching procedure. In Ikonen and Toivanen [2004] they note that finite difference methods which only invert M-matrices has better stability properties. By making use of their observation and the discussions of the previous sections we can deduce that the following fitted Yanenko scheme will show better stability properties than that of the classical scheme[5]

$$\frac{\widetilde{u}_{i,\,j} - u^k_{i,\,j}}{\Delta t} = f(a_{i,\,j}, d_{i,\,j}, h_x)\Delta^2_x\widetilde{u}_{i,\,j} + d_{i,\,j}\Delta_x\widetilde{u}_{i,\,j} + b_{i,\,j}\Delta^\omega_{xy}u^k_{i,\,j} \tag{5.50}$$

$$\frac{u^{k\,+\,1}_{i,\,j} - \widetilde{u}_{i,\,j}}{\Delta t} = f(c_{i,\,j}, e_{i,\,j}, h_y)\Delta^2_y u^{k\,+\,1}_{i,\,j} + e_{i,\,j}\Delta_y u^{k\,+\,1}_{i,\,j} + b_{i,\,j}\Delta^\omega_{xy}\widetilde{u}_{i,\,j} \tag{5.51}$$

After rearranging we obtain

$$\begin{aligned}
&(-\lambda_{xx}f(a_{i,\,j}, d_{i,\,j}, h_x) + \lambda_x d_{i,\,j})u^{k\,+\,1}_{i\,-\,1,\,j} + (1 + 2f(a_{i,\,j}, d_{i,\,j}, h_x)\lambda_{xx})u^{k\,+\,1}_{i,\,j}\\
&\quad + (-\lambda_{xx}f(a_{i,\,j}, d_{i,\,j}, h_x) - \lambda_x d_{i,\,j})u^{k\,+\,1}_{i\,+\,1,\,j}\\
&= \lambda_{xy}b_{i,\,j}(1+\omega)[u^k_{i\,+\,1,\,j\,+\,1} + u^k_{i\,-\,1,\,j\,-\,1}] - \lambda_{xy}b_{i,\,j}(1-\omega)[u^k_{i\,-\,1,\,j\,+\,1} + u^k_{i\,+\,1,\,j\,-\,1}]\\
&\quad + 4\lambda_{xy}b_{i,\,j}\omega u^k_{i,\,j} - 2\lambda_{xy}b_{i,\,j}\omega[u^k_{i\,-\,1,\,j} + u^k_{i\,+\,1,\,j} + u^k_{i,\,j\,-\,1} + u^k_{i,\,j\,+\,1}]
\end{aligned} \tag{5.52}$$

and

$$\begin{aligned}
&(-\lambda_{yy}f(c_{i,\,j}, e_{i,\,j}, h_y) + \lambda_y e_{i,\,j})u^{k\,+\,1}_{i,\,j\,-\,1} + (1 + 2f(c_{i,\,j}, e_{i,\,j}, h_y)\lambda_{yy})u^{k\,+\,1}_{i,\,j}\\
&\quad + (-\lambda_{yy}f(c_{i,\,j}, e_{i,\,j}, h_y) - \lambda_y e_{i,\,j})u^{k\,+\,1}_{i,\,j\,+\,1}\\
&= \lambda_{xy}b_{i,\,j}(1+\omega)[u^k_{i\,+\,1,\,j\,+\,1} + u^k_{i\,-\,1,\,j\,-\,1}] - \lambda_{xy}b_{i,\,j}(1-\omega)[u^k_{i\,-\,1,\,j\,+\,1} + u^k_{i\,+\,1,\,j\,-\,1}]\\
&\quad + 4\lambda_{xy}b_{i,\,j}\omega u^k_{i,\,j} - 2\lambda_{xy}b_{i,\,j}\omega[u^k_{i\,-\,1,\,j} + u^k_{i\,+\,1,\,j} + u^k_{i,\,j\,-\,1} + u^k_{i,\,j\,+\,1}]
\end{aligned} \tag{5.53}$$

for $i = 1,\ 2,\ \ldots,\ m-1$, $j = 1,\ 2,\ \ldots,\ n-1$ and $k = 0,\ 1,\ \ldots,\ l-1$, where the function $f : \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+$ is defined by equation (5.45).

5.6.1 Boundary conditions

To obtain boundary conditions for the Yanenko scheme we simply use boundary conditions that coincide with those of the original problem. It is well known that the time-dependance of such boundary

[5] This follows from the observation of the previous section that exponential fitting allows for stability under the maximum norm.

conditions might result in a lower rate of convergence near the boundary than the rate of convergence on the interior of the grid, see Fairweather and Mitchell [1967], Khaliq and Twizell [1986] and Kiškis and C̃iegis [1997]. All the boundary conditions that we are going to use are either time-independent or smoothing conditions. The Dirichlet condition posed in equation (5.26) can be split as follows

$$\begin{aligned} \widetilde{u}_{0,\,j} &= c_l \quad \text{for} \quad j = 0,\, 1,\, \ldots,\, n \\ u_{0,\,j}^{k\,+\,1} &= c_l \quad \text{for} \quad j = 0,\, 1,\, \ldots,\, n \end{aligned} \tag{5.54}$$

where upwinding is used to ensure stability under the maximum norm. we From the fact that c_r is time independent it follows that the von Neumann condition in equation (5.27) can be made locally one dimensional as follows

$$\begin{aligned} \frac{\widetilde{u}_{m,\,j} - \widetilde{u}_{m\,-\,1,\,j}}{h_x^-} &= c_r \quad \text{for } j = \quad 0,\, 1,\, \ldots,\, n \\ u_{m,\,j}^{k\,+\,1} &= \widetilde{u}_{m,\,j} \quad \text{for } j = \quad 0,\, 1,\, \ldots,\, n. \end{aligned} \tag{5.55}$$

The smoothing condition at y_0 results in a two dimensional convection equation. For such problems we have the following first order, consistent and convergent locally one dimensional scheme

$$\begin{aligned} \frac{\widetilde{u}_{i,\,0} - u_{i,\,0}^k}{\Delta t} &= d_{i,\,0}\Delta_x^+\widetilde{u}_{i,\,0} \quad \text{for} \quad i = 1,\, 2,\, \ldots,\, m-1 \\ \frac{u_{i,\,0}^{k\,+\,1} - \widetilde{u}_{i,\,0}}{\Delta t} &= e_{i,\,0}\Delta_y^+ u_{i,\,0}^{k\,+\,1} \quad \text{for} \quad i = 1,\, 2,\, \ldots,\, m-1. \end{aligned} \tag{5.56}$$

As noted in section 3.7 we need to make use of upwinding to ensure the stability of this scheme under the maximum norm. The smoothing condition at y_n results in a one dimensional convection diffusion equation. As shown in chapter 3 such problems can be solved with the following stable and convergent scheme

$$\begin{aligned} \frac{\widetilde{u}_{i,\,n} - u_{i,\,n}^k}{\Delta t} &= a_{i,\,n}\Delta_x^2\widetilde{u}_{i,\,n} + d_{i,\,n}\Delta_x\widetilde{u}_{i,\,n} \quad \text{for} \quad i = 1,\, 2,\, \ldots,\, m-1 \\ u_{i,\,n}^{k\,+\,1} &= \widetilde{u}_{i,\,n} \quad i = 1,\, 2,\, \ldots,\, m-1. \end{aligned} \tag{5.57}$$

As noted in section 3.5.2 we can make use of exponential fitting to ensure that this boundary condition is stable under the maximum norm. Note that the proposed boundary conditions do not remove the tri-diagonal property from the relevant matrices. We conclude the motivation for these boundary conditions with the following quote from Kiškis and C̃iegis [1997]

> Note that purely implicit locally one dimensional methods do not require any correction of the boundary conditions, since they are unconditionally stable and preserve the optimal accuracy order.

5.7 Consistency

By adding the two fractional steps in the Yanenko scheme, (5.50) and (5.51), we obtain

$$\begin{aligned} \frac{u_{i,\,j}^{k\,+\,1} - u_{i,\,j}^k}{\Delta t} &= f(a_{i,\,j}, d_{i,\,j}, h_x)\Delta_x^2\widetilde{u}_{i,\,j} + d_{i,\,j}\Delta_x\widetilde{u}_{i,\,j} \\ &+ b_{i,\,j}\Delta_{xy}^{\omega}\widetilde{u}_{i,\,j} + b_{i,\,j}\Delta_{xy}^{\omega}u_{i,\,j}^k \\ &+ f(c_{i,\,j}, e_{i,\,j}, h_x)\Delta_x^2 u_{i,\,j}^{k\,+\,1} + e_{i,\,j}\Delta_y u_{i,\,j}^{k\,+\,1}. \end{aligned} \tag{5.58}$$

Subtracting equation (5.50) from (5.51) results in

$$\begin{aligned}\frac{u_{i,j}^{k+1} - 2\widetilde{u}_{i,j} + u_{i,j}^{k}}{\Delta t} &= -f(a_{i,j}, d_{i,j}, h_x)\Delta_x^2 \widetilde{u}_{i,j} - d_{i,j}\Delta_x \widetilde{u}_{i,j} \\ &+ b_{i,j}\Delta_{xy}^{\omega} \widetilde{u}_{i,j} - b_{i,j}\Delta_{xy}^{\omega} u_{i,j}^{k} \\ &+ f(c_{i,j}, e_{i,j}, h_x)\Delta_x^2 u_{i,j}^{k+1} + e_{i,j}\Delta_y u_{i,j}^{k+1}\end{aligned}$$

which implies that

$$\begin{aligned}\widetilde{u}_{i,j} &= \tfrac{u_{i,j}^{k+1} + u_{i,j}^{k}}{2} & -\frac{\Delta t}{2}\,[-f(a_{i,j}, d_{i,j}, h_x)\Delta_x^2 \widetilde{u}_{i,j} - d_{i,j}\Delta_x \widetilde{u}_{i,j} \\ && +b_{i,j}\Delta_{xy}^{\omega} \widetilde{u}_{i,j} - b_{i,j}\Delta_{xy}^{\omega} u_{i,j}^{k} \\ && +f(c_{i,j}, e_{i,j}, h_x)\Delta_x^2 u_{i,j}^{k+1} + e_{i,j}\Delta_y u_{i,j}^{k+1}] \\ &= \tfrac{u_{i,j}^{k+1} + u_{i,j}^{k}}{2} & +O(\Delta t).\end{aligned}$$

Substitution in equation (5.58) shows that the Yanenko scheme is equivalent to

$$\begin{aligned}\frac{u_{i,j}^{k+1} - u_{i,j}^{k}}{\Delta t} &= f(a_{i,j}, d_{i,j}, h_x)\Delta_x^2 \left(\frac{u_{i,j}^{k+1} + u_{i,j}^{k}}{2}\right) + d_{i,j}\Delta_x \left(\frac{u_{i,j}^{k+1} + u_{i,j}^{k}}{2}\right) \\ &+ b_{i,j}\Delta_{xy}^{\omega} \left(\frac{u_{i,j}^{k+1} + u_{i,j}^{k}}{2}\right) + b_{i,j}\Delta_{xy}^{\omega} u_{i,j}^{k} \\ &+ f(c_{i,j}, e_{i,j}, h_x)\Delta_x^2 u_{i,j}^{k+1} + e_{i,j}\Delta_y u_{i,j}^{k+1} + O(\Delta t)\end{aligned}$$

The truncation error becomes

$$\begin{aligned}L_{h_x h_y}^{\Delta t} v(x_i, y_j, \tau_k) &= \frac{v(x_i, y_j, \tau_{k+1}) - v(x_i, y_j, \tau_k)}{\Delta t} \\ &- \Bigg[f(a_{i,j}, d_{i,j}, h_x)\Delta_x^2 \left(\frac{v(x_i, y_j, \tau_{k+1}) + v(x_i, y_j, \tau_k)}{2}\right) \\ &+ d_{i,j}\Delta_x \left(\frac{v(x_i, y_j, \tau_{k+1}) + v(x_i, y_j, \tau_k)}{2}\right) \\ &+ b_{i,j}\Delta_{xy}^{\omega} \left(\frac{v(x_i, y_j, \tau_{k+1}) + v(x_i, y_j, \tau_k)}{2}\right) \\ &+ f(c_{i,j}, e_{i,j}, h_x)\Delta_x^2 v(x_i, y_j, \tau_{k+1}) + e_{i,j}\Delta_y v(x_i, y_j, \tau_{k+1}) \\ &+ b_{i,j}\Delta_{xy}^{\omega} v(x_i, y_j, \tau_k) \Bigg] + O(\Delta t)\end{aligned}$$

where the last term is known as the *splitting error*. Substituting the Taylor approximations derived in section 5.3 results in

$$
\begin{aligned}
&L^{\Delta t}_{h_x,\,h_y} v(x_i, y_j, \tau_k) = \\
&\quad \frac{\partial v}{\partial \tau}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \frac{1}{24}\Delta t^2 \frac{\partial^3 v}{\partial \tau^3}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \ldots \\
&\quad - f(a_{i,\,j}, d_{i,\,j}, h_x)\left[\frac{\partial^2 v}{\partial x^2}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x^2}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \ldots\right] \\
&\quad - d_{i,\,j}\left[\frac{\partial v}{\partial x}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial \tau^2 \partial x}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \ldots\right] \\
&\quad - f(c_{i,\,j}, e_{i,\,j}, h_y)\left[\frac{\partial^2 v}{\partial y^2}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^3 v}{\partial \tau^2 \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial y^2}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \ldots\right] \\
&\quad - e_{i,\,j}\left[\frac{\partial v}{\partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \tfrac{1}{2}\Delta t \frac{\partial^2 v}{\partial \tau \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^3 v}{\partial \tau^2 \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \ldots\right] \\
&\quad - 2b_{i,\,j}\left[\frac{\partial^2 v}{\partial x \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) - \tfrac{1}{4}\Delta t \frac{\partial^3 v}{\partial \tau \partial x \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \tfrac{1}{2}(\tfrac{1}{2}\Delta t)^2 \frac{\partial^4 v}{\partial \tau^2 \partial x \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + \ldots\right] \\
&\quad + O(h_x^2, h_x h_y, h_y^2, \Delta t) \\
&= \left[\frac{\partial v}{\partial \tau}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) - a_{i,\,j}\frac{\partial^2 v}{\partial x^2}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) - d_{i,\,j}\frac{\partial v}{\partial x}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}})\right. \\
&\quad \left. - c_{i,\,j}\frac{\partial^2 v}{\partial y^2}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) - e_{i,\,j}\frac{\partial v}{\partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) - 2b_{i,\,j}\frac{\partial^2 v}{\partial x \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}})\right] \\
&\quad - \tfrac{1}{2}\Delta t\left[c_{i,\,j}\frac{\partial^3 v}{\partial \tau^2 \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) + e_{i,\,j}\frac{\partial^2 v}{\partial \tau \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}}) - 2b_{i,\,j}\frac{\partial^3 v}{\partial \tau \partial x \partial y}(x_i, y_j, \tau_{k\,+\,\frac{1}{2}})\right] \qquad (5.59) \\
&\quad + O(h_x^2, h_x h_y, h_y^2, \Delta t) \\
&= O(h_x^2, h_x h_y, h_y^2, \Delta t)
\end{aligned}
$$

where we used the fact that $f(x, y, \epsilon) = x + O(\epsilon^2)$ to obtain the second equation and the fact that v is the solution of (5.1) to obtain the last equation. From the last equation it is clear that $L^{\Delta t}_{h_x,\,h_y} v(x_i, y_j, \tau_k) \to 0$ as $\Delta t, h_x, h_y \to 0$ hence we can deduce that the Yanenko scheme is consistent.

5.8 Stability

In this section we will discuss the stability of the fitted Yanenko scheme by making use of von Neumann stability analysis and the matrix method of analysis.

5.8.1 von Neumann stability analysis

Again assuming constant coefficients we use von Neumann stability analysis to obtain necessary conditions for stability. Substitution of (5.34) in (5.50) and (5.51) results in

$$
\frac{\widetilde{\gamma}}{\gamma^k} = \frac{1 - b_{ij}\frac{\Delta t}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y)}{1 + 4f(a_{i,\,j}, d_{i,\,j}, h_x)\frac{\Delta t}{h_x^2}\sin^2(\frac{\alpha h_x}{2}) - I d_{i,\,j}\frac{\Delta t}{h_x}\sin(\alpha h_x)}
$$

and

$$\frac{\gamma^{k\,+\,1}}{\widetilde{\gamma}} = \frac{1 - b_{i,\,j}\frac{\Delta t}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y)}{1 + 4f(c_{i,\,j}, e_{i,\,j}, h_y)\frac{\Delta t}{h_y^2}\sin^2(\frac{\beta h_y}{2}) - Ie_{i,\,j}\frac{\Delta t}{h_y}\sin(\beta h_y)}$$

respectively. By multiplying these two equations we obtain

$$\frac{\gamma^{k\,+\,1}}{\gamma^k} = \frac{\alpha}{\omega_1 - \omega_2 - I(\beta_1 + \beta_2)}$$

where $I = \sqrt{-1}$ and

$$\begin{aligned}\alpha &= \left(1 - b_{i,\,j}\frac{\Delta t}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y)\right)^2\\ &= 1 - 2b_{ij}\frac{\Delta t}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y) + b_{ij}^2\left(\frac{\Delta t}{h_x h_y}\right)^2\sin^2(\alpha h_x)\sin^2(\beta h_y)\\ &= 1 - 2b_{ij}\frac{\Delta t}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y) + 16b_{ij}^2\left(\frac{\Delta t}{h_x h_y}\right)^2\sin^2\left(\frac{\alpha h_x}{2}\right)\sin^2\left(\frac{\beta h_y}{2}\right)\cos^2\left(\frac{\alpha h_x}{2}\right)\cos^2\left(\frac{\beta h_y}{2}\right),\end{aligned}$$

$$\beta_1 = e_{i,\,j}\frac{\Delta t}{h_y}\sin(\beta h_y)\left(1 + 4f(a_{i,\,j}, d_{i,\,j}, h_x)\frac{\Delta t}{h_x^2}\sin^2\left(\frac{\beta h_y}{2}\right)\right),$$

$$\beta_2 = d_{i,\,j}\frac{\Delta t}{h_x}\sin(\alpha h_x)\left(1 + 4f(c_{i,\,j}, e_{i,\,j}, h_y)\frac{\Delta t}{h_y^2}\sin^2\left(\frac{\beta h_y}{2}\right)\right),$$

$$\begin{aligned}\omega_1 = 1 &+ 4f(a_{i,\,j}, d_{i,\,j}, h_x)\frac{\Delta t}{h_x^2}\sin^2\left(\frac{\alpha h_x}{2}\right) + 4f(c_{i,\,j}, e_{i,\,j}, h_y)\frac{\Delta t}{h_y^2}\sin^2\left(\frac{\beta h_y}{2}\right)\\ &+ 16f(a_{i,\,j}, d_{i,\,j}, h_x)f(c_{i,\,j}, e_{i,\,j}, h_y)\left(\frac{\Delta t}{h_x h_y}\right)^2\sin^2\left(\frac{\alpha h_x}{2}\right)\sin^2\left(\frac{\beta h_y}{2}\right),\end{aligned}$$

and

$$\omega_2 = e_{i,\,j}d_{i,\,j}\frac{\Delta t^2}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y).$$

By making use of the following identity

$$\left|\frac{a + Ib}{c + Id}\right|^2 = \frac{a^2 + b^2}{c^2 + d^2}$$

and the fact that $\omega_1\omega_2 = \beta_1\beta_2$ we deduce that

$$\left|\frac{\gamma^{k\,+\,1}}{\gamma^k}\right|^2 = \frac{\alpha^2}{\omega_1^2 + \omega_2^2 + \beta_1^2 + \beta_2^2} \leq \frac{\alpha^2}{\omega_1^2}.$$

The inequality above arises from the fact that $\omega_2^2 + \beta_1^2 + \beta_2^2 \geq 0$. Since $\alpha \geq 0$ and $\omega_1 \geq 0$ we can rewrite the equation above as

$$\left|\frac{\gamma^{k\,+\,1}}{\gamma^k}\right| \leq \frac{\alpha}{\omega_1}.$$

It follows that the scheme will be proven stable if we can show that

$$\frac{\alpha}{\omega_1} \leq 1.$$

Note that this will be true if we can prove

$$16\left(\frac{\Delta t}{h_x h_y}\right)^2 \sin^2\left(\frac{\alpha h_x}{2}\right)\sin^2\left(\frac{\beta h_y}{2}\right)\left[f(a_{i,j},d_{i,j},h_x)f(c_{i,j},e_{i,j},h_y) - b_{ij}^2\cos^2\left(\frac{\alpha h_x}{2}\right)\cos^2\left(\frac{\beta h_y}{2}\right)\right] \geq 0 \tag{5.60}$$

and

$$4f(a_{i,j},d_{i,j},h_x)\frac{\Delta t}{h_x^2}\sin^2\left(\frac{\alpha h_x}{2}\right) + 4f(c_{i,j},e_{i,j},h_y)\frac{\Delta t}{h_y^2}\sin^2\left(\frac{\beta h_y}{2}\right) + 2b_{ij}\frac{\Delta t}{h_x h_y}\sin(\alpha h_x)\sin(\beta h_y) \geq 0. \tag{5.61}$$

To show this we will need to make use of the following lemma.

Lemma 5.8.1.

$$f(a_{i,j},d_{i,j},h_x)f(c_{i,j},e_{i,j},h_y) - b_{ij}^2 > 0$$

Proof. If we can show that $f(x,y,\epsilon) \geq x$ for all $x, y, \epsilon \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^+$ then we have that

$$\begin{aligned} f(a_{i,j},d_{i,j},h_x)f(c_{i,j},e_{i,j},h_y) - b_{ij}^2 &\geq a_{i,j}c_{i,j} - b_{ij}^2 \\ &> 0 \end{aligned}$$

where the final inequity follows from the assumption that the PDE considered is parabolic. Consider the following

$$\begin{aligned} f(x,y,\epsilon) - x &= \frac{y\epsilon}{2}\coth\frac{y\epsilon}{2x} - x \\ &= x\left[\frac{y\epsilon}{2x}\coth\frac{y\epsilon}{2x} - 1\right] \\ &= x\left[\theta\frac{\cosh\theta}{\sinh\theta} - 1\right] \\ &= x\frac{\theta}{\sinh\theta}\left[\cosh\theta - \frac{\sinh\theta}{\theta}\right] \end{aligned}$$

where $\theta = \frac{y\epsilon}{2x} \in \mathbb{R}$. The following Taylor expansions of $\cosh\theta$ and $\sinh\theta$ are valid for all $\theta \in \mathbb{R}$

$$\begin{aligned} \cosh\theta &= 1 + \frac{\theta^2}{2!} + \frac{\theta^4}{4!} + \frac{\theta^6}{6!} + \dots \\ \sinh\theta &= \theta + \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \frac{\theta^7}{7!} + \dots \end{aligned}$$

By back substitution we obtain

$$\begin{aligned} f(x,y,\epsilon) - x &= x\frac{\theta}{\sinh\theta}\left[\left(\frac{1}{2!} - \frac{1}{3!}\right)\theta^2 + \left(\frac{1}{4!} - \frac{1}{5!}\right)\theta^4 + \left(\frac{1}{6!} - \frac{1}{7!}\right)\theta^6 + \dots\right] \\ &\geq 0. \end{aligned} \tag{5.62}$$

The last line follows from the fact that $\frac{\theta}{\sinh\theta} \geq 0$ and $\frac{1}{n!} - \frac{1}{(n+1)!} > 0$ for all $n \in \mathbb{N}$. □

To prove (5.60) consider the following

$$\begin{aligned} &16\left(\frac{\Delta t}{h_x h_y}\right)^2 \sin^2\left(\frac{\alpha h_x}{2}\right)\sin^2\left(\frac{\beta h_y}{2}\right)\left[f(a_{i,j},d_{i,j},h_x)f(c_{i,j},e_{i,j},h_y) - b_{ij}^2\cos^2\left(\frac{\alpha h_x}{2}\right)\cos^2\left(\frac{\beta h_y}{2}\right)\right] \\ &\geq 16\left(\frac{\Delta t}{h_x h_y}\right)^2 \sin^2\left(\frac{\alpha h_x}{2}\right)\sin^2\left(\frac{\beta h_y}{2}\right)\left[f(a_{i,j},d_{i,j},h_x)f(c_{i,j},e_{i,j},h_y) - b_{ij}^2\right] \\ &\geq 0 \end{aligned}$$

where the first inequality follows from the fact that $\cos^2\left(\frac{\alpha h_x}{2}\right)\cos^2\left(\frac{\beta h_y}{2}\right) \leq 1$ and the second from lemma 5.8.1. To see that (5.61) is true, note that it is exactly the same as equation (5.36) in lemma 5.4.1 if we make the following substitutions

$$a = 4f(a_{i,\,j}, d_{i,\,j}, h_x)\frac{\Delta t}{h_x^2}, \qquad c = 4f(c_{i,\,j}, e_{i,\,j}, h_y)\frac{\Delta t}{h_y^2} \quad \text{and} \quad b = 8b_{ij}\frac{\Delta t}{h_x h_y}.$$

From the fact that

$$\begin{aligned} 4ac - b^2 &= 64\left(\frac{\Delta t}{h_x h_y}\right)^2 \left(f(a_{i,\,j}, d_{i,\,j}, h_x)f(c_{i,\,j}, e_{i,\,j}, h_y) - b_{ij}^2\right) \\ &> 0 \end{aligned}$$

it follows that we can apply lemma 5.4.1 to deduce that equation (5.61) holds.

5.8.2 Matrix formulation

Equations (5.50) and (5.51) can be rewritten in matrix form as

$$(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))\mathbf{u}_A^{k+\frac{1}{2}} = (\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)\mathbf{u}_A^k + \mathbf{b}_{1,\,A} \tag{5.63}$$

$$(\mathbf{I}_{\text{im},C} - \Delta t(\overline{\mathbf{C}}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}))\mathbf{u}_C^{k+1} = (\mathbf{I}_{\text{ex},C} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+\frac{1}{2}} + \mathbf{b}_{2,\,C} \tag{5.64}$$

where $\overline{\mathbf{A}}_{\text{diff,A}}$ and $\overline{\mathbf{C}}_{\text{diff,C}}$ are the exponentially fitted versions of the matrices defined in section 5.4.2. Let the boundary vectors $\mathbf{b}_{1,\,A}$ and $\mathbf{b}_{2,\,C}$ be defined as

$$\begin{aligned} \mathbf{b}_{1,\,A} &= (c_l, 0, 0, c_r h_x^-; c_l, 0, 0, c_r h_x^-; c_l, 0, 0, c_r h_x^-)^T \\ \mathbf{b}_{2,\,C} &= (c_l, c_l, c_l; 0, 0, 0; 0, 0, 0; 0, 0, 0)^T \end{aligned}$$

and $\mathbf{I}_{\text{im},C}$ and $\mathbf{I}_{\text{ex},C}$ redefined as

$$|\mathbf{I}_{\text{im},C}||\mathbf{u}_C^k| = \left|\begin{array}{c|c|} 1 & u_{0,\,0}^k \\ 1 & u_{0,\,1}^k \\ 1 & u_{0,\,2}^k \\ \hline 1 & u_{1,\,0}^k \\ 1 & u_{1,\,1}^k \\ 1 & u_{1,\,2}^k \\ \hline 1 & u_{2,\,0}^k \\ 1 & u_{2,\,1}^k \\ 1 & u_{2,\,2}^k \\ \hline 1 & u_{3,\,0}^k \\ 1 & u_{3,\,1}^k \\ 1 & u_{3,\,2}^k \end{array}\right| \qquad \text{and} \qquad |\mathbf{I}_{\text{ex},C}||\mathbf{u}_C^k| = \left|\begin{array}{c|c|} 0 & u_{0,\,0}^k \\ 0 & u_{0,\,1}^k \\ 0 & u_{0,\,2}^k \\ \hline 1 & u_{1,\,0}^k \\ 1 & u_{1,\,1}^k \\ 1 & u_{1,\,2}^k \\ \hline 1 & u_{2,\,0}^k \\ 1 & u_{2,\,1}^k \\ 1 & u_{2,\,2}^k \\ \hline 1 & u_{3,\,0}^k \\ 1 & u_{3,\,1}^k \\ 1 & u_{3,\,2}^k \end{array}\right|$$

The fundamental motivation of splitting becomes clear in equations (5.63) and (5.64). For the classical schemes considered in the previous section a single matrix equation, with five non zero diagonals, needs to be solved. Splitting allows to us solve two tri-diagonal system of equations instead. The fact that tri-diagonal systems can be solved very efficiently is one of the main reasons why splitting methods became very popular. As an aside we will give a more intuitive explanation of the error that arises when splitting is implemented.

