AARHUS
UNIVERSITY
SCIENCE AND TECHNOLOGY

Numerical Solution of Parabolic Equations

Department of Computer Science

Ole Østerby

Monograph

ISBN 978-87-7507-313-9

DOI 10.7146/aul.5.5

Department of Computer Science
Aarhus University

April 30, 2014

Numerical Solution of Parabolic Equations

Ole Østerby

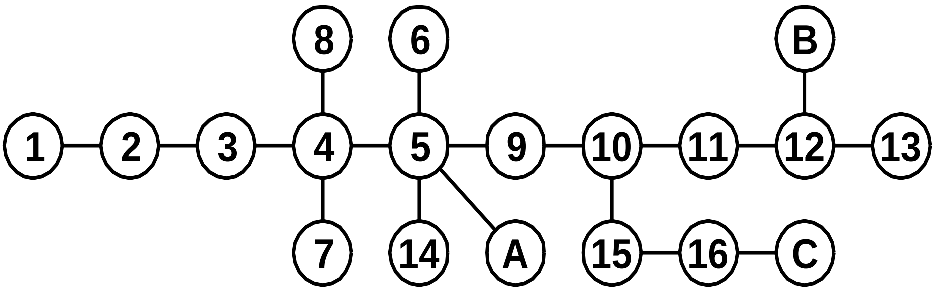
April 2014

Preface

These lecture notes are designed for a one semester course (10 ECTS). They deal with a rather specific topic: finite difference methods for parabolic partial differential equations as they occur in many areas such as Physics (diffusion, heat conduction), Geology (ground water flow), and Economics (finance theory).

The reader is supposed to have a basic knowledge of calculus (Taylor's formula), complex analysis ($\exp(it)$), linear algebra (eigenvalues and eigenvectors for matrices), and some familiarity with programming. A deeper knowledge in the theory of partial differential equations is, however, not required.

The text is divided into 16 chapters and 3 appendices, some long, some short, and some very short – which does not necessarily reflect the importance. If one is planning a shorter (5 ECTS) course one possibility is to omit Chapters 6, 7, and 8 and stop at (or after) Chapter 12. The interconnection between the various chapters and appendices is illustrated in the reading pattern:



Chapter 1 contains the basic introduction to parabolic equations (existence, uniqueness, well-posedness) and to the finite difference schemes which the book is all about. Two sections deal with the solution of (almost) tridiagonal linear systems of equations, and the Lax-Richtmyer equivalence theorem is briefly mentioned.

In Chapter 2 we discuss stability using the Fourier (or von Neumann) method, and in Chapter 3 we define the local truncation error as a means to discuss accuracy of the finite difference schemes. In Chapter 4 we look closely at boundary conditions which, if they contain derivatives, deserve special attention.

In Chapter 5 we study equations with a significant first order (convection) term. These equations are somewhat related to the one-way wave equation which is discussed in Appendix A. The next three Chapters can be omitted if time requires. Chapter 6 studies an alternate approach to stability analysis using matrix eigenvalues. Chapter 7 explains why we seldom use two-step methods, and Chapter 8 gives various suggestions on how to fight the adverse effects of discontinuities in the initial and/or boundary conditions.

A very important – and difficult – topic is to estimate the global error, i.e. the difference between the true and the computed solution. Chapter 9 provides one method to study this global error including the effect of boundary conditions (with a derivative).

The method of Chapter 9 gives some answers but is not ideal in practice. In Chapter 10 we introduce a practical way of estimating error which works in many cases – and issues warning signals when the results may not be trustworthy. Having a reliable error estimate enables us to choose a reasonable set of step sizes, small enough to meet a given error tolerance, but not excessively small such that we waste computer time. A reliable error estimate also opens possibilities for extrapolation thereby gaining extra accuracy at a minimal effort. This chapter is probably the most important one, containing material which is not readily found elsewhere.

In Chapter 11 we take up problems with two space variables. We want to take advantage of the good stability properties of implicit methods but not pay the price incurred by solving large systems of equations. The answer is ADI-methods where we solve for one space variable at a time using tridiagonal systems. ADI-methods cannot be used directly on equations with a mixed derivative term. We show in Chapter 12 how to modify our methods to take care of this without sacrificing stability or efficiency. Chapter 13 is devoted to two examples from Finance Theory involving two-factor models.

One should stay away from ill-posed problems. What may happen if we don't is illustrated in Chapter 14.

Chapter 15 treats the Stefan problem, an example of a moving boundary problem where the solution region is not known beforehand, but a boundary curve must be found together with the solution. For this particular problem we can avoid (some of the) interpolation problems by varying the time step size. Another example of a moving boundary problem is the American option of Chapter 16 where the exercise boundary is not known in advance but must be determined together with the solution.

In Appendix A we evaluate a number of difference schemes for the one-way wave equation which is related to the convection-diffusion equation of Chapter 5.

Apendix B introduces two classes of test problems with two space variables and in Appendix C we discuss some side effects of interpolation with importance for our methods for moving boundary problems.

Among the special features in this book we can mention

- An efficient way of estimating the order and accuracy of a method on a particular problem.
- The possibility of extrapolating to get higher order and better accuracy.
- An efficient way of determining (near-)optimal step sizes to meet a prescribed error tolerance (Chapter 10).
- A systematic way of incorporating inhomogeneous terms and boundary conditions in ADI-methods (Chapter 11).
- A systematic way of incorporating mixed derivative terms in ADI-methods (Chapter 12).

I should like to express my thanks to colleagues at the Danish Technical University and to several students at Aarhus University in Computer Science, Geology, and especially Finance Theory who have been exposed to more or less preliminary versions of these lecture notes. The questions that arose in various discussions have been the inspiration for many of the topics I have taken up and where answers are not readily found in common text books.

Contents

1 Basics 1

1.1 Differential equations 1

1.2 Main types of PDEs 2

1.3 Separation of variables 2

1.4 Side conditions 3

1.5 Well-posed and ill-posed problems 4

1.6 Two test problems 4

1.7 Difference operators 5

1.8 Difference schemes 7

1.8.1 The explicit method 7

1.8.2 The implicit method 8

1.8.3 Crank-Nicolson 8

1.8.4 The general θ-method 9

1.8.5 The operators P, $P_{k,h}$ and $R_{k,h}$ 9

1.9 Two-step schemes 10

1.10 Error norms 11

1.11 Tridiagonal systems of equations 11

1.12 Almost tridiagonal systems 13

1.13 Convergence 14

1.14 Exercises 16

2 Stability **19**

2.1 Fourier analysis 19

2.2 Two examples 21

2.3 Fourier analysis and differential equations 22

2.4 Von Neumann analysis 23

2.5 Implicit methods 27

2.6 Two kinds of stability 27

2.7 Exercise 28

3 Accuracy **29**

3.1 Taylor expansions 29

3.2 Order 30

3.3 Symbols of operators 32

3.4 The local error 34

4 Boundary Conditions **35**

4.1 A Dirichlet condition 35

4.2 Derivative boundary conditions 36

4.3 A third test problem 36

4.4 The explicit method 37

4.4.1 First order approximation 37

4.4.2 Asymmetric second order 37

4.4.3 Symmetric second order 38

4.5 The implicit method 38

4.5.1 First order approximation 38

4.5.2 Asymmetric second order 38

4.5.3 Symmetric second order 39

4.6 The θ-method 39

4.7 Exercises 39

5 The Convection-Diffusion Equation 41

5.1 Introduction 41

5.2 Maximum principle 42

5.3 The explicit method 42

5.4 The upwind scheme 44

5.5 The implicit method 47

5.6 Crank-Nicolson 48

5.7 Comparing the methods 48

5.8 Exercises 49

6 The Matrix Method 51

6.1 Notation 51

6.2 The explicit method 51

6.3 The implicit method 53

6.4 The θ-method 53

6.5 Stability by the matrix method 54

6.6 Eigenvalues of tridiagonal matrices 55

6.7 The influence of boundary values 58

6.8 A derivative boundary condition 59

6.9 Exercises 60

7 Two-step Methods 61

7.1 The central-time central-space scheme 61

7.2 The DuFort-Frankel scheme 62

7.3 Exercise 63

8 Discontinuities 65

8.1 Stability and damping 65
8.2 The growth factor 66
8.3 Reducing the Crank-Nicolson oscillations 67
8.3.1 AV – the moving average 67
8.3.2 IM1 – One step with IM 68
8.3.3 SM – Small steps at the beginning 68
8.3.4 Pearson 68
8.3.5 EI – Exponentially increasing steps 68
8.4 Discussion 69
8.5 A discontinuous corner 69
8.6 Exercises 70

9 The Global Error – Theoretical Aspects 71

9.1 The local error 71
9.2 The global error 71
9.3 The explicit method 72
9.4 The initial condition 73
9.5 Dirichlet boundary conditions 74
9.6 The error for the explicit method 74
9.7 The implicit method 75
9.7.1 An example 76
9.8 Crank-Nicolson 76
9.8.1 Example continued 78
9.9 Upwind schemes 78
9.10 Boundary conditions with a derivative 79
9.10.1 First order approximation 80
9.10.2 Asymmetric second order 80

9.10.3 Symmetric second order 81
9.10.4 Test problem 3 revisited 82
9.11 Exercise 84

10 Estimating the Global Error and Order 85

10.1 Introduction 85
10.2 The global error 85
10.3 Can we trust these results? 88
10.4 Further improvements of the error estimate 90
10.5 Two independent variables 92
10.6 Limitations of the technique. 94
10.7 Test problem 3 – once again 96
10.8 Which method to choose 99
10.9 Literature 100
10.10 Exercises 101

11 Two Space Dimensions 103

11.1 Introduction 103
11.2 The explicit method 104
11.3 Implicit methods 105
11.4 ADI methods 106
11.5 The Peaceman-Rachford method 107
11.6 Practical considerations 108
11.7 Stability of Peaceman-Rachford 110
11.8 D'Yakonov 111
11.9 Douglas-Rachford 111
11.10 Stability of Douglas-Rachford 113
11.11 The local truncation error 113

11.12 The global error 116

11.13 Exercises 120

12 Equations with Mixed Derivative Terms 121

12.1 Practical considerations 122

12.2 Stability with mixed derivative 123

12.3 Stability of ADI-methods 126

12.4 Summary 130

12.5 Exercises 131

13 Two-Factor Models – two examples 133

13.1 The Brennan-Schwartz model 133

13.2 Practicalities 136

13.3 A Traditional Douglas-Rachford step 139

13.4 The Peaceman-Rachford method 140

13.5 Fine points on efficiency 141

13.6 Convertible bonds 142

14 Ill-Posed Problems 143

14.1 Theory 143

14.2 Practice 144

14.3 Variable coefficients – An example 148

15 A Free Boundary Problem 149

15.1 The Stefan problem 149

15.2 The Douglas-Gallie method 150

15.3 The global error 153

15.4 Estimating the global error 157

16 The American Option 161

16.1 Introduction . . . 161

16.2 The mathematical model . . . 161

16.3 The boundary condition at infinity . . . 164

16.4 Finite difference schemes . . . 168

16.5 Varying the time steps . . . 171

16.6 The implicit method . . . 172

16.7 Crank-Nicolson . . . 177

16.8 Determining the order . . . 182

16.9 Efficiency of the methods . . . 185

A The One-Way Wave Equation 187

A.1 Forward-time forward-space . . . 187

A.2 Forward-time backward-space . . . 188

A.3 Forward-time central-space . . . 189

A.4 Central-time central-space or leap-frog . . . 189

A.5 Lax-Friedrichs . . . 190

A.6 Backward-time forward-space . . . 191

A.7 Backward-time backward-space . . . 191

A.8 Backward-time central-space . . . 192

A.9 Lax-Wendroff . . . 193

A.10 Crank-Nicolson . . . 193

A.11 Overview of the methods . . . 194

B A Class of Test Problems 197

C Interpolation and the Order Ratio 199

C.1 Introduction . . . 199

C.2 Linear interpolation 199

C.3 Three-point interpolation 201

Bibliography **203**

Chapter 1

Basics

1.1 Differential equations

An ordinary differential equation (ODE) is a relation between a function, y, of one independent variable, x, and its derivative, and possibly derivatives of higher order, e.g.

$$\frac{d^3y}{dx^3} + a\frac{d^2y}{dx^2} + b\frac{dy}{dx} \; = \; f(x).$$

A solution to the differential equation is a differentiable function $y(x)$ which satisfies this equation.

A partial differential equation (PDE) is the generalization to functions of two or more independent variables, e.g.

$$\frac{\partial u}{\partial t} - b\frac{\partial^2 u}{\partial x^2} + a\frac{\partial u}{\partial x} - \kappa u \; = \; \nu(t,x).$$

A solution is a differentiable function $u(t,x)$ which satisfies the equation.

Notation We shall in the following use subscripts to denote partial derivatives, e.g.

$$u_t \; = \; \frac{\partial u}{\partial t}, \;\; u_x \; = \; \frac{\partial u}{\partial x}, \;\; u_{xx} \; = \; \frac{\partial^2 u}{\partial x^2}, \qquad \text{etc.}$$

The partial differential equation above now reads

$$u_t - bu_{xx} + au_x - \kappa u \; = \; \nu(t,x)$$

The coefficients a, b, κ may be functions of t and x but in most of our theoretical investigations we shall treat them as constants.

1.2 Main types of PDEs

The most commonly occurring PDEs are divided into three groups:

type	simple example	
Elliptic	$u_{xx} + u_{yy} = 0$	Laplace's equation
Parabolic	$u_t - u_{xx} = 0$	Heat equation
Hyperbolic	$u_{tt} - u_{xx} = 0$	Wave equation

Parabolic and hyperbolic equations describe phenomena which evolve in time whereas elliptic equations describe steady state situations. In physics the elliptic equations model an electrostatic field, the parabolic equations model diffusion or heat conduction problems, and the hyperbolic equations model wave motion. In recent years parabolic equations have been used increasingly to model systems in mathematical economy and finance theory. The remainder of this book will be aimed exclusively at methods for solving parabolic equations.

1.3 Separation of variables

Some of the simpler PDEs such as the simple heat equation:

$$u_t = u_{xx} \tag{1.1}$$

can be solved using separation of variables, i.e. we assume the solution $u(t, x)$ can be written as a product of a function of t and a function of x:

$$u(t, x) = T(t)X(x). \tag{1.2}$$

Inserting in the differential equation we get

$$T'(t)X(x) = T(t)X''(x).$$

If $X(x) \equiv 0$ then $u(t, x) \equiv 0$ which is a (trivial) solution to (1.1). Since we are interested in non-trivial solutions we can assume that there is an x_1 such that $X(x_1) \neq 0$ and therefore $X(x) \neq 0$ in a neighbourhood around x_1. Similarly we can find a t_1 such that $T(t) \neq 0$ in a neighbourhood around t_1. In a neighbourhood around (t_1, x_1) we can therefore divide by $T(t)X(x)$ and get

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)},$$

but since the left-hand-side is only a function of t and the right-hand-side is only a function of x the result must be a constant which could be either positive or negative. Taking the latter case first and setting the constant equal to $-\omega^2$ where ω is a real number we end up with two ODEs, one for T and one for X. The general solutions are

$$T(t) \;=\; c_t \exp(-\omega^2 t), \qquad X(x) \;=\; c_x \cos(\omega x + \varphi).$$

Combining these we find that

$$u(t,x) \;=\; c \exp(-\omega^2 t) \cos(\omega x + \varphi) \tag{1.3}$$

is a general solution to the simple heat equation. As this equation is linear and homogeneous any linear combination of two or more solutions (e.g. with different values of ω) is also a solution. We need extra information to determine the values of c, ω, and φ. If for instance the initial value at $t = 0$ is a cosine or a linear combination of cosines then we get a solution by multiplying each cosine by the appropriate exponential factor. We shall return to this in sections 1.4 and 1.6.

If the constant above is positive we put it on the form $+\omega^2$ and we are in a similar manner led to a general solution of (1.1) in the form

$$u(t,x) \;=\; c \exp(\omega^2 t) \cosh(\omega x + \varphi) \tag{1.4}$$

Functions of type (1.4) grow without bound with increasing x and t. If we are interested in bounded solutions, functions of this type must have zero weight.

1.4 Side conditions

As illustrated by (1.3) solutions to differential equations are not unique unless we impose extra conditions. For parabolic equations it is customary to specify an *initial condition*, i.e. to specify the solution at some initial time, usually $t = 0$:

$$u(0,x) \;=\; u_0(x) \tag{1.5}$$

where $u_0(x)$ is a given function. If this is supposed to hold for $-\infty < x < \infty$ then we speak of an *initial value problem (IVP)*. More typical in practical situations, however, is to specify the initial value on a finite interval $X_1 \le x \le X_2$ in which case we must also specify *boundary conditions* for $x = X_1$ and $x = X_2$. On the left boundary such a boundary condition may be of the general form

$$\alpha_1 u(t, X_1) - \beta_1 u_x(t, X_1) \;=\; \gamma_1, \qquad\qquad t > 0 \tag{1.6}$$

where α_1, β_1, and γ_1 may depend on t. If $\beta_1 = 0$ we speak of a *Dirichlet* condition (J.P.G.L. Dirichlet, 1805 – 1859). If $\alpha_1 = 0$ we speak of a *Neumann* condition (Carl Neumann, 1832 – 1925). If both α_1 and β_1 are different from 0 we speak of a boundary condition of the third kind or a *Robin* condition (Gustave Robin, 1855 – 1897). On the right boundary we have a similar condition:

$$\alpha_2 u(t, X_2) + \beta_2 u_x(t, X_2) \quad = \quad \gamma_2, \qquad\qquad t > 0. \qquad (1.7)$$

In this case we speak of an *initial-boundary value problem (IBVP).*

The choice of signs in (1.6) and (1.7) may seem a bit strange at first sight. If β_1 and β_2 are non-zero then we can assume without loss of generality that they are equal to 1. In this case positive values for α_1 and α_2 will ensure that the solutions of the differential equation are not exponentially growing. We refer the reader to [7], [18], and [25] for a further discussion of the effect of derivative boundary conditions on the qualitative nature of the solutions, but here it shall be our general assumption that α_1, α_2, β_1, and β_2 are all non-negative.

1.5 Well-posed and ill-posed problems

If in an IVP we change the initial function by a small amount we hope and expect that the effect on the solution function at some later time will also be small. This is indeed the case for the simple heat equation and is related to the fact that a *maximum principle* (cf. section 5.2) exists for this equation. We say that the problem is *well-posed* and we shall discuss this property in more detail in Chapter 2. If on the other hand we would attempt to solve the heat equation in the opposite time direction – or what amounts to the same – would consider the equation $u_t = -u_{xx}$, then arbitrarily small changes in the initial condition (e.g. with large values of ω) would imply large deviations in the solution at later times. We say that this problem is *ill-posed.* Such problems are not well suited as mathematical models and attempts to solve them numerically are doomed to disaster. The reason for this is that when we attempt to solve a differential equation numerically we invariably introduce small (rounding) errors, typically with large frequencies (ω) already in the first time step. In an ill-posed problem these perturbations will be amplified in the following time steps thereby distorting the solution function beyond recognition. We shall discuss this further in Chapters 2 and 14.

1.6 Two test problems

The following two test problems will be used extensively in examples and exercises to demonstrate the behaviour of various techniques and methods.

Problem 1

$$\begin{aligned} u_t &= u_{xx}, & -1 \le x \le 1, \quad & t > 0, \\ u(0,x) = u_0(x) &= \cos x, & -1 \le x \le 1, \quad & \\ u(t,-1) = u(t,1) &= e^{-t}\cos 1, & & t > 0. \end{aligned}$$

This is an IBVP with boundary conditions of Dirichlet type. It is easily seen from formula (1.3) that the true solution is $u(t,x) = e^{-t}\cos x$.

Problem 2

$$u_t = u_{xx}, \qquad -1 \le x \le 1, \qquad t > 0,$$

$$u(0,x) = u_0(x) = \begin{cases} 1, & |x| < \frac{1}{2}, \\ 0, & |x| = \frac{1}{2}, \\ -1, & |x| > \frac{1}{2}. \end{cases}$$

We now have a discontinuous initial condition. The solution will, however, be continuous and infinitely often differentiable for $t > 0$. The true solution can be found by taking the Fourier cosine series for the initial function and appending the corresponding exponential factors as given by formula (1.3):

$$u(t,x) = \frac{4}{\pi}\sum_{j=0}^{\infty}(-1)^j \frac{\cos((2j+1)\pi x)}{2j+1} e^{-(2j+1)^2\pi^2 t}. \tag{1.8}$$

If we take the boundary values at $x = -1$ og $x = 1$ from this series we obtain a Dirichlet problem.

Note that we only need rather few terms in the infinite sum in order to achieve any specific finite accuracy when $t > 0$ (cf. Exercise 2).

1.7 Difference operators

In most practically occurring cases an analytical solution to an IBVP for a parabolic equation cannot be obtained and we must resort to numerical techniques.

We choose a step size in the t-direction, k, and a step size in the x-direction, h, usually chosen as $h = (X_2 - X_1)/M$ for some integer M, and we wish to find approximations

$$v_m^n = v(nk, X_1 + mh)$$

to the true solution $u(t,x)$ at all grid points (cf. Fig. 1.1)

$$(t,x) = (nk, X_1 + mh), \quad m = 0,1,\dots M, \quad n = 1,2,\dots,N$$

where $T = Nk$ is the maximum time. We do this by approximating the partial derivatives with difference quotients.

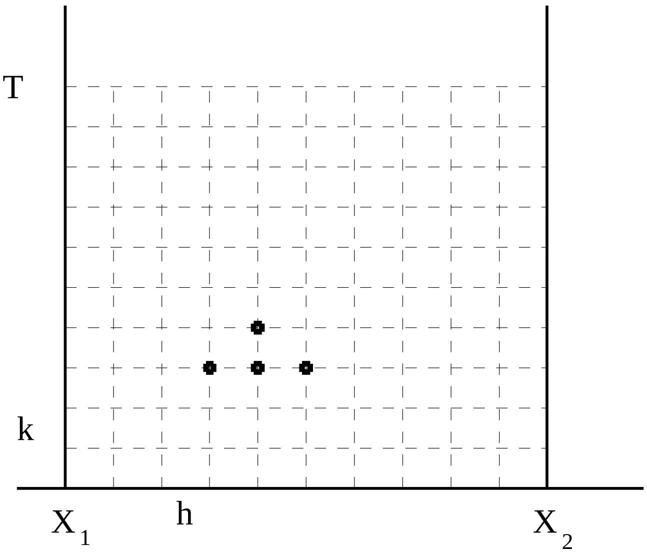


Figure 1.1: The (t,x) grid.

We first introduce the *shift* operator in the x-direction, E:

$$Ev_m^n = v_{m+1}^n$$

and the *mean value* operator, $\tilde{\mu}$:

$$\tilde{\mu} = (E^{1/2} + E^{-1/2})/2.$$

Remark. In the literature the mean value operator usually appears without the tilde. We have introduced the tilde here to avoid confusion with another μ to be introduced in the next section. □

We now have three different approximations to D, the operator which denotes the partial derivative w.r.t. x:

Forward difference:	Δ	$=$	$(E-1)/h,$
Backward difference:	∇	$=$	$(1-E^{-1})/h,$
Central difference:	δ	$=$	$(E^{1/2}-E^{-1/2})/h.$

δ and $\tilde{\mu}$ refer to half-way points and are most useful in combinations such as:

$$\begin{aligned}\tilde{\mu}\delta &= (E - E^{-1})/(2h), \\ \delta^2 &= (E - 2 + E^{-1})/h^2.\end{aligned}$$

The former is another approximation to D, the latter is an approximation to D^2, the second partial derivative w.r.t. x.

To approximate $D_t = \frac{\partial}{\partial t}$ we have a choice between

$$\begin{aligned}\Delta_t &= (E_t - 1)/k, \\ \nabla_t &= (1 - E_t^{-1})/k,\end{aligned}$$

and

$$\tilde{\mu}_t\delta_t = (E_t - E_t^{-1})/(2k)$$

where the shift operator in the t-direction is

$$E_t v_m^n = v_m^{n+1}.$$

1.8 Difference schemes

1.8.1 The explicit method

To approximate the heat equation

$$u_t - bu_{xx} = 0 \tag{1.9}$$

we can thus suggest

$$\Delta_t v_m^n - b\,\delta^2 v_m^n = 0 \tag{1.10}$$

or written out

$$\frac{v_m^{n+1} - v_m^n}{k} = b\,\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}. \tag{1.11}$$

Introducing the step ratio

$$\mu = \frac{k}{h^2} \tag{1.12}$$

this equation can be rewritten as

$$v_m^{n+1} = v_m^n + b\mu(v_{m+1}^n - 2v_m^n + v_{m-1}^n) \tag{1.13}$$

and thus provides a value for v_m^{n+1} explicitly from values at the previous time level. The method is therefore called *the explicit method.* In Fig. 1.1 we have marked the *stencil* for the explicit method, i.e. the geometric pattern describing the points which enter into the basic formula (1.11). If values for the solution function are given at the initial time level, v_m^0, $m = 0, 1, \ldots, M$, then (1.13) can be used to provide values for v_m^1 for $m = 1, 2, \ldots, M-1$, i.e. for all internal points at time level 1. If Dirichlet boundary values are given for $x = X_1$ and $x = X_2$ then we have values for v_m^1 also for $m = 0$ and $m = M$. We now have a complete set of values at time level 1 and can proceed from here to time level 2, 3, etc.

1.8.2 The implicit method

Another numerical formula for (1.9) is

$$\nabla_t v_m^{n+1} - b\,\delta^2 v_m^{n+1} \;=\; 0 \tag{1.14}$$

or written out

$$\frac{v_m^{n+1} - v_m^n}{k} - b\,\frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2} \;=\; 0 \tag{1.15}$$

or

$$v_m^{n+1} - b\mu(v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}) \;=\; v_m^n. \tag{1.16}$$

This formula expresses an implicit relation between three neighbouring function values at the advanced time level and is therefore generally known as *the implicit method* [19] although strictly speaking it is not a method until we have specified a technique for solving the resulting tridiagonal set of linear equations. We shall do this in the next section.

1.8.3 Crank-Nicolson

A third important formula is *Crank-Nicolson* [8]:

$$\Delta_t v_m^n - \frac{1}{2}b(\delta^2 v_m^n + \delta^2 v_m^{n+1}) \;=\; 0. \tag{1.17}$$

Remark. Another way of writing Crank-Nicolson which better expresses the symmetric nature of the formula is

$$(\delta_t - b\tilde{\mu}_t\delta^2)v_m^{n+\frac{1}{2}} \;=\; 0. \qquad \square$$

1.8.4 The general θ-method

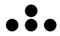
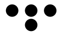
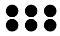
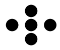
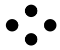
The three formulae above are special cases of the general θ-*method*:

$$\Delta_t v_m^n - b((1-\theta)\delta^2 v_m^n + \theta\delta^2 v_m^{n+1}) \quad = \quad 0 \tag{1.18}$$

corresponding to $\theta = 0$, 1, and $\frac{1}{2}$, respectively. The linear equations in the general case look like

$$\begin{aligned} v_m^{n+1} - \theta b\mu(v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}) \quad &= \\ & v_m^n + (1-\theta)b\mu(v_{m+1}^n - 2v_m^n + v_{m-1}^n). \end{aligned} \tag{1.19}$$

Table 1.1: Stencils for the methods

Method	Stencil
Explicit	
Implicit	
Crank-Nicolson	
Richardson	
DuFort Frankel	

In Table 1.1 we have given the stencils for the methods we have just defined together with two more which we shall discuss in Chapter 7.

1.8.5 The operators P, $P_{k,h}$ and $R_{k,h}$

For the general parabolic equation

$$Pu \equiv u_t - bu_{xx} + au_x - \kappa u = \nu$$

the θ-method approximates the differential operator P by the difference operator $P_{k,h}$:

$$P_{k,h}v_m^n = \Delta_t v_m^n - ((1-\theta)I + \theta E_t)(b\delta^2 - a\tilde{\mu}\delta + \kappa)v_m^n$$

where I denotes the identity, and for the right-hand-side we suggest

$$R_{k,h}\nu_m^n = ((1-\theta)I + \theta E_t)\nu_m^n.$$

So the differential equation

$$Pu \quad = \quad \nu \tag{1.20}$$

is approximated by the difference scheme

$$P_{k,h}v \;=\; R_{k,h}\nu. \tag{1.21}$$

We mention in passing that there are other options for approximating the u_x-term. We shall return to these in Chapter 5.

Since the operator $R_{k,h}$ is an approximation to the identity it has a well-defined inverse $R_{k,h}^{-1}$ and if we apply this to (1.21) we get

$$R_{k,h}^{-1}P_{k,h}v \;=\; \nu. \tag{1.22}$$

Comparing (1.20) and (1.22) we see that

$$R_{k,h}^{-1}P_{k,h} \;\approx\; P$$

or, since we really don't want to work with the inverse:

$$P_{k,h} - R_{k,h}P \;\approx\; 0.$$

We shall return to this operator in section 1.13 and Chapter 3.

1.9 Two-step schemes

All the above-mentioned difference schemes are examples of *one-step* schemes in the sense that they span one time step. They take data from one time level (n) in order to compute values for the succeeding time level ($n+1$). As an example of a *two-step* scheme for (1.9) we mention

$$\tilde{\mu}_t\delta_t v_m^n \;=\; b\delta^2 v_m^n$$

or

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} \;=\; b\,\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}. \tag{1.23}$$

This scheme spans two time steps taking data from both time level n and $n-1$ in order to compute values at time level $n+1$. We shall return to this scheme in Chapter 7 and here just mention that two-step methods require a special starting procedure since initial values are usually only specified at one time level ($n=0$).

1.10 Error norms

The difference between the true solution $u(t,x)$ and the numerical solution v_m^n $(t = nk, x = X_1 + mh)$ is the error. It is only defined at the points where we have a numerical solution, i.e. at the grid points as defined by the step sizes k and h, although we shall find ways to extend the error function as a differentiable function between the grid points in Chapter 9.

To study the behaviour of the error as a function of the step sizes k and h (and the time t) we shall use various norms. It may be important for the user that the error at any specific (grid-)point does not exceed a given tolerance. The user will therefore be interested in the *max-norm* (or sup-norm or ∞-norm) at time $t = nk$:

$$e_\infty(t) \;=\; ||u^n - v^n||_\infty \;=\; \max_{0\le m\le M} |u(nk, X_1 + mh) - v_m^n|. \qquad (1.24)$$

The max-norm is rather difficult to analyze mathematically and we shall therefore often study the *2-norm*:

$$e_2(t) \;=\; ||u^n - v^n||_2 \;=\; \left[h \sum_{m=0}^{M} (u(nk, X_1 + mh) - v_m^n)^2\right]^{\frac{1}{2}}. \qquad (1.25)$$

Remark. Notice the difference from the usual vector 2-norm in that we in (1.25) have a factor $h = (X_2 - X_1)/M$ in front of the summation. This factor originates from the Fourier transform (cf. Chapter 2) but comes in handy because we shall wish to compare the errors corresponding to different values of h, and therefore the norms from vector spaces of different dimensions. □

Remark. The max-norm and the 2-norm of the error will usually exhibit the same behaviour when the solution is a smooth function in the closed region $\{0 \le t \le T, X_1 \le x \le X_2\}$ but they will often differ considerably when there is a discontinuity in the initial function. □

Remark. When we have Dirichlet boundary conditions the error for $m = 0$ and $m = M$ are 0, and the max and the summation only apply to the internal grid points. □

1.11 Tridiagonal systems of equations

When $\theta > 0$ the formula (1.18) is implicit. For each internal point (i.e. $m = 1$, 2, $\ldots$,$M - 1$) we have a linear expression involving the unknowns v_{m-1}^{n+1}, v_m^{n+1}, and v_{m+1}^{n+1}. We have thus $M - 1$ equations in the $M + 1$ unknowns v_m^{n+1}, $m = 0$, 1,

…, M. If Dirichlet boundary values are specified then v_0^{n+1} and v_M^{n+1} are given and we are left with $M-1$ equations in $M-1$ unknowns (or $M+1$ equations where the first and the last are trivial).

Systems of linear equations are often solved using Gaussian elimination. For a general set of $M-1$ equations in $M-1$ unknowns the computational cost is about $\frac{1}{3}M^3$ additions and multiplications, but the coefficient matrix of systems resulting from using the θ-method is tridiagonal, i.e. the only non-zero coefficients appear in the diagonal and the immediate neighbours above and below, and this implies a considerable reduction in computing time.

The general equation is written

$$a_m v_{m-1}^{n+1} + b_m v_m^{n+1} + c_m v_{m+1}^{n+1} \;=\; d_m, \qquad m = 1, 2, \ldots, M-1. \tag{1.26}$$

Example. For the implicit method on $u_t = bu_{xx}$ we have

$$a_m \;=\; c_m \;=\; -b\mu, \quad b_m \;=\; 1+2b\mu, \quad d_m \;=\; v_m^n. \qquad \square$$

In the first equation (for $m=1$) the value of v_0^{n+1} on the left-hand-side is known from the Dirichlet boundary condition. We can therefore move the corresponding term to the right-hand-side:

$$d_1' \;=\; d_1 - a_1 v_0^{n+1}$$

and similarly for v_M^{n+1} in the last equation.

The system of equations now looks like

$$\left\{\begin{array}{ccccc} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \cdot & \cdot & \cdot & \\ & & a_{M-2} & b_{M-2} & c_{M-2} \\ & & & a_{M-1} & b_{M-1} \end{array}\right\} \left\{\begin{array}{c} v_1^{n+1} \\ v_2^{n+1} \\ \cdot \\ v_{M-2}^{n+1} \\ v_{M-1}^{n+1} \end{array}\right\} = \left\{\begin{array}{c} d_1' \\ d_2 \\ \cdot \\ d_{M-2} \\ d_{M-1}' \end{array}\right\} \tag{1.27}$$

Using Gaussian elimination we zero out the a_m thereby modifying the b_m and the d_m:

$$z = a_m / b_{m-1}'; \quad b_m' = b_m - z c_{m-1}; \quad d_m' = d_m - z d_{m-1}'. \tag{1.28}$$

Starting with $b_1' = b_1$ and executing (1.28) for $m = 2, 3, \ldots, M-1$ we end up with a triangular set of equations which can be solved from the bottom up (the back substitution):

$$v_{M-1}^{n+1} = d_{M-1}' / b_{M-1}', \quad v_m^{n+1} = (d_m' - c_m v_{m+1}^{n+1}) / b_m', \quad m = M-2, \ldots, 1. \tag{1.29}$$

The process will break down if any of the calculated b'_m become equal to 0 and will be numerically unstable if they come close to 0, but this cannot happen for the systems we consider here.

Example. For the implicit method on $u_t = bu_{xx}$ we have

$$b'_2 = b_2 - \frac{a_2}{b_1}c_1 = 1 + 2b\mu - \frac{(b\mu)^2}{1+2b\mu} \geq 1 + 2b\mu - \frac{(b\mu)^2}{2b\mu} = 1 + \frac{3}{2}b\mu \geq 1 + b\mu$$

$$b'_3 = b_3 - \frac{a_3}{b'_2}c_2 = 1 + 2b\mu - \frac{(b\mu)^2}{b'_2} \geq 1 + 2b\mu - \frac{(b\mu)^2}{1+b\mu} \geq 1 + b\mu$$

By induction we can show that $b'_m \geq 1 + b\mu$ for $m > 3$.

The result for the general θ-method is the topic of Exercise 13. □

Remark. The process will break down on $u_t = bu_{xx} + \kappa u$ if $\kappa k = 1 + 2b\mu$, so we must be careful when $\kappa > 0$. □

Altogether the computational cost of solving the system is roughly
$3M$ additions, $3M$ multiplications, and $2M$ divisions
for a total of $8M$ simple arithmetic operations (SAO).

The computational cost of advancing the solution one time step with the implicit formula is therefore linear in the number of unknowns and actually comparable to the cost of using the explicit method which is roughly $5M$ SAO for a trivial set of equations with a complicated right-hand-side.

For the general θ-method, $0 < \theta < 1$, (and in particular $\theta = 1/2$) we have both a system to solve and a complicated right-hand-side so the cost amounts to $13M$ SAO per time step. We shall see later that Crank-Nicolson is well worth this extra cost.

The main observation is that the computational cost for all these schemes, whether explicit or implicit, is linear in the number of grid points

1.12 Almost tridiagonal systems

In some cases we encounter systems of equations which are not completely tridiagonal but still easy to solve. This happens for example in connection with derivative boundary conditions to be discussed in Chapter 4.

A typical system may look like

$$\left\{ \begin{array}{cccccc} b_0 & c_0 & e_0 & & & \\ a_1 & b_1 & c_1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & a_{M-1} & b_{M-1} & c_{M-1} \\ & & f_M & a_M & b_M \end{array} \right\} \left\{ \begin{array}{c} v_0^{n+1} \\ v_1^{n+1} \\ \cdot \\ v_{M-1}^{n+1} \\ v_M^{n+1} \end{array} \right\} = \left\{ \begin{array}{c} d_0 \\ d_1 \\ \cdot \\ d_{M-1} \\ d_M \end{array} \right\} \tag{1.30}$$

Using Gaussian elimination we notice that when zeroing out a_1 we must also change c_1 using

$$c_1' = c_1 - ze_0 \quad \text{where} \quad z = a_1/b_0.$$

Likewise before zeroing out a_M we must eliminate f_M using equation $M-2$ which causes a change in a_M (and d_M):

$$a_M' = a_M - zc_{M-2} \quad \text{where} \quad z = f_M/b_{M-2}'.$$

The back substitution is performed as in the tridiagonal case except for the last step (equation 0) where an extra term involving e_0 appears:

$$v_0^{n+1} = (d_0 - c_0 v_1^{n+1} - e_0 v_2^{n+1})/b_0.$$

These extra calculations will complicate the programming slightly; but they do not affect the overall computational cost which is still $8M$ SAO for the implicit method and $13M$ for the general θ-method.

1.13 Convergence

The calculated values v_m^n are supposed to be approximations to the true solution function $u(t, x)$ at $t = nk$ and $x = X_1 + mh$. As we make the step sizes k and h smaller we should like these approximations to become better and better. But this will not always be the case.

Using Taylor series it is fairly easy to verify that our various difference expressions become better and better approximations to the partial derivatives as the step sizes become smaller. But at each time level we base our computations on previously computed values with inherent errors. And as h and k become smaller we must take more steps to get to a specific point (t, x).

And it is not obvious – and indeed not always the case – that the net effect is a better approximation. But for a numerical method to be useful this must be the case. This property is captured in the definition of *convergence*:

Definition. A difference scheme is called *convergent* if when applied to a well-posed IBVP it produces approximations $v_{h,k}(t,x)$ which converge to the true solution $u(t,x)$ for any point (t,x) in the region when $h \to 0, k \to 0$. □

We shall of course assume that the side conditions are applied correctly. It is not required that $v_{h,k}(0,x)$ coincides exactly with $u(0,x)$ for all h and k but only that $\lim v_{h,k}(0,x) = u(0,x)$ where the limit is for $h \to 0, k \to 0$ (and similarly for boundary conditions).

It is often a difficult task to prove that a numerical scheme converges for any well-posed IBVP. Luckily it is not necessary either. An important theorem by Peter Lax [21], [31] breaks the task into two much easier ones:

- to show that the numerical scheme is *consistent* with the differential equation and
- to show that the numerical scheme is *stable* i.e. that errors are not amplified (too much).

Definition. A difference scheme $P_{k,h}v = R_{k,h}\nu$ is *consistent* with a differential equation $Pu = \nu$ iff

$$P_{k,h}\psi - R_{k,h}P\psi \to 0 \quad \text{as} \quad h \to 0,\ k \to 0$$

for all smooth functions ψ. □

Since the solution to a parabolic equation is infinitely often differentiable for $t > 0$ it is not unreasonable to invoke smooth functions here. Since we prefer our results to have wide applicability we shall require the above convergence to 0 to happen for any smooth function ψ and not just for particular solution functions $u(t,x)$ where cancellations may occur and secure a convergence which is not generally available. We shall return to discuss the concept of consistency in Chapter 3.

Definition. A one-step difference scheme $P_{k,h}v = 0$ is *stable* iff

$$\forall T,\ \exists C_T \quad ||v^n||_2 \ \leq \ C_T||v^0||_2, \qquad\qquad nk \leq T \tag{1.31}$$

for $h \leq h_0$ and $k \leq k_0$. □

We note that the inhomogeneous term does not play any role for the question of stability, and neither does the differential operator. Since the difference between two solutions to the difference scheme also satisfies the homogeneous scheme, the stability condition implies that errors do not grow without bound.

Remark If (1.31) is satisfied for *all* $h \leq h_0$ and $k \leq k_0$ then we talk of *unconditional stability.* In many cases (1.31) is only satisfied in a subset, e.g. characterized by $k \leq h^2/2$. For a method to be useful there must be a path in this subset leading all the way from (h_0, k_0) to $(0,0)$. In such a case we talk of *conditional stability.*

1.14 Exercises

1. Solve problem 1 on page 5 with the explicit method (1.13) from $t = 0$ to $t = 0.5$ with $h = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$, and with $\mu(= k/h^2) = 0.5$.
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$.

2. How many terms are needed in the sum (1.8) of problem 2 to make the remainder less than 10^{-6} when $t = \frac{1}{200}$?
   And when $t = \frac{1}{3200}$?

3. Solve problem 2 with the explicit method from $t = 0$ to $t = 0.5$ with $h = \frac{1}{10}$ and $k = \frac{1}{200}$.
   Draw the true solution and the numerical solution as functions of x for $t = 0.005,\ 0.01,\ 0.015,\ 0.02,\ 0.025$.
   Draw the error as a function of x for $t = 0.005,\ 0.01,\ \ldots,\ 0.025$.
   Draw the error as a function of t for $x = 0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$.

4. Solve problem 2 with the explicit method from $t = 0$ to $t = 0.5$ with
   $h = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$, and with $\mu(= k/h^2) = 0.5$.
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$.

5. Solve problem 1 with the implicit method (1.15) from $t = 0$ to $t = 0.5$ with
   $h = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$, and with $\mu(= k/h^2) = 0.5$.
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$.

6. Solve problem 2 with the implicit method from $t = 0$ to $t = 0.5$ with $h = \frac{1}{10}$ and $k = \frac{1}{200}$.
   Draw the error as a function of x for $t = 0.005,\ 0.01,\ \ldots,\ 0.025$.
   Draw the error as a function of t for $x = 0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$.

7. Solve problem 2 with the implicit method from $t = 0$ to $t = 0.5$ with
   $h = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$, and with $\mu(= k/h^2) = 0.5$.
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$.

8. Solve problem 1 with Crank-Nicolson (1.17) from $t = 0$ to $t = 0.5$ with
   $h = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$, and with $\mu(= k/h^2) = 0.5$.

Compute the max-norm and the 2-norm of the error for
$t = 0.1, 0.2, 0.3, 0.4, 0.5$.

9. Solve problem 2 with Crank-Nicolson from $t = 0$ to $t = 0.5$ with $h = \frac{1}{10}$ and $k = \frac{1}{200}$.
Draw the error as a function of x for $t = 0.005, 0.01, \ldots, 0.025$.
Draw the error as a function of t for $x = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

10. Solve problem 2 with Crank-Nicolson from $t = 0$ to $t = 0.5$ with
$h = \frac{1}{10}, \frac{1}{20}$, and $\frac{1}{40}$, and with $\mu(= k/h^2) = 0.5$.
Compute the max-norm and the 2-norm of the error for
$t = 0.1, 0.2, 0.3, 0.4, 0.5$.

11. Solve problem 1 with the explicit method from $t = 0$ to $t = 0.5$ with
$h = k = \frac{1}{10}$.
Draw the error as a function of x for $t = 0.1, 0.2, 0.3, 0.4, 0.5$.
Draw the error as a function of t for $x = 0, 0.1, 0.2, 0.3, 0.4, 0.5$.

12. Solve problems 1 and 2 with Crank-Nicolson and the implicit method from
$t = 0$ to $t = 0.5$ with $h = k = \frac{1}{10}, \frac{1}{20}$, and $\frac{1}{40}$.
Compute the max-norm and the 2-norm of the error for
$t = 0.1, 0.2, 0.3, 0.4, 0.5$.

13. Show that $b'_m \geq 1$ in the tridiagonal system of equations (cf. section 1.11) which arises when we use the general θ-method (1.18) on $u_t = bu_{xx}$.

Chapter 2

Stability

2.1 Fourier analysis

A very important tool in the study of stability of difference schemes as well as the behaviour of differential equations is Fourier analysis. We begin with the continuous case. If $u(x)$ is a real function defined on the real line then the *Fourier transform* of u is

$$\hat{u}(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega x}\, u(x)\, dx. \tag{2.1}$$

It is possible to recreate $u(x)$ from $\hat{u}(\omega)$ by the inversion formula

$$u(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega x}\, \hat{u}(\omega)\, d\omega. \tag{2.2}$$

$\hat{u}(\omega)$ is a function of a real variable ω but it may assume complex values. From the inversion formula we may deduce that $\hat{u}(\omega)$ is uniquely determined by $u(x)$ and vice versa. $\hat{u}(\omega)$ is just an alternate representation of $u(x)$ just like a Fourier series is an alternate representation of a periodic function.

If v is a grid function, i.e. v_m is defined for all integers m then we define the discrete transform

$$\hat{v}(\xi) \;=\; \frac{1}{\sqrt{2\pi}} \sum_{-\infty}^{\infty} e^{-im\xi}\, v_m$$

where $\xi \in [-\pi, \pi]$. The inversion formula reads

$$v_m \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{im\xi}\, \hat{v}(\xi)\, d\xi.$$

The more useful case is where the distance between grid points is h. We change variable and define

$$\hat{v}(\xi) \;=\; \frac{1}{\sqrt{2\pi}} \sum_{-\infty}^{\infty} e^{-imh\xi}\, v_m\, h \tag{2.3}$$

where $h\xi \in [-\pi, \pi]$. The inversion formula now reads

$$v_m \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi}\, \hat{v}(\xi)\, d\xi. \tag{2.4}$$

The L^2-norm of $u(x)$ is defined by

$$||u||_2 \;=\; \left(\int_{-\infty}^{\infty} |u(x)|^2\, dx\right)^{1/2} \tag{2.5}$$

and is equal to the L^2-norm of $\hat{u}$:

$$||\hat{u}||_2 \;=\; \left(\int_{-\infty}^{\infty} |\hat{u}(\omega)|^2\, d\omega\right)^{1/2}. \tag{2.6}$$

This relation which is named after *Parseval* (Marc-Antoine Parseval des Chênes, 1755 – 1836) is proved by the following calculations:

$$\begin{aligned}
||u||_2^2 \;&=\; \int_{-\infty}^{\infty} u(x)\, \overline{u(x)}\, dx \;=\; \int_{-\infty}^{\infty} u(x) \overline{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega x}\, \hat{u}(\omega) d\omega\, dx} \\
&=\; \int_{-\infty}^{\infty} u(x) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega x}\, \overline{\hat{u}(\omega)}\, d\omega\, dx \\
&=\; \int_{-\infty}^{\infty} \overline{\hat{u}(\omega)} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega x}\, u(x)\, dx\, d\omega \\
&=\; \int_{-\infty}^{\infty} \overline{\hat{u}(\omega)}\, \hat{u}(\omega)\, d\omega \;=\; ||\hat{u}||_2^2.
\end{aligned}$$

The crucial point of this derivation is the interchange of the order of integration. This is allowed if and only if u (and $\hat{u}$) are L^2-functions, and this is also the condition for the Fourier transform to be well-defined.

The 2-norm of v_m was defined in section 1.10:

$$||v||_2 \;=\; \left(h \sum_{-\infty}^{\infty} |v_m|^2\right)^{1/2} \tag{2.7}$$

and also here we have a Parseval relation which states that

$$||v||_2 \;=\; ||\hat{v}||_2. \tag{2.8}$$

Remark. Comparing (2.7) with the usual definition of the 2-norm we note an extra factor h. This is introduced because we shall wish to compare grid functions with different values of h with each other and with the continuous function which they are supposed to be approximations to. □

2.2 Two examples

Example 1: The square. We should like to compute the Fourier transform of the grid function defined by

$$v_m = \begin{cases} 1 & \text{if } |m| < M \\ \frac{1}{2} & \text{if } |m| = M \\ 0 & \text{if } |m| > M \end{cases} \qquad \text{where} \quad Mh = 1$$

For $\xi \neq 0$ we have

$$\begin{aligned} \hat{v}(\xi) &= \frac{1}{\sqrt{2\pi}} \sum_{-\infty}^{\infty} e^{-imh\xi} v_m h = \frac{h}{\sqrt{2\pi}} \left\{ \frac{1}{2}(e^{-i\xi} + e^{i\xi}) + \sum_{-M+1}^{M-1} e^{-imh\xi} \right\} \\ &= \frac{h}{\sqrt{2\pi}} \left\{ \cos(\xi) + e^{i(M-1)h\xi} \frac{1 - e^{-i(2M-1)h\xi}}{1 - e^{-ih\xi}} \right\} \\ &= \frac{h}{\sqrt{2\pi}} \left\{ \cos(\xi) + \frac{\sin(\xi - \frac{1}{2}h\xi)}{\sin(\frac{1}{2}h\xi)} \right\} = \frac{h}{\sqrt{2\pi}} \sin(\xi) \cot(\frac{1}{2}h\xi). \end{aligned}$$

For $\xi = 0$ we have

$$\hat{v}(\xi) = \frac{h}{\sqrt{2\pi}}(1 + 2M - 1) = \frac{2}{\sqrt{2\pi}}$$

which is also the limit of the first expression as $\xi \to 0$.

In Fig. 2.1 we have to the left shown the Fourier transform of the square for $M = 11$. The abscissa is $h\xi$ and we have only shown the interval $0 \leq h\xi \leq \pi$ as the function is symmetric around 0. We notice a substantial weight near 0. □

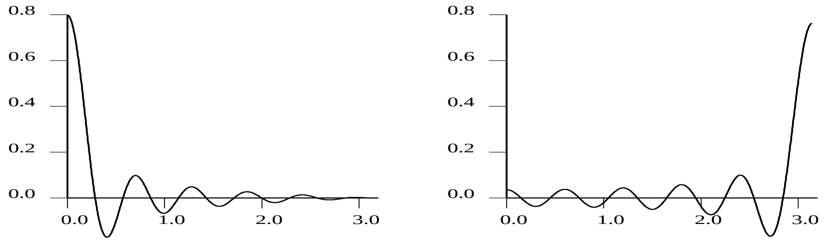


Figure 2.1: Fourier transform of square (left) and oscillation (right).

Example 2. The oscillation. We next consider an oscillating grid function

$$v_m = \begin{cases} (-1)^m & \text{if } |m| < M \\ 0 & \text{if } |m| \geq M \end{cases} \qquad \text{where} \quad Mh = 1$$

We now have

$$\begin{aligned}
\hat{v}(\xi) &= \frac{h}{\sqrt{2\pi}} \sum_{-M+1}^{M-1} e^{-imh\xi} e^{-im\pi} = \frac{h}{\sqrt{2\pi}} \sum_{-M+1}^{M-1} e^{-im(\pi+h\xi)} \\
&= \frac{h}{\sqrt{2\pi}} \frac{\sin(M-\frac{1}{2})(\pi+h\xi)}{\sin(\frac{1}{2}(\pi+h\xi))} = \frac{(-1)^{M-1}h}{\sqrt{2\pi}} \frac{\cos(\xi-\frac{1}{2}h\xi)}{\cos(\frac{1}{2}h\xi)} \\
&= \frac{(-1)^{M-1}h}{\sqrt{2\pi}} (\cos(\xi) + \sin(\xi)\tan(\frac{1}{2}h\xi)).
\end{aligned}$$

In Fig. 2.1 we have to the right shown the Fourier transform of the oscillation for $M = 11$. The abscissa is $h\xi$ and we have only shown the interval $0 \le h\xi \le \pi$ as the function is symmetric around 0. We notice a substantial weight near π. □

From Example 1 we see that the Fourier transform of the discrete function which is equal to 1 in an interval around 0 has a significant contribution for $h\xi$ close to 0. The remaining wiggles originate from the sudden drop to 0 which must happen somewhere for the function to have a finite norm.

From Example 2 we see that the Fourier transform of the oscillation has a significant contribution near $h\xi = \pi$ (and $-\pi$). We therefore talk about values around $h\xi = \pi$ as corresponding to *high-frequency* components and values around $h\xi = 0$ as corresponding to *low-frequency* components of the function in question.

2.3 Fourier analysis and differential equations

If we differentiate (2.2) w.r.t. x we get

$$\frac{du}{dx} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega x}\, i\omega\, \hat{u}(\omega)\, d\omega. \tag{2.9}$$

From this and the uniqueness of the Fourier transform it follows that the Fourier transform of the derivative is obtained by multiplication with $i\omega$:

$$(\widehat{\frac{du}{dx}})(\omega) = i\omega\hat{u}(\omega). \tag{2.10}$$

If we now consider functions, $u(t,x)$, of two variables we can still perform the Fourier transform in the x-variable. Applying this to the IVP for the simple heat equation

$$u_t = bu_{xx}, \qquad u(0,x) = u_0(x) \tag{2.11}$$

we obtain

$$\hat{u}_t(t,\omega) \;=\; b(i\omega)^2\hat{u}(t,\omega) \;=\; -b\omega^2\hat{u}(t,\omega). \tag{2.12}$$

Using the fact that the Fourier transform of the time derivative is the same as the time derivative of the Fourier transform we have obtained an ordinary differential equation in t for the Fourier transform of u with initial condition $\hat{u}(0,\omega) = \hat{u}_0(\omega)$. Since the solution to the IVP

$$y' \;=\; -b\omega^2 y, \;\; y(0) \;=\; y_0$$

is

$$y(t) \;=\; y_0 e^{-b\omega^2 t}$$

we deduce that

$$\hat{u}(t,\omega) \;=\; e^{-b\omega^2 t}\,\hat{u}_0(\omega). \tag{2.13}$$

Using Parseval we can now get a bound for the L^2-norm of $u(t,.)$ at an arbitrary time $t > 0$:

$$\begin{aligned} ||u(t,.)||_2^2 \;&=\; \int_{-\infty}^{\infty} |u(t,x)|^2\,dx \;=\; \int_{-\infty}^{\infty} |\hat{u}(t,\omega)|^2\,d\omega \qquad (2.14)\\ &=\; \int_{-\infty}^{\infty} e^{-2b\omega^2 t}|\hat{u}_0(\omega)|^2\,d\omega \;\leq\; \int_{-\infty}^{\infty} |\hat{u}_0(\omega)|^2\,d\omega \;=\; ||u_0(x)||^2 \end{aligned}$$

provided $b > 0$.