Splitting error

Even when Crank-Nicolson time marching is used in each fractional step this splitting method will still only be first order accurate, see Ikonen and Toivanen [2005a]. To see this consider (5.1) with $b(x,y) \equiv 0$ and homogeneous Dirichlet boundary conditions, this problem can be solved by applying the classical fully implicit scheme

$$(\mathbf{I} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}) - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))\mathbf{u}_A^{k+1} = \mathbf{u}_A^k \tag{5.65}$$

for $k = 0,\ 1,\ \ldots,\ l-1$. To first order we can approximate (5.65) by

$$\begin{aligned}
&(\mathbf{I} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}) - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}) \\
&\qquad + \Delta t^2(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}})(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))\mathbf{u}_A^{k+1} = \mathbf{u}_A^k \\
&\Rightarrow (\mathbf{I} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))(\mathbf{I} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))\mathbf{u}_A^{k+1} = \mathbf{u}_A^k
\end{aligned} \tag{5.66}$$

for $k = 0,\ 1,\ \ldots,\ l-1$. The last equation above can be rewritten as

$$\begin{aligned}
(\mathbf{I} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))\mathbf{u}_A^{k+\frac{1}{2}} &= \mathbf{u}_A^k \\
(\mathbf{I} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))\mathbf{u}_A^{k+1} &= \mathbf{u}_A^{k+\frac{1}{2}}.
\end{aligned}$$

This shows that the Yanenko scheme is a first order approximation of the fully implicit scheme. The second matrix equation above is not a tri-diagonal matrix equation. By reordering the elements of the solution vector we can obtain the following tri-diagonal system of equations

$$(\mathbf{I} - \Delta t(\overline{\mathbf{C}}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}))\mathbf{u}_C^{k+1} = \mathbf{u}_C^{k+\frac{1}{2}}.$$

5.8.3 Stability under the maximum norm

Elimination of the intermediate time step in equations (5.63) and (5.64) results in

$$\begin{aligned}
\mathbf{u}_A^{k+1} = {} & (\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))^{-1}(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))^{-1}(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)\mathbf{u}_A^k \\
& + (\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))^{-1}(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))^{-1}\mathbf{b}_{1,\,A} \\
& + (\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))^{-1}\mathbf{b}_{2,\,A}.
\end{aligned}$$

From this it is clear that the Yanenko method will be stable under the maximum norm if we can show that

$$||(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))^{-1}(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))^{-1}(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)||_\infty \leq 1$$

The fitting procedure ensures that $(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))$ and $(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))$ are invertible M-matrices. It follows from definition 3.5.1 with $\mathbf{x} = (1, \ldots, 1)^T$ that

$$||(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))^{-1}||_\infty \leq 1$$

and

$$||(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))^{-1}||_\infty \leq 1$$

From the fact that

$$\begin{aligned}
&||(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))^{-1}(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))^{-1}(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)||_\infty \\
\leq &||(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{C}}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}))^{-1}||_\infty ||(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)||_\infty \\
&\quad ||(\mathbf{I}_{\text{im},A} - \Delta t(\overline{\mathbf{A}}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}))^{-1}||_\infty ||(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)||_\infty \\
\leq &||(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)||_\infty^2
\end{aligned} \tag{5.67}$$

it follows that we only need to prove that

$$||(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)||_\infty^2 \leq 1$$

for the scheme to be stable. From the proof of theorem 5.4.1 it follows that all the off-diagonal elements must be non-negative for the equation above to hold. Equations (5.52) or (5.53) can be used to obtain the off-diagonal elements of $(\mathbf{I}_{\text{ex},A} + \Delta t\mathbf{B}_A)$. From these equations it is clear that the off-diagonal elements will only be non-negative when $b_{i,j} = 0, \quad \forall i, j$. Hence we are not able to prove stability under the maximum norm with the proposed method.

We have shown that the Yanenko method is unconditionally stable if we make use of von Neumann stability analysis. Since von Neumann stability analysis is, strictly speaking, not applicable to problems with non-constant coefficients, we have introduced an alternative method to prove stability namely, the matrix method of analysis under the maximum norm. Using the matrix method of analysis we have only been able to show stability for the case when $b(x, y) \equiv 0$. Nonetheless, extensive experiments with $b(x, y) \neq 0$ have as yet not resulted in a single case of instability.

5.8.4 Convergence

As done in the previous section the fictitious time step can be eliminated to obtain the desired form for the Yanenko scheme (3.37). Using the Lax equivalence theorem we deduce that the Yanenko scheme is convergent whenever it is stable.

Chapter 6

Alternative Approaches for the Two Dimensional FDM

In this chapter we extend the methods discussed in chapter 4 to two dimensions. We will attempt to derive higher order L_0-stable schemes to solve PDEs of the form

$$\frac{\partial u}{\partial \tau} = a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} \tag{6.1}$$

on the domain $(x, y) \in \Omega = [l_x, r_x] \times [l_y, r_y] \times \mathbb{R}^+$ where all the coefficients are real constants and satisfies the following inequalities

$$ac - b^2 > 0, \quad a > 0 \quad \text{and} \quad c > 0.$$

For the problem to be well posed the unknown function $u(x, y, \tau)$ must satisfy an initial condition $u(x, y, 0) = u_0(x, y) = \Psi(x, y)$ and four boundary conditions

$$\begin{aligned} u(l_x, y, \tau) &= c_l \\ u(r_x, y, \tau) &= c_r \\ u(x, l_y, \tau) &= c_b \\ u(x, r_y, \tau) &= c_t. \end{aligned}$$

Note that we simplified the problem posed in chapter 5 by assuming constant coefficients and Dirichlet boundary conditions.

6.1 Reduction to a system of ordinary differential equations

To obtain a system of ODEs we only discretisize Ω in the spatial direction and keep the time axis continuous. After truncating the domain Ω such that $(x, \tau) \in \overline{\Omega} = [x_{\min}, x_{\max}] \times [0, T]$ we obtain the following semi-discrete mesh

$$\widehat{\Omega} = \{(x_i, y_j, \tau) | i = 0,\ 1,\ \ldots,\ m, j = 0,\ 1,\ \ldots,\ n, \tau \in [0, T]\}.$$

By making use of the second order approximations derived in section 5.1 we can rewrite (6.1) as a system of ordinary differential equations

$$\frac{d\mathbf{u}_A(t)}{d\tau} = (\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)\mathbf{u}_A(\tau) + \mathbf{b}_A \tag{6.2}$$

where the solution vector $\mathbf{u}_A$ and the boundary vector, $\mathbf{b}_A$ are defined in section 5.4.2. The choice of the ordering of $\mathbf{u}$ is completely arbitrary and we could have chosen to use $\mathbf{u}_C$ instead of $\mathbf{u}_A$. The matrices $\mathbf{A}_A$ and $\mathbf{C}_A$ are given by

$$\mathbf{A}_A = \mathbf{A}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}$$
$$\mathbf{C}_A = \mathbf{C}_{\text{diff,A}} + \mathbf{C}_{\text{cov,A}}$$

where $\mathbf{A}_{\text{diff,A}}$, $\mathbf{A}_{\text{cov,A}}$, $\mathbf{C}_{\text{diff,A}}$, $\mathbf{C}_{\text{cov,A}}$ and $\mathbf{B}_A$ have the same form as in section 5.4.2 if we make the simplifying assumptions of constant coefficients and Dirichlet boundary conditions. We also assume that $\mathbf{A}_A$ and $\mathbf{C}_A$ are row reduced such that the Dirichlet boundary conditions are no longer present, as we have done for the matrix $\mathbf{A}$ of section 4.1.

6.2 A derivation of the Yanenko scheme

As shown in section 4.1.1 the solution of (6.2) is given by

$$\mathbf{u}(\tau) = -(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A + e^{\tau(\mathbf{A}_A+2\mathbf{B}_A+\mathbf{C}_A)}\left(\mathbf{\Psi}_A + (\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A\right). \tag{6.3}$$

where $\mathbf{\Psi}_A$ is the vector containing the initial data. Consider the usual uniform partition of $[0, T]$

$$0 = \tau_0 < \tau_1 < \ldots < \tau_l = T$$

where $\tau_k = k\Delta t$ and $\Delta t = \frac{T}{l}$. If the solution at time τ is known then the solution at time $\tau + \Delta t$ can be obtained as follows

$$\begin{aligned}\mathbf{u}_A(\tau + \Delta t) &= -(A_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A \\ &\quad + e^{\Delta t(\mathbf{A}_A+2\mathbf{B}_A+\mathbf{C}_A)}e^{\tau(\mathbf{A}_A+2\mathbf{B}_A+\mathbf{C}_A)}\left(\mathbf{\Psi}_A + (\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A\right) \\ &= -(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A + e^{\Delta t(\mathbf{A}_A+2\mathbf{B}_A+\mathbf{C}_A)}\left(\mathbf{u}(\tau) + (\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A\right). \end{aligned} \tag{6.4}$$

Assuming that $\mathbf{A}$ and $\mathbf{C}$ do not commute results in the following first order approximations, see Khaliq and Twizell [1986][1]

$$e^{\Delta t(\mathbf{A}+\mathbf{C})} = e^{\Delta t\mathbf{C}}e^{\Delta t\mathbf{A}} + O(\Delta t^2) \tag{6.5}$$
$$e^{\Delta t(\mathbf{A}+\mathbf{C})} = e^{\Delta t\mathbf{A}}e^{\Delta t\mathbf{C}} + O(\Delta t^2). \tag{6.6}$$

The error introduced in the equation above is known as the splitting error. We can use equation (6.6) to rewrite (6.4) in the following form

$$\begin{aligned}\mathbf{u}_A(\tau + \Delta t) &= -(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A \\ &\quad + e^{\Delta t(\mathbf{A}_A+\mathbf{B}_A)}e^{\Delta t(\mathbf{B}_A+\mathbf{C}_A)}(\mathbf{u}(\tau) + (\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A) + O(\Delta t^2).\end{aligned}$$

[1] This result can be obtained by direct substitution of $\Delta t(\mathbf{A} + \mathbf{C})$ in the definition of the exponential of a matrix.

Let $R(\Delta t\mathbf{A}, \Delta t\mathbf{C})$ be a rational approximation of $e^{\Delta t(\mathbf{A}+\mathbf{C})}$. We can obtain the solution at time τ_{k+1} if the solution at time τ_k is known via

$$\begin{aligned}\mathbf{u}_A^{k+1} = &-(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A \\ &+ R(\Delta t\mathbf{A}_A, \Delta t\mathbf{B}_A)R(\Delta t\mathbf{C}_A, \Delta t\mathbf{B}_A)(\mathbf{u}_A^k + (\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A)\end{aligned} \tag{6.7}$$

which can be rewritten as

$$\mathbf{u}_A^{k+\frac{1}{2}} = R(\Delta t\mathbf{C}_A, \Delta t\mathbf{B}_A)\mathbf{u}_A^k \tag{6.8}$$

$$\begin{aligned}\mathbf{u}_A^{k+1} = &-(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A + R(\Delta t\mathbf{A}_A, \Delta t\mathbf{B}_A)R(\Delta t\mathbf{C}_A, \Delta t\mathbf{B}_A)(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A \\ &+ R(\Delta t\mathbf{A}_A, \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}}.\end{aligned} \tag{6.9}$$

By making use of the Padé approximations of section 4.2.1 it is easy to see that

$$e^{\theta+\beta} = \frac{1+\beta}{1-\theta} + O(\theta^2, \beta^2).$$

From this we can deduce the following rational approximation for $e^{\Delta t(\mathbf{A}+\mathbf{C})}$

$$e^{\Delta t(\mathbf{A}+\mathbf{C})} = R(\Delta t\mathbf{A}, \Delta t\mathbf{C}) + O(\Delta t^2) = [\mathbf{I} - \Delta t\mathbf{A}]^{-1}[\mathbf{I} + \Delta t\mathbf{C}] + O(\Delta t^2).$$

By substituting into (6.8) and (6.9) we obtain

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{C}_A)\mathbf{u}_A^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^k \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= [-(\mathbf{I} - \Delta t\mathbf{A}_A) + (\mathbf{I} + \Delta t\mathbf{B}_A)(\mathbf{I} - \Delta t\mathbf{C}_A)^{-1}(\mathbf{I} + \Delta t\mathbf{B}_A)]\,(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)^{-1}\mathbf{b}_A \\ &\quad + (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}}.\end{aligned} \tag{6.10}$$

The coefficient of the boundary vector, $\mathbf{b}_A$, requires the inversion of $(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)$ which is a lengthy procedure and defies the point of splitting. The problem can be avoided by making use of the following first order approximation

$$\begin{aligned}&-\mathbf{I} + (\mathbf{I} - \Delta t\mathbf{A}_A)^{-1}(\mathbf{I} + \Delta t\mathbf{B}_A)(\mathbf{I} - \Delta t\mathbf{C}_A)^{-1}(\mathbf{I} + \Delta t\mathbf{B}_A) \\ =&-\mathbf{I} + e^{\Delta t(\mathbf{A}_A+2\mathbf{B}_A+\mathbf{C}_A)} + O(\Delta t^2) \\ =&\Delta t(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A) + O(\Delta t^2).\end{aligned} \tag{6.11}$$

Substituting back into (6.10) results in the Yanenko scheme

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{C}_A)\mathbf{u}_A^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^k \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}} + (\mathbf{I} - \Delta t\mathbf{A}_A)\Delta t\mathbf{b}_A.\end{aligned}$$

As done in previous chapters we can rewrite the scheme in the following form

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^k \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}} + (\mathbf{I} - \Delta t\mathbf{A}_A)\Delta t\mathbf{b}_A.\end{aligned}$$

to solve only tri-diagonal matrices.

6.2.1 L_0-stability of the Yanenko method

By making use of Gerschgorin's theorem 4.3.3 it follows that $\mathbf{B}_A$ might have non-negative eigenvalues[2]. Thus we can not apply definition 4.3.3 to deduce the stability of the Yanenko scheme whenever there is a cross-derivative term present. The explicitness of the cross derivative term will also have an adverse effect on the A-stability and L-stability of the Yanenko scheme scheme. In this section we will show that the Yanenko scheme is L_0-stable when there is no cross-derivative term. When the coefficient of the cross-derivative term is zero the Yanenko scheme can be written in compact form as

$$\mathbf{u}_A^{k+1} = [\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}\mathbf{u}_A^k + \Delta t\mathbf{b}_A. \tag{6.12}$$

By making use of exponential fitting, as done in section 4.3.1, we can ensure that the eigenvalues of $\mathbf{C}_A$ and $\mathbf{A}_A$ are real and non-positive[3]. Which implies that the symbol of the Yanenko method is given by

$$R_{\text{Yanenko}}(-z_A, -z_C) = R_{1,0}(-z_A)R_{1,0}(-z_C) = \frac{1}{(1+z_a)(1+z_c)} \tag{6.13}$$

from which it it easy to see that the Yanenko scheme is L_0-stable. The same results are obtained in Khaliq and Twizell [1986] and Ayati et al. [2006].

6.3 Extrapolation methods

In this section we extend the discussion of section 4.4 to two dimensions. L_0-stability will be proved for the case when there is no cross derivative term. Although the approximation made in (6.11) allows us to dramatically increase the speed of computation it is only first order accurate. Whenever the boundary conditions are non-zero we will not be able use extrapolation methods to increase the order of accuracy near the boundary. In this section we assume zero Dirichlet boundary conditions, i.e. $\mathbf{b}_A = \mathbf{0}$. If the solution at time τ_k is known, then we can use the exponentially fitted Yanenko scheme to proceed to time τ_{k+1}

$$\mathbf{u}_A^{k+1} = L_{\Delta t}\mathbf{u}_A^k \tag{6.14}$$

where $L_{\Delta t} = [\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} + \Delta t\mathbf{B}_A][\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}[\mathbf{I} + \Delta t\mathbf{B}_A]$ and, $\mathbf{A}_A$ and $\mathbf{C}_A$ are the exponentially fitted versions of the matrices defined in section 5.4.2 after we made the simplifying assumptions of Dirichlet boundary conditions.

6.3.1 Third order extrapolation scheme

In Khaliq and Twizell [1986] a third and fourth order accurate scheme is proposed for the simple two dimensional heat equation. In this section and the following section we will give a derivation of their

[2] Since $\mathbf{B}_A$ is a real symmetric matrix, all the eigenvalues of $\mathbf{B}_A$ are real.

[3] With exactly the same arguments as given in section 4.3.1 we can deduce that all the eigenvalues of $\mathbf{C}_C$ are real and non-positive. Let λ be an arbitrary eigenvalue of $\mathbf{C}_A$ and be $\mathbf{v}_A$ the corresponding eigenvector, ie. $\mathbf{C}_A\mathbf{v}_A = \lambda\mathbf{v}_A$. Clearly $\mathbf{C}_C\mathbf{v}_C = \lambda\mathbf{v}_C$. This shows that λ is also an eigenvalue of $\mathbf{C}_C$, hence all eigenvalues of $\mathbf{C}_A$ real and non-positive.

extensions on the Gourlay-Morris scheme and show that these modifications can be applied to two dimensional convection diffusion equations with cross-derivative terms. Say the solution at time τ_k is known then the solution at time $\tau_{k\,+\,3}$ can be obtained as follows

$$\begin{aligned}
\mathbf{u}_{A(1)}^{k\,+\,3} &= L_{\Delta t}^3 \mathbf{u}_A^k && (6.15)\\
\mathbf{u}_{A(2)}^{k\,+\,3} &= L_{2\Delta t} L_{\Delta t} \mathbf{u}_A^k && (6.16)\\
\mathbf{u}_{A(3)}^{k\,+\,3} &= L_{\Delta t} L_{2\Delta t} \mathbf{u}_A^k && (6.17)\\
\mathbf{u}_{A(4)}^{k\,+\,3} &= L_{3\Delta t} \mathbf{u}_A^k. && (6.18)
\end{aligned}$$

Define

$$\begin{aligned}
N(k, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3) &:= (\mathbf{I} + k\Delta t \mathbf{\Theta}_1 + k^2 \Delta t^2 \mathbf{\Theta}_2 + k^3 \Delta t^3 \mathbf{\Theta}_3)\\
&\quad \cdot (\mathbf{I} + k\Delta t \mathbf{\Gamma}_1 + k^2 \Delta t^2 \mathbf{\Gamma}_2 + k^3 \Delta t^3 \mathbf{\Gamma}_3)\\
&= \mathbf{I} + k\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + k^2 \Delta t^2 (\mathbf{\Theta}_2 + \mathbf{\Theta}_1 \mathbf{\Gamma}_1 + \mathbf{\Gamma}_2)\\
&\quad + k^3 \Delta t^3 (\mathbf{\Theta}_3 + \mathbf{\Theta}_1 \mathbf{\Gamma}_2 + \mathbf{\Theta}_2 \mathbf{\Gamma}_1 + \mathbf{\Gamma}_3) + O(\Delta t^4)
\end{aligned}$$

where $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$, $\mathbf{\Theta}_3$, $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$ and $\mathbf{\Gamma}_3$ are square matrices. After some tedious manipulations we obtain

$$\begin{aligned}
&N(k_1, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3) N(k_2, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3) N(k_3, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3)\\
&= \mathbf{I} + (k_1 + k_2 + k_3)\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1)\\
&\quad + \Delta t^2 [(k_3^2 + k_2^2 + k_1^2)\mathbf{\Theta}_2 + (k_1 k_3 + k_2 k_3 + k_1 k_2)\mathbf{\Gamma}_1 \mathbf{\Theta}_1 + (k_3^2 + k_2^2 + k_1^2 + k_1 k_3 + k_2 k_3 + k_1 k_2)\mathbf{\Theta}_1 \mathbf{\Gamma}_1\\
&\qquad + (k_3^2 + k_2^2 + k_1^2)\mathbf{\Gamma}_2 + (k_1 k_3 + k_2 k_3 + k_1 k_2){\mathbf{\Theta}_1}^2 + (k_1 k_3 + k_2 k_3 + k_1 k_2){\mathbf{\Gamma}_1}^2]\\
&\quad + \Delta t^3 [(k_3^3 + k_2^3 + k_1^3)\mathbf{\Theta}_3 + (k_1^2 k_2 + k_2^2 k_3 + k_1^2 k_3)\mathbf{\Theta}_2 \mathbf{\Theta}_1 + (k_1 k_2^2 + k_2 k_3^2 + k_1 k_3^2)\mathbf{\Theta}_1 \mathbf{\Theta}_2 + k_1 k_2 k_3 {\mathbf{\Theta}_1}^3\\
&\qquad + (k_3^3 + k_2^3 + k_1^3 + k_1 k_2^2 + k_2 k_3^2 + k_1 k_3^2)\mathbf{\Theta}_1 \mathbf{\Gamma}_2 + (k_1^2 k_2 + k_2^2 k_3 + k_1^2 k_3 + k_1 k_2 k_3)\mathbf{\Theta}_1 {\mathbf{\Gamma}_1}^2\\
&\qquad + (k_3^3 + k_2^3 + k_1^3 + k_1^2 k_2 + k_2^2 k_3 + k_1^2 k_3)\mathbf{\Theta}_2 \mathbf{\Gamma}_1 + (k_1 k_2^2 + k_2 k_3^2 + k_1 k_3^2 + k_1 k_2 k_3){\mathbf{\Theta}_1}^2 \mathbf{\Gamma}_1\\
&\qquad + (k_3^3 + k_2^3 + k_1^3)\mathbf{\Gamma}_3 + (k_1^2 k_2 + k_2^2 k_3 + k_1^2 k_3)\mathbf{\Gamma}_2 \mathbf{\Gamma}_1 + (k_1 k_2^2 + k_2 k_3^2 + k_1 k_3^2)\mathbf{\Gamma}_1 \mathbf{\Gamma}_2 + k_1 k_2 k_3 {\mathbf{\Gamma}_1}^3\\
&\qquad + (k_1 k_2^2 + k_2 k_3^2 + k_1 k_3^2 + k_1 k_2 k_3)\mathbf{\Gamma}_1 \mathbf{\Theta}_1 \mathbf{\Gamma}_1 + (k_1 k_2^2 + k_2 k_3^2 + k_1 k_3^2)\mathbf{\Gamma}_1 \mathbf{\Theta}_2 + k_1 k_2 k_3 \mathbf{\Gamma}_1 {\mathbf{\Theta}_1}^2\\
&\qquad + (k_1^2 k_2 + k_1^2 k_3 + k_2^2 k_3)\mathbf{\Gamma}_2 \mathbf{\Theta}_1 + k_1 k_2 k_3 {\mathbf{\Gamma}_1}^2 \mathbf{\Theta}_1 + (k_1^2 k_2 + k_1^2 k_3 + k_2^2 k_3 + k_1 k_2 k_3)\mathbf{\Theta}_1 \mathbf{\Gamma}_1 \mathbf{\Theta}_1]\\
&\quad + O(\Delta t^4)\\
&:= \Lambda_{k_1,\,k_2,\,k_3} + O(\Delta t^4)
\end{aligned}$$

where $\Lambda_{k_1,\,k_2,\,k_3} = \Lambda(k_1, k_2, k_3, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3)$. By making use of the binomial expansions in (4.12) we obtain

$$\begin{aligned}
L_{k\Delta t} &= (\mathbf{I} + k\Delta t(\mathbf{A}_A + \mathbf{B}_A) + k^2 \Delta t^2 \mathbf{A}_A(\mathbf{A}_A + \mathbf{B}_A) + k^3 \Delta t^3 \mathbf{A}_A^2 (\mathbf{A}_A + \mathbf{B}_A) + k^4 \Delta t^4 \mathbf{A}_A^3 (\mathbf{A}_A + \mathbf{B}_A))\\
&\quad \cdot (\mathbf{I} + k\Delta t(\mathbf{C}_A + \mathbf{B}_A) + k^2 \Delta t^2 \mathbf{C}_A(\mathbf{C}_A + \mathbf{B}_A) + k^3 \Delta t^3 \mathbf{C}_A^2 (\mathbf{C}_A + \mathbf{B}_A) + k^4 \Delta t^4 \mathbf{C}_A^3 (\mathbf{C}_A + \mathbf{B}_A))\\
&\quad + O(\Delta t^5) \qquad (6.19)\\
&= N(k, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3) + O(\Delta t^4)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{\Theta}_1 = \mathbf{A}_A + \mathbf{B}_A, \quad \mathbf{\Theta}_2 = \mathbf{A}_A(\mathbf{A}_A + \mathbf{B}_A), \quad \mathbf{\Theta}_3 = \mathbf{A}_A^2(\mathbf{A}_A + \mathbf{B}_A)\\
\mathbf{\Gamma}_1 = \mathbf{C}_A + \mathbf{B}_A, \quad \mathbf{\Gamma}_2 = \mathbf{C}_A(\mathbf{C}_A + \mathbf{B}_A), \quad \mathbf{\Gamma}_3 = \mathbf{C}_A^2(\mathbf{C}_A + \mathbf{B}_A).
\end{aligned}$$

Equations (6.15) to (6.18) can be written in terms of the Λ-function as follows

$$\begin{aligned}
\mathbf{u}_{A(1)}^{k\,+\,3} &= \Lambda_{1,\,1,\,1}\mathbf{u}_A^k + O(\Delta t^4)\\
\mathbf{u}_{A(2)}^{k\,+\,3} &= \Lambda_{2,\,1,\,0}\mathbf{u}_A^k + O(\Delta t^4)\\
\mathbf{u}_{A(3)}^{k\,+\,3} &= \Lambda_{1,\,2,\,0}\mathbf{u}_A^k + O(\Delta t^4)\\
\mathbf{u}_{A(4)}^{k\,+\,3} &= \Lambda_{3,\,0,\,0}\mathbf{u}_A^k + O(\Delta t^4).
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbf{u}_{A(1)}^{k\,+\,3} &= [\mathbf{I} + 3\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + \Delta t^2(3\mathbf{\Theta}_2 + 6\mathbf{\Theta}_1\mathbf{\Gamma}_1 + 3\mathbf{\Gamma}_2 + 3\mathbf{\Gamma}_1\mathbf{\Theta}_1 + 3{\mathbf{\Theta}_1}^2 + 3{\mathbf{\Gamma}_1}^2)\\
&\quad + \Delta t^3(3\mathbf{\Theta}_3 + 6\mathbf{\Theta}_2\mathbf{\Gamma}_1 + 6\mathbf{\Theta}_1\mathbf{\Gamma}_2 + 3\mathbf{\Gamma}_3 + 3\mathbf{\Theta}_1\mathbf{\Theta}_2 + 4{\mathbf{\Theta}_1}^2\mathbf{\Gamma}_1 + 3\mathbf{\Gamma}_1\mathbf{\Theta}_2 + 4\mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Gamma}_1 + 3\mathbf{\Gamma}_1\mathbf{\Gamma}_2\\
&\quad + 3\mathbf{\Theta}_2\mathbf{\Theta}_1 + 4\mathbf{\Theta}_1\mathbf{\Gamma}_1\mathbf{\Theta}_1 + 3\mathbf{\Gamma}_2\mathbf{\Theta}_1 + \mathbf{\Gamma}_1{\mathbf{\Theta}_1}^2 + {\mathbf{\Theta}_1}^3 + {\mathbf{\Gamma}_1}^2\mathbf{\Theta}_1 + {\mathbf{\Gamma}_1}^3 + 4\mathbf{\Theta}_1{\mathbf{\Gamma}_1}^2 + 3\mathbf{\Gamma}_2\mathbf{\Gamma}_1)]\mathbf{u}_A^k\\
&\quad + O(\Delta t^4)\\
\mathbf{u}_{A(2)}^{k\,+\,3} &= [\mathbf{I} + 3\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + \Delta t^2(5\mathbf{\Theta}_2 + 7\mathbf{\Theta}_1\mathbf{\Gamma}_1 + 5\mathbf{\Gamma}_2 + 2\mathbf{\Gamma}_1\mathbf{\Theta}_1 + 2{\mathbf{\Theta}_1}^2 + 2{\mathbf{\Gamma}_1}^2)\\
&\quad + \Delta t^3(9\mathbf{\Theta}_3 + 13\mathbf{\Theta}_2\mathbf{\Gamma}_1 + 11\mathbf{\Theta}_1\mathbf{\Gamma}_2 + 9\mathbf{\Gamma}_3 + 2\mathbf{\Theta}_1\mathbf{\Theta}_2 + 2{\mathbf{\Theta}_1}^2\mathbf{\Gamma}_1 + 2\mathbf{\Gamma}_1\mathbf{\Theta}_2 + 2\mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Gamma}_1 + 2\mathbf{\Gamma}_1\mathbf{\Gamma}_2\\
&\quad + 4\mathbf{\Theta}_2\mathbf{\Theta}_1 + 4\mathbf{\Theta}_1\mathbf{\Gamma}_1\mathbf{\Theta}_1 + 4\mathbf{\Gamma}_2\mathbf{\Theta}_1 + 4\mathbf{\Theta}_1{\mathbf{\Gamma}_1}^2 + 4\mathbf{\Gamma}_2\mathbf{\Gamma}_1)]\mathbf{u}_A^k\\
&\quad + O(\Delta t^4)\\
\mathbf{u}_{A(3)}^{k\,+\,3} &= [\mathbf{I} + 3\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + \Delta t^2(5\mathbf{\Theta}_2 + 7\mathbf{\Theta}_1\mathbf{\Gamma}_1 + 5\mathbf{\Gamma}_2 + 2\mathbf{\Gamma}_1\mathbf{\Theta}_1 + 2{\mathbf{\Theta}_1}^2 + 2{\mathbf{\Gamma}_1}^2)\\
&\quad + \Delta t^3(9\mathbf{\Theta}_3 + 11\mathbf{\Theta}_2\mathbf{\Gamma}_1 + 13\mathbf{\Theta}_1\mathbf{\Gamma}_2 + 9\mathbf{\Gamma}_3 + 4\mathbf{\Theta}_1\mathbf{\Theta}_2 + 4{\mathbf{\Theta}_1}^2\mathbf{\Gamma}_1 + 4\mathbf{\Gamma}_1\mathbf{\Theta}_2 + 4\mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Gamma}_1 + 4\mathbf{\Gamma}_1\mathbf{\Gamma}_2\\
&\quad + 2\mathbf{\Theta}_2\mathbf{\Theta}_1 + 2\mathbf{\Theta}_1\mathbf{\Gamma}_1\mathbf{\Theta}_1 + 2\mathbf{\Gamma}_2\mathbf{\Theta}_1 + 2\mathbf{\Theta}_1{\mathbf{\Gamma}_1}^2 + 2\mathbf{\Gamma}_2\mathbf{\Gamma}_1)]\mathbf{u}_A^k\\
&\quad + O(\Delta t^4)\\
\mathbf{u}_{A(4)}^{k\,+\,3} &= [\mathbf{I} + 3\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + \Delta t^2(9\mathbf{\Theta}_2 + 9\mathbf{\Theta}_1\mathbf{\Gamma}_1 + 9\mathbf{\Gamma}_2)\\
&\quad + \Delta t^3(27\mathbf{\Theta}_3 + 27\mathbf{\Theta}_2\mathbf{\Gamma}_1 + 27\mathbf{\Theta}_1\mathbf{\Gamma}_2 + 27\mathbf{\Gamma}_3)]\mathbf{u}_A^k + O(\Delta t^4).
\end{aligned}$$

From the definition of the exponential of a matrix

$$\begin{aligned}
e^{3\Delta t(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)} &= e^{3\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1)}\\
&= \mathbf{I} + 3\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + \tfrac{9}{2}\Delta t^2({\mathbf{\Theta}_1}^2 + \mathbf{\Gamma}_1\mathbf{\Theta}_1 + \mathbf{\Theta}_1\mathbf{\Gamma}_1 + {\mathbf{\Gamma}_1}^2)\\
&\quad + \tfrac{9}{2}\Delta t^3({\mathbf{\Theta}_1}^2\mathbf{\Gamma}_1 + \mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Gamma}_1 + \mathbf{\Theta}_1\mathbf{\Gamma}_1\mathbf{\Theta}_1 + \mathbf{\Gamma}_1{\mathbf{\Theta}_1}^2\\
&\quad + {\mathbf{\Theta}_1}^3 + {\mathbf{\Gamma}_1}^2\mathbf{\Theta}_1 + {\mathbf{\Gamma}_1}^3 + \mathbf{\Theta}_1{\mathbf{\Gamma}_1}^2) + O(\Delta t^4)
\end{aligned}$$

we see that the correct approximation to solution at time $\tau_{k\,+\,3}$ is given by

$$\begin{aligned}
\mathbf{u}_A^{k\,+\,3} &= e^{3\Delta t(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)}\mathbf{u}_A^k\\
&= [\mathbf{I} + 3\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + \tfrac{9}{2}\Delta t^2({\mathbf{\Theta}_1}^2 + \mathbf{\Gamma}_1\mathbf{\Theta}_1 + \mathbf{\Theta}_1\mathbf{\Gamma}_1 + {\mathbf{\Gamma}_1}^2)\\
&\quad + \tfrac{9}{2}\Delta t^3({\mathbf{\Theta}_1}^2\mathbf{\Gamma}_1 + \mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Gamma}_1 + \mathbf{\Theta}_1\mathbf{\Gamma}_1\mathbf{\Theta}_1 + \mathbf{\Gamma}_1{\mathbf{\Theta}_1}^2\\
&\quad + {\mathbf{\Theta}_1}^3 + {\mathbf{\Gamma}_1}^2\mathbf{\Theta}_1 + {\mathbf{\Gamma}_1}^3 + \mathbf{\Theta}_1{\mathbf{\Gamma}_1}^2) + \ldots]\mathbf{u}_A^k.
\end{aligned}$$

From this it is clear that none of the methods match the second or third order terms correctly. But we might be able to match the second and third order terms by making use of the following linear combination

$$\mathbf{u}_A^{k\,+\,3} = \eta_1\mathbf{u}_{A(1)}^{k\,+\,3} + \eta_2\mathbf{u}_{A(2)}^{k\,+\,3} + \eta_3\mathbf{u}_{A(3)}^{k\,+\,3} + \eta_4\mathbf{u}_{A(4)}^{k\,+\,3}$$

where η_1, η_2, η_3 and η_4 satisfies

$$\begin{aligned}
\eta_1 + \eta_2 + \eta_3 + \eta_4 &= 1 \\
3\eta_1 + 5\eta_2 + 5\eta_3 + 9\eta_4 &= 0 \\
6\eta_1 + 7\eta_2 + 7\eta_3 + 9\eta_4 &= \tfrac{9}{2} \\
3\eta_1 + 2\eta_2 + 2\eta_3 &= \tfrac{9}{2} \\
3\eta_1 + 9\eta_2 + 9\eta_3 + 27\eta_4 &= 0 \\
6\eta_1 + 13\eta_2 + 11\eta_3 + 27\eta_4 &= 0 \\
6\eta_1 + 11\eta_2 + 13\eta_3 + 27\eta_4 &= 0 \\
3\eta_1 + 2\eta_2 + 4\eta_3 &= 0 \\
4\eta_1 + 2\eta_2 + 4\eta_3 &= \tfrac{9}{2} \\
3\eta_1 + 4\eta_2 + 2\eta_3 &= 0 \\
4\eta_1 + 4\eta_2 + 2\eta_3 &= \tfrac{9}{2} \\
\eta_1 &= \tfrac{9}{2}.
\end{aligned}$$

This system of equations is in fact not over specified and has the following unique solution

$$\eta_1 = \tfrac{9}{2}, \quad \eta_2 = \eta_3 = -\tfrac{9}{4} \quad \text{and} \quad \eta_4 = 1.$$

The third order accurate extrapolation algorithm follows

$$\begin{aligned}
(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\
(\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}} \\
(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{3}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+1} \\
(\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+2} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{3}{2}} \\
(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{5}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+2} \\
(\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_{A(1)}^{k+3} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{5}{2}}
\end{aligned}$$

$$\begin{aligned}
(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\
(\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}} \\
(\mathbf{I} - 2\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+2} &= (\mathbf{I} + 2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+1} \\
(\mathbf{I} - 2\Delta t\mathbf{A}_A)\mathbf{u}_{A(2)}^{k+3} &= (\mathbf{I} + 2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+2}
\end{aligned}$$

$$\begin{aligned}
(\mathbf{I} - 2\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+1} &= (\mathbf{I} + 2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\
(\mathbf{I} - 2\Delta t\mathbf{A}_A)\mathbf{u}_A^{k+2} &= (\mathbf{I} + 2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+1} \\
(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{5}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+2} \\
(\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_{A(3)}^{k+3} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{5}{2}}
\end{aligned}$$

$$
\begin{aligned}
(\mathbf{I} - 3\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{3}{2}} &= (\mathbf{I} + 3\Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\
(\mathbf{I} - 3\Delta t\mathbf{A}_A)\mathbf{u}_{A(4)}^{k+3} &= (\mathbf{I} + 3\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{3}{2}}
\end{aligned}
$$

and

$$
\mathbf{u}_A^{k+3} = \tfrac{9}{2}\mathbf{u}_{A(1)}^{k+3} - \tfrac{9}{4}\mathbf{u}_{A(2)}^{k+3} - \tfrac{9}{4}\mathbf{u}_{A(3)}^{k+3} + \mathbf{u}_{A(4)}^{k+3}.
$$

L_0-stability of the third order Khaliq-Twizwell scheme

By removing the cross derivative term we obtain a version Khaliq-Twizwell scheme that uses exponential fitting to handle the convection terms. It is easy to see that the third order Khaliq-Twizwell scheme can be written in compact form as

$$
\mathbf{u}_A^{k+3} = R_{\text{KT3}}(\Delta t\mathbf{A}_A, \Delta t\mathbf{C}_A)\mathbf{u}_A^{k}
$$

where

$$
\begin{aligned}
R_{\text{KT3}}(\Delta t\mathbf{A}_A, \Delta t\mathbf{C}_A) = {} & \tfrac{9}{2}\left([\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}\right)^3 \\
& - \tfrac{9}{4}[\mathbf{I} - 2\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 2\Delta t\mathbf{C}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1} \\
& - \tfrac{9}{4}[\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}[\mathbf{I} - 2\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 2\Delta t\mathbf{C}_A]^{-1} \\
& + [\mathbf{I} - 3\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 3\Delta t\mathbf{C}_A]^{-1}.
\end{aligned}
$$

From this we can deduce that the symbol of the third order Khaliq-Twizwell scheme is given by

$$
\begin{aligned}
R_{\text{KT3}}(-z_A, -z_C) = {} & \frac{9}{2(1+z_A)^3(1+z_C)^3} - \frac{9}{2(1+2z_A)(1+2z_C)(1+z_A)(1+z_C)} \\
& + \frac{1}{(1+3z_A)(1+3z_C)}.
\end{aligned}
$$

From figure 6.1 we see that

$$
\max_{z_A,\, z_C \,\geq\, 0} |R_{\text{KT3}}(-z_A, -z_C)| \leq 1
$$

and

$$
\lim_{z_A,\, z_C \,\to\, \infty} R_{\text{KT3}}(-z_A, -z_C) = 0.
$$

Hence we can deduce that this scheme is indeed L_0-stable.

6.3.2 Fourth order extrapolation scheme

The derivation of the fourth order accurate scheme will have a similar form to that of the third order scheme in section 6.3.1. Say the solution at time τ_k is known then the solution at time τ_{k+4} can be

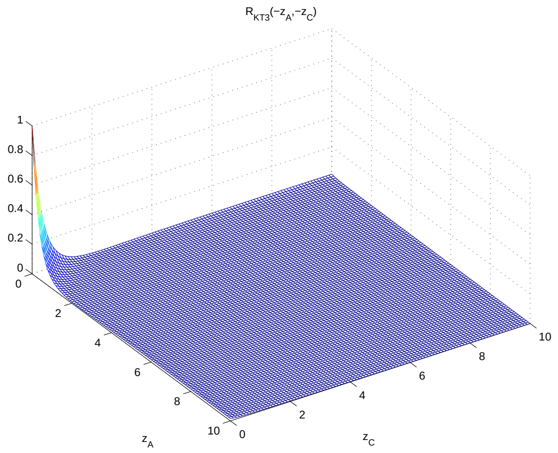


Figure 6.1: The symbol of the third order Khaliq-Twizwell scheme, $R_{\text{KT3}}(-z_A, -z_C)$.

obtained as follows

$$\mathbf{u}_{A(1)}^{k+4} = L_{\Delta t}^4 \mathbf{u}_A^k \tag{6.20}$$

$$\mathbf{u}_{A(2)}^{k+4} = L_{\Delta t} L_{3\Delta t} \mathbf{u}_A^k \tag{6.21}$$

$$\mathbf{u}_{A(3)}^{k+4} = L_{3\Delta t} L_{\Delta t} \mathbf{u}_A^k \tag{6.22}$$

$$\mathbf{u}_{A(4)}^{k+4} = L_{\Delta t}^2 L_{2\Delta t} \mathbf{u}_A^k \tag{6.23}$$

$$\mathbf{u}_{A(5)}^{k+4} = L_{2\Delta t} L_{\Delta t}^2 \mathbf{u}_A^k \tag{6.24}$$

$$\mathbf{u}_{A(6)}^{k+4} = L_{2\Delta t}^2 \mathbf{u}_A^k \tag{6.25}$$

$$\mathbf{u}_{A(7)}^{k+4} = L_{\Delta t} L_{2\Delta t} L_{\Delta t} \mathbf{u}_A^k \tag{6.26}$$

$$\mathbf{u}_{A(8)}^{k+4} = L_{4\Delta t} \mathbf{u}_A^k. \tag{6.27}$$

Define

$$\begin{aligned} N(k, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Theta}_4, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3, \mathbf{\Gamma}_4) &:= (\mathbf{I} + k\Delta t \mathbf{\Theta}_1 + k^2\Delta t^2 \mathbf{\Theta}_2 + k^3 \Delta t^3 \mathbf{\Theta}_3 + k^4 \Delta t^4 \mathbf{\Theta}_4) \\ &\quad \cdot (\mathbf{I} + k\Delta t \mathbf{\Gamma}_1 + k^2 \Delta t^2 \mathbf{\Gamma}_2 + k^3 \Delta t^3 \mathbf{\Gamma}_3 + k^4 \Delta t^4 \mathbf{\Gamma}_4) \\ &= \mathbf{I} + k\Delta t(\mathbf{\Theta}_1 + \mathbf{\Gamma}_1) + k^2 \Delta t^2 (\mathbf{\Theta}_2 + \mathbf{\Theta}_1 \mathbf{\Gamma}_1 + \mathbf{\Gamma}_2) \\ &\quad + k^3 \Delta t^3 (\mathbf{\Theta}_3 + \mathbf{\Theta}_1 \mathbf{\Gamma}_2 + \mathbf{\Theta}_2 \mathbf{\Gamma}_1 + \mathbf{\Gamma}_3) \\ &\quad + k^4 \Delta t^4 (\mathbf{\Theta}_4 + \mathbf{\Theta}_1 \mathbf{\Gamma}_3 + \mathbf{\Theta}_3 \mathbf{\Theta}_3 + \mathbf{\Theta}_3 \mathbf{\Gamma}_1 + \mathbf{\Gamma}_4) + O(\Delta t^5) \end{aligned}$$

where $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$, $\mathbf{\Theta}_3$, $\mathbf{\Theta}_4$, $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, $\mathbf{\Gamma}_3$ and $\mathbf{\Gamma}_4$ are square matrices. Define a new Λ-function which is accurate to fourth order via the following[4]

$$\begin{aligned} &N(k_1, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Theta}_4, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3, \mathbf{\Gamma}_4) N(k_2, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Theta}_4, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3, \mathbf{\Gamma}_4) \\ &\quad \cdot N(k_3, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Theta}_4, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3, \mathbf{\Gamma}_4) N(k_4, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \mathbf{\Theta}_3, \mathbf{\Theta}_4, \mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \mathbf{\Gamma}_3, \mathbf{\Gamma}_4) \\ &= \Lambda_{k_1, k_2, k_3, k_4} + O(\Delta t^5) \end{aligned} \tag{6.28}$$

[4] We will not give the explicit form of the fourth order accurate Λ-function as we did for the third order accurate Λ-function in section 6.3.1, since the function expression is extremely long for this case and does not give extra intuition.

where $\Lambda_{k_1,\,k_2,\,k_3,\,k_4} = \Lambda(k_1, k_2, k_2, k_3, k_4, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3, \boldsymbol{\Theta}_4, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_3, \boldsymbol{\Gamma}_4)$. By making use of equation (6.19) we obtain

$$L_{k\Delta t} = N(k, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3, \boldsymbol{\Theta}_4, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\Gamma}_3, \boldsymbol{\Gamma}_4) + O(\Delta t^5)$$

where

$$\begin{aligned} \boldsymbol{\Theta}_1 = \mathbf{A}_A + \mathbf{B}_A, \quad \boldsymbol{\Theta}_2 = \mathbf{A}_A(\mathbf{A}_A + \mathbf{B}_A), \quad \boldsymbol{\Theta}_3 = \mathbf{A}_A^2(\mathbf{A}_A + \mathbf{B}_A), \quad \boldsymbol{\Theta}_4 = \mathbf{A}_A^3(\mathbf{A}_A + \mathbf{B}_A) \\ \boldsymbol{\Gamma}_1 = \mathbf{C}_A + \mathbf{B}_A, \quad \boldsymbol{\Gamma}_2 = \mathbf{C}_A(\mathbf{C}_A + \mathbf{B}_A), \quad \boldsymbol{\Gamma}_3 = \mathbf{C}_A^2(\mathbf{C}_A + \mathbf{B}_A), \quad \boldsymbol{\Gamma}_4 = \mathbf{C}_A^3(\mathbf{C}_A + \mathbf{B}_A). \end{aligned}$$

Equations (6.20) to (6.27) can be written in terms of the Λ-function as follows

$$\begin{aligned} \mathbf{u}_{A(1)}^{k\,+\,4} &= \Lambda_{1,\,1,\,1,\,1}\mathbf{u}_A^k + O(\Delta t^5) \\ \mathbf{u}_{A(2)}^{k\,+\,4} &= \Lambda_{1,\,3,\,0,\,0}\mathbf{u}_A^k + O(\Delta t^5) \\ \mathbf{u}_{A(3)}^{k\,+\,4} &= \Lambda_{3,\,1,\,0,\,0}\mathbf{u}_A^k + O(\Delta t^5) \\ \mathbf{u}_{A(4)}^{k\,+\,4} &= \Lambda_{1,\,1,\,2,\,0}\mathbf{u}_A^k + O(\Delta t^5) \\ \mathbf{u}_{A(5)}^{k\,+\,4} &= \Lambda_{2,\,1,\,1,\,0}\mathbf{u}_A^k + O(\Delta t^5) \\ \mathbf{u}_{A(6)}^{k\,+\,4} &= \Lambda_{2,\,2,\,0,\,0}\mathbf{u}_A^k + O(\Delta t^5) \\ \mathbf{u}_{A(7)}^{k\,+\,4} &= \Lambda_{1,\,2,\,1,\,0}\mathbf{u}_A^k + O(\Delta t^5) \\ \mathbf{u}_{A(8)}^{k\,+\,4} &= \Lambda_{4,\,0,\,0,\,0}\mathbf{u}_A^k + O(\Delta t^5). \end{aligned}$$

The approximation of the solution at time $\tau_{k\,+\,4}$ with the correct higher order terms is given by

$$\mathbf{u}_A^{k\,+\,4} = e^{4\Delta t(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)}\mathbf{u}_A^k.$$

The following tables give the coefficients of the first, second, third and fourth order terms of the relevant Λ-functions and $e^{4\Delta t(\mathbf{A}_A + 2\mathbf{B}_A + \mathbf{C}_A)}$

$O(\cdot)$		$\Lambda_{1,1,1,1}$	$\Lambda_{1,3,0,0}$	$\Lambda_{3,1,0,0}$	$\Lambda_{1,1,2,0}$	$\Lambda_{2,1,1,0}$	$\Lambda_{2,2,0,0}$	$\Lambda_{1,2,1,0}$	$\Lambda_{4,0,0,0}$	$e^{4\Delta t(\mathbf{A}_A+2\mathbf{B}_A+\mathbf{C}_A)}$
	$\mathbf{I}$	1	1	1	1	1	1	1	1	1
Δt	$\boldsymbol{\Gamma}_1 + \boldsymbol{\Theta}_1$	4	4	4	4	4	4	4	4	4
Δt^2	$\boldsymbol{\Gamma}_2$	4	10	10	6	6	8	6	16	0
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_1$	10	13	13	11	11	12	11	16	8
	$\boldsymbol{\Theta}_2$	4	10	10	6	6	8	6	16	0
	${\boldsymbol{\Gamma}_1}^2$	6	3	3	5	5	4	5	0	8
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_1$	6	3	3	5	5	4	5	0	8
	${\boldsymbol{\Theta}_1}^2$	6	3	3	5	5	4	5	0	8

$O(\cdot)$		$\Lambda_{1,1,1,1}$	$\Lambda_{1,3,0,0}$	$\Lambda_{3,1,0,0}$	$\Lambda_{1,1,2,0}$	$\Lambda_{2,1,1,0}$	$\Lambda_{2,2,0,0}$	$\Lambda_{1,2,1,0}$	$\Lambda_{4,0,0,0}$	$e^{4\Delta t(\cdot)}$
Δt^3	$\boldsymbol{\Gamma}_3$	4	28	28	10	10	16	10	64	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_2$	10	37	31	19	15	24	17	64	0
	$\boldsymbol{\Theta}_2\boldsymbol{\Gamma}_1$	10	31	37	15	19	24	17	64	0
	$\boldsymbol{\Theta}_3$	4	28	28	10	10	16	10	64	0
	$\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_2$	6	9	3	9	5	8	7	0	0
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1$	10	9	3	11	7	8	9	0	$\frac{32}{3}$
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_2$	6	9	3	9	5	8	7	0	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2$	6	9	3	9	5	8	7	0	0
	$\boldsymbol{\Theta}_1{}^2\boldsymbol{\Gamma}_1$	10	9	3	11	7	8	9	0	$\frac{32}{3}$
	$\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_1$	6	3	9	5	9	8	7	0	0
	$\boldsymbol{\Gamma}_2\boldsymbol{\Theta}_1$	6	3	9	5	9	8	7	0	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1{}^2$	10	3	9	7	11	8	9	0	$\frac{32}{3}$
	$\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_1$	10	3	9	7	11	8	9	0	$\frac{32}{3}$
	$\boldsymbol{\Theta}_2\boldsymbol{\Theta}_1$	6	3	9	5	9	8	7	0	0
	$\boldsymbol{\Gamma}_1{}^3$	4	0	0	2	2	0	2	0	$\frac{32}{3}$
	$\boldsymbol{\Gamma}_1{}^2\boldsymbol{\Theta}_1$	4	0	0	2	2	0	2	0	$\frac{32}{3}$
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_1{}^2$	4	0	0	2	2	0	2	0	$\frac{32}{3}$
	$\boldsymbol{\Theta}_1{}^3$	4	0	0	2	2	0	2	0	$\frac{32}{3}$
Δt^4	$\boldsymbol{\Gamma}_4$	4	82	82	18	18	32	18	256	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_3$	10	109	85	35	23	48	29	256	0
	$\boldsymbol{\Theta}_2\boldsymbol{\Gamma}_2$	10	91	91	27	27	48	27	256	0
	$\boldsymbol{\Theta}_3\boldsymbol{\Gamma}_1$	10	85	109	23	35	48	29	256	0
	$\boldsymbol{\Theta}_4$	4	82	82	18	18	32	18	256	0
	$\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_3$	6	27	3	17	5	16	11	0	0
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_2$	10	27	3	21	7	16	13	0	0
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_2\boldsymbol{\Gamma}_1$	10	27	3	19	7	16	15	0	0
	$\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_3$	6	27	3	17	5	16	11	0	0
	$\boldsymbol{\Theta}_1{}^2\boldsymbol{\Gamma}_2$	10	27	3	21	7	16	13	0	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2\boldsymbol{\Gamma}_1$	10	27	3	19	7	16	15	0	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Theta}_3$	6	27	3	17	5	16	11	0	0
	$\boldsymbol{\Gamma}_2{}^2$	6	9	9	9	9	16	9	0	0
	$\boldsymbol{\Gamma}_2\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1$	10	9	9	11	13	16	11	0	0
	$\boldsymbol{\Gamma}_2\boldsymbol{\Theta}_2$	6	9	9	9	9	16	9	0	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1\boldsymbol{\Gamma}_2$	10	9	9	13	11	16	11	0	0
	$\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1$	15	9	9	15	15	16	13	0	$\frac{32}{3}$
	$\boldsymbol{\Theta}_1\boldsymbol{\Gamma}_1\boldsymbol{\Theta}_2$	10	9	9	13	11	16	11	0	0