We note that for positive b we have a bound on the norm of the solution for $t > 0$: the IVP is *well-posed*. On the other hand, if $b < 0$ (or $t < 0$) the problem is *ill-posed* and we may expect unbounded growth. When $\hat{u}_0(\omega) \neq 0$ for some value of $\omega \neq 0$ then these components will be amplified with the factor $e^{-b\omega^2 t}$. The higher the frequency ω the higher the amplification factor.

2.4 Von Neumann analysis

Consider the explicit method (1.11) on the simple heat equation $u_t = bu_{xx}$. Solving for v_m^{n+1} we get

$$v_m^{n+1} \;=\; (1-2b\mu)v_m^n + b\mu(v_{m+1}^n + v_{m-1}^n). \tag{2.15}$$

We now take the Fourier inversion formula

$$v_m^n \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h} e^{imh\xi}\,\hat{v}^n(\xi)\,d\xi \tag{2.16}$$

and apply it on the right-hand-side of (2.15) for m, $m+1$, and $m-1$ to get

$$v_m^{n+1} \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \left[1 - 2b\mu + b\mu(e^{ih\xi} + e^{-ih\xi})\right] \hat{v}^n(\xi)\, d\xi. \tag{2.17}$$

Using the uniqueness of the Fourier transform we deduce

$$\hat{v}^{n+1}(\xi) \;=\; g(h\xi)\hat{v}^n(\xi) \tag{2.18}$$

where the *amplification factor* or *growth factor*, $g(h\xi)$, is given by the square bracket above:

$$\begin{aligned} g(h\xi) &= 1 - 2b\mu + b\mu(e^{ih\xi} + e^{-ih\xi}) \;=\; 1 - 2b\mu + 2b\mu\cos(h\xi) \\ &= 1 - 4b\mu \sin^2 \frac{h\xi}{2}. \end{aligned} \tag{2.19}$$

So when we advance the numerical solution from one time step to the next the Fourier transform of the solution is multiplied by $g(h\xi)$. By induction we then have

$$\hat{v}^n(\xi) \;=\; (g(h\xi))^n\; \hat{v}^0(\xi). \tag{2.20}$$

Looking at the norms we have

$$\begin{aligned} ||v^n||_2^2 &= h\sum_m |v_m^n|^2 \;=\; \int_{-\pi/h}^{\pi/h} |\,\hat{v}^n(\xi)|^2\, d\xi \\ &= \int_{-\pi/h}^{\pi/h} |g(h\xi)|^{2n}\, |\hat{v}^0(\xi)|^2\, d\xi. \end{aligned} \tag{2.21}$$

For stability we would like $||v^n||_2$ to be bounded in relation to $||v^0||_2$. We see that we can achieve this if $|g(h\xi)^{2n}|$ is suitably bounded. In the considerations to follow, the variable, ξ, will usually appear together with the step size, h, so to simplify the notation we introduce $\varphi = h\xi$.

For the explicit method on $u_t = bu_{xx}$ the growth factor now reads

$$g(\varphi) \;=\; 1 - 4b\mu \sin^2 \frac{\varphi}{2}, \qquad\qquad -\pi \le \varphi \le \pi. \tag{2.22}$$

We note that $g(\varphi)$ is always real and ≤ 1. If $2b\mu \le 1$ then $4b\mu\sin^2\frac{\varphi}{2} \le 2$ and $g(\varphi) \ge -1$. In this case $|g(\varphi)| \le 1$ and therefore $|g(\varphi)|^{2n} \le 1$ showing that we have stability with $C_T = 1$ for all T:

$$||v^n||_2 \;\le\; ||v^0||_2.$$

If $2b\mu > 1$ then $g(\varphi) < -1$ in an interval around π. Therefore $|g(\varphi)|^{2n}$ will grow without bound in this interval. We can provide no bound for $||v^n||_2$ and if $\hat{v}^0 \ne 0$

somewhere in this interval the corresponding components of the solution will be magnified: we have instability.

We conclude that the explicit method (2.15) for the heat equation $u_t = bu_{xx}$ is stable iff $2b\mu \leq 1$ or $k \leq h^2/(2b)$. We also note that instability shows up first in the high frequency components of the solution.

Remark. This technique for analyzing the stability of finite difference schemes is named after John von Neumann (or Neumann János Lajos, 1903 – 1957). □

It is not necessary to invoke the Fourier transform and equate integral expressions every time we want to investigate the stability properties of a difference scheme for a differential equation. Looking at equations (2.17) and (2.20) we conclude that the essential features are captured if we replace v_m^n by $g^n e^{im\varphi}$ in the difference scheme and this is also the way the technique was first presented in [5]. This is sometimes interpreted in the way that we are seeking solutions to the difference scheme on the form $v_m^n = g^n e^{im\varphi}$ with obvious parallels to the solution by separation of variables for PDEs as discussed in section 1.3. This is also fine for memorizing as long as we keep in mind that through the Fourier transform we have the sound mathematical basis for our arguments.

Example. Consider now the explicit method on the equation $u_t = bu_{xx} + \kappa u$. The difference equation is

$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + \kappa v_m^n. \tag{2.23}$$

Replacing v_m^n by $g^n e^{im\varphi}$ and dividing by $g^n e^{im\varphi}$ gives

$$\frac{g-1}{k} = -\frac{4b}{h^2}\sin^2\frac{\varphi}{2} + \kappa, \qquad -\pi \leq \varphi \leq \pi$$

or

$$g(\varphi) = 1 - 4b\mu\sin^2\frac{\varphi}{2} + \kappa k, \qquad -\pi \leq \varphi \leq \pi. \tag{2.24}$$

□

If $\kappa > 0$ then we have $g(\varphi) > 1$ for small φ and so it seems that we may have problems with stability for $\kappa > 0$ (and stricter bounds on $b\mu$ for $\kappa < 0$). But note the following three points:

- $u(t,x) = (\alpha + \beta x)e^{\kappa t}$ is a solution to the differential equation and if $\kappa > 0$ then this solution exhibits exponential growth. We should expect the same from a good numerical solution.
- As $k \to 0$ (which it does when we consider convergence) g will approach the usual value from when κu was not present in the equation.

- The condition $|g(\varphi)| \leq 1$ actually gave more than we demanded. We wanted bounded growth and we got $C_T = 1$ for all T.

We may allow $|g(\varphi)| > 1$ under certain conditions as the following theorem shows:

Theorem. A one-step difference scheme is *stable* if

$$\exists K, h_0, k_0 \text{ such that } |g(\varphi, h, k)| \leq 1 + Kk,\ \forall \varphi,\ 0 < k \leq k_0,\ 0 < h \leq h_0.$$

Proof:

$$|g(\varphi, h, k)|^{2n} \leq (1 + Kk)^{2n} < (e^{Kk})^{2n} = e^{2Knk} = e^{2Kt} \leq e^{2KT} = C_T^2$$

so for $nk = t \leq T$ the norm of v^n is bounded by $C_T = e^{KT}$ times the norm of the initial function. □

Example. Now consider the explicit method on the equation $u_t = bu_{xx} - au_x$. The difference equation is

$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} - a\frac{v_{m+1}^n - v_{m-1}^n}{2h}. \qquad (2.25)$$

Replacing v_m^n by $g^n e^{im\varphi}$ and dividing by $g^n e^{im\varphi}$ gives

$$\frac{g-1}{k} = -\frac{4b}{h^2}\sin^2\frac{\varphi}{2} - a\frac{e^{i\varphi} - e^{-i\varphi}}{2h}, \qquad -\pi \leq \varphi \leq \pi$$

or

$$g(\varphi) = 1 - 4b\mu \sin^2\frac{\varphi}{2} - ia\frac{k}{h}\sin\varphi, \qquad -\pi \leq \varphi \leq \pi. \qquad (2.26)$$

Now the growth factor is a complex number and we find the square of the absolute value

$$|g(\varphi)|^2 = (1 - 4b\mu \sin^2\frac{\varphi}{2})^2 + a^2\frac{k^2}{h^2}\sin^2\varphi. \qquad (2.27)$$

If $2b\mu \leq 1$ then the first parenthesis is ≤ 1 and we have

$$|g(\varphi)|^2 \leq 1 + a^2\mu k, \qquad -\pi \leq \varphi \leq \pi$$

and therefore

$$|g(\varphi)| \leq 1 + \frac{1}{2}a^2\mu k + O(k^2), \qquad -\pi \leq \varphi \leq \pi.$$

Once again the stability condition is the same as for the simple heat equation ($2b\mu \leq 1$) irrespective of the lower order terms. □

2.5 Implicit methods

The implicit method for $u_t = bu_{xx} - au_x + \kappa u$ is

$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2} - a\frac{v_{m+1}^{n+1} - v_{m-1}^{n+1}}{2h} + \kappa v_m^{n+1}. \quad (2.28)$$

Replacing v_m^n by $g^n e^{im\varphi}$ and dividing by $g^n e^{im\varphi}$ gives

$$\frac{g-1}{k} = -g\frac{4b}{h^2}\sin^2\frac{\varphi}{2} - ag\frac{e^{i\varphi} - e^{-i\varphi}}{2h} + \kappa g$$

or

$$g(\varphi) = \frac{1}{1 + 4b\mu\sin^2\frac{\varphi}{2} + ia\frac{k}{h}\sin\varphi - \kappa k}. \quad (2.29)$$

If $\kappa \leq 0$ then already the real part of the denominator is ≥ 1 and therefore $|g(\varphi)| \leq 1$ irrespective of $b\mu$ and a. If $\kappa > 0$ then $|g(\varphi)|$ may be greater than 1 for small values of φ but by no more than $O(k)$ so we can conclude that the implicit method is unconditionally stable. If $a = \kappa = 0$ then $0 \leq g(\varphi) \leq 1$ and the smallest values are attained when $\varphi \approx \pi$ implying that high frequency components are damped most.

For the Crank-Nicolson method the growth factor becomes

$$g(\varphi) = \frac{1 - 2b\mu\sin^2\frac{\varphi}{2} - ia\frac{k}{2h}\sin\varphi + \frac{1}{2}\kappa k}{1 + 2b\mu\sin^2\frac{\varphi}{2} + ia\frac{k}{2h}\sin\varphi - \frac{1}{2}\kappa k}. \quad (2.30)$$

Once again we have unconditional stability with $|g(\varphi)| \leq 1$ in all cases except possibly when $\kappa > 0$ in which case we still have $|g(\varphi)| \leq 1 + O(k)$. If $a = \kappa = 0$ then $-1 \leq g(\varphi) \leq 1$ and values close to -1 are attained when $\varphi \approx \pi$ implying that high frequency components are damped little when $b\mu \gg 1$. So a solution or error component which oscillates in the x-direction will also oscillate in the t-direction with a slowly diminishing amplitude.

2.6 Two kinds of stability

As indicated in the theorem in section 2.4 a certain growth is allowed for a stable difference scheme. This is in accordance with the definition of stability which allows a growth in the norm with a factor C_T times the norm of the initial function. This kind of stability is what is needed in the Lax equivalence theorem where stability is needed to ensure *convergence* as the step sizes tend to 0. This kind of stability is sometimes called *numerical stability* [35] or *pointwise stability*

[25] or with a term borrowed from ODEs *0-stability.* In practice we shall often be concerned with problems with decaying solutions and we shall use a numerical scheme with fixed positive step sizes and for many steps. In this case any kind of growth is unwanted. We shall in these cases want the condition $|g(\varphi)| \leq 1$ and we speak of *dynamic stability* [35] or *stepwise stability* [25] or with a term from ODEs *absolute stability.*

2.7 Exercise

1. Prove formula (2.30).

Chapter 3

Accuracy

In Chapter 1 we introduced the concept of consistency which together with stability implies convergence of the numerical solution to the true solution as the step sizes tend to 0. In this chapter we shall also be concerned with how fast this convergence is as a means to compare various difference schemes. In other words we shall study the *local truncation error*: $P_{k,h}\psi - R_{k,h}P\psi$ and how fast it converges to 0 as the step sizes h and k tend to 0. We shall use Taylor expansions in this study of the difference schemes.

3.1 Taylor expansions

If $f(x)$ is a smooth function of x then it can be expanded in a Taylor series

$$f(x+h) \;=\; f(x) + hf'(x) + \frac{1}{2}h^2 f''(x) + \cdots + \frac{1}{p!}h^p f^{(p)}(x) + O(h^{p+1}). \quad (3.1)$$

That f is smooth will in this case mean that f is p times continuously differentiable in an interval around x. If $\psi(t,x)$ is a smooth function of t and x, meaning that it possesses continuous partial derivatives of a suitable high order, then it likewise can be expanded in a Taylor series with two variables. We shall usually refrain from doing so because less complicated expressions are produced, and cancellations easier detected, when we expand first in one coordinate and later in the other. The end result will be the same but the risk of committing errors is greatly reduced this way. The basic expansions are the following

$$\psi_m^{n+1} \;=\; \psi_m^n + k\psi_t + \frac{1}{2}k^2\psi_{tt} + \cdots + \frac{1}{p!}k^p\psi_{pt} + O(k^{p+1}), \quad (3.2)$$

$$\psi_{m+1}^{n} \;=\; \psi_m^n + h\psi_x + \frac{1}{2}h^2\psi_{xx} + \cdots + \frac{1}{q!}h^q\psi_{qx} + O(h^{q+1}). \quad (3.3)$$

The common expansion point for the function and all derivatives is indicated in the leading term on the right-hand-side but otherwise omitted. The expansion for ψ_{m-1}^n is easily obtained from (3.3) by changing sign for all the odd terms. For symmetric expressions we can exploit cancellations such as

$$\begin{aligned}\psi_{m+1}^n + \psi_{m-1}^n &= 2\psi_m^n + \frac{2}{2}h^2\psi_{xx} + \cdots + \frac{2}{q!}h^q\psi_{qx} + O(h^{q+2}), &(3.4)\\ \psi_{m+1}^n - \psi_{m-1}^n &= 2h\psi_x + \frac{2}{6}h^3\psi_{xxx} + \cdots + \frac{2}{q!}h^q\psi_{qx} + O(h^{q+2}), &(3.5)\end{aligned}$$

where q is even and odd, respectively.

3.2 Order

We shall illustrate the use of Taylor expansions with the explicit method on the simple heat equation $u_t - bu_{xx} = \nu$. The differential operator is $P = D_t - bD^2$, the difference operator is $P_{k,h} = \Delta_t - b\delta^2$, and the right-hand-side operator $R_{k,h}$ is the identity. Using Taylor expansions we get

$$\begin{aligned}\Delta_t\psi_m^n &= \psi_t + \frac{1}{2}k\psi_{tt} + O(k^2),\\ \delta^2\psi_m^n &= \psi_{xx} + \frac{1}{12}h^2\psi_{xxxx} + O(h^4),\end{aligned}$$

and

$$P_{k,h}\psi - R_{k,h}P\psi = \frac{1}{2}k\psi_{tt} - \frac{1}{12}bh^2\psi_{xxxx} + O(k^2 + h^4) \qquad (3.6)$$

As the step sizes tend to 0 the whole expression on the right-hand-side tends to 0, so we immediately conclude that the explicit method is consistent. But the expression also reveals the rate of convergence.

Definition. A difference scheme $P_{k,h}v = R_{k,h}\nu$ which is consistent with the equation $Pu = \nu$ is *accurate of order* p in time and *order* q in space iff

$$P_{k,h}\psi - R_{k,h}P\psi = O(k^p) + O(h^q) \qquad (3.7)$$

for smooth functions ψ. We say that the scheme is accurate of order (p, q). □

Remark. If $p = q$ we say that the method is of order p. □

Using this definition we say that the explicit method on the simple heat equation is accurate of order $(1, 2)$ meaning that it is first order in time and second order in space.

We have already seen that the explicit method requires $2b\mu \le 1$ or equivalently $k \le h^2/(2b)$ to be stable. It is therefore customary with the explicit method to use a step size k which is proportional to h^2. Therefore the following

Definition. A difference scheme $P_{k,h}v = R_{k,h}\nu$ with $k = \Lambda(h)$ is *accurate of order* r iff

$$P_{k,h}\psi - R_{k,h}P\psi \;\; = \;\; O(h^r) \tag{3.8}$$

for smooth functions ψ. □

If we use the explicit method with $k = h^2/(2b)$ then it is accurate of order 2.

Example. For the implicit method on the simple heat equation we have $P_{k,h} = \nabla_t - b\delta^2$, where the evaluation point now is at the advanced time level, and

$$\nabla_t \psi_m^n \;\; = \;\; \psi_t - \frac{1}{2}k\psi_{tt} + O(k^2) \tag{3.9}$$

and therefore

$$P_{k,h}\psi - R_{k,h}P\psi \;\; = \;\; -\frac{1}{2}k\psi_{tt} - \frac{1}{12}bh^2\psi_{xxxx} + O(k^2 + h^4). \tag{3.10}$$

We note that the implicit method is accurate of order $(1,2)$ just like the explicit method. Since the implicit method is unconditionally stable we are free to choose k proportional to h as far as stability is concerned. For reasons of accuracy this might not be such a great idea since the first order term in k will probably dominate the error.

Note also that we in (3.9) and (3.10) implicitly have assumed that we evaluate the inhomogeneous term at the advanced time level. This will probably not cause any problems since the right-hand-side function is supposed to be known, but if for some reason we want to evaluate $\nu(t,x)$ at the time level where we know the solution function, i.e. $\nu(t-k,x)$ then the right-hand-side operator $R_{k,h}$ is no longer the identity but an inverse time shift: $R_{k,h} = E_t^{-1}$ such that

$$\begin{aligned} R_{k,h}P\psi_m^n &= E_t^{-1}(\psi_t - b\psi_{xx}) \\ &= \psi_t - b\psi_{xx} - k\psi_{tt} + bk\psi_{xxt} + O(k^2) \end{aligned}$$

and therefore

$$P_{k,h}\psi - R_{k,h}P\psi \;\; = \;\; \frac{1}{2}k\psi_{tt} - bk\psi_{xxt} - \frac{1}{12}bh^2\psi_{xxxx} + O(k^2 + h^4).$$

The order is still $(1,2)$. □

3.3 Symbols of operators

The result of using Taylor expansions on the various (combinations of) differential and difference operators are polynomials or power series in k and h whose coefficients in addition to numerical constants contain partial derivatives w.r.t. t and x of the smooth function ψ. Any smooth function will do as long as none of its partial derivatives vanish. We can simplify our investigations by a suitable choice of ψ, a choice inspired by our considerations in section 1.3 and the test functions we introduced for stability in section 2.4 where we used the product of an exponential function in time and a trigonometric (or complex exponential) function in space. So we choose

$$\psi(t,x) \;=\; e^{st}e^{i\xi x} \;=\; e^{snk}e^{i\xi mh} \tag{3.11}$$

the first expression to be used with differential operators, the second one with difference operators.

Example.

$$\psi_t = s\psi, \;\; \psi_x = i\xi\psi, \;\; \psi_{xx} = -\xi^2\psi$$

$$E\psi = e^{snk}e^{i\xi(m+1)h} = e^{i\xi h}\psi = (1 + i\xi h - \frac{1}{2}\xi^2 h^2 + \cdots)\psi$$

$$\Delta\psi = \frac{1}{h}(E-1)\psi = (i\xi - \frac{1}{2}\xi^2 h + \cdots)\psi \qquad \square$$

When we apply a differential operator P on ψ the result is a polynomial $p(s,\xi)$ in s and ξ times ψ. This polynomial is called the *symbol* of the differential operator. Similarly when we apply a difference operator $P_{k,h}$ on ψ the result is a power series $p_{k,h}(s,\xi)$ times ψ. $p_{k,h}(s,\xi)$ is called the *symbol* of the difference operator $P_{k,h}$.

Example. Since $\psi_t - b\psi_{xx} = (s + b\xi^2)\psi$ the symbol of the differential operator $P = D_t - bD^2$ for the simple heat equation is $p(s,\xi) = s + b\xi^2$. □

The definitions of order can now be reformulated in terms of the symbols of the operators:

Theorem. A difference scheme $P_{k,h}v = R_{k,h}\nu$ which is consistent with the equation $Pu = \nu$ is *accurate of order* p in time and *order* q in space iff

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) \;=\; O(k^p) + O(h^q) \tag{3.12}$$

where $p(s,\xi)$, $p_{k,h}(s,\xi)$, and $r_{k,h}(s,\xi)$ are the symbols of the operators P, $P_{k,h}$, and $R_{k,h}$, respectively. □

Theorem. A difference scheme $P_{k,h}v = R_{k,h}\nu$ with $k = \Lambda(h)$ is *accurate of order* r iff

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) \;\; = \;\; O(h^r). \tag{3.13}$$

□

Example. Crank-Nicolson's method on the simple heat equation $u_t - bu_{xx} = \nu$ can be written

$$(\delta_t - b\tilde{\mu}_t\delta^2)v_m^{n+\frac{1}{2}} = \tilde{\mu}_t\nu_m^{n+\frac{1}{2}}$$

with the evaluation point located midway between the present and the advanced time level. The symbol of the differential operator is $p(s,\xi) = s+b\xi^2$ as mentioned above. The symbol of the left-hand-side difference operator $P_{k,h} = \delta_t - b\tilde{\mu}_t\delta^2$ is

$$\begin{aligned}p_{k,h}(s,\xi) &= \frac{1}{k}(e^{\frac{1}{2}sk} - e^{-\frac{1}{2}sk}) - \frac{b}{2h^2}(e^{\frac{1}{2}sk} + e^{-\frac{1}{2}sk})(e^{i\xi h} - 2 + e^{-i\xi h})\\
&= s + \frac{1}{24}s^3k^2 + O(k^4) - b(1 + \frac{1}{8}s^2k^2 + O(k^4))(-\xi^2 + \frac{1}{12}\xi^4h^2 + O(h^4))\\
&= s + b\xi^2 + \frac{1}{24}s^3k^2 + \frac{b}{8}s^2\xi^2k^2 - \frac{b}{12}\xi^4h^2 + O(h^4 + h^2k^2 + k^4)\end{aligned}$$

and the symbol for the right-hand-side operator is

$$r_{k,h}(s,\xi) = \frac{1}{2}(e^{\frac{1}{2}sk} + e^{-\frac{1}{2}sk}) = 1 + \frac{1}{8}s^2k^2 + O(k^4)$$

such that

$$r_{k,h}(s,\xi)p(s,\xi) = s + b\xi^2 + \frac{1}{8}s^3k^2 + \frac{b}{8}s^2\xi^2k^2 + O(k^4)$$

and

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = -\frac{1}{12}s^3k^2 - \frac{b}{12}\xi^4h^2 + O(h^4 + h^2k^2 + k^4)$$

showing that Crank-Nicolson is second order accurate in both time and space. □

Remark. The contribution of the right-hand-side operator $R_{k,h}$ is important in producing second order accuracy. If we only evaluate the right-hand-side function ν at time $t = nk$ or $t = (n+1)k$ the method will be first order accurate in time. □

Remark. If the right-hand-side function is known for intermediate values of t then it is O.K. to evaluate it for $t = (n + \frac{1}{2})k$. In this case $r_{k,h}(s,\xi) = 1$, since this is the evaluation point, and the local truncation error becomes

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = \frac{1}{24}s^3k^2 + \frac{b}{8}s^2\xi^2k^2 - \frac{b}{12}\xi^4h^2 + O(h^4 + h^2k^2 + k^4)$$

and the method retains second order accuracy. □

3.4 The local error

The local truncation error was defined earlier in this chapter as $P_{k,h}\psi - R_{k,h}P\psi$. Another important concept is the *local error* which is defined as the difference between the true solution $u(t,x)$ and the computed solution $v(t,x)$, calculated from correct initial values at time $t-k$ and boundary values at (t, X_1) and (t, X_2). If we consider the explicit method on $u_t = bu_{xx}$ we have

$$\begin{aligned} v_m^{n+1} &= u_m^n + b\mu(u_{m-1}^n - 2u_m^n + u_{m+1}^n) \\ &= u_m^n + bku_{xx} + \frac{1}{12}bkh^2u_{xxxx} + \cdots \\ &= u_m^n + ku_t + \frac{1}{12b}kh^2u_{tt} + \cdots \qquad (3.14) \\ u_m^{n+1} &= u_m^n + ku_t + \frac{1}{2}k^2u_{tt} + \cdots \qquad (3.15) \end{aligned}$$

such that the

$$\text{local error} = (\frac{1}{2}k^2 - \frac{1}{12b}kh^2)u_{tt} + \cdots = O(k^2 + kh^2). \qquad (3.16)$$

Comparing with (3.6) we notice that the local error contains an extra factor k on each term compared to the local truncation error. This is a general result although trickier to show for implicit methods.

Chapter 4

Boundary Conditions

As mentioned in section 1.4 boundary conditions are necessary in order to specify a unique solution to a parabolic differential equation on a finite space interval. They also come in handy in supplying the extra equations needed to solve for the numerical solution. In the sections below we shall specify these extra equations for the various difference schemes we have introduced. Dirichlet conditions are easy to accomodate whereas conditions involving a (normal) derivative present new challenges. We shall often just treat conditions at the left end point, $x = X_1$, since the considerations for $x = X_2$ are quite similar.

4.1 A Dirichlet condition

A Dirichlet boundary conditon

$$u(t, X_1) = \gamma(t), \qquad t > 0. \tag{4.1}$$

is straightforward to apply.

For the explicit method the formulae (1.13) specify the values for the numerical solution, v_m^{n+1}, at all internal points, $m = 1, 2, \ldots, M-1$ at time level $n+1$. The boundary condition (4.1) is then used to supply the boundary value, v_0^{n+1}, and the boundary condition at $x = X_2$ is used in a similar fashion to provide v_M^{n+1}, such that we have determined the solution at all points at time level $n+1$.

For the general θ-method the formulae (1.20) provide a set of $M-1$ equations in the $M+1$ unknowns v_0^{n+1}, v_1^{n+1}, $\ldots$, v_M^{n+1}. From the Dirichlet conditions we have values for v_0^{n+1} and v_M^{n+1} and we are ready to solve for the remaining $M-1$ unknowns.

4.2 Derivative boundary conditions

If one of the boundary conditions involves a derivative then the discretization of this has an effect on the accuracy and stability of the numerical solution as well as on the solution process. Assume that the condition on the left boundary is

$$\alpha u(t, X_1) - \beta u_x(t, X_1) \quad = \quad \gamma, \qquad\qquad t > 0 \tag{4.2}$$

where α, β and γ may depend on t. A similar condition might be imposed on the other boundary (cf. section 1.4) and the considerations would be completely similar so we shall just consider a derivative condition on one boundary. We shall in turn study three different discretizations of the derivative in (4.2):

$$\frac{v_1^n - v_0^n}{h} \qquad \text{(first order)} \tag{4.3}$$

$$\frac{-v_2^n + 4v_1^n - 3v_0^n}{2h} \qquad \text{(second order, asymmetric)} \tag{4.4}$$

$$\frac{v_1^n - v_{-1}^n}{2h} \qquad \text{(second order, symmetric)} \tag{4.5}$$

We have similar expressions for u_x on the right boundary. (4.3) and (4.5) are easily adapted while (4.4) is slightly more tricky and is therefore given here:

$$\frac{v_{M-2}^n - 4v_{M-1}^n + 3v_M^n}{2h}$$

These approximations and their respective orders are easily determined using Taylor series. The effects of the discretization on the overall accuracy of the computed solution will be considered in Chapter 9. The effects on the stability of the overall method will be treated in Chapter 6. In the subsequent sections we shall focus on the practical considerations around the sets of linear equations to be solved at each time step.

4.3 A third test problem

As an example of a problem with a derivative boundary condition at one of the boundaries we consider the following which is closely related to test problem 1 on page 5:

Problem 3

$$u_t \quad = \quad u_{xx}, \qquad\qquad 0 \leq x \leq 1, \qquad t > 0,$$

$$\begin{aligned} u(0,x) = u_0(x) &= \cos x, && 0 \le x \le 1, \\ u(t,1) &= e^{-t}\cos 1, && t > 0, \\ u_x(t,0) &= 0, && t > 0. \end{aligned}$$

It is easily seen that the true solution is the same as for test problem 1: $u(t,x) = e^{-t}\cos x$.

4.4 The explicit method

4.4.1 First order approximation

Using the first order approximation (4.3) to u_x in (4.2) results in

$$\alpha v_0^{n+1} - \beta \frac{v_1^{n+1} - v_0^{n+1}}{h} = \gamma$$

or

$$(h\alpha + \beta) v_0^{n+1} = h\gamma + \beta v_1^{n+1} \tag{4.6}$$

or

$$v_0^{n+1} = \frac{h\gamma + \beta v_1^{n+1}}{h\alpha + \beta} \tag{4.7}$$

which is used to compute v_0^{n+1}. Since we have assumed that α and β have the same sign there are no problems with a zero denominator.

4.4.2 Asymmetric second order

The approximation (4.3) is only first order accurate and this will have an adverse effect on the overall accuracy of the method as we shall see in Chapter 9. Using the second order approximation (4.4) in (4.2) we get

$$\alpha v_0^{n+1} - \beta \frac{-v_2^{n+1} + 4v_1^{n+1} - 3v_0^{n+1}}{2h} = \gamma$$

or

$$(2h\alpha + 3\beta) v_0^{n+1} = 2h\gamma - \beta v_2^{n+1} + 4\beta v_1^{n+1} \tag{4.8}$$

or

$$v_0^{n+1} = \frac{2h\gamma - \beta v_2^{n+1} + 4\beta v_1^{n+1}}{2h\alpha + 3\beta} \tag{4.9}$$

which is used to compute v_0^{n+1}. Once again a zero denominator is not possible.

4.4.3 Symmetric second order

Symmetric difference approximations are often more accurate than asymmetric ones so we should like to investigate the merits of such a formula. The symmetric second order approximation (4.5) refers to a point outside the region where the differential equation is defined. We call this point a *fictitious point* and no physical significance should be attached to the value assigned to it. It is merely a computational quantity. The basic assumption is that the solution function can be extended slightly beyond the boundary as a smooth function obeying the same differential equation as in the interior. We then apply the difference scheme also at the boundary. In order to calculate v_0^{n+1} we need information on v_1^n, v_0^n, and the fictitious value v_{-1}^n. This is obtained by applying (4.5) to (4.2):

$$\alpha v_0^n - \beta \frac{v_1^n - v_{-1}^n}{2h} \quad = \quad \gamma$$

or

$$\beta(v_{-1}^n - v_1^n) \quad = \quad 2h\gamma - 2h\alpha v_0^n. \tag{4.10}$$

If $\beta = 0$ we have a Dirichlet condition, v_0^n is defined from (4.10), and there is no reason to incorporate v_{-1}^n in the first place. If $\beta \neq 0$ we get

$$v_{-1}^n \quad = \quad v_1^n + \frac{2h}{\beta}(\gamma - \alpha v_0^n). \tag{4.11}$$

4.5 The implicit method

4.5.1 First order approximation

The system of linear equations for v^{n+1} contains $M-1$ equations in $M+1$ unknowns. One extra equation at the beginning is supplied by (4.6):

$$(h\alpha + \beta)v_0^{n+1} - \beta v_1^{n+1} \quad = \quad h\gamma \tag{4.12}$$

and a similar one is supplied at the other end from the boundary condition at X_2.

4.5.2 Asymmetric second order

The extra equation is now supplied by (4.8):

$$(2h\alpha + 3\beta)v_0^{n+1} - 4\beta v_1^{n+1} + \beta v_2^{n+1} \quad = \quad 2h\gamma. \tag{4.13}$$

The resulting system of equations is no longer tridiagonal because of the extra coefficient in the first (and possibly the last) equation but a Gaussian elimination can still be done without introducing new non-zero values in the coefficient matrix and without affecting the linear complexity of the solution (cf. section 1.12).

4.5.3 Symmetric second order

We apply the difference scheme at the boundary point arriving at a linear equation involving v_{-1}^{n+1}, v_0^{n+1}, and v_1^{n+1}. The extra equation is obtained from (4.10):

$$\beta v_{-1}^{n+1} + 2h\alpha v_0^{n+1} - \beta v_1^{n+1} = 2h\gamma. \tag{4.14}$$

Once again the system of equations is almost tridiagonal in the terminology of section 1.12, and a Gaussian elimination can be performed without any real difficulties.

4.6 The θ-method

For the general θ-method and in particular for Crank-Nicolson we use the formulas of the preceding sections to give the extra equations needed.

4.7 Exercises

1. Solve problem 3 with the implicit method from $t = 0$ to $t = \frac{1}{2}$ with $h = k = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$.
   Use each of the three approximations (4.3) – (4.5) to approximate the derivative at the boundary.
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1, 0.2, 0.3, 0.4, 0.5$.

2. Solve problem 3 with Crank-Nicolson from $t = 0$ to $t = \frac{1}{2}$ with
   $h = k = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$.
   Use each of the three approximations (4.3) – (4.5) to approximate the derivative at the boundary.
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1, 0.2, 0.3, 0.4, 0.5$.

Chapter 5

The Convection-Diffusion Equation

5.1 Introduction

The simplest convection-diffusion equation is

$$u_t = bu_{xx} - au_x \tag{5.1}$$

or as we sometimes prefer

$$u_t + au_x = bu_{xx}. \tag{5.2}$$

If we begin with

$$u_t + au_x = 0 \tag{5.3}$$

we can easily see that the solution can be written

$$u(t, x) = u_0(x - at) \tag{5.4}$$

where $u_0(x)$ is the initial value at $t = 0$. (5.3) is called the *one-way wave equation* and according to (5.4) describes transport in the x-direction with velocity a. Inspired by (5.4) we introduce

$$w(t, y) = u(t, y + at) = u(t, x) \tag{5.5}$$

and find that

$$w_t = u_t + au_x = bu_{xx} = bw_{yy}. \tag{5.6}$$

So $w(t, y)$ is the solution to the simple heat equation and

$$u(t, x) = w(t, x - at). \tag{5.7}$$

The equation (5.1) thus describes simultaneous transport and diffusion.

5.2 Maximum principle

The solutions to the simple heat equation

$$u_t \quad = \quad b u_{xx} \tag{5.8}$$

satisfy a *maximum principle*:

$$\max_{\Omega} |u(t,x)| \quad \leq \quad \max_{\partial\Omega} |u(t,x)| \tag{5.9}$$

where Ω is the open region $\Omega = \{(t,x) |\ 0 < t < T,\ X_1 < x < X_2\}$ and $\partial\Omega$ is the *parabolic boundary* of Ω consisting of the three straight lines

$$\{t = 0,\ X_1 \leq x \leq X_2\}, \{0 \leq t \leq T,\ x = X_1\}, \{0 \leq t \leq T,\ x = X_2\}.$$

To see this assume that $u(t_0, x_0)$ is a local maximum for some (t_0, x_0) satisfying $0 < t_0 \leq T,\ X_1 < x_0 < X_2$. Since it is a maximum in the x-direction we must have $u_x = 0,\ u_{xx} < 0$ and since it is a maximum in the t-direction (possibly at T) we must have $u_t \geq 0$ which is incompatible with (5.8).

Remark. We might have $u_{xx} = 0$ in which case $u_t = 0$ and some higher order even derivative, e.g. u_{xxxx}, must be negative and all lower order derivatives w.r.t. x must be 0. By differentiating (5.8) we get $u_{tt} = (bu_{xx})_t = bu_{txx} = b^2 u_{xxxx}$ and a similar contradiction arises. □

Since the solutions to (5.8) satisfy a maximum principle so do the solutions to (5.1) by the relations (5.6) and (5.7).

In Appendix A we study a number of difference schemes which can be proposed for the solution of the one-way wave equation. Here we continue with the convection-diffusion equation

5.3 The explicit method

For the convection-diffusion equation we would prefer to use a central difference approximation to u_x such that our method can remain of order 2 in x:

$$\frac{v_m^{n+1} - v_m^n}{k} - b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} \quad = \quad 0 \tag{5.10}$$

or

$$\begin{aligned} v_m^{n+1} &= (b\mu + \frac{1}{2}a\lambda)v_{m-1}^n + (1 - 2b\mu)v_m^n + (b\mu - \frac{1}{2}a\lambda)v_{m+1}^n \\ &= b\mu(1+\alpha)v_{m-1}^n + (1 - 2b\mu)v_m^n + b\mu(1-\alpha)v_{m+1}^n \end{aligned} \tag{5.11}$$

where we have introduced

$$\lambda = \frac{k}{h}, \quad \mu = \frac{k}{h^2}, \quad \alpha = \frac{a\lambda}{2b\mu} = \frac{ah}{2b}. \tag{5.12}$$

As we noticed in section 2.4 the low order term has no influence on 0-stability so we still have the well-known condition

$$2b\mu \leq 1. \tag{5.13}$$

For this problem it might be relevant to ask for absolute stability, i.e. $|g(\varphi)| \leq 1$. From section 2.4 we have

$$g(\varphi) = 1 - 4b\mu \sin^2\frac{\varphi}{2} - ia\lambda \sin\varphi, \qquad -\pi \leq \varphi \leq \pi. \tag{5.14}$$

and

$$\begin{aligned} |g(\varphi)|^2 &= (1 - 4b\mu \sin^2\frac{\varphi}{2})^2 + a^2\frac{k^2}{h^2}\sin^2\varphi \qquad (5.15)\\ &= 1 - 8b\mu\sin^2\frac{\varphi}{2} + 16b^2\mu^2\sin^4\frac{\varphi}{2} + 4a^2k\mu\sin^2\frac{\varphi}{2}\cos^2\frac{\varphi}{2}. \end{aligned}$$

$$\begin{aligned} |g(\varphi)| &\leq 1 \\ \Leftrightarrow \quad a^2k\cos^2\frac{\varphi}{2} + 4b^2\mu\sin^2\frac{\varphi}{2} &\leq 2b, \qquad -\pi \leq \varphi \leq \pi \\ \Leftrightarrow \quad k \leq \frac{2b}{a^2} = k_0 \qquad \text{and} \qquad 2b\mu &\leq 1. \qquad (5.16) \end{aligned}$$

We see that for absolute stability we have in addition to the usual condition (5.13) an upper bound k_0 on the allowable time step, a bound which might be rather strict if we have convection dominated diffusion ($a > b$).

Since the true solution obeys a maximum principle we might consider a similar requirement on the numerical solution. The condition for this is that all coefficients in (5.11) be non-negative, i.e.

$$\alpha \leq 1 \qquad \text{and} \qquad 2b\mu \leq 1. \tag{5.17}$$

That the conditions are sufficient follows from the fact that v_m^{n+1} is a weighted average of values from time step n.
That $\alpha \leq 1$ is necessary is seen by assuming $\alpha > 1$ and taking
$v_m^0 = 1$ for $m \leq 0$ and $v_m^0 = 0$ for $m > 0$. Then

$$v_0^1 = b\mu(1+\alpha) + 1 - 2b\mu = 1 + (\alpha - 1)b\mu > 1.$$

The condition $\alpha \leq 1$ is equivalent to

$$h \leq \frac{2b}{a} = h_0 \tag{5.18}$$

so for a discrete maximum principle we have a bound on the maximum allowable x-step. If we choose $h = h_0$ then we must have

$$k \;\leq\; \frac{h_0^2}{2b} \;=\; \frac{2b}{a^2} \;=\; k_0$$

where k_0 is the same as in the condition for absolute stability.

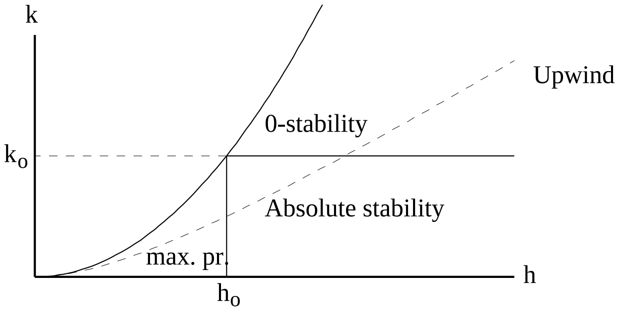


Figure 5.1: Regions for maximum principle, absolute and 0-stability for the explicit and upwind schemes

We have illustrated the three conditions in an h-k-diagram in Fig. 5.1 for the case $b = 0.1$ and $a = 10$. In this case $h_0 = 0.02$ and $k_0 = 0.002$. For 0-stability k must be below the parabola $h^2/2b$. For absolute stability k must also be smaller than k_0, and for a maximum principle to hold h must be smaller than h_0. If h is larger then the solution will display (bounded) oscillations.

Remark. If we choose $h = h_0$ (i.e. $\alpha = 1$) and $k = k_0$ (i.e. $2b\mu = 1$) then (5.11) degenerates into

$$v_m^{n+1} \;=\; v_{m-1}^n.$$

In this case we represent the transport part perfectly (because $ak_0 = h_0$) and neglect the diffusion part completely. □

The ratio a/b is called the *Reynolds number* in fluid dynamics literature and the *Peclet number* in heat conduction literature. If this number is large it imposes strict limits on the step size h to avoid oscillations with the explicit method. One way to circumvent this problem is to use a first order approximation to u_x.

5.4 The upwind scheme

When $a > 0$ then it is possible to approximate u_x with a backward difference leading to

$$\frac{v_m^{n+1} - v_m^n}{k} - b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + a\frac{v_m^n - v_{m-1}^n}{h} \;=\; 0 \qquad (5.19)$$

or

$$\begin{aligned} v_m^{n+1} &= (b\mu + a\lambda)v_{m-1}^n + (1 - 2b\mu - a\lambda)v_m^n + b\mu v_{m+1}^n \\ &= b\mu(1+2\alpha)v_{m-1}^n + (1 - 2b\mu(1+\alpha))v_m^n + b\mu v_{m+1}^n. \end{aligned} \tag{5.20}$$

The approximation will only be first order accurate in x and may thus require small values of h, but this was necessary anyway to avoid oscillations, so this scheme may be worth a try.

For the growth factor we find

$$\begin{aligned} g(\varphi) &= 1 - 4b\mu\sin^2\frac{\varphi}{2} - a\lambda(1 - e^{-i\varphi}) \qquad (5.21) \\ &= 1 - 2(2b\mu + a\lambda)\sin^2\frac{\varphi}{2} - ia\lambda\sin\varphi, \qquad -\pi \le \varphi \le \pi. \end{aligned}$$

The condition for 0-stability is now

$$2b\mu + a\lambda \le 1. \tag{5.22}$$

For absolute stability we consider

$$\begin{aligned} |g(\varphi)|^2 &= 1 - 8b\mu\sin^2\frac{\varphi}{2} + 16b^2\mu^2\sin^4\frac{\varphi}{2} \qquad (5.23) \\ &\quad -4a\lambda\sin^2\frac{\varphi}{2}(1 - 4b\mu\sin^2\frac{\varphi}{2} - a\lambda\sin^2\frac{\varphi}{2}) + a^2\lambda^2\sin^2\varphi \end{aligned}$$

and

$$\begin{aligned} &|g(\varphi)| \le 1 \Leftrightarrow \\ &2b\mu(1 - (2b\mu + a\lambda)\sin^2\frac{\varphi}{2}) + a\lambda(1 - 2b\mu\sin^2\frac{\varphi}{2} - a\lambda) \ge 0. \end{aligned} \tag{5.24}$$

This inequality must hold for $\varphi = \pi$ so we must have

$$(2b\mu + a\lambda)(1 - 2b\mu - a\lambda) \ge 0$$

so (5.22) is a necessary condition. That it is also sufficient is easily seen.

From (5.20) we see that when $2b\mu + a\lambda \le 1$ then all coefficients are non-negative and v_m^{n+1} is a weighted average of values at time n, so the condition (5.22) also guarantees a discrete maximum principle.

So how does condition (5.22) compare to our previous requirements for the explicit method. First we observe that there is no upper limit on h. Once we have decided on the step size h then (5.22) puts a limit on k:

$$k \le \frac{1}{\frac{2b}{h^2} + \frac{a}{h}} = \frac{h^2}{2b(1+\alpha)} = \frac{h_0^2\alpha^2}{2b(1+\alpha)} = \frac{\alpha^2}{1+\alpha}k_0. \tag{5.25}$$

Table 5.1: Step size limits for the upwind scheme

α	$2b\mu$	h/h_0	k/k_0	assessment
1	1/2	1	1/2	worse
3	1/4	3	9/4	better
9	1/10	9	81/10	'very good'

This limit is shown with the dashed curve in Fig. 5.1. For a given value of h the bound on k is stricter, but on the other hand we have a numerical maximum principle without restrictions on h. In Table 5.1 we show the upper limit for $2b\mu$ and k for various choices of $\alpha = h/h_0$ for the special case $b = 0.1$ and $a = 10$ where $h_0 = 0.02$ and $k_0 = 0.002$.

Formula (5.19) can be rewritten as

$$\frac{v_m^{n+1} - v_m^n}{k} - (b + \frac{ah}{2})\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0 \qquad (5.26)$$

showing that the upwind scheme is an $O(h^2)$ approximation to a convection-diffusion equation with diffusion coefficient

$$b + \frac{ah}{2} = b(1 + \alpha) \qquad (5.27)$$

The upwind difference scheme (5.19) is consistent with equation (5.1) and also with the modified convection-diffusion equation where the diffusion coefficient b is replaced by $b(1 + \alpha)$. It is a first order approximation to (5.1) but a second order (in x) approximation to the modified equation. In the limit when h (and k) tend to 0, α will tend to 0 and the two equations become equal.

When using the upwind scheme we introduce *numerical diffusion* or *artificial viscosity* into the system, and even with $\alpha = 1$ this extra contribution is of the same magnitude as the original. The effect is that the upwind scheme will tend to smoothe everything too much and the assessment *very good* in Table 5.1 should be taken with a fair amount of irony. This issue has been addressed by Gresho and Lee in a paper entitled: 'Don't suppress the wiggles. They are telling you something' [13].

5.5 The implicit method

In order to avoid the stability limitations of the explicit method we might instead consider the implicit method

$$\frac{v_m^{n+1}-v_m^n}{k} - b\frac{v_{m+1}^{n+1}-2v_m^{n+1}+v_{m-1}^{n+1}}{h^2} + a\frac{v_{m+1}^{n+1}-v_{m-1}^{n+1}}{2h} \quad = \quad 0 \qquad (5.28)$$

or

$$-b\mu(1+\alpha)v_{m-1}^{n+1} + (1+2b\mu)v_m^{n+1} - b\mu(1-\alpha)v_{m+1}^{n+1} \quad = \quad v_m^n. \qquad (5.29)$$

For the growth factor we now get

$$g - 1 + 4gb\mu\sin^2\frac{\varphi}{2} + iga\lambda\sin\varphi \quad = \quad 0, \qquad -\pi \le \varphi \le \pi \qquad (5.30)$$

or

$$g(\varphi) \quad = \quad \frac{1}{1+4b\mu\sin^2\frac{\varphi}{2}+ia\lambda\sin\varphi}, \qquad -\pi \le \varphi \le \pi \qquad (5.31)$$

from which we immediately deduce that $|g(\varphi)| \le 1$, i.e. we have absolute and unconditional stability.

If $\alpha = 1$, i.e. $h = h_0$, then (5.29) reduces to

$$-2b\mu v_{m-1}^{n+1} + (1+2b\mu)v_m^{n+1} \quad = \quad v_m^n \qquad (5.32)$$

which can be solved from left to right:

$$v_m^{n+1} \quad = \quad \frac{v_m^n + 2b\mu v_{m-1}^{n+1}}{1+2b\mu} \qquad (5.33)$$

showing that v_m^{n+1} is a weighted average of v_m^n and v_{m-1}^{n+1} and therefore no larger than the largest of these, thus proving that a maximum principle holds for the numerical solution.

For other values of α we must solve a tridiagonal system of equations with
(cf. section 1.11)

$$a_m = -b\mu(1+\alpha), \quad b_m = 1+2b\mu, \quad c_m = -b\mu(1-\alpha).$$

This system can be solved using the procedure of section 1.11 with no fear of numerical instability since

$$b_2' = b_2 - \frac{a_2}{b_1}c_1 = 1+2b\mu - \frac{b\mu(1+\alpha)}{1+2b\mu}b\mu(1-\alpha) = 1+2b\mu - \frac{b\mu}{1+2b\mu}b\mu(1-\alpha^2).$$

If $\alpha \ge 1$ then $b_2' \ge b_2$ and in general $b_m' \ge b_m$.
If $\alpha < 1$ the $b_2' > 1+b\mu$ and in general $b_m' > 1+b\mu$.

5.6 Crank-Nicolson

The Crank-Nicolson method can be written

$$\begin{aligned}\frac{v_m^{n+1}-v_m^n}{k}-b\frac{v_{m+1}^{n+1}-2v_m^{n+1}+v_{m-1}^{n+1}}{2h^2}+a\frac{v_{m+1}^{n+1}-v_{m-1}^{n+1}}{4h} &= \\ b\frac{v_{m+1}^{n}-2v_m^{n}+v_{m-1}^{n}}{2h^2}-a\frac{v_{m+1}^{n}-v_{m-1}^{n}}{4h}\end{aligned} \tag{5.34}$$

or

$$\begin{aligned}-\frac{1}{2}b\mu(1+\alpha)v_{m-1}^{n+1}+(1+b\mu)v_m^{n+1}-\frac{1}{2}b\mu(1-\alpha)v_{m+1}^{n+1} &= \\ \frac{1}{2}b\mu(1+\alpha)v_{m-1}^{n+1}+(1-b\mu)v_m^{n+1}-\frac{1}{2}b\mu(1-\alpha)v_{m+1}^{n+1}.\end{aligned} \tag{5.35}$$

For the growth factor we now get

$$g(\varphi) = \frac{1-2b\mu\sin^2\frac{\varphi}{2}-\frac{1}{2}ia\lambda\sin\varphi}{1+2b\mu\sin^2\frac{\varphi}{2}+\frac{1}{2}ia\lambda\sin\varphi}, \qquad -\pi \le \varphi \le \pi \tag{5.36}$$

and it is easily seen that $|g(\varphi)| \le 1$, proving absolute and unconditional stability.

If $\alpha = 1$, i.e. $h = h_0$, then (5.35) reduces to

$$-b\mu v_{m-1}^{n+1}+(1+b\mu)v_m^{n+1} = b\mu v_{m-1}^n+(1-b\mu)v_m^n \tag{5.37}$$

which can be solved from left to right:

$$v_m^{n+1} = \frac{b\mu v_{m-1}^n+(1-b\mu)v_m^n+b\mu v_{m-1}^{n+1}}{1+b\mu} \tag{5.38}$$

showing that a maximum principle holds for the numerical solution if and only if $b\mu \le 1$ or $k \le 2k_0$.

For other values of α we must again solve a tridiagonal system of equations and also here there is no fear of numerical instability (cf. exercise 1).

5.7 Comparing the methods

We have compared the various methods on two test examples based on the equation $u_t - bu_{xx} + au_x = 0$ with $b = 0.1$ and $a \ge 0$ and with an initial function either a sawtooth which increases linearly from 0 to 1 on $[-1, 0]$ and decreases linearly to 0 on $[0, 1]$ or a smooth bump defined as $(1 + cos(\pi x))/2$ on $[-1, 1]$. Outside $[-1, 1]$ the initial function is set to 0 and the interval is chosen large enough that we can use 0 as boundary values. The time interval was chosen to $[0, 0.5]$.

As time passes the amount of matter represented by the bump will diffuse and be transported to the right with velocity a, but nothing will disappear so the area under the curve will remain constant. Numerically the area is defined as $h\sum_m v_m$ and it is easy to show that all methods considered here will conserve the area (up to rounding errors and as long as the bump does not reach the boundaries, cf. exercise 3).

Because of the diffusion term the maximum value of the bump will become smaller and the half-width (measured as the width of the bump at half the maximum value) will widen. Because of the transport term the top of the bump will move with a velocity close to a.

We have used x-steps around h_0 and time steps equal to k_0, $5k_0$, and $25k_0$. $a = 0$ was chosen to get a reference value for the maximum and the half-width and $a = 9$ for the real computation. With the smaller values of k Crank-Nicolson meets these values quite well. The implicit method tends to smoothe too much, and the explicit method too little. Time steps of $25k_0$ cannot be recommended as the implicit method gives a very wide and low maximum and Crank-Nicolson tends to produce waves with negative function values upstream. In no case does the upwind scheme produce acceptable results.

5.8 Exercises

1. Show that we have numerical stability when solving the linear equations for Crank-Nicolson's method by showing that $b'_m > 1 + \frac{1}{2}b\mu$.

2. Take a single step with the explicit method, the implicit method, and Crank-Nicolson with $h = h_0$ and $k = k_0, 2k_0$ and $3k_0$ from an initial function which is equal to 1 when $x \leq 0$ and equal to 0 when $x > 0$. Compare the results.

3. Show that the explicit and the implicit methods preserve the area under the curves in the two test examples in section 5.7.