$O(\cdot)$		$\Lambda_{1,\,1,\,1,\,1}$	$\Lambda_{1,\,3,\,0,\,0}$	$\Lambda_{3,\,1,\,0,\,0}$	$\Lambda_{1,\,1,\,2,\,0}$	$\Lambda_{2,\,1,\,1,\,0}$	$\Lambda_{2,\,2,\,0,\,0}$	$\Lambda_{1,\,2,\,1,\,0}$	$\Lambda_{4,\,0,\,0,\,0}$	$e^{4\Delta t(\cdot)}$
Δt^4	$\mathbf{\Theta}_2\mathbf{\Theta}_1\mathbf{\Gamma}_1$	10	9	9	11	13	16	11	0	0
	$\mathbf{\Theta}_2{}^2$	6	9	9	9	9	16	9	0	0
	$\mathbf{\Gamma}_3\mathbf{\Gamma}_1$	6	3	27	5	17	16	11	0	0
	$\mathbf{\Gamma}_3\mathbf{\Theta}_1$	6	3	27	5	17	16	11	0	0
	$\mathbf{\Theta}_1\mathbf{\Gamma}_2\mathbf{\Gamma}_1$	10	3	27	7	19	16	15	0	0
	$\mathbf{\Theta}_1\mathbf{\Gamma}_2\mathbf{\Theta}_1$	10	3	27	7	19	16	15	0	0
	$\mathbf{\Theta}_2\mathbf{\Gamma}_1{}^2$	10	3	27	7	21	16	13	0	0
	$\mathbf{\Theta}_2\mathbf{\Gamma}_1\mathbf{\Theta}_1$	10	3	27	7	21	16	13	0	0
	$\mathbf{\Theta}_3\mathbf{\Theta}_1$	6	3	27	5	17	16	11	0	0
	$\mathbf{\Gamma}_1{}^2\mathbf{\Gamma}_2$	4	0	0	4	2	0	2	0	0
	$\mathbf{\Gamma}_1{}^2\mathbf{\Theta}_1\mathbf{\Gamma}_1$	5	0	0	4	2	0	2	0	$\frac{32}{3}$
	$\mathbf{\Gamma}_1{}^2\mathbf{\Theta}_2$	4	0	0	4	2	0	2	0	0
	$\mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Theta}_2$	4	0	0	4	2	0	2	0	0
	$\mathbf{\Theta}_1{}^2\mathbf{\Theta}_2$	4	0	0	4	2	0	2	0	0
	$\mathbf{\Gamma}_1\mathbf{\Theta}_1{}^2\mathbf{\Gamma}_1$	5	0	0	4	2	0	2	0	$\frac{32}{3}$
	$\mathbf{\Theta}_1{}^3\mathbf{\Gamma}_1$	5	0	0	4	2	0	2	0	$\frac{32}{3}$
	$\mathbf{\Gamma}_1\mathbf{\Gamma}_2\mathbf{\Gamma}_1$	4	0	0	2	2	0	4	0	0
	$\mathbf{\Gamma}_1\mathbf{\Gamma}_2\mathbf{\Theta}_1$	4	0	0	2	2	0	4	0	0
	$\mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Gamma}_1{}^2$	5	0	0	2	2	0	4	0	$\frac{32}{3}$
	$\mathbf{\Theta}_1{}^2\mathbf{\Gamma}_1{}^2$	5	0	0	2	2	0	4	0	$\frac{32}{3}$
	$\mathbf{\Gamma}_1\mathbf{\Theta}_1\mathbf{\Gamma}_1\mathbf{\Theta}_1$	5	0	0	2	2	0	4	0	$\frac{32}{3}$
	$\mathbf{\Theta}_1{}^2\mathbf{\Gamma}_1\mathbf{\Theta}_1$	5	0	0	2	2	0	4	0	$\frac{32}{3}$
	$\mathbf{\Gamma}_1\mathbf{\Theta}_2\mathbf{\Theta}_1$	4	0	0	2	2	0	4	0	0
	$\mathbf{\Theta}_1\mathbf{\Theta}_2\mathbf{\Theta}_1$	4	0	0	2	2	0	4	0	0
	$\mathbf{\Gamma}_2\mathbf{\Gamma}_1{}^2$	4	0	0	2	4	0	2	0	0
	$\mathbf{\Gamma}_2\mathbf{\Gamma}_1\mathbf{\Theta}_1$	4	0	0	2	4	0	2	0	0
	$\mathbf{\Gamma}_2\mathbf{\Theta}_1{}^2$	4	0	0	2	4	0	2	0	0
	$\mathbf{\Theta}_1\mathbf{\Gamma}_1{}^3$	5	0	0	2	4	0	2	0	$\frac{32}{3}$
	$\mathbf{\Theta}_1\mathbf{\Gamma}_1{}^2\mathbf{\Theta}_1$	5	0	0	2	4	0	2	0	$\frac{32}{3}$
	$\mathbf{\Theta}_1\mathbf{\Gamma}_1\mathbf{\Theta}_1{}^2$	5	0	0	2	4	0	2	0	$\frac{32}{3}$
	$\mathbf{\Theta}_2\mathbf{\Theta}_1{}^2$	4	0	0	2	4	0	2	0	0
	$\mathbf{\Gamma}_1{}^4$	1	0	0	0	0	0	0	0	$\frac{32}{3}$
	$\mathbf{\Gamma}_1{}^3\mathbf{\Theta}_1$	1	0	0	0	0	0	0	0	$\frac{32}{3}$
	$\mathbf{\Gamma}_1{}^2\mathbf{\Theta}_1{}^2$	1	0	0	0	0	0	0	0	$\frac{32}{3}$
	$\mathbf{\Gamma}_1\mathbf{\Theta}_1{}^3$	1	0	0	0	0	0	0	0	$\frac{32}{3}$
	$\mathbf{\Theta}_1{}^4$	1	0	0	0	0	0	0	0	$\frac{32}{3}$

From these tables it is clear that none of the methods match the second, third and fourth order terms correctly. By solving the resulting system of equations we see that the following linear combination matches the second, third and fourth order terms

$$\begin{aligned}\mathbf{u}_A^{k+4} &= \eta_1\mathbf{u}_{A(1)}^{k+4} + \eta_2\mathbf{u}_{A(2)}^{k+4} + \eta_3\mathbf{u}_{A(3)}^{k+4} + \eta_4\mathbf{u}_{A(4)}^{k+4} \\ &\quad + \eta_5\mathbf{u}_{A(5)}^{k+4} + \eta_6\mathbf{u}_{A(6)}^{k+4} + \eta_7\mathbf{u}_{A(7)}^{k+4} + \eta_8\mathbf{u}_{A(8)}^{k+4}\end{aligned}$$

where

$$\eta_1 = \frac{32}{3}, \quad \eta_2 = \eta_3 = \frac{8}{3}, \quad \eta_4 = \eta_5 = \eta_7 = -\frac{16}{3}, \quad \eta_6 = 2 \quad \text{and} \quad \eta_8 = -1.$$

The fourth order accurate extrapolation algorithm follows

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}} \\ (\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{3}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+1} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+2} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{3}{2}} \\ (\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{5}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+2} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+3} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{5}{2}} \\ (\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{7}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+3} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_{A(1)}^{k+4} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{7}{2}}\end{aligned}$$

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}} \\ (\mathbf{I} - 3\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{5}{2}} &= (\mathbf{I} + 3\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+1} \\ (\mathbf{I} - 3\Delta t\mathbf{A}_A)\mathbf{u}_{A(2)}^{k+4} &= (\mathbf{I} + 3\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{5}{2}}\end{aligned}$$

$$\begin{aligned}(\mathbf{I} - 3\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{3}{2}} &= (\mathbf{I} + 3\Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\ (\mathbf{I} - 3\Delta t\mathbf{A}_A)\mathbf{u}_A^{k+3} &= (\mathbf{I} + 3\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{3}{2}} \\ (\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{7}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+3} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_{A(3)}^{k+4} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{7}{2}}\end{aligned}$$

$$\begin{aligned}(\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{1}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}} \\ (\mathbf{I} - \Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{3}{2}} &= (\mathbf{I} + \Delta t\mathbf{B}_C)\mathbf{u}_C^{k+1} \\ (\mathbf{I} - \Delta t\mathbf{A}_A)\mathbf{u}_A^{k+2} &= (\mathbf{I} + \Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{3}{2}} \\ (\mathbf{I} - 2\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+3} &= (\mathbf{I} + 2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+2} \\ (\mathbf{I} - 2\Delta t\mathbf{A}_A)\mathbf{u}_{A(4)}^{k+4} &= (\mathbf{I} + 2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+3}\end{aligned}$$

$$\begin{aligned}
(\mathbf{I}-2\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+1} &= (\mathbf{I}+2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k}\\
(\mathbf{I}-2\Delta t\mathbf{A}_A)\mathbf{u}_A^{k+2} &= (\mathbf{I}+2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+1}\\
(\mathbf{I}-\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{5}{2}} &= (\mathbf{I}+\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+2}\\
(\mathbf{I}-\Delta t\mathbf{A}_A)\mathbf{u}_A^{k+3} &= (\mathbf{I}+\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{5}{2}}\\
(\mathbf{I}-\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{7}{2}} &= (\mathbf{I}+\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+3}\\
(\mathbf{I}-\Delta t\mathbf{A}_A)\mathbf{u}_{A(5)}^{k+4} &= (\mathbf{I}+\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{7}{2}}
\end{aligned}$$

$$\begin{aligned}
(\mathbf{I}-2\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+1} &= (\mathbf{I}+2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k}\\
(\mathbf{I}-2\Delta t\mathbf{A}_A)\mathbf{u}_A^{k+2} &= (\mathbf{I}+2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+1}\\
(\mathbf{I}-2\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+3} &= (\mathbf{I}+2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+2}\\
(\mathbf{I}-2\Delta t\mathbf{A}_A)\mathbf{u}_{A(6)}^{k+4} &= (\mathbf{I}+2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+3}
\end{aligned}$$

$$\begin{aligned}
(\mathbf{I}-\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{1}{2}} &= (\mathbf{I}+\Delta t\mathbf{B}_C)\mathbf{u}_C^{k}\\
(\mathbf{I}-\Delta t\mathbf{A}_A)\mathbf{u}_A^{k+1} &= (\mathbf{I}+\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{1}{2}}\\
(\mathbf{I}-2\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+2} &= (\mathbf{I}+2\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+1}\\
(\mathbf{I}-2\Delta t\mathbf{A}_A)\mathbf{u}_A^{k+3} &= (\mathbf{I}+2\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+2}\\
(\mathbf{I}-\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+\frac{7}{2}} &= (\mathbf{I}+\Delta t\mathbf{B}_C)\mathbf{u}_C^{k+3}\\
(\mathbf{I}-\Delta t\mathbf{A}_A)\mathbf{u}_{A(7)}^{k+4} &= (\mathbf{I}+\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+\frac{7}{2}}
\end{aligned}$$

$$\begin{aligned}
(\mathbf{I}-4\Delta t\mathbf{C}_C)\mathbf{u}_C^{k+2} &= (\mathbf{I}+4\Delta t\mathbf{B}_C)\mathbf{u}_C^{k}\\
(\mathbf{I}-4\Delta t\mathbf{A}_A)\mathbf{u}_{A(8)}^{k+4} &= (\mathbf{I}+4\Delta t\mathbf{B}_A)\mathbf{u}_A^{k+2}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{u}_A^{k+4} = {} & \tfrac{32}{3}\mathbf{u}_{A(1)}^{k+4} + \tfrac{8}{3}\mathbf{u}_{A(2)}^{k+4} + \tfrac{8}{3}\mathbf{u}_{A(3)}^{k+4} - \tfrac{16}{3}\mathbf{u}_{A(4)}^{k+4}\\
& - \tfrac{16}{3}\mathbf{u}_{A(5)}^{k+4} + 2\mathbf{u}_{A(6)}^{k+4} - \tfrac{16}{3}\mathbf{u}_{A(7)}^{k+4} - \mathbf{u}_{A(8)}^{k+4}.
\end{aligned}$$

L_0-stability of the fourth order scheme

As done in section 6.3.1 we remove the cross-derivative term and write the fourth order extrapolation scheme as

$$\mathbf{u}_A^{k+3} = R_{\text{KT4}}(\Delta t\mathbf{A}_A, \Delta t\mathbf{C}_A)\mathbf{u}_A^{k}$$

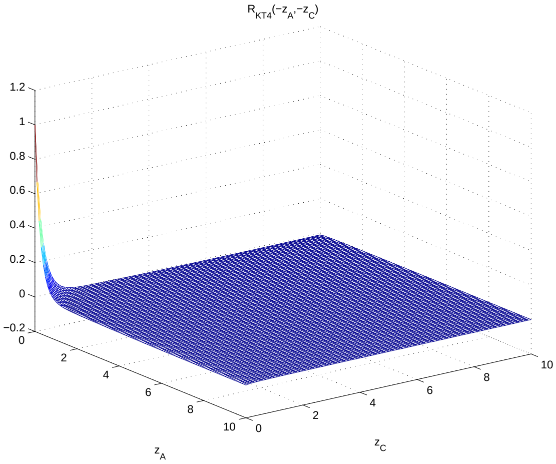


Figure 6.2: The symbol of the fourth order extrapolation scheme, $R_{\text{KT4}}(-z_A, -z_C)$.

where

$$\begin{aligned} R_{\text{KT3}}(\Delta t\mathbf{A}_A, \Delta t\mathbf{C}_A) = {} & \tfrac{32}{3}\left([\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}\right)^4 \\ & + \tfrac{8}{3}[\mathbf{I} - 3\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 3\Delta t\mathbf{C}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1} \\ & + \tfrac{8}{3}[\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}[\mathbf{I} - 3\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 3\Delta t\mathbf{C}_A]^{-1} \\ & - \tfrac{16}{3}\left([\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}\right)^2 [\mathbf{I} - 2\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 2\Delta t\mathbf{C}_A]^{-1} \\ & - \tfrac{16}{3}[\mathbf{I} - 2\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 2\Delta t\mathbf{C}_A]^{-1}\left([\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}\right)^2 \\ & + 2\left([\mathbf{I} - 2\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 2\Delta t\mathbf{C}_A]^{-1}\right)^2 \\ & - \tfrac{16}{3}[\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}\left([\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1}\right)^2 [\mathbf{I} - \Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - \Delta t\mathbf{C}_A]^{-1} \\ & - [\mathbf{I} - 4\Delta t\mathbf{A}_A]^{-1}[\mathbf{I} - 4\Delta t\mathbf{C}_A]^{-1}. \end{aligned}$$

From this we can deduce that the symbol of this fourth order scheme is given by

$$\begin{aligned} R_{\text{KT3}}(-z_A, -z_C) = {} & \frac{32}{3(1+z_A)^4(1+z_C)^4} + \frac{16}{3(1+3z_A)(1+3z_C)(1+z_A)(1+z_C)} \\ & - \frac{16}{(1+z_A)^2(1+z_C)^2(1+2z_A)(1+2z_C)} + \frac{2}{(1+2z_A)^2(1+2z_C)^2} \\ & - \frac{1}{(1+4z_A)(1+4z_C)}. \end{aligned}$$

From figure 6.2 we see that

$$\max_{z_A,\, z_C \,\geq\, 0} |R_{\text{KT4}}(-z_A, -z_C)| \leq 1$$

and

$$\lim_{z_A,\, z_C \,\to\, \infty} R_{\text{KT4}}(-z_A, -z_C) = 0.$$

Hence we can deduce that this scheme is indeed L_0-stable.

Comparison of the efficiency between the third and fourth order extrapolation schemes

Suppose we want to solve (6.1) by making use of extrapolation schemes. Let l_{KT3} denote the number of time steps used for the third order extrapolation scheme and l_{KT4} denote the number of time steps used for the fourth order extrapolation scheme. In this section we will determine if there exists a l_{KT3} such that the following holds:

- The third order scheme has a similar order of accuracy than that of the fourth order scheme
- The third order scheme is computationally less demanding than the fourth order scheme.

Clearly if there exists a range for l_{KT3} in which it satisfies the criteria above, the third order scheme has preference over the fourth order scheme in this range. The third and fourth order extrapolation schemes will have a similar order of convergence if

$$O(\Delta t_{\text{KT3}}^3) = O(\Delta t_{\text{KT4}}^4) \tag{6.29}$$

where $\Delta t_{\text{KT3}} = \frac{T}{l_{\text{KT3}}}$ and $\Delta t_{\text{KT4}} = \frac{T}{l_{\text{KT4}}}$. Hence the third and fourth order extrapolation schemes will approximately have the same order of convergence if

$$l_{\text{KT4}} = T^{1/4} l_{\text{KT3}}^{3/4}. \tag{6.30}$$

The third order extrapolation scheme solves 16 tri-diagonal matrices to advance three time increments, this is equivalent to solving $\frac{16}{3}$ tri-diagonal matrices to advance one time increment. The fourth order extrapolation scheme solves 40 tri-diagonal matrices to advance four time increments, this is equivalent to solving 10 tri-diagonal matrices to advance one time increment. From this we can deduce that the fourth order extrapolation scheme takes $\frac{15}{8}$ times as long as the third order scheme to advance one time step. Assume it takes γ seconds for the third order scheme to advance one time step. A certain level of accuracy can be obtained with the third order extrapolation scheme using l_{KT3} time steps in $l_{\text{KT3}}\gamma$ seconds. Using equation (6.30) and the arguments above we see that the same level of accuracy can be obtained with the fourth order scheme in $\frac{15}{8} T^{1/4} l_{\text{KT3}}^{3/4} \gamma$ seconds. From this we can deduce that whenever

$$l_{\text{KT3}}^{3/4}\left(l_{\text{KT3}}^{1/4} - \tfrac{15}{8} T^{1/4}\right) < 0$$

the third order scheme will be more efficient than the fourth order scheme. The inequality above can be simplified to obtain the following condition

$$l_{\text{KT3}} < \left(\tfrac{15}{8}\right)^4 T \approx 12.36T. \tag{6.31}$$

Note that this analysis is done simply to show that there might be cases where the third order extrapolation scheme is preferable to the fourth order scheme. Equation (6.31) might not be a very sharp bound since (6.29) is not a very effective method to compare the order of convergence of different schemes. A better method would be to equate the exact leading error terms of the third and fourth order extrapolation schemes.

Chapter 7

Extensions on the Finite Difference Method

In chapter 5 we discussed two different finite difference methods that can be implemented to approximate the solution of two dimensional parabolic partial differential equations. Although IMEX-methods can give second order accuracy they are computationally inefficient. Since the Yanenko method only solves tri-diagonal systems it can be implemented very efficiently, the downfall is that the Yanenko method is only first order accurate. In chapter 6 we showed how extrapolation methods can be used to obtain a computationally efficient scheme that is third or fourth order accurate in time. In this chapter we will discuss different methods of further improving finite difference schemes.

The first modification will be to replace the uniform grid with a non-uniform grid that is more dense at a point of interest. The non-uniform grid can be chosen such that local error is minimized, see Kluge [2002]. Non-uniform grids increase the accuracy in spatial direction whereas extrapolation increases the accuracy in the time direction.

7.1 Non-uniform grids

Up and till now we defined our partition of the x-axis and y-axis to be uniform. In the next section we will show how a uniform partition can be mapped onto a non-uniform partition by making use of a grid generating function.

7.1.1 Grid generating functions

This section is based on a section from Kluge [2002]. In section 3.1 we defined a uniform mesh by an ordered evenly spaced sequence of numbers

$$x_{\text{min}} = x_0 < x_1 < \ldots < x_m = x_{\text{max}}$$

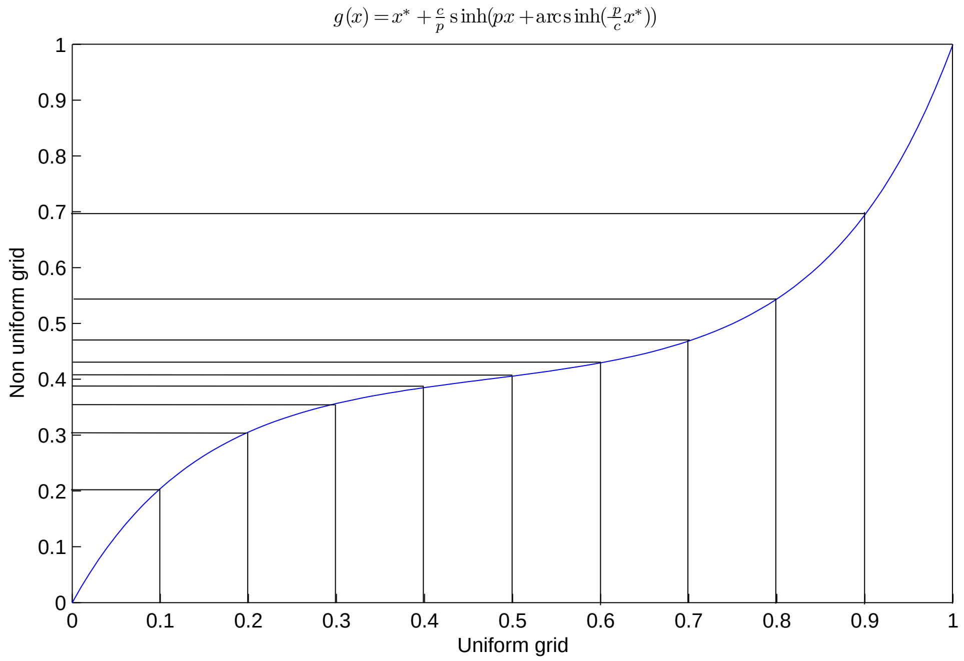


Figure 7.1: Grid generating function with $y^* = 0.4$, $c = 0.2$, $p = 7.1$ and $x_{\max} = 1$

where $x_i - x_{i-1} = h$ for $i = 1,\ 2,\ \ldots,\ m$. To obtain more accurate solutions at the point $y^* \in [x_{\min}, x_{\max}]$ the number of grid points can be increased, at the expense of computational time, or restructured such that the grid is more dense at y^*. A grid generating function, $g : [x_{\min}, x_{\max}] \rightarrow [x_{\min}, x_{\max}]$, is a continuously differentiable, bijective and strictly monotone increasing function that satisfies the following compatibility equations

$$\begin{aligned} g(x_{\min}) &= x_{\min} \\ g(x_{\max}) &= x_{\max}. \end{aligned}$$

The non-uniform grid can be generated via the following relationship

$$\widetilde{x}_i = g(x_i)$$

for $i = 0,\ 1,\ \ldots,\ m$. For the case when $x_{\min} = 0$ we can use the following grid generating function, see Kluge [2002]

$$g(x) = y^* + \frac{c}{p}\,\mathrm{sinh}\left(px + \mathrm{arcsinh}\left(-\frac{p}{c}y^*\right)\right) \tag{7.1}$$

where c is the density of the non-uniform grid at x^* and p is chosen such that $g(x_{\max}) = x_{\max}$[1]. Figure 7.1 shows how the density at $y^* = 0.4$ is increased, without increasing the number of grid points, using this grid generating function. For some problems it might be the case that $x_{\min} \neq 0$, for example a down-and-out barrier option. In such cases we can use a third degree polynomial as a grid generating function

$$g(x) = a_3(x - x^*)^3 + a_2(x - x^*)^2 + a_1(x - x^*) + a_0 \tag{7.2}$$

[1] It is easy to implement any one dimensional numerical solver to obtain the correct value of p.

where $x^* \in [x_{\text{min}}, x_{\text{max}}]$ is a constant with the property that $g(x^*) = y^*$, see Kluge [2002]. The five unknown parameters, a_0, a_1, a_2, a_3 and x^*, can be solved if we can obtain five equations. From the definition of x^* and the compatibility requirements at x_{min} and x_{max} it follows that

$$\begin{aligned} g(x^*) &= y^* \\ g(x_{\text{min}}) &= x_{\text{min}} \\ g(x_{\text{max}}) &= x_{\text{max}}. \end{aligned}$$

The gradient of $g(x)$ is directly related to the density of the mapped grid compared to the uniform grid. If $g'(x^*) > 1$ (resp. $g'(x^*) < 1$) then the density of the grid at y^* will be lower (resp. higher) than that of the uniform grid. An extra equation can be obtained from the fact that the solution of $g''(x) = 0$ will give the unique point where the grid is most dense[2]. Using the additional requirement that the grid points are $\frac{1}{c}$ times as dense at the concentration point as in the uniform case we can obtain the following equations

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

Substituting (7.2) into these equations results in

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

By rearranging we obtain the following non-linear system of equations

$$\begin{aligned} a_3(x_{\text{min}} - x^*)^3 + c(x_{\text{min}} - x^*) + y^* &= x_{\text{min}} \\ a_3(x_{\text{max}} - x^*)^3 + c(x_{\text{max}} - x^*) + y^* &= x_{\text{max}} \end{aligned}$$

which can be solved using a numerical method. Figure 7.2 shows how the density at $y^* = 3$ is increased for the case when $x_{\text{min}} = 1$ and $x_{\text{max}} = 5$, without increasing the number of grid points, using the second generating function.

7.1.2 Divided differences

Let h_x^- and h_x^+ be the non-uniform step sizes adjacent to a reference node x_i

$$\begin{aligned} h_x^+ &= x_{i+1} - x_i \\ h_x^- &= x_i - x_{i-1} \end{aligned}$$

[2] This is based on the observation that the solution of $f'(x) = 0$ will be a local extreme value of $f(x)$.