Chapter 6

The Matrix Method

6.1 Notation

The process of advancing the solution from one time step to the next can be formulated in linear algebra terms. We arrange the function values at time step n, $\{v_0^n, v_1^n, \dots, v_{M-1}^n, v_M^n\}$ as an $(M+1)$-dimensional column vector, $\underline{v}^n$, and the internal function values $\{v_1^n, \dots, v_{M-1}^n\}$ as an $(M-1)$-dimensional column vector, v^n. We shall also use this underline convention for matrices such that a matrix name with no underline shall refer to an $(M-1) \times (M-1)$ matrix such as the one representing the operator which takes v^n into v^{n+1}, whereas a matrix name with an underline refers to an $(M+1) \times (M+1)$ matrix which also takes boundary values into account. A double underline shall signify a rectangular $(M-1) \times (M+1)$ matrix. We shall return to the precise definition of these matrices in the following sections.

It should be stressed here that we never in practice construct the matrices which we are about to introduce. They are used in the analysis of the computational process and serve merely as a means to study and understand the general behaviour of our numerical solutions.

6.2 The explicit method

A time step with the explicit method on the simple heat equation $u_t = bu_{xx}$ can be written

$$v^{n+1} = \underline{\underline{A}}\,\underline{v}^n \tag{6.1}$$

where

$$\underline{\underline{A}} \;=\; \left\{ \begin{array}{ccccccc} c & d & e & & & & \\ & c & d & e & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & c & d & e & \\ & & & & c & d & e \end{array} \right\} \tag{6.2}$$

with $c = e = b\mu$ and $d = 1 - 2b\mu$. We shall also write $\underline{\underline{A}} = \underline{\underline{I}} - b\mu\,\underline{\underline{T}}$ introducing

$$\underline{\underline{I}} = \left\{ \begin{array}{ccccccc} 0 & 1 & 0 & & & & \\ & 0 & 1 & 0 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 0 & 1 & 0 & \\ & & & & 0 & 1 & 0 \end{array} \right\} \text{ and } \underline{\underline{T}} = \left\{ \begin{array}{ccccccc} -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \end{array} \right\}. \tag{6.3}$$

We prefer to work with square matrices, so we usually treat the boundary values separately thus removing the first and last column from $\underline{\underline{A}}$ to form $A = I - b\mu\,T$ where

$$I \;=\; \left\{ \begin{array}{ccccc} 1 & 0 & & & \\ 0 & 1 & 0 & & \\ & \cdot & \cdot & \cdot & \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \end{array} \right\} \quad \text{and} \quad T \;=\; \left\{ \begin{array}{ccccc} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \cdot & \cdot & \cdot & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{array} \right\} \tag{6.4}$$

and the explicit time step now reads

$$v^{n+1} \;=\; Av^n + q^n \tag{6.5}$$

where q^n is the $(M-1)$-dimensional vector $q^n = b\mu\{v_0^n, 0, \ldots, 0, v_M^n\}^T$.

Remark. q^n is really $(M-1)$-dimensional. Component number 1 is $b\mu v_0^n$ and component number $M-1$ is $b\mu v_M^n$. □

Using relation (6.5) repeatedly from $n = 0$ we get

$$v^n \;=\; A^n v^0 + A^{n-1} q^0 + \cdots + q^{n-1}. \tag{6.6}$$

Remark. The superscripts on the vectors v and q are indices referring to the step number whereas the superscripts on A indicate powers of A. □

The behaviour of the solution at time n is thus to a large extent governed by the behaviour of the powers of matrix A. We shall return to this in section 6.6.

6.3 The implicit method

For the implicit method on $u_t = bu_{xx}$ we can in a similar way express the step from time n to time $n+1$ as

$$\underline{\underline{B}}\,\underline{v}^{n+1} \;=\; \underline{v}^n \tag{6.7}$$

where $\underline{\underline{B}} = \underline{\underline{I}} + b\,\mu\,\underline{\underline{T}}$.

In case of Dirichlet boundary conditions the known values of v_0^{n+1} and v_M^{n+1} can be inserted and we arrive at

$$Bv^{n+1} \;=\; v^n + q^n \tag{6.8}$$

where $B = I + b\,\mu\,T$ and q^n is the $(M-1)$-dimensional vector

$$q^n \;=\; b\mu\{v_0^{n+1}, 0, \ldots, 0, v_M^{n+1}\}^T.$$

Equation (6.8) can be reformulated as

$$v^{n+1} \;=\; Av^n + Aq^n \tag{6.9}$$

where $A = B^{-1}$.

6.4 The θ-method

The general θ-method can be formulated as

$$\underline{\underline{B}}\,\underline{v}^{n+1} \;=\; \underline{\underline{C}}\,\underline{v}^n \tag{6.10}$$

where $\underline{\underline{B}} = \underline{\underline{I}} + \theta\,b\,\mu\,\underline{\underline{T}}$ and $\underline{\underline{C}} = \underline{\underline{I}} - (1-\theta)\,b\,\mu\,\underline{\underline{T}}$.

Taking the (Dirichlet) boundary values separately we can remove the first and the last columns of $\underline{\underline{B}}$ and $\underline{\underline{C}}$ and arrive at

$$Bv^{n+1} \;=\; Cv^n + q^n \tag{6.11}$$

where $B = I + \theta\,b\,\mu\,T$ and $C = I - (1-\theta)\,b\,\mu\,T$ and q^n is the $(M-1)$-dimensional vector

$$q^n = b\mu\{\theta v_0^{n+1} + (1-\theta)v_0^n, 0, \ldots, 0, \theta v_M^{n+1} + (1-\theta)v_M^n\}^T$$

or

$$v^{n+1} \;=\; Av^n + B^{-1}q^n \tag{6.12}$$

where $A = B^{-1}C$.

6.5 Stability by the matrix method

If we have homogeneous boundary conditions we have $q^n = 0$ and the transformation from time step n to $n+1$ is in all cases

$$v^{n+1} = Av^n \tag{6.13}$$

and from the beginning to time step n

$$v^n = A^n v^0 \tag{6.14}$$

where the superscript on v indicates the step number and the superscript on A indicates a power.

Introducing a vector norm, $||.||$, and a compatible matrix norm (for which we shall use the same symbol) we then have

$$||v^{n+1}|| \leq ||A||\,||v^n|| \tag{6.15}$$

and

$$||v^n|| \leq ||A||^n\,||v^0||. \tag{6.16}$$

We note that $||A|| \leq 1$ implies absolute stability in the given vector norm.

Example. If we choose the ∞-norm for vectors

$$||v||_\infty = \max_m |v_m|$$

a compatible matrix norm is the maximum row sum

$$||A||_\infty = \max_i \sum_j |a_{ij}|.$$

For the explicit method we have

$$|c| + |d| + |e| = 2b\mu + |1 - 2b\mu| = 1 \qquad \text{if} \qquad 2b\mu \leq 1.$$

We therefore conclude that the explicit method on $u_t = bu_{xx}$ is absolutely stable in the ∞-norm provided $2b\mu \leq 1$. □

There are many different norms to choose from. What happens if one matrix norm measures $||A||$ less than 1 and another greater than 1? Since any two vector norms in a finite-dimensional space are equivalent in the sense that there exist constants γ and δ such that

$$\gamma||v||_\alpha \leq ||v||_\beta \leq \delta||v||_\alpha \tag{6.17}$$

it follows that if a scheme is 0-stable in one norm it is 0-stable in any other norm.

Remark. But it is not necessarily true that a scheme is *absolutely* stable in one norm if it is *absolutely* stable in another since the constants γ and δ allow for a (limited) growth (cf. Exercise 1). □

So it becomes interesting to look for matrix norms which produce small values when applied to matrix A. And to search for the smallest possible value of $||A||$. We have the following two results from matrix theory:

1.

$$||A|| \geq \rho(A)$$

where $||.||$ is any matrix norm and $\rho(A)$ is the *spectral radius* of A, i.e. the maximum absolute value of the eigenvalues of A. □

2. For any matrix A and any $\varepsilon > 0$ there is a matrix norm such that

$$||A|| \leq \rho(A) + \varepsilon$$

For a proof of **2** see [34, p. 284]. □

As a consequence we have the following stability results:

A. $\rho(A) < 1 \Longrightarrow$ 0-stability.

B. $\rho(A) > 1 \Longrightarrow$ instability.

C. $\rho(A) \leq 1$ and A symmetric $\Longrightarrow$ absolute stability in the 2-norm.

Remark. If A is not symmetric and $\rho(A) = 1$ the situation is undecided. □

Remark. When $\rho(A) < 1$ we might still encounter considerable (although bounded) error growth when measured in one of the norms we should like to use such as the 2-norm or the ∞-norm. □

6.6 Eigenvalues of tridiagonal matrices

In order to investigate the stability of numerical solutions of $u_t = bu_{xx}$ it is thus important to study the eigenvalues and eigenvectors of Toeplitz matrices of the form $A = I + \alpha T$ of dimension $M - 1$. If w is an eigenvector with corresponding eigenvalue τ for T then w is also an eigenvector for A with corresponding eigenvalue $\lambda = 1 + \alpha\tau$.

For a vector $w = \{w_1, w_2, \ldots, w_{M-1}\}$ to be an eigenvector for T with corresponding eigenvalue τ we must have

$$-w_{m-1} + 2w_m - w_{m+1} \;\; = \;\; \tau w_m, \qquad m = 2, \ldots, M-2. \qquad (6.18)$$

It is usually not easy to find eigenvalues and eigenvectors for a matrix, but if we have a candidate then it is very easy to check whether it fits. A good suggestion for w is to take

$$w_m = \sin(m\varphi), \qquad m = 1, \ldots, M-1. \tag{6.19}$$

Since

$$\begin{aligned} w_{m-1} + w_{m+1} &= \sin(m-1)\varphi + \sin(m+1)\varphi \\ &= 2\sin(m\varphi)\cos\varphi \; = \; 2w_m \cos\varphi \end{aligned} \tag{6.20}$$

(6.18) now gives

$$\tau = 2 - 2\cos\varphi = 4\sin^2\frac{\varphi}{2}. \tag{6.21}$$

In addition to the $M-3$ equations (6.18) we must also have the similar relations for $m = 1$ and $m = M-1$:

$$\begin{aligned} 2w_1 - w_2 &= \tau w_1, \\ -w_{M-2} + 2w_{M-1} &= \tau w_{M-1}. \end{aligned}$$

These are fulfilled automatically if we can manage to have $w_0 = w_M = 0$.
$w_0 = 0$ comes naturally out of (6.19). For w_M we must require

$$w_M = \sin(M\varphi) = 0$$

or

$$M\varphi = p\pi, \qquad p = 1, 2, \ldots \tag{6.22}$$

We therefore define

$$\varphi_p = \frac{p\pi}{M}, \qquad p = 1, 2, \ldots, M-1 \tag{6.23}$$

and with these $M-1$ values of φ we have a set of $M-1$ orthogonal eigenvectors and corresponding eigenvalues for T:

$$\tau_p = 4\sin^2\frac{p\pi}{2M}, \qquad p = 1, 2, \ldots, M-1. \tag{6.24}$$

For the explicit method on $u_t = bu_{xx}$ stability is governed by the eigenvalues of matrix $A = I - b\,\mu\,T$ such that

$$\lambda_p = 1 - 4b\mu\sin^2\frac{\varphi_p}{2}, \qquad p = 1, 2, \ldots, M-1. \tag{6.25}$$

and the condition for stability is that all eigenvalues are ≤ 1 in absolute magnitude.

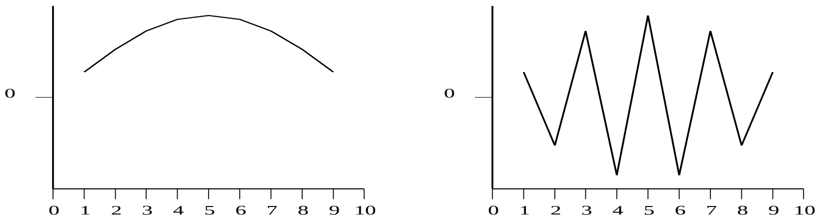


Figure 6.1: Eigenvectors corresponding to $p = 1$ and $p = M-1 = 9$

Notice the close similarity between (6.25) and (2.22). The extra information we get out of (6.25) is that with a given choice of h (or M) only a discrete and finite set of frequencies, φ_p, are applicable. In Fig. 6.1 we show the components of the eigenvectors corresponding to the lowest ($p = 1$) and the highest ($p = 9$) frequency for the case $M = 10$.

Remark. We have absolute stability iff $|\lambda_p| \leq 1$, $p = 1, 2, \ldots, M-1$. If we choose $M = 10$, $h = 0.1$ then $\varphi_9 = 9\pi/10$ and $\sin^2(\varphi_9/2) \approx 0.97553$. If $b = 1$ we can actually have absolute stability with $k = 0.005125$, even though $b\mu = 0.5125 > 0.5$.
But for the explicit method to be absolutely stable for *arbitrary* h (or M) we must still require $b\mu \leq 0.5$. □

For the implicit method on $u_t = bu_{xx}$ the eigenvalues of $A = B^{-1}$ are the reciprocals of the eigenvalues of matrix $B = I + b\,\mu\,T$. We therefore have

$$\lambda_p \;=\; \frac{1}{1 + 4b\mu \sin^2 \frac{\varphi_p}{2}}, \quad p = 1, 2, \ldots, M-1. \tag{6.26}$$

Once again there is a close similarity between (6.26) and (2.29) (with $a = \kappa = 0$) with the former emphasizing the discrete nature of a finite-dimensional problem.

For the general θ-method we use the fact that the eigenvectors for matrices B and C in (6.11) are the same, such that the eigenvalues of $A = B^{-1}C$ in (6.12) are the ratios of corresponding eigenvalues from C and B.

If we apply the explicit method on $u_t = bu_{xx} - au_x$ then matrix A has components $c = b\mu + \frac{1}{2}ak/h$, $e = b\mu - \frac{1}{2}ak/h$, and $d = 1 - 2b\mu$ (cf. (6.2)). The matrix is no longer symmetric, but as long as $e > 0$, i.e. $a < \frac{2b}{h}$ then c and e have the same sign and A has real eigenvalues which are given by:

$$\lambda_p \;=\; 1 - 2b\mu + 2\sqrt{ce}\cos\varphi_p. \tag{6.27}$$

The condition on a is equivalent to (5.18) which together with (5.13) secures a discrete maximum principle in accordance with the fact that λ_p in (6.27) are less than 1 in absolute magnitude.

Remark. But the analogy with the von Neumann results is not complete since the eigenvalues of A remain real for $0 < a < 2b/h$ where the growth factors are complex. □

6.7 The influence of boundary values

The stability considerations so far have disregarded the boundary conditions (or assumed them to be homogeneous) and can therefore in certain cases be slightly deceiving. One might for instance get the impression that all numerical solutions to the simple heat equation are decreasing in time if the scheme is stable.

To study the effect of Dirichlet boundary conditions on the explicit scheme for the simple heat equation we first look at equation (6.6) which can be interpreted to state that the effect of the initial function will diminish in time. If the boundary function is increasing with time then the effect of the latter terms will tend to dominate the expression for v^n.

To put it in matrix terms we add two rows to the rectangular matrix $\underline{\underline{A}}$ to form the quadratic matrix

$$\underline{A} \;=\; \left\{ \begin{array}{ccccccc} a_{00}^n & & & & & & \\ c & d & e & & & & \\ & c & d & e & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & c & d & e & \\ & & & & c & d & e \\ & & & & & & a_{MM}^n \end{array} \right\} \tag{6.28}$$

where c, d, and e are as in (6.2) and $a_{00}^n = v_0^{n+1}/v_0^n$ and $a_{MM}^n = v_M^{n+1}/v_M^n$. The step from time n to $n+1$ now reads

$$\underline{v}^{n+1} \;=\; \underline{A}\,\underline{v}^n \tag{6.29}$$

Remark. Since the first and last component of $\underline{A}$ depend on n there really ought to be a superscript, n, on $\underline{A}$. We have chosen to omit it because superscripts on matrices indicate powers of the matrix. □

Remark. If we have Dirichlet boundary conditions then we know the values of v_0 and v_M. If these values happen to be 0 at one or more points this analysis must be modified. □

Since we know the eigenvalues and eigenvectors of matrix A it is easy to find the eigenvalues and eigenvectors of $\underline{A}$. If we augment the previous eigenvectors with $w_0 = w_M = 0$ we have a set of $M-1$ eigenvectors and corresponding eigenvalues for $\underline{A}$. The remaining two eigenvalues for $\underline{A}$ are a_{00} and a_{MM} and the corresponding two eigenvectors are shown graphically in Fig. 6.2 for the case $M = 10$. The main observation here is that when v_0^n is increasing with n the eigenvalue a_{00} is greater than 1, meaning that we have an increasing solution component, and from Fig. 6.2 we see that the effect at the internal points diminishes as we get further into the region.

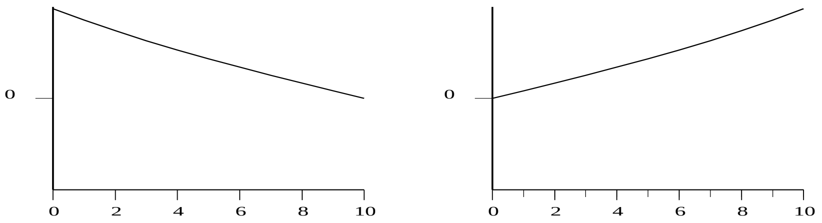


Figure 6.2: Eigenvectors corresponding to a_{00} and a_{MM} for $M = 10$.

Remark. The above analysis holds only for exponentially increasing boundary functions such that a_{00} is constant throughout the computation. But it gives an idea of the kind of linear transformation which takes the numerical solution from one time step to the next. □

6.8 A derivative boundary condition

When the boundary condition involves a derivative we use one of the formulas from Chapter 4 to give the necessary extra equation for v_0^{n+1} and/or v_M^{n+1}.

As a simple example we consider the Neumann condition $u_x = 0$ at the left boundary and assume that we want to use the implicit method. If we use the first order approximation to the derivative we get $v_0^{n+1} = v_1^{n+1}$ which can be used in the first equation which then reads

$$(1 + b\mu)v_1^{n+1} - b\mu v_2^{n+1} \;=\; v_1^n$$

We see that that matrix B is changed to $B' = I - b\,\mu\,T'$ where T' is obtained from T by changing the first diagonal element from 2 to 1. T' is still symmetric and therefore has real eigenvalues which by Gerschgorin's theorem still lie in the interval $(0, 4)$.

If we instead use the symmetric second order approximation we get $v_{-1}^{n+1} = v_1^{n+1}$ which can be used to eliminate v_{-1}^{n+1} from the equation at $m = 0$ which now reads

$$(1 + 2b\mu)v_0^{n+1} - 2b\mu v_1^{n+1} \;=\; v_0^n$$

Now the transformation matrix is $B'' = I - b\,\mu\,T''$ where T'' (which now has dimension M) is obtained from T by changing the second element in the first row from -1 to -2. T'' is no longer symmetric but since it is similar to a symmetric matrix (by the similarity transformation using $D = \text{diag}\{\sqrt{2}, 1, 1, \ldots\}$) the eigenvalues are still real, and by Gershgorin's theorem they are also still in the interval $(0, 4)$.

For a further analysis of problems involving the general θ-method and general boundary conditions we refer to [25] and [18].

6.9 Exercises

1. Prove that if a scheme is 0-stable in the α-norm then it is 0-stable in the β-norm (cf. (6.17)).
   If the scheme is absolutely stable in the α-norm how much can $||v^n||_\beta$ grow.

2. Prove that $||A|| \geq \rho(A)$ for any matrix norm.

3. Prove that $2b\mu \leq 1$ and $a < 2b/h$ imply $|\lambda_p| \leq 1$ for λ_p in (6.27).

4. Show that the (skewsymmetric) matrix with 1 in the upper bidiagonal, -1 in the lower bidiagonal, and 0 everywhere else, has an eigenvalue

   $$\lambda_p \;=\; 2\,i\;\cos\varphi_p$$

   corresponding to the eigenvector with components

   $$w_k \;=\; i^{k+1}\;\sin(k\varphi_p)$$

   with φ_p given by (6.23).

Chapter 7

Two-step Methods

All the methods we have considered so far have been *one-step methods* in the sense that they take information from one time step in order to produce values for the succeeding time step. Many of these methods are second order accurate in space but only first order accurate in time.

7.1 The central-time central-space scheme

In order to balance things better we might consider the scheme

$$\tilde{\mu}\delta_t v_m^n - b\,\delta^2 v_m^n \quad = \quad 0 \tag{7.1}$$

or written out

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} \quad = \quad b\,\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} \tag{7.2}$$

or

$$v_m^{n+1} - v_m^{n-1} - 2b\mu(v_{m+1}^n - 2v_m^n + v_{m-1}^n) \quad = \quad 0. \tag{7.3}$$

This scheme which is also known as *leap-frog* has been used by Richardson [29] and is expected to be second order accurate in both space and time (cf. Exercise 1). It will require a special starting procedure since we only have information available at one beginning time level.

We shall begin with an analysis of the stability properties of the scheme. We proceed as in section 2.4 using the Fourier inversion formula (2.16) on (7.3) obtaining

$$\int_{-\pi/h}^{\pi/h} e^{imh\xi}\left[\hat{v}^{n+1}(\xi) - \hat{v}^{n-1}(\xi) - 2b\mu(e^{ih\xi} - 2 + e^{-ih\xi})\hat{v}^n(\xi)\right]\,d\xi \quad = \quad 0. \tag{7.4}$$

By uniqueness of the Fourier transform the integrand must be 0 leading to

$$\hat{v}^{n+1}(\xi) - \hat{v}^{n-1}(\xi) + 8b\mu \sin^2 \frac{h\xi}{2} \hat{v}^n(\xi) \;=\; 0. \tag{7.5}$$

This reminds us very much of a second order difference equation and we are therefore lead to suggest a solution of the form

$$\hat{v}^n(\xi) \;=\; g^n \tag{7.6}$$

where the right-hand-side is the n-th power of the growth factor g which is supposed to be a function of $\varphi = h\xi$. Inserting in (7.5) and dividing by g^{n-1} we arrive at the quadratic equation.

$$g^2 + 8b\mu \sin^2 \frac{\varphi}{2}\, g - 1 \;=\; 0. \tag{7.7}$$

We note that the suggestion on page 25 on how to avoid invoking the Fourier transform by expressing v_m^n as $g^n e^{im\varphi}$ also applies in the two-step case and leads us directly to (7.7).

The main difference is that for a two-step scheme we have two growth factors and they must both be ≤ 1 or $1 + O(k)$ for the method to be (0-)stable. The two growth factors are the two roots of the quadratic (7.7):

$$g \;=\; -4b\mu \sin^2 \frac{\varphi}{2} \pm \sqrt{1 + 16b^2\mu^2 \sin^4 \frac{\varphi}{2}}. \tag{7.8}$$

The two values for g in (7.8) are real and their product is -1. Therefore if one root is less than 1 in absolute magnitude the other must be larger than 1. The only exception is for $\varphi = 0$ where the roots are $+1$ and -1. We conclude that the difference scheme is always unstable and therefore not useful in practice.

Remark. Richardson used the method with success in [29] but only for a very small number of time steps. Since the high frequency components are spawned from rounding errors they are very small in the beginning and it takes a number of steps for them to build up to an appreciable size. □

7.2 The DuFort-Frankel scheme

In order to remedy this lack of stability, DuFort and Frankel [11] suggested replacing v_m^n in (7.2) by an average leading to

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} \;=\; b\, \frac{v_{m+1}^n - v_m^{n+1} - v_m^{n-1} + v_{m-1}^n}{h^2} \tag{7.9}$$

or

$$(1+2b\mu)v_m^{n+1} - 2b\mu(v_{m+1}^n + v_{m-1}^n) - (1-2b\mu)v_m^{n-1} \;=\; 0. \qquad (7.10)$$

Replacing v_m^n by $g^n e^{im\varphi}$ and dividing by $g^{n-1}e^{im\varphi}$ we get

$$(1+2b\mu)g^2 - 4b\mu\cos\varphi\, g - (1-2b\mu) \;=\; 0 \qquad (7.11)$$

with roots

$$\begin{aligned} g &= \frac{4b\mu\cos\varphi \pm \sqrt{16b^2\mu^2\cos^2\varphi + 4(1-4b^2\mu^2)}}{2(1+2b\mu)} \qquad (7.12)\\ &= \frac{2b\mu\cos\varphi \pm \sqrt{1-4b^2\mu^2\sin^2\varphi}}{1+2b\mu}. \end{aligned}$$

If the discriminant is negative the product of the (absolute values of the complex conjugate) roots is $(1-2b\mu)/(1+2b\mu)$ which is less than 1 in absolute magnitude implying the same for the roots. If the discriminant is positive the roots are real and satisfy

$$|g| \le \frac{2b\mu|\cos\varphi| + 1}{1+2b\mu} \le \frac{1+2b\mu}{1+2b\mu} = 1$$

so we conclude that the DuFort-Frankel scheme is unconditionally stable. This is unusual for an explicit scheme so something else must be wrong.

In order to check the accuracy of the scheme we determine the symbol of the difference operator. So we write $e^{snk}e^{imh\xi}$ for v_m^n in (7.9) and get after division

$$p_{k,h}(s,\xi) = \frac{e^{sk} - e^{-sk}}{2k} - \frac{b}{h^2}(e^{i\xi h} + e^{-i\xi h} - (e^{sk} + e^{-sk}))$$

$$= s + \frac{1}{6}s^3k^2 + O(k^4) - \frac{b}{h^2}(2 - \xi^2h^2 + \frac{1}{12}\xi^4h^4 + O(h^6) - (2 + s^2k^2 + \frac{1}{12}s^4k^4 + O(k^6)))$$

$$= s + b\xi^2 + \frac{1}{6}s^3k^2 - \frac{1}{12}b\xi^4h^2 + bs^2\frac{k^2}{h^2} + O(h^4 + k^4 + \frac{k^4}{h^2}).$$

We recognize the symbol of $u_t - bu_{xx}$ in the first two terms. For the scheme to be consistent the remaining terms must tend to 0, but for this to happen k must tend to 0 faster than h. In particular, if $k = \alpha h^2$ then the scheme is second order accurate but then we are back to restrictions on the time step which are comparable to those for the explicit method.

7.3 Exercise

1. Show that the central-time central-space scheme is second order accurate in both time and space.

Chapter 8

Discontinuities

The solutions to the simple heat equation and related parabolic equations are smooth, i.e. infinitely often differentiable in the interior of the region where the equation applies. But it happens frequently in the mathematical model that there is a jump discontinuity between the initial function and a boundary function at a corner point such as $(0, X_1)$ or $(0, X_2)$, or that there is a discontinuity in the initial function or in one of its derivatives (cf. problem 2 on page 5). If we exclude a small neighbourhood around the singular point(s) the solution will still be smooth in the remaining region. But our numerical methods may respond in various ways to such discontinuities.

8.1 Stability and damping

When we study absolute stability our principal object is that the growth factor satisfies $|g| \leq 1$ and we are not particularly concerned about whether g is positive or negative or close to $+1$ or -1. This is fine for smooth initial conditions where our main concern is that the (small) errors we commit, be they rounding or truncation errors, stay small. Rounding errors often have considerable high-frequency parts but since their absolute magnitude is small we can even allow a certain growth as long as it is bounded. The situation is different with a discontinuous initial function where significant high-frequency components are present from the beginning. In the continuous problem these components are damped effectively. The higher the frequency the more effective the damping. Not necessarily so for the numerical schemes which we shall study in the following sections when applied to the simple heat equation $u_t = bu_{xx}$.

8.2 The growth factor

Recall that the growth factor (damping factor might be a more appropriate name here) for the explicit method is

$$g(\varphi) \quad = \quad 1 - 4b\mu \sin^2 \frac{\varphi}{2}, \qquad\qquad -\pi \leq \varphi \leq \pi. \tag{8.1}$$

The explicit method is absolutely stable when $2b\mu \leq 1$ because we then have $|g(\varphi)| \leq 1$ for all φ. But for high frequency components φ is close to π, and if $2b\mu = 1$ then g will be negative and close to -1. These components will give rise to slowly damped oscillations in time. Because of the discrete nature of the problem for a given choice of step sizes there is a limit to how close φ can get to π (cf. section 6.5) and there is a guaranteed minimum damping per time step. And since there is a strict bound on the time step, the oscillations will usually not be a serious problem.

The growth factor for the implicit method **(IM)** is

$$g(\varphi) \quad = \quad \frac{1}{1 + 4b\mu \sin^2 \frac{\varphi}{2}}, \qquad\qquad -\pi \leq \varphi \leq \pi \tag{8.2}$$

and we immediately notice that $0 \leq g \leq 1$ independently of $b\mu$ and φ. We see furthermore that g becomes smaller when φ approaches π so we are in the ideal situation where the numerical method mimics the continuous case rather closely. But the implicit method is only first order accurate in time and although the solution *looks* good and smooth it might be rather inaccurate.

For Crank-Nicolson **(CN)** the growth factor is

$$g(\varphi) \quad = \quad \frac{1 - 2b\mu \sin^2 \frac{\varphi}{2}}{1 + 2b\mu \sin^2 \frac{\varphi}{2}}, \qquad\qquad -\pi \leq \varphi \leq \pi \tag{8.3}$$

and $|g| \leq 1$ for all values of $b\mu$ and φ. If $k \approx h$ then $b\mu$ may be rather large and when $\varphi \approx \pi$, g can be very close to -1. The consequence is that high frequency components are damped very slowly and we observe oscillations in time at certain points in space. Crank-Nicolson produces solutions which are quite accurate when measured in the 2-norm, but these oscillations which occur near the points of discontinuity can be rather annoying as also noted by Wood and Lewis [38]. We shall in the next sections discuss ways of damping these oscillations.

8.3 Reducing the Crank-Nicolson oscillations

8.3.1 AV – the moving average

A device for coping with damped oscillations known from physics and used by Lindberg [23] for a system of ordinary differential equations is the moving average. If the numerical solution, v_m^n is oscillating in time then we might instead use

$$w_m^n = \frac{v_m^{n+1} + 2v_m^n + v_m^{n-1}}{4}. \tag{8.4}$$

as an approximation to the solution at $(t,x) = (nk, mh)$. A main difference between our approach and Lindberg's is that he proposes to continue the calculations based on the average. If (8.4) is used in connection with Crank-Nicolson the growth factor becomes

$$g_{av} = \frac{g + 2 + g^{-1}}{4} = \frac{1}{1 - 4b^2\mu^2 \sin^4 \frac{\varphi}{2}} \tag{8.5}$$

indicating that this method must never be used when $b\mu$ is small but that it will have good performance for oscillatory components when $b\mu$ is large. It is also seen that we may encounter difficulties with small (or even zero) values in the denominator for small values of φ. These unfavourable growth phenomena for slowly varying components is our reason for not continuing the calculations with the average value. Instead we propose to compute with a straight Crank-Nicolson to the end (and one step beyond) and perform the average only at the points where a solution value shall be recorded.

Table 8.1: Growth factors for IM, CN, and AV at $\varphi = \pi$ and various $b\mu$.

$b\mu$	IM	CN	AV
0.1	0.7143	0.6667	1.0417
0.5	0.3333	0.0000	–
1	0.2000	−0.3333	−0.3333
10	0.0244	−0.9048	−0.0025
100	0.0025	−0.9900	−0.000025

In Table 8.1 we have given the growth factors for the implicit method, Crank-Nicolson, and the moving average method for $\varphi = \pi$ and for various values of $b\mu$ illustrating the good damping effect of **AV** for large values of $b\mu$ and the possible problems for small values of $b\mu$ and/or small values of φ.

8.3.2 IM1 – One step with IM

As $b\mu$ gets bigger the high-frequency components receive less and less damping from CN but more and more from IM as seen in Table 8.1. At $b\mu = 100$ the damping is only 1% per Crank-Nicolson step but one step with the implicit method will reduce the amplitude of the high frequency component by 0.0025. Furthermore since the local error of the implicit method is second order in time (cf. section 3.4) a single step with IM (one ping only) will not affect the second order accuracy of CN (cf. section 9.8). It will affect the magnitude of the global truncation error so it is a matter of balancing the effects.

8.3.3 SM – Small steps at the beginning

As seen in Table 8.1 Crank-Nicolson itself can eliminate high frequency components if $b\mu$ is small enough. So we propose an initial time step k_1 such that the corresponding $b\mu_1$ becomes equal to 0.5 and the high frequency component is annihilated altogether. In practice we should not expect a dramatic effect since there are also other solution components corresponding to values of φ smaller than π and these will not be reduced to zero. We might therefore consider taking more than one small step, say s small steps where s could be 5 or 10 or 20. In this way other solution components will be reduced by the appropriate growth factor raised to the s-th power. In order to get back to the 'normal' step size, k, we may have to take an extra step of length $k - sk_1$.

8.3.4 Pearson

It is not necessary to aim at a complete annihilation of the oscillations in one step. If the first step is subdivided into s equal steps of length $k_1 = k/s$ as suggested by Pearson [28] then the cumulative damping will be $g^* = g^s$ and a larger value of $b\mu_1$ such as 2 or 5 is acceptable.

8.3.5 EI – Exponentially increasing steps

One problem with both **SM** and **Pearson** is that the change in time step from k_1 to k may itself produce unwanted high frequency effects. We might therefore suggest another way of subdividing the first interval, namely by exponentially increasing subintervals (cf. [6]) where the subintervals are given by $k_i = \beta k_{i-1}$, $i = 2, \ldots, s$ for some $\beta > 1$ and with $\sum_1^s k_i = k$. This gives a smoother transition from the subintervals to the regular intervals, especially when β is large, i.e. near 2. The **Pearson** method can be viewed as a special case when $\beta = 1$.

8.4 Discussion

We refer the reader to [41] for a more detailed treatment and comparison of the five different proposals. Here we shall just summarize the results by mentioning that **AV** and **IM1** are very economical but also limited in how large a reduction of the oscillations they can achieve. Furthermore they perform worse for other than the highest frequency component and therefore show results which are poorer than what theory predicts. If the reduction achieved with **AV** or **IM1** is not sufficient we must resort to one of the other three methods where any amount of reduction is theoretically possible but at a higher computational expense. The latter two methods perform better for lower frequency components and are therefore better than what theory predicts.

The implicit method avoids the problem with oscillations completely but was ruled out because it is only first order accurate in time. Using the extrapolation techniques of Chapter 10 we can raise the order of the implicit method and thus get the better of both worlds as reported in [20] and [12].

8.5 A discontinuous corner

Consider the following example

$$\begin{array}{rcllr} u_t & = & u_{xx}, & 0 \le x \le 1, & t > 0, \\ u(0,x) & = & 1, & 0 \le x \le 1, & \\ u(t,0) & = & 0, & & t > 0, \\ u(t,1) & = & 1, & & t > 0. \end{array}$$

There is a jump discontinuity between the initial value and the boundary value at (0,0), and we are faced with a decision about which value to choose for v_0^0, a decision which depends on which numerical method we have chosen.

If we use the explicit method the corner point enters in a second difference of initial value points, and it would seem natural to use the initial value also at the corner point, but in fact this only delays the effect one time step.

If we use the implicit method the corner point is not used at all and the problem disappears. This makes the implicit method seem like an ideal choice despite the fact that it is only first order accurate in time.

The behaviour of Crank-Nicolson is studied in exercise 1. With the considerations of the previous sections in mind the best option might be to begin with one implicit step and then switch to Crank-Nicolson in order to keep the overall second order accuracy.

8.6 Exercises

1. Solve the problem in the previous example with the implicit method and Crank-Nicolson (with various choices of v_0^0) with $h = k = 0.1$ up to $t = 0.5$.
2. Solve problem 2 with $h = k = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$ using Crank-Nicolson and **AV** and **IM1** (cf. section 8.3).
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1$, 0.2, 0.3, 0.4, 0.5.

Chapter 9

The Global Error – Theoretical Aspects

9.1 The local error

Information about the error of a finite difference scheme for solving a partial differential equation is often stated in terms of the *local error* which is the error committed in one step given correct starting values, or more frequently as the *local truncation error* expressed in terms of a Taylor expansion, again for a single step and with presumed correct starting values. Rather than giving numerical values one often resorts to giving the *order* of the scheme in terms of the step size such as $O(h)$ or $O(h^2)$. The interesting issues, however, are the magnitude of the *global error*, i.e. the difference between the true solution and the computed value at a specified point, in a sense the cumulated value of all the local errors up to this point, and the order of this error in terms of the step size used.

9.2 The global error

We study the linear, parabolic equation

$$u_t \;=\; bu_{xx} - au_x + \kappa u + \nu \tag{9.1}$$

or as we prefer to write it here

$$Pu \;=\; u_t - bu_{xx} + au_x - \kappa u \;=\; \nu \tag{9.2}$$

using the partial differential operator P. The coefficients b, a, κ, and ν may depend on t and x. We produce a numerical solution $v(t, x)$ and our basic assumption is that the global error can be expressed in terms of a series expansion in the step sizes k and h

$$v(t,x) \;=\; u(t,x) - hc - kd - hke - h^2 f - k^2 g - \cdots \tag{9.3}$$

The auxiliary functions c, d, e, f, and g are functions of t and x but do not depend on the step sizes h and k. They need not all be present in any particular situation. Often we shall observe that c or e or d are identically zero such that the numerical solution is second order accurate in one or both of the step sizes.

Strictly speaking $v(t,x)$ is only defined on a discrete set of grid points but it is possible to extend it in a differentiable manner to the whole region. Actually this can be done in many ways. The same considerations apply to the auxiliary functions and we shall in the following see a concrete way of extending these.

The formula (9.3) expresses an assumption or a hypothesis and as such can not be proved, but it leads to predictions which can be verified computationally, thereby identifying those situations where the hypothesis can be assumed to hold. The hypothesis expresses the notion that the computed solution contains information, not only about the true solution, but also about the truncation error.

We can get information on the auxiliary functions by studying the difference equations and by using Taylor expansions. We first look at the explicit scheme.

9.3 The explicit method

We use the difference operators from section 1.7

$$\delta^2 v_m^n \;=\; \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}, \tag{9.4}$$

$$\tilde{\mu}\delta v_m^n \;=\; \frac{v_{m+1}^n - v_{m-1}^n}{2h}. \tag{9.5}$$

The explicit scheme for (9.2) can now be written as

$$\frac{v_m^{n+1} - v_m^n}{k} - b_m^n \delta^2 v_m^n + a_m^n \tilde{\mu}\delta v_m^n - \kappa_m^n v_m^n \;=\; \nu_m^n. \tag{9.6}$$

We apply (9.3) and Taylor expand around $(nk, X_1 + mh)$:

$$\begin{aligned}
\frac{v_m^{n+1} - v_m^n}{k} \;=\;& u_t + \frac{1}{2}ku_{tt} + \frac{1}{6}k^2 u_{ttt} - hc_t - \frac{1}{2}hkc_{tt} - kd_t - \frac{1}{2}k^2 d_{tt} \\
&- hke_t - h^2 f_t - k^2 g_t + O(k^3 + k^2 h + kh^2 + h^3) \qquad (9.7)\\
\delta^2 v_m^n \;=\;& u_{xx} + \frac{1}{12}h^2 u_{4x} - hc_{xx} - kd_{xx} - hke_{xx} \qquad (9.8)\\
&- h^2 f_{xx} - k^2 g_{xx} + O(\cdots)
\end{aligned}$$

$$\begin{aligned}\tilde{\mu}\delta v_m^n &= u_x + \frac{1}{6}h^2 u_{xxx} - hc_x - kd_x - hke_x \qquad (9.9)\\ &\quad - h^2 f_x - k^2 g_x + O(\cdots)\\ v_m^n &= u - hc - kd - hke - h^2 f - k^2 g + O(\cdots) \qquad (9.10)\end{aligned}$$

We insert (9.7) – (9.10) in (9.6) and equate terms with the same powers of h and k:

$$\begin{aligned}1: &\qquad Pu = \nu \qquad (9.11)\\ h: &\qquad Pc = 0 \qquad (9.12)\\ k: &\qquad Pd = \frac{1}{2}u_{tt} \qquad (9.13)\\ hk: &\qquad Pe = -\frac{1}{2}c_{tt} \qquad (9.14)\\ h^2: &\qquad Pf = -\frac{1}{12}bu_{4x} + \frac{1}{6}au_{xxx} \qquad (9.15)\\ k^2: &\qquad Pg = \frac{1}{6}u_{ttt} - \frac{1}{2}d_{tt} \qquad (9.16)\end{aligned}$$

The first thing we notice is that we in (9.11) recover the original equation for u indicating that the difference scheme (and our assumption (9.3)) is consistent. The auxiliary functions are actually only defined on the grid points but inspired by (9.12) – (9.16) it seems natural to extend them between the gridpoints such that these differential equations are satisfied at all points in the region. We note that each of the auxiliary functions should satisfy a differential equation very similar to the original one, the only difference lying in the inhomogeneous terms.

9.4 The initial condition

In order to secure a unique solution to (9.1) we must impose some side conditions. One of these is an initial condition, typically of the form

$$u(0,x) = u_0(x), \qquad X_1 \le x \le X_2, \qquad (9.17)$$

where $u_0(x)$ is a given function of x. It is natural to expect that we start our numerical solution as accurately as possible, i.e. we set $v_m^0 = v(0, X_1 + mh) = u_0(X_1 + mh)$ for all grid points between X_1 and X_2. But we would like to extend v between grid points as well, and the natural thing would be to set $v(0,x) = u_0(x), X_1 \le x \le X_2$. With this assumption we see from (9.3) that

$$c(0,x) = d(0,x) = e(0,x) = f(0,x) = g(0,x) = \cdots = 0, \quad X_1 \le x \le X_2. \qquad (9.18)$$

In section 8.3.2 we discussed the positive effects on discontinuities which we can achieve by using the implicit method for one step and then switch to Crank-Nicolson. To study the effects on accuracy we note that we shall be solving with Crank-Nicolson for $t > k$ with an initial condition at $t = k$ given by the values from the implicit method. Since the local error for the implicit method is $O(k^2 + kh^2)$ (cf. section 3.4) we have

$$c(k,x) = d(k,x) = e(k,x) = f(k,x) = 0, \qquad X_1 \le x \le X_2 \tag{9.19}$$

in addition to a nonzero value for $g(k,x)$. As we shall see in section 9.8 this has no effect on the order of the global error for Crank-Nicolson since the differential equation (9.41) for g is inhomogeneous anyway. It might have an effect on the magnitude of the global error though.

9.5 Dirichlet boundary conditions

In order to secure uniqueness we must in addition to the initial condition impose two boundary conditions which could look like

$$u(t, X_1) = u_1(t), \qquad u(t, X_2) = u_2(t), \qquad t > 0, \tag{9.20}$$

where $u_1(t)$ and $u_2(t)$ are two given functions of t. Just like for the initial condition it is natural to require $v(t,x)$ to satisfy these conditions not only at the grid points on the boundary but on the whole boundary and as a consequence the auxiliary functions will all assume the value 0 on the boundary:

$$c(t,X_1) = d(t,X_1) = e(t,X_1) = f(t,X_1) = g(t,X_1) = \cdots = 0, \;\; t > 0, \tag{9.21}$$
$$c(t,X_2) = d(t,X_2) = e(t,X_2) = f(t,X_2) = g(t,X_2) = \cdots = 0, \;\; t > 0. \tag{9.22}$$

9.6 The error for the explicit method

If we have an initial-boundary value problem for (9.1) with Dirichlet boundary conditions, and if we use the explicit method for the numerical solution then we have the following results for the auxiliary functions:

The differential equation (9.12) for $c(t,x)$ is homogeneous and so are the side conditions according to (9.18), (9.21), and (9.22). $c(t,x) \equiv 0$ is a solution, and by uniqueness the only one. It follows that $c(t,x) \equiv 0$ and therefore that there is no h-contribution to the global error in (9.3).

The differential equation (9.14) for $e(t,x)$ is apparently inhomogeneous, but since $c(t,x) \equiv 0$ so is c_{tt} and the equation is homogeneous after all. So are the side conditions and we can conclude that $e(t,x) \equiv 0$.

The global error expression (9.3) for the explicit method therefore takes the form

$$v(t,x) \;=\; u(t,x) - kd - h^2 f - k^2 g - \cdots \tag{9.23}$$

and we deduce that the explicit method is indeed first order in time and second order in space.

For d we have from (9.13) that $Pd = \frac{1}{2}u_{tt}$ so we must require the problem to be such that u is twice differentiable w.r.t. t. This is usually no problem except possibly in small neighbourhoods around isolated points on the boundary.

9.7 The implicit method

We write the implicit method as

$$\frac{v_m^n - v_m^{n-1}}{k} - b_m^n \delta^2 v_m^n + a_m^n \tilde{\mu}\delta v_m^n - \kappa_m^n v_m^n \;=\; \nu_m^n. \tag{9.24}$$

where time step $n-1$ now contains the known values and time step n the values we are about to calculate. Equations (9.8) – (9.10) still hold while equation (9.7) is replaced by:

$$\begin{aligned}\frac{v_m^n - v_m^{n-1}}{k} \;=\;& u_t - \frac{1}{2}ku_{tt} + \frac{1}{6}k^2 u_{ttt} - hc_t + \frac{1}{2}hkc_{tt} - kd_t + \frac{1}{2}k^2 d_{tt} \\ & - hke_t - h^2 f_t - k^2 g_t + O(k^3 + k^2 h + kh^2 + h^3)\end{aligned} \tag{9.25}$$

Equating terms as before we get a set of equations rather similar to (9.11) – (9.16). (9.11) and (9.12) are unchanged, there is a single sign change in (9.14) and we can still conclude that $c(t,x) \equiv e(t,x) \equiv 0$. The remaining equations are

$$\begin{aligned} k: &\qquad Pd \;=\; -\frac{1}{2}u_{tt} && (9.26)\\ h^2: &\qquad Pf \;=\; -\frac{1}{12}bu_{4x} + \frac{1}{6}au_{xxx} && (9.27)\\ k^2: &\qquad Pg \;=\; \frac{1}{6}u_{ttt} + \frac{1}{2}d_{tt} && (9.28)\end{aligned}$$

and the error expansion for the implicit method has the same form as (9.23). Since there is a sign change in (9.26) as compared to (9.13) we can conclude that $d_{Im}(t,x) = -d_{Ex}(t,x)$. The right-hand-side of (9.27) is the same as in (9.15) and the sign change in the right-hand-side of (9.28) is compensated by d being of opposite sign. We therefore have that $f(t,x)$ and $g(t,x)$ are the same for the explicit and the implicit method.

9.7.1 An example

Consider test problem 1:

$$\begin{array}{rcll} u_t &=& u_{xx}, & -1 \le x \le 1, \quad t > 0, \\ u(0,x) = u_0(x) &=& \cos x, & -1 \le x \le 1, \\ u(t,-1) = u(t,1) &=& e^{-t}\cos 1, & \qquad\qquad\quad t > 0. \end{array}$$

with the true solution $u(t,x) = e^{-t}\cos x$.

For the explicit method we have

$$Pd = d_t - d_{xx} = \frac{1}{2}u_{tt}.$$

For f we have similarly

$$Pf = f_t - f_{xx} = -\frac{1}{12}u_{4x}.$$

Since $u_t = u_{xx}$ we have $u_{tt} = u_{txx} = u_{4x}$ such that $d(t,x) = -6f(t,x)$.

For the explicit method we must have $k = \mu h^2$ with $\mu \le \frac{1}{2}$. Keeping μ fixed the leading terms in the error expansion are

$$kd + h^2 f = -6\mu h^2 f + h^2 f = (1-6\mu)h^2 f.$$

If we choose $\mu = \frac{1}{2}$ as is common we get the leading term of the error to be $-2h^2 f(t,x)$. There is an obvious advantage in choosing $\mu = \frac{1}{6}$ in which case we obtain fourth order accuracy in h.

If we use the implicit method f stays the same and d changes sign and the leading terms of the error expansion become

$$6\mu h^2 f + h^2 f = (1+6\mu)h^2 f.$$

With $\mu = \frac{1}{2}$ the error becomes $4h^2 f(t,x)$ i.e. twice as big (and of opposite sign) as for the explicit method (cf. the solutions to exercises 1 and 5 of Chapter 1). There is no value for μ that will secure a higher order accuracy.

9.8 Crank-Nicolson

The Crank-Nicolson method can be written as

$$\begin{array}{rcl} \frac{v_m^{n+1} - v_m^n}{k} - \frac{1}{2}b_m^{n+1}\delta^2 v_m^{n+1} - \frac{1}{2}b_m^n \delta^2 v_m^n & & \qquad (9.29) \\ + \frac{1}{2}a_m^{n+1}\tilde{\mu}\delta v_m^{n+1} + \frac{1}{2}a_m^n\tilde{\mu}\delta v_m^n - \frac{1}{2}\kappa_m^{n+1}v_m^{n+1} - \frac{1}{2}\kappa_m^n v_m^n &=& \frac{1}{2}(\nu_m^{n+1} + \nu_m^n). \end{array}$$

The optimal expansion point is now $((n+\frac{1}{2})k, mh)$. To take full advantage of the even/odd cancellations we split the expansion in two stages. First we do the expansions (9.8) and (9.9) for time step n and $n+1$ and then we combine the results using the formula

$$\frac{1}{2}u^{n+1} + \frac{1}{2}u^n \;=\; u^{n+\frac{1}{2}} + \frac{1}{8}k^2 u_{tt} + O(k^4) \tag{9.30}$$

on all the individual terms. The resulting equations are

$$\begin{aligned}
\frac{v_m^{n+1} - v_m^n}{k} &= u_t + \frac{1}{24}k^2 u_{ttt} - hc_t - kd_t - hke_t && (9.31)\\
&\quad - h^2 f_t - k^2 g_t + O(k^3 + k^2 h + kh^2 + h^3) \\
\frac{1}{2}(b_m^{n+1}\delta^2 v_m^{n+1} + b_m^n \delta^2 v_m^n) &= b_m^{n+\frac{1}{2}}\{u_{xx} + \frac{1}{12}h^2 u_{4x} - hc_{xx} - kd_{xx} - && (9.32)\\
&\quad hke_{xx} - h^2 f_{xx} - k^2 g_{xx}\} + \frac{1}{8}k^2(bu_{xx})_{tt} + O(\cdots) \\
\frac{1}{2}(a_m^{n+1}\tilde{\mu}\delta v_m^{n+1} + a_m^n \tilde{\mu}\delta v_m^n) &= a_m^{n+\frac{1}{2}}\{u_x + \frac{1}{6}h^2 u_{xxx} - hc_x - kd_x - hke_x && (9.33)\\
&\quad - h^2 f_x - k^2 g_x\} + \frac{1}{8}k^2(au_x)_{tt} + O(\cdots) \\
\frac{1}{2}(\kappa_m^{n+1} v_m^{n+1} + \kappa_m^n v_m^n) &= \kappa_m^{n+\frac{1}{2}}\{u - hc - kd - hke - h^2 f - k^2 g\} && (9.34)\\
&\quad + \frac{1}{8}k^2(\kappa u)_{tt} + O(\cdots) \\
\frac{1}{2}(\nu_m^{n+1} + \nu_m^n) &= \nu + \frac{1}{8}k^2 \nu_{tt} + O(\cdots) && (9.35)
\end{aligned}$$

We insert (9.31) – (9.35) in (9.29) and equate terms with the same powers of h and k:

$$\begin{aligned}
1: &\quad Pu = \nu && (9.36)\\
h: &\quad Pc = 0 && (9.37)\\
k: &\quad Pd = 0 && (9.38)\\
hk: &\quad Pe = 0 && (9.39)\\
h^2: &\quad Pf = -\frac{1}{12}bu_{4x} + \frac{1}{6}au_{xxx} && (9.40)\\
k^2: &\quad Pg = \frac{1}{24}u_{ttt} - \frac{1}{8}(bu_{xx})_{tt} + \frac{1}{8}(au_x)_{tt} - \frac{1}{8}(\kappa u)_{tt} - \frac{1}{8}\nu_{tt} && (9.41)
\end{aligned}$$

The right-hand-side in (9.41) looks rather complicated but if the solution to (9.1) is smooth enough such that we can differentiate (9.1) twice w.r.t. t then we can combine the last four terms in (9.41) to $-\frac{1}{8}u_{ttt}$ and the equation becomes

$$k^2: \quad Pg = -\frac{1}{12}u_{ttt} \tag{9.42}$$

If the inhomogeneous term $\nu(t,x)$ in the equation (9.1) can be evaluated at the mid-points $((n+\frac{1}{2})k, mh)$ then it is tempting to use $\nu_m^{n+\frac{1}{2}}$ instead of $\frac{1}{2}(\nu_m^{n+1}+\nu_m^n)$ in (9.29). We shall then miss the term with $\frac{1}{8}\nu_{tt}$ in (9.41) and therefore not have complete advantage of the reduction leading to (9.42). Instead we shall have

$$k^2: \qquad Pg \;=\; -\frac{1}{12}u_{ttt}+\frac{1}{8}\nu_{tt}.$$

It is impossible to say in general which is better, but certainly (9.42) is simpler.

Looking at equations (9.36) – (9.42) we again recognize the original equation for u in (9.36), and from (9.37) – (9.39) we may conclude that $c(t,x)\equiv d(t,x)\equiv e(t,x)\equiv 0$ showing that Crank-Nicolson is indeed second order in both k and h. We also note from (9.40) that $f(t,x)$ for Crank-Nicolson is the same function as for the explicit and the implicit method.

9.8.1 Example continued

For our example we have

$$Pg = g_t - g_{xx} = -\frac{1}{12}u_{ttt}.$$

For this particular problem we have $u_{ttt}=-u_{tt}=-u_{4x}$ such that
$f(t,x)=-g(t,x)$ and that the leading terms of the error are

$$h^2 f + k^2 g = (h^2-k^2)f.$$

There is a distinct advantage in choosing $k=h$ in which case the second order terms in the error expansion will cancel (cf. the answer to exercise 12 of Chapter 1), but we must stress that this holds for this particular example and is not a general result for Crank-Nicolson.