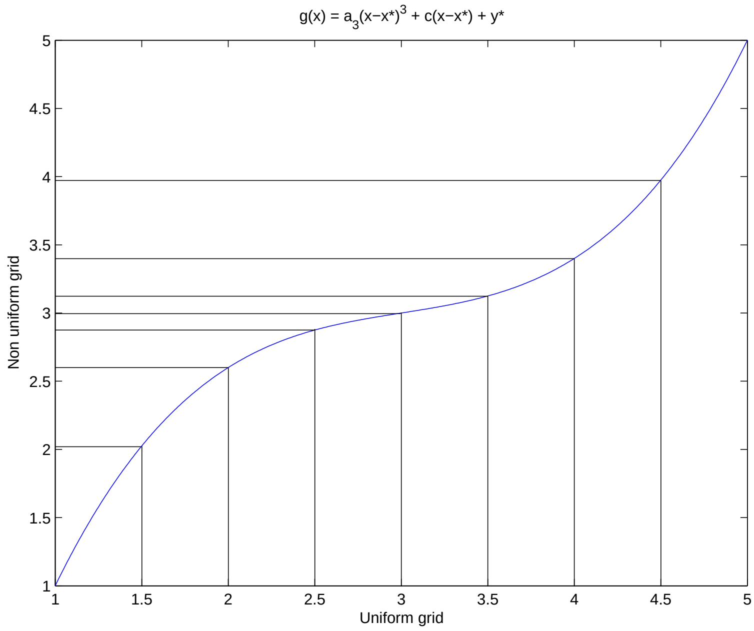


Figure 7.2: Grid generating function with $y^* = 3$, $x^* = 3$, $c = 0.2$, $a_3 = 0.2$, $x_{\text{min}} = 1$ and $x_{\text{max}} = 5$

and h_y^- and h_y^+ be the non-uniform step sizes adjacent to a reference node y_j

$$h_y^+ = y_{j+1} - y_j$$
$$h_y^- = y_j - y_{j-1}.$$

The finite difference approximations at the respective grid points are denoted by

$$u_{ij}^k \approx u(x_i, y_j, \tau_k).$$

where $\tau_k = k\Delta t$ for $k = 0,\ 1,\ \ldots,\ l$. The spacial nodes x_i and y_j are generated by a grid generating function for $i = 0,\ 1,\ \ldots,\ m$ and $j = 0,\ 1,\ \ldots,\ n$. The relevant forward and backward finite difference approximations for a reference grid point u_{ij}^k are given by

$$\Delta_x^- u_{i,j}^k = \frac{u_{i,j}^k - u_{i-1,j}^k}{h_x^-} \tag{7.3}$$

$$\Delta_y^- u_{i,j}^k = \frac{u_{i,j}^k - u_{i,j-1}^k}{h_y^-} \tag{7.4}$$

$$\Delta_x^+ u_{i,j}^k = \frac{u_{i+1,j}^k - u_{i,j}^k}{h_x^+} \tag{7.5}$$

$$\Delta_y^+ u_{i,j}^k = \frac{u_{i,j+1}^k - u_{i,j}^k}{h_y^+}. \tag{7.6}$$

We will need the following Taylor approximations to derive the central difference approximations for the case when the mesh is non-uniform.

$$\begin{aligned} u(x_{i+1}, y_{j+1}, \tau_k) = {} & u + h_x^+ \frac{\partial u}{\partial x} + h_y^+ \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^{+2} \frac{\partial^2 u}{\partial x^2} + h_x^+ h_y^+ \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^{+2} \frac{\partial^2 u}{\partial y^2} \\ & + \frac{1}{3!} h_x^{+3} \frac{\partial^3 u}{\partial x^3} + \frac{3}{3!} h_x^{+2} h_y^+ \frac{\partial^3 u}{\partial x^2 \partial y} + \frac{3}{3!} h_x^+ h_y^+ 2 \frac{\partial^3 u}{\partial x \partial y^2} + \frac{1}{3!} h_y^{+3} \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^{+4}, h_x^{+3} h_y^+, h_x^{+2} h_y^{+2}, h_x^+ h_y^{+3}, h_y^{+4}) \end{aligned} \tag{7.7}$$

$$\begin{aligned} u(x_{i-1}, y_{j+1}, \tau_k) = {} & u - h_x^- \frac{\partial u}{\partial x} + h_y^+ \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^{-2} \frac{\partial^2 u}{\partial x^2} - h_x^- h_y^+ \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^{+2} \frac{\partial^2 u}{\partial y^2} \\ & - \frac{1}{3!} h_x^{-3} \frac{\partial^3 u}{\partial x^3} + \frac{3}{3!} h_x^{-2} h_y^+ \frac{\partial^3 u}{\partial x^2 \partial y} - \frac{3}{3!} h_x^- h_y^{+2} \frac{\partial^3 u}{\partial x \partial y^2} + \frac{1}{3!} h_y^{+3} \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^{-4}, h_x^{-3} h_y^+, h_x^{-2} h_y^{+2}, h_x^- h_y^{+3}, h_y^{+4}) \end{aligned} \tag{7.8}$$

$$\begin{aligned} u(x_{i+1}, y_{j-1}, \tau_k) = {} & u + h_x^+ \frac{\partial u}{\partial x} - h_y^- \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^{+2} \frac{\partial^2 u}{\partial x^2} - h_x^+ h_y^- \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^{-2} \frac{\partial^2 u}{\partial y^2} \\ & + \frac{1}{3!} h_x^{+3} \frac{\partial^3 u}{\partial x^3} - \frac{3}{3!} h_x^{+2} h_y^- \frac{\partial^3 u}{\partial x^2 \partial y} + \frac{3}{3!} h_x^+ h_y^{-2} \frac{\partial^3 u}{\partial x \partial y^2} - \frac{1}{3!} h_y^{-3} \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^{+4}, h_x^{+3} h_y^-, h_x^{+2} h_y^{-2}, h_x^+ h_y^{-3}, h_y^{-4}) \end{aligned} \tag{7.9}$$

$$\begin{aligned} u(x_{i-1}, y_{j-1}, \tau_k) = {} & u - h_x^- \frac{\partial u}{\partial x} - h_y^- \frac{\partial u}{\partial y} + \tfrac{1}{2} h_x^{-2} \frac{\partial^2 u}{\partial x^2} + h_y^- h_y^- \frac{\partial^2 u}{\partial x \partial y} + \tfrac{1}{2} h_y^{-2} \frac{\partial^2 u}{\partial y^2} \\ & - \frac{1}{3!} h_x^{-3} \frac{\partial^3 u}{\partial x^3} - \frac{3}{3!} h_x^{-2} h_y^- \frac{\partial^3 u}{\partial x^2 \partial y} - \frac{3}{3!} h_x^- h_y^{-2} \frac{\partial^3 u}{\partial x \partial y^2} - \frac{1}{3!} h_y^{-3} \frac{\partial^3 u}{\partial y^3} \\ & + O(h_x^{-4}, h_x^{-3} h_y^-, h_x^{-2} h_y^{-2}, h_x^- h_y^{-3}, h_y^{-4}) \end{aligned} \tag{7.10}$$

$$u(x_{i-1}, y_j, \tau_k) = u - h_x^- \frac{\partial u}{\partial x} + \tfrac{1}{2} h_x^{-2} \frac{\partial^2 u}{\partial x^2} - \frac{1}{3!} h_x^{-3} \frac{\partial^3 u}{\partial x^3} + O(h_x^{-4}) \tag{7.11}$$

$$u(x_{i+1}, y_j, \tau_k) = u + h_x^+ \frac{\partial u}{\partial x} + \tfrac{1}{2} h_x^{+2} \frac{\partial^2 u}{\partial x^2} + \frac{1}{3!} h_x^{+3} \frac{\partial^3 u}{\partial x^3} + O(h_x^{+4}) \tag{7.12}$$

$$u(x_i, y_{j-1}, \tau_k) = u - h_y^- \frac{\partial u}{\partial y} + \tfrac{1}{2} h_y^{-2} \frac{\partial^2 u}{\partial y^2} - \frac{1}{3!} h_y^{-3} \frac{\partial^3 u}{\partial y^3} + O(h_y^{-4}) \tag{7.13}$$

$$u(x_i, y_{j+1}, \tau_k) = u + h_y^+ \frac{\partial u}{\partial y} + \tfrac{1}{2} h_y^{+2} \frac{\partial^2 u}{\partial y^2} + \frac{1}{3!} h_y^{+3} \frac{\partial^3 u}{\partial y^3} + O(h_y^{+4}). \tag{7.14}$$

We define the following parameters to shorten the formulas in the derivation

$$\begin{aligned} \zeta_{-1} &:= h_x^- (h_x^- + h_x^+) \\ \zeta_0 &:= h_x^+ h_x^- \\ \zeta_1 &:= h_x^+ (h_x^- + h_x^+) \\ \varphi_{-1} &:= h_y^- (h_y^- + h_y^+) \\ \varphi_0 &:= h_y^+ h_y^- \\ \varphi_1 &:= h_y^+ (h_y^- + h_y^+). \end{aligned}$$

By subtracting h_x^{-2} times equation (7.12) from h_x^{+2} times equation (7.11) and rearranging we obtain

$$\frac{\partial u}{\partial x} = -\frac{h_x^+}{\zeta_{-1}} u(x_{i-1}, y_j, \tau_k) + \frac{(h_x^+ - h_x^-)}{\zeta_0} u(x_i, y_j, \tau_k) + \frac{h_x^-}{\zeta_1} u(x_{i+1}, y_j, \tau_k) + O(h_x^+ h_x^-).$$

Similarly by subtracting h_y^{-2} times equation (7.14) from h_y^{+2} times equation (7.13) we obtain

$$\frac{\partial u}{\partial y} = -\frac{h_y^+}{\varphi_{-1}} u(x_i, y_{j-1}, \tau_k) + \frac{(h_y^+ - h_y^-)}{\varphi_0} u(x_i, y_j, \tau_k) + \frac{h_y^-}{\varphi_1} u(x_i, y_{j+1}, \tau_k) + O(h_y^+ h_y^-).$$

Subtracting h_x^- times equation (7.12) from h_x^+ times equation (7.11) we obtain

$$\frac{\partial^2 u}{\partial x^2} = \frac{2}{\zeta_{-1}} u(x_{i-1}, y_j, \tau_k) - \frac{2}{\zeta_0} u(x_i, y_j, \tau_k) + \frac{2}{\zeta_1} u(x_{i+1}, y_j, \tau_k) + O\left(\frac{h_x^{-3} + h_x^{+3}}{h_x^- + h_x^+}\right).$$

Similarly by subtracting h_y^- times equation (7.14) from h_y^+ times equation (7.13) we obtain

$$\frac{\partial^2 u}{\partial y^2} = \frac{2}{\varphi_{-1}} u(x_i, y_{j-1}, \tau_k) - \frac{2}{\varphi_0} u(x_i, y_j, \tau_k) + \frac{2}{\varphi_1} u(x_i, y_{j+1}, \tau_k) + O\left(\frac{h_y^{-3} + h_y^{+3}}{h_y^- + h_y^+}\right).$$

To obtain the approximation of the cross derivative term we consider the following linear combination of (7.7) to (7.10) that eliminates the $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ terms

$$\begin{aligned}
&h_x^{+2} h_y^{+2} u(x_{i-1}, y_{j-1}, \tau_k) - h_x^{-2} h_y^{+2} u(x_{i+1}, y_{j-1}, \tau_k) - h_x^{+2} h_y^{-2} u(x_{i-1}, y_{j+1}, \tau_k) + h_x^{-2} h_y^{-2} u(x_{i+1}, y_{j+1}, \tau_k) \\
&= (h_x^{+2} - h_x^{-2})(h_y^{+2} - h_y^{-2}) u \\
&\quad - h_x^+ h_x^- (h_y^{+2} - h_y^{-2})(h_x^- + h_x^+) \frac{\partial u}{\partial x} - h_y^+ h_y^- (h_x^{+2} - h_x^{-2})(h_y^- + h_y^+) \frac{\partial u}{\partial y} \\
&\quad + h_x^+ h_x^- h_y^+ h_y^- (h_x^- + h_x^+)(h_y^- + h_y^+) \frac{\partial^2 u}{\partial x \partial y} \\
&\quad - \frac{1}{3!} h_x^{+2} h_x^{-2} (h_y^{+2} - h_y^{-2})(h_x^- + h_x^+) \frac{\partial^3 u}{\partial x^3} - \frac{1}{3!} h_y^{+2} h_y^{-2} (h_x^{+2} - h_x^{-2})(h_y^- + h_y^+) \frac{\partial^3 u}{\partial y^3} \\
&\quad + \text{Higher order terms} \qquad (7.15)
\end{aligned}$$

The $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ terms can also be eliminated from (7.11) to (7.14) with the following linear combinations

$$\begin{aligned}
&-h_x^{+2} u(x_{i-1}, y_j, \tau_k) + h_x^{-2} u(x_{i+1}, y_j, \tau_k) \\
&\qquad = -(h_x^{+2} - h_x^{-2}) u + h_x^+ h_x^- (h_x^+ + h_x^-) \frac{\partial u}{\partial x} + \frac{1}{3!} h_x^{+2} h_x^{-2} (h_x^- + h_x^+) \frac{\partial^3 u}{\partial x^3} \\
&\qquad\quad + \text{Higher order terms}
\end{aligned}$$

$$\begin{aligned}
&-h_y^{+2} u(x_i, y_{j-1}, \tau_k) + h_y^{-2} u(x_i, y_{j+1}, \tau_k) \\
&\qquad = -(h_y^{+2} - h_y^{-2}) u + h_y^+ h_y^- (h_y^+ + h_y^-) \frac{\partial u}{\partial y} + \frac{1}{3!} h_y^{+2} h_y^{-2} (h_y^- + h_y^+) \frac{\partial^3 u}{\partial y^3} \\
&\qquad\quad + \text{Higher order terms.}
\end{aligned}$$

We are only left to eliminate the first order terms from (7.15), this can be achieved with the following linear combination

$$\begin{aligned}
&h_x^{+2} h_y^{+2} u(x_{i-1}, y_{j-1}, \tau_k) - h_x^{-2} h_y^{+2} u(x_{i+1}, y_{j-1}, \tau_k) - h_x^{+2} h_y^{-2} u(x_{i-1}, y_{j+1}, \tau_k) + h_x^{-2} h_y^{-2} u(x_{i+1}, y_{j+1}, \tau_k) \\
&\quad - h_x^{+2} (h_y^{+2} - h_y^{-2}) u(x_{i-1}, y_j, \tau_k) + h_x^{-2} (h_y^{+2} - h_y^{-2}) u(x_{i+1}, y_j, \tau_k) \\
&\quad - h_y^{+2} (h_x^{+2} - h_x^{-2}) u(x_i, y_{j-1}, \tau_k) + h_y^{-2} (h_x^{+2} - h_x^{-2}) u(x_i, y_{j+1}, \tau_k) \\
&= -(h_x^{+2} - h_x^{-2})(h_y^{+2} - h_y^{-2}) u + h_x^+ h_x^- h_y^+ h_y^- (h_x^- + h_x^+)(h_y^- + h_y^+) \frac{\partial^2 u}{\partial x \partial y} + \text{Higher order terms.}
\end{aligned}$$

Rearranging results in

$$
\begin{aligned}
\frac{\partial^2 u}{\partial x \partial y} = {} & \frac{h_x^+ h_y^+}{\zeta_{-1}\varphi_{-1}} u(x_{i-1}, y_{j-1}, \tau_k) - \frac{(h_x^+ - h_x^-)h_y^+}{\zeta_0 \varphi_{-1}} u(x_i, y_{j-1}, \tau_k) - \frac{h_x^- h_y^+}{\zeta_1 \varphi_{-1}} u(x_{i+1}, y_{j-1}, \tau_k) \\
& - \frac{h_x^+(h_y^+ - h_y^-)}{\zeta_{-1}\varphi_0} u(x_{i-1}, y_j, \tau_k) + \frac{(h_x^+ - h_x^-)(h_y^+ - h_y^-)}{\zeta_0 \varphi_0} u(x_i, y_j, \tau_k) + \frac{h_x^-(h_y^+ - h_y^-)}{\zeta_1 \varphi_0} u(x_{i+1}, y_j, \tau_k) \\
& - \frac{h_x^+ h_y^-}{\zeta_{-1}\varphi_1} u(x_{i-1}, y_{j+1}, \tau_k) + \frac{(h_x^+ - h_x^-)h_y^-}{\zeta_0 \varphi_1} u(x_i, y_{j+1}, \tau_k) + \frac{h_x^- h_y^-}{\zeta_1 \varphi_1} u(x_{i+1}, y_{j+1}, \tau_k) \\
& + \text{Higher order terms.}
\end{aligned}
$$

The same result is obtained in Kluge [2002] by making use of matrix equations. The difference approximations are then given by

$$
\Delta_x u_{i,j}^k = -\frac{h_x^+}{\zeta_{-1}} u_{i-1,j}^k + \frac{(h_x^+ - h_x^-)}{\zeta_0} u_{i,j}^k + \frac{h_x^-}{\zeta_1} u_{i+1,j}^k \tag{7.16}
$$

$$
\Delta_y u_{i,j}^k = -\frac{h_y^+}{\varphi_{-1}} u_{i,j-1}^k + \frac{(h_y^+ - h_y^-)}{\varphi_0} u_{i,j}^k + \frac{h_y^-}{\varphi_1} u_{i,j+1}^k \tag{7.17}
$$

$$
\Delta_x^2 u_{i,j}^k = \frac{2}{\zeta_{-1}} u_{i-1,j}^k - \frac{2}{\zeta_0} u_{i,j}^k + \frac{2}{\zeta_1} u_{i+1,j}^k \tag{7.18}
$$

$$
\Delta_y^2 u_{i,j}^k = \frac{2}{\varphi_{-1}} u_{i,j-1}^k - \frac{2}{\varphi_0} u_{i,j}^k + \frac{2}{\varphi_1} u_{i,j+1}^k \tag{7.19}
$$

$$
\begin{aligned}
\Delta_{xy} u_{i,j}^k = {} & \frac{h_x^+ h_y^+}{\zeta_{-1}\varphi_{-1}} u_{i-1,j-1}^k - \frac{(h_x^+ - h_x^-)h_y^+}{\zeta_0 \varphi_{-1}} u_{i,j-1}^k - \frac{h_x^- h_y^+}{\zeta_1 \varphi_{-1}} u_{i+1,j-1}^k \\
& - \frac{h_x^+(h_y^+ - h_y^-)}{\zeta_{-1}\varphi_0} u_{i-1,j}^k + \frac{(h_x^+ - h_x^-)(h_y^+ - h_y^-)}{\zeta_0 \varphi_0} u_{i,j}^k + \frac{h_x^-(h_y^+ - h_y^-)}{\zeta_1 \varphi_0} u_{i+1,j}^k \\
& - \frac{h_x^+ h_y^-}{\zeta_{-1}\varphi_1} u_{i-1,j+1}^k + \frac{(h_x^+ - h_x^-)h_y^-}{\zeta_0 \varphi_1} u_{i,j+1}^k + \frac{h_x^- h_y^-}{\zeta_1 \varphi_1} u_{i+1,j+1}^k.
\end{aligned} \tag{7.20}
$$

7.1.3 Matrix formulation

It is easy to see that we can apply the non-uniform grid to the IMEX and Yanenko schemes defined in chapter 5 and the extrapolated Yanenko schemes in chapter 6 by simply redefining the difference operators in equations (5.5) to (5.12) and (5.23) by those given in (7.3) to (7.6) and (7.16) to (7.20). As in chapter 5 the structure of the matrices involved can be made clear if we give a simple example. We give the matrices, in bandwidth form, for the case when $m = 3$ and $n = 2$ (see figure 5.1). The 1-1-1 bandwidth form of $\mathbf{A}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}$ is given by

$$|\mathbf{A}_{\text{diff,A}}+\mathbf{A}_{\text{cov,A}}||\mathbf{u}^k_A| = \left|\begin{array}{c|c|c||c}
\times & 0 & & u^k_{0,\,0} \\
 & -\frac{d_{1,\,0}}{h_x^+} & \frac{d_{1,\,0}}{h_x^+} & u^k_{1,\,0} \\
 & -\frac{d_{2,\,0}}{h_x^+} & \frac{d_{2,\,0}}{h_x^+} & u^k_{2,\,0} \\
 & 0 & & u^k_{3,\,0} \\
\hline
 & 0 & & u^k_{0,\,1} \\
\frac{2}{\zeta_{-1}}\left(a_{1,\,1}-\frac{h_x^+}{2}d_{1,\,1}\right) & -\frac{2}{\zeta_0}\left(a_{1,\,1}-\frac{(h_x^+-h_x^-)}{2}d_{1,\,1}\right) & \frac{2}{\zeta_1}\left(a_{1,\,1}+\frac{h_x^-}{2}d_{1,\,1}\right) & u^k_{1,\,1} \\
\frac{2}{\zeta_{-1}}\left(a_{2,\,1}-\frac{h_x^+}{2}d_{2,\,1}\right) & -\frac{2}{\zeta_0}\left(a_{2,\,1}-\frac{(h_x^+-h_x^-)}{2}d_{2,\,1}\right) & \frac{2}{\zeta_1}\left(a_{2,\,1}+\frac{h_x^-}{2}d_{2,\,1}\right) & u^k_{2,\,1} \\
 & 0 & & u^k_{3,\,1} \\
\hline
 & 0 & & u^k_{0,\,2} \\
\frac{2}{\zeta_{-1}}\left(a_{1,\,2}-\frac{h_x^+}{2}d_{1,\,2}\right) & -\frac{2}{\zeta_0}\left(a_{1,\,2}-\frac{(h_x^+-h_x^-)}{2}d_{1,\,2}\right) & \frac{2}{\zeta_1}\left(a_{1,\,2}+\frac{h_x^-}{2}d_{1,\,2}\right) & u^k_{1,\,2} \\
\frac{2}{\zeta_{-1}}\left(a_{2,\,2}-\frac{h_x^+}{2}d_{2,\,2}\right) & -\frac{2}{\zeta_0}\left(a_{2,\,2}-\frac{(h_x^+-h_x^-)}{2}d_{2,\,2}\right) & \frac{2}{\zeta_1}\left(a_{2,\,2}+\frac{h_x^-}{2}d_{2,\,2}\right) & u^k_{2,\,2} \\
 & 0 & \times & u^k_{3,\,2}
\end{array}\right|$$

$\mathbf{C}_{\text{diff,A}}+\mathbf{C}_{\text{cov,A}}$ can be written in $(m+1)$-1-$(m+1)$ bandwidth form as

$$|\mathbf{C}_{\text{diff,A}}+\mathbf{C}_{\text{cov,A}}||\mathbf{u}^k_A| =$$

$$\left|\begin{array}{c|c|c||c}
\times \quad \overbrace{\cdots}^{m\text{ columns}} & 0 & & u^k_{0,\,0} \\
\times & -\frac{e_{1,\,0}}{h_y^+} & \frac{e_{1,\,0}}{h_y^+} & u^k_{1,\,0} \\
\times & -\frac{e_{2,\,0}}{h_y^+} & \frac{e_{2,\,0}}{h_y^+} & u^k_{2,\,0} \\
\times & 0 & & u^k_{3,\,0} \\
\hline
 & 0 & & u^k_{0,\,1} \\
\frac{2}{\varphi_{-1}}\left(c_{1,\,1}-\frac{h_y^+}{2}e_{1,\,1}\right) & -\frac{2}{\varphi_0}\left(c_{1,\,1}-\frac{(h_y^+-h_y^-)}{2}e_{1,\,1}\right) & \frac{2}{\varphi_1}\left(c_{1,\,1}+\frac{h_y^-}{2}e_{1,\,1}\right) & u^k_{1,\,1} \\
\frac{2}{\varphi_{-1}}\left(c_{2,\,1}-\frac{h_y^+}{2}e_{2,\,1}\right) & -\frac{2}{\varphi_0}\left(c_{2,\,1}-\frac{(h_y^+-h_y^-)}{2}e_{2,\,1}\right) & \frac{2}{\varphi_1}\left(c_{2,\,1}+\frac{h_y^-}{2}e_{2,\,1}\right) & u^k_{2,\,1} \\
 & 0 & & u^k_{3,\,1} \\
\hline
 & 0 & \times & u^k_{0,\,2} \\
 & 0 & \times & u^k_{1,\,2} \\
 & 0 & \times & u^k_{2,\,2} \\
 & 0 & \underbrace{\cdots}_{m\text{ columns}} \quad \times & u^k_{3,\,2}
\end{array}\right|$$

and $\mathbf{B}_A$ can be written in $(m+2)$-1-$(m+2)$ bandwidth form as

$$
|\mathbf{B}_A||\mathbf{u}_A^k| =
\left|\begin{array}{ccccc|c|ccccc||c}
\times & \times & \times & \overbrace{\cdots}^{m-2\,\text{col}} & \times & 0 & & & & & & u_{0,\,0}^k \\
\times & \times & \times & & & 0 & & & & & & u_{1,\,0}^k \\
\times & \times & \times & & & 0 & & & & & & u_{2,\,0}^k \\
\times & \times & & & & 0 & & & & & & u_{3,\,0}^k \\
\hline
\times & & & & & 0 & & & & & & u_{0,\,1}^k \\
\frac{h_x^+ h_y^+ b_{1,\,1}}{\zeta_{-1}\varphi_{-1}} & -\frac{(h_x^+ - h_x^-) h_y^+ b_{1,\,1}}{\zeta_0\varphi_{-1}} & -\frac{h_x^- h_y^+ b_{1,\,1}}{\zeta_1\varphi_{-1}} & & -\frac{h_x^+ (h_y^+ - h_y^-) b_{1,\,1}}{\zeta_{-1}\varphi_0} & \frac{(h_x^+ - h_x^-)(h_y^+ - h_y^-) b_{1,\,1}}{\zeta_0\varphi_0} & \frac{h_x^- (h_y^+ - h_y^-) b_{1,\,1}}{\zeta_1\varphi_0} & & -\frac{h_x^+ h_y^- b_{1,\,1}}{\zeta_{-1}\varphi_1} & \frac{(h_x^+ - h_x^-) h_y^- b_{1,\,1}}{\zeta_0\varphi_1} & \frac{h_x^- h_y^- b_{1,\,1}}{\zeta_1\varphi_1} & u_{1,\,1}^k \\
\frac{h_x^+ h_y^+ b_{2,\,1}}{\zeta_{-1}\varphi_{-1}} & -\frac{(h_x^+ - h_x^-) h_y^+ b_{2,\,1}}{\zeta_0\varphi_{-1}} & -\frac{h_x^- h_y^+ b_{2,\,1}}{\zeta_1\varphi_{-1}} & & -\frac{h_x^+ (h_y^+ - h_y^-) b_{2,\,1}}{\zeta_{-1}\varphi_0} & \frac{(h_x^+ - h_x^-)(h_y^+ - h_y^-) b_{2,\,1}}{\zeta_0\varphi_0} & \frac{h_x^- (h_y^+ - h_y^-) b_{2,\,1}}{\zeta_1\varphi_0} & & -\frac{h_x^+ h_y^- b_{2,\,1}}{\zeta_{-1}\varphi_1} & \frac{(h_x^+ - h_x^-) h_y^- b_{2,\,1}}{\zeta_0\varphi_1} & \frac{h_x^- h_y^- b_{2,\,1}}{\zeta_1\varphi_1} & u_{2,\,1}^k \\
 & & & & & 0 & & & & & \times & u_{3,\,1}^k \\
\hline
 & & & & & 0 & & & & \times & \times & u_{0,\,2}^k \\
 & & & & & 0 & & & \times & \times & \times & u_{1,\,2}^k \\
 & & & & & 0 & & & \times & \times & \times & u_{2,\,2}^k \\
 & & & & & 0 & \times & \underbrace{\cdots}_{m-2\,\text{col}} & \times & \times & \times & u_{3,\,2}^k
\end{array}\right|
$$

The *modified* identity matrices $\mathbf{I}_{\text{im},A}$ and $\mathbf{I}_{\text{ex},A}$ structure is the same as in the case for the uniform grid. The boundary vector is given by

$$\mathbf{b}_A = (c_l, 0, 0, c_r h_x^-; c_l, 0, 0, c_r h_x^-; c_l, 0, 0, c_r h_x^-)^T.$$

For the second ordering of the elements in $\mathbf{u}$ it is easy to see that $\mathbf{A}_{\text{diff,C}} + \mathbf{A}_{\text{cov,C}}$ can be written in $(n+1)$-1-$(n+1)$ bandwidth form as

$$|\mathbf{A}_{\text{diff,C}} + \mathbf{A}_{\text{cov,C}}||\mathbf{u}_C^k| =$$

$$\begin{array}{|c|c|c||c|}
\times \quad \overbrace{\cdots}^{n\text{ col}} & 0 & & u_{0,\,0}^k \\
\times & 0 & & u_{0,\,1}^k \\
\times & 0 & & u_{0,\,2}^k \\
\hline
 & -\frac{d_{1,\,0}}{h_x^+} & \frac{d_{1,\,0}}{h_x^+} & u_{1,\,0}^k \\
\frac{2}{\zeta_{-1}}\left(a_{1,\,1} - \frac{h_x^+}{2} d_{1,\,1}\right) & -\frac{2}{\zeta_0}\left(a_{1,\,1} - \frac{(h_x^+ - h_x^-)}{2} d_{1,\,1}\right) & \frac{2}{\zeta_1}\left(a_{1,\,1} + \frac{h_x^-}{2} d_{1,\,1}\right) & u_{1,\,1}^k \\
\frac{2}{\zeta_{-1}}\left(a_{1,\,2} - \frac{h_x^+}{2} d_{1,\,2}\right) & -\frac{2}{\zeta_0}\left(a_{1,\,2} - \frac{(h_x^+ - h_x^-)}{2} d_{1,\,2}\right) & \frac{2}{\zeta_1}\left(a_{1,\,2} + \frac{h_x^-}{2} d_{1,\,2}\right) & u_{1,\,2}^k \\
\hline
 & -\frac{d_{2,\,0}}{h_x^+} & \frac{d_{2,\,0}}{h_x^+} & u_{2,\,0}^k \\
\frac{2}{\zeta_{-1}}\left(a_{2,\,1} - \frac{h_x^+}{2} d_{2,\,1}\right) & -\frac{2}{\zeta_0}\left(a_{2,\,1} - \frac{(h_x^+ - h_x^-)}{2} d_{2,\,1}\right) & \frac{2}{\zeta_1}\left(a_{2,\,1} + \frac{h_x^-}{2} d_{2,\,1}\right) & u_{2,\,1}^k \\
\frac{2}{\zeta_{-1}}\left(a_{2,\,2} - \frac{h_x^+}{2} d_{2,\,2}\right) & -\frac{2}{\zeta_0}\left(a_{2,\,2} - \frac{(h_x^+ - h_x^-)}{2} d_{2,\,1}\right) & \frac{2}{\zeta_1}\left(a_{2,\,2} + \frac{h_x^-}{2} d_{2,\,2}\right) & u_{2,\,2}^k \\
\hline
 & 0 & \times & u_{3,\,0}^k \\
 & 0 & \times & u_{3,\,1}^k \\
 & 0 & \underbrace{\cdots}_{n\text{ col}} \quad \times & u_{3,\,2}^k
\end{array}$$

The tri-diagonal matrix $\mathbf{C}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}$ can be written in 1-1-1 bandwidth form as

$$|\mathbf{C}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}||\mathbf{u}_C^k| = \begin{array}{|c|c|c||c|}
\times & 0 & & u_{0,\,0}^k \\
 & 0 & & u_{0,\,1}^k \\
 & 0 & & u_{0,\,2}^k \\
\hline
 & -\frac{e_{1,\,0}}{h_y^+} & \frac{e_{1,\,0}}{h_y^+} & u_{1,\,0}^k \\
\frac{2}{\varphi_{-1}}\left(c_{1,\,1} - \frac{h_y^+}{2} e_{1,\,1}\right) & -\frac{2}{\varphi_0}\left(c_{1,\,1} - \frac{(h_y^+ - h_y^-)}{2} e_{1,\,1}\right) & \frac{2}{\varphi_1}\left(c_{1,\,1} + \frac{h_y^-}{2} e_{1,\,1}\right) & u_{1,\,1}^k \\
 & 0 & & u_{1,\,2}^k \\
\hline
 & -\frac{e_{2,\,0}}{h_y^+} & \frac{e_{2,\,0}}{h_y^+} & u_{2,\,0}^k \\
\frac{2}{\varphi_{-1}}\left(c_{2,\,1} - \frac{h_y^+}{2} e_{2,\,1}\right) & -\frac{2}{\varphi_0}\left(c_{2,\,1} - \frac{(h_y^+ - h_y^-)}{2} e_{2,\,1}\right) & \frac{2}{\varphi_1}\left(c_{2,\,1} + \frac{h_y^-}{2} e_{2,\,1}\right) & u_{2,\,1}^k \\
 & 0 & & u_{2,\,2}^k \\
\hline
 & 0 & & u_{3,\,0}^k \\
 & 0 & & u_{3,\,1}^k \\
 & 0 & \times & u_{3,\,2}^k
\end{array}$$

and $\mathbf{B}_C$ can be written in $(n+2)$-1-$(n+2)$ bandwidth form as

$$
|\mathbf{B}_C||\mathbf{u}^k_C| =
\left|
\begin{array}{ccccc|c|ccccc||c}
\times & \times & \times & \overbrace{\cdots}^{n-2\,\text{col}} & \times & 0 & & & & & & u^k_{0,\,0} \\
\times & \times & \times & & & 0 & & & & & & u^k_{0,\,1} \\
\times & \times & & & & 0 & & & & & & u^k_{0,\,2} \\
\hline
\times & & & & & 0 & & & & & & u^k_{1,\,0} \\
\frac{h^+_x h^+_y b_{1,\,1}}{\zeta_{-1}\varphi_{-1}} & -\frac{h^+_x (h^+_y - h^-_y)}{\zeta_{-1}\varphi_0} & -\frac{h^+_x h^-_y b_{1,\,1}}{\zeta_{-1}\varphi_1} & & -\frac{(h^+_x - h^-_x) h^+_y b_{1,\,1}}{\zeta_0\varphi_{-1}} & \frac{(h^+_x - h^-_x)(h^+_y - h^-_y)}{\zeta_0\varphi_0} & \frac{(h^+_x - h^-_x) h^-_y b_{1,\,1}}{\zeta_0\varphi_1} & & -\frac{h^-_x h^+_y b_{1,\,1}}{\zeta_1\varphi_{-1}} & \frac{h^-_x (h^+_y - h^-_y)}{\zeta_1\varphi_0} & \frac{h^-_x h^-_y b_{1,\,1}}{\zeta_1\varphi_1} & u^k_{1,\,1} \\
 & & & & & 0 & & & & & & u^k_{1,\,2} \\
\hline
 & & & & & 0 & & & & & & u^k_{2,\,0} \\
\frac{h^+_x h^+_y b_{2,\,1}}{\zeta_{-1}\varphi_{-1}} & -\frac{h^+_x (h^+_y - h^-_y)}{\zeta_{-1}\varphi_0} & -\frac{h^+_x h^-_y b_{2,\,1}}{\zeta_{-1}\varphi_1} & & -\frac{(h^+_x - h^-_x) h^+_y b_{2,\,1}}{\zeta_0\varphi_{-1}} & \frac{(h^+_x - h^-_x)(h^+_y - h^-_y)}{\zeta_0\varphi_0} & \frac{(h^+_x - h^-_x) h^-_y b_{2,\,1}}{\zeta_0\varphi_1} & & -\frac{h^-_x h^+_y b_{2,\,1}}{\zeta_1\varphi_{-1}} & \frac{h^-_x (h^+_y - h^-_y)}{\zeta_1\varphi_0} & \frac{h^-_x h^-_y b_{2,\,1}}{\zeta_1\varphi_1} & u^k_{2,\,1} \\
 & & & & & 0 & & & & & \times & u^k_{2,\,2} \\
\hline
 & & & & & 0 & & & & \times & \times & u^k_{3,\,0} \\
 & & & & & 0 & & & \times & \times & \times & u^k_{3,\,1} \\
 & & & & & 0 & \times & \underbrace{\cdots}_{n-2\,\text{col}} & \times & \times & \times & u^k_{3,\,2}
\end{array}
\right|
$$

The *modified* identity matrices $\mathbf{I}_{\text{im},C}$ and $\mathbf{I}_{\text{ex},C}$ structure is the same as in the case of the uniform grid. The boundary vector is given by

$$\mathbf{b}_C = (c_l, c_l, c_l; 0, 0, 0; 0, 0, 0; c_r h_x^-, c_r h_x^-, c_r h_x^-)^T.$$

Exponential fitting

By making use of the fact that

$$f(x, y, \epsilon) = \frac{y\epsilon}{2} \coth\left(\frac{y\epsilon}{2x}\right)$$

is a strictly increasing function of ϵ, we see that by substituting $a_{i,\,j}$ and $c_{i,\,j}$ with $f(a_{i,\,j}, d_{i,\,j}, \max(h_x^+, h_x^-))$ and $f(c_{i,\,j}, e_{i,\,j}, \max(h_y^+, h_y^-))$ the off-diagonal elements of $\mathbf{A}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}$ and $\mathbf{C}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}$ will be non-negative[3]. Furthermore, from the fact that $\max(h_y^+, h_y^-) \geq \{h_y^+, h_y^-, |h_y^+ - h_y^-|\}$ it follows that the diagonals of $\mathbf{A}_{\text{diff,A}} + \mathbf{A}_{\text{cov,A}}$ and $\mathbf{C}_{\text{diff,C}} + \mathbf{C}_{\text{cov,C}}$ will be negative.

7.2 Removing the cross-derivative term

In this section we will show that the cross-derivative term of the Heston-model can be removed by making use of the appropriate transformations[4]. Consider the following transformations

$$S \to \alpha(S, \sigma) \quad \text{and} \quad \sigma \to \beta(S, \sigma)$$

such that

$$V(S, \sigma) = V(\alpha(S, \sigma), \beta(S, \sigma))$$

in equation (2.28). We obtain

$$\begin{aligned}
\frac{\partial V}{\partial S} &= \frac{\partial \alpha}{\partial S}\frac{\partial V}{\partial \alpha} + \frac{\partial \beta}{\partial S}\frac{\partial V}{\partial \beta} \\
\frac{\partial V}{\partial \sigma} &= \frac{\partial \alpha}{\partial \sigma}\frac{\partial V}{\partial \alpha} + \frac{\partial \beta}{\partial \sigma}\frac{\partial V}{\partial \beta} \\
\frac{\partial^2 V}{\partial S^2} &= \frac{\partial^2 \alpha}{\partial S^2}\frac{\partial V}{\partial \alpha} + \left(\frac{\partial \alpha}{\partial S}\right)^2 \frac{\partial^2 V}{\partial \alpha^2} + \frac{\partial^2 \beta}{\partial S^2}\frac{\partial V}{\partial \beta} + \left(\frac{\partial \beta}{\partial S}\right)^2 \frac{\partial^2 V}{\partial \beta^2} + 2\frac{\partial \alpha}{\partial S}\frac{\partial \beta}{\partial S}\frac{\partial^2 V}{\partial \alpha \partial \beta} \\
\frac{\partial^2 V}{\partial \sigma^2} &= \frac{\partial^2 \alpha}{\partial \sigma^2}\frac{\partial V}{\partial \alpha} + \left(\frac{\partial \alpha}{\partial \sigma}\right)^2 \frac{\partial^2 V}{\partial \alpha^2} + \frac{\partial^2 \beta}{\partial \sigma^2}\frac{\partial V}{\partial \beta} + \left(\frac{\partial \beta}{\partial \sigma}\right)^2 \frac{\partial^2 V}{\partial \beta^2} + 2\frac{\partial \alpha}{\partial \sigma}\frac{\partial \beta}{\partial \sigma}\frac{\partial^2 V}{\partial \alpha \partial \beta} \\
\frac{\partial^2 V}{\partial S \partial \sigma} &= \frac{\partial^2 \alpha}{\partial S \partial \sigma}\frac{\partial V}{\partial \alpha} + \frac{\partial \alpha}{\partial S}\frac{\partial \alpha}{\partial \sigma}\frac{\partial^2 V}{\partial \alpha^2} + \frac{\partial^2 \beta}{\partial S \partial \sigma}\frac{\partial V}{\partial \beta} + \frac{\partial \beta}{\partial S}\frac{\partial \beta}{\partial \sigma}\frac{\partial^2 V}{\partial \beta^2} + \left(\frac{\partial \alpha}{\partial S}\frac{\partial \beta}{\partial \sigma} + \frac{\partial \alpha}{\partial \sigma}\frac{\partial \beta}{\partial S}\right)\frac{\partial^2 V}{\partial \alpha \partial \beta}.
\end{aligned}$$

[3] To see that $f(x, y, \epsilon)$ is a strictly increasing function of ϵ, consider

$$\begin{aligned}
\frac{\partial f}{\partial \epsilon} &= \frac{y}{2\sinh^2\theta}(\cosh\theta \sinh\theta - \theta) \\
&= \frac{y}{4\sinh^2\theta}(\sinh 2\theta - 2\theta) \\
&> 0
\end{aligned}$$

where $\theta = \frac{y\epsilon}{2x}$. The inequality follows from the fact that $\frac{\sinh 2\theta}{2\theta} > 1$ whenever $\theta \neq 0$.

[4] These calculations was inspired by conversations with Daniel J. Duffy.

Substituting into (2.28) and rearranging results in

$$\begin{aligned}\frac{\partial V}{\partial \tau} &= \left[rS\frac{\partial \alpha}{\partial S} + \kappa^*(\theta^* - \sigma)\frac{\partial \alpha}{\partial \sigma} + \tfrac{1}{2}\sigma S^2\frac{\partial^2 \alpha}{\partial S^2} + \rho\nu\sigma S\frac{\partial^2 \alpha}{\partial S \partial \sigma} + \tfrac{1}{2}\nu^2\sigma\frac{\partial^2 \alpha}{\partial \sigma^2}\right]\frac{\partial V}{\partial \alpha}\\ &+ \left[rS\frac{\partial \beta}{\partial S} + \kappa^*(\theta^* - \sigma)\frac{\partial \beta}{\partial \sigma} + \tfrac{1}{2}\sigma S^2\frac{\partial^2 \beta}{\partial S^2} + \rho\nu\sigma S\frac{\partial^2 \beta}{\partial S \partial \sigma} + \tfrac{1}{2}\nu^2\sigma\frac{\partial^2 \beta}{\partial \sigma^2}\right]\frac{\partial V}{\partial \beta}\\ &+ \left[\tfrac{1}{2}\sigma S^2\left(\frac{\partial \alpha}{\partial S}\right)^2 + \rho\nu\sigma S\frac{\partial \alpha}{\partial S}\frac{\partial \alpha}{\partial \sigma} + \tfrac{1}{2}\nu^2\sigma\left(\frac{\partial \alpha}{\partial \sigma}\right)^2\right]\frac{\partial^2 V}{\partial \alpha^2}\\ &+ \left[\tfrac{1}{2}\sigma S^2\left(\frac{\partial \beta}{\partial S}\right)^2 + \rho\nu\sigma S\frac{\partial \beta}{\partial S}\frac{\partial \beta}{\partial \sigma} + \tfrac{1}{2}\nu^2\sigma\left(\frac{\partial \beta}{\partial \sigma}\right)^2\right]\frac{\partial^2 V}{\partial \beta^2}\\ &+ \left[\sigma S\frac{\partial \alpha}{\partial S}\left(S\frac{\partial \beta}{\partial S} + \rho\nu\frac{\partial \beta}{\partial \sigma}\right) + \nu\sigma\frac{\partial \alpha}{\partial \sigma}\left(\nu\frac{\partial \beta}{\partial \sigma} + \rho S\frac{\partial \beta}{\partial S}\right)\right]\frac{\partial^2 V}{\partial \alpha \partial \beta}.\end{aligned}$$

Note that the coefficient of the cross-derivative term will be zero if

$$\sigma S\frac{\partial \alpha}{\partial S} = \nu\sigma\frac{\partial \alpha}{\partial \sigma} \quad \text{and} \quad S(1+\rho)\frac{\partial \beta}{\partial S} = -\nu(1+\rho)\frac{\partial \beta}{\partial \sigma}.$$

It is easy to show that

$$\alpha(S,\sigma) = \ln S + \frac{1}{\nu}\sigma \quad \text{and} \quad \beta(S,\sigma) = \ln S - \frac{1}{\nu}\sigma \tag{7.21}$$

satisfies these equations. Thus we have obtained transformations that will remove the cross-derivative term from the Heston PDE. This is in contrast to Zvan et al. [2003] where it is said that such transformations do not appear to be possible. The non-zero coefficient of the cross-derivative term was the main reason why we could not prove the extrapolated Yanenko scheme L_0-stable in section 6.3. Now that we have removed the cross-derivative term we can simply apply the L_0-stable Khaliq-Twizwell scheme, which we modified such that it incorporates convection terms as well, to solve the transformed Heston PDE. The main reason why we do not pursue this idea further is the complications that arise with the boundary conditions. Suppose that we want to solve the Heston PDE on the rectangular domain $[\ln S_{\text{min}}, \ln S_{\text{max}}] \times [0, \sigma]$ with the boundary conditions given by

$$\begin{aligned}\Delta_{\text{BC1}}V(\ln S_{\text{min}}, \sigma) &= 0, \quad \forall \sigma\\ \Delta_{\text{BC2}}V(\ln S_{\text{max}}, \sigma) &= 0, \quad \forall \sigma\\ \Delta_{\text{BC3}}V(\ln S, \sigma_{\text{min}}) &= 0, \quad \forall \ln S\\ \Delta_{\text{BC4}}V(\ln S, \sigma_{\text{max}}) &= 0, \quad \forall \ln S\end{aligned}$$

where Δ_{BC1}, Δ_{BC2}, Δ_{BC3} and Δ_{BC4} are linear differential operators. Using (7.21) to remove the cross-derivative term we see that the transformed boundary conditions are given on the boundaries of a non-rectangular domain, see figure 7.3

$$\begin{aligned}\widetilde{\Delta}_{\text{BC1}}V\Big|_{\alpha+\beta=\ln S_{\text{min}}} &= 0\\ \widetilde{\Delta}_{\text{BC2}}V\Big|_{\alpha+\beta=\ln S_{\text{max}}} &= 0\\ \widetilde{\Delta}_{\text{BC3}}V\Big|_{\alpha-\beta=\frac{2}{\nu}\sigma_{\text{min}}} &= 0\\ \widetilde{\Delta}_{\text{BC4}}V\Big|_{\alpha-\beta=\frac{2}{\nu}\sigma_{\text{max}}} &= 0\end{aligned}$$

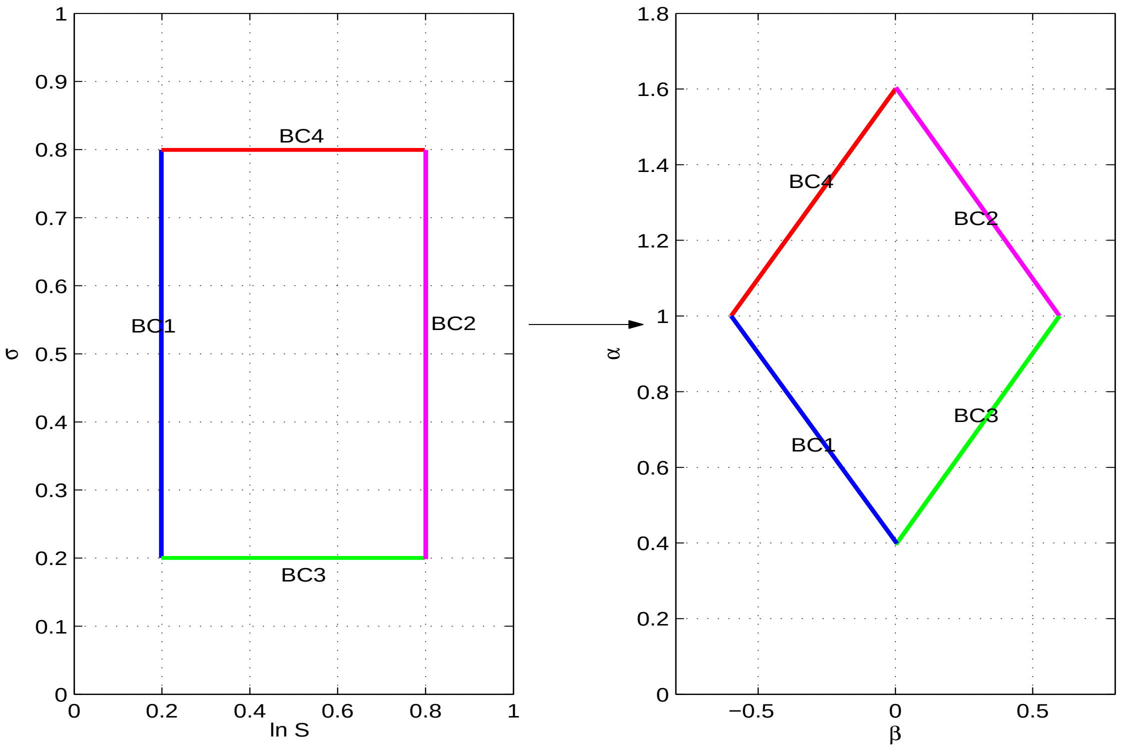


Figure 7.3: The original and transformed domain, where $\nu = 2$, $\ln S_{\text{min}} = 0.2$, $\ln S_{\text{max}} = 0.8$, $\sigma_{\text{min}} = 0.2$ and $\sigma_{\text{max}} = 0.8$.

for all α and β where $\widetilde{\Delta}_{\text{BC1}}$, $\widetilde{\Delta}_{\text{BC2}}$, $\widetilde{\Delta}_{\text{BC3}}$ and $\widetilde{\Delta}_{\text{BC4}}$ are the transformed boundary conditions. This non-rectangular domain will have an adverse effect on the ordered structure of the matrices, we feel that the transformed problem becomes more complicated than the untransformed problem.

7.3 Non-Smooth payoff functions

In financial mathematics we almost always work with functions that are not *smooth*, i.e. either the payoff function or the derivative of the payoff function is discontinuous. These discontinuities can lower the order of convergence of a scheme and have an adverse effect on the stability of the implied hedging parameters, see Heston and Zhou [2000] and Pooley et al. [2003]. In this section we will give three algorithms to handle these problems, namely: Rannacher time-marching, the averaging of initial conditions, and grid shifting.

7.3.1 Rannacher time-marching

Rannacher time marching has been implemented to smooth the initial data for Heston's stochastic volatility model in Ikonen and Toivanen [2005b], Ikonen and Toivanen [2005c] and Giles and Carter

[2006]. Say we have chosen a finite difference scheme, Crank-Nicolson for example, with a time increment of Δt. The Rannacher algorithm is then given by:

- Set the time increment to $\frac{\Delta t}{2}$ and make $2n$ steps with the classical fully implicit method.
- Change the time increment back to Δt and continue with the chosen scheme.

where $n \in \mathbb{N}$. In Giles and Carter [2006] it is shown that $n = 2$ is optimal. For $n > 2$ the first order accurate fully implicit scheme can have an adverse effect on the accuracy of the scheme and for $n < 2$ the initial data will not be smooth enough. The intuition behind this scheme is that a robust scheme must be used for the startup procedure which can "handle" the discontinuities. After a few time steps with the first order fully implicit scheme the initial data is smooth enough for the Crank-Nicolson scheme to be stable and second order accurate.

7.3.2 Averaging of initial conditions

In Thomee and Wahlbin [1974] and Pooley et al. [2003] they propose the following method to smooth the initial data, $\Psi(S, \sigma)$

$$u^0_{i,j} = \frac{1}{S_{i+\frac{1}{2}} - S_{i-\frac{1}{2}}} \int_{S_{i-\frac{1}{2}}}^{S_{i+\frac{1}{2}}} \Psi(S_i - y, \sigma_j) dy$$

where $S_{i+\frac{1}{2}}$ denotes the point halfway between S_i and S_{i+1}. Note that we only need to smooth the payoff function in the stock direction since the payoff functions of the contingent claims we consider will not be dependent on volatility.

7.3.3 Shifting the mesh

In Tavella and Randall [2000] they shift the grid such that the discontinuity of the initial condition (or the derivative of the initial condition) falls exactly between two adjacent nodes. It is suggested in Tavella and Randall [2000] and Pooley et al. [2003] that this simple remedy improves the rate of convergence.

Chapter 8

Numerical solution of stochastic volatility PDEs: Heston, Hull & White, and SABR

In this chapter we apply the two dimensional PDE solvers derived in chapters 5, 6 and 7 to obtain numerical solutions of the PDEs derived in sections 2.2.1, 2.2.2 and 2.2.3.

8.1 Stochastic volatility models

In this section the FDM schemes derived in the previous chapters are applied to obtain numerical solutions of the PDEs arising in stochastic volatility models. For illustrative purposes the valuation of European call and put options will be considered. For all stochastic volatility models the PDE governing the value of the derivatives can be written in the following form

$$\begin{aligned}\frac{\partial V}{\partial \tau} =& rS\frac{\partial V}{\partial S} + (p_\sigma - q_\sigma \lambda_\sigma(S,\sigma,t))\frac{\partial V}{\partial \sigma} \\ &+ \tfrac{1}{2}q_S^2\frac{\partial^2 V}{\partial S^2} + \rho q_S q_\sigma \frac{\partial^2 V}{\partial S \partial \sigma} + \tfrac{1}{2}q_\sigma^2\frac{\partial^2 V}{\partial \sigma^2} - rV \end{aligned} \tag{8.1}$$

on the truncated domain $(S,\sigma) \in [0, S_{\text{max}}] \times [0, \sigma_{\text{max}}]$, with the initial condition determined by the payoff function

$$V(S,\sigma,0) = \Psi(S,\sigma).$$

In order to apply the FDMs from the previous chapters we need to *remove* the discounting term from this PDE and prove that the PDE is parabolic. We can remove the discounting term by making use of the transformation $V(S,\sigma,\tau) = u(S,\sigma,\tau)e^{-r\tau}$. Thus the pricing problem becomes

$$\frac{\partial u}{\partial \tau} = a(S,\sigma)\frac{\partial^2 u}{\partial S^2} + 2b(S,\sigma)\frac{\partial^2 u}{\partial S \partial \sigma} + c(S,\sigma)\frac{\partial^2 u}{\partial \sigma^2} + d(S)\frac{\partial u}{\partial S} + e(S,\sigma)\frac{\partial u}{\partial \sigma} \tag{8.2}$$

with the initial condition

$$u(S,\sigma,0)=\Psi(S,\sigma)$$

where

$$\begin{array}{ll} a(S,\sigma)=\frac{1}{2}q_S^2 & d(S)=rS \\ c(S,\sigma)=\frac{1}{2}q_\sigma^2 & e(S,\sigma)=(p_\sigma-q_\sigma\lambda_\sigma(S,\sigma,t)) \\ b(S,\sigma)=\frac{1}{2}\rho q_S q_\sigma. & \end{array}$$

We will also have to transform the original boundary conditions using $V=ue^{-r\tau}$. The final step is to confirm the inequalities. It is trivial to see that $a(S,\sigma)>0$ and $c(S,\sigma)>0$, for the final inequality we have

$$ac-b^2=\frac{1}{4}q_S^2q_\sigma^2(1-\rho^2)>0.$$

The inequality follows from the fact that $\rho\in(-1,1)$.