9.9 Upwind schemes

When $|a|$ is large compared to b we occasionally observe oscillations in the numerical solution. One remedy is to reduce the step sizes but this costs computer time. Another option is to use an upwind scheme such as in the explicit case (for $a>0$):

$$\frac{v_m^{n+1}-v_m^n}{k} - b_m^n\delta^2 v_m^n + a_m^n\frac{v_m^n - v_{m-1}^n}{h} - \kappa_m^n v_m^n \;=\; \nu_m^n. \qquad (9.43)$$

To analyze the effect on the error we use

$$\begin{aligned}\frac{v_m^n - v_{m-1}^n}{h} &= u_x - \frac{1}{2}hu_{xx} + \frac{1}{6}h^2u_{xxx} - hc_x + \frac{1}{2}h^2c_{xx} \\ &\quad - kd_x + \frac{1}{2}hkd_{xx} - hke_x - h^2f_x - k^2g_x + O(\cdots)\end{aligned} \tag{9.44}$$

together with (9.7), (9.8), and (9.10). Equating terms with the same powers of h and k now gives

$$1: \qquad Pu = \nu \tag{9.45}$$

$$h: \qquad Pc = -\frac{1}{2}au_{xx} \tag{9.46}$$

$$k: \qquad Pd = \frac{1}{2}u_{tt} \tag{9.47}$$

$$hk: \qquad Pe = -\frac{1}{2}c_{tt} + \frac{1}{2}ad_{xx} \tag{9.48}$$

From (9.46) and (9.48) we conclude that $c(t,x)$ and $e(t,x)$ are no longer identically zero and that the method is now first order in both k and h. A similar result holds for the implicit scheme. For Crank-Nicolson the order in h is also reduced to 1 but we keep second order accuracy in k.

9.10 Boundary conditions with a derivative

If one of the boundary conditions involves a derivative then the discretization of this has an effect on the global error of the numerical solution. Assume that the condition on the left boundary is

$$\alpha u(t, X_1) - \beta u_x(t, X_1) = \gamma, \qquad t > 0 \tag{9.49}$$

where α, β and γ may depend on t. A similar condition might be imposed on the other boundary and the considerations would be completely similar so we shall just consider a derivative condition on one boundary. We shall in turn study three different discretizations of the derivative in (9.49):

$$\frac{v_1^n - v_0^n}{h} \qquad \text{(first order)} \tag{9.50}$$

$$\frac{-v_2^n + 4v_1^n - 3v_0^n}{2h} \qquad \text{(second order, asymmetric)} \tag{9.51}$$

$$\frac{v_1^n - v_{-1}^n}{2h} \qquad \text{(second order, symmetric)} \tag{9.52}$$

9.10.1 First order approximation

We first use the approximation (9.50) in (9.49). If the coefficients α, β and γ depend on t they should be evaluated at $t = nk$:

$$\alpha v_0^n - \beta \frac{v_1^n - v_0^n}{h} = \gamma, \qquad t > 0. \tag{9.53}$$

We now use the assumption (9.3) and Taylor-expand v_1^n around (nk, X_1):

$$\begin{aligned} &\alpha\{u - hc - kd - hke - h^2 f - k^2 g\} - \beta\{u_x + \frac{1}{2}hu_{xx} + \frac{1}{6}h^2 u_{xxx} - hc_x \\ &- \frac{1}{2}h^2 c_{xx} - kd_x - \frac{1}{2}hkd_{xx} - hke_x - h^2 f_x - k^2 g_x\} - \gamma = O(\cdots) \end{aligned} \tag{9.54}$$

Collecting terms with 1, h, k, hk, h^2, and k^2 as before we get

$$\begin{aligned} 1: &\quad \alpha u - \beta u_x = \gamma &(9.55)\\ h: &\quad \alpha c - \beta c_x = -\frac{1}{2}\beta u_{xx} &(9.56)\\ k: &\quad \alpha d - \beta d_x = 0 &(9.57)\\ hk: &\quad \alpha e - \beta e_x = \frac{1}{2}\beta d_{xx} &(9.58)\\ h^2: &\quad \alpha f - \beta f_x = -\frac{1}{6}\beta u_{xxx} + \frac{1}{2}\beta c_{xx} &(9.59)\\ k^2: &\quad \alpha g - \beta g_x = 0 &(9.60) \end{aligned}$$

We recognize the condition (9.49) for u in (9.55). As for c the boundary condition (9.56) is no longer homogeneous and we shall expect c to be nonzero. This holds independently of which method is used for the discretization of the equation (9.1). So if we use a first order boundary approximation we get a global error which is first order in h.

9.10.2 Asymmetric second order

We now apply the approximation (9.51) in (9.49):

$$\alpha v_0^n - \beta \frac{-v_2^n + 4v_1^n - 3v_0^n}{2h} = \gamma, \qquad t > 0. \tag{9.61}$$

We again use the assumption (9.3) and Taylor-expand v_1^n and v_2^n around (nk, X_1):

$$\begin{aligned} &\alpha\{u - hc - kd - hke - h^2 f - k^2 g\} - \beta\{u_x - \frac{1}{3}h^2 u_{xxx} - hc_x \\ &- kd_x - hke_x - h^2 f_x - k^2 g_x\} - \gamma = O(h^3 + h^2 k + hk^2 + k^3) \end{aligned} \tag{9.62}$$

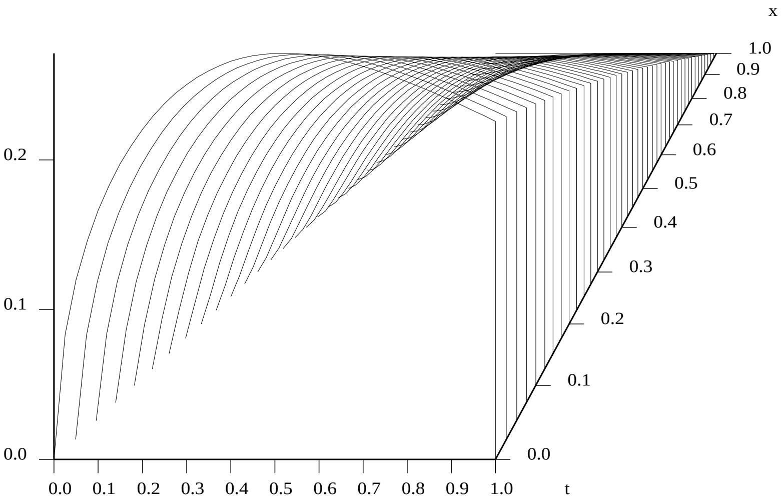


Figure 9.1: $c(t, x)$.

Collecting terms with 1, h, k, hk, h^2, and k^2 as before we get

$$\begin{aligned}
1: \quad & \alpha u - \beta u_x &=& \ \gamma & (9.63)\\
h: \quad & \alpha c - \beta c_x &=& \ 0 & (9.64)\\
k: \quad & \alpha d - \beta d_x &=& \ 0 & (9.65)\\
hk: \quad & \alpha e - \beta e_x &=& \ 0 & (9.66)\\
h^2: \quad & \alpha f - \beta f_x &=& \ \frac{1}{3}\beta u_{xxx} & (9.67)\\
k^2: \quad & \alpha g - \beta g_x &=& \ 0 & (9.68)
\end{aligned}$$

We recognize the condition (9.49) for u in (9.63). We now have a homogeneous condition (9.64) for c and this will assure that $c(t, x) \equiv 0$ when we combine (9.61) with the explicit, the implicit, or the Crank-Nicolson method. We also have $e(t, x) \equiv 0$, but in order to have $d(t, x) \equiv 0$ we must use the Crank-Nicolson method. One disadvantage with this asymmetric approximation which does not show in equations (9.63) – (9.68) is that the next h-term is third order and therefore can be expected to interfere more than the fourth order term which is present in the symmetric case below.

9.10.3 Symmetric second order

We finally apply the symmetric approximation (9.52) in (9.49):

$$\alpha v_0^n - \beta \frac{v_1^n - v_{-1}^n}{2h} \ = \ \gamma, \qquad t > 0. \qquad (9.69)$$

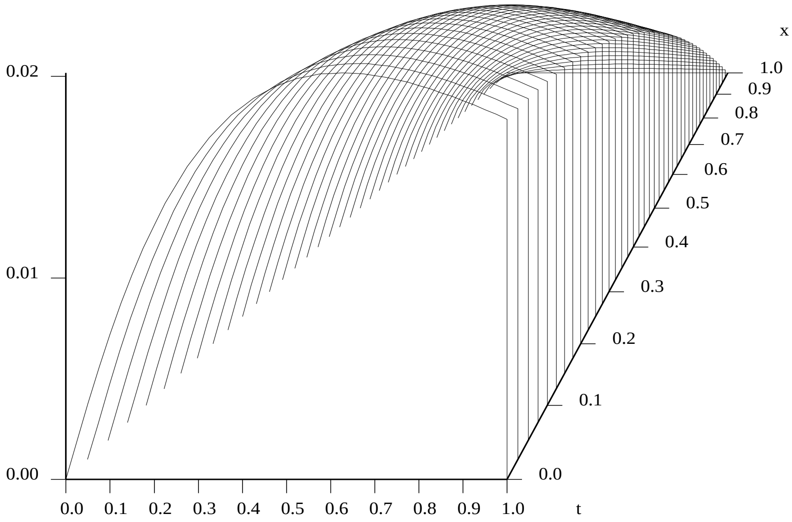


Figure 9.2: $g(t,x)$.

We again use the assumption (9.3) and Taylor-expand v_1^n and v_{-1}^n around (nk, X_1):

$$\alpha\{u - hc - kd - hke - h^2 f - k^2 g\} - \beta\{u_x + \frac{1}{6}h^2 u_{xxx} - hc_x$$
$$- kd_x - hke_x - h^2 f_x - k^2 g_x\} - \gamma = O(h^3 + h^2 k + hk^2 + k^3) \qquad (9.70)$$

Collecting terms with 1, h, k, hk, h^2, and k^2 as before we get

$$1: \quad \alpha u - \beta u_x = \gamma \qquad (9.71)$$
$$h: \quad \alpha c - \beta c_x = 0 \qquad (9.72)$$
$$k: \quad \alpha d - \beta d_x = 0 \qquad (9.73)$$
$$hk: \quad \alpha e - \beta e_x = 0 \qquad (9.74)$$
$$h^2: \quad \alpha f - \beta f_x = -\frac{1}{6}\beta u_{xxx} \qquad (9.75)$$
$$k^2: \quad \alpha g - \beta g_x = 0 \qquad (9.76)$$

All the same conclusions as for the asymmetric case will hold also in this symmetric case.

9.10.4 Test problem 3 revisited

To illustrate the above analyses let us look at test problem 3 on page 36:

$$u_t = u_{xx}, \qquad 0 \le x \le 1, \qquad t > 0,$$

$$\begin{aligned} u(0,x) \;=\; u_0(x) &= \cos x, && 0 \le x \le 1, \\ u(t,1) &= e^{-t}\cos 1, && t > 0, \\ u_x(t,0) &= 0, && t > 0. \end{aligned}$$

with the true solution $u(t,x) = e^{-t}\cos x$.

We wish to solve numerically using Crank-Nicolson and want to study the behaviour of the global error using various discretizations of the derivative boundary condition.

Using the first order boundary approximation we have $d = e = 0$ and the global error will be on the form

$$h\,c + h^2 f + k^2 g + \cdots$$

We have solved the initial-boundary value problems for the functions $c(t,x)$ and $g(t,x)$ (using Crank-Nicolson and $h = k = 0.025$) and show the results graphically in Fig. 9.1 and Fig. 9.2.

It is clear from the figures that the first order contribution to the error is considerable. The values of $c(t,x)$ lie between 0 and 0.28 and those of $g(t,x)$ between 0 and 0.022 and from this we could estimate the truncation error for given values of the step sizes h and k. Or we could suggest step sizes in order to make the truncation error smaller than a given tolerance.

With the second order boundary approximations $c(t,x)$ is expected to be identically 0 and the accuracy correspondingly better. The derivative boundary condition for f reduces to $f_x(t,0) = 0$ for this particular case since $u_{xxx}(t,0) = e^{-t}\sin 0 = 0$. We can therefore again conclude that $f(t,x) = -g(t,x)$.

In more general situations it is not so easy to gain information about the auxiliary functions in this way. In the next chapter we shall see how we can let the computer do the work and verify our assumptions about the order of the error and at the same time gain information about the magnitude of the error.

9.11 Exercise

1. Solve problem 1 on page 5 with the explicit method, the implicit method and Crank-Nicolson from $t = 0$ to $t = 0.5$ with $h = \frac{1}{10}, \frac{1}{20}$, and $\frac{1}{40}$, and with $\mu (= k/h^2) = 1/6$.
   Compute the max-norm and the 2-norm of the error for each method for $t = 0.1, 0.2, 0.3, 0.4, 0.5$.

Chapter 10

Estimating the Global Error and Order

10.1 Introduction

In the previous chapter we introduced the basic hypothesis that the global error could be expressed as a power series in h and k and with the auxiliary functions c, d, $\ldots$, and we found differential equations defining these functions. In this chapter we shall see how we can get the computer to help us acquiring information on the order of the method and the magnitude of the auxiliary functions at the grid points.

10.2 The global error

We shall begin our analysis in one dimension and later extend it to functions of two or more variables. We shall first define what we mean by the global error being of order say $O(h)$. Let $u(x)$ be the true solution, and let $v(x)$ be the computed solution. Our basic hypothesis is (as in Chapter 9) that the computed solution can be written as

$$v(x) = u(x) - hc(x) - h^2 d(x) - h^3 f(x) - h^4 g(x) - \cdots \qquad (10.1)$$

where $c(x)$, $d(x)$, $f(x)$, and $g(x)$ are differentiable functions of x alone, the dependence of v on the step size h being expressed through the power series in h. This is a hypothesis and as such can not be proved, but it leads to predictions which can be verified computationally, thereby identifying those situations where the hypothesis can be assumed to hold.

If the function $c(x)$ happens to be identically 0 then the method is (at least) of second order, otherwise it is of first order. Even if $c(x)$ is not identically 0 then it might very well have isolated zeroes. At such places our analysis might give results which are difficult to interpret correctly. Therefore the analysis should always be performed for a substantial set of grid points in order to give trustworthy results.

In the following we shall show how we by performing calculations with various values of the step size, h, can extract information not only about the true solution but also about the order and magnitude of the error.

A calculation with step size h will yield

$$v_1 \;=\; u - hc - h^2 d - h^3 f - h^4 g - \cdots \tag{10.2}$$

A second calculation with twice as large a step size gives

$$v_2 \;=\; u - 2hc - 4h^2 d - 8h^3 f - 16h^4 g - \cdots \tag{10.3}$$

We can now eliminate u by subtraction:

$$v_1 - v_2 \;=\; hc + 3h^2 d + 7h^3 f + 15h^4 g + \cdots \tag{10.4}$$

A third calculation with $4h$ is necessary to retrieve information about the order

$$v_3 \;=\; u - 4hc - 16h^2 d - 64h^3 f - 256h^4 g - \cdots \tag{10.5}$$

whence

$$v_2 - v_3 \;=\; 2hc + 12h^2 d + 56h^3 f + 240h^4 g + \cdots \tag{10.6}$$

and a division gives the *order-ratio*:

$$q \;=\; \frac{v_2 - v_3}{v_1 - v_2} \;=\; 2\frac{c + 6hd + 28h^2 f + 120h^3 g + \cdots}{c + 3hd + 7h^2 f + 15h^3 g + \cdots}. \tag{10.7}$$

This ratio can be computed in all those points where we have information from all three calculations, i.e. all grid points corresponding to the last calculation with step size $4h$.

If $c \neq 0$ and h is suitably small we shall observe numbers in the neighbourhood of 2 in all points, and this would indicate that the method is of first order. If $c = 0$ and $d \neq 0$, then the quotient will assume values close to 4 and if this happens for many points and not just at isolated spots then we can deduce that c is identically 0 and that the method is of second order. The smaller h the smaller influence for the next terms in the numerator and the denominator, and the picture should become clearer.

The error in the first calculation, v_1, is given by

$$e_1 \;=\; u - v_1 \;=\; hc + h^2 d + h^3 f + h^4 g + \cdots \tag{10.8}$$

If we observe many values of the order-ratio (10.7) in the neighbourhood of 2 indicating that $|c|$ is substantially larger than $h|d|$, and that the method therefore is of first order, then e_1 is represented reasonably well by $v_1 - v_2$:

$$e_1 \;=\; v_1 - v_2 - 2h^2 d - 6h^3 f - \cdots \tag{10.9}$$

and $v_1 - v_2$ can be used as an estimate of the error in v_1.

One could choose to add $v_1 - v_2$ to v_1 and thereby get more accurate results. This process is called *Richardson extrapolation* and can be done for all grid points involved in the calculation of v_2.

$$v_1' \;=\; v_1 + (v_1 - v_2) \;=\; u + 2h^2 d + 6h^3 f + 14h^4 g + \cdots \tag{10.10}$$

If the error (estimate) behaves nicely we might even consider interpolating to the intermediate points and thus get extrapolated values with spacing h. Interpolation or not, we cannot at the same time, i.e. without doing some extra work, get a realistic estimate of the error in this improved value. The old estimate can of course still be used but it is expected to be rather pessimistic.

If in contrast we observe many numbers in the neighbourhood of 4 then $|c|$ is substantially less than $h|d|$ and is probably 0. At the same time $|d|$ will be larger than $h|f|$, and the method would be of second order with an error

$$e_2 \;=\; u - v_1 \;=\; h^2 d + h^3 f + h^4 g + \cdots \tag{10.11}$$

This error will be estimated nicely by $(v_1 - v_2)/3$:

$$e_2 \;=\; \frac{1}{3}(v_1 - v_2) - \frac{4}{3}h^3 f - 4h^4 g - \cdots \tag{10.12}$$

It is thus important to check the order before calculating an estimate of the error and certainly before making any corrections using this estimate. If in doubt it is usually safer to estimate the order on the low side. If the order is 2 and the correct error estimate therefore $(v_1 - v_2)/3$, then misjudging the order to be 1 and using $v_1 - v_2$ for the error estimate would not be terribly bad, and actually on the safe side. But if we want to attempt Richardson extrapolation it is very important to have the right order.

If our task is to compute function values with a prescribed error tolerance then the error estimates can also be used to predict a suitable step size which would satisfy this requirement and in the second round to check that the ensuing calculations are satisfactory.

How expensive are these extra calculations which are needed to gain information on the error? We shall compare with the computational work for v_1 under the assumption that the work is proportional to the number of grid points. Therefore v_2 costs half as much as v_1, and v_3 costs one fourth. The work involved in calculating $v_1 - v_2$, $v_2 - v_3$ and their quotient which is done for $1/4$ of the grid points will not be considered since it is assumed to be considerably less than the fundamental difference calculations.

The work involved in finding v_1, v_2 and v_3 is therefore 1.75, i.e. an extra cost of 75%, and that is actually very inexpensive for an error estimate. Getting information on the magnitude of the error enables us to choose a realistic step size and thus meet the requirements without performing too many unnecessary calculations. If the numbers allow an extrapolation then the result of this is expected to be much better than a calculation with half the step size and we are certainly better off. If the computational work increases faster than the number of grid points then the result is even more in favour of the present method.

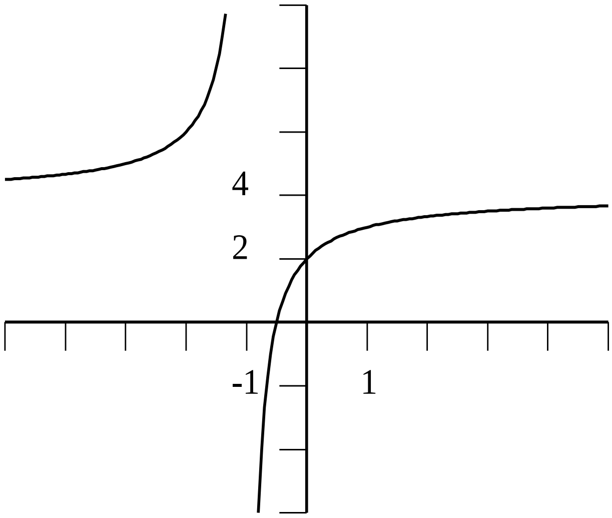


Figure 10.1: The function $w(y) = 2\frac{1+2y}{1+y}$.

10.3 Can we trust these results?

Yes, if we really observe values of the order-ratio (10.7) between say 1.8 and 2.2 for all relevant grid points then the method is of first order and the first term in the remainder series dominates the rest. Discrepancies from this pattern in small areas are also allowed. They may be due to the fact that $c(x)$ has an isolated zero. This can be checked by observing the values of $v_1 - v_2$ in a neighbourhood. These numbers which are usually dominated by the term hc will then become smaller and display a change of sign indicating that $c(x)$ has a zero somewhere in the neighbourhood. The zero of $c(x)$ and that of $v_1 - v_2$ will usually not coincide, since the latter will correspond to $c(x) \approx -3hd(x)$. The global error itself will also display a sign change and thus be small and pass through zero somewhere close. We don't know precisely where, and the exact location is also of academic interest only, since we only have information on the computed solution at a discrete set of points. In a small neighbourhood around this zero the error estimate, $v_1 - v_2$ may not even reproduce the sign of the error correctly, but as long as the absolute value is small this is of lesser significance. The important thing is that the error estimate is reliable in sign and magnitude when the error is large, and this will be the case as long as the order ratio stays close to 2.

If a method is of first order and we choose to exploit the error estimate to adjust the calculated value (i.e. to perform Richardson extrapolation) then it might be reasonable to assume that the resulting method is of second order as indicated in (10.10). This of course can be tested by repeating the above process. We shall need a fourth calculation v_4 (with step size $8h$), such that we can compute three extrapolated values, $v'_q = v_q + (v_q - v_{q+1})$, $q = 1, 2, 3$, on the basis of which we can get information about the (new) order. We of course expect the order to be at least 2, but it is important to have this extra assurance that our basic hypothesis is valid. If the results do not confirm this then it might be an idea to review the calculations.

What will actually happen if we perform a Richardson extrapolation based on a wrong assumption about the order? Usually not too much. If we attempt to eliminate a second order term in a first order calculation then the result will still be of first order; and if we attempt to eliminate a first order term in a second order process then the absolute value of the error will double but the result will retain its high order.

If we want to understand in detail what might happen to the order-ratio (10.7) in the strange areas, i.e. how the ratio might vary when $h|d|$ is not small compared to $|c|$, then we can consider the behaviour of the function

$$w(y) \quad = \quad 2\frac{1 + 2y}{1 + y} \tag{10.13}$$

where $y = 3h\frac{d}{c}$ (see Fig. 10.1).

If y is positive, then $2 < w(y) < 4$, and $w(y) \to 4$, when $y \to \infty$.
This corresponds to c and d having the same sign.
If y is small then $w(y) \approx 2$.
If y is large and negative then $w(y) > 4$, and $w(y) \to 4$ when $y \to -\infty$.

The situation $y \to \pm\infty$ corresponds to $c = 0$, i.e. that the method is of second order.
The picture becomes rather blurred when y is close to -1, i.e. when c and d have opposite sign and $c \approx -3hd$:

$$\begin{aligned} y \uparrow -1 \quad &\Rightarrow \quad w \to +\infty \\ y \downarrow -1 \quad &\Rightarrow \quad w \to -\infty \\ -1 < y < -\frac{1}{2} \quad &\Rightarrow \quad w < 0 \end{aligned}$$

But in these cases we are far away from $|c| \gg h|d|$.

Reducing the step size by one half corresponds to reducing y by one half.

If	$0 < w(y) < 4$	then	$w(\frac{y}{2})$ will be closer to 2.
If	$w(y) < 0$	then	$0 < w(\frac{y}{2}) < 2$.
If	$6 < w(y)$	then	$w(\frac{y}{2}) < 0$.
If	$4 < w(y) < 6$	then	$w(\frac{y}{2}) > w(y)$.

If c and d have opposite sign and c is not dominant, the picture will be rather chaotic, but a suitable reduction of h will result in a clearer picture if the fundamental assumptions are valid.

We have been rather detailed in our analysis of first order methods with a non-vanishing second order term. Quite similar analyses can be made for second and third order, or for second and fourth order or for higher orders. If the ratio (10.7) is close to 2^p then our method is of order p.

10.4 Further improvements of the error estimate

The error estimates we compute are just estimates and not upper bounds on the magnitude of the error. They will often be very realistic, but they may sometimes underestimate the error, and it would be useful to identify those situations. The following analysis will show that this happens when the order ratio is consistently smaller than 2^p. On the other hand, if the order ratio is larger than 2^p then the error estimate is usually a (slight) overestimate.

If the method is first order ($c \neq 0$) we expect the next term in the error expansion to be second order ($d \neq 0$) and we have

$$q = \frac{v_2 - v_3}{v_1 - v_2} \approx 2\frac{1+2y}{1+y} \approx 2(1+y) \qquad \text{with} \qquad y = 3h\tfrac{d}{c} \tag{10.14}$$

If we observe values $q = 2(1+\varepsilon)$ then we have $y \approx \varepsilon = \frac{q-2}{2}$.

From (10.4) and (10.9) we have

$$v_1 - v_2 \;=\; hc(1 + 3h\frac{d}{c} + \cdots) \;=\; hc(1 + y + \cdots) \qquad (10.15)$$

and

$$e_1 \;=\; (v_1 - v_2)(1 - \frac{2h\frac{d}{c} + \cdots}{1 + y + \cdots}) \;\approx\; (v_1 - v_2)(1 - \frac{2}{3}y). \qquad (10.16)$$

If $\varepsilon > 0$ (i.e. $q > 2$) then the error is smaller than the estimate $v_1 - v_2$, and if $\varepsilon < 0$ (i.e. $q < 2$) then the error is larger than the estimate.

Since $y \approx \varepsilon = (q - 2)/2$ we can even compensate for the effect taking as our improved error estimate the value

$$est_{12} \;=\; (v_1 - v_2)(1 - \frac{2}{3}\varepsilon) \;=\; (v_1 - v_2)(1 - \frac{q-2}{3}). \qquad (10.17)$$

A direct calculation reveals that

$$est_{12} \;=\; e_1 - 8h^3 f - \cdots \qquad (10.18)$$

showing that this improved estimate takes both the first and the second order term into account.

We must of course be careful with these calculations. They should only be used when ε is small and varies slowly over the region in question.

If the method is second order ($c = 0,\ d \neq 0$) then the next term in the error expansion might be third order ($f \neq 0$) or fourth order ($f = 0,\ g \neq 0$). In any case

$$q \;=\; \frac{v_2 - v_3}{v_1 - v_2} \;\approx\; 4\frac{1 + 2y + 4z}{1 + y + z} \qquad (10.19)$$

with

$$y \;=\; \frac{7}{3}h\frac{f}{d}, \qquad z \;=\; 5h^2\frac{g}{d} \qquad (10.20)$$

From (10.4) and (10.12) we have

$$v_1 - v_2 = 3h^2 d(1 + \frac{7}{3}h\frac{f}{d} + 5h^2\frac{g}{d} + \cdots) \;=\; 3h^2 d(1 + y + z + \cdots) \qquad (10.21)$$

and

$$e_2 = \frac{v_1 - v_2}{3}(1 - \frac{4h^3 f + 12h^4 g + \cdots}{3h^2 d + \cdots}) \;\approx\; \frac{v_1 - v_2}{3}(1 - \frac{4}{7}y - \frac{4}{5}z). \qquad (10.22)$$

If we have a second and a third order term then y will probably dominate z and

$$q \;\approx\; 4\frac{1+2y}{1+y} \;\approx\; 4(1+y)$$

and if we observe values $q = 4(1+\varepsilon)$ then we have $\varepsilon \approx y$ and the error estimate should be

$$est_{23} \;=\; \frac{v_1 - v_2}{3}(1 - \frac{4}{7}\varepsilon) \;=\; \frac{v_1 - v_2}{3}(1 - \frac{4}{7}\frac{q-4}{4}). \qquad (10.23)$$

If the next term is fourth order then $y = 0$ and

$$q \;\approx\; 4\frac{1+4z}{1+z} \;\approx\; 4(1+3z)$$

and if we observe values $q = 4(1+\varepsilon)$ then we have $\varepsilon \approx 3z$ and the error estimate should be

$$est_{24} \;=\; \frac{v_1 - v_2}{3}(1 - \frac{4}{15}\varepsilon) \;=\; \frac{v_1 - v_2}{3}(1 - \frac{4}{15}\frac{q-4}{4}). \qquad (10.24)$$

Although we might have our suspicions it is not easy to *know* whether the next term is third or fourth order and this is important in order to decide which correction to apply. We can therefore not recommend using (10.23) or (10.24) directly for error estimation or extrapolation. To be on the safe side we instead suggest the following guidelines for error estimation:
If $\varepsilon > 0$ (i.e. $q > 4$) then $(v_1 - v_2)/3$ is probably larger than the error and can safely be used as an error estimate.
If $\varepsilon < 0$ (i.e. $q < 4$) then the error is larger than $(v_1 - v_2)/3$ and we recommend using $(1 - \frac{4}{7}\varepsilon)(v_1 - v_2)/3$ as the error estimate. If the next term is fourth order we shall be on the safe side; if it is third order the estimate will probably be more realistic, but it might be a slight underestimate.

10.5 Two independent variables

If u is a function of two or more variables then we can perform similar analyses taking one variable at a time. If say $u(t,x)$ is a function of two variables, t and x, and v is a numerical approximation based on step sizes k and h then our basic hypothesis would be

$$v_1 \;=\; u - hc - kd - hke - h^2 f - k^2 g - \cdots \qquad (10.25)$$

A calculation with $2h$ and k gives

$$v_2 \;=\; u - 2hc - kd - 2hke - 4h^2 f - k^2 g - \cdots$$

such that

$$v_1 - v_2 \quad = \quad hc + hke + 3h^2 f + \cdots \tag{10.26}$$

To check the order (in h) we need a third calculation with step sizes $4h$ and k:

$$v_3 = u - 4hc - kd - 4hke - 16h^2 f - k^2 g - \cdots$$

and we have

$$v_2 - v_3 = 2hc + 2hke + 12h^2 f + \cdots$$

and the order-ratio

$$\frac{v_2 - v_3}{v_1 - v_2} \quad = \quad 2\frac{c + ke + 6hf + \cdots}{c + ke + 3hf + \cdots}. \tag{10.27}$$

For the k-dependence we compute with h and $2k$:

$$v_4 = u - hc - 2kd - 2hke - h^2 f - 4k^2 g - \cdots$$

and with h and $4k$:

$$v_5 = u - hc - 4kd - 4hke - h^2 f - 16k^2 g - \cdots$$

such that

$$v_1 - v_4 \quad = \quad kd + hke + 3k^2 g + \cdots \tag{10.28}$$

and

$$\frac{v_4 - v_5}{v_1 - v_4} \quad = \quad 2\frac{d + he + 6kg + \cdots}{d + he + 3kg + \cdots}. \tag{10.29}$$

Using (10.27) and (10.29) we can check the order in h and k of our approximation and through (10.26) and (10.28) we can get information on the leading error terms.

If the method is first order in h we can estimate the h-component of the error by $v_1 - v_2$ and if the method is first order in k we can estimate the k-component of the error by $v_1 - v_4$. We can use this information to reduce either or both the step sizes in order to meet specific error tolerances or we can use Richardson-extrapolation in order to get higher order and (hopefully) more accurate results. More specificly

$$v_1 + (v_1 - v_2) + (v_1 - v_4) \quad = \quad u + hke + 2h^2 f + 2k^2 g + \cdots \tag{10.30}$$

If the method is first order in k and second order in h as is typical for the implicit method then the h-component of the error is estimated by $(v_1 - v_2)/3$ and the

k-component by $v_1 - v_4$. The latter will often be dominant and it will be natural to perform Richardson-extrapolation in the k-direction only, arriving at

$$v_1 + (v_1 - v_4) \quad = \quad u - h^2 f + 2k^2 g + \cdots \tag{10.31}$$

In order to check that these extrapolations give the expected results, it is again necessary to supplement with further calculations (and with a more advanced numbering system for these v's).

When u is a function of two variables with two independent step sizes then the cost of the five necessary calculations is 2.5 times the cost of v_1. This is still a reasonable price to pay. Knowing the magnitude of the error and its dependence on the step sizes enables us to choose near-optimal combinations of these and thus avoid redundant calculations, and a possible extrapolation might improve the results considerably more than halving the step sizes and quadrupling the work.

Table 10.1: h-ratio for first order boundary condition.

$t \setminus x$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.1	2.05	2.05	2.06	2.07	2.07	2.08	2.08	2.08	2.09	2.09
0.2	2.03	2.03	2.04	2.04	2.04	2.04	2.05	2.05	2.05	2.05
0.3	2.02	2.03	2.03	2.03	2.03	2.03	2.03	2.03	2.03	2.03
0.4	2.02	2.02	2.02	2.02	2.02	2.02	2.02	2.02	2.02	2.02
0.5	2.01	2.01	2.01	2.02	2.02	2.02	2.02	2.02	2.02	2.02
0.6	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01
0.7	2.00	2.00	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01
0.8	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
0.9	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
1.0	1.99	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00

10.6 Limitations of the technique.

It is essential for the technique to give satisfactory results that the leading term in the error expansion is the dominant one. This will always be the case when the step size is small, but how can we know that the step size is small enough?

This will be revealed by a study of the order-ratio and how it behaves in the region in question. A picture like the one seen in Table 10.1 is a clear witness of a first order process where the first order term clearly dominates the rest. The

Table 10.2: k-ratio for first order boundary condition.

$t \setminus x$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.1	1.89	14.69	3.87	1.20	4.30	5.25	4.52	3.42	2.81	2.76
0.2	1.82	17.13	2.65	4.11	4.16	3.99	3.96	4.00	4.03	4.02
0.3	1.68	6.15	3.63	4.13	4.00	3.98	4.01	4.01	4.00	4.00
0.4	1.51	4.19	3.98	4.05	3.98	4.00	4.00	4.00	4.00	4.00
0.5	1.33	3.65	4.06	4.01	3.99	4.00	4.00	4.00	4.00	4.00
0.6	1.15	3.50	4.07	3.99	4.00	4.00	4.00	4.00	4.00	4.00
0.7	1.00	3.49	4.06	3.99	4.00	4.00	4.00	4.00	4.00	4.00
0.8	0.88	3.54	4.05	3.99	4.00	4.00	4.00	4.00	4.00	4.00
0.9	0.82	3.60	4.03	3.99	4.00	4.00	4.00	4.00	4.00	4.00
1.0	0.84	3.67	4.02	4.00	4.00	4.00	4.00	4.00	4.00	4.00

error estimate will be very reliable (and a slight overestimate) and we can expect good results from an extrapolation.

A behaviour like in Table 10.2 is more difficult to interpret. For $x > 0.1$ and $t > 0.1$ the method is clearly second order (in k) and we should be able to trust the estimate of the corresponding error component. For small values of t and especially x our basic hypothesis (10.25) does not seem to quite capture the situation. A reduction of the step size, k, might help, but the problems may partly be due to the fact that the contribution to the error from k is so much smaller than the contribution from h. A comparison of the differences (10.26) and (10.28) will shed light on this.

Isolated deviations from the pattern such as seen in Table 16.3 at $(t, x) = (1.2, 175)$ and $(0.8, 150)$ and $(0.4, 130)$ are allowed and can be explained by referring to Fig. 10.1. The second order term in (10.26) has opposite sign and the same absolute magnitude as the next in line on a curve in (t, x)-space, and small values, negative values, and very large values of the order ratio will occur depending on how close the grid points lie to this curve. A study of the differences (10.26) will reveal very small numbers because of this cancellation. An extrapolation will have little effect because of these small differences. An error estimate based on these small differences cannot be trusted. It is safer to assume that the error is about the same as in surrounding points where the order ratio warrants a better determination. (The error is probably very small somewhere in the neighbourhood, but we don't know exactly where.)

The oscillations which often occur when using Crank-Nicolson (cf. Chapter 8) can also confuse the picture. These oscillations typically have a period of 2 times the step size such that for example the values at odd steps are large and the values

Table 10.3: h-ratio for asymmetric second order.

$t \setminus x$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.1	3.24	3.43	3.60	3.71	3.80	3.86	3.90	3.93	3.95	3.96
0.2	3.44	3.54	3.63	3.69	3.75	3.79	3.82	3.85	3.87	3.88
0.3	3.51	3.58	3.64	3.69	3.73	3.76	3.78	3.80	3.82	3.83
0.4	3.54	3.60	3.65	3.68	3.72	3.74	3.76	3.78	3.80	3.81
0.5	3.56	3.61	3.65	3.68	3.71	3.73	3.75	3.77	3.78	3.79
0.6	3.57	3.62	3.65	3.68	3.71	3.73	3.74	3.76	3.77	3.78
0.7	3.58	3.62	3.65	3.68	3.70	3.72	3.74	3.75	3.76	3.77
0.8	3.59	3.63	3.66	3.68	3.70	3.72	3.73	3.74	3.76	3.76
0.9	3.59	3.63	3.66	3.68	3.70	3.72	3.73	3.74	3.75	3.76
1.0	3.60	3.63	3.66	3.68	3.70	3.71	3.73	3.74	3.75	3.76

at even steps are small. If we use all values with step size $4k$ they are alternately large and small. At step size $2k$ and k we only use values at even step numbers, i.e. small values, and the order ratios will tend to oscillate. It may be a good idea to use every other value to get a smoother picture.

If the order ratios do not show any easily explainable pattern (cf. Table 16.3 for $x \leq 100$), then a reduction of the step size(s) may solve the problem. If not the necessary conclusion is that our basic hypothesis (10.25) does not hold in this region for this problem.

How much can we expect to gain from extrapolations? If we assume that the auxiliary functions have roughly the same magnitude then going from a first order to a second order result may almost double the number of correct decimals. A similar gain can be expected going from second to fourth order (when there is no third order term present). Going from second to third order will ‘only’ give 50 % more and only if the corresponding auxiliary function is well-behaved. So the main area of application is to first (and second) order methods, but of course here the need is also the greatest.

10.7 Test problem 3 – once again

We shall illustrate the techniques on test problem 3:

$$\begin{array}{rcll} u_t & = & u_{xx}, & 0 \leq x \leq 1, \quad t > 0, \\ u(0,x) = u_0(x) & = & \cos x, & 0 \leq x \leq 1, \\ u(t,1) & = & e^{-t}\cos 1, & \qquad\qquad t > 0, \end{array}$$

Table 10.4: k-ratio for asymmetric second order.

$t \setminus x$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.1	3.99	3.99	3.99	3.99	4.00	4.02	4.07	4.18	4.40	4.86
0.2	4.01	4.01	4.02	4.03	4.04	4.05	4.04	4.01	3.94	3.85
0.3	4.02	4.02	4.02	4.02	4.02	4.01	4.01	4.01	4.04	4.17
0.4	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.02	4.00	3.94
0.5	4.02	4.02	4.02	4.02	4.02	4.01	4.01	4.01	4.02	4.08
0.6	4.01	4.01	4.01	4.01	4.01	4.01	4.02	4.02	4.01	3.96
0.7	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.05
0.8	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.01	3.98
0.9	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.00	4.03
1.0	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.01	4.01	3.98

Table 10.5: h-ratio for symmetric second order with $h = k$.

$t \setminus x$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.1	15.94	15.94	15.94	15.95	15.98	16.05	16.21	16.55	17.22	18.58
0.2	16.02	16.03	16.05	16.07	16.11	16.13	16.12	16.03	15.84	15.53
0.3	16.07	16.06	16.06	16.05	16.04	16.03	16.02	16.03	16.12	16.47
0.4	16.05	16.05	16.05	16.05	16.05	16.06	16.07	16.06	16.00	15.82
0.5	16.05	16.05	16.05	16.05	16.05	16.04	16.03	16.03	16.05	16.22
0.6	16.04	16.04	16.04	16.04	16.04	16.04	16.05	16.06	16.03	15.91
0.7	16.04	16.04	16.04	16.04	16.04	16.04	16.03	16.03	16.03	16.14
0.8	16.04	16.04	16.04	16.04	16.04	16.04	16.04	16.05	16.04	15.95
0.9	16.03	16.03	16.03	16.03	16.03	16.03	16.03	16.02	16.02	16.10
1.0	16.03	16.03	16.03	16.03	16.03	16.03	16.03	16.04	16.04	15.96

$$u_x(t,0) \quad = \quad 0, \qquad\qquad t > 0.$$

We solve the problem numerically using Crank-Nicolson and begin with the first order boundary approximation. Using formulas (10.27) and (10.29) we check the order of the method calculating the ratios on a 10×10 grid using step sizes that are 16 times smaller. The results are shown in Table 10.1 and Table 10.2 for h and k respectively.

The method is clearly first order in h with only few values deviating slightly from 2.0. We deduce that the first order contribution to the error clearly dominates the other h-terms and we conclude that using a first order boundary approximation degrades the performance of Crank-Nicolson. The picture is more confusing for k where the second order is only convincing for larger values of t or x. The

Table 10.6: h-ratio for asymmetric second order with $h = k$.

$t \setminus x$	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.1	9.91	9.10	7.55	6.97	6.86	7.00	7.30	7.67	8.04	8.30
0.2	7.90	7.68	8.28	8.54	8.45	8.19	7.89	7.63	7.44	7.32
0.3	8.41	8.12	7.83	7.91	8.03	8.13	8.20	8.24	8.25	8.25
0.4	7.95	7.86	8.17	8.16	8.06	8.01	8.00	8.02	8.05	8.07
0.5	8.19	8.03	7.91	7.98	8.04	8.07	8.07	8.05	8.03	8.02
0.6	7.97	7.91	8.11	8.05	8.00	7.99	8.01	8.03	8.04	8.05
0.7	8.10	8.00	7.94	7.99	8.03	8.03	8.02	8.01	8.00	8.00
0.8	7.97	7.93	8.07	8.01	7.98	7.99	8.00	8.01	8.02	8.02
0.9	8.06	7.99	7.95	7.99	8.01	8.01	8.00	7.99	7.99	7.99
1.0	7.97	7.94	8.05	7.99	7.98	7.99	8.00	8.00	8.00	8.00

h-component of the error estimate is two orders of magnitude larger than the k-component and thus dominates the error. A comparison with the actual error shows that the estimate is less than 3 % over the actual error which has a maximum value of about 0.002.

Table 10.1 indicates that Richardson extrapolation is possible and the dominance of the h-component of the error estimate promises that it will be useful. The maximum error after extrapolation is 0.00003.

The k-component of the error estimate exhibits a strange behaviour for small x where also the order ratio indicated trouble. With a maximum value of 0.000005 this of no real concern, but it still indicates that we are doing something wrong. This something is probably the first order boundary approximation which not only reduces the h-order to 1 but also introduces a discontinuity in $c_x(t, x)$ at $(0, 0)$ since $c(0, x) = 0$ according to (9.18) and $c_x(t, 0) = -\frac{1}{2}e^{-t}$ according to (9.56) for our example. This will have an effect on some of the higher auxiliary functions such that $v_1 - v_4$ in (10.28) is not dominated by kd when x is close to 0.

The values for $c(t, x)$ as determined by $v_1 - v_2$ (see 10.26) with $h = k = 0.00625$ agree within 7 % with those obtained from solving the differential equation for c (which we did in Chapter 9) and a better agreement can be obtained using smaller step sizes. The corresponding determination of $g(t, x)$ is reasonably good when t and x are not too close to 0. In the regions where we have difficulty determining the order (see Table 10.2) we can of course have little trust in an application of formula (10.28) but in regions where the ratio (10.29) is between 3.0 and 5.0 the agreement is within 8 % with the step sizes chosen.

For the asymmetric second order boundary approximation (9.61) the second order in k is clearly detectable on the 10×10 grid using $h = 0.00625$ and $k = 0.025$ (see Table 10.4) and reasonably so for h (see Table 10.3). The values obtained here for $g(t, x)$ agree within 1 % with those previously obtained whereas the values for $f(t, x)$ are 10 - 20 % too small in agreement with the h-ratio being determined consistently 10 - 20 % too small. The presence of an interfering h^3-term in the error expansion is clearly noticeable here. The error estimate is dominated by the k-term (because of the larger step size) and the maximum error is now less than 0.000013. The number of grid points and thus the number of arithmetic operations is four times smaller than before and yet the error is much smaller (even after extrapolation of the foremr result).

For the symmetric second order approximation (9.69) we also expect second order accuracy. The order ratio (10.27) for h gives values between 3.99 and 4.01 and for k between 3.8 and 4.8 when using $h = k = 0.025$. These good results are due to the fact that the next terms in the error expansion (10.25) are h^4 and k^4 because of the symmetry and therefore interfere little with the second order terms. The number of grid points is again reduced by a factor 4 and the two components of the error estimate are both smaller than 0.000013. We know that $f = -g$ in theory. In practice they agree within 2 % with each other and with the values obtained from the independent solution of the differential equation.

Since the two error components thus almost cancel out the results are better than the estimates tell us. To check this we try equal step sizes with the second order boundary approximations. Using $h = k = 0.025$ we can confirm that the symmetric approximation now leads to a method which is fourth order (see Table 10.5) and that the asymmetric approximation (9.61) leads to a method which is third order in the common step size (see Table 10.6). The error estimate (and the error) is less than 2.2×10^{-6} for the asymmetric approximation and 0.6×10^{-9} for the symmetric one.

10.8 Which method to choose

This is a difficult one to answer because there is no method which is best in all situations; but we can issue some general guidelines. Usually the choice is between the explicit method, the implicit method, and Crank-Nicolson. The explicit method is the easiest one to program, but the stability condition, $2b\mu \leq 1$, will often imply such restrictions on the time step, as to render this option impractical. Crank-Nicolson seems optimal, being unconditionally stable and second order accurate. But in some cases it gives rise to oscillations which are damped very slowly. In such cases the implicit method becomes a viable alternative. It is, however, only first order accurate in t and in many cases it tends

to produce solutions that are too smooth. Using the extrapolation techniques in this chapter we can suggest the following alternative:
1. Use the implicit method.
2. Check the order in t, i.e. compute the order ratio.
3. If the results are first order then estimate the error.
4. Extrapolate to second order in t.
5. Check the orders in t and x.
6. Estimate the two error components.

In this way we can take advantage of the very stable behaviour of the implicit method and still get a second order result, provided the order ratio allows it.

As for boundary conditions with a derivative the theory in Chapter 9 as well as the previous example shows that we should prefer second order approximations to first order and symmetric approximations to asymmetric ones whenever possible. The work as measured by the number of arithmetic operations only grows marginally and the extra programming effort is well rewarded in better accuracy.

10.9 Literature

The idea to perform extrapolation and thus achieve a higher order goes back to the british mathematician Lewis F. Richardson who introduced it for the special case of $p = 2$ [30]. An application to parabolic equations is reported by Hartree and Womersley [16]. Extrapolation has also been used in [20] and [12] but they propose to extrapolate step by step and continue the difference scheme with the extrapolated values. The formula for determining or checking the order is given in lecture notes from the Technical University of Denmark [2] but the idea is probably older. The history of extrapolation processes in numerical analysis is given in the survey paper [17] which also contains an extensive bibliography.

10.10 Exercises

1. Solve problem 1 with the implicit method from $t = 0$ to $t = 0.5$ with $h = k = \frac{1}{10}, \frac{1}{20}, \frac{1}{40}$.

   For each point in the grid corresponding to the largest step size compute the order-ratio (10.7) and produce a table similar to Table 10.1.
   Deduce the order of the method.
   Estimate the leading term of the error for each point in the above grid and compare to the actual error.

2. Same as exercise 1 but with Crank-Nicolson's method.

3. Solve problem 1 with the implicit method from $t = 0$ to $t = 0.5$ with $h = k = \frac{1}{40}$.

   Using calculations with $(2k, h)$, $(4k, h)$, $(k, 2h)$ and $(k, 4h)$ you should gain information about the order in k and h as well as about the first terms in the expression for the error in all grid points corresponding to $4k$ and $4h$.

   What can we do to obtain an error which in every grid point is less than 10^{-6}?

4. Same as exercise 3 but with Crank-Nicolson's method.

5. Same as exercise 3 but with problem 2.

6. Same as exercise 5 but with Crank-Nicolson's method.

7. Using the results from exercise 3 perform Richardson extrapolation in k.
   Under the assumption that the order in time is now 2, what is the k-contribution to the error.
   How small should k be to reduce this error term to less than $\frac{1}{2}10^{-6}$ in every grid point?

8. Same as exercise 7 but with problem 2.

Chapter 11

Two Space Dimensions

11.1 Introduction

A general, linear, parabolic equation in two space dimensions:

$$u_t = b_1 u_{xx} + 2b_{12} u_{xy} + b_2 u_{yy} - a_1 u_x - a_2 u_y + \kappa u + \nu \tag{11.1}$$

is well-posed (cf. page 23) if

$$b_1 > 0 \quad \text{and} \quad b_1 b_2 > b_{12}^2. \tag{11.2}$$

The solution $u(t,x,y)$ is a function of t, x, and y, and the coefficients may also depend on t, x, and y.

To ensure a unique solution (11.1) must be supplemented by an initial condition

$$u(0,x,y) = u_0(x,y), \qquad X_1 \le x \le X_2, \quad Y_1 \le y \le Y_2 \tag{11.3}$$

and boundary conditions which might be of Dirichlet type such as

$$u(t,X_1,y) = u_{11}(t,y), \qquad t > 0 \tag{11.4}$$

or involving a normal derivative such as

$$\alpha u(t,X_1,y) - \beta u_x(t,X_1,y) = \gamma, \qquad t > 0 \tag{11.5}$$

and similar relations for $x = X_2$, $y = Y_1$, and $y = Y_2$.

We shall begin treating the case where there is no mixed derivative term in (11.1), i.e. that $b_{12} = 0$. We can then write the equation as

$$u_t = P_1 u + P_2 u + \nu \tag{11.6}$$

where

$$\begin{aligned} P_1 u &= b_1 u_{xx} - a_1 u_x + \theta \kappa u, && (11.7) \\ P_2 u &= b_2 u_{yy} - a_2 u_y + (1-\theta)\kappa u, && (11.8) \end{aligned}$$

and $0 \le \theta \le 1$. While symmetry considerations might speak for an even distribution of the κu-term ($\theta = \frac{1}{2}$) it is computationally simpler to use $\theta = 0$ or $\theta = 1$ i.e. to include the κu-term fully in one of the two operators.

For the numerical solution of (11.1) we choose a step size k in the t-direction and step sizes $h_1 = (X_2 - X_1)/L$ and $h_2 = (Y_2 - Y_1)/M$ in the x- and the y-direction, respectively, and seek the numerical solution v_{lm}^n on the discrete set of points $(nk, X_1 + lh_1, Y_1 + mh_2)$, $(l = 0, 1, \ldots, L;\ m = 0, 1, \ldots, M;\ n = 1, 2, \ldots, N)$.

11.2 The explicit method

In the simplest case $P_1 u = b_1 u_{xx}$ and $P_2 u = b_2 u_{yy}$ such that

$$u_t = b_1 u_{xx} + b_2 u_{yy}. \tag{11.9}$$

The explicit method for (11.9) looks like

$$\frac{v_{lm}^{n+1} - v_{lm}^n}{k} = b_1 \delta_x^2 v_{lm}^n + b_2 \delta_y^2 v_{lm}^n. \tag{11.10}$$

To study the stability of (11.10) we take v_{lm}^n on the form

$$v_{lm}^n = g^n e^{i\xi_1 l h_1} e^{i\xi_2 m h_2} = g^n e^{il\varphi_1} e^{im\varphi_2} \tag{11.11}$$

and insert in (11.10):

$$\begin{aligned} g &= 1 + b_1\mu_1(e^{i\varphi_1} - 2 + e^{-i\varphi_1}) + b_2\mu_2(e^{i\varphi_2} - 2 + e^{-i\varphi_2}) \\ &= 1 - 4b_1\mu_1 \sin^2\frac{\varphi_1}{2} - 4b_2\mu_2 \sin^2\frac{\varphi_2}{2} && (11.12) \end{aligned}$$

where $\mu_1 = k/h_1^2$ and $\mu_2 = k/h_2^2$. The stability requirement is $|g| \le 1$ and since g is real and clearly less than 1 the critical condition is $g \ge -1$ or

$$2b_1\mu_1 \sin^2\frac{\varphi_1}{2} + 2b_2\mu_2 \sin^2\frac{\varphi_2}{2} \le 1. \tag{11.13}$$

Since this must hold for all φ_1 and φ_2 the requirement for stability is

$$b_1\mu_1 + b_2\mu_2 \le \frac{1}{2}. \tag{11.14}$$

11.3 Implicit methods

Formula (11.14) puts severe restrictions on the step size k and it is very tempting to study generalizations of the implicit or the Crank-Nicolson methods to two space dimensions. A similar analysis will show that they are both unconditionally stable, i.e. they are stable for any choice of k, h_1, and h_2 (cf. exercise 1).

The use of an implicit method requires the solution of a set of linear equations for each time step. In one space dimension the coefficient matrix for these equations is tridiagonal, and the equations can be solved with a number of simple arithmetic operations (SAO) proportional to the number of internal grid points. The situation is less favourable in two or more space dimensions.

If we have $L-1$ internal points in the x-direction and $M-1$ internal points in the y-direction then we have $(L-1)(M-1)$ grid points and the same number of equations per time step. There are at most 5 non-zero coefficients in each equation. If the grid points are ordered lexicographically then the coefficient matrix will have a band structure such as shown to the left in Fig. 11.1 where the non-zero coefficients are marked with squares. During a Gaussian elimination the region between the outer bands will fill in as shown to the right in Fig. 11.1 resulting in a number of non-zeroes which is approximately $L^2 \cdot M$ and a number of SAO proportional to $L^3 \cdot M$.