8.1.1 European put

To find the value of a European put option we need to numerically solve (8.1) with the initial condition

$$V(S,\sigma,0)=\Psi(S,\sigma)=\max(K-S,0)$$

and the boundary conditions

$$\begin{array}{ll} V(0,\sigma,\tau)=Ke^{-r\tau} & \\ \frac{\partial V}{\partial S}(S_{\text{max}},\sigma,\tau)=0 & \\ \frac{\partial V}{\partial \tau}(S,\sigma,\tau)=d(S)\frac{\partial V}{\partial S}+e(S,\sigma)\frac{\partial V}{\partial \sigma}-rV, & \sigma=0 \\ \frac{\partial V}{\partial \tau}(S,\sigma,\tau)=a(S,\sigma)\frac{\partial^2 V}{\partial S^2}+d(S)\frac{\partial V}{\partial S}-rV, & \sigma=\sigma_{\text{max}}. \end{array}$$

To see where the first boundary condition comes from, note that when the stock price is zero then the European put option will definitely give its owner a cash inflow of K at maturity. The value of this cash flow at time t is $Ke^{-r\tau}$. The second boundary conditions comes from the assumption that the price of a put will be independent of the underlying for very large values of the underlying. Neither Dirichlet nor Neumann boundary conditions are posed at $\sigma=0$ and $\sigma=\sigma_{\text{max}}$, instead we require that the PDE itself must be satisfied on these boundaries, this is known as a smoothing condition. The third boundary condition arises from the fact the the parabolic part of the PDE on the boundary $\sigma=0$ is zero for all the stochastic volatility models that we are going to consider. Similar boundary conditions to those discussed here can be found in the literature, see Heston [1993], Zvan et al. [1998], Ikonen and Toivanen [2005a] and Duffy [2006] for example. Equivalently (8.2) can be solved with the following boundary conditions

$$\begin{aligned}
&u(0,\sigma,\tau) = K\\
&\frac{\partial u}{\partial S}(S_{\text{max}},\sigma,\tau) = 0\\
&\frac{\partial u}{\partial \tau}(S,\sigma,\tau) = d(S)\frac{\partial u}{\partial S} + e(S,\sigma)\frac{\partial u}{\partial \sigma}, \qquad \sigma = 0\\
&\frac{\partial u}{\partial \tau}(S,\sigma,\tau) = a(S,\sigma)\frac{\partial^2 u}{\partial S^2} + d(S)\frac{\partial u}{\partial S}, \qquad \sigma = \sigma_{\text{max}}.
\end{aligned}$$

8.1.2 European call

Similarly to a European put, (8.1) needs to be solved with the following initial condition

$$V(S,\sigma,0) = \Psi(S,\sigma) = \max(S-K,0)$$

to obtain the correct value surface. The boundary conditions are given by

$$\begin{aligned}
&V(0,\sigma,\tau) = 0\\
&\frac{\partial V}{\partial S}(S_{\text{max}},\sigma,\tau) = e^{-r\tau}\\
&\frac{\partial V}{\partial \tau}(S,\sigma,\tau) = d(S)\frac{\partial V}{\partial S} + e(S,\sigma)\frac{\partial V}{\partial \sigma} - rV, \qquad \sigma = 0\\
&\frac{\partial V}{\partial \tau}(S,\sigma,\tau) = a(S,\sigma)\frac{\partial^2 V}{\partial S^2} + d(S)\frac{\partial V}{\partial S} - rV, \qquad \sigma = \sigma_{\text{max}}.
\end{aligned}$$

Equivalently (8.2) can be solved with the following boundary conditions

$$\begin{aligned}
&u(0,\sigma,\tau) = 0\\
&\frac{\partial u}{\partial S}(S_{\text{max}},\sigma,\tau) = 1\\
&\frac{\partial u}{\partial \tau}(S,\sigma,\tau) = d(S)\frac{\partial u}{\partial S} + e(S,\sigma)\frac{\partial u}{\partial \sigma}, \qquad \sigma = 0\\
&\frac{\partial u}{\partial \tau}(S,\sigma,\tau) = a(S,\sigma)\frac{\partial^2 u}{\partial S^2} + d(S)\frac{\partial u}{\partial S}, \qquad \sigma = \sigma_{\text{max}}.
\end{aligned}$$

8.2 The SABR model

When we consider European options we need not concern ourselves with the dynamic SABR model. By comparing (2.21) with (2.12) we see that we can use the arguments in the previous section after we make the following substitutions,

$$\begin{aligned}
&q_S = \sigma F^\beta \qquad && q_\sigma = \nu\sigma\\
&S = F && D = -rS\\
&p_\sigma - q_\sigma\lambda_\sigma = 0.
\end{aligned}$$

To find the value of a European option in the SABR world with a payoff $\Psi(F,\sigma)$ at maturity, we need to solve (2.21) with the payoff function as a terminal condition. Using the arguments above we deduce that we can obtain the solution of the problem above by solving

$$\frac{\partial u}{\partial \tau} = \tfrac{1}{2}\sigma^2 F^{2\beta}\frac{\partial^2 u}{\partial F^2} + \rho\nu\sigma^2 F^{\beta}\frac{\partial^2 u}{\partial F\partial\sigma} + \tfrac{1}{2}\nu^2\sigma^2\frac{\partial^2 u}{\partial \sigma^2} \tag{8.3}$$

with the initial condition

$$u(F,\sigma,0) = \Psi(F,\sigma)$$

on the truncated domain $(F,\sigma) \in [0, F_{\max}] \times [0, \sigma_{\max}]$. We can transform back to financial variables, from u to V, via $V(F,\sigma,\tau) = u(F,\sigma,\tau)e^{-r\tau}$.

8.3 The Heston model

By comparing (2.28) with (2.12) we see that we can use the arguments in section 8.1 after we make the following substitutions,

$$\begin{aligned} q_S &= \sqrt{\sigma}S & q_\sigma &= \xi\sigma \\ p_\sigma - q_\sigma\lambda_\sigma &= \mu\sigma & D &= 0. \end{aligned}$$

To find the value of a European option in the Heston model with a payoff $\Psi(F,\sigma)$ at maturity, we need to solve (2.28) with the payoff function as a terminal condition. Using the arguments in section 8.1 we deduce that we can obtain the solution of the problem above by solving

$$\begin{aligned} \frac{\partial u}{\partial \tau} =& rS\frac{\partial u}{\partial S} + \kappa^*(\theta^* - \sigma)\frac{\partial u}{\partial \sigma} \\ &+ \tfrac{1}{2}\sigma S^2\frac{\partial^2 u}{\partial S^2} + \rho\nu\sigma S\frac{\partial^2 u}{\partial S\partial\sigma} + \tfrac{1}{2}\nu^2\sigma\frac{\partial^2 u}{\partial \sigma^2} \end{aligned}$$

with the initial condition

$$u(S,\sigma,0) = \Psi(S,\sigma)$$

on the truncated domain $(S,\sigma) \in [0, S_{\max}] \times [0, \sigma_{\max}]$. We can transform back to financial variables, from u to V, via $V(S,\sigma,\tau) = u(S,\sigma,\tau)e^{-r\tau}$.

8.4 The Hull & White model

By comparing (2.29) with (2.12) we see that we can use the arguments in section 8.1 after we make the following substitutions,

$$\begin{aligned} q_S &= \sqrt{\sigma}S & q_\sigma &= \xi\sigma \\ p_\sigma - q_\sigma\lambda_\sigma &= \mu\sigma & D &= 0. \end{aligned}$$

To find the value of a European option in the Hull & White model with a payoff $\Psi(F,\sigma)$ at maturity, we need to solve (2.29) with the payoff function as a terminal condition. Using the arguments in section 8.1 we deduce that we can obtain the solution of the problem above by solving

$$\begin{aligned} \frac{\partial u}{\partial \tau} =& rS\frac{\partial u}{\partial S} + \mu\sigma\frac{\partial u}{\partial \sigma} \\ &+ \tfrac{1}{2}\sigma S^2\frac{\partial^2 u}{\partial S^2} + \rho\xi\sigma^{3/2}S\frac{\partial^2 u}{\partial S\partial\sigma} + \tfrac{1}{2}\xi^2\sigma^2\frac{\partial^2 u}{\partial \sigma^2} \end{aligned} \tag{8.4}$$

with the initial condition

$$u(S, \sigma, 0) = \Psi(S, \sigma)$$

on the truncated domain $(S, \sigma) \in [0, S_{\text{max}}] \times [0, \sigma_{\text{max}}]$. We can transform back to financial variables, from u to V, via $V(S, \sigma, \tau) = u(S, \sigma, \tau)e^{-r\tau}$.

Chapter 9

Numerical Results

In this chapter we compare the numerical solutions obtained with the finite difference method for a European call in the Heston and SABR model with the known solutions. We will conclude this chapter by giving an example to illustrate how exponential fitting improves the fully implicit method for the evaluation of one dimensional convection diffusion equations.

9.1 The algorithm

The numerical results, regarding stochastic volatility models, in this chapter are obtained with the two dimensional finite difference methods discussed in chapters 5, 6 and 7. For the two dimensional case we use non-uniform grids, structured such that they are most dense near the strike of the call options and the spot volatility, see figure 9.1. The non-uniform grid is generated with the grid generating function defined in equation (7.1). Let $S_{\hat{i}}$ and $S_{\hat{i}\,+\,1}$ be the nodes immediately adjacent the strike K. To ensure that the strike of the option is placed precisely in the middle of $S_{\hat{i}}$ and $S_{\hat{i}\,+\,1}$, we make use of a two dimensional optimization procedure on S_{max} and p to ensure that the following equalities hold:

$$\begin{aligned} K &= \frac{S_{\hat{i}\,+\,1} + S_{\hat{i}}}{2} \\ g(S_{\text{max}}) &= S_{\text{max}}. \end{aligned}$$

We make use of exponential fitting, as shown in section 7.1, whenever there are non-zero convection terms. For the implementation of splitting methods we have written a procedure that dynamically reorders the solution vector such that only tri-diagonal matrices are solved.

Since two dimensional finite difference schemes require the iterative solution of large matrices, these schemes must be implemented in a computer language that can handle large data sets. We chose to implement the finite difference schemes in MATLAB since large sparse systems of equations are easy to construct and solve in MATLAB. We used MATLAB's `sparse` function to construct the relevant matrices. Once the matrices are constructed it is easy to implement the different time marching procedures: fully implicit, Crank-Nicolson, Yanenko and extrapolation schemes. The engine for the Yanenko scheme

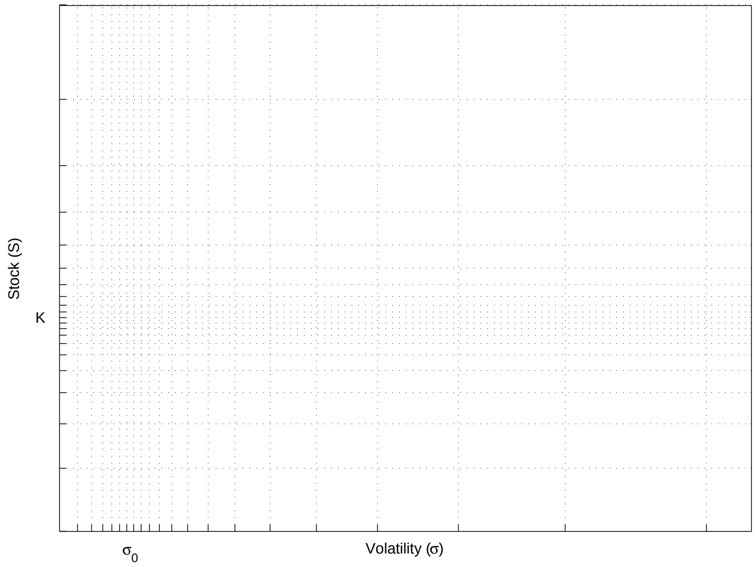


Figure 9.1: A structured non-uniform grid with a concentration point (σ_0, K).

is given by the following piece of MATLAB code[1]

```
% Preliminary definitions

% Iim1ForA, Iex1ForA, Iim2ForC, Iex2ForC, boundaryVec1ForA and
% boundaryVec2ForC are defined in section 5.4.2.
ThetaA1 = (Iim1ForA - dt*A1);
ThetaA1B = (Iex1ForA + dt*B1);

ThetaC2 = (Iim2ForC - dt*C2);
ThetaC2B = (Iex2ForC + dt*B2);

for i = 1:l % l = number of time steps, m = number of steps in the stock
            % direction and n = number steps in the volatility direction.
    u = ThetaA1\(ThetaA1B*u + boundaryVec1ForA);
    u = order1to2(u,m,n); % Change the ordering such that we can use the
                          % tri-diagonal form of C. This is done such that
                          % only tri-diagonal matrices are inverted
                          % (See Chapter 5).
    u = ThetaC2\(ThetaC2B*u + boundaryVec2ForC);
    u = order2to1(u,m,n); % Change the ordering such that we can use the
                          % tri-diagonal form of A.
end
```

[1] We omit the engines for the extrapolation schemes since they are lengthy and do not give any extra intuition.

```
function u2 = order1to2(u1,m,n)
    % Change the ordering of the elements in u from order 1 to order 2.
    u1 = reshape(u1,m+1,n+1);
    u1 = u1';
    u2 = reshape(u1,(m+1)*(n+1),1);

function u1 = order2to1(u2,m,n)
    % Change the ordering of the elements in u from order 2 to order 1.
    u2 = reshape(u2,n+1,m+1);
    u2 = u2';
    u1 = reshape(u2,(m+1)*(n+1),1);
```

The code was executed on a personal computer with an AMD 2800+ CPU and 1GB RAM.

9.2 Heston model

The benchmark solution for the value a European option in the Heston model is obtained with the VBA code given in Vogt [2004]. For the numerical results of this section we used the following parameter values[2]

$$\begin{aligned}&K = 123.4, \quad \sigma_0 = 0.02, \quad \tau = 1, \quad r = 0.1, \quad \rho = -0.9\\ &\kappa^* = 1.988937, \quad \theta^* = 0.011876, \quad \nu = 0.15\\ &\sigma_{\min} = 0, \quad \sigma_{\max} = 0.2\\ &S_{\min} = 0, \quad S_{\max} = 300\\ &m = 200\\ &n = 100\\ &l = 81\end{aligned}$$

where m, n and l are the number of grid points in the x, y and τ direction respectively. The boundaries $S_{\max}$ and $\sigma_{\max}$ must be chosen such that the boundary conditions at these boundaries do not have an adverse effect on the interior point of interest. Ideally we would like to compare the finite difference approximations to the benchmark solutions on a whole range of volatilities $[\sigma_{\min}, \sigma_{\max}]$. Since the benchmark solution given in Vogt [2004] is computationally too intense for the construction of value surfaces we only compare the finite difference solution to the benchmark solution at a single volatility value. Classical Alternating Direction Implicit schemes (ADI-schemes) are problematic when the coefficient of the cross-derivative term is large, see Kluge [2002] and Duffy [2006]. In Hout and Welfert [2006] an unconditionally stable ADI scheme is proposed for the solution of two dimensional convection diffusion

[2] Note that the spot volatility can be quite small, this means that the boundary condition at $\sigma_{\min}$ may have an adverse effect on the accuracy of the scheme at this point. We can work around this problem my making the grid more dense near the boundary and imposing appropriate boundary conditions. It turns out that the smoothing condition at $\sigma_{\min}$ is appropriate.

equations with mixed derivatives[3]. Although we could not prove the extrapolated schemes L_0-stable for the case when $\rho \neq 0$, we were unable to find a parameter set for which these schemes are unstable. Figure 9.2, 9.3 and 9.4 shows the the error at $\sigma = 0.02$ when we use the fully implicit, Crank-Nicolson and Yanenko schemes respectively to value a European call in the Heston model.

	ATM-value	ATM relative error (%)	Average absolute error	Elapsed time (seconds)
Exact solution	13.8571			
Fully Explicit	-3.7935×10^{169}	-2.7376×10^{170}	6.3716×10^{168}	6.52
Fully Implicit	13.8588	0.01243	0.0096	48.32
CN	14.1853	2.3683	0.0603	49.85
Craig-Sneyd	13.8598	0.01926	0.0020	18.05
Yanenko	13.8617	0.0327	0.0106	7.14
KT3	13.8569	-0.0015	3.8468×10^{-4}	18.05
KT4	13.8570	-0.0011	3.6191×10^{-4}	32.39

From the table above we see that although the fully implicit scheme is stable, it is computationally inefficient. The Crank-Nicolson scheme is computationally inefficient and unstable, hence we conclude that the Crank-Nicolson scheme is not appropriate for this problem. The Craig-Sneyd scheme has better convergence than the Yanenko method but worse than the extrapolation schemes. Yanenko splitting is stable with a slightly higher error than the fully implicit scheme due to splitting error. Since we solve only tri-diagonal systems with the Yanenko scheme it is almost as computationally efficient as the fully explicit scheme where no matrix inversions are required. The lowest error is obtained with the extrapolation type schemes at a reasonable computational cost.

From figure 9.3 we see that the instability is greatest at the strike, this supports the idea that the discontinuous derivative of the payoff function has an adverse effect on the stability of the Crank-Nicolson scheme.

[3] Stability is proven with von Neumann stability analysis, which is not a suitable method if the initial data is not smooth.

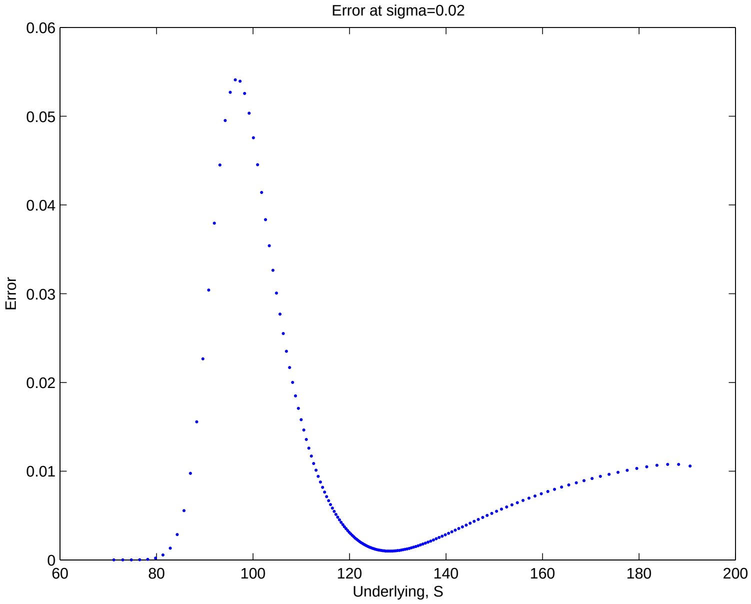


Figure 9.2: The difference between the true Heston solution and the solution obtained with the fully implicit method at $\sigma = 0.02$.

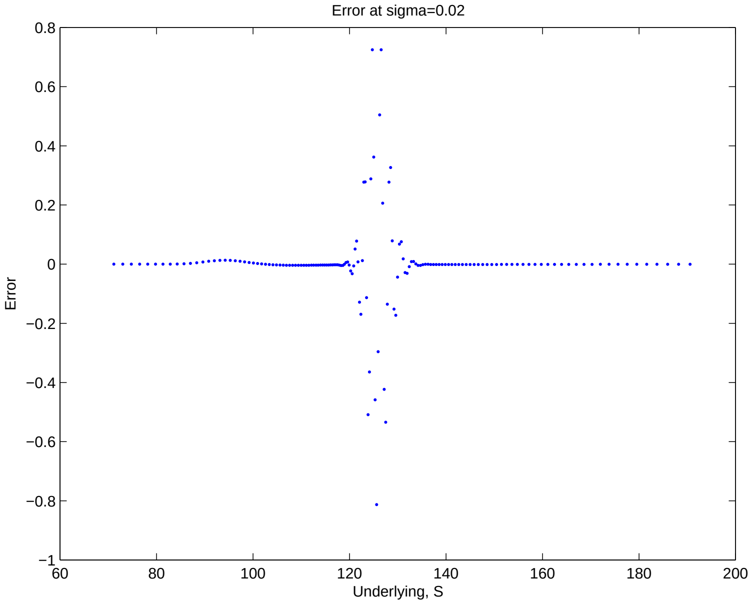


Figure 9.3: The difference between the true Heston solution and the solution obtained with the Crank-Nicolson at $\sigma = 0.02$.

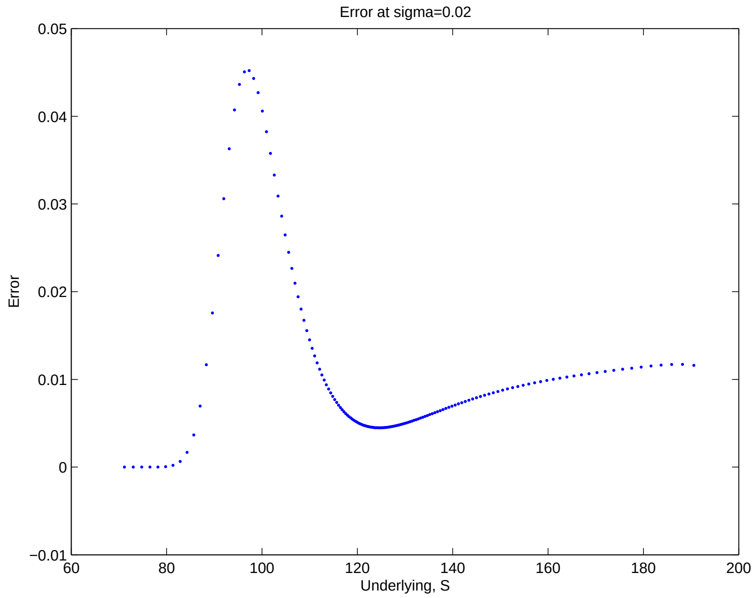


Figure 9.4: The difference between the true Heston solution and the solution obtained with the Yanenko scheme $\sigma = 0.02$.

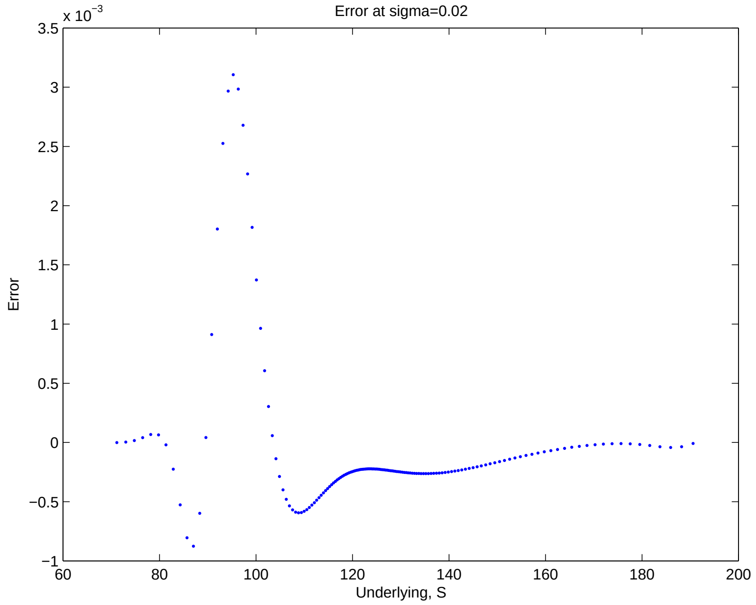


Figure 9.5: The difference between the true Heston solution and the solution obtained with the third order extrapolated Yanenko scheme at $\sigma = 0.02$.

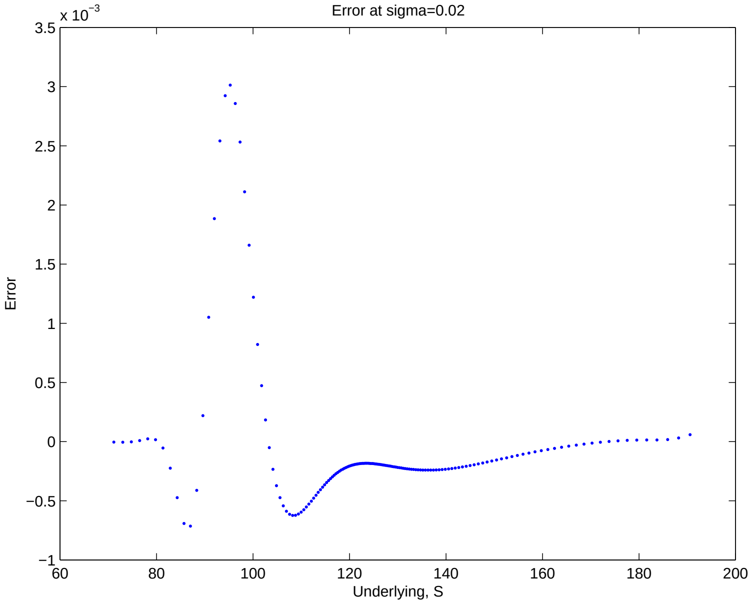


Figure 9.6: The difference between the true Heston solution and the solution obtained with the fourth order extrapolated Yanenko scheme at $\sigma = 0.02$.

The following tables show the results for different correlations in the Craig-Sneyd scheme, Yanenko scheme, extrapolated Yenenko scheme and the Crank-Nicolson scheme respectively

Correlation	Craig-Sneyd	Exact	Absolute difference
-0.9000	13.8598	13.8571	0.0027
-0.8000	13.8197	13.8173	0.0024
-0.7000	13.7785	13.7763	0.0022
-0.6000	13.7360	13.7341	0.0019
-0.5000	13.6922	13.6906	0.0016
-0.4000	13.6470	13.6457	0.0013
-0.3000	13.6002	13.5993	0.0009
-0.2000	13.5518	13.5512	0.0006
-0.1000	13.5016	13.5013	0.0003
0	13.4494	13.4495	0.0001

Correlation	Yanenko	Exact	Absolute difference
-0.9000	13.8617	13.8571	0.0045
-0.8000	13.8217	13.8173	0.0044
-0.7000	13.7805	13.7763	0.0042
-0.6000	13.7382	13.7341	0.0040
-0.5000	13.6945	13.6906	0.0039
-0.4000	13.6494	13.6457	0.0037
-0.3000	13.6028	13.5993	0.0035
-0.2000	13.5545	13.5512	0.0034
-0.1000	13.5045	13.5013	0.0032
0	13.4524	13.4495	0.0029

Correlation	KT4	Exact	Absolute difference
-0.9000	13.8570	13.8571	0.0002
-0.8000	13.8171	13.8173	0.0002
-0.7000	13.7762	13.7763	0.0002
-0.6000	13.7340	13.7341	0.0002
-0.5000	13.6905	13.6906	0.0002
-0.4000	13.6455	13.6457	0.0002
-0.3000	13.5991	13.5993	0.0002
-0.2000	13.5510	13.5512	0.0002
-0.1000	13.5011	13.5013	0.0002
0	13.4493	13.4495	0.0002

Correlation	Crank-Nicolson	Exact	Absolute difference
-0.9000	13.4857	13.8571	0.3714
-0.8000	13.8035	13.8173	0.0138
-0.7000	13.7750	13.7763	0.0013
-0.6000	13.7331	13.7341	0.0011
-0.5000	13.6896	13.6906	0.0011
-0.4000	13.6447	13.6457	0.0010
-0.3000	13.5983	13.5993	0.0009
-0.2000	13.5504	13.5512	0.0008
-0.1000	13.5007	13.5013	0.0006
0	13.4491	13.4495	0.0004

From these tables we can see that the Craig-Sneyd scheme as well as Crank-Nicolson scheme's accuracy decreases as the correlation becomes more negative. Both the Yanenko scheme and the extrapolated Yanenko scheme's accuracy are almost independent of the size of the correlation.

9.3 SABR model

From the previous section we can deduce that the third order extrapolated Yanenko scheme gives acceptable results. It is well known that the perturbation expansion of the SABR model requires the volatility of the volatility, ν, to be small and the time to maturity, T, to be relatively short, see Skabelin [2005] for example. Thus we cannot compare the finite difference solutions to the perturbation solutions for arbitrary parameter sets. Consider the following parameter set

$$\begin{aligned} K &= 100, \quad T = 0.5, \quad r = 0.02, \quad \beta = 0.7, \quad \rho = 0, \quad \nu = 0.1 \\ \sigma_{\text{min}} &= 0, \quad \sigma_{\text{max}} = 0.2 \\ F_{\text{min}} &= 0, \quad F_{\text{max}} = 300 \\ m &= 250 \\ n &= 128 \\ l &= 81 \end{aligned}$$

where m, n and l are the number of grid points in the x, y and τ direction respectively. Unlike for the benchmark solution of the Heston model in section 9.2, the pricing formulae given in Hagan et al. [2002] is instantaneous, hence we can compute option values for a whole range of volatility like parameters. The spot value of the second stochastic process in the SABR model is not the spot volatility of the underlying but is linked to the at-the-money implied volatility and the other calibration parameters, see West [2005]

$$\sigma_0 = f(F, \sigma_{\text{ATM}}, \rho, \beta, \nu, T).$$

In this parameter set we chose ν and T relatively small to ensure that the closed form approximations of European options in the SABR model, derived in Hagan et al. [2002], gives accurate results. We chose $\rho = 0$ to be certain that the finite difference scheme is L_0-stable and convergent for the test case. For the results that follow we used the extrapolated Yanenko scheme on a grid with $m = 200$, $n = 100$ and $l = 81$. Figure 9.7 shows the error obtained when we use the *ideal* parameter set. From the figure we see that the error spikes near the strike of the option, this is not too surprising since the first derivative of the payoff function is discontinuous at this point.