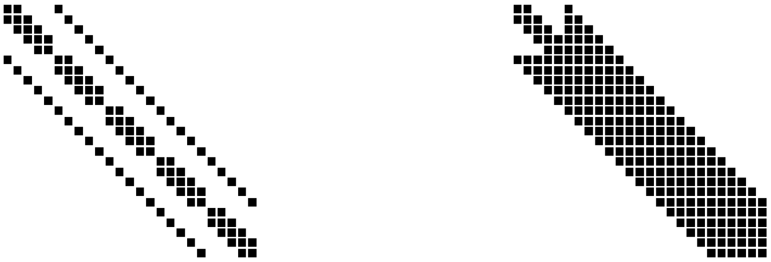


Figure 11.1: A coefficient matrix before and after elimination.

Example.
Consider $u_t = u_{xx} + u_{yy}$ on the unit square with $h_1 = h_2 = 0.01$, $L = M = 100$. If we want to integrate to $t = 1$ using the explicit method then stability requires $k \leq \frac{1}{4 \cdot 100^2}$ and a number of time steps at least $K = 4 \cdot 100^2$. The number of SAO is proportional to the number of grid points which is approximately $K \cdot L \cdot M = 4L^4 = 4 \cdot 10^8$.

If we want to use an implicit method we may choose $k = h_1 = h_2$, $K = L = M$ so the number of grid points is equal to L^3 but the number of SAO is proportional to $K \cdot L^3 \cdot M = L^5 = 10^{10}$.

We can conclude that there is no advantage in using implicit methods directly since the time involved in solving the linear equations outweighs the advantage of using larger step sizes in time. There are more elaborate ways to order the equations and to perform the solution process, but none that will make a significant difference in favour of implicit methods.

11.4 ADI methods

When the differential operator can be split as in (11.6) – (11.8) there are ways to avoid the L^3-factor. Such methods are called time-splitting methods or Locally One-Dimensional (LOD) or Alternating Direction Implicit (ADI) and the general idea is to split a time step in two and to take one operator or one space coordinate at a time.

Taking our inspiration from the Crank-Nicolson method we begin discretizing (11.6) in the time-direction:

$$\begin{aligned} u_t((n+\frac{1}{2})k, x, y) &= \frac{u^{n+1}-u^n}{k} + O(k^2), && (11.15)\\ P_1u + P_2u + \nu &= \frac{1}{2}P_1(u^{n+1}+u^n) + \frac{1}{2}P_2(u^{n+1}+u^n) && (11.16)\\ &\qquad + \frac{1}{2}(\nu^{n+1}+\nu^n) + O(k^2). \end{aligned}$$

Insert in (11.6), multiply by k, and rearrange:

$$\begin{aligned} (I - \frac{1}{2}kP_1 - \frac{1}{2}kP_2)u^{n+1} &= (I + \frac{1}{2}kP_1 + \frac{1}{2}kP_2)u^n && (11.17)\\ &\qquad + \frac{1}{2}k(\nu^{n+1}+\nu^n) + O(k^3). \end{aligned}$$

If we add $\frac{1}{4}k^2P_1P_2u^{n+1}$ on the left side and $\frac{1}{4}k^2P_1P_2u^n$ on the right side then we commit an error which is $O(k^3)$ and therefore can be included in that term:

$$\begin{aligned} (I - \frac{1}{2}kP_1)(I - \frac{1}{2}kP_2)u^{n+1} &= (I + \frac{1}{2}kP_1)(I + \frac{1}{2}kP_2)u^n && (11.18)\\ &\qquad + \frac{1}{2}k(\nu^{n+1}+\nu^n) + O(k^3). \end{aligned}$$

We now discretize in the space coordinates replacing P_1 by P_{1h}, P_2 by P_{2h}, and u by v:

$$\begin{aligned} (I - \frac{1}{2}kP_{1h}^{n+1})(I - \frac{1}{2}kP_{2h}^{n+1})v^{n+1} &= (I + \frac{1}{2}kP_{1h}^{n})(I + \frac{1}{2}kP_{2h}^{n})v^n \\ &\qquad + \frac{1}{2}k(\nu^{n+1}+\nu^n) && (11.19) \end{aligned}$$

and this gives rise to the Peaceman-Rachford method [27]:

$$
\begin{aligned}
(I - \frac{1}{2}kP_{1h}^{n+1})\tilde{v} &= (I + \frac{1}{2}kP_{2h}^{n})v^n + \alpha, &(11.20)\\
(I - \frac{1}{2}kP_{2h}^{n+1})v^{n+1} &= (I + \frac{1}{2}kP_{1h}^{n})\tilde{v} \ + \beta. &(11.21)
\end{aligned}
$$

The operators P_{1h} and P_{2h} involve the respective coefficients of the differential equation and may therefore depend on time. This dependence is indicated by superscripts n and $n+1$. The intermediate value, $\tilde{v}$, has no special relation to any particular intermediate time value. It has been introduced for reasons of computational efficiency and has no particular significance otherwise.

We have introduced the values α and β to take into account the inhomogeneous term ν because it is not evident how this term should be split. We shall attend to this matter shortly.

11.5 The Peaceman-Rachford method

In order to check whether the solution v^{n+1} to (11.20) and (11.21) is also the solution to (11.19) we start with v^{n+1} from (11.21) and apply the difference operators from the left side of (11.19):

$$
\begin{aligned}
(I - \frac{1}{2}kP_{1h}^{n+1})(I - \frac{1}{2}kP_{2h}^{n+1})v^{n+1} &= (I - \frac{1}{2}kP_{1h}^{n+1})\{(I + \frac{1}{2}kP_{1h}^{n})\tilde{v} + \beta\}\\
&= (I + \frac{1}{2}kP_{1h}^{n})(I - \frac{1}{2}kP_{1h}^{n+1})\tilde{v} + (I - \frac{1}{2}kP_{1h}^{n+1})\beta\\
&= (I + \frac{1}{2}kP_{1h}^{n})(I + \frac{1}{2}kP_{2h}^{n})v^n + (I + \frac{1}{2}kP_{1h}^{n})\alpha \qquad (11.22)\\
&\qquad + (I - \frac{1}{2}kP_{1h}^{n+1})\beta
\end{aligned}
$$

The first equal sign follows from (11.21), the third one from (11.20), and the second one requires that the operators P_{1h}^{n+1} and P_{1h}^{n} commute.

This is not always the case when the coefficients depend on t and x. A closer analysis reveals that we have commutativity if the coefficients b_1, a_1, and κ are either constant or only depend on t and y or only depend on x and y. If κ depends on all of t, x, and y we may incorporate it in the operator P_2. If a_1 and κ are 0 then b_1 may depend on both t and x (and y) if it can be written as a product of a function of t (and y) and a function of x (and y).

The operators P_1 and P_2 do not enter in a symmetric fashion. Therefore it may happen that we do not have commutativity for one but we do for the other. In this case we may switch freely between the x- and y-coordinates.

The main consequence of non-commutativity is that the ADI method (11.20) – (11.21) becomes first order in time instead of the expected second order.

Once commutativity is established we take a closer look at the inhomogeneous term. From (11.19) and (11.22) we have the requirement that

$$(I + \frac{1}{2}kP_{1h}^n)\alpha + (I - \frac{1}{2}kP_{1h}^{n+1})\beta \;=\; \frac{1}{2}k(\nu^{n+1} + \nu^n) \tag{11.23}$$

where a discrepancy of order $O(k^3)$ may be allowed with reference to a similar term in (11.18). There are three possible choices for α and β that will satisfy this:

$$\alpha \;=\; \frac{1}{2}k\nu^n, \qquad \beta \;=\; \frac{1}{2}k\nu^{n+1}, \tag{11.24}$$

$$\alpha \;=\; \beta \;=\; \frac{1}{4}k(\nu^{n+1} + \nu^n), \tag{11.25}$$

$$\alpha \;=\; \beta \;=\; \frac{1}{2}k\nu^{n+\frac{1}{2}}. \tag{11.26}$$

11.6 Practical considerations

The system of equations (11.20) for $\tilde{v}$ contains one equation for each interior grid point in the xy-region. The operator P_{1h} refers to neighbouring points in the x-direction and the resulting coefficient matrix becomes tridiagonal and we can therefore solve the system with a number of SAO proportional to the number of interior grid points. Similarly the system of equations (11.21) for v^{n+1} is effectively tridiagonal and can be solved at a similar cost irrespective of whether we reorder the grid points or not.

We shall now take a detailed look at how we set up and solve the two systems (11.20) and (11.21).

1. To compute the right-hand-side of (11.20) we need the values of v^n at all interior grid points and at all interior grid points on the boundaries $y = Y_1$ and $y = Y_2$. If we have Dirichlet conditions on these boundaries we know these values directly. If the boundary conditions involve the y-derivative on one or both of these boundary line segments then we use a (preferably second order) difference approximation to the derivative.
Cost: $\sim 5LM$ SAO.

2. To complete the system (11.20) we need information on $\tilde{v}$ for interior grid points on the boundary line segments $x = X_1$ and $x = X_2$. If P_{1h} does not depend on time then rearranging (11.21) and adding to (11.20) gives

$$2\tilde{v} \;=\; (I + \frac{1}{2}kP_{2h}^n)v^n + (I - \frac{1}{2}kP_{2h}^{n+1})v^{n+1} + \alpha - \beta. \tag{11.27}$$

If we have Dirichlet boundary conditions on $x = X_1$ and $x = X_2$ then we have information on v^n and v^{n+1} here and we can apply P_{2h}.
If the boundary conditions on one or both these lines involve the x-derivative but those on $y = Y_1$ and $y = Y_2$ are of Dirichlet type, then we might consider interchanging P_1 and P_2.
If the boundary conditions, however, are on the form $u_x(t, X_j, y) = f_j(t)$, ($j = 1$ and/or 2) where f_j does not depend on y then we can differentiate (11.27) w.r.t. x, and since the terms with P_{2h} vanish we end up with

$$\tilde{v}_x = \frac{1}{2}(f(t_n) + f(t_{n+1})) \qquad (11.28)$$

plus possible α- and β-terms.

Remark. Since $\tilde{v}$ is an intermediate value it is sometimes suggested to use

$$\tilde{v} = \frac{1}{2}(v^n + v^{n+1}) \qquad (11.29)$$

on the boundary. We cannot in general recommend this since $\tilde{v}$ has no particular relation to the intermediate time level $n + \frac{1}{2}$, but we note that it is after all an $O(k^2)$ approximation to (11.27) and although this is not quite good enough (11.29) might still come in handy in special cases. □

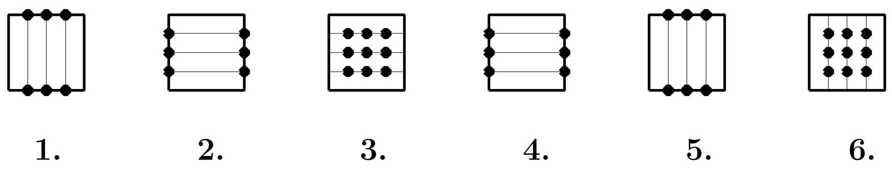
3. Solve system (11.20).
Cost: $\sim 8LM$ SAO.

4. To compute the right-hand-side of (11.21) we need the same values of $\tilde{v}$ for $x = X_1$ and $x = X_2$ as we discussed in **2.**
Cost: $\sim 5LM$ SAO.

5. To complete the system (11.21) we need information on v^{n+1} for $y = Y_1$ and $y = Y_2$. As in **1.** if we have Dirichlet boundary conditions we know these values directly. Otherwise we include equations involving difference approximations to the derivatives.

6. Solve system (11.21).
Cost: $\sim 8LM$ SAO.

Total cost: $\sim 26LM$ SAO per time step.



In the figure above we have visualized the considerations. The horizontal or vertical lines indicate which operator we are concerned with (P_{1h} or P_{2h}) and the bullets indicate which function values we are considering. In **1.** we are computing right-hand-side values based on v^n. In **2.** and **3.** we are computing $\tilde{v}$-values and in **4.** right-hand-side values based on these. Finally in **5.** and **6.** we compute values for v^{n+1}.

11.7 Stability of Peaceman-Rachford

We have derived the Peaceman-Rachford method on the basis of ideas from Crank-Nicolson so we expect unconditional stability, but we have also made a few minor alterations along the way so it is probably a good idea to perform an independent check. We shall do this for the special case $u_t = b_1 u_{xx} + b_2 u_{yy}$ with constant coefficients b_1 and b_2. Inserting

$$v_{lm}^n = g^n e^{il\varphi_1} e^{im\varphi_2} \quad \text{and} \quad \tilde{v}_{lm} = \tilde{g} v_{lm}^n \tag{11.30}$$

in (11.20) and (11.21) gives

$$(1 + 2b_1\mu_1 \sin^2 \frac{\varphi_1}{2})\tilde{g} = 1 - 2b_2\mu_2 \sin^2 \frac{\varphi_2}{2}, \tag{11.31}$$

$$(1 + 2b_2\mu_2 \sin^2 \frac{\varphi_2}{2})g = (1 - 2b_1\mu_1 \sin^2 \frac{\varphi_1}{2})\tilde{g}, \tag{11.32}$$

such that

$$g = \frac{1 - 2b_1\mu_1 \sin^2 \frac{\varphi_1}{2}}{1 + 2b_1\mu_1 \sin^2 \frac{\varphi_1}{2}} \cdot \frac{1 - 2b_2\mu_2 \sin^2 \frac{\varphi_2}{2}}{1 + 2b_2\mu_2 \sin^2 \frac{\varphi_2}{2}}. \tag{11.33}$$

For simplicity we introduce

$$x_1 = b_1\mu_1 \sin^2 \frac{\varphi_1}{2}, \qquad x_2 = b_2\mu_2 \sin^2 \frac{\varphi_2}{2} \tag{11.34}$$

and the formula for the growth factor now takes the simpler form

$$g = \frac{(1 - 2x_1)(1 - 2x_2)}{(1 + 2x_1)(1 + 2x_2)}. \tag{11.35}$$

Since $x_1 \geq 0$ and $x_2 \geq 0$ it is easily seen that $-1 \leq g \leq 1$ such that we indeed have unconditional stability. We also note that components with high frequency in both directions $\varphi_1 \sim \pi$, $\varphi_2 \sim \pi$ will have $g \sim 1$ (if $b_1\mu_1$ and $b_2\mu_2$ are large) so these components will not be damped very much and they will not alternate from one time step to the next. The growth factor might still take values close to -1 due to components with high frequency in one direction and low frequency in the other, and the well-known Crank-Nicolson oscillations will be observed when Peaceman-Rachford is used on problems with discontinuities in the initial condition.

11.8 D'Yakonov

There are other ways of splitting equation (11.19). D'Yakonov [39] has suggested

$$
\begin{aligned}
(I - \frac{1}{2}kP_{1h}^{n+1})\tilde{v} &= (I + \frac{1}{2}kP_{1h}^{n})(I + \frac{1}{2}kP_{2h}^{n})v^n + \alpha, \qquad (11.36)\\
(I - \frac{1}{2}kP_{2h}^{n+1})v^{n+1} &= \tilde{v} + \beta. \qquad (11.37)
\end{aligned}
$$

To check the equivalence we take the solution v^{n+1} from (11.37) and apply the difference operators from the left side of (11.19):

$$
\begin{aligned}
(I - \frac{1}{2}kP_{1h}^{n+1})(I - \frac{1}{2}kP_{2h}^{n+1})v^{n+1} &= (I - \frac{1}{2}kP_{1h}^{n+1})(\tilde{v} + \beta) \qquad (11.38)\\
= (I + \frac{1}{2}kP_{1h}^{n})(I + \frac{1}{2}kP_{2h}^{n})v^n &+ \alpha + (I - \frac{1}{2}kP_{1h}^{n+1})\beta
\end{aligned}
$$

In this case we have no problem with commutativity of the operators. As for the inhomogeneous term an obvious choice is

$$\beta = 0, \qquad \alpha = \frac{1}{2}k(\nu^{n+1} + \nu^n). \qquad (11.39)$$

When calculating the right-hand-side of (11.36) we must know v^n on *all* grid points including the boundaries and the corners. In addition we have in general a sum of 9 terms for each equation possibly with different coefficients so the cost for step **1.** is $\sim 17LM$ SAO.

Setting up system (11.36) requires $\tilde{v}$ on the interior points on the boundary segments $x = X_1$ and $x = X_2$. These values can be found by solving (11.37) from right to left if we have Dirichlet conditions on these boundary segments.

Solving equations (11.36) now costs $\sim 8LM$ SAO.

The right-hand-side of (11.37) is easy and so are the necessary boundary values of v^{n+1} on $y = Y_1$ and $y = Y_2$. The solution of (11.37) then costs another $\sim 8LM$ SAO, and the total cost of a time step with D'Yakonov amounts to $\sim 33LM$ SAO making D'Yakonov slightly more expensive than Peaceman-Rachford.

11.9 Douglas-Rachford

Other ADI methods can be derived from other basic schemes. If we for example take our inspiration from the implicit method and discretize in time we get

$$\frac{u^{n+1} - u^n}{k} = P_1u^{n+1} + P_2u^{n+1} + \nu^{n+1} + O(k) \qquad (11.40)$$

or

$$(I - kP_1 - kP_2)u^{n+1} \quad = \quad u^n + k\nu^{n+1} + O(k^2) \tag{11.41}$$

or

$$(I - kP_1)(I - kP_2)u^{n+1} \quad = \quad (I + k^2 P_1 P_2)u^n + k\nu^{n+1} + O(k^2) \tag{11.42}$$

where we in the last equation have incorporated $k^2 P_1 P_2(u^{n+1} - u^n)$ in the $O(k^2)$-term. (11.42) is now discretized in the space directions to

$$(I - kP_{1h}^{n+1})(I - kP_{2h}^{n+1})v^{n+1} \quad = \quad (I + k^2 P_{1h}^n P_{2h}^n)v^n + k\nu^{n+1} \tag{11.43}$$

and this formula can be split into the following two which are known as the Douglas-Rachford method [10]:

$$\begin{aligned} (I - kP_{1h}^{n+1})\tilde{v} \quad &= \quad (I + kP_{2h}^n)v^n + \alpha & (11.44) \\ (I - kP_{2h}^{n+1})v^{n+1} \quad &= \quad \tilde{v} - kP_{2h}^n v^n + \beta & (11.45) \end{aligned}$$

To check that v^{n+1} in (11.45) is also the solution to (11.43) we take v^{n+1} from (11.45) and apply the difference operators from (11.43):

$$\begin{aligned} (I - kP_{1h}^{n+1})(I - kP_{2h}^{n+1})v^{n+1} \quad &= \quad (I - kP_{1h}^{n+1})\{\tilde{v} - kP_{2h}^n v^n + \beta\} \\ &= \quad (I + kP_{2h}^n)v^n + \alpha - (I - kP_{1h}^{n+1})(kP_{2h}^n v^n - \beta) \\ &= \quad (I + k^2 P_{1h}^{n+1} P_{2h}^n)v^n + \alpha + (I - kP_{1h}^{n+1})\beta. & (11.46) \end{aligned}$$

The term with v^n on the right-hand-side is not exactly what it should be if P_1 depends on time, but the difference is $O(k^3)$ which is allowed.

In order to match the inhomogeneous term in (11.43) a natural choice for α and β would be $\alpha = k\nu^{n+1}$, $\beta = 0$.

One could question the relevance of the first order terms on the right-hand-side of (11.44) and (11.45). Actually the k^2-term in (11.42) could easily be incorporated in the $O(k^2)$-term and the result would be a simpler version of formula (11.43) which could be split into

$$\begin{aligned} (I - kP_{1h}^{n+1})\tilde{v} \quad &= \quad v^n + \alpha & (11.47) \\ (I - kP_{2h}^{n+1})v^{n+1} \quad &= \quad \tilde{v} + \beta & (11.48) \end{aligned}$$

where we again would suggest $\alpha = k\nu^{n+1}$, $\beta = 0$ in order to match a possible inhomogeneous term.

The practical considerations are dealt with as for Peaceman-Rachford or D'Yakonov. We just summarize the results for the computational work which is similar to Peaceman-Rachford for (11.44) – (11.45) and slightly less ($\sim 18LM$ SAO) for the simpler scheme (11.47) – (11.48).

11.10 Stability of Douglas-Rachford

Again we expect to inherit the unconditional stability of the implicit method but we had better check it directly. We again look at the special case $u_t = b_1 u_{xx} + b_2 u_{yy}$, and we use (11.30) and (11.34). From (11.44) – (11.45) we get

$$(1+4x_1)\tilde{g} \; = \; 1-4x_2, \qquad (1+4x_2)g \; = \; \tilde{g}+4x_2, \tag{11.49}$$

such that

$$g \; = \; \frac{1-4x_2+(1+4x_1)4x_2}{(1+4x_1)(1+4x_2)} \; = \; \frac{1+16x_1x_2}{1+4x_1+4x_2+16x_1x_2}. \tag{11.50}$$

Since $x_1 > 0$ and $x_2 > 0$ we have $0 < g < 1$ just like we hoped. For the simpler scheme (11.47) – (11.48) the result is

$$g \; = \; \frac{1}{(1+4x_1)(1+4x_2)} \tag{11.51}$$

which also ensures $0 < g < 1$. We mention in passing that the original implicit method would have given

$$g \; = \; \frac{1}{1+4x_1+4x_2} \tag{11.52}$$

For large values of x_1 and x_2, i.e. large values of $b_i\mu_i$ and $\varphi_i \approx \pi$, formula (11.50) gives $g \approx 1$ corresponding to weak damping whereas (11.51) gives $g \approx 0$ corresponding to strong damping and even stronger than with the implicit scheme (11.52). This may speak in favour of the simpler scheme (11.47) – (11.48).

11.11 The local truncation error

In order to check the local truncation error we use the symbols of the differential and difference operators (cf. section 3.3). We consider the simple equation

$$u_t - b_1 u_{xx} - b_2 u_{yy} = \nu \tag{11.53}$$

with constant coefficients and use the test functions

$$v_{lm}^n \; = \; e^{snk} e^{ilh_1\xi_1} e^{imh_2\xi_2} \; = \; e^{st} e^{ix\xi_1} e^{iy\xi_2}, \qquad \tilde{v}_{lm} \; = \; \tilde{g} v_{lm}^n. \tag{11.54}$$

The symbol for the differential operator in (11.53) is

$$p(s,\xi_1,\xi_2) \; = \; s + b_1\xi_1^2 + b_2\xi_2^2. \tag{11.55}$$

We first look at the simple scheme (11.47) – (11.48) where we get

$$\begin{aligned}(1+4b_1\frac{k}{h_1^2}\sin^2\frac{h_1\xi_1}{2})\tilde{g} &= 1+k\nu^{n+1}, &(11.56)\\ (1+4b_2\frac{k}{h_2^2}\sin^2\frac{h_2\xi_2}{2})e^{sk} &= \tilde{g} &(11.57)\end{aligned}$$

or

$$(1+4b_1\frac{k}{h_1^2}\sin^2\frac{h_1\xi_1}{2})(1+4b_2\frac{k}{h_2^2}\sin^2\frac{h_2\xi_2}{2})e^{sk}-1 \;=\; k\nu^{n+1}. \qquad (11.58)$$

The left-hand-side contains the terms which originate from the operator $P_{k,h}$ and the right-hand-side refers to $R_{k,h}$. Since Douglas-Rachford is derived from the implicit method the natural expansion point is at $t=(n+1)k$. With this choice the right-hand-side operator $R_{k,h}$ becomes the identity and the corresponding symbol

$$r_{k,h}(s,\xi_1,\xi_2) \;=\; 1. \qquad (11.59)$$

This also means that we should divide the left-hand-side of (11.58) by e^{sk} before Taylor expansion which then gives

$$(1+kb_1\xi_1^2+O(kh_1^2))(1+kb_2\xi_2^2+O(kh_2^2))-(1-sk+\frac{1}{2}s^2k^2+O(k^3)). \quad (11.60)$$

Before we begin checking orders we should remember that we have multiplied by k in order to get formula (11.41) and the following formulae. Therefore we must divide (11.60) by k in order to get back to the standard form and we now get

$$p_{k,h}(s,\xi_1,\xi_2) \;=\; b_1\xi_1^2+b_2\xi_2^2+kb_1b_2\xi_1^2\xi_2^2+s-\frac{1}{2}s^2k+O(k^2+h_1^2+h_2^2). \quad (11.61)$$

We now combine (11.55), (11.59) and (11.61) in

$$p_{k,h}-r_{k,h}p \;=\; k(b_1b_2\xi_1^2\xi_2^2-\frac{1}{2}s^2)+O(k^2+h_1^2+h_2^2). \qquad (11.62)$$

Formula (11.62) shows that the Simple Douglas-Rachford (SDR) scheme (11.47) – (11.48) is indeed first order in time and second order in the space variables as we would expect for a scheme derived from the implicit method. We note that w.r.t. the order of the local truncation error it is not important whether we compute the inhomogeneous term at $t=(n+1)k$ or at $t=nk$. It might have an effect on the size of the error, though.

For the Traditional Douglas-Rachford (TDR) scheme we have instead of (11.58)

$$\begin{aligned}&(1+4b_1\frac{k}{h_1^2}\sin^2\frac{h_1\xi_1}{2})(1+4b_2\frac{k}{h_2^2}\sin^2\frac{h_2\xi_2}{2})e^{sk}-(1-4b_2\frac{k}{h_2^2}\sin^2\frac{h_2\xi_2}{2})\\ &\quad-(1+4b_1\frac{k}{h_1^2}\sin^2\frac{h_1\xi_1}{2})\cdot 4b_2\frac{k}{h_2^2}\sin^2\frac{h_2\xi_2}{2} \;=\; k\nu^{n+1}. \qquad (11.63)\end{aligned}$$

With the expansion point at $t = (n+1)k$ we again get $r_{k,h} = 1$ and using a Taylor expansion of the left hand side:

$$\begin{aligned}&(1 + kb_1\xi_1^2 + O(kh_1^2))(1 + kb_2\xi_2^2 + O(kh_2^2)) \\ &- (1 + (kb_1\xi_1^2 + O(kh_1^2))kb_2\xi_2^2 + O(kh_2^2))(1 - sk + \frac{1}{2}s^2k^2 + O(k^3)).\end{aligned} \tag{11.64}$$

The symbol of the difference operator becomes

$$\begin{aligned}p_{k,h}(s,\xi_1,\xi_2) &= s + b_1\xi_1^2 + b_2\xi_2^2 + k(b_1b_2\xi_1^2\xi_2^2 - \frac{1}{2}s^2 - b_1b_2\xi_1^2\xi_2^2) \\ &\qquad + O(k^2 + h_1^2 + h_2^2)\end{aligned} \tag{11.65}$$

and the local truncation error is

$$p_{k,h} - r_{k,h}p = -\frac{1}{2}ks^2 + O(k^2 + h_1^2 + h_2^2). \tag{11.66}$$

This result looks more elegant than (11.62) but whether the error becomes smaller is quite a different matter.

For Peaceman-Rachford (PR) and D'Yakonov the formula corresponding to (11.60) and (11.64) is

$$\begin{aligned}&(1 + \frac{1}{2}kb_1\xi_1^2 + O(kh_1^2))(1 + \frac{1}{2}kb_2\xi_2^2 + O(kh_2^2))(1 + sk + \frac{1}{2}s^2k^2 + O(k^3)) \\ &\qquad - (1 - \frac{1}{2}kb_1\xi_1^2 + O(kh_1^2))(1 - \frac{1}{2}kb_2\xi_2^2 + O(kh_2^2))\end{aligned} \tag{11.67}$$

such that

$$\begin{aligned}p_{k,h}(s,\xi_1,\xi_2) &= s + \frac{1}{2}b_1\xi_1^2 + \frac{1}{2}b_2\xi_2^2 + \frac{1}{2}b_1\xi_1^2 + \frac{1}{2}b_2\xi_2^2 \\ &\quad + k(\frac{1}{2}b_1s\xi_1^2 + \frac{1}{2}b_2s\xi_2^2 + \frac{1}{4}b_1b_2\xi_1^2\xi_2^2 + \frac{1}{2}s^2 - \frac{1}{4}b_1b_2\xi_1^2\xi_2^2) \\ &\quad + O(k^2 + h_1^2 + h_2^2).\end{aligned} \tag{11.68}$$

The exact expression for $r_{k,h}(s,\xi_1,\xi_2)$ depends on which one of the choices (11.24) – (11.26) we select, but up to $O(k^2)$, and $O(h_1^2)$ in case of (11.24), we get

$$r_{k,h}(s,\xi_1,\xi_2) = 1 + \frac{1}{2}sk + O(k^2 + h_1^2). \tag{11.69}$$

We now combine (11.55), (11.68), and (11.69):

$$\begin{aligned}p_{k,h} - r_{k,h}p &= s + b_1\xi_1^2 + b_2\xi_2^2 + \frac{1}{2}k(b_1s\xi_1^2 + b_2s\xi_2^2 + s^2) \\ &\quad - (1 + \frac{1}{2}sk)(s + b_1\xi_1^2 + b_2\xi_2^2) + O(k^2 + h_1^2 + h_2^2) \\ &= O(k^2 + h_1^2 + h_2^2)\end{aligned} \tag{11.70}$$

showing that Peaceman-Rachford is indeed second order accurate at least for the simple equation (11.53) with constant coefficients. Extending the result to lower order terms presents no problem but if the coefficients are allowed to vary with space and time we have a more complicated picture as mentioned in the discussion on page 107.

11.12 The global error

In order to study the global error we introduce a set of auxiliary functions which may depend on (t, x, y) but not on (k, h_1, h_2), and we assume that the numerical solution can be written as

$$v \quad = \quad u - h_1c_1 - h_2c_2 - kd - h_1ke_1 - h_2ke_2 - h_1^2f_1 - h_2^2f_2 - k^2g - \cdots \qquad (11.71)$$

We need a similar assumption for the intermediate values

$$\tilde{v} \quad = \quad \tilde{u} - h_1\tilde{c}_1 - h_2\tilde{c}_2 - k\tilde{d} - h_1k\tilde{e}_1 - h_2k\tilde{e}_2 - h_1^2\tilde{f}_1 - h_2^2\tilde{f}_2 - k^2\tilde{g} - \cdots \qquad (11.72)$$

and we shall seek information on these auxiliary functions. We shall assume Dirichlet boundary conditions and therefore have homogeneous side conditions for all the auxiliary functions. Beginning with the simple version of Douglas-Rachford (11.47) – (11.48) we have

$$\begin{aligned}(I - kP_{1h}^{n+1})\tilde{v} \quad &= \quad v^n + k\nu^{n+1} &(11.73)\\ (I - kP_{2h}^{n+1})v^{n+1} \quad &= \quad \tilde{v} &(11.74)\end{aligned}$$

where P_{1h} and P_{2h} are discretized versions of P_1 and P_2 from (11.7) – (11.8). Using Taylor expansion we have

$$\begin{aligned}P_{2h}u \quad &= \quad (b_2\delta_y^2 - a_2\tilde{\mu}\delta_y + (1-\theta)\kappa)u\\ &= \quad b_2u_{yy} - a_2u_y + (1-\theta)\kappa u + \frac{1}{12}b_2h_2^2u_{4y} - \frac{1}{6}a_2h_2^2u_{yyy} + O(h_2^4)\\ &= \quad P_2u + \frac{1}{12}b_2h_2^2u_{4y} - \frac{1}{6}a_2h_2^2u_{yyy} + O(h_2^4) \qquad (11.75)\end{aligned}$$

and similarly for the auxiliary functions and for P_1.
Inserting (11.71) on the left-hand-side of (11.74) and applying (11.75) we have with $t = (n+1)k$ as expansion point

$$\begin{aligned}(I - kP_{2h})v^{n+1} \quad &= \quad (I - kP_{2h})(u - h_1c_1 - \cdots - k^2g) + O(\cdots)\\ &= \quad (I - kP_2)(u - h_1c_1 - \cdots - k^2g) \qquad (11.76)\\ &\quad - \frac{1}{12}b_2kh_2^2u_{4y} + \frac{1}{6}a_2kh_2^2u_{yyy} + O(\cdots).\end{aligned}$$

$O(\cdots)$ shall here and in the following indicate third order terms in k, h_1, and h_2 and will therefore include the two terms with u_{4y} and u_{yyy}.
Inserting (11.72) on the right-hand-side of (11.74) and equating terms we get

$$\tilde{u} = u, \quad \tilde{c}_1 = c_1, \quad \tilde{c}_2 = c_2, \quad \tilde{f}_1 = f_1, \quad \tilde{f}_2 = f_2 \tag{11.77}$$

together with

$$\begin{aligned} \tilde{d} &= d + P_2 u, & (11.78)\\ \tilde{e}_1 &= e_1 - P_2 c_1, & (11.79)\\ \tilde{e}_2 &= e_2 - P_2 c_2, & (11.80)\\ \tilde{g} &= g - P_2 d. & (11.81) \end{aligned}$$

Next we insert (11.72) on the left-hand-side of (11.73)

$$\begin{aligned} (I - kP_{1h})\tilde{v} &= (I - kP_{1h})(\tilde{u} - h_1\tilde{c}_1 - \cdots - k^2\tilde{g}) + O(\cdots) \\ &= (I - kP_1)(\tilde{u} - h_1\tilde{c}_1 - \cdots - k^2\tilde{g}) \qquad (11.82) \\ &\quad - \frac{1}{12} b_1 k h_1^2 u_{4x} + \frac{1}{6} a_1 k h_1^2 u_{xxx} + O(\cdots) \end{aligned}$$

For the right-hand-side of (11.73) we must remember that the expansion point is at $t = (n+1)k$ and that v^n therefore is one time step earlier:

$$\begin{aligned} v^n + k\nu^{n+1} &= (u - h_1 c_1 - \cdots - k^2 g) + k\nu \qquad (11.83) \\ &\quad - k u_t + h_1 k c_{1t} + h_2 k c_{2t} + k^2 d_t + \frac{1}{2} k^2 u_{tt} + O(\cdots) \end{aligned}$$

Equating terms in (11.82) and (11.83) confirms (11.77) and adds

$$\begin{aligned} \tilde{d} + P_1\tilde{u} &= d + u_t - \nu, & (11.84)\\ \tilde{e}_1 - P_1\tilde{c}_1 &= e_1 - c_{1t}, & (11.85)\\ \tilde{e}_2 - P_1\tilde{c}_2 &= e_2 - c_{2t}, & (11.86)\\ \tilde{g} - P_1\tilde{d} &= g - d_t - \frac{1}{2}u_{tt}. & (11.87) \end{aligned}$$

Comparing (11.78) and (11.84) and remembering that $\tilde{u} = u$ we have

$$d + P_2 u = d + u_t - P_1 u - \nu$$

or

$$u_t - P_1 u - P_2 u = \nu \tag{11.88}$$

confirming the consistency of our assumptions.
(11.79) and (11.85) together with (11.77) gives

$$e_1 - P_2 c_1 = e_1 - c_{1t} + P_1 c$$

or

$$c_{1t} - P_1 c_1 - P_2 c_1 \; = \; 0. \tag{11.89}$$

From (11.80) and (11.86) we get a similar equation for c_2 and since the side conditions are also homogeneous we may conclude that $c_1 \equiv c_2 \equiv 0$.
From (11.81) and (11.87) we get using (11.78)

$$g - P_2 d \; = \; g - d_t + P_1 \tilde{d} - \frac{1}{2} u_{tt} \; = \; g - d_t + P_1 d - \frac{1}{2} u_{tt} + P_1 P_2 u$$

or

$$d_t - P_1 d - P_2 d \; = \; -\frac{1}{2} u_{tt} + P_1 P_2 u \tag{11.90}$$

showing that $d(t,x,y)$ is not identically 0 and that the error therefore is first order in k. If we continue in this way we shall see that f_1 and f_2 are also different from 0 and that the error therefore is $O(k + h_1^2 + h_2^2)$ as we would expect.

We might note here that we have multiplied by k in order to get to equations (11.73) – (11.74) and this is the reason why we only get information about the auxiliary functions corresponding to the first order terms in (11.71) even though we compare terms up to and including second order.

For the traditional Douglas-Rachford scheme (11.44) – (11.45) the equations are

$$\begin{aligned} (I - kP_{1h})\tilde{v} \; &= \; (I + kP_{2h})v^n + k\nu^{n+1}, && (11.91) \\ (I - kP_{2h})v^{n+1} \; &= \; \tilde{v} - kP_{2h}v^n. && (11.92) \end{aligned}$$

The extra term which has been added in (11.91) and subtracted in (11.92) is

$$\begin{aligned} kP_{2h}v^n \; &= \; kP_{2h}(u^n - h_1 c_1 - h_2 c_2 - kd) + O(\cdots) \\ &= \; kP_2(u^n - h_1 c_1 - h_2 c_2 - kd) + O(\cdots) \\ &= \; kP_2(u^{n+1} - h_1 c_1 - h_2 c_2 - kd) - k^2 P_2 u_t + O(\cdots) && (11.93) \end{aligned}$$

where the last term is due to the expansion point being at $t = (n+1)k$.
Equating terms in (11.92) gives the equalities (11.77) together with

$$\tilde{d} = d, \;\; \tilde{e}_1 = e_1, \;\; \tilde{e}_2 = e_2, \;\; \tilde{g} = g + P_2 u_t \tag{11.94}$$

and from (11.91) we get

$$\begin{aligned} \tilde{d} \; + P_1 \tilde{u} \; &= \; d \; + u_t \; - P_2 u - \nu, && (11.95) \\ \tilde{e}_1 - P_1 \tilde{c}_1 \; &= \; e_1 - c_{1t} + P_2 c_1, && (11.96) \\ \tilde{e}_2 - P_1 \tilde{c}_2 \; &= \; e_2 - c_{2t} + P_2 c_2, && (11.97) \\ \tilde{g} \; - P_1 \tilde{d} \; &= \; g \; - d_t \; + P_2 d - \frac{1}{2} u_{tt} + P_2 u_t. && (11.98) \end{aligned}$$

The first three imply (11.88), (11.89), and its analogue such that we again can conclude that $c_1 \equiv c_2 \equiv 0$.
From (11.94) and (11.98) we finally deduce that

$$d_t - P_1 d - P_2 d \;=\; -\frac{1}{2} u_{tt} \tag{11.99}$$

in accordance with our expectations that Douglas-Rachford is first order in time. We note that $\tilde{v}$ now is a rather good approximation to v^{n+1}, but whether v^{n+1} now is a better or worse approximation to u is impossible to decide.

A similar analysis can be performed for Peaceman-Rachford and D'Yakonov to show that these methods are indeed second order in all three step sizes.

In practice we can check the orders and estimate the various contributions to the error using the methods from Chapter 10 taking each step size separately.

11.13 Exercises

1. Calculate the growth factor, g, for the implicit method and Crank-Nicolson on (11.9) and show that both methods are unconditionally stable.

2. Investigate the stability of D'Yakonov's method (11.36) – (11.37) when applied to (11.9).

3. Solve (11.9) with $b_1 = b_2 = 1$ on the unit square $0 \leq x \leq 1$, $0 \leq y \leq 1$ and for $0 \leq t \leq 0.5$ using the Simple Douglas-Rachford (SDR) method (11.47) – (11.48) with $h_1 = h_2 = k = \frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{40}$.
   Initial and boundary values are taken from the true solution
   $u(t, x, y) = e^{-4t} \sin(x - y) \cos(x + y)$.
   Compute the max-norm and the 2-norm of the error for
   $t = 0.1$, 0.2, 0.3, 0.4, 0.5.

4. Solve the problem from the previous exercise using the Traditional Douglas-Rachford (TDR) method (11.44) – (11.45).

5. Solve the problem from the previous exercise using the Peaceman-Rachford (PR) method.

6. Solve the problem from exercise 3 with SDR and with $h_1 = h_2 = k = \frac{1}{40}$.

   Using calculations with $2h_1$, $4h_1$, resp. $2h_2$ and $4h_2$, resp. $2k$, $4k$ you should gain information about the order in h_1, h_2, and k as well as about the first terms in the expression for the error in all grid points corresponding to $4h_1$, $4h_2$, $4k$ at $t = 1$.

7. Same as above with TDR.

8. Same as above with PR.

9. Using the results from exercise 6 perform Richardson extrapolation in k. Under the assumption that the order in time is now 2, what is the k-contribution to the error.

10. Same as above but with TDR.

Chapter 12

Equations with Mixed Derivative Terms

We now return to the general equation (11.1). As a difference approximation to u_{xy} we shall use

$$\begin{aligned} \delta^2_{xy} v_{lm} &= \tilde{\mu}_x \delta_x (\tilde{\mu}_y \delta_y v_{lm}) \;=\; \tilde{\mu}_x \delta_x \Big(\frac{v_{l,m+1} - v_{l,m-1}}{2h_2}\Big) \\ &= \frac{1}{4h_1 h_2}(v_{l+1,m+1} - v_{l-1,m+1} - v_{l+1,m-1} + v_{l-1,m-1}) \qquad (12.1) \\ &= \tilde{\mu}_y \delta_y (\tilde{\mu}_x \delta_x v_{lm}) \;=\; \delta^2_{yx} v_{lm}. \end{aligned}$$

There is no obvious way of splitting the mixed derivative or difference operator between the two operators P_1 and P_2 in (11.6) so we shall instead treat the mixed derivative term in a way analogous to what we did for the inhomogeneous term.

The first scheme we consider is the Simple Douglas-Rachford scheme (11.47) – (11.48) where α and β now should be chosen to take care of the mixed derivative term (in addition to a possible inhomogeneous term which we shall disregard here).

Following the analysis on page 112 we shall select α and β such that

$$\alpha + (I - kP^{n+1}_{1h})\beta \;=\; 2kb_{12}\delta^2_{xy}v^{n+1} + O(k^2). \qquad (12.2)$$

There are three straightforward choices for α and β which will satisfy (12.2):

$$\alpha \;=\; \beta \;=\; kb_{12}\delta^2_{xy}v^n, \qquad (12.3)$$

$$\alpha \;=\; kb_{12}\delta^2_{xy}v^n, \qquad \beta \;=\; kb_{12}\delta^2_{xy}\tilde{v}, \qquad (12.4)$$

$$\alpha \;=\; 2kb_{12}\delta^2_{xy}v^n, \qquad \beta \;=\; 0. \qquad (12.5)$$

For the Traditional Douglas-Rachford scheme the condition is the same so the same three possibilities for α and β apply.

For the Peaceman-Rachford scheme (11.20) – (11.21) we would aim at

$$(I + \frac{1}{2}kP_{1h}^{n})\alpha + (I - \frac{1}{2}kP_{1h}^{n+1})\beta \;=\; kb_{12}\delta_{xy}^{2}(v^{n+1} + v^{n}) + O(k^{3}). \tag{12.6}$$

This is a bit more difficult to achieve. Two obvious suggestions are (12.3) and (12.4) but they are not quite accurate enough and the resulting method becomes only first order in time.

Formula (12.3) would be good enough if we could replace v^n by an intermediate value $v^{n+\frac{1}{2}}$. An approximation to a value at time $t = (n + \frac{1}{2})k$ can be obtained by extrapolation from values at $t = (n-1)k$ and $t = nk$:

$$\hat{v}^{n+\frac{1}{2}} \;=\; v^{n} + \frac{1}{2}(v^{n} - v^{n-1}) \tag{12.7}$$

and a good suggestion for α and β is now

$$\alpha \;=\; \beta \;=\; kb_{12}\delta_{xy}^{2}\hat{v}^{n+\frac{1}{2}}. \tag{12.8}$$

In the same manner (12.4) would be good if we could replace $\tilde{v}$ by v^{n+1}. An extrapolated value would here be $v^n + (v^n - v^{n-1})$ such that we have

$$\alpha \;=\; kb_{12}\delta_{xy}^{2}v^{n}, \qquad \beta \;=\; kb_{12}\delta_{xy}^{2}(2v^{n} - v^{n-1}) \tag{12.9}$$

and both (12.8) and (12.9) would lead to schemes which are second order in time.

12.1 Practical considerations

We now have about a dozen different combinations but they are not all equally good. The first point to consider is how to get boundary values at $x = X_1$ and $x = X_2$ for $\tilde{v}$. For the Simple Douglas-Rachford scheme (11.47) – (11.48) we solve (11.48) to get

$$\tilde{v} \;=\; (I - kP_{2h}^{n+1})v^{n+1} - \beta \tag{12.10}$$

If we choose formula (12.5) then $\beta = 0$ and (12.10) can be used as it stands.

If we choose formula (12.3) then the β-term involves $\delta_{xy}^{2}v^{n}$ which cannot be calculated at the boundary points. There are two ways around this.

1. Take the difference at the nearest neighbour point, e.g.

$$\delta_{xy}^{2}v_{0m}^{n} \;:=\; \delta_{xy}^{2}v_{1m}^{n}$$

This will introduce an $O(h_1)$-error and that is not ideal.

2. Use a linear extrapolation in the x-direction, e.g.

$$\delta_{xy}^2 v_{0m}^n \;:=\; 2\delta_{xy}^2 v_{1m}^n - \delta_{xy}^2 v_{2m}^n .$$

Formula (12.4) presents even bigger problems since we cannot calculate $\delta_{xy}^2 \tilde{v}$ at the neighbouring points to any of the boundaries. We suggest either to use extrapolation as above or to use (11.29) which is accurate enough here.

For the Traditional Douglas-Rachford scheme (11.44) – (11.45) the formula for $\tilde{v}$ at the x-boundaries is

$$\tilde{v} \;=\; (I - kP_{2h}^{n+1})v^{n+1} + kP_{2h}^n v^n - \beta \tag{12.11}$$

and the same considerations apply.

For the Peaceman-Rachford scheme (11.20)-(11.21) the formula for $\tilde{v}$ is (11.27) and this clearly favours the case where $\alpha = \beta$ so that (12.3) and (12.8) are ideal choices. (12.4), (12.5), and (12.9) can be tackled using the suggestions above.

12.2 Stability with mixed derivative

We shall study stability requirements in relation to the differential equation

$$u_t \;=\; b_1 u_{xx} + 2b_{12}u_{xy} + b_2 u_{yy} \tag{12.12}$$

with the discretization

$$\begin{aligned} \frac{v_{lm}^{n+1} - v_{lm}^n}{k} \;=\; & (1-\theta)(b_1\delta_x^2 + 2b_{12}\delta_{xy}^2 + b_2\delta_y^2)v_{lm}^n \\ & +\theta(b_1\delta_x^2 + 2b_{12}\delta_{xy}^2 + b_2\delta_y^2)v_{lm}^{n+1} \end{aligned} \tag{12.13}$$

where $\theta = 0$, 0.5, and 1 corresponds to the explicit, the Crank-Nicolson, and the implicit method, respectively. We put the discretized solution on the form

$$v_{lm}^n \;=\; e^{snk}e^{i\xi_1 lh_1}e^{i\xi_2 mh_2} \;=\; g^n e^{il\varphi_1}e^{im\varphi_2} \tag{12.14}$$

and use the abbreviations

$$x_1 \;=\; b_1\mu_1 \sin^2\frac{\varphi_1}{2}, \tag{12.15}$$

$$x_2 \;=\; b_2\mu_2 \sin^2\frac{\varphi_2}{2}, \tag{12.16}$$

$$x_{12} \;=\; b_{12}\mu_{12}\sin\varphi_1 \sin\varphi_2. \tag{12.17}$$

Remember that the condition for (12.12) to be well-posed is

$$b_{12}^2 \; < \; b_1 b_2 \tag{12.18}$$

together with $b_1 > 0$ and $b_2 > 0$. It follows that

$$b_{12}^2 \mu_{12}^2 \; = \; b_{12}^2 \frac{k^2}{h_1^2 h_2^2} \; < \; b_1 b_2 \frac{k^2}{h_1^2 h_2^2} \; = \; b_1 \mu_1 b_2 \mu_2 \tag{12.19}$$

or

$$|b_{12}| \mu_{12} \; < \; \sqrt{b_1 \mu_1 b_2 \mu_2}. \tag{12.20}$$

We also have

$$0 \; \leq \; (\sqrt{b_1 \mu_1} - \sqrt{b_2 \mu_2})^2 \; = \; b_1 \mu_1 + b_2 \mu_2 - 2\sqrt{b_1 \mu_1 b_2 \mu_2}. \tag{12.21}$$

Combining (12.20) and (12.21) we get

$$2|b_{12}|\mu_{12} \; < \; 2\sqrt{b_1 \mu_1 b_2 \mu_2} \; \leq \; b_1 \mu_1 + b_2 \mu_2. \tag{12.22}$$

Similarly we have

$$\begin{aligned} 0 \; &\leq \; (\frac{\sqrt{b_1}}{h_1} \sin \frac{\varphi_1}{2} - \frac{\sqrt{b_2}}{h_2} \sin \frac{\varphi_2}{2})^2 \qquad (12.23) \\ &= \; \frac{b_1}{h_1^2} \sin^2 \frac{\varphi_1}{2} + \frac{b_2}{h_2^2} \sin^2 \frac{\varphi_2}{2} - 2 \frac{\sqrt{b_1 b_2}}{h_1 h_2} \sin \frac{\varphi_1}{2} \sin \frac{\varphi_2}{2}. \end{aligned}$$

Use (12.18), multiply by k and rearrange

$$2|b_{12}| \; \mu_{12} \; |\sin \frac{\varphi_1}{2} \sin \frac{\varphi_2}{2}| \; \leq \; b_1 \mu_1 \sin^2 \frac{\varphi_1}{2} + b_2 \mu_2 \sin^2 \frac{\varphi_2}{2}. \tag{12.24}$$

Now multiply by $4|\cos \frac{\varphi_1}{2} \cos \frac{\varphi_2}{2}| \leq 4$ and get

$$2|b_{12}| \; \mu_{12} \; |\sin \varphi_1 \sin \varphi_2| \; \leq \; 4 b_1 \mu_1 \sin^2 \frac{\varphi_1}{2} + 4 b_2 \mu_2 \sin^2 \frac{\varphi_2}{2} \tag{12.25}$$

or

$$2|x_{12}| \; \leq \; 4x_1 + 4x_2. \tag{12.26}$$

Furthermore we have

$$\begin{aligned} x_{12}^2 \; &= \; b_{12}^2 \mu_{12}^2 \sin^2 \varphi_1 \sin^2 \varphi_2 \\ &\leq \; 16 b_1 b_2 \mu_1 \mu_2 \sin^2 \frac{\varphi_1}{2} \cos^2 \frac{\varphi_1}{2} \sin^2 \frac{\varphi_2}{2} \cos^2 \frac{\varphi_2}{2} \\ &= \; 16 x_1 x_2 \cos^2 \frac{\varphi_1}{2} \cos^2 \frac{\varphi_2}{2} \; \leq \; 16 x_1 x_2. \qquad (12.27) \end{aligned}$$

For the explicit scheme the growth factor becomes

$$g \;=\; 1 - 4x_1 - 4x_2 - 2x_{12} \tag{12.28}$$

and for stability we require $-1 \leq g \leq 1$. $g \leq 1$ is equivalent to (12.26) and is therefore satisfied for a well-posed problem. $g \geq -1$ is equivalent to

$$2x_1 + 2x_2 + x_{12} \;\leq\; 1 \tag{12.29}$$

This inequality must be fulfilled for all φ_1 and φ_2, in particular for $\varphi_1 = \varphi_2 = \pi$ where it reduces to

$$2b_1\mu_1 + 2b_2\mu_2 \;\leq\; 1 \tag{12.30}$$

so this relation which we recognize from the equation without the mixed term is a necessary condition for stability. The mixed term is significant for $\varphi_1 = \varphi_2 = \pi/2$ where (12.29) reduces to

$$b_1\mu_1 + b_2\mu_2 + b_{12}\mu_{12} \;\leq\; 1 \tag{12.31}$$

This inequality follows from (12.30) and (12.22) (cf. exercise 1).

If we put $\varphi_1 = \pi - \varepsilon_1$ and $\varphi_2 = \pi - \varepsilon_2$ then $\sin\varphi_1 = \sin\varepsilon_1$ and

$$\sin^2\frac{\varphi_1}{2} \;=\; \cos^2\frac{\varepsilon_1}{2} \;=\; 1 - \sin^2\frac{\varepsilon_1}{2}$$

and similarly with index 2. We therefore have

$$2x_1+2x_2+x_{12} \;=\; 2b_1\mu_1+2b_2\mu_2-2b_1\mu_1\sin^2\frac{\varepsilon_1}{2}-2b_2\mu_2\sin^2\frac{\varepsilon_2}{2}+b_{12}\mu_{12}\sin\varepsilon_1\sin\varepsilon_2.$$

Since by (12.30) the sum of the first two terms on the right-hand-side is ≤ 1 we just need to show that the sum of the last three terms is non-positive, which is seen by the following

$$\begin{aligned}
0 \;&\leq\; (2\sqrt{b_1\mu_1}\sin\frac{\varepsilon_1}{2} - 2\sqrt{b_2\mu_2}\sin\frac{\varepsilon_2}{2})^2 \\
&=\; 4b_1\mu_1\sin^2\frac{\varepsilon_1}{2} + 4b_2\mu_2\sin^2\frac{\varepsilon_2}{2} - 8\sqrt{b_1\mu_1b_2\mu_2}\sin\frac{\varepsilon_1}{2}\sin\frac{\varepsilon_2}{2} \\
&\leq\; 4b_1\mu_1\sin^2\frac{\varepsilon_1}{2} + 4b_2\mu_2\sin^2\frac{\varepsilon_2}{2} - 2b_{12}\mu_{12}\sin\varepsilon_1\sin\varepsilon_2.
\end{aligned}$$

For the last inequality we have used (12.19) and

$$\sin\varepsilon_1 \;=\; 2\sin\frac{\varepsilon_1}{2}\cos\frac{\varepsilon_1}{2} \;\leq\; 2\sin\frac{\varepsilon_1}{2}.$$

We conclude that (12.30) is also a sufficient condition for stability of the explicit method applied to (12.12).

For the implicit scheme the growth factor is

$$g = \frac{1}{1+4x_1+4x_2+2x_{12}} \tag{12.32}$$

and because of (12.26) we always have $0 \le g \le 1$.
For Crank-Nicolson we have

$$g = \frac{1-2x_1-2x_2-x_{12}}{1+2x_1+2x_2+x_{12}} \tag{12.33}$$

and because of (12.26) we always have $-1 \le g \le 1$.

Altogether the mixed derivative term does not alter the basic stability properties of the explicit, Crank-Nicolson, or the implicit method. But in practice we do not wish to use any of these. We would rather prefer an ADI-method.

12.3 Stability of ADI-methods

We first look at the Simple Douglas-Rachford scheme (11.47)-(11.48) together with $\alpha = \beta$ as given by (12.3). The equations for the growth factor are

$$(1+4x_1)\tilde{g} = 1-x_{12}, \tag{12.34}$$

$$(1+4x_2)g = \tilde{g}-x_{12}, \tag{12.35}$$

$$g = \frac{1-x_{12}}{(1+4x_1)(1+4x_2)} - \frac{x_{12}}{1+4x_2} = \frac{1-2x_{12}-4x_1x_{12}}{(1+4x_1)(1+4x_2)}. \tag{12.36}$$

A necessary condition for stability is $g \le 1$ or

$$-x_{12}(1+2x_1) \le 2x_1+2x_2+8x_1x_2$$

or

$$-x_{12} \le 2x_2+2x_1\frac{1+2x_2}{1+2x_1}.$$

Comparing with (12.26) we suspect that we may be in trouble when $x_2 < x_1$ and μ is large. Actually the inequality is violated when $b_1 = b_2 = 1$, $b_{12} = 0.9$, $h_1 = h_2$, $\varphi_1 = \pi/2$, $\mu > 10$, and φ_2 is small and negative.

If we combine the Simple Douglas-Rachford scheme with (12.4) the equations for the growth factor are

$$\begin{aligned}(1+4x_1)\tilde{g} &= 1-x_{12},\\ (1+4x_2)g &= \tilde{g}(1-x_{12}),\\ g &= \frac{1-x_{12}}{1+4x_2}\cdot\frac{1-x_{12}}{1+4x_1}.\end{aligned} \tag{12.37}$$

We notice immediately that $g \geq 0$ and the condition $g \leq 1$ is equivalent to

$$2|x_{12}| + x_{12}^2 \;\leq\; 4x_1 + 4x_2 + 16x_1x_2$$

which follows from (12.26) and (12.27). We conclude that this combination is unconditionally stable.

If we combine the Simple Douglas-Rachford scheme with (12.5) the equations for the growth factor are

$$\begin{aligned}
(1+4x_1)\tilde{g} &= 1 - 2x_{12}, \\
(1+4x_2)g &= \tilde{g}, \\
g &= \frac{1-2x_{12}}{(1+4x_1)(1+4x_2)}. \qquad (12.38)
\end{aligned}$$

From (12.26) it follows readily that $-1 \leq g \leq 1$ and that we therefore have unconditional stability which makes this scheme very interesting indeed.

If we combine the Traditional Douglas-Rachford scheme (11.44) – (11.45) with (12.3) the equations for the growth factor are

$$\begin{aligned}
(1+4x_1)\tilde{g} &= 1 - 4x_2 - x_{12}, \\
(1+4x_2)g &= \tilde{g} + 4x_2 - x_{12}, \\
g &= \frac{1-4x_2-x_{12}}{(1+4x_1)(1+4x_2)} + \frac{4x_2 - x_{12}}{1+4x_2} \\
&= \frac{1-2x_{12}-4x_1x_{12}+16x_1x_2}{(1+4x_1)(1+4x_2)}. \qquad (12.39)
\end{aligned}$$

Comparing with Simple Douglas-Rachford it is apparent that we have even greater problems with the stability condition $g \leq 1$.

If we combine the Traditional Douglas-Rachford scheme with (12.4) the equations for the growth factor are

$$\begin{aligned}
(1+4x_1)\tilde{g} &= 1 - 4x_2 - x_{12}, \\
(1+4x_2)g &= \tilde{g}(1-x_{12}) + 4x_2, \\
g &= \frac{1-x_{12}}{1+4x_2} \cdot \frac{1-x_{12}-4x_2}{1+4x_1} + \frac{4x_2}{1+4x_2} \\
&= \frac{(1-x_{12})^2 + 4x_2x_{12} + 16x_1x_2}{(1+4x_1)(1+4x_2)}. \qquad (12.40)
\end{aligned}$$

If we supplement our earlier counterexample with $\varphi_2 = \pi/2$, and $\mu > 10$ then we have $g > 1$ violating the stability requirement.

If we combine the Traditional Douglas-Rachford scheme with (12.5) the equations for the growth factor are

$$\begin{aligned}(1+4x_1)\tilde{g} &= 1-4x_2-2x_{12},\\ (1+4x_2)g &= \tilde{g}+4x_2,\\ g &= \frac{1-4x_2-2x_{12}}{(1+4x_1)(1+4x_2)}+\frac{4x_2}{1+4x_2}\\ &= \frac{1+16x_1x_2-2x_{12}}{(1+4x_1)(1+4x_2)}. \qquad (12.41)\end{aligned}$$

From (12.26) it follows readily that $-1 \le g \le 1$ and once again we have a useful combination.

If we combine the Peaceman-Rachford scheme (11.20)-(11.21) with (12.3) the equations for the growth factor are

$$\begin{aligned}(1+2x_1)\tilde{g} &= 1-2x_2-x_{12},\\ (1+2x_2)g &= \tilde{g}(1-2x_1)-x_{12},\\ g &= \frac{(1-2x_1)(1-2x_2-x_{12})}{(1+2x_1)(1+2x_2)}-\frac{x_{12}}{1+2x_2}\\ &= \frac{(1-2x_1)(1-2x_2)-2x_{12}}{(1+2x_1)(1+2x_2)}. \qquad (12.42)\end{aligned}$$

$g \le 1$ follows directly from (12.26), and $g \ge -1$ is equivalent to

$$1+4x_1x_2-2x_{12} \;\ge\; -1-4x_1x_2$$

or

$$1+4x_1x_2-x_{12} \;\ge\; 0.$$

If φ_1 and φ_2 have different signs or if $b_{12}<0$ then $x_{12}<0$ and we are done. We can therefore assume $0<\varphi_1,\ \varphi_2<\pi$ and $b_{12}>0$.

$$\begin{aligned}0 &\le (1-2\sqrt{b_1b_2}\frac{k}{h_1h_2}\sin\frac{\varphi_1}{2}\sin\frac{\varphi_2}{2})^2\\ &= 1+4b_1\mu_1b_2\mu_2\sin^2\frac{\varphi_1}{2}\sin^2\frac{\varphi_1}{2}-4\sqrt{b_1b_2}\mu_{12}\sin\frac{\varphi_1}{2}\sin\frac{\varphi_2}{2}\\ &\le 1+4x_1x_2-4b_{12}\mu_{12}\sin\frac{\varphi_1}{2}\sin\frac{\varphi_2}{2}\cos\frac{\varphi_1}{2}\cos\frac{\varphi_2}{2}\\ &= 1+4x_1x_2-x_{12}\end{aligned}$$

thus proving stability.

If we combine the Peaceman-Rachford scheme with (12.4) the equations for the growth factor are

$$(1+2x_1)\tilde{g} \;=\; 1-2x_2-x_{12},$$

$$\begin{aligned}(1+2x_2)g &= \tilde{g}(1-2x_1-x_{12}),\\ g &= \frac{(1-2x_1-x_{12})(1-2x_2-x_{12})}{(1+2x_1)(1+2x_2)}.\end{aligned} \tag{12.43}$$

If $x_{12} > 2$ which can easily happen when μ is large, then it is clear that the numerator is greater than the denominator, implying $g > 1$ and thus instability. A specific example is $b_1 = b_2 = 1,\ b_{12} = 0.5$,
$k = h_1 = h_2 = 0.1 \Rightarrow \mu_1 = \mu_2 = \mu_{12} = 10,\ \varphi_1 = \varphi_2 = \pi/2$
leading to $x_1 = x_2 = x_{12} = 5$ and $g = (14/11)^2 > 1$.

If we combine the Peaceman-Rachford scheme with (12.5) the equations for the growth factor are

$$\begin{aligned}(1+2x_1)\tilde{g} &= 1-2x_2-2x_{12},\\ (1+2x_2)g &= \tilde{g}(1-2x_1),\\ g &= \frac{(1-2x_1)(1-2x_2-2x_{12})}{(1+2x_1)(1+2x_2)}.\end{aligned} \tag{12.44}$$

Taking the same example as above we get $g = (9\cdot 19)/11^2 > 1$ proving instability.

If we combine the Peaceman-Rachford scheme with (12.8) the equations for the growth factor are

$$\begin{aligned}(1+2x_1)\tilde{g} &= 1-2x_2-x_{12}(\frac{3}{2}-\frac{1}{2}g^{-1}),\\ (1+2x_2)g &= \tilde{g}(1-2x_1)-x_{12}(\frac{3}{2}-\frac{1}{2}g^{-1}),\\ (1+2x_1)(1+2x_2)g &= (1-2x_1)(1-2x_2-x_{12}(\frac{3}{2}-\frac{1}{2}g^{-1}))\\ &\qquad -x_{12}(\frac{3}{2}-\frac{1}{2}g^{-1})(1+2x_1)\\ &= (1-2x_1)(1-2x_2)-x_{12}(3-g^{-1}).\end{aligned}$$

We thus have a quadratic equation for g:

$$(1+2x_1)(1+2x_2)g^2-((1-2x_1)(1-2x_2)-3x_{12})g-x_{12} = 0. \tag{12.45}$$

With $b_1 = b_2 = 1,\ b_{12} = 0.9,\ \mu_1 = \mu_2 = \mu_{12} = 10,\ \varphi_1 = \varphi_2 = \pi/5$ one of the roots of this equation is -1.29 and this scheme is therefore only conditionally stable. When $b_{12} \leq 0.5$ it appears that the roots are always ≤ 1 (independent of μ) in absolute value, and we propose that this scheme can be used whenever the mixed term has a small weight.

If we combine the Peaceman-Rachford scheme with (12.9) the equations for the growth factor are

$$(1+2x_1)\tilde{g} = 1-2x_2-x_{12},$$

$$(1+2x_2)g \;=\; \tilde{g}(1-2x_1) - x_{12}(2-g^{-1}),$$

$$(1+2x_1)(1+2x_2)g^2 - ((1-2x_1)(1-2x_2) - x_{12}(3+2x_1))g - (1+2x_1)x_{12} \;=\; 0.$$

The product of the two roots is

$$\frac{x_{12}}{1+2x_2} \;=\; \frac{b_{12}\mu_{12}\sin\varphi_1\sin\varphi_2}{1+2b_2\mu_2\sin^2\frac{\varphi_2}{2}}.$$

With $b_2 = 1,\ b_{12} = 0.5,\ \mu_2 = \mu_{12} = 10,\ \varphi_1 = \pi/2,\ \varphi_2 = 0.5$ we get $|g_1 \cdot g_2| \approx 1.486 > 1$. If the product of the roots is greater than one then at least one of the roots must be greater than one in absolute magnitude thus implying instability.

12.4 Summary

Table 12.1: A comparison of methods with mixed derivative

	SDR	TDR	PR		PR
12.3	−	−	++	12.8	(+)
12.4	+	−	−	12.9	−
12.5	++	++	−		

An assessment of the various combinations of ADI-schemes with suggestions for treating the mixed derivative term is given in Table 12.1. A − indicates conditional stability, a + indicates unconditional stability, and a ++ indicates a recommended combination where the practicalities also can be solved. Formula (12.5) is recommended with either the simple or the traditional Douglas-Rachford method. Experiments indicate that SDR together with (12.4) is stable provided (11.29) is used, but not together with extrapolation. The Peaceman-Rachford method plays well together with (12.3) although the result will only be first order in t. For a second order method we recommend (12.8) when the mixed term is suitably small.

12.5 Exercises

1. Show that (12.30) and (12.22) imply (12.31).

2. Equation (12.12) with $b_1 = 1$, $b_2 = 1$, $b_{12} = 0.5$ has the solution (cf. Appendix B)

$$u(t, x, y) \;=\; e^{-t} \sin(2x - 2y) \cosh(x + y).$$

   Solve the equation for $0 < t, x, y < 1$ with SDR and one or more of (12.3) – (12.9). Initial and boundary conditions are taken from the true solution. Use $h_1 = h_2 = k = \frac{1}{10}, \frac{1}{20}, \frac{1}{40}$, and $\frac{1}{80}$ and compute the 2-norm of the error at $t = 1$.

3. Same as above with TDR.

4. Same as above with PR.

Chapter 13

Two-Factor Models – two examples

13.1 The Brennan-Schwartz model

A model for the determination of prices for bonds suggested by Brennan og Schwartz [3] can be formulated in the following way:

$$u_t = (\mu_r - \lambda\beta_r)u_r + (\frac{\beta_l^2}{l} + (l-r)l)u_l - ru + \frac{1}{2}\beta_r^2 u_{rr} + \rho\beta_r\beta_l u_{rl} + \frac{1}{2}\beta_l^2 u_{ll}$$

where
u is the price of the bond, r is the short interest, l is the long interest, and
$\mu_r = a_r + b_r(l-r)$, $\beta_r = r\sigma_r$, and $\beta_l = l\sigma_l$.
The coefficients have been estimated to
$a_r = -0.00622$, $b_r = 0.2676$, $\sigma_r = 0.10281$, $\sigma_l = 0.02001$, $\rho = -0.0022$, $\lambda = -0.9$.

We transform the r-interval $[0, \infty)$ to $(0, 1]$ using

$$x = \frac{1}{1+\pi_r r}, \qquad r = \frac{1-x}{\pi_r x},$$

and similarly for l:

$$y = \frac{1}{1+\pi_l l}, \qquad l = \frac{1-y}{\pi_l y},$$

where the transformation coefficients π_r and π_l are chosen properly, often between 10 and 13. An interest interval from 10% to 1% will with $\pi = 10$ be transformed into $[0.5, 0.91]$ and with $\pi = 13$ into $[0.43, 0.88]$.

We now have

$$\begin{aligned}
\frac{\partial u}{\partial r} &= \frac{\partial u}{\partial x}\frac{dx}{dr} = \frac{-\pi_r}{(1+\pi_r r)^2}\frac{\partial u}{\partial x} = -\pi_r x^2 u_x \\
\frac{\partial u}{\partial l} &= -\pi_l y^2 u_y \\
\frac{\partial^2 u}{\partial r^2} &= -\frac{\partial}{\partial x}(\pi_r x^2 u_x)\frac{dx}{dr} = 2\pi_r^2 x^3 u_x + \pi_r^2 x^4 u_{xx} \\
\frac{\partial^2 u}{\partial l^2} &= 2\pi_l^2 y^3 u_y + \pi_l^2 y^4 u_{yy} \\
\frac{\partial^2 u}{\partial r \partial l} &= -\frac{\partial}{\partial x}(\pi_l y^2 u_y)\frac{dx}{dr} = \pi_r \pi_l x^2 y^2 u_{xy}
\end{aligned}$$

and the differential equation becomes

$$u_t = b_1 u_{xx} + 2b_{12} u_{xy} + b_2 u_{yy} - a_1 u_x - a_2 u_y + \kappa u,$$

where

$$\begin{aligned}
b_1 = b_1(x) &= \frac{1}{2}\beta_r^2 \pi_r^2 x^4 = \frac{1}{2}\sigma_r^2 (1-x)^2 x^2 \\
b_{12} = b_{12}(x,y) &= \frac{1}{2}\rho\beta_r\beta_l\pi_r\pi_l x^2 y^2 = \frac{1}{2}\rho\sigma_r\sigma_l(1-x)(1-y)xy \\
b_2 = b_2(y) &= \frac{1}{2}\beta_l^2\pi_l^2 y^4 = \frac{1}{2}\sigma_l^2(1-y)^2 y^2 \\
a_1 = a_1(x,y) &= -\beta_r^2\pi_r^2 x^3 + (\mu_r - \lambda\beta_r)\pi_r x^2 \\
&= -x((1-x)(\sigma_r^2(1-x) + b_r + \lambda\sigma_r) - a_r\pi_r x - b_r\frac{\pi_r}{\pi_l}\frac{x}{y}(1-y)) \\
a_2 = a_2(x,y) &= (\sigma_l^2 + l - r)l\pi_l y^2 - \sigma_l^2(1-y)^2 y \\
&= y(1-y)(\sigma_l^2 y + \frac{1-y}{\pi_l y} - \frac{1-x}{\pi_r x}) \\
\kappa = \kappa(x) &= -r = -\frac{1-x}{\pi_r x}
\end{aligned}$$

We note in passing that the differential equation is well-posed since

$$b_1 b_2 - b_{12}^2 = \frac{1}{4}\sigma_r^2\sigma_l^2(1-x)^2(1-y)^2 x^2 y^2(1-\rho^2) > 0.$$

The initial condition at $t = 0$ is the value of the bond at expiry, i.e.

$$u(0, x, y) = 1$$

The boundary condition at $x = 0$ is found by multiplying the differential equation by x and then let $x \to 0$, which gives

$$0 = (1 - y)\frac{y}{\pi_r}u_y - \frac{1}{\pi_r}u$$

or

$$(1 - y)y\frac{du}{dy} = u$$

or

$$\frac{1}{u}du = \frac{1}{(1 - y)y}dy = (\frac{1}{1 - y} + \frac{1}{y})dy.$$

Integration gives

$$\ln u - \ln u_0 = \ln y - \ln y_0 - \ln(1 - y) + \ln(1 - y_0)$$

or

$$u = u_0\frac{y}{1 - y}\frac{1 - y_0}{y_0}.$$

We wish u to be bounded, also when $y \to 1$, and therefore we must have $u_0 = 0$, and thus

$$u(t, 0, y) = 0.$$

This is a so-called *natural* boundary condition, i.e. a condition which follows naturally from the equation (when we are interested in bounded solutions). It also fits well to our intuitive understanding that if $r \to \infty$ then the back value of the bond will not be particularly high.

The boundary condition at $y = 0$ is found in a similar way by multiplying with y and then letting $y \to 0$. We then find

$$0 = -b_r\frac{\pi_r}{\pi_l}x^2u_x$$

i.e. u must be constant, and since $u(t, 0, 0) = 0$ according to the first boundary condition we must have

$$u(t, x, 0) = 0;$$

but this is also in agreement with our intuition about the case $l \to \infty$.

The boundary condition at $y = 1$ is found by inserting $y = 1$ in the differential equation. We then have

$$b_{12}(x, 1) = b_2(1) = a_2(x, 1) = 0$$

and

$$u_t = \frac{1}{2}\sigma_r^2(1-x)^2x^2u_{xx} + ((1-x)(\sigma_r^2(1-x) + b_r + \lambda\sigma_r) - a_r\pi_r x)xu_x - \frac{1-x}{\pi_r x}u.$$

This is a parabolic equation in one space dimension which can be solved beforehand, or concurrently with the solution in the interior. This equation has the initial condition $u(0, x, 1) = 1$ from the general initial condition and the boundary conditions $u(t, 0, 1) = 0$ from the boundary condition for $x = 0$, and $u(t, 1, 1) = 1$ from the argument that when both the interests are 0, then the bond will retain its value.

In order to find a boundary condition at $x = 1$ we likewise put $x = 1$ in the differential equation and find

$$b_1(1) = b_{12}(1, y) = \kappa(1) = 0$$

and

$$u_t = \frac{1}{2}\sigma_l^2(1-y)^2y^2u_{yy} - (a_r\pi_r + b_r\frac{\pi_r}{\pi_l}\frac{1-y}{y})u_x - (1-y)(\sigma_l^2y^2 + \frac{1-y}{\pi_l})u_y$$

This is in principle a parabolic differential equation in t and y with an extra term involving u_x and therefore referring to u-values in the interior. This equation cannot be solved beforehand but must be solved concurrently with the solution in the interior.

The initial condition is as before $u(0, 1, y) = 1$, and the boundary conditions are $u(t, 1, 0) = 0$ and $u(t, 1, 1) = 1$.

13.2 Practicalities

We should like to implement an ADI method for the solution of this problem. One small detail in the practical considerations is that we need values for $\tilde{v}$ on two of the boundaries of the region. Because of the difficulties mentioned above getting boundary values at $x = 1$ it seems convenient to reverse the order of the operators P_1 and P_2 from the usual order in Chapter 11. Thus we wish to solve for $\tilde{v}$ in the y-direction and therefore we shall need information on $\tilde{v}$ at $y = 0$ and $y = 1$ where information is more readily available.

We select a step size, h_1, in the x-direction, or rather we select an integer, L, and set $h_1 = 1/L$. Similarly the step size in the y-direction is given through the integer, M, by $h_2 = 1/M$, and the time step by $k = 1/N$. Including boundary nodes we thus have $(L+1)(M+1)N$ function values to compute. With small step sizes we might not have storage space for all these numbers at the same time, but then again we don't need to. At any particular time step we only need information corresponding to two consecutive time levels (or three for Peaceman-Rachford) and we can therefore make do with two (or three) $(L+1)(M+1)$ arrays. If solution values are needed at intermediate times these can be recorded along the way. Such values are usually only required at coarser intervals, $\bar{h}_1 > h_1$, $\bar{h}_2 > h_2$, and $\bar{k} > k$, and therefore require smaller arrays.

Because of the discontinuity between the initial values and the boundary values at $x = 0$ and $y = 0$ it may be convenient to use or at least begin with the Douglas-Rachford method. Using the simple version (11.47 – 11.48) and (12.5) for the mixed derivative term the equations become

$$(I - kP_{2h})\tilde{v} = v^n + 2kb_{12}(x,y)\delta^2_{xy}v^n \tag{13.1}$$

$$(I - kP_{1h})v^{n+1} = \tilde{v} \tag{13.2}$$

The time step from nk to $(n+1)k$ is now divided into a number of subtasks numbered like in section 11.6 except that we have added one subtask at the beginning.

0. Advance the solution on $y = 1$ using

$$u_t = b_1(x)u_{xx} - a_1(x,1)u_x + \kappa(x)u \tag{13.3}$$

discretized using the implicit method

$$v^{n+1}_{l,M} - v^n_{l,M} = b_1\mu_1(v^{n+1}_{l+1,M} - 2v^{n+1}_{l,M} + v^{n+1}_{l-1,M}) - \frac{a_1}{2}\lambda_1(v^{n+1}_{l+1,M} - v^{n+1}_{l-1,M}) + \kappa k v^{n+1}_{l,M}$$

or

$$\begin{aligned} -(b_1\mu_1 + \frac{1}{2}a_1\lambda_1)v^{n+1}_{l-1,M} + (1 + 2b_1\mu_1 - \kappa k)v^{n+1}_{l,M} & \\ -(b_1\mu_1 - \frac{1}{2}a_1\lambda_1)v^{n+1}_{l+1,M} &= v^n_{l,M} \end{aligned} \tag{13.4}$$

where we have introduced $\lambda_1 = k/h_1$.

This tridiagonal system of equations supplemented with the boundary conditions $v^{n+1}_{0,M} = 0$ and $v^{n+1}_{L,M} = 1$ can now be solved using Gaussian elimination.

1. The right-hand-side of (13.1) requires v^n at all interior points which is no problem and $\delta^2_{xy}v^n$ at the same points which means v^n at all points including those on the boundary. The only problem arises at the first time step because of the discontinuity between the initial condition and the boundary conditions at $x = 0$ and $y = 0$. We recommend using the initial value throughout and thereby avoid divided differences of the order $1/h_1h_2$. In the present case the b_{12}-coefficient is rather small because of the small numerical value of ρ so the effect of a different choice is minimal.

2. We next compute $\tilde{v}$ for $y = 0$ and $y = 1$ using (13.2) and get $\tilde{v}_{l,0} = 0$ and

$$\tilde{v}_{l,M} \;=\; (I - kP_{1h})v^{n+1}_{l,M}. \tag{13.5}$$

Comparing (13.5) with (13.3) and (13.4) we note that there might be an advantage in including the κu-term in the P_1-operator because then (13.5) takes the simpler form of

$$\tilde{v}_{l,M} \;=\; v^{n}_{l,M}. \tag{13.6}$$

In the general case with $\theta\kappa u$ in P_1 and $(1-\theta)\kappa u$ in P_2 the formula for $\tilde{v}$ becomes

$$\tilde{v}_{l,M} \;=\; v^{n}_{l,M} + (1-\theta)\kappa k v^{n+1}_{l,M} \tag{13.7}$$

3. The system of equations (13.1) can now be solved for $\tilde{v}$ at all interior points. The system consists of $L - 1$ tridiagonal systems of $M - 1$ unknowns each and they can be solved independently of each other.

4. The right-hand-side of system (13.2) consists of $\tilde{v}$ at all interior points which we have just computed in **3.**

5. On each horizontal line (13.2) gives rise to one equation for each internal node, i.e. a total of $L - 1$ equations in $L + 1$ unknowns, the extra unknowns being the values of v^{n+1} at $x = 0$ and $x = 1$. The former is equal to 0, and for the latter we must resort to the boundary equation

$$u_t \;=\; b_2(y)u_{yy} - a_2(1, y)u_y - a_1(1, y)u_x \tag{13.8}$$

An implicit discretization of (13.8) could be

$$\begin{aligned} v^{n+1}_{L,m} - v^{n}_{L,m} \;=\; & b_2\mu_2(v^{n+1}_{L,m+1} - 2v^{n+1}_{L,m} + v^{n+1}_{L,m-1}) - \frac{1}{2}a_2\lambda_2(v^{n+1}_{L,m+1} - v^{n+1}_{L,m-1}) \\ & -\frac{1}{2}a_1\lambda_1(v^{n+1}_{L-2,m} - 4v^{n+1}_{L-1,m} + 3v^{n+1}_{L,m}) \end{aligned} \tag{13.9}$$

where we have used the asymmetric second order difference approximation for u_x on the boundary. A simpler formula would result from replacing the last parenthesis in (13.9) by $(-2v^{n+1}_{L-1,m} + 2v^{n+1}_{L,m})$ but since this is only a first order approximation of u_x the resulting v^{n+1} would be only first order correct in x.

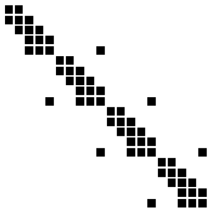


Figure 13.1: The coefficient matrix corresponding to $L = M = 5$.

Equation (13.9) supplies the extra information we need about $v_{L,m}^{n+1}$, but now the various rows are no longer independent of each other.

6. The total system which is outlined in Fig. 13.1 in the case $L = M = 5$ is tridiagonal with three exceptions all due to equation (13.9): In each block there is an element two places left of the diagonal in the last row (the coefficient of $v_{L-2,m}^{n+1}$). In each block but the first there is an element L places left of the diagonal in the last row (the coefficient of $v_{L,m-1}^{n+1}$). In each block but the last there is an element L places right of the diagonal in the last row (the coefficient of $v_{L,m+1}^{n+1}$).

On the right-hand-side of the system we must remember the effect of the boundary value at (1,1) in the last equation. The other boundary value at (1,0) is 0 so no correction is needed here.

Although the system of linear equations is not tridiagonal it still can be solved using Gaussian elimination without introducing new non-zero elements, and the solution process requires a number of simple arithmetic operations which is linear in the number of unknowns and only marginally larger than that of a tridiagonal system.

13.3 A Traditional Douglas-Rachford step

If one prefers to use the Traditional Douglas-Rachford method then the equations to be solved instead of (13.1) and (13.2) are

$$(I - kP_{2h})\tilde{v} = (I + kP_{1h})v^n + 2kb_{12}(x,y)\delta_{xy}^2 v^n \tag{13.10}$$

$$(I - kP_{1h})v^{n+1} = \tilde{v} - kP_{1h}v^n \tag{13.11}$$

Most of the considerations of the preceding section are still applicable so we shall just focus our attention on the differences which occur in **1.** and **4.**

1. The right-hand-side of (13.10) now also includes $P_{1h}v^n$ which means that it requires knowledge of v^n not only at all interior points but also for $x = 0$ and

$x = 1$. The only difficulty lies at the very first time step where we still prefer to settle the discontinuity issue by adopting the initial value throughout.

4. Similar considerations apply for the P_{1h}-term on the right-hand-side of (13.11).

13.4 The Peaceman-Rachford method

The Douglas-Rachford method is only first order in time and therefore we might prefer to use Peaceman-Rachford, possibly after an initial DR step.

The Peaceman-Rachford equations, augmented with (12.8) are

$$(I - \frac{1}{2}kP_{2h})\tilde{v} \quad = \quad (I + \frac{1}{2}kP_{1h})v^n + kb_{12}(x,y)\delta^2_{xy}\hat{v}^{n+\frac{1}{2}} \tag{13.12}$$

$$(I - \frac{1}{2}kP_{1h})v^{n+1} \quad = \quad (I + \frac{1}{2}kP_{2h})\tilde{v} + kb_{12}(x,y)\delta^2_{xy}\hat{v}^{n+\frac{1}{2}} \tag{13.13}$$

where

$$\hat{v}^{n+\frac{1}{2}} \quad = \quad v^n + \frac{1}{2}(v^n - v^{n-1}) \qquad\qquad n \geq 1. \tag{13.14}$$

Formula (13.14) can not be used in the first step but here it is OK to replace it by

$$\hat{v}^{n+\frac{1}{2}} \quad = \quad v^n. \tag{13.15}$$

Since $b_{12}(x,y)$ in this example is so small there is actually little difference between the results obtained with (13.14) and with (13.15).

Once again we divide the time step from nk to $(n+1)k$ into subtasks with the same numbering as before.

0. On the boundary $y = 1$ it is now appropriate to discretize (13.3) using Crank-Nicolson:

$$-(\frac{1}{2}b_1\mu_1 + \frac{1}{4}a_1\lambda_1)v^{n+1}_{l-1,M} + (1 + b_1\mu_1 - \frac{1}{2}\kappa k)v^{n+1}_{l,M} - (\frac{1}{2}b_1\mu_1 - \frac{1}{4}a_1\lambda_1)v^{n+1}_{l+1,M} \quad =$$
$$(\frac{1}{2}b_1\mu_1 + \frac{1}{4}a_1\lambda_1)v^{n}_{l-1,M} + (1 - b_1\mu_1 + \frac{1}{2}\kappa k)v^{n}_{l,M} + (\frac{1}{2}b_1\mu_1 - \frac{1}{4}a_1\lambda_1)v^{n}_{l+1,M}.$$

This is a system of the same structure as (13.4) although with a more complicated right-hand-side where previous considerations concerning the jump between the initial and boundary values at $(t,x) = (0,0)$ apply at the first step.

1. The right-hand-side of (13.12) is very similar to that of (13.10) and previous comments apply.

2. $\tilde{v}$ for $y = 0$ and $y = 1$ are now given by (11.27) which gives $\tilde{v}_{l,0} = 0$ and

$$\tilde{v}_{l,M} \quad = \quad \frac{1}{2}(I + \frac{1}{2}kP_{1h})v^n_{l,M} + \frac{1}{2}(I - \frac{1}{2}kP_{1h})v^{n+1}_{l,M}. \tag{13.16}$$

Again there is a computational advantage in including the κu-term in the P_1-operator in which case (13.16) reduces to

$$\tilde{v}_{l,M} \quad = \quad (I + \frac{1}{2}kP_{1h})v^n_{l,M}. \tag{13.17}$$

In the general case with $\theta\kappa u$ in P_1 and $(1-\theta)\kappa u$ in P_2 the formula for $\tilde{v}$ becomes

$$\tilde{v}_{l,M} \quad = \quad (I + \frac{1}{2}kP_{1h})v^n_{l,M} + \frac{1}{4}(1-\theta)\kappa k(v^{n+1}_{l,M} - v^n_{l,M}). \tag{13.18}$$

3. The system of equations (13.12) can now be solved for $\tilde{v}$ at all interior points. The system consists of $L - 1$ tridiagonal systems of $M - 1$ unknowns each and they can be solved independently of each other.

4. The right-hand-side of system (13.13) requires knowledge of $\tilde{v}$ at all interior points in addition to the boundary values from **2.** The values needed for $\hat{v}$ are the same as in **1.**

5. Equation (13.13) now gives rise to a set of tridiagonal equations which must be supplemented by the Crank-Nicolson equivalent of (13.9) to form a system of equations with the same pattern of nonzeroes as before.

6. This system is now solved for v^{n+1} at all interior grid points as well as at all interior points on the boundary line $x = 1$.

13.5 Fine points on efficiency

Efficiency often amounts to a trade-off between storage space and computation time. Readability of the program can also tip the scale in favour of one particular strategy. Since the coefficients do not depend on time many things can be computed once and reused in each time step. Also the Gaussian elimination can be performed once and the components of the LU factors stored for later use.

The coefficient functions $(b_1, \ldots, \kappa)$ may be supplied as subroutines or they may be computed ahead of time at all grid points and stored in arrays. $b_{12}(x, y)$, $a_1(x, y)$, and $a_2(x, y)$ require two-dimensional $(L + 1) \cdot (M + 1)$ arrays, $b_1(x)$, $\kappa(x)$, and $b_2(y)$ need one-dimensional vectors with $L + 1$, resp. $M + 1$ elements.

13.6 Convertible bonds

In a model by Longstaff and Schwartz [24] the two independent variables are the interest, r, and the volatility, V. The differential equation can be written as

$$\begin{aligned} u_t &= \frac{1}{2}Vu_{rr} + ((\alpha+\beta)V - \alpha\beta r)u_{rV} + \frac{1}{2}((\alpha^2+\alpha\beta+\beta^2)V - \alpha\beta(\alpha+\beta)r)u_{VV} \\ &\quad +(\alpha\gamma + \beta\eta + \frac{\xi\alpha - \delta\beta}{\beta-\alpha}r + \frac{\delta-\xi}{\beta-\alpha}V)u_r \\ &\quad +(\alpha^2\gamma + \beta^2\eta + \frac{\alpha\beta(\xi-\delta)}{\beta-\alpha}r + \frac{\alpha\delta-\beta\xi}{\beta-\alpha}V)u_V - ru, \end{aligned}$$

where the parameters have been estimated to
$\alpha = 0.001149$, $\beta = 0.1325$, $\gamma = 3.0493$, $\delta = 0.05658$, $\eta = 0.1582$, $\xi = 3.998$.

The conditions for this problem to be well-posed are

$$V > 0$$

and

$$V((\alpha^2+\alpha\beta+\beta^2)V - \alpha\beta(\alpha+\beta)r) - ((\alpha+\beta)V - \alpha\beta r)^2 > 0.$$

The last condition can be rewritten to

$$\alpha\beta V^2 + \alpha^2\beta^2 r^2 - \alpha\beta(\alpha+\beta)rV < 0$$

or

$$V^2 - (\alpha+\beta)rV + \alpha\beta r^2 < 0$$

or

$$\alpha r < V < \beta r$$

since $\alpha < \beta$ in our case.

The equation is therefore only well-posed in part of the region $(r > 0, V > 0)$ and ill-posed in the rest. If an equation is ill-posed, then the norm of the solution at any particular time is not guaranteed to be bounded in terms of the initial condition (cf. section 2.3). Or, small changes in the initial condition may produce large changes in the solution at a later time. In principle the solution can become very large in a very short time although in practice we may be able to retain a limited accuracy for small time intervals. And it doesn't help much that the problem is well-posed in part of the region. The disturbances which originate in the ill-posed part will quickly spread to the rest (cf. section 14.3).

The main advice for ill-posed problems is not to touch them. It is far better to search for another model with reasonable mathematical properties. If an ill-posed problem must be solved then approach it very carefully. And be prepared that our numerical methods may deceive us when they are used outside their usual area of application. A more detailed analysis is given in the next chapter.

Chapter 14

Ill-Posed Problems

14.1 Theory

For the simple parabolic problem

$$u_t \quad = \quad bu_{xx} \tag{14.1}$$

it is essential that b is positive.

From Fourier analysis (cf. Chapter 2) we know that

$$\hat{u}_t(t,\omega) \quad = \quad -b\omega^2\hat{u}(t,\omega) \tag{14.2}$$

and therefore that the Fourier transform of u can be written

$$\hat{u}(t,\omega) \quad = \quad e^{-b\omega^2 t}\hat{u}(0,\omega) \tag{14.3}$$

and by Parseval's theorem we have that

$$\int |u(t,x)|^2\, dx \;=\; \int |\hat{u}(t,\omega)|^2\, d\omega \;=\; \int |e^{-2b\omega^2 t}|\, |\hat{u}(0,\omega)|^2 d\omega. \tag{14.4}$$

When $b > 0$ we know that the exponential factor is ≤ 1 for $t > 0$ and therefore that the norm of the solution at any time t is bounded by the norm of the initial value function. It also follows that small changes in the initial value will produce small changes in the solution at later times.

If $b < 0$, or if we try to solve the heat equation backwards in time, the situation is quite different. Since $-2b\omega^2 t > 0$ we shall now observe a magnification of the various components of the solution, and the higher the frequency, ω, the higher the magnification.

If the initial condition is smooth, consisting only of low frequency components then the effect of the magnification is limited for small values of t. But if the initial function contains high frequency components, or equivalently that $\hat{u}(0, \omega)$ is different from 0 for large values of ω, then the corresponding components of the solution will exhibit a large magnification. The solution will be extremely sensitive to small variations in the initial value if these variations have high frequency. In mathematical terms the solution will not depend continuously on the initial data.

The main advice is: Stay away from such problems.
Even if the initial function is smooth, the unavoidable rounding errors connected with numerical computations will introduce high frequency perturbations and although small at the beginning they will by nature of the equation be magnified.

14.2 Practice

But it is difficult to restrain our curiosity. How will our difference schemes react if we try to solve such a problem numerically with a finite difference method.

The components of the numerical solution are governed by the growth factor which for the general θ-method is (cf. section 2.4)

$$g(\varphi) \quad = \quad \frac{1 - 4(1-\theta)b\mu \sin^2 \frac{\varphi}{2}}{1 + 4\theta b\mu \sin^2 \frac{\varphi}{2}}. \tag{14.5}$$

For a given step size, $h = (X_2 - X_1)/M$, not all frequencies, φ, will occur. Because of the finite and discrete nature of the problem, only a finite number of frequencies are possible, given by (6.23):

$$\varphi_p \quad = \quad \frac{p\pi}{M}, \qquad p = 1, 2, \ldots M-1. \tag{14.6}$$

For the explicit method we have

$$g(\varphi) \quad = \quad 1 - 4b\mu \sin^2 \frac{\varphi}{2}. \tag{14.7}$$

When $b\mu < 0$ we notice immediately that $g(\varphi) > 1$ for all φ. All components will be magnified, and the high frequency components will be magnified most. This is fine for it reflects the behaviour of the true solution, at least qualitatively.

The largest magnification at time $t = nk$ is

$$(1 + 4|b\mu|)^n \;=\; (1 + 4\frac{|bk|}{h^2})^n \;\approx\; e^{4\frac{|bnk|}{h^2}} \;=\; e^{4\frac{|bt|}{h^2}},$$

so with a given step size h there is a limit to the magnification at a given time, t, independent of the time step k. This is also in accordance with the mathematical properties of the solution since the value of h defines an upper limit on the possible frequencies.

For the Crank-Nicolson method the growth factor is

$$g(\varphi) \quad = \quad \frac{1 - 2b\mu \sin^2 \frac{\varphi}{2}}{1 + 2b\mu \sin^2 \frac{\varphi}{2}}. \tag{14.8}$$

We may experience infinite magnification at the finite frequency φ given by

$$\sin^2 \frac{\varphi}{2} \quad = \quad -\frac{1}{2b\mu},$$

a situation which is possible if $b\mu < -1/2$. We may not observe infinite magnification in practice if the corresponding frequency is not among those given by (14.6).

The largest magnification is given by the value of φ_p which maximizes (14.8). On the other hand it is easily seen from (14.8) that all components are magnified, just as for the explicit method, but the magnification becomes rather small for high frequency components when $|b\mu|$ is large.

For the implicit method the growth factor is

$$g(\varphi) \quad = \quad \frac{1}{1 + 4b\mu \sin^2 \frac{\varphi}{2}}. \tag{14.9}$$

Again we may experience infinite magnification when $b\mu < -1/4$ but we may not observe it in practice because of the discrete set of applicable frequencies. If $|b\mu|$ is large we observe from (14.9) that high frequency components of the numerical solution will be damped. This may result in a pleasantly looking solution, but it is a deception. Since all components of the true solution are magnified, a damping of some is really an unwanted effect.

A further complication associated with negative values of $b\mu$ is that we may encounter zeroes in the diagonal in the course of the Gaussian elimination, even in cases where the tridiagonal matrix is non-singular. It may therefore be necessary to introduce pivoting.

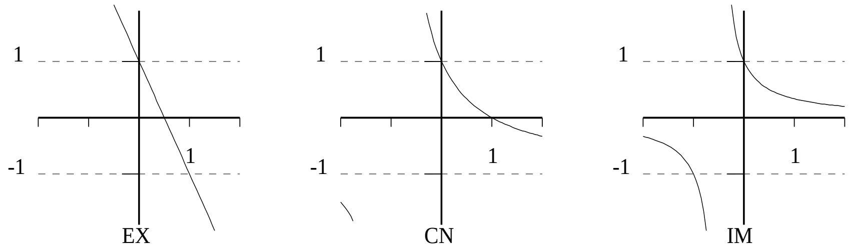


Figure 14.1: The growth factors for EX, CN, and IM as functions of $x = 2b\mu \sin^2 \frac{\varphi}{2}$.

The behaviour of the explicit method (EX), Crank-Nicolson (CN), and the implicit method (IM) is visualized in Fig. 14.1 using the graphs of $1 - 2x$ for EX, $(1-x)/(1+x)$ for CN, and $1/(1+2x)$ for IM, where $x = 2b\mu \sin^2 \frac{\varphi}{2}$. We have stability (damping) when the respective functions lie in the strip between -1 and 1, so for positive $b\mu$, (positive x), we have unconditional stability with CN and IM, and we require $2b\mu \leq 1$, $(x \leq 1)$ with EX. For negative $b\mu$ we always have instability with EX and CN, but we note that the magnification is rather small for large $|b\mu|$ with CN. For IM we have stability for large $|b\mu|$ (and φ). These observations should be compared with the fact that the true solution exhibits large growth for large values of ω.

When $b\mu < -1/2$ the numerical results from CN and IM will not be influenced most by the high frequency components. Larger growth factors will appear due to the singularity of $g(\varphi)$ for intermediate values of $\varphi_p = p\pi/M$. The dominant factor will occur for the value of p which makes $2b\mu \sin^2(\varphi_p/2)$ closest to -1, respectively $-1/2$, and the behaviour will be somewhat erratic when we vary the step sizes.

Table 14.1: Error growth for the negative heat equation.

	k	$b\mu$	n	t	g	p
EX	0.1	-10	4	0.4	40	19
	0.01	-1	11	0.11	5	19
	0.005	-0.5	16	0.08	3	19
	0.0025	-0.25	27	0.0675	2	19
	0.00125	-0.125	47	0.05875	1.5	19
CN	0.1	-10	7	0.7	-23	3
	0.01	-1	2	0.02	∞	10
	0.005	-0.5	4	0.02	300	19
	0.0025	-0.25	19	0.0475	3	19
	0.00125	-0.125	39	0.04875	1.7	19
IM	0.1	-10	10	1.0	45	2
	0.01	-1	3	0.03	-10	7
	0.005	-0.5	2	0.01	∞	10
	0.0025	-0.25	5	0.0125	150	19
	0.00125	-0.125	31	0.03875	2	19

Example. The equation $u_t = -u_{xx}$ has the solution $u(t,x) = e^t \cos x$ and this is used to define initial and boundary values. We have solved the equation numerically using EX, CN, and IM on $x \in [-1,1]$ with $M = 20$ corresponding to $h = 1/10$ and a range of time steps from $k = 1/10$ to $k = 1/800$ giving values of $b\mu$ from -10 to $-1/8$. We have continued the numerical solution until the error exceeded 5 and have recorded the number of time steps, the final value of t, and the observed growth which in most cases was in good agreement with the theoretical value from (14.5) and (14.6).

The results are given in Table 14.1. For small values of $b\mu$ the growth factor approaches 1 and the worst growth is always associated with the highest frequency component. As the reduction in g is coupled with a reduction in the time step we notice that the time interval of integration becomes smaller as the time step is reduced. As $b\mu$ gets larger (in absolute value) we may observe significant growth with CN and IM for low frequency components because of the singularity in the expression for g. □

14.3 Variable coefficients – An example

What happens when an equation is ill-posed in part of its domain. Can we trust the solution in the rest of the domain where the equation is well posed. This question can be illustrated by the following

Example. Consider the equation

$$u_t = \frac{1}{2} x u_{xx}. \tag{14.10}$$

The coefficient of u_{xx} is negative when $x < 0$ and the equation is therefore ill-posed here. A solution to (14.10) is

$$u(t,x) = tx + x^2$$

and this is used to define initial and boundary values. If we solve (14.10) on an interval such that 0 becomes a grid point then the system of equations will decouple in two, because of zeroes in the side diagonal, and bad vibrations from the negative part will have no influence on the positive side. To avoid this decoupling we therefore choose to solve in the interval $x \in [-0.983, 1.017]$. We have solved the equation numerically using CN and IM with $M = 20$ corresponding to $h = 1/10$, and a range of time steps from $k = 1/10$ to $k = 1/200$. We have continued the numerical solution until the error exceeded 5 and have recorded the number of time steps and the final value of t and give the results in Table 14.2.

In all cases we observed severe error growth originating from the negative part of the interval and eventually spreading to the whole interval. □

Table 14.2: Range of integration until error exceeds 5.

	k	n	t
CN	1/10	7	0.7
	1/40	14	0.35
	1/100	25	0.25
	1/200	56	0.28
IM	1/10	8	0.8
	1/40	22	0.55
	1/100	10	0.1
	1/200	34	0.17

Chapter 15

A Free Boundary Problem

15.1 The Stefan problem

Free boundary problems arise in the mathematical modelling of systems involving heat conduction together with a phase change such as the freezing of a liquid or the melting of a solid. The original paper by J. Stefan [33] was a study of the thickness of ice in arctic waters, and since then these problems have often been called Stefan problems.

To illustrate the one-dimensional one-phase Stefan problem consider the following system: A horisontal rod of ice, enclosed in an insulated tube, is kept initially at the freezing point, 0° C. We now supply heat to one end of the rod. The problem is to determine the position of the ice-water interface as a function of time and to find the temperature distribution in the water as a function of time and distance from the heat source. Mathematically this can be formulated as

$$\begin{aligned}
u_t - u_{xx} &= 0, && t > 0,\ 0 < x < y(t), && (15.1)\\
u_x(t,0) &= -1, && t > 0, && (15.2)\\
u(t,y(t)) &= 0, && t \geq 0, && (15.3)\\
u_x(t,y(t)) &= -y'(t), && t > 0, && (15.4)\\
y(0) &= 0. && && (15.5)
\end{aligned}$$

$y(t)$ denotes the position of the ice-water interface at time t, and $u(t,x)$ is the temperature of the water at time t and distance x. At time $t = 0$ there is no water, (15.5), the supply of heat is constant in time, (15.2), the temperature of the water at the interface is 0° C (15.3), and the rate of melting of ice is proportional to the heat flux at the interface (15.4). The temperature of the water is governed by the simple heat equation (15.1), and by suitable linear transformations all physical constants are set equal to 1.

If all the heat supplied was used to melt ice, the interface would be at $y(t) = t$. But some of it is used to heat the water, so at time t we have

$$y(t) \quad = \quad t - \int_0^{y(t)} u(t,x)dx, \qquad\qquad t > 0 \tag{15.6}$$

a relation which is equivalent to (15.4), given (15.1) – (15.3) and often used instead of (15.4) in the numerical calculations.

The first question one should be concerned with is that of existence and uniqueness of a solution to (15.1) – (15.5). We shall not address this question here but be content with the fact that our numerical schemes seem to converge as the step sizes become smaller, and this might be used as a basis of an existence proof.

It is intuitively clear that as time passes more and more ice will melt, and the temperature of the water at any particular point will increase, and also that the temperature of the water at any particular time will decrease with x. We shall therefore expect to find that $y'(t) > 0$ for $t > 0$ and that $u_t(t,x) > 0$ and $u_x(t,x) < 0$ for $t > 0$ and $0 < x < y(t)$.

We shall first note some immediate consequences of equations (15.1) – (15.6). From (15.2), (15.4), and the continuity of $y(t)$ and $y'(t)$ at $t = 0$ we get

$$y'(0) \quad = \quad 1. \tag{15.7}$$

From (15.3) and (15.2) and the maximum principle for (15.1) we deduce that

$$u(t,x) \quad > \quad 0, \qquad\qquad t > 0,\ 0 < x < y(t), \tag{15.8}$$

and

$$u_x(t,y(t)) \quad < \quad 0, \qquad\qquad t > 0 \tag{15.9}$$

and from (15.4) we then have

$$y'(t) \quad \geq \quad 0, \qquad\qquad t > 0 \tag{15.10}$$

and from (15.6)

$$y(t) \quad < \quad t, \qquad\qquad t > 0. \tag{15.11}$$

15.2 The Douglas-Gallie method

Among the various numerical methods proposed for the Stefan problem we have chosen the difference scheme of Douglas and Gallie [9]. They choose a fixed step

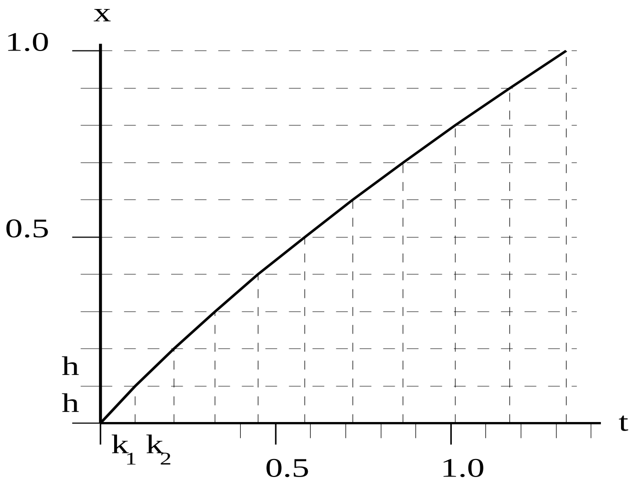


Figure 15.1: The Douglas-Gallie grid

size, h, in the x-direction and a variable step size in the t-direction with steps k_1, k_2, ... determined such that the computed boundary curve $y(t)$ passes through grid points and such that there is precisely one extra grid point when going from time t_{n-1} to t_n. In Fig. 15.1 we have shown how the grid might look.

We shall use the following notation

$$x_m = mh; \quad t_n = \sum_{i=1}^{n} k_i; \quad v_m^n = v(t_n, x_m); \quad m = 0, 1, \ldots, n; \quad n = 0, 1, \ldots$$

$v(t_n, x_m)$ is the numerical approximation to the temperature distribution $u(t, x)$ and x_n is the numerical approximation to $y(t_n)$.

Since we have one extra grid point at time t_n compared to the previous time t_{n-1}, the implicit formula looks like an ideal choice. Douglas and Gallie propose the following equations at time n:

$$\begin{array}{rcll} \dfrac{v_m^n - v_m^{n-1}}{k_n} & = & \dfrac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}, \quad m = 1, 2, \ldots, n-1, & (15.12) \\ v_0^n - v_1^n & = & h, & (15.13) \\ v_n^n & = & 0, & (15.14) \\ k_n & = & h \displaystyle\sum_{m=1}^{n-1} v_m^n + nh - t_{n-1}. & (15.15) \end{array}$$

These equations are straightforward discretizations of (15.1), (15.2), (15.3), and (15.6). They comprise $n+2$ equations in the $n+2$ unknowns, v_m^n, $(m = 0, 1, \ldots, n)$ plus k_n. However, the equations are non-linear so the solution process is not completely straightforward.

Remark. A simpler alternative to (15.15) is to discretize (15.4) to

$$k_n \;=\; h^2/v_{n-1}^n \tag{15.16}$$

which then can be used to correct the time step k_n. □

We shall first see how to get the process started. For $n = 0$ we get from (15.14) that

$$v_0^0 \;=\; 0 \tag{15.17}$$

and this is the only value for $n = 0$. For $n = 1$ we have two values. From (15.14), (15.13), and (15.15) (or (15.16)) we get

$$v_1^1 \;=\; 0, \quad v_0^1 \;=\; h, \quad k_1 \;=\; h. \tag{15.18}$$

For $n = 2$ (15.14), (15.13), and (15.12) reduce to

$$\frac{v_1^2}{k_2} \;=\; \frac{h - v_1^2}{h^2}$$

and (15.15) gives together with (15.18)

$$k_2 \;=\; hv_1^2 + h.$$

Combined we have a quadratic equation for v_1^2 where the positive root is

$$v_1^2 \;=\; -\frac{1}{2} + \sqrt{\frac{1}{4} + h} \qquad \Rightarrow \qquad k_2 \;=\; h(\frac{1}{2} + \sqrt{\frac{1}{4} + h}). \tag{15.19}$$

Remark. The same solution is obtained if we use (15.16) instead of (15.15). □

Remark. The superscript '2' on v indicates the time step, The same superscript on h indicates the second power. □

For $n \geq 3$ we solve the equations (15.12) – (15.15) iteratively, guessing a starting value $k_n^{(0)}$, solving (15.12) – (15.14) for $v_m^{n(0)}$, $m = 0, 1, \ldots, n$ and then using (15.15) (or possibly (15.16)) to produce an improved value $k_n^{(1)}$. The process is then repeated until $k_n^{(r)} - k_n^{(r-1)}$ is smaller than some predetermined tolerance, ε.