Figures 9.8, 9.9 and 9.10 shows the error surfaces obtained if we perturb ρ, ν and T respectively. From these figures we see that the perturbations do not have a dramatic effect on the error surface in the domain of importance. Figure 9.11 shows the error surface if we perturb all the parameters at once. We deduce that the closed form SABR formulae fails to give accurate solutions for a case when all the parameters are perturbed.

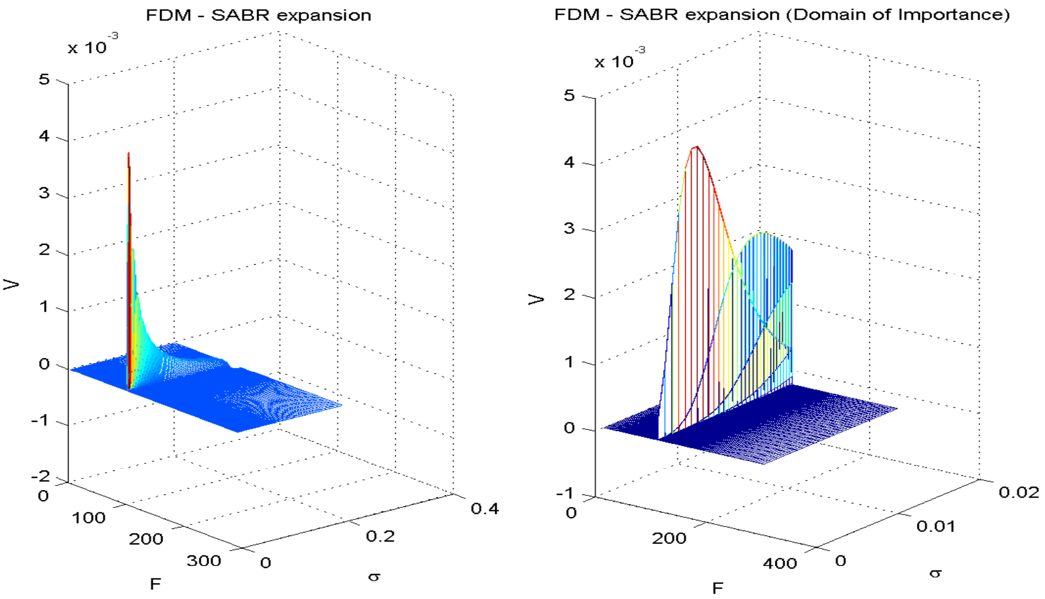


Figure 9.7: The error obtained on both the original domain and a truncated domain of importance for the case when, $K = 100$, $T = 0.5$, $r = 0.02$, $\beta = 0.7$, $\rho = 0$ and $\nu = 0.1$.

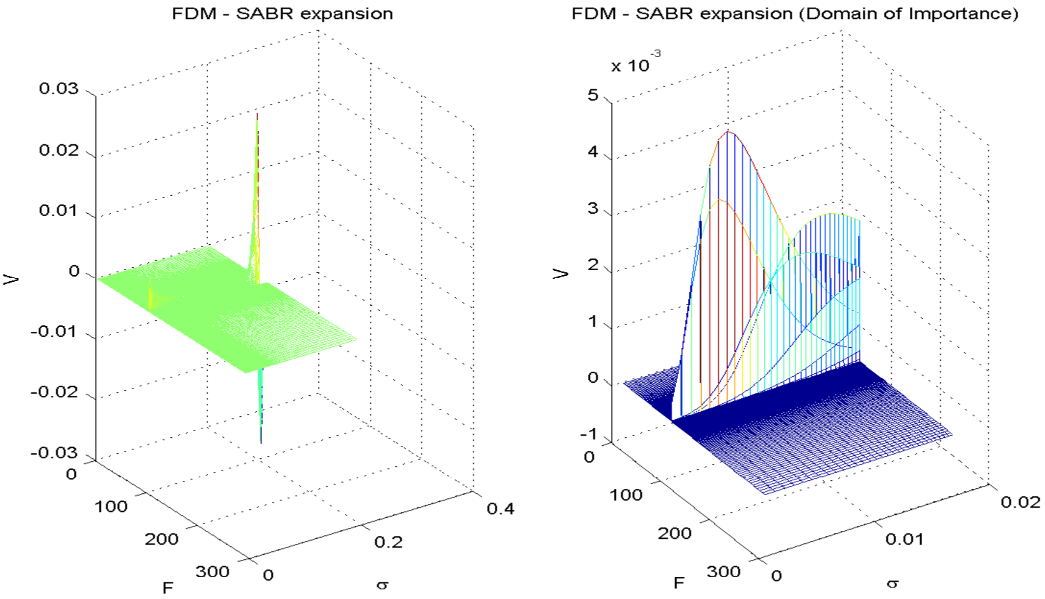


Figure 9.8: The error obtained on both the original domain and a truncated domain of importance for the case when, $K = 100$, $T = 0.5$, $r = 0.02$, $\beta = 0.7$, $\rho = -0.9$ and $\nu = 0.1$.

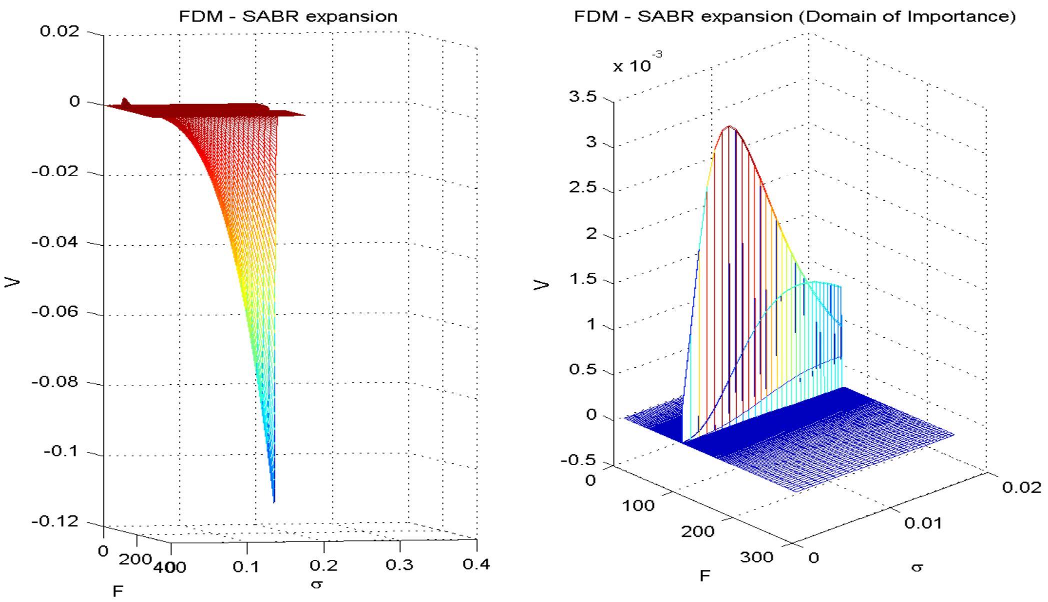


Figure 9.9: The error obtained on both the original domain and a truncated domain of importance for the case when, $K = 100$, $T = 0.5$, $r = 0.02$, $\beta = 0.7$, $\rho = 0$ and $\nu = 1$.

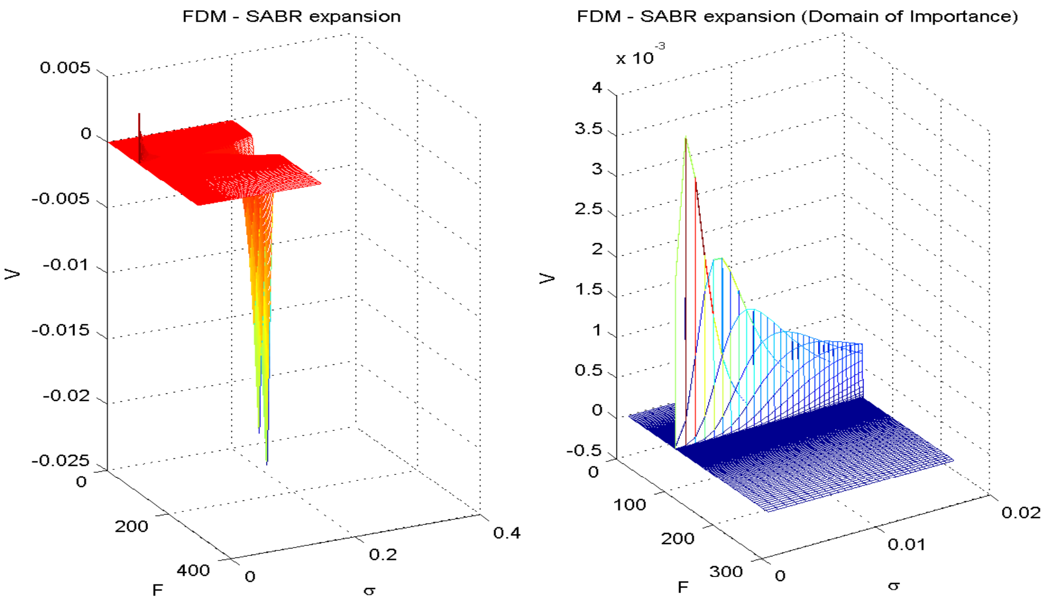


Figure 9.10: The error obtained on both the original domain and a truncated domain of importance for the case when, $K = 100$, $T = 4$, $r = 0.02$, $\beta = 0.7$, $\rho = 0$ and $\nu = 0.1$.

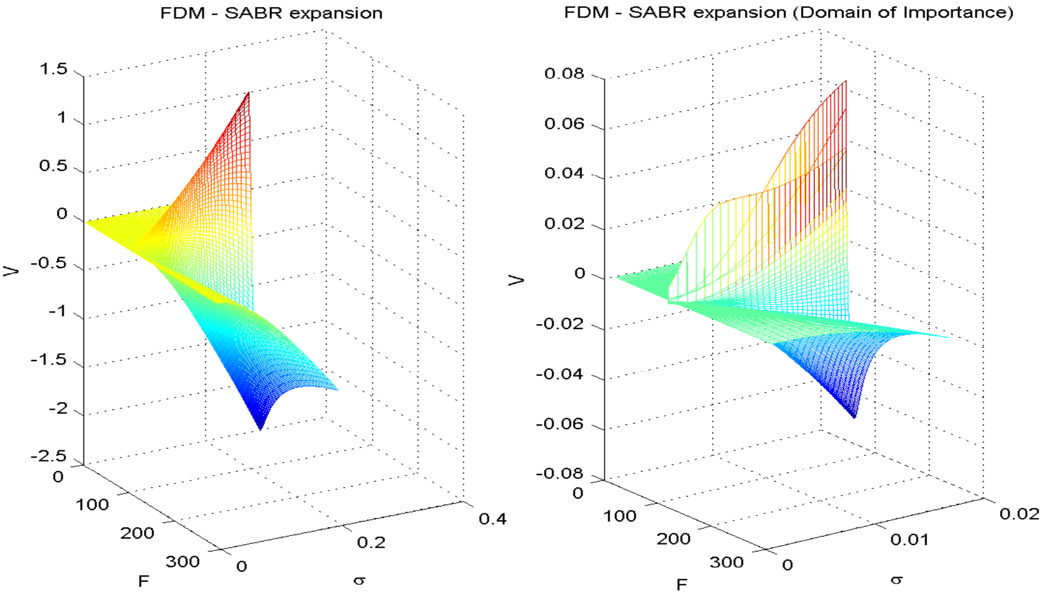
9.4 Exponential fitting

Consider an at-the-money European call option with a strike $K = 50$, maturity $T = 5/12$, volatility $\sigma = 0.07$ and risk-free rate $r = 0.46$. We compute the price and the delta of the option with both the closed form Black-Scholes equation and the one dimensional fully implicit method, see figure 9.12. We used the following parameters for the finite difference method, $m = 80$, $l = 50$ and $S_{\max} = 200$. From figure 9.12 we see that although the fully implicit method gives a good approximation of the price, it does not give a stable solution for the delta of the call option. This follows from the fact that the unfitted fully implicit method is not L_0-stable for convection-diffusion PDEs, see section 4.3.1.

In section 4.3.1 we showed that the exponential fully implicit scheme is unconditionally L_0-stable. Figure 9.13 shows that the exponentially fitted fully implicit method gives stable solutions for both the price and the delta of the option.

9.5 Conclusion

The accurate prices obtained for the Heston model with the VBA code given in Vogt [2004] produces value surfaces at a high computational cost. The proposed extrapolated Yanenko scheme can be used to obtain value surfaces at a much lower computational cost. Value surfaces obtained with the FDM can in turn be used to give approximations of the delta, gamma and vega surfaces. Although we where unable to prove the extrapolated Yanenko scheme L_0-stable whenever $\rho \neq 0$, extensive experiments with $\rho \neq 0$ have as yet not resulted in a single case of instability.



FDM - SABR expansion

FDM - SABR expansion (Domain of Importance)

Figure 9.11: The error obtained on both the original domain and a truncated domain of importance for the case when, $K = 100$, $T = 4$, $r = 0.02$, $\beta = 0.7$, $\rho = -0.9$ and $\nu = 1$

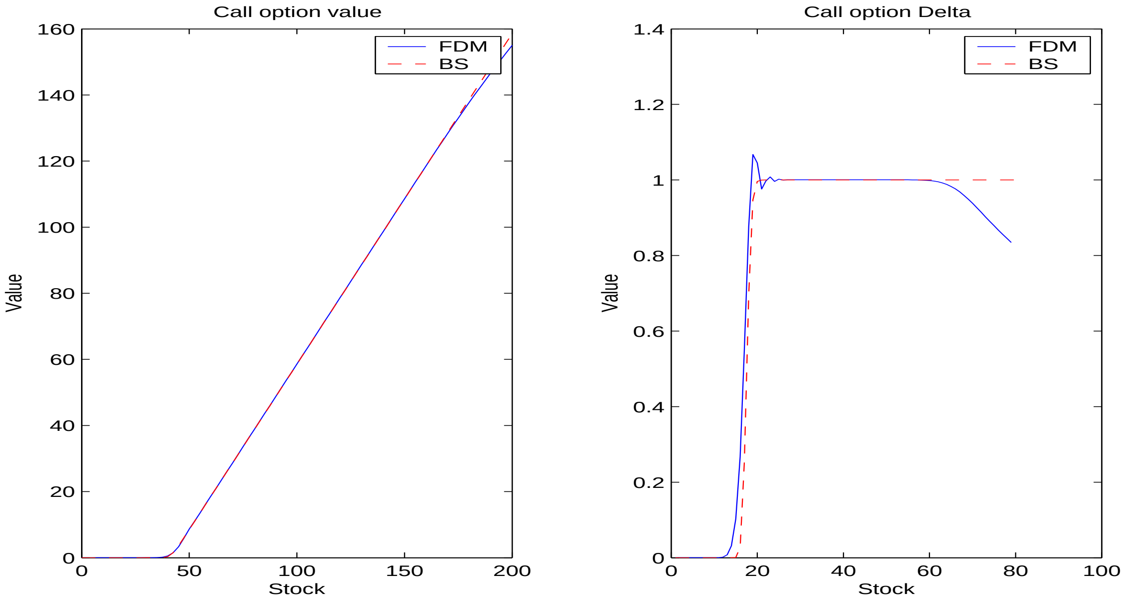


Figure 9.12: The price and the delta of a European call option in the Black-Scholes world computed with the closed form solution and the classical fully implicit method.

Although the analytical solutions of the SABR model derived in Hagan et al. [2002] gives instantaneous results, these solutions are perturbation expansions. Hence there might be parameter values for which the analytical solutions are not appropriate. If one calibrates the SABR model and obtain a *bad* parameter set, it would be inconsistent to price non-vanilla European options with any other approximation method[4].

From section 9.4 we see that the fitting method proposed in Duffy [2006] improves the, already robust, one dimensional fully implicit method. We used this fitting procedure to improve the stability properties of our two dimensional schemes.

[4] Calibration can be done as discussed in West [2005].

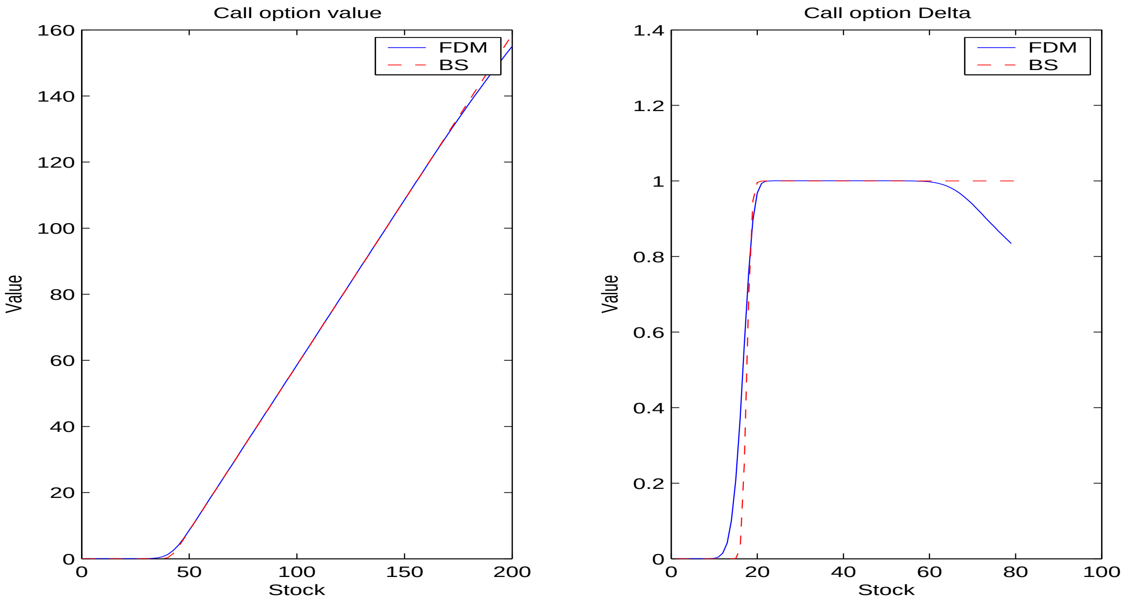


Figure 9.13: The price and the delta of an European call option in the Black-Scholes world computed with the closed form solution and the exponentially fitted fully implicit method.

Bibliography

Bruce P. Ayati, Glenn F. Webb, and Alexander R. A. Anderson. Computational methods and results for structured multiscale models of tumor invasion. *Multiscale Modeling & Simulation: A SIAM Interdisciplinary Journal*, 5(1):1–20, 2006.

Tomas Björk. *Arbitrage Theory in Continuous Time*. Oxford University Press, 1998.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 73(May-June):637–659, 1973.

T. D. Bui. A note on the Rosenbrock procedure. *Mathematics of Computation,*, 33(147):971–975, July 1979.

J. R. Cash. Two new finite difference schemes for parabolic equations. *SIAM Journal on Numerical Analysis*, 21(3):433–446, June 1984. URL `http://www.jstor.org/view/00361429/di976236/97p03064/0`.

I. J. D. Craig and A. D. Sneyd. An alternating direction implicit scheme for parabolic equations with mixed derivatives. *Computers and Mathematics with Applications*, 16(4):341–350, 1988.

Germund G. Dahlquist. A special stability problem for linear multistep methods. *BIT Numerical Mathematics*, 3(1):27–43, March 1963.

Emanuel Derman and Iraj Kani. Riding on a smile. *Risk*, 7(2), 1994.

Daniel J. Duffy. *Finite difference methods in financial engineering : A Partial Differential Equation Approach*. John Wiley and Sons Ltd., 2006.

Daniel J. Duffy. *Financial Instrument Pricing Using C++*. John Wiley and Sons Ltd., 2005.

Daniel J. Duffy. Robust and accurate finite difference methods in option pricing one factor models, 2003. URL `http://www.datasim-component.com/downloads/financial/daniel3.pdf`.

Bruno Dupire. Pricing with a smile. *Risk*, 7(1), January 1994.

G. Fairweather and A. R. Mitchell. A new computational procedure for a.d.i. methods. *SIAM Journal on Numerical Analysis,*, 4(2):163–170, June 1967.

Jean-Pierre Fouque and Tracey Andrew Tullie. Variance reduction for monte carlo simulation in a stochastic volatility environment. *Quantitative Finance*, 2(1):24–30, February 2002.

Michael Giles and Rebecca Carter. Convergence analysis of Crank-Nicolson and Rannacher time-marching. *Journal of Computational Finance*, 9(4):89–112, 2006.

A. R. Gourlay and J. Ll. Morris. The extrapolation of first order methods for parabolic partial differential equations, ii. *SIAM Journal on Numerical Analysis*, 17(5):641–655, 1980.

Patrick S. Hagan and Graeme West. Interpolation methods for curve construction. *Applied Mathematical Finance*, 13(2):89–129, 2006.

Patrick S. Hagan, Deep Kumar, Andrew S. Lesniewski, and Diana E. Woodward. Managing smile risk. *WILMOTT Magazine*, September:84–108, 2002. URL `http://www.wilmott.com/pdfs/021118_smile.pdf`.

David Heath and Martin Schweizer. Martingales versus pdes in finance: An equivalence result with examples. *Journal of Applied Probability*, 37:947–957, 2000.

Steve Heston and Guofu Zhou. On the rate of convergence of discrete-time contingent claims. *Mathematical Finance*, 10(1):53–75, January 2000.

Steven L. Heston. A closed form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2), 1993.

R. Horváth. Finite difference solution of linear second order elliptic partial differential equations, February 2004. URL `http://www.cs.elte.hu/~faragois/phdcourse/phdcourse.html`.

K. J. Hout and B. D. Welfert. Stability of ADI schemes applied to convection-diffusion equations with mixed derivative terms, October 2006. URL `http://www.win.ua.ac.be/~kihout/adik.pdf`.

John Hull and Alan White. The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 42:281–300, 1987.

Hardy Hulley and Grant Lotter. Numerical Methods in Finance, 2004. University of the Witwatersrand.

Samuli Ikonen and Jari Toivanen. Componentwise splitting methods for pricing American options under stochastic volatility. Technical report, Department of Mathematical Information Technology, University of Jyvskyl, November 2005a. URL `http://www.mit.jyu.fi/tene/papers/index.html`.

Samuli Ikonen and Jari Toivanen. Efficient numerical methods for pricing American options under stochastic volatility. Technical report, Department of Mathematical Information Technology, University of Jyvskyl, December 2005b. URL `http://www.mit.jyu.fi/tene/papers/index.html`.

Samuli Ikonen and Jari Toivanen. Pricing American options using LU decomposition. Technical report, Department of Mathematical Information Technology, University of Jyvskyl, November 2005c. URL `http://www.mit.jyu.fi/tene/papers/index.html`.

Samuli Ikonen and Jari Toivanen. Operator splitting methods for American options with stochastic volatility. *Applied Mathematics Letters*, 7:809–814, 2004. URL `http://www.mit.jyu.fi/tene/papers/index.html`.

L. W. Johanson, R. D. Riess, and J. T. Arnold. *Introduction to Linear Algebra*. Addison Wesley, 5 edition, 2002.

Christian Kahl and Peter Jäckel. Fast strong approximation monte carlo schemes for stochastic volatility models. *Quantitative Finance*, 6(6):513–536, 2006.

A. Q. M. Khaliq and E. H. Twizell. L_0-stable splitting methods for the simple heat equation in two space dimensions with homogeneous boundary conditions. *SIAM Journal on Numerical Analysis*, 23 (3):473–484, 1986.

K. Kis̃kis and R. C̃iegis. On the stability of splitting difference schemes with respect to boundary conditions. *Lithuanian Mathematical Journal*, 37(4):364–373, 1997.

Tino Kluge. Pricing derivatives in stochastic volatility environment using the finite difference method. Technical report, Technische Universitat Chemintz, 2002.

Erwin Kreyszig. *Advanced Engineering Mathematics*. Wiley, eigth edition, 1999.

J. D. Lawson and J. Ll. Morris. The extrapolation of first order methods for parabolic partial differential equations, II. *SIAM Journal on Numerical Analysis*, 15:641–655, 1978.

Alan Lewis. *Option Valuation under Stochastic Volatility : with Mathematica Code*. Finance Press, 2000.

S. McKee, D.P. Wall, and S.K. Wilson. An Alternating Direction Implicit Scheme for Parabolic Equations with Mixed Derivative and Convective Terms. *Journal of Computational Physics*, 126:64–76, February 1996.

Cameron McLeman. Matrix exponential, 2006. URL planetmath.org_/?op=authorlist&from=objects&id=4162.

K. W. Morton and D. F. Mayers. *Numerical Solution of Partial Differential Equations*. Cambridge University Press, 1996.

D. M. Pooley, P. A. Forsyth, and K. R. Vetzal. Convergence remedies for non-smooth payoffs in option pricing. *Journal of Computational Finance*, 6(4):25–40, 2003.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2 edition, 1992.

Alexander Skabelin. Pricing of options on assets with level dependent stochastic volatility. In Derek Abbott, Jean-Philippe Bouchaud, Xavier Gabaix, and Joseph L. McCauley, editors, *Noise and Fluctuations in Econophysics and Finance*, volume 5848, pages 110–124. SPIE, 2005. URL http://www.fma.org/Chicago/Papers/AlexanderSkabelin_StochasticVolatility_v2.pdf.

G.D. Smith. *Numerical Solutions of Partial Differential Equations*. Oxford Applied Mathematics and Computing Science Series. Oxford University Press, third edition, 1985.

Domingo A Tavella and Curt Randall. *Pricing Financial Instruments: The Finite Difference Method*. John Wiley and Sons, 2000.

J.W. Thomas. *Numerical Partial Differential Equations*. Springer-Verlag, New York, 1995.

Vidar Thomee and Lars Wahlbin. Convergence rates of parabolic difference schemes for non-smooth data. *Mathematics of Computation*, 28(125):1–13, January 1974.

Axel Vogt. Optimized Excel solution for Heston's model using Gauss integration, 2004. URL `http://www.axelvogt.de/axalom/Heston93_opt.txt`.

Graeme West. Calibration of the SABR model in illiquid markets. *Applied Mathematical Finance*, 12(4): 371–385, 2005.

Paul Wilmott. *Paul Wilmott on Quantitative Finance*. John Wiley and Sons Ltd., 2000.

N. N. Yanenko. *The Method of Fractional Steps*. Springer-Verlag, Berlin, 1971.

R. Zvan, P. A. Forsyth, and K. R. Vetzal. Penalty methods for American options with stochastic volatility. *Journal of Computational and Applied Mathematics*, 91:199–218, 1998.

R. Zvan, P. A. Forsyth, and K. R. Vetzal. Negative coefficients in two-factor option pricing models. *Journal of Computational Finance*, 7(1):37–73, 2003.