Several questions can be raised at this point:
Is there a (useful) solution to (15.12) – (15.15)?
Does the iteration converge to this solution?
Does it converge fast enough to be useful?
How do we get good starting values for k_n?
And more specificly:
Should we include v_0^n in the sum in (15.15)?

Or maybe with weight 0.5 (the trapezoidal rule)?
Should we use second order approximations instead of (15.13) (and (15.16))?
Possibly symmetric ones with fictitious points?
Is it possible to use Crank-Nicolson?

We observe convergence in practice and this assures us that the equations have a solution. The convergence (in r) appears to be linear and can therefore be accelerated using Aitken's device [1]. A good starting value for k_n is the previous time step, k_{n-1}, and an even better value is obtained by extrapolation from the previous two time steps: $k_n^{(0)} = 2k_{n-1} - k_{n-2}$. The iterations using (15.16) appears to have slower convergence than those using (15.15) but after one or two Aitken extrapolations the difference is minimal. The limit value appears to be the same.

We have only one independent stepsize (h), since the time steps are determined from h. The implicit method is first order in k and we shall therefore expect the overall method to be first order in h. Therefore there seems to be no immediate demand for second order approximations to the derivative boundary conditions (15.13) and (15.16), or the integral (15.15).

The boundary curve $y(t)$ is an increasing function of t and therefore has an inverse function which we shall call $t(x)$. The computed value of $t(x)$ for $x = n \cdot h$ is the sum of the first n time steps. It is therefore straightforward to experimentally determine the order of the method w.r.t. the determination of $t(x)$.

Computer experiments confirm our assumptions. They also seem to indicate that use of the trapezoidal rule instead of (15.15) (i.e. adding $v_0^n/2$ to the sum) give more, and use of (15.16) less, accurate results, although still first order. After two Richardson extrapolations the results differ by less then 0.0000005. The results in Section 15.4 were computed using (15.16).

15.3 The global error

Along the lines of Chapter 9 we shall assume that the computed solution $v(t,x)$ can be expressed in a power series in h:

$$v(t,x) \quad = \quad u(t,x) - hc - h^2 d - h^3 f - \cdots \qquad (15.20)$$

where $u(t,x)$ is the true solution and c, d, and f are auxiliary functions of t and x. Likewise we shall assume that the computed boundary function $Y(t)$ can be expressed as

$$Y(t) \quad = \quad y(t) - h\gamma - h^2\delta - h^3\varphi - \cdots \qquad (15.21)$$

where $y(t)$ is the true boundary function and γ, δ, and φ are auxiliary functions of t.

The functions, $v(t,x)$ and $Y(t)$, are actually known only at grid points $t = t_n$, $x = x_m$, but we shall assume (as we have done previously) that they can be extended in a differentiable manner to all $t > 0$ and $x \in (0, Y(t))$

The computed step sizes, k_n, can be viewed as instances of a step size function $k(t)$ which also can be written as a power series

$$k(t) \;=\; h\kappa + h^2\lambda + h^3\mu + \cdots \tag{15.22}$$

where κ, λ, and μ are auxiliary functions of t. These functions are closely related to the functions y, γ, δ, and φ above since

$$\begin{aligned} h &= Y_n - Y_{n-1} \;=\; Y(t_n) - Y(t_n - k_n) \\ &= k_n Y'(t_n) - \frac{1}{2}k_n^2 Y'' + \frac{1}{6}k_n^3 Y''' - \cdots \\ &= (h\kappa + h^2\lambda)(y' - h\gamma' - h^2\delta') - \frac{1}{2}h^2\kappa^2 y'' + \cdots \\ &= h\kappa y' + h^2(\lambda y' - \kappa\gamma' - \frac{1}{2}\kappa^2 y'') + \cdots \end{aligned}$$

using (15.21) and (15.22). Equating terms with equal powers of h we get

$$\begin{aligned} 1 &= \kappa y' \\ 0 &= \lambda y' - \kappa\gamma' - \frac{1}{2}\kappa^2 y'' \end{aligned}$$

When we know $y(t)$ and have established that $y' > 0$ we have

$$\kappa(t) \;=\; \frac{1}{y'(t)} \tag{15.23}$$

and when we also know $\gamma(t)$ then we find

$$\begin{aligned} \lambda(t) &= (\kappa\gamma' + \frac{1}{2}\kappa^2 y'')/y' \\ &= \frac{\gamma'}{(y')^2} + \frac{1}{2}\frac{y''}{(y')^3} \end{aligned} \tag{15.24}$$

From the left-hand-side of (15.12) we get using (15.20) and (15.22)

$$\begin{aligned} \frac{v_m^n - v_m^{n-1}}{k_n} &= u_t - hc_t - h^2 d_t - \frac{1}{2}k_n u_{tt} + \frac{1}{2}k_n h c_{tt} + \frac{1}{6}k_n^2 u_{ttt} + \cdots \\ &= u_t - hc_t - h^2 d_t - \frac{1}{2}(h\kappa + h^2\lambda)(u_{tt} - hc_{tt}) + \frac{1}{6}h^2\kappa^2 u_{ttt} + \cdots \\ &= u_t - h(c_t + \frac{1}{2}\kappa u_{tt}) - h^2(d_t + \frac{1}{2}\lambda u_{tt} - \frac{1}{2}\kappa c_{tt} - \frac{1}{6}\kappa^2 u_{ttt}) + \cdots \end{aligned}$$

From the right-hand-side we get

$$u_{xx} - hc_{xx} - h^2 d_{xx} + \frac{1}{12} h^2 u_{4x} + \cdots$$

Equating terms with equal powers of h we get the differential equations to be satisfied by the auxiliary functions:

$$\begin{aligned} u_t - u_{xx} &= 0, && (15.25)\\ c_t - c_{xx} &= -\frac{1}{2}\kappa u_{tt}, && (15.26)\\ d_t - d_{xx} &= -\frac{1}{2}\lambda u_{tt} + \frac{1}{2}\kappa c_{tt} + \frac{1}{6}\kappa^2 u_{ttt} - \frac{1}{12} u_{4x}. && (15.27) \end{aligned}$$

From (15.13) we get

$$h \;=\; v_0^n - v_1^n \;=\; -hu_x - \frac{1}{2}h^2 u_{xx} - \frac{1}{6}h^3 u_{xxx} + h^2 c_{xx} + \frac{1}{2}h^3 c_{xx} + h^3 d_x + \cdots$$

and equating terms we get the following boundary conditions at $x = 0$ for the auxiliary functions

$$\begin{aligned} u_x(t,0) &= -1, && (15.28)\\ c_x(t,0) &= \frac{1}{2} u_{xx}(t,0), && (15.29)\\ d_x(t,0) &= \frac{1}{6} u_{xxx} - \frac{1}{2} c_{xx}. && (15.30) \end{aligned}$$

The boundary condition (15.14) involves $v(t_n, nh)$ whereas we have information about u at $(t_n, y(t_n))$. The difference in the x-coordinate is

$$y(t_n) - nh \;=\; y(t_n) - Y(t_n) \;=\; h\gamma + h^2\delta + h^3\varphi + \cdots$$

and we therefore have

$$\begin{aligned} 0 \;=\; v_n^n \;&=\; v(t_n, y(t_n)) + (nh - y(t_n))v_x + \frac{1}{2}(nh - y(t_n))^2 v_{xx} + \cdots\\ &=\; u(t_n, y(t_n)) - hc - h^2 d - (h\gamma + h^2\delta)(u_x - hc_x) + \frac{1}{2}h^2\gamma^2 u_{xx} + \cdots \end{aligned}$$

leading to

$$\begin{aligned} u(t_n, y(t_n)) &= 0, && (15.31)\\ c(t_n, y(t_n)) &= -\gamma u_x \;=\; \gamma(t) y'(t), && (15.32)\\ d(t_n, y(t_n)) &= -\delta u_x + \gamma c_x + \frac{1}{2}\gamma^2 u_{xx}. && (15.33) \end{aligned}$$

As the second condition at $y(t)$ we take (15.16) which we rewrite to

$$\begin{aligned}
-h &= -k_n \frac{v_{n-1}^n - v_n^n}{h} = k_n[v_x(t_n, nh) - \frac{1}{2}hv_{xx} + \frac{1}{6}h^2 v_{xxx} + \cdots] \\
&= (h\kappa + h^2\lambda)[u_x(t_n, y(t_n)) - hc_x - h^2 d_x - (h\gamma + h^2\delta)(u_{xx} - hc_{xx}) \\
&\qquad + \frac{1}{2}h^2\gamma^2 u_{xxx} - \frac{1}{2}h(u_{xx} - hc_{xx} - h\gamma u_{xxx}) + \frac{1}{6}h^2 u_{xxx}] + \cdots \\
&= h\kappa u_x + h^2(\lambda u_x - \kappa(c_x + (\gamma + \frac{1}{2})u_{xx})) + \cdots
\end{aligned}$$

Equating powers of h we get

$$\kappa(t)u_x(t, y(t)) = -1, \tag{15.34}$$

$$\lambda(t)u_x(t, y(t)) = \kappa(c_x + \gamma u_{xx} + \frac{1}{2}u_{xx}). \tag{15.35}$$

At $t = 0$ we have $y(0) = 0$ and $u(0, 0) = 0$ and since we begin with $Y(0) = 0$ and $v(0, 0) = 0$ we have the initial values

$$\gamma(0) = \delta(0) = c(0, 0) = d(0, 0) = 0. \tag{15.36}$$

From (15.25), (15.28), (15.31), (15.23), and (15.34) we recover the original Stefan problem indicating that our difference scheme is consistent and that our basic assumptions (15.20), (15.21), and (15.22) are not completely unrealistic. We now assume $u(t, x)$ and $y(t)$ known such that the upcoming problems are with a fixed boundary. In equation (15.35) we have κ, λ, and γ appearing and they must first be eliminated using (15.23), (15.24), and (15.32). Differentiating (15.32) we get

$$c_t(t, y(t)) + c_x(t, y(t))y'(t) = \gamma' y' + \gamma y'' \tag{15.37}$$

and using this in (15.35) we end up with

$$2y'(t)c_x(t, y(t)) + c_t(t, y(t)) + (y'(t)^2 - \frac{y''(t)}{y'(t)})c(t, y(t)) = -\frac{1}{2}y''(t) - \frac{1}{2}y'(t)^2. \tag{15.38}$$

(15.38), together with (15.26), (15.29), and (15.36), defines a boundary value problem for the auxiliary function $c(t, x)$. The boundary condition (15.38) is unusual since it involves c_t in addition to c and c_x and standard existence and uniqueness theorems do not apply. Uniqueness of solutions is not difficult to prove, and convergence of a difference scheme can be used to establish existence of a solution function $c(t, x)$. Once $c(t, x)$ is known, (15.37) supplies an ordinary differential equation for the determination of $\gamma(t)$.

The differential equation for $c(t, x)$ is inhomogeneous and so are the boundary conditions, so we expect c to be different from 0 and our difference approximation accordingly to be first order in h.

15.4 Estimating the global error

In practice we do not wish to solve extra differential equations in order to gain information on the order and discretization error of our difference schemes. Instead we would use the techniques of Chapter 10 and let the computer do the work.

We only have one independent step size, h, so we perform calculations with three values, h, $2h$, and $4h$ and compare results.

The inverse function $t(x)$ to the boundary function $y(t)$ is the easier one. The calculated values are:

$$t(x_n) \;=\; t(nh) \;=\; t_n \;=\; \sum_1^n k_i .$$

Based on h-values of 1/80, 1/40, 1/20, and 1/10 we have calculated values for $x = 0.1,\ 0.2,\ \ldots,\ 1.0$. In Table 15.1 we supply in column 2 and 3 the order ratios corresponding to the three small step sizes and the three large step sizes, respectively. It is clearly seen that the results are first order and furthermore that they follow the scheme of $2 + \varepsilon$ and $2 + 2\varepsilon$ as formula (10.14) would prescribe. Richardson extrapolation can be performed and the order ratios in column 4 indicate, that the extrapolated results are indeed second order. In columns 5 and 6 we give extrapolated values for $t(x)$ to second and third order, respectively. The error estimate on the values in column 6 indicate that the error is positive and at most 2 units in the last figure.

Table 15.1: Order ratios and extrapolated values for $t(x)$.

x	Order ratios			Extrapolated	
0.1	2.016	2.035	4.442	0.1047219	0.1047188
0.2	2.011	2.023	4.268	0.2180340	0.2180301
0.3	2.008	2.017	4.193	0.3390453	0.3390411
0.4	2.006	2.013	4.150	0.4671460	0.4671419
0.5	2.005	2.010	4.122	0.6018826	0.6018785
0.6	2.004	2.008	4.102	0.7428988	0.7428949
0.7	2.003	2.007	4.086	0.8899047	0.8899009
0.8	2.003	2.006	4.074	1.0426579	1.0426543
0.9	2.003	2.005	4.064	1.2009517	1.2009482
1.0	2.002	2.004	4.056	1.3646068	1.3646035

Because of the variable step sizes in the t-direction we have little control over the t-values where we calculate approximations to $u(t, x)$ and $y(t)$. In order to get function values for specific values of t which we must for order and error estimation, we resort to interpolation.

Since our basic approximations are first order it would seem that linear interpolation would be adequate. For reasons to be elaborated in Appendix C we prefer to go one step further and use 3-point interpolation in order to minimize the effect of the erratic error components introduced by interpolation.

We first look at the boundary function, $y(t)$. Like before we calculate with four h-values from 1/80 and up to 1/10 and supply in Table 15.2 in column 2 and 3 the order ratios corresponding to the three small step sizes and the three large step sizes, respectively. It is clearly seen that the results are first order but the $(2+\varepsilon)$-effect is sometimes drowned by the interference of the interpolation error. Richardson extrapolation can be performed and the order ratios in column 4 indicate, that the extrapolated results are probably second order although the effect of the interpolation error blurs the picture. In column 5 we give extrapolated values for $y(t)$ and based on the error estimates we claim that the error is at most one unit in the last figure.

Table 15.2: Order ratios and extrapolated values for $y(t)$.

t	Order ratios			$y(t)$
0.1	2.045	2.103	4.697	0.09567
0.2	2.018	2.076	8.538	0.18454
0.3	2.042	2.021	1.025	0.26840
0.4	2.027	2.028	2.049	0.34825
0.5	2.019	2.073	7.905	0.42483
0.6	2.022	2.047	4.281	0.49864
0.7	2.017	2.039	4.680	0.57003
0.8	2.019	2.028	3.050	0.63932
0.9	2.018	2.038	4.246	0.70673
1.0	2.014	2.033	4.587	0.77244

For $u(t,x)$ the picture is similar. We refer to Appendix C for sample values of the order ratio and to Table 15.3 for extrapolated values which according to the error estimate are correct up to one unit in the last figure.

Table 15.3: Extrapolated values of $u(t, x)$ using 3-point interpolation.

	$u(t,x)$							
$t \setminus x$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	0.0918							
0.2	0.1717	0.0755						
0.3	0.2437	0.1472	0.0575					
0.4	0.3099	0.2131	0.1226	0.0384				
0.5	0.3715	0.2745	0.1834	0.0982	0.0188			
0.6	0.4294	0.3322	0.2407	0.1547	0.0741			
0.7	0.4843	0.3869	0.2950	0.2083	0.1268	0.0505		
0.8	0.5365	0.4391	0.3467	0.2594	0.1771	0.0998	0.0272	
0.9	0.5865	0.4889	0.3963	0.3085	0.2254	0.1472	0.0736	0.0045
1.0	0.6345	0.5368	0.4439	0.3556	0.2720	0.1929	0.1183	0.0481

Chapter 16

The American Option

16.1 Introduction

Another free boundary problem arises when modeling the price of an American option. Because of the early exercise property we do not know beforehand the extent of the region where the differential equation must be solved. The position of the boundary must be calculated along with the solution. The following is joint work with Asbjørn Trolle Hansen as presented in [15] which is also part of Asbjørn's Ph.D.-thesis [14].

16.2 The mathematical model

The differential equation for the price function for an American put option is

$$u_t = \frac{1}{2}\sigma^2 x^2 u_{xx} + rxu_x - ru, \qquad t > 0, \ \ x > y(t) \tag{16.1}$$

where t is the time to expiry and x is the price of the underlying risky asset. The initial condition is

$$u(0,x) = 0, \qquad x \geq K \tag{16.2}$$

and the boundary conditions are

$$\begin{aligned} \lim_{x\to\infty} u(t,x) &= 0, & t &> 0, & (16.3)\\ u(t,y(t)) &= K - y(t), & t &> 0, & (16.4)\\ u_x(t,y(t)) &= -1, & t &> 0. & (16.5) \end{aligned}$$

The function $y(t)$ is called *the exercise boundary* and this function is not known beforehand except for the information that

$$\lim_{t \to 0} y(t) \quad = \quad K \tag{16.6}$$

where K is the exercise price. The region

$$C \quad = \quad \{(t,x) \mid t > 0, x > y(t)\} \tag{16.7}$$

is called *the continuation region.* It is in this region we seek the solution of (16.1), and it is characterized by the condition that $u(t,x) > K - x$. The region

$$S \quad = \quad \{(t,x) \mid t > 0, x < y(t)\} \tag{16.8}$$

is called *the stopping region.* We can assign the price

$$u(t,x) \quad = \quad K - x, \qquad (t,x) \in S \tag{16.9}$$

but we must emphasize that this is not a solution to (16.1).

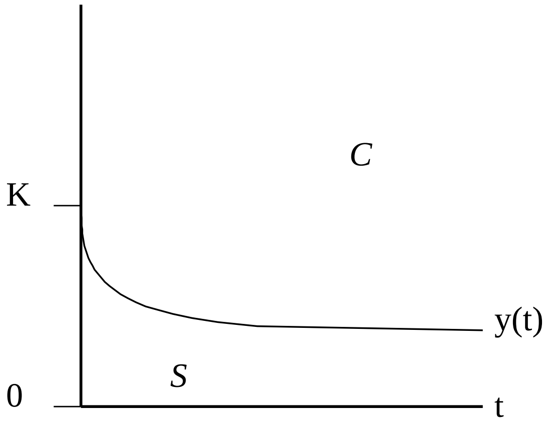


Figure 16.1: The continuation region and the stopping region.

The initial condition is only given for $x \geq K$ because the American option will never expire in the money due to the early exercise feature. If one wishes, one can extend (16.9) to $t = 0$.

The boundary condition (16.3) expresses that the value of the option approaches 0 as the price of the underlying asset approaches infinity.

The boundary condition (16.4) expresses that we are at the exercise boundary, and (16.5) is the *smooth fit* condition expressing that the partial derivative of u

with respect to x is continuous across the boundary if we assume the price from (16.9) in the stopping region.

In [26] the above problem has been studied extensively, and the existence, uniqueness, and differentiability of $u(t, x)$ and $y(t)$ has been shown. We shall now study the behaviour near the boundary further.

Theorem.
a. $y(t) < K$ for $t > 0$.
b. $y'(t) < 0$ for $t > 0$.
c. u_t is continuous across the boundary. More specifically

$$\lim_{s\downarrow t} u_t(s, y(t)) \;=\; \lim_{s\uparrow t} u_t(s, y(t)) \;=\; 0. \tag{16.10}$$

d. u_{xx} is discontinuous across the boundary. More specifically

$$\lim_{s\downarrow t} u_{xx}(s, y(t)) \;=\; \frac{2rK}{(\sigma y(t))^2} \tag{16.11}$$

whereas

$$\lim_{s\uparrow t} u_{xx}(s, y(t)) = 0.$$

Proof. If $y(t) > K$ for some $t > 0$ then $u(t, y(t)) < 0$ by (16.4) which is counterintuitive.
If $y(t) = K$ for some $t > 0$ then $u(t, y(t)) = 0$ and because of (16.5) we would have $u(t, y(t) + \varepsilon) < 0$ for small, positive ε which again is counterintuitive.
Now consider for $k > 0$

$$\begin{aligned} & u(t+k, y(t)) - u(t, y(t)) \\ &\quad = u(t+k, y(t+k)) - u(t+k, y(t+k)) + u(t+k, y(t)) - u(t, y(t)) \\ &\qquad = K - y(t+k) - K + y(t) - (y(t+k) - y(t))u_x(t+k, z) \end{aligned}$$

for some z between $y(t)$ and $y(t+k)$. If we also use the mean value theorem on the very first expression then there is a $\theta \in (0, 1)$ such that

$$ku_t(t + \theta k, y(t)) = (y(t) - y(t+k))(1 + u_x(t+k, z)).$$

From the definition of the continuation region we know that $u(t, x) > K - x$ for $x > y(t)$ and it follows that $u_x(t, x) > -1$ for $x - y(t)$ small and positive. It also follows that $u_t(t, x) > 0$ for $x - y(t)$ small and positive. We therefore must have $y(t) > y(t+k)$ and therefore $y'(t) < 0$ for $t > 0$. Applying the mean value theorem to y, dividing by k, and letting $k \to 0$ gives

$$\lim_{s\downarrow t} u_t(s, y(t)) = 0.$$

The limit from the other side is also 0 since $u(t,x)$ is independent of t in the stopping region and we have established (16.10).
We therefore have

$$\lim_{s\downarrow t}\{\frac{1}{2}\sigma^2x^2u_{xx}(s,y(t))+rxu_x(s,y(t))-ru(s,y(t))\}=0.$$

By (16.4) and (16.5) and the continuity of u and u_x in C (16.11) follows. That the limit from the other side is 0 follows from the fact that u is a linear function in S. □

We conclude from (16.1) and (16.2) that

$$\lim_{t\downarrow 0}u_t(t,x)=0,\qquad x>K$$

but we note that $\lim_{t\to 0}u_t(t,K)$ is undefined. For reasons of monotonicity we expect to have

$$u_t(t,x)>0,\ u_x(t,x)<0,\ u_{xx}(t,x)>0,\quad \text{for}\quad (t,x)\in C.$$

16.3 The boundary condition at infinity

A boundary condition at infinity is impractical when implementing a finite difference method. A commonly used technique to avoid it is to pick some large $L>K$ and replace (16.3) by

$$u(t,L) \;=\; 0,\qquad t>0. \tag{16.12}$$

Since the solution $u(t,x)$ is usually very small for large x the error in using (16.12) instead of (16.3) should be small. But how small is it and what is the effect on the boundary curve.

Let us first look at the steady-state solution to the original problem, i.e. $u(x)=\lim_{t\to\infty}u(t,x)$. Since $\lim_{t\to\infty}u_t(t,x)=0$ we have the following ordinary differential equation problem for $u(x)$

$$\begin{aligned}
\frac{1}{2}\sigma^2x^2u''+rxu'-ru &= 0,\qquad y\le x<\infty, && (16.13)\\
\lim_{x\to\infty}u(x) &= 0, && (16.14)\\
u(y) &= K-y, && (16.15)\\
u'(y) &= -1. && (16.16)
\end{aligned}$$

where $y=\lim_{t\to\infty}y(t)$.
To find the general solution to (16.13) we try with the power function $u(x)=x^z$

and get the characteristic equation

$$\frac{1}{2}\sigma^2 z(z-1) + rz - r = 0$$

or

$$\frac{1}{2}\sigma^2 z^2 + (r - \frac{1}{2}\sigma^2)z - r = 0.$$

The discriminant of this quadratic is

$$\text{disc} = (r - \frac{1}{2}\sigma^2)^2 + 2r\sigma^2 = (r + \frac{1}{2}\sigma^2)^2$$

such that

$$z = \frac{-(r - \frac{1}{2}\sigma^2) \pm (r + \frac{1}{2}\sigma^2)}{\sigma^2}$$

and the two roots are $z = 1$ and $z = -\gamma = -\frac{2r}{\sigma^2}$. The general solution can therefore be written

$$u(x) \quad = \quad A\frac{x}{K} + B(\frac{x}{K})^{-\gamma}. \qquad (16.17)$$

The boundary condition at infinity gives $A = 0$, and from (16.15) and (16.16) we get

$$\begin{aligned} u(y) = B(\frac{y}{K})^{-\gamma} = K - y \;\; &\Rightarrow \;\; B = (K-y)(\frac{y}{K})^{\gamma}, \\ u'(y) = -\gamma\frac{B}{K}(\frac{y}{K})^{-\gamma-1} &= -\gamma\frac{K-y}{K}\frac{K}{y} = -1 \\ \Rightarrow \;\; -\gamma K + \gamma y = -y \;\; &\Rightarrow \;\; y = \frac{\gamma}{\gamma+1}K \end{aligned}$$

$$\Rightarrow \;\; B \quad = \quad K\frac{1}{\gamma+1}(\frac{\gamma}{\gamma+1})^{\gamma}. \qquad (16.18)$$

If we replace the upper boundary condition (16.14) with $u(L) = 0$ the general solution is still (16.17) but the determination of A and B becomes a bit more complicated.

$$u(L) = 0 \;\; \Rightarrow \;\; A\frac{L}{K} + B(\frac{L}{K})^{-\gamma} = 0 \;\; \Rightarrow \;\; B = -A(\frac{L}{K})^{\gamma+1}.$$

The boundary conditions (16.15) and (16.16) now give

$$\begin{aligned} A\frac{y}{K} - A(\frac{L}{K})^{\gamma+1}(\frac{y}{K})^{-\gamma} \quad &= \quad K - y, \\ \frac{A}{K} + \gamma\frac{A}{K}(\frac{L}{K})^{\gamma+1}(\frac{y}{K})^{-\gamma-1} \quad &= \quad -1. \end{aligned}$$

The second equation gives

$$(\frac{y}{L})^{-\gamma-1} = \frac{K+A}{-\gamma A} \quad \Rightarrow \quad y = L(\frac{-\gamma A}{K+A})^{1/(\gamma+1)}$$

and the first equation now gives

$$\frac{A}{K}(1+\frac{A}{K})^{\gamma} \;\; = \;\; -\frac{1}{\gamma}(\frac{\gamma}{\gamma+1}\frac{K}{L})^{\gamma+1} \;\; = \;\; -\alpha. \tag{16.19}$$

We cannot give a closed form solution for A from (16.19) but when α is small, A/K will also be small, and $(1 + A/K)^{\gamma}$ will be close to 1, and an approximate value is $A^{(1)} = -\alpha K$. A better value can be obtained by Newton-iteration where the next iterate will be

$$A^{(2)} = -(\alpha + \beta)K$$

with

$$\beta = \alpha \frac{(1-\alpha)((1-\alpha)^{-\gamma} - 1)}{1 - \alpha(\gamma+1)}.$$

Table 16.1: Corresponding values of L, α, β, A, B, and y.

L	α	β	A	B	y
120	0.022 431	0.003 044	−2.5527	7.6224	85.516
140	0.008 896	0.000 426	−0.9322	7.0191	84.117
200	0.001 047	0.000 006	−0.1052	6.7333	83.421
∞				6.6980	83.333

In Table 16.1 we supply values for α, β, A, B, and y for various values of L and corresponding to $K = 100$, $\sigma = 0.2$, $r = 0.1$ and therefore $\gamma = 5$.

Denote the solution function and the boundary function corresponding to a finite value of L by $u_L(t, x)$ and $y_L(t)$, respectively. We note that the limit value $\lim_{t\to\infty} y_L(t)$ moves upwards as the value for L decreases and we conclude that this holds for the whole boundary curve since otherwise the boundary curves for two different values of L would intersect.

We want to estimate the error in u and y when we use a finite value, L, so we define the error function

$$w(t, x) = u(t, x) - u_L(t, x).$$

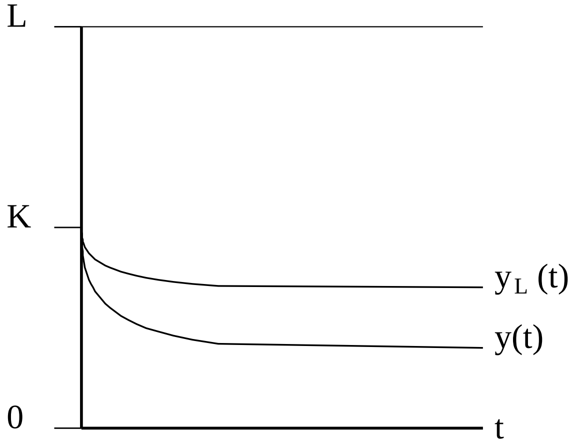


Figure 16.2: $y(t)$ and $y_L(t)$.

It is defined in the same region as u_L since $y_L(t) > y(t)$ and here it satisfies the same differential equation:

$$w_t = \frac{1}{2}\sigma^2 x^2 w_{xx} + rxw_x - rw, \quad t \geq 0, \ \ y_L(t) \leq x \leq L. \tag{16.20}$$

The initial condition is

$$w(0,x) = 0, \qquad K \leq x \leq L, \tag{16.21}$$

and the boundary conditions are

$$\begin{aligned} w(t,L) &= u(t,L), & t &\geq 0, & (16.22)\\ w(t,y_L(t)) &= u(t,y_L(t)) - K + y_L(t), & t &> 0, & (16.23)\\ w_x(t,y_L(t)) &= u_x(t,y_L(t)) + 1, & t &> 0. & (16.24) \end{aligned}$$

Using Taylor expansion we get

$$\begin{aligned} u(t,y_L(t)) &= u(t,y(t)) + (y_L(t)-y(t))u_x + \frac{1}{2}(y_L(t)-y(t))^2 u_{xx} + \cdots \\ &= K - y(t) - y_L(t) + y(t) + \frac{1}{2}(y_L(t)-y(t))^2 \frac{2rK}{\sigma^2 y(t)^2} + \cdots \\ &= K - y_L(t) + (\frac{y_L(t)}{y(t)} - 1)^2 \frac{rK}{\sigma^2} + \cdots, \\ u_x(t,y_L(t)) &= -1 + (\frac{y_L(t)}{y(t)} - 1)\frac{2rK}{\sigma^2 y(t)} + \cdots \end{aligned}$$

so the conditions (16.23) and (16.24) read

$$\begin{aligned} w(t,y_L(t)) &= (\frac{y_L(t)}{y(t)}-1)^2\frac{rK}{\sigma^2}+\cdots, & t>0, \\ w_x(t,y_L(t)) &= (\frac{y_L(t)}{y(t)}-1)\frac{2rK}{\sigma^2 y(t)}+\cdots, & t>0. \end{aligned}$$

We note that $w(t,L)>0$, $w(t,y_L(t))>0$, $w_x(t,y_L(t))>0$, $w_t(t,y_L(t))>0$, and $w_t(t,L)>0$. This suggests that $w(t,x)>0$ and is increasing with t and x. We therefore have

$$0 \le w(t,x) < \lim_{t\to\infty} w(t,x) = u(x)-u_L(x) \le u(L) = B(\frac{K}{L})^\gamma$$

with B given by (16.18). With the above parameter values $K=100$, $\sigma=0.2$, $r=0.1$, and $L=200$, the error for at the money options is bounded by $u(K)-u_L(K)=0.07$, and the maximum error for any x is bounded by $u(L)=0.21$.

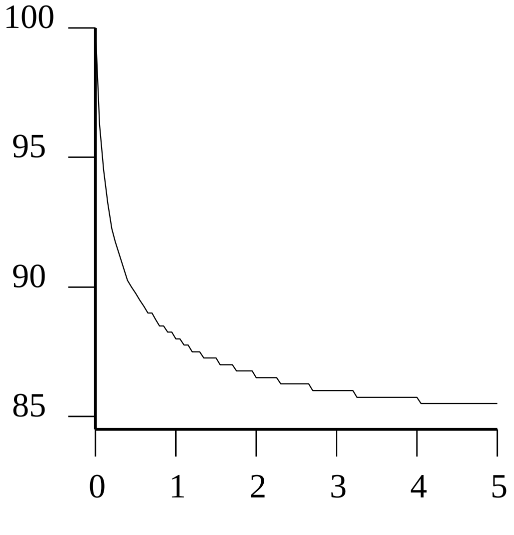


Figure 16.3: The boundary curve calculated with Brennan-Schwartz.

16.4 Finite difference schemes

If we introduce a traditional grid with fixed step sizes k and h then we face the problem that the boundary curve, $y(t)$, typically passes between grid points. There are various ways to deal with this difficulty.

Table 16.2: The price function calculated with Brennan-Schwartz.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	4.0
195	0.000	0.000	0.001	0.002	0.005	0.008	0.011	0.014	0.019
190	0.000	0.000	0.002	0.005	0.011	0.018	0.024	0.031	0.042
185	0.000	0.000	0.003	0.009	0.019	0.029	0.040	0.050	0.068
180	0.000	0.001	0.005	0.014	0.028	0.044	0.059	0.074	0.099
175	0.000	0.001	0.008	0.022	0.041	0.062	0.083	0.103	0.136
170	0.000	0.002	0.012	0.032	0.058	0.086	0.113	0.139	0.181
165	0.000	0.004	0.019	0.047	0.081	0.117	0.152	0.184	0.237
160	0.000	0.007	0.031	0.069	0.113	0.159	0.202	0.241	0.305
155	0.000	0.012	0.048	0.100	0.158	0.214	0.267	0.313	0.390
150	0.001	0.022	0.076	0.146	0.219	0.288	0.351	0.407	0.497
145	0.002	0.041	0.119	0.211	0.303	0.387	0.462	0.528	0.632
140	0.006	0.073	0.185	0.306	0.419	0.521	0.609	0.685	0.805
135	0.016	0.129	0.287	0.442	0.580	0.700	0.803	0.890	1.025
130	0.040	0.226	0.442	0.636	0.803	0.943	1.060	1.158	1.308
125	0.097	0.390	0.676	0.915	1.111	1.271	1.402	1.511	1.675
120	0.223	0.661	1.026	1.311	1.535	1.714	1.859	1.977	2.154
115	0.488	1.101	1.545	1.872	2.122	2.316	2.471	2.596	2.781
110	1.009	1.795	2.306	2.665	2.931	3.135	3.295	3.423	3.611
105	1.962	2.868	3.411	3.780	4.048	4.252	4.410	4.535	4.718
100	3.519	4.456	4.988	5.340	5.591	5.780	5.926	6.040	6.207
95	6.142	6.836	7.249	7.530	7.731	7.885	8.004	8.096	8.232
90		10.222	10.430	10.589	10.708	10.803	10.877	10.934	11.022

At a given time level we can artificially move the boundary curve to the nearest grid point. We hereby introduce an error in the x-direction of order h and this is undesirable.

We can also introduce difference approximations to u_x and u_{xx} based on uneven steps. This is a viable approach when the boundary curve is known beforehand, but things become complicated when $y(t)$ has to be determined along with $u(t,x)$.

A third method was proposed by Brennan & Schwartz in 1977 [4]. They start with the initial values from (16.2) augmented with values from (16.9) for $t = 0$ and $x < K$. They choose a step size k in the time direction and a step size h in the x-direction such that K/h is an integer, and a value $L = Mh$ where M is another integer. They then perform a Crank-Nicolson step to $t = k$ computing a set of auxiliary values ν_m^1, $m = 0, 1, \ldots, M = L/h$ with boundary values $\nu_0^1 = K$ and $\nu_M^1 = 0$. The solution values are now determined as $v_m^1 = \max(K - mh, \nu_m^1)$ and the position of the exercise boundary is given by $\bar{y}_L(k) = h \cdot \max\{m | \nu_m^1 \leq K - mh\}$. Once v^1 has been determined we can move on to v^2, v^3, etc. This

Table 16.3: Order ratios corresponding to h.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
195			4.7	2.8	3.5	3.7	3.7	3.8	3.9	3.8
190		4.1	4.9	2.9	3.5	3.7	3.8	3.8	3.9	3.8
185		4.1	5.1	3.1	3.5	3.7	3.8	3.8	3.9	3.8
180		4.1	6.0	3.2	3.6	3.7	3.8	3.8	3.9	3.8
175		4.2	12.7	3.3	3.6	3.7	3.8	3.8	3.9	3.8
170		4.2	0.5	3.4	3.6	3.7	3.8	3.9	3.8	3.8
165	4.0	4.4	2.5	3.5	3.6	3.7	3.8	3.9	3.8	3.8
160	4.0	4.6	3.1	3.5	3.7	3.7	3.8	3.9	3.8	3.8
155	4.0	5.5	3.3	3.6	3.7	3.8	3.8	3.9	3.8	3.8
150	4.0	-2.4	3.4	3.6	3.7	3.8	3.9	3.9	3.8	3.8
145	4.0	2.8	3.5	3.7	3.7	3.8	3.9	3.9	3.8	3.8
140	4.1	3.3	3.6	3.7	3.8	3.8	3.9	3.8	3.8	3.8
135	4.3	3.5	3.6	3.7	3.8	3.8	3.9	3.8	3.8	3.9
130	-4.0	3.6	3.7	3.8	3.8	3.9	3.9	3.8	3.8	3.9
125	3.4	3.6	3.7	3.8	3.8	4.0	3.9	3.7	3.8	4.0
120	3.6	3.7	3.7	3.8	3.8	4.1	3.8	3.7	3.8	4.1
115	3.6	3.7	3.7	3.9	3.8	4.2	3.7	3.7	3.9	4.2
110	3.6	3.7	3.7	3.8	3.9	4.3	3.4	3.7	3.9	4.5
105	3.6	3.7	3.7	3.6	4.1	4.4	3.2	3.9	4.0	4.9
100	2.7	2.0	1.4	1.1	0.8	0.6	0.5	0.4	0.3	0.2
95	-2.5	-30.2	11.4	5.2	4.3	2.4	2.3	2.4	2.6	3.0
90		-1.1	-1.1	-0.8	-1.1	-0.5	-0.7	-1.3	-1.9	-1.4

may seem like a rather harsh treatment of the problem, but the results seem reasonable at a first glance.

In Fig. 16.3 we show a plot of the computed boundary curve for the parameter values $K = 100$, $\sigma = 0.2$, $r = 0.1$, and $L = 200$, and calculated with $h = 0.25$ and $k = 0.0625$. In Table 16.2 we give values for the price function for a selection of points in the continuation region.

In order to estimate the error we apply the techniques of chapter 10. In Table 16.3 we supply the order ratios which should be close to 4.0 if the method is second order in h. This looks fairly reasonable for $x > K$. Occasional isolated deviations correspond to small values of the associated error estimate which is shown in Table 16.4. For $x \leq K$ the order determination is far from reliable and the error estimate due to the x-discretisation is unfortunately also much larger here, close to the exercise boundary.

Table 16.4: Error estimate *1000 corresponding to h.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
195			-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
190		-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
185		-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
180		-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
175		-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
170		-0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
165	-0.00	-0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01
160	-0.00	-0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
155	-0.00	-0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
150	-0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.02
145	-0.00	0.00	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02
140	-0.00	0.00	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03
135	-0.00	0.01	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03
130	0.00	0.02	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04
125	0.01	0.03	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05
120	0.03	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
115	0.07	0.08	0.08	0.08	0.08	0.07	0.08	0.07	0.07	0.08
110	0.13	0.12	0.11	0.10	0.10	0.09	0.09	0.08	0.09	0.09
105	0.15	0.17	0.15	0.13	0.12	0.10	0.11	0.09	0.12	0.11
100	-6.74	-8.53	-9.31	-9.57	-9.52	-9.30	-8.98	-8.59	-8.12	-7.71
95	-0.50	-0.06	0.18	0.30	0.42	0.50	0.45	0.47	0.51	0.40
90		-1.90	-1.72	-1.57	-1.43	-1.07	-1.17	-0.91	-0.81	-0.91

In Table 16.5 we supply the similar ratios corresponding to the time discretisation. Values close to 2.0 indicate that the method is first order in k contrary to our expectations of a method based on Crank-Nicolson. Also here the order determination leaves a lot to be desired for $x \leq K$, and the error estimate due to the time discretisation is also much larger here as seen in Table 16.6. We must conclude that the Brennan-Schwartz approach is not ideal.

16.5 Varying the time steps

For the American option problem where the boundary curve is known to be monotonic we can suggest an alternate approach. Since the finite difference methods which we are usually considering for parabolic problems are one-step methods there is no need to keep the step size in time constant throughout the calcu-

Table 16.5: Order ratios corresponding to k.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
195	5.8	3.6	0.7	2.2	2.1	2.0	1.9	1.9	1.9	1.9
190	5.5	3.7	1.0	2.2	2.1	2.0	1.9	1.9	1.9	1.9
185	4.9	3.9	1.3	2.2	2.1	2.0	1.9	1.9	1.9	1.9
180	4.4	4.3	1.7	2.2	2.1	2.0	1.9	1.9	1.9	1.9
175	4.0	5.1	1.9	2.2	2.0	2.0	1.9	1.9	1.9	1.9
170	3.6	7.7	2.1	2.1	2.0	1.9	1.9	1.9	1.9	1.9
165	3.3	-25.1	2.2	2.1	2.0	1.9	1.9	1.9	1.9	1.9
160	3.0	-0.3	2.2	2.1	2.0	1.9	1.9	1.9	1.9	1.9
155	2.9	1.5	2.2	2.0	1.9	1.9	1.9	1.9	1.9	1.9
150	3.0	2.2	2.2	2.0	1.9	1.9	1.9	1.9	1.9	1.9
145	3.5	2.6	2.1	1.9	1.9	1.9	1.9	1.9	1.9	1.9
140	6.0	2.7	2.0	1.9	1.9	1.9	1.9	1.9	1.9	1.9
135	-2.7	2.5	2.0	1.9	1.8	1.8	1.9	1.9	1.9	1.9
130	1.0	2.2	1.9	1.8	1.8	1.8	1.8	1.9	1.9	1.9
125	2.4	1.8	1.8	1.8	1.8	1.8	1.9	1.9	1.9	1.9
120	4.3	1.6	1.7	1.8	1.8	1.9	1.9	1.9	1.9	1.9
115	5.5	2.2	1.7	1.7	1.8	1.8	1.9	1.9	1.9	1.9
110	-0.1	2.4	2.3	2.0	1.9	1.9	1.8	1.8	1.9	1.9
105	-11.0	-0.6	0.6	1.3	1.7	2.0	2.1	2.1	2.1	2.1
100	1.8	1.9	2.0	2.0	1.9	1.9	1.9	1.9	1.8	1.8
95	-5.4	-1.5	-0.5	-0.0	0.3	0.5	0.6	0.7	0.8	0.9
90			1.2	1.2	1.2	1.3	1.4	1.4	1.4	1.5

lation. Instead we propose to choose the step sizes k_n such that the boundary curve will pass exactly through grid points. This idea was proposed for the original Stefan Problem by Douglas & Gallie [9]. In our case the boundary curve is decreasing so we shall choose k_n such that there is precisely one extra grid point at the next time level, see Fig. 16.4. This will imply that the initial time steps will be very small and then keep increasing, but as we shall see that is not such a bad idea from other points of view as well.

16.6 The implicit method

Since there is exactly one extra grid point on the next time level the situation is ideally suited for the implicit method which will never have to refer to points outside the continuation region.

Table 16.6: Error estimate *100 corresponding to k.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	4.0
195	-0.00	-0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02
190	-0.00	-0.00	0.00	0.01	0.01	0.02	0.02	0.03	0.04
185	-0.00	-0.00	0.00	0.01	0.02	0.03	0.04	0.05	0.06
180	-0.00	-0.00	0.01	0.02	0.03	0.05	0.06	0.07	0.09
175	-0.00	-0.00	0.01	0.03	0.05	0.07	0.08	0.10	0.12
170	-0.00	-0.00	0.02	0.04	0.07	0.09	0.11	0.13	0.15
165	-0.00	0.00	0.03	0.06	0.09	0.12	0.15	0.17	0.20
160	-0.01	0.01	0.05	0.09	0.12	0.16	0.19	0.22	0.26
155	-0.01	0.02	0.07	0.12	0.17	0.21	0.25	0.28	0.33
150	-0.02	0.04	0.11	0.17	0.23	0.28	0.32	0.36	0.41
145	-0.02	0.08	0.16	0.24	0.30	0.36	0.42	0.46	0.52
140	-0.02	0.13	0.23	0.33	0.41	0.48	0.54	0.59	0.66
135	0.03	0.22	0.34	0.45	0.55	0.63	0.70	0.75	0.83
130	0.15	0.34	0.49	0.63	0.74	0.83	0.91	0.96	1.05
125	0.36	0.51	0.71	0.87	1.00	1.10	1.18	1.24	1.33
120	0.54	0.77	1.01	1.21	1.35	1.46	1.54	1.60	1.69
115	0.66	1.15	1.48	1.68	1.83	1.94	2.02	2.08	2.16
110	1.67	1.65	2.06	2.33	2.49	2.59	2.66	2.70	2.77
105	0.60	3.36	3.56	3.48	3.43	3.43	3.45	3.48	3.54
100	19.86	15.82	13.94	12.90	12.22	11.72	11.32	10.99	10.46
95	4.33	7.09	7.09	6.94	6.86	6.82	6.81	6.80	6.77
90		9.50	9.30	9.07	8.85	8.63	8.50	8.45	8.36

Like Brennan and Schwartz we choose a step size in the x-direction, $h = K/M_0$, where M_0 is a positive integer, and such that $M = L/h$ is also an integer. The step size h is kept fixed during the computation. The number of grid points above the exercise boundary is thus $M - M_0$ at time $t = 0$ which corresponds to the expiration time. For each time step we add one extra grid point in the x-direction such that at the end of the n-th time step we have $M - M_n$ grid points above the boundary where $M_n = M_0 - n$. The time steps will be denoted k_1, k_2, ... and we define $t_n = \sum_{i=1}^{n} k_i$. We now compute a grid function

$$v_m^n = v(t_n, mh) \approx u(t_n, mh)$$

satisfying

$$\begin{aligned} \frac{v_m^n - v_m^{n-1}}{k_n} &= \frac{\sigma^2}{2} m^2 (v_{m+1}^n - 2v_m^n + v_{m-1}^n) \qquad (16.25) \\ &\quad + \frac{r}{2} m (v_{m+1}^n - v_{m-1}^n) - r v_m^n, \qquad m = M_n + 1, \ldots, M - 1, \end{aligned}$$

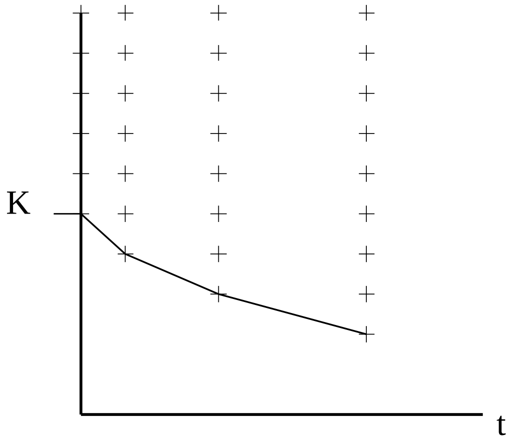


Figure 16.4: The first few grid lines.

$$\begin{aligned}
v_m^0 &= 0, & m = M_0, \ldots, M, && (16.26)\\
v_M^n &= 0, & && (16.27)\\
v_{M_n}^n &= K - M_n h \ = \ nh, & && (16.28)\\
-1 &= \frac{-v_{m+2}^n + 4v_{m+1}^n - 3v_m^n}{2h}, & m = M_n, && (16.29)
\end{aligned}$$

$n = 1, 2, \ldots$. The first order approximation to the boundary derivative $v_{m+1}^n - v_m^n = -h$ is abandoned in favour of (16.29) for reasons of accuracy and because of problems with the initial steps.

The equations (16.25) – (16.29) are non-linear so we propose an iterative approach. We guess a value for k_n, solve the linear tridiagonal system (16.25), (16.27), and (16.28), and use formula (16.29) to correct the time step. Alternatively we could solve the (almost) tridiagonal system determined by (16.25), (16.27), and (16.29), and use formula (16.28) to correct the time step. Numerical experiments, however, suggest that the former method is computationally more efficient in that it requires fewer iterations when searching for the size of the next time step.

Getting Started

For $n = 1$ equations (16.28) and (16.29) give

$$-v_{m+2}^1 + 4v_{m+1}^1 = h, \qquad (m = M_1 = M_0 - 1).$$

When h is small we expect v^1_{m+2} to be very small. We put $v^1_{m+2} = \varepsilon h$ and get

$$v^1_{m+1} = \frac{h}{4}(1+\varepsilon), \qquad\qquad m = M_1.$$

Applying (16.25) with $m = M_1 + 1 = M_0 = K/h$ we then get

$$\begin{aligned} \frac{h(1+\varepsilon)}{4k_1} &= \frac{\sigma^2K^2}{2h^2}(\varepsilon - \frac{1+\varepsilon}{2} + 1)h + \frac{rK}{2h}(\varepsilon - 1)h - r\frac{h}{4}(1+\varepsilon) \\ \Rightarrow \frac{1}{k_1} &= \frac{\sigma^2K^2}{h^2} - \frac{1-\varepsilon}{1+\varepsilon}\frac{2rK}{h} - r \\ \Rightarrow k_1 &= \frac{h^2}{\sigma^2K^2 - 2rKh\frac{1-\varepsilon}{1+\varepsilon} - rh^2} \approx \frac{h^2}{\sigma^2K^2 - 2rKh - rh^2} \end{aligned} \qquad (16.30)$$

and we have a good initial guess for the size of the first time step. Note further that

$$k_1 \approx \frac{h^2}{\sigma^2K^2}$$

so it appears that the boundary curve $y(T)$ starts out from $(0, K)$ with a vertical tangent and with a shape much like a parabola with its apex at $(0, K)$.

For $n > 1$ we could as a starting value for k_n use k_{n-1}, but for $n > 2$ it turns out to be more efficient to use

$$k^1_n \;=\; 2k_{n-1} - k_{n-2} \qquad (16.31)$$

based on the assumption that the second derivative of $y_L(t)$ is small. Given the initial value k^1_n, a good second value can be obtained as follows. First we solve equations (16.25) – (16.28) and obtain tentative values for $v^n_m, m = M_n, \ldots, M$. We then evaluate the accuracy of k^1_n by calculating the error term

$$e \;=\; -1 - \frac{-v^n_{M_n+2} + 4v^n_{M_n+1} - 3v^n_{M_n}}{2h}. \qquad (16.32)$$

If $e = 0$ then k^1_n is the right size of the time step. Otherwise we would like to change k_n such that $e = 0$. First notice that

$$\Delta e = -\frac{4\Delta v^n_{M_n+1} - \Delta v^n_{M_n+2}}{2h} \approx -\frac{3\Delta v^n_{M_n+1}}{2h} \approx -\frac{3\frac{\partial v^n_{M_n+1}}{\partial t}\Delta t}{2h}.$$

An approximation to the time derivative at grid point (n, M_n+1) can be obtained by

$$\frac{\partial v^n_{M_n+1}}{\partial t} \approx \frac{v^n_{M_n+1} - v^{n-1}_{M_n+1}}{k^1_n}$$

and since we want $\Delta e = -e$ our second value for k_n becomes

$$k^2_n \;=\; k^1_n + \frac{2}{3}\frac{hk^1_ne}{v^n_{M_n+1} - v^{n-1}_{M_n+1}}. \qquad (16.33)$$

Iterating k_n

With a proposed value for k_n we can rewrite equation (16.25) as

$$v_m^n - \alpha_m k_n (v_{m+1}^n - 2v_m^n + v_{m-1}^n) - \beta_m k_n (v_{m+1}^n - v_{m-1}^n) + r k_n v_m^n = v_m^{n-1}$$

with $\alpha_m = \frac{1}{2}\sigma^2 m^2$ and $\beta_m = \frac{1}{2} rm$. Collecting terms we get

$$(\beta_m - \alpha_m) k_n v_{m-1}^n + (1 + (2\alpha_m + r)k_n) v_m^n - (\beta_m + \alpha_m) k_n v_{m+1}^n \quad = \quad v_m^{n-1}, \quad (16.34)$$

$m = M_n + 1, \ldots, M-1$. With the two boundary conditions (16.27) and (16.28) we have as many equations as unknowns and we can solve the resulting system of equations. The calculated values are then checked with equation (16.29). If they do not fit, and they seldom do the first time around, the value for k_n is adjusted, and the equations are solved again until a satisfactory agreement with (16.29) is achieved.

The first and the second value for k_n have been discussed above. The general way of calculating k_n^{i+1} from k_n^{i-1} and k_n^i for $i > 1$ is by using the secant method on formula (16.32):

$$k_n^{i+1} \quad = \quad k_n^i - e^i \frac{k_n^i - k_n^{i-1}}{e^i - e^{i-1}} \qquad (16.35)$$

where e^i is calculated from (16.32) with v-values calculated with k_n^i. The iteration is continued until two successive values of k_n^i differ by less than a predetermined tolerance.

Table 16.7 illustrates the time step iterations corresponding to the usual set of parameter values $K = 100$, $\sigma = 0.2$, $r = 0.1$, and $L = 200$, and calculated with $h = 1$. We note that the initial guess for k_1 from (16.30) is reasonably good, and so is the second guess from (16.33). k_1 is a reasonable initial guess for k_2 which turns out to be only slightly larger than k_1, thus not supporting the parabola hypothesis. For the subsequent steps (16.31) is a reasonable first guess although it always undershoots. In all cases the subsequent secant method displays rapid convergence. We have chosen a tolerance of 10^{-10} because even when only a limited accuracy is demanded in the final results it is important that the time steps are correct. And since the secant method has superlinear convergence the last decimals are inexpensive (cf. Table 16.14).

Because the boundary curve has a horizontal asymptote the time steps must eventually increase without bound as we approach the magical value of y_L. We stop the calculations when the proposed time step exceeds a predetermined large value (or becomes negative).

Table 16.7: Time step iterations with the implicit method.

n	1	2	3	4	5	6
1	0.00263227	0.00262001	0.00261359	0.00261361	0.00261361	
2	0.00261361	0.00290627	0.00305600	0.00306050	0.00306055	0.00306055
3	0.00350749	0.00387369	0.00404826	0.00404955	0.00404955	
4	0.00503856	0.00595513	0.00624391	0.00624057	0.00624058	
5	0.00843160	0.01026408	0.01041317	0.01041053	0.01041053	
6	0.01458048	0.01695680	0.01683730	0.01683827	0.01683827	
7	0.02326601	0.02603689	0.02574487	0.02574540		
8	0.03465253	0.03849420	0.03797056	0.03796966	0.03796966	
9	0.05019391	0.05598995	0.05510357	0.05509914	0.05509915	
10	0.07222864	0.08125947	0.07982226	0.07980969	0.07980971	
11	0.10452027	0.11917606	0.11686413	0.11683260	0.11683268	
12	0.15385565	0.17916186	0.17537148	0.17529225	0.17529254	
13	0.23375241	0.28168447	0.27523543	0.27502091	0.27502202	
14	0.37475150	0.47960268	0.46820027	0.46753680	0.46754150	0.46754149
15	0.66006096	0.95513540	0.93700802	0.93475183	0.93476999	0.93476997
16	1.40199846	2.89947139	2.99960311	3.04665061	3.04805724	3.04807080

16.7 Crank-Nicolson

When attempting to use the Crank-Nicolson method on this problem there is a minor complication. Since the boundary curve is decreasing, the first interior grid point at a particular time level corresponds to a boundary point at the previous time level. If we want to use a Crank-Nicolson approximation here we shall refer to a point outside the continuation region at the previous time level. Such a point is usually called a fictitious point and it must be treated separately. There are (at least) three suggestions:

1. Use the value $K - x$ at the fictitious point. This is not correct but reasonably close.

2. Extrapolate from the boundary point and its neighbour using the central difference approximation to u_x which we know is -1.

3. Replace the difference approximations to u_x and u_{xx} on the boundary by the 'exact' values of -1 and $\frac{2rK}{\sigma^2 y(t)^2}$ from (16.5) and (16.11).

In practice these three approaches perform equally well. We shall therefore concentrate on the method described in **3.** and note that the details in implementing

1. and **2.** can be filled in similarly. We do make one exception from the methodology described in **3.** At the initial step neither $u_x(0,K)$ nor $u_{xx}(0,K)$ is defined so we need to deal with the initial step differently. Instead we look at suggestion **1.** and approximate

$$u_x(0,K) \quad \text{by} \quad \frac{u(0,K+h)-u(0,K-h)}{2h} \;=\; -\frac{1}{2}$$

and

$$u_{xx}(0,K) \quad \text{by} \quad \frac{u(0,K+h)-2u(0,K)+u(0,K-h)}{h^2} \;=\; \frac{1}{h}.$$

With notation as in the previous section we compute a grid function

$$v_m^n = v(t_n, mh) \approx u(t_n, mh)$$

satisfying

$$\begin{aligned}
\frac{v_m^1 - v_m^0}{k_n} &= \frac{\sigma^2}{4}m^2(v_{m+1}^1 - 2v_m^1 + v_{m-1}^1 + h) && (16.36)\\
&\quad + \frac{r}{4}m(v_{m+1}^1 - v_{m-1}^1 - h) - \frac{r}{2}v_m^1, \qquad m = M_0, n = 1, &&\\
\frac{v_m^n - v_m^{n-1}}{k_n} &= \frac{\sigma^2}{4}m^2(v_{m+1}^n - 2v_m^n + v_{m-1}^n) + \frac{rK}{2} && (16.37)\\
&\quad + \frac{r}{4}m(v_{m+1}^n - v_{m-1}^n - 2h) - \frac{r}{2}(v_m^n + v_m^{n-1}), \qquad m = M_{n-1}, n > 1, &&\\
\frac{v_m^n - v_m^{n-1}}{k_n} &= \frac{\sigma^2}{4}m^2(v_{m+1}^n - 2v_m^n + v_{m-1}^n + v_{m+1}^{n-1} - 2v_m^{n-1} + v_{m-1}^{n-1}) && (16.38)\\
&\quad + \frac{r}{4}m(v_{m+1}^n - v_{m-1}^n + v_{m+1}^{n-1} - v_{m-1}^{n-1}) - \frac{r}{2}(v_m^n + v_m^{n-1}), &&\\
&\qquad m = M_n + 2, \ldots, M-1, &&\\
v_m^0 &= 0, \qquad m = M_0, \ldots, M, && (16.39)\\
v_M^n &= 0, && (16.40)\\
v_{M_n}^n &= K - M_n h \;=\; nh, && (16.41)\\
-1 &= \frac{-v_{m+2}^n + 4v_{m+1}^n - 3v_m^n}{2h}, \qquad m = M_n, && (16.42)
\end{aligned}$$

$n = 1, 2, \ldots$. Just as for the implicit method we guess a value for k_n, solve the tridiagonal system of equations given by (16.36) – (16.41), and use (16.42) to correct the time step. Alternatively (16.42) could be incorporated in the system of equations and (16.41) be used for the correction. We prefer the former since it appears to lead to fewer iterations.

Getting Started

For $n = 1$ equations (16.41) and (16.42) give

$$-v_{m+2}^1 + 4v_{m+1}^1 = h, \qquad (m = M_1 = M_0 - 1).$$

When h is small we expect v^1_{m+2} to be very small. We put $v^1_{m+2} = \varepsilon h$ and get

$$v^1_{m+1} = \frac{h}{4}(1+\varepsilon), \qquad\qquad m = M_1.$$

Applying (16.36) with $m = M_1 + 1 = M_0 = K/h$ we then get

$$\begin{aligned}
\frac{h(1+\varepsilon)}{4k_1} &= \frac{\sigma^2 m^2}{4}(\varepsilon - \frac{1+\varepsilon}{2} + 2)h + \frac{rm}{4}(\varepsilon - 2)h - \frac{r}{2}\frac{h}{4}(1+\varepsilon) \\
\Rightarrow \frac{1}{k_1} &= \sigma^2 m^2(\frac{1}{2} + \frac{1}{1+\varepsilon}) - \frac{2-\varepsilon}{1+\varepsilon} rm - \frac{1}{2}r \\
\Rightarrow k_1 &= \frac{h^2}{\sigma^2 K^2(\frac{1}{2} + \frac{1}{1+\varepsilon}) - rKh\frac{2-\varepsilon}{1+\varepsilon} - \frac{1}{2}rh^2}
\end{aligned}$$

and we have a good initial guess for the size of the first time step. Note further that

$$k_1 \approx \frac{h^2}{\frac{3}{2}\sigma^2 K^2 - 2rKh - \frac{1}{2}rh^2} \approx \frac{h^2}{\frac{3}{2}\sigma^2 K^2} \tag{16.43}$$

Table 16.8: Time step iterations with Crank-Nicolson.

n	1	2	3	4	5	6
1	0.00172429	0.00181050	0.00176141	0.00176115	0.00176116	
2	0.00528347	0.00534501	0.00575238	0.00575488	0.00575490	
3	0.00974864	0.00803102	0.00654763	0.00659399	0.00659352	0.00659352
4	0.00743214	0.00779111	0.00804438	0.00804408	0.00804408	
5	0.00949463	0.01017074	0.01067041	0.01066699	0.01066699	
6	0.01328991	0.01456567	0.01543735	0.01542461	0.01542468	
7	0.02018236	0.02222535	0.02337563	0.02335643	0.02335653	
8	0.03128838	0.03377432	0.03489219	0.03488019	0.03488023	
9	0.04640393	0.04936500	0.05056048	0.05055332	0.05055333	
10	0.06622644	0.07076255	0.07257623	0.07256957	0.07256957	
11	0.09458581	0.10195371	0.10499982	0.10499678		
12	0.13742398	0.14955464	0.15495471	0.15496947	0.15496949	
13	0.20494220	0.22635535	0.23713380	0.23722496	0.23722525	
14	0.31948101	0.36112641	0.38639870	0.38687650	0.38688015	
15	0.53653504	0.63033291	0.70622891	0.70946632	0.70953139	0.70953144
16	1.03218273	1.30246815	1.65738437	1.70025643	1.70343860	1.70346436

Table 16.9: Order ratios for the implicit method.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
195	6.9	1.7	2.0	1.9	1.5	-	5.7	2.0	1.7	1.6
190	6.4	1.7	2.0	1.9	1.5	-	5.6	2.0	1.7	1.6
185	5.8	1.7	2.0	1.9	1.5	6.6	5.3	2.0	1.7	1.7
180	5.2	1.7	2.0	1.9	1.5	4.5	5.1	2.0	1.7	1.7
175	4.5	1.6	2.0	1.9	1.5	3.4	4.8	2.0	1.7	1.7
170	4.0	1.6	2.0	1.9	1.3	2.7	4.5	2.0	1.7	1.7
165	3.5	1.6	1.9	1.8	0.9	2.3	4.3	2.1	1.8	1.7
160	3.1	1.5	1.9	1.7	7.1	2.1	4.1	2.1	1.8	1.7
155	2.8	1.5	1.9	1.3	2.5	1.9	3.9	2.1	1.8	1.8
150	2.5	1.5	1.9	-2.4	2.2	1.8	3.7	2.1	1.8	1.8
145	2.3	1.5	1.9	3.7	2.1	1.7	3.6	2.2	1.9	1.9
140	2.1	1.5	1.8	2.9	2.1	1.7	3.5	2.2	1.9	1.9
135	2.0	1.5	2.2	2.7	2.1	1.6	3.5	2.2	2.0	1.9
130	1.9	1.6	2.1	2.7	2.1	1.6	3.5	2.3	2.0	2.0
125	1.8	1.5	2.2	2.7	2.2	1.6	3.4	2.3	2.1	2.0
120	1.6	1.6	2.2	2.7	2.2	1.6	3.4	2.4	2.1	2.1
115	2.3	1.7	2.3	2.7	2.3	1.6	3.4	2.4	2.1	2.1
110	2.1	1.7	2.3	2.7	2.3	1.6	3.4	2.5	2.2	2.2
105	2.1	1.8	2.3	2.8	2.3	1.6	3.4	2.5	2.2	2.2
100	2.2	1.8	2.4	2.8	2.3	1.6	3.4	2.5	2.2	2.2
95	2.1	1.8	2.3	2.7	2.3	1.7	3.4	2.5	2.2	2.2
90	1.6	1.7	2.1	2.5	2.2	1.7	3.3	2.4	2.2	2.2
85						8.0	0.6	0.7	1.0	1.1
y	2.1	1.6	2.3	2.8	2.3	1.5	3.5	2.5	2.2	2.2

so just like for the implicit method it appears that the boundary curve $y(T)$ starts out from $(0, K)$ with a vertical tangent and with a shape much like a parabola with its apex at $(0, K)$, although this time a slightly different parabola.

For $n = 2$ practical experience shows that it is efficient to exploit the fact that the boundary curve at the beginning looks like a parabola with its apex at $(0, K)$. When $f(x) = \alpha x^2$ then $f(2h)/f(h) = 4$ so that $f(2h) - f(h) = 3f(h)$. This indicates that it might be a good idea to put $k_2^1 = 3k_1$. This is in contrast to the implicit method where the second step turns out to be of the same order of magnitude as the first. Thus it looks like the parabola conjecture fits better to Crank-Nicolson than to the implicit method, at least for the first two steps. The third step with Crank-Nicolson is, however, not large enough to fit the same pattern.

For $n > 2$ we proceed like with the implicit method and put

$$k_n^1 = 2k_{n-1} - k_{n-2}. \tag{16.44}$$

For the second guess we use (16.33) and for subsequent values the secant method (16.35) is used just as with the implicit method producing better and better values for the time step until the tolerance is met.

Table 16.8 illustrates the time step iterations corresponding to the usual set of parameter values $K = 100$, $\sigma = 0.2$, $r = 0.1$, and $L = 200$, and calculated with $h = 1$. Most comments from the previous table can be taken verbatim to this one. The main differences are that k_2 now is close to $3k_1$ whereas k_3 is close to k_2. The initial guess for k_3 therefore overshoots. So much for the parabola hypothesis.

Table 16.10: Error estimate *10 for the implicit method.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	4.0
195	-0.000	-0.001	-0.002	-0.003	-0.002	0.000	0.002	0.008	0.013
190	-0.000	-0.002	-0.005	-0.006	-0.004	0.001	0.005	0.017	0.027
185	-0.000	-0.003	-0.007	-0.009	-0.006	0.002	0.008	0.028	0.044
180	-0.000	-0.005	-0.011	-0.013	-0.008	0.004	0.013	0.041	0.064
175	-0.000	-0.007	-0.014	-0.016	-0.009	0.008	0.019	0.057	0.087
170	-0.001	-0.010	-0.018	-0.018	-0.008	0.015	0.028	0.077	0.115
165	-0.001	-0.014	-0.022	-0.019	-0.005	0.026	0.040	0.102	0.147
160	-0.002	-0.018	-0.025	-0.018	0.001	0.041	0.056	0.132	0.185
155	-0.004	-0.024	-0.027	-0.013	0.012	0.063	0.078	0.170	0.230
150	-0.007	-0.030	-0.027	-0.003	0.029	0.093	0.106	0.215	0.283
145	-0.012	-0.035	-0.021	0.014	0.055	0.134	0.141	0.269	0.343
140	-0.020	-0.037	-0.009	0.041	0.090	0.185	0.184	0.333	0.411
135	-0.031	-0.031	0.015	0.079	0.136	0.249	0.236	0.406	0.486
130	-0.042	-0.011	0.052	0.132	0.195	0.326	0.296	0.486	0.566
125	-0.044	0.028	0.107	0.199	0.264	0.412	0.361	0.571	0.648
120	-0.024	0.094	0.179	0.277	0.341	0.502	0.428	0.654	0.726
115	0.040	0.185	0.261	0.359	0.417	0.588	0.489	0.726	0.790
110	0.152	0.287	0.342	0.432	0.481	0.655	0.535	0.776	0.830
105	0.279	0.370	0.399	0.477	0.514	0.683	0.551	0.785	0.830
100	0.333	0.391	0.403	0.468	0.498	0.651	0.523	0.735	0.770
95	0.242	0.312	0.330	0.386	0.411	0.535	0.432	0.604	0.633
90	0.043	0.134	0.174	0.218	0.242	0.319	0.267	0.376	0.399
85						0.002	0.024	0.040	0.057
y	-0.579	-0.654	-0.642	-0.734	-0.765	-1.003	-0.772	-1.092	-1.127

16.8 Determining the order

As usual we should like to determine the order of the methods and to estimate the error. There is only one independent step size, h, and we can easily perform calculations with h, $2h$, and $4h$, but now we are faced with the same problem as in Chapter 15. The step sizes in the t-direction are determined in the course of the calculations, and there is no way of ensuring that we have comparable grid function values at the same point in time corresponding to two different values of h.

Table 16.11: Order ratios for Crank-Nicolson.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
195	-	4.0	-	3.8	5.1	5.5	7.1	4.2	6.8	3.1
190	-	4.0	-	3.8	5.1	5.5	7.1	4.2	7.0	3.2
185	-	4.0	-0.2	3.8	5.2	5.5	7.1	4.2	7.3	3.4
180	-	4.1	3.5	3.9	5.2	5.5	7.2	4.2	8.0	3.5
175	3.9	4.1	4.4	3.9	5.2	5.5	7.2	4.2	-	3.7
170	3.9	4.2	4.8	3.9	5.2	5.6	7.2	4.2	-	3.8
165	3.9	4.5	5.0	4.0	5.2	5.6	7.3	4.3	-5.7	3.9
160	4.0	-	5.1	4.0	5.2	5.6	7.5	4.4	1.1	3.9
155	4.0	3.5	5.2	4.0	5.2	5.7	7.7	7.0	2.6	4.0
150	4.1	3.8	5.3	4.1	5.2	5.8	8.8	3.8	3.3	4.0
145	4.3	3.9	5.3	4.1	5.2	6.2	1.6	4.0	3.7	4.0
140	5.9	3.9	5.3	4.1	5.2	-	6.0	4.0	3.9	4.1
135	3.5	4.0	5.2	4.2	4.4	5.0	6.5	4.1	4.1	4.1
130	3.7	4.0	5.0	4.2	5.5	5.3	6.7	4.1	4.2	4.1
125	3.8	3.9	4.2	4.1	5.4	5.4	6.8	4.1	4.3	4.1
120	3.8	3.8	-	4.1	5.4	5.5	6.8	4.1	4.3	4.1
115	3.7	2.0	6.6	4.2	5.4	5.5	6.9	4.1	4.4	4.1
110	3.4	4.8	6.2	4.2	5.4	5.6	6.9	4.1	4.4	4.1
105	2.4	4.5	6.0	4.2	5.4	5.6	6.9	4.1	4.4	4.1
100	-	4.4	5.9	4.2	5.3	5.5	6.8	4.1	4.4	4.1
95	-0.2	4.3	5.7	4.1	5.3	5.4	6.7	4.1	4.4	4.1
90	3.8	4.2	4.7	4.1	4.9	5.1	6.2	4.1	4.3	4.1
85					1.5	-	-3.3	3.7	2.2	3.9
y	7.5	4.2	5.3	4.1	5.0	5.2	6.2	4.1	4.4	4.1

The solution is again to interpolate between the time values that we actually compute. In Table 16.9 we give the order ratios and in Table 16.10 the error function multiplied by 10 for a computation with the parameter values $K = 100$,

Table 16.12: Error function *1000 for Crank-Nicolson.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	4.0
195	-	-0.000	-	0.001	0.002	0.003	0.002	0.003	-0.001
190	-	-0.000	-	0.002	0.004	0.006	0.005	0.005	-0.002
185	-	-0.001	0.000	0.003	0.006	0.009	0.008	0.008	-0.004
180	-	-0.001	0.000	0.005	0.009	0.012	0.010	0.010	-0.006
175	-0.000	-0.001	0.001	0.007	0.011	0.015	0.013	0.012	-0.011
170	-0.000	-0.001	0.002	0.009	0.014	0.017	0.014	0.012	-0.017
165	-0.000	-0.001	0.003	0.012	0.017	0.019	0.015	0.011	-0.025
160	-0.000	-0.000	0.005	0.014	0.018	0.020	0.014	0.008	-0.037
155	-0.001	0.001	0.008	0.017	0.019	0.019	0.011	0.001	-0.052
150	-0.001	0.003	0.011	0.019	0.019	0.016	0.006	-0.009	-0.072
145	-0.001	0.007	0.014	0.020	0.016	0.010	-0.002	-0.024	-0.096
140	-0.000	0.011	0.016	0.018	0.011	0.000	-0.014	-0.044	-0.125
135	0.002	0.015	0.016	0.013	0.002	-0.013	-0.029	-0.068	-0.160
130	0.007	0.019	0.014	0.004	-0.011	-0.031	-0.048	-0.097	-0.198
125	0.014	0.019	0.008	-0.010	-0.028	-0.052	-0.069	-0.129	-0.238
120	0.022	0.014	-0.003	-0.028	-0.048	-0.076	-0.092	-0.162	-0.278
115	0.026	0.003	-0.017	-0.049	-0.070	-0.100	-0.115	-0.193	-0.313
110	0.021	-0.013	-0.033	-0.069	-0.089	-0.120	-0.133	-0.217	-0.337
105	0.009	-0.028	-0.046	-0.084	-0.102	-0.132	-0.142	-0.228	-0.343
100	0.001	-0.034	-0.050	-0.086	-0.102	-0.129	-0.138	-0.218	-0.323
95	0.003	-0.027	-0.041	-0.070	-0.085	-0.108	-0.116	-0.181	-0.267
90	0.005	-0.008	-0.021	-0.037	-0.049	-0.064	-0.072	-0.111	-0.168
85						-0.001	-0.009	-0.008	-0.022
y	0.007	0.121	0.159	0.231	0.256	0.302	0.311	0.443	0.598

$\sigma = 0.2$, $r = 0.1$, and $L = 200$, and calculated with $h = 0.25$, 0.5, and 1.0 and using linear interpolation. The last line in each table refers to the boundary curve. We expect the implicit method to be of first order and have therefore elected to use linear interpolation. Furthermore 3-point interpolation produces bad results for large values of t because of the large time steps here. The first order is clearly demonstrated by the order ratios. which vary slowly with x (with occasional deviations which can be explained by small values of the auxiliary function). But the effect of the interpolation error is evident in the t-dependence which displays a somewhat erratic behaviour. As discussed in Chapter 15 and Appendix C we can still have confidence in the error estimates.

In contrast to Brennan-Schwartz the order ratios behave nicely around $x = K$ and the error function does not have an excessive maximum there. The behaviour is much smoother with a 'soft' maximum attained near $x = K$.

Table 16.13: The price function computed with Crank-Nicolson.

x \ t	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	4.0
195	0.000	0.000	0.001	0.002	0.005	0.008	0.011	0.014	0.020
190	0.000	0.000	0.002	0.005	0.011	0.018	0.025	0.031	0.042
185	0.000	0.000	0.003	0.009	0.019	0.030	0.041	0.051	0.069
180	0.000	0.001	0.005	0.015	0.029	0.044	0.060	0.075	0.100
175	0.000	0.001	0.008	0.022	0.041	0.063	0.084	0.104	0.138
170	0.000	0.002	0.012	0.033	0.059	0.087	0.115	0.140	0.183
165	0.000	0.004	0.020	0.048	0.082	0.119	0.154	0.186	0.239
160	0.000	0.007	0.031	0.070	0.115	0.161	0.204	0.243	0.308
155	0.000	0.012	0.049	0.102	0.160	0.217	0.269	0.317	0.394
150	0.001	0.023	0.077	0.148	0.221	0.291	0.355	0.411	0.502
145	0.002	0.041	0.121	0.214	0.306	0.391	0.467	0.533	0.638
140	0.006	0.074	0.188	0.310	0.424	0.526	0.615	0.692	0.812
135	0.016	0.132	0.291	0.447	0.587	0.708	0.811	0.898	1.034
130	0.041	0.230	0.448	0.644	0.812	0.953	1.070	1.169	1.320
125	0.100	0.396	0.684	0.925	1.122	1.283	1.416	1.525	1.690
120	0.229	0.671	1.038	1.325	1.551	1.731	1.877	1.995	2.173
115	0.498	1.116	1.563	1.892	2.143	2.339	2.494	2.620	2.806
110	1.026	1.818	2.332	2.693	2.960	3.164	3.325	3.454	3.642
105	1.992	2.901	3.447	3.818	4.087	4.290	4.449	4.574	4.758
100	3.636	4.528	5.047	5.394	5.644	5.831	5.977	6.091	6.258
95	6.234	6.911	7.320	7.599	7.800	7.952	8.070	8.163	8.299
90	10.068	10.320	10.526	10.681	10.799	10.892	10.965	11.023	11.110
85					15.000	15.006	15.014	15.022	15.038
y	88.528	86.792	85.875	85.294	84.891	84.597	84.374	84.201	83.953

Crank-Nicolson is expected to be of second order and we therefore switch to 3-point interpolation. Table 16.11 gives the order ratios and Table 16.12 gives the error function multiplied by 1000 for a computation with the same parameter values as before but now with step sizes $h = 0.0625$, 0.125, and 0.25. The last line in each table refers to the boundary curve. The general picture is that of a second order method but it is clear that 3-point interpolation is not really good enough for the order ratios. On the positive side we notice that the behaviour on the line $x = K$ is pretty much like in the rest of the region and that the error function is very small, and again with a 'soft' maximum in the neighbourhood of $x = K$. That the singularity at $(0, K)$ has no significant effect on the numbers is explained by the fact that the time steps in the beginning are very small, leading to small values of $b\mu$ and therefore efficient damping of high frequency components by the Crank-Nicolson method.

We conclude this section by giving in Table 16.13 the values of the price function and the boundary curve as calculated by Crank-Nicolson with $h = 0.0625$.

Table 16.14: Average number of iterations per time step,
and total number (N) of time steps.

h	Tolerance										N
	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-11}	10^{-12}	
1/1	3.06	3.19	3.75	4.19	4.56	4.81	5.06	5.19	5.56	5.69	16
1/2	1.94	3.12	3.21	3.58	4.12	4.36	4.58	4.82	5.12	5.27	33
1/4	1.44	2.08	2.79	3.20	3.36	3.62	4.20	4.30	4.36	4.61	66
1/8	1.33	1.50	2.21	2.58	3.16	3.23	3.36	3.69	4.17	4.21	132
1/16	1.26	1.34	1.51	2.18	2.48	3.11	3.15	3.23	3.39	3.94	265
1/32	-	1.27	1.36	1.48	2.14	2.38	3.07	3.09	3.15	3.34	530
1/64	-	1.27	1.30	1.35	1.55	2.12	2.31	3.05	3.07	-	1061
1/128	-	1.30	1.31	1.32	1.36	1.57	2.10	2.24	2.88	-	2122

16.9 Efficiency of the methods

Comparing Brennan-Schwartz to Crank-Nicolson with variable time steps we note that the latter is a second order method with error estimates that can be trusted, also in the interesting region where $x \leq K$. The price we have to pay to achieve this is very small time steps in the beginning and several iterations per time step. To take the last point first we supply in Table 16.14 the average number of iterations per time step as a function of h and the tolerance as well as the total number (N) of time steps for a given value of h. As expected the number of iterations increase with decreasing values of the tolerance but not very much. On the other hand the number of iterations decrease with decreasing values of h. Typically we can expect 3-4 iterations per time step for small h. For comparison we should remember that Brennan-Schwartz computes all the way down to 0 which amounts to 1.7-2 times the number of valid grid points.

When discussing the time step variations it is essential to point out that we use very small time steps for t close to 0 (time close to expiry) where most of the 'action' is, as measured by large values of the time derivatives of u near $x = K$. One property to strive for is a constant value of $k_n \cdot u_t(t_n, K)$ as t_n increases. In Table 16.15 we have given values of $v_m^n - v_m^{n-1}$ which can be taken as approximations of $k_n \cdot u_t(t_n, mh)$ for values of $x = mh$ around K and for all the 16 time steps we take when $h = 1$. It is seen that for constant x, $k_n \cdot u_t(t_n, x)$ displays

Table 16.15: Values of $k_n \cdot u_t$ near $x = K$.

$n \setminus x$	93	94	95	96	97	98	99	100	101	102	103	104
1								0.26	0.03	0.00	0.00	0.00
2							0.16	0.36	0.22	0.08	0.03	0.01
3						0.10	0.24	0.21	0.21	0.14	0.07	0.03
4					0.07	0.17	0.18	0.22	0.19	0.16	0.11	0.07
5				0.05	0.13	0.16	0.22	0.22	0.21	0.18	0.14	0.10
6			0.04	0.11	0.15	0.22	0.24	0.26	0.25	0.23	0.19	0.15
7		0.03	0.10	0.15	0.22	0.25	0.29	0.30	0.30	0.28	0.25	0.22
8	0.03	0.10	0.15	0.22	0.26	0.31	0.33	0.35	0.35	0.34	0.31	0.28
9	0.09	0.15	0.21	0.26	0.31	0.34	0.37	0.38	0.39	0.38	0.36	0.34
10	0.15	0.21	0.26	0.31	0.34	0.38	0.40	0.42	0.42	0.42	0.41	0.39
11	0.21	0.26	0.31	0.35	0.39	0.41	0.43	0.44	0.45	0.45	0.45	0.44
12	0.26	0.31	0.35	0.39	0.42	0.45	0.46	0.48	0.49	0.49	0.49	0.49
13	0.31	0.35	0.40	0.43	0.46	0.48	0.50	0.51	0.53	0.53	0.53	0.53
14	0.36	0.40	0.43	0.47	0.49	0.52	0.54	0.56	0.57	0.58	0.58	0.58
15	0.41	0.44	0.48	0.51	0.54	0.56	0.58	0.60	0.61	0.62	0.63	0.64
16	0.46	0.49	0.52	0.56	0.58	0.61	0.63	0.65	0.66	0.68	0.69	0.70

a slow increase, possibly after some initial fluctuations. With this behaviour in mind we can of course modify Brennan-Schwartz and incorporate varying time steps, possibly making this method more competitive.

Appendix A

The One-Way Wave Equation

In this appendix we shall analyze various difference schemes for the one-way wave equation

$$u_t + au_x \;=\; 0 \tag{A.1}$$

using the tools developed in Chapters 2 and 3.

To ensure uniqueness we must impose an initial condition at $t = 0$, and a boundary condition at one end of the interval in question. Which end depends on the sign of the coefficient a. If a is positive, such that the movement is from left to right, the boundary information should be supplied at the left end. If a is negative the boundary condition should be given at the right end. Some of the numerical schemes below come in pairs, such as A.1 and A.2, or A.6 and A.7, where one can be used for negative a, the other for positive a.

In order to check the accuracy (and other properties) of the difference schemes we note that the symbol of the differential operator in (A.1) is

$$p(s,\xi) \;=\; s + ia\xi. \tag{A.2}$$

A.1 Forward-time forward-space

As suggested by the name this method can be written

$$(\Delta_t + a\Delta)v = \nu$$

or

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_m^n}{h} \;=\; \nu_m^n \tag{A.3}$$

or

$$v_m^{n+1} = v_m^n - a\lambda(v_{m+1}^n - v_m^n) + k\nu_m^n \tag{A.4}$$

using the step ratio

$$\lambda = \frac{k}{h}.$$

The growth factor is

$$\begin{aligned} g(\varphi) &= 1 - a\lambda(e^{i\varphi} - 1) \\ &= 1 + a\lambda(1 - \cos\varphi) - ia\lambda\sin\varphi. \end{aligned} \tag{A.5}$$

The condition for stability is $-1 \leq a\lambda \leq 0$, so this method should only be used with negative a.

The symbol (cf. section 3.3) for the difference operator is

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{k}(e^{sk} - 1) + \frac{a}{h}(e^{i\xi h} - 1) \\ &= s + \frac{1}{2}s^2k + ia\xi - \frac{1}{2}a\xi^2h + \cdots \end{aligned} \tag{A.6}$$

showing that this method is first order accurate in both k and h.

A.2 Forward-time backward-space

This method can be written

$$(\Delta_t + a\nabla)v = \nu$$

or

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = \nu_m^n \tag{A.7}$$

or

$$v_m^{n+1} = v_m^n - a\lambda(v_m^n - v_{m-1}^n) + k\nu_m^n. \tag{A.8}$$

The growth factor is

$$\begin{aligned} g(\varphi) &= 1 - a\lambda(1 - e^{-i\varphi}) \\ &= 1 - a\lambda(1 - \cos\varphi) - ia\lambda\sin\varphi. \end{aligned} \tag{A.9}$$

The condition for stability is $0 \leq a\lambda \leq 1$ and this condition also guarantees a maximum principle.

The symbol for the difference operator is

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{k}(e^{sk}-1) + \frac{a}{h}(1-e^{-i\xi h}) \\ &= s + \frac{1}{2}s^2k + ia\xi + \frac{1}{2}a\xi^2h + \cdots \end{aligned} \tag{A.10}$$

showing that this method is first order accurate in both k and h.
Better accuracy is obtained using

A.3 Forward-time central-space

This method can be written

$$(\Delta_t + a\tilde{\mu}\delta)v = \nu$$

or

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = \nu_m^n \tag{A.11}$$

or

$$v_m^{n+1} = v_m^n - \frac{1}{2}a\lambda(v_{m+1}^n - v_{m-1}^n) + k\nu_m^n. \tag{A.12}$$

The growth factor is

$$\begin{aligned} g(\varphi) &= 1 - \frac{1}{2}a\lambda(e^{i\varphi} - e^{-i\varphi}) \\ &= 1 - ia\lambda\sin\varphi. \end{aligned} \tag{A.13}$$

The method is 0-stable if $k = O(h^2)$ but it is never absolutely stable. We shall therefore disregard it in favour of

A.4 Central-time central-space or leap-frog

This method can be written

$$(\tilde{\mu}_t\delta_t + a\tilde{\mu}\delta)v = \nu$$

or

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = \nu_m^n \tag{A.14}$$

or

$$v_m^{n+1} = v_m^{n-1} - a\lambda(v_{m+1}^n - v_{m-1}^n) + 2k\nu_m^n. \tag{A.15}$$

For the growth factor we have

$$g^2 + 2ia\lambda \sin\varphi \cdot g - 1 = 0 \tag{A.16}$$

and the condition for stability is $|a\lambda| \leq 1$.

The symbol for the difference operator is

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{2k}(e^{sk} - e^{-sk}) + \frac{a}{2h}(e^{i\xi h} - e^{-i\xi h}) \\ &= s + ia\xi + \frac{1}{6}s^3k^2 - \frac{1}{6}ia\xi^3h^2 + \cdots \end{aligned} \tag{A.17}$$

showing that this method is second order accurate in both k and h. The main drawback is that it is a two-step method. The following modification results in a similar one-step method:

A.5 Lax-Friedrichs

This method can be written

$$\frac{v_m^{n+1} - \frac{1}{2}(v_{m+1}^n + v_{m-1}^n)}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = \nu_m^n \tag{A.18}$$

or

$$v_m^{n+1} = \frac{1}{2}(v_{m+1}^n + v_{m-1}^n) - \frac{1}{2}a\lambda(v_{m+1}^n - v_{m-1}^n) + k\nu_m^n. \tag{A.19}$$

For the growth factor we have

$$g(\varphi) = \cos\varphi - ia\lambda\sin\varphi \tag{A.20}$$

and the condition for stability is again $|a\lambda| \leq 1$ and this condition also guarantees a maximum principle.

The symbol for the difference operator is

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{k}(e^{sk} - \cos\xi h) + \frac{a}{2h}(e^{i\xi h} - e^{-i\xi h}) \\ &= s + ia\xi + \frac{1}{2}s^2k - \frac{1}{6}ia\xi^3h^2 + \frac{1}{2}\xi^2\frac{h^2}{k} + \cdots \end{aligned} \tag{A.21}$$

If $\mu = k/h^2$ is constant then Lax-Friedrichs is not consistent with (A.1), but if $\lambda = k/h$ is constant which is more customary for this problem and in line with the stability condition above, then the method is first order accurate.

A.6 Backward-time forward-space

This method can be written

$$\frac{v_m^n - v_m^{n-1}}{k} + a\frac{v_{m+1}^n - v_m^n}{h} = \nu_m^n \tag{A.22}$$

or

$$v_m^n(1 - a\lambda) = v_m^{n-1} - a\lambda v_{m+1}^n + k\nu_m^n. \tag{A.23}$$

The growth factor is

$$g(\varphi) = \frac{1}{1 - a\lambda(1 - e^{i\varphi})} \tag{A.24}$$

showing that this method is stable for $a\lambda \leq 0$ and thus can be used when a is negative.

The method is implicit but since a is negative we have a boundary condition on the right boundary and we can arrange the calculations in an explicit manner as shown in (A.23), since we always know the value to the right at the advanced time step (which is here step n):

Remark. We also have $|g(\varphi)| \leq 1$ when $a\lambda \geq 1$ but this is less useful. □

The symbol for the difference operator is

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{k}(1 - e^{-sk}) + \frac{a}{h}(e^{i\xi h} - 1) \\ &= s - \frac{1}{2}s^2k + ia\xi - \frac{1}{2}a\xi^2 h + \cdots \end{aligned} \tag{A.25}$$

showing that this method is first order accurate in both k and h.

A.7 Backward-time backward-space

This method can be written

$$\frac{v_m^n - v_m^{n-1}}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = \nu_m^n \tag{A.26}$$

or

$$v_m^n(1 + a\lambda) = v_m^{n-1} + a\lambda v_{m-1}^n + k\nu_m^n. \tag{A.27}$$

The growth factor is

$$g(\varphi) \;=\; \frac{1}{1 + a\lambda(1 - e^{-i\varphi})} \tag{A.28}$$

showing that this method is stable for $a\lambda > 0$ and this condition also guarantees a maximum principle.

This method is also implicit but again we can arrange the calculations in an explicit manner (as shown) since we always know the value to the left at the advanced time step.

The symbol for the difference operator is

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{k}(1 - e^{-sk}) + \frac{a}{h}(1 - e^{-i\xi h}) \\ &= s - \frac{1}{2}s^2k + ia\xi + \frac{1}{2}a\xi^2h + \cdots \end{aligned} \tag{A.29}$$

showing that this method is first order accurate in both k and h.

A.8 Backward-time central-space

This method can be written

$$\frac{v_m^n - v_m^{n-1}}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} \;=\; \nu_m^n. \tag{A.30}$$

This method is truly implicit and we must solve the tridiagonal system of equations

$$-\frac{1}{2}a\lambda v_{m-1}^n + v_m^n + \frac{1}{2}a\lambda v_{m+1}^n \;=\; v_m^{n-1} + k\nu_m^n. \tag{A.31}$$

This may present some problems since we usually only have one boundary condition in the one-way wave equation.

The growth factor is

$$g(\varphi) \;=\; \frac{1}{1 + ia\lambda\sin\varphi} \tag{A.32}$$

showing that this method is unconditionally stable.

The symbol for the difference operator is

$$p_{k,h}(s,\xi) \;=\; s - \frac{1}{2}s^2k + ia\xi - \frac{1}{6}ia\xi^3h^2 + \cdots \tag{A.33}$$

showing that this method is first order accurate in k and second order accurate in h.

A.9 Lax-Wendroff

This method can be viewed as a modification of forward-time central-space

$$\begin{aligned} \frac{v_m^{n+1}-v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} - \frac{a^2k}{2}\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} & \\ = \frac{1}{2}(\nu_m^{n+1}+\nu_m^n) - \frac{ak}{4h}(\nu_{m+1}^n - \nu_{m-1}^n) & \end{aligned} \tag{A.34}$$

or

$$\begin{aligned} v_m^{n+1} &= v_m^n - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n) + \frac{1}{2}a^2\lambda^2(v_{m+1}^n - 2v_m^n + v_{m-1}^n) \\ &\quad + \frac{k}{2}(\nu_m^{n+1}+\nu_m^n) - \frac{ak^2}{4h}(\nu_{m+1}^n - \nu_{m-1}^n). \end{aligned} \tag{A.35}$$

The growth factor is

$$g(\varphi) = 1 - ia\lambda\sin\varphi - 2a^2\lambda^2\sin^2\frac{\varphi}{2}. \tag{A.36}$$

and the method is therefore stable if $|a\lambda| \le 1$, but it does not satisfy a maximum principle.

The symbol for the difference operator is

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{k}(e^{sk}-1) + ia\frac{\sin\xi h}{h} + 2a^2\frac{k}{h^2}\sin^2\frac{\xi h}{2} \\ &= s + ia\xi + \frac{1}{2}s^2k + \frac{1}{6}s^3k^2 - \frac{1}{6}ia\xi^3h^2 + \frac{1}{2}a^2\xi^2k + \cdots \end{aligned} \tag{A.37}$$

and for the right-hand-side operator

$$\begin{aligned} r_{k,h}(s,\xi) &= \frac{1}{2}(e^{sk}+1) - \frac{1}{2}iak\frac{\sin\xi h}{h} \\ &= 1 + \frac{1}{2}sk + \frac{1}{4}s^2k^2 - \frac{1}{2}ia\xi k + \cdots \end{aligned} \tag{A.38}$$

such that

$$p_{k,h} - r_{k,h}p = -\frac{1}{12}s^3k^2 - \frac{1}{4}ia\xi s^2k^2 - \frac{1}{6}ia\xi^3h^2 + \cdots \tag{A.39}$$

showing that this method is second order accurate in both k and h.

A.10 Crank-Nicolson

This method can be written

$$\frac{v_m^{n+1}-v_m^n}{k} + a\frac{v_{m+1}^{n+1} - v_{m-1}^{n+1} + v_{m+1}^n - v_{m-1}^n}{4h} = \frac{\nu_m^{n+1}+\nu_m^n}{2}. \tag{A.40}$$

Crank-Nicolson is also implicit and the tridiagonal system of equations now looks like

$$\begin{aligned} -\frac{1}{4}a\lambda v_{m-1}^{n+1} + v_m^{n+1} + \frac{1}{4}a\lambda v_{m+1}^{n+1} & \qquad\qquad \text{(A.41)} \\ = \frac{1}{4}a\lambda v_{m-1}^{n} + v_m^{n} - \frac{1}{4}a\lambda v_{m+1}^{n} + k\frac{\nu_m^{n+1}+\nu_m^n}{2}. & \end{aligned}$$

The growth factor is

$$g(\varphi) = \frac{1-\frac{1}{2}ia\lambda\sin\varphi}{1+\frac{1}{2}ia\lambda\sin\varphi} \qquad \text{(A.42)}$$

showing that the method is unconditionally stable.

The symbol for the difference operator is

$$p_{k,h}(s,\xi) = \frac{1}{k}(e^{sk}-1) + ia\frac{e^{sk}+1}{2}\frac{\sin\xi h}{h} \qquad \text{(A.43)}$$

and for the right-hand-side operator

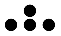
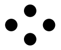
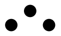
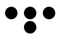
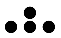
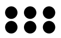
$$r_{k,h}(s,\xi) = \frac{1}{2}(e^{sk}+1) \qquad \text{(A.44)}$$

such that

$$p_{k,h} - r_{k,h}p = -\frac{1}{12}s^3k^2 - \frac{1}{6}ia\xi^3h^2 + \cdots \qquad \text{(A.45)}$$

showing that this method is second order accurate in both k and h.

A.11 Overview of the methods

Method		Stencil	Order	Stability	Comments
1			1,1	$-1 \le a\lambda \le 0$	$a < 0$
2			1,1	$0 \le a\lambda \le 1$	$a > 0$
3			1,2	$-$	
4			2,2	$-1 \le a\lambda \le 1$	2-step
5	LF		1,1	$-1 \le a\lambda \le 1$	
6			1,1	$a\lambda \le 0$	$a < 0$
7			1,1	$a\lambda \ge 0$	$a > 0$
8			1,2	$+$	bdry.
9	LW		2,2	$-1 \le a\lambda \le 1$	–max
10	CN		2,2	$+$	bdry.

Method A.3 is never stable and is disregarded in the following. Method A.4 is a two-step method and methods A.8 and A.10 require an extra boundary condition in order to solve the system of linear equations. Of the remaining methods, only A.9 is second order, but it does not satisfy a maximum principle. If we use a first order method together with Richardson extrapolation we can often get second order results anyway.

Remark. A stability condition of the form $|a\lambda| \leq 1$ requires $k \leq |a|h$ but this is not unreasonable for a method which has the same order in t and x, where we for reasons of accuracy probably should choose k proportional to h anyway. □

It should be mentioned here that if a is constant and if we choose $k = ah$, i.e. $a\lambda = 1$ then methods A.2, A.4, A.5, A.6, and A.9 are all exact in the sense that they all reduce to $v_m^{n+1} = v_{m-1}^n$ and thus reproduce the transport with no error. In this special case the methods do not exhibit their more general behaviour which usually includes a certain amount of dissipation.

We have compared the various methods on the two test examples from section 5.7, now with $b = 0$ and $a = 9$. It is again easy to show that all methods considered will conserve the area. The true solution is just a translation of the initial function with velocity a. The numerical solution will exhibit various degrees of dissipation, which means that certain components will be damped ($|g(\varphi)| < 1$), and dispersion, which means that some components will travel with a velocity different from a. As a result sharp corners will be rounded, the maximum will be lowered, and the half-width will be wider.

Since $a > 0$ we do not test methods A.1 and A.6. Method A.3 is disregarded because of instability and method A.4 because it is a 2-step method.

For the remaining six methods we have checked the height of the maximum, the position of the maximum (using 3-point interpolation around the maximum v-value) and the half-width of the bump. We have used $h = 0.02$ together with $k = 0.002$, and $h = 0.1$ together with $k = 0.01$ and $k = 0.05$. The two former combinations give $a\lambda = 0.9$, the latter $a\lambda = 0.45$. The general conclusion from this limited test is that method A.7 always shows a large amount of dissipation/dispersion (low max, large half-width), closely followed by A.8. Methods A.5 and A.2 perform reasonably well for $a\lambda = 0.9$ but not for the smaller value. Methods A.9 and A.10 always perform well except that A.10 sometimes has problems with the right speed. The tests also show that methods A.8, A.9, and A.10 produce 'waves' with occasional negative function values upstream.

Appendix B

A Class of Test Problems

When testing algorithms it is useful to have a number of test problems for which we know the true solution. Two such sets are introduced here for the two-dimensional equation

$$u_t = b_1 u_{xx} + 2b_{12} u_{xy} + b_2 u_{yy}. \tag{B.1}$$

This equation has solutions of the form

$$u(t, x, y) = e^{\alpha t} \sin(\beta x - \gamma y) \cosh(\delta x + \epsilon y) \tag{B.2}$$

provided the constants α, β, γ, δ, and ϵ satisfy some simple relations. Differentiating (B.2) we get

$$\begin{aligned}
u_{xx} &= (\delta^2 - \beta^2)u + 2\beta\delta e^{\alpha t} \cos(\beta x - \gamma y) \sinh(\delta x + \epsilon y), && \text{(B.3)}\\
u_{yy} &= (\epsilon^2 - \gamma^2)u - 2\gamma\epsilon e^{\alpha t} \cos(\beta x - \gamma y) \sinh(\delta x + \epsilon y), && \text{(B.4)}\\
u_{xy} &= (\beta\gamma + \delta\epsilon)u + (\beta\epsilon - \delta\gamma)e^{\alpha t} \cos(\beta x - \gamma y) \sinh(\delta x + \epsilon y). && \text{(B.5)}
\end{aligned}$$

If we choose β, γ, δ, and ϵ such that

$$b_1\beta\delta - b_2\gamma\epsilon + b_{12}(\beta\epsilon - \delta\gamma) = 0 \tag{B.6}$$

then the last terms on the right-hand-side of (B.3) - (B.5) will cancel, and if we furthermore define

$$\alpha = b_1(\delta^2 - \beta^2) + b_2(\epsilon^2 - \gamma^2) + 2b_{12}(\beta\gamma + \delta\epsilon) \tag{B.7}$$

then (B.2) is a solution to (B.1).

We should avoid combinations with $\beta\gamma + \delta\epsilon = 0$ and $\beta\epsilon - \delta\gamma = 0$ because this leads to $u_{xy} = 0$ and we shall not see the effect of the mixed term.

If we try solutions of the form

$$u(t,x,y) \;=\; e^{\alpha t}\sin(\beta x-\gamma y)\cos(\delta x+\epsilon y) \tag{B.8}$$

then we get

$$\begin{aligned} u_{xx} &= -(\delta^2+\beta^2)u-2\beta\delta e^{\alpha t}\cos(\beta x-\gamma y)\sin(\delta x+\epsilon y), && \text{(B.9)}\\ u_{yy} &= -(\epsilon^2+\gamma^2)u+2\gamma\epsilon e^{\alpha t}\cos(\beta x-\gamma y)\sin(\delta x+\epsilon y), && \text{(B.10)}\\ u_{xy} &= (\beta\gamma-\delta\epsilon)u-(\beta\epsilon-\delta\gamma)e^{\alpha t}\cos(\beta x-\gamma y)\sin(\delta x+\epsilon y). && \text{(B.11)}\end{aligned}$$

If (B.6) is satisfied then (B.8) is a solution to (B.1) when

$$\alpha \;=\; -b_1(\delta^2+\beta^2)-b_2(\epsilon^2+\gamma^2)+2b_{12}(\beta\gamma-\delta\epsilon). \tag{B.12}$$

Here we should avoid $\beta\gamma-\delta\epsilon=0$ and $\beta\epsilon-\delta\gamma=0$ if we want to see the effect of the mixed term.

Appendix C

Interpolation and the Order Ratio

C.1 Introduction

When we use variable step sizes in the t-direction as we do in the moving boundary problems of Chapters 15 and 16 it becomes necessary to interpolate to get values of the computed solution function at specified time values. When choosing an interpolation formula we should have in mind that the interpolation error does not interfere too much with the order ratios and the error estimation. If we use the implicit method we expect to have first order results and a linear interpolation should suffice, but as we shall see in the following section this might not be quite enough.

C.2 Linear interpolation

If we want a function value $v(t)$ at a specific value t, and $t_n < t < t_{n+1}$ then we shall interpolate between t_n and t_{n+1} and the result is a function value

$$\begin{aligned}
w(t) &= \frac{t_{n+1}-t}{t_{n+1}-t_n}v(t_n) + \frac{t-t_n}{t_{n+1}-t_n}v(t_{n+1}) \\
&= \frac{t_{n+1}-t}{t_{n+1}-t_n}v(t_n) + \frac{t-t_n}{t_{n+1}-t_n}(v(t_n) + (t_{n+1}-t_n)v' + \frac{1}{2}(t_{n+1}-t_n)^2v'' + \cdots) \\
&= v(t_n) + (t-t_n)v' + \frac{1}{2}(t_{n+1}-t_n)(t-t_n)v'' + \cdots \\
&= v(t) + \frac{1}{2}(t-t_n)(t_{n+1}-t)v'' + \cdots
\end{aligned}$$

$$= \quad v(t) + \frac{1}{2}\alpha(1-\alpha)k_{n+1}^2 v'' + \cdots \tag{C.1}$$

with

$$k_{n+1} \;=\; t_{n+1} - t_n \quad \text{and} \quad \alpha \;=\; \frac{t - t_n}{t_{n+1} - t_n} \tag{C.2}$$

Since $k_n = \kappa h + \cdots$ we see that the difference between $w(t)$ and $v(t)$ is of second order, and linear interpolation is accurate enough because the interpolation error is of higher order than the truncation error.

Table C.1: Order ratios for $y(t)$ and $u(t,x)$ using linear interpolation.

	$y(t)$	$u(t,x)$							
$t \setminus x$		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	2.174	2.287							
0.2	2.131	2.193	2.192						
0.3	2.048	2.096	2.077	2.062					
0.4	1.862	1.917	1.870	1.825	1.787				
0.5	1.991	2.032	2.016	2.001	1.989	1.982			
0.6	2.013	2.046	2.038	2.030	2.023	2.018	2.016		
0.7	2.072	2.094	2.094	2.095	2.095	2.096	2.096		
0.8	1.921	1.957	1.941	1.927	1.914	1.903	1.897	1.897	
0.9	2.011	2.035	2.030	2.026	2.021	2.018	2.015	2.014	2.014
1.0	2.066	2.082	2.083	2.084	2.084	2.085	2.085	2.084	2.083

But there is another consideration. The truncation error and the various terms in the error expansion (15.20) tend to vary smoothly with the independent variables. The interpolation error on the other hand depends on the position of the interpolation point relative to the calculated t_n-values and will therefore exhibit an erratic behaviour as t varies. The effect is seen most clearly in the order ratios. Since they are ratios of differences of nearly equal quantities they are especially sensitive to erratic changes. In Table C.1 we show the order ratios for $y(t)$ (the second column) and $u(t,x)$ (the triangular scheme) when linear interpolation has been used in connection with numerical solution of the Stefan problem (cf. section 15.4). There is no doubt about the first order behaviour but we notice that as t varies the ratio is sometimes bigger, sometimes smaller than 2. The variation in x for a fixed value of t is much smoother because we use the same interpolation coefficients for all x.

In our basic assumptions (15.20) and (15.21) about the truncation error it is understood that the auxiliary functions depend only on t and x but not on the step size. The interpolation error, however, depends on the step size, not only through the k^2-factor in (C.1) but implicitly through α which most likely takes on different values as h is varied. We use calculations with three different values of h in order to determine the order and it is necessary to keep the interpolation error small relative to the relevant components of the truncation error in order to get a reliable error determination.

Table C.2: Error estimates*1000 for $y(t)$ and $u(t, x)$ using linear interpolation.

	$y(t)$	$u(t,x)$							
$t \setminus x$		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	-1.048	-1.946							
0.2	-1.706	-2.917	-2.007						
0.3	-2.235	-3.606	-2.770	-2.114					
0.4	-2.834	-4.343	-3.565	-2.923	-2.418				
0.5	-3.154	-4.662	-3.931	-3.316	-2.815	-2.425			
0.6	-3.592	-5.128	-4.436	-3.839	-3.337	-2.928	-2.608		
0.7	-3.827	-5.323	-4.665	-4.091	-3.600	-3.191	-2.862		
0.8	-4.240	-5.734	-5.105	-4.545	-4.057	-3.640	-3.291	-3.007	
0.9	-4.421	-5.861	-5.258	-4.718	-4.242	-3.830	-3.480	-3.191	-2.957
1.0	-4.632	-6.025	-5.444	-4.920	-4.454	-4.046	-3.695	-3.399	-3.154

The interpolation error will on the average increase by a factor 4 when we double the step size, although the actual value also depends on α. Therefore the interpolation error will tend to be largest when the step size is $4h$ (and it increases with a higher rate than the truncation error). Since results with all three step sizes enter in the calculation of the order ratio, this result is very sensitive. The difference between function values corresponding to h and $2h$ which we use as error estimate or correction term in the extrapolation is less sensitive as demonstrated in Table C.2.

C.3 Three-point interpolation

We now propose a more accurate interpolation formula. For a given value t we use three consecutive time-values t_{n-1}, t_n, and t_{n+1} with t_n being the closer to t:

$$|t - t_n| \; < \; \min\{t - t_{n-1}, t_{n+1} - t\}.$$

Repeating the calculations of the previous section we arrive at

$$w(t) \;\; = \;\; v(t_n) + (t - t_n)v' + \frac{1}{2}(t - t_n)^2 v'' + O(k_n^3)v''' \;\; = \;\; v(t) + O(h^3).$$

Table C.3: Order ratios for $y(t)$ and $u(t,x)$ using 3-point interpolation.

	$y(t)$	$u(t,x)$							
$t \setminus x$		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	2.103	2.217							
0.2	2.076	2.144	2.124						
0.3	2.021	2.062	2.038	2.019					
0.4	2.028	2.064	2.050	2.038	2.031				
0.5	2.073	2.112	2.112	2.111	2.110	2.046			
0.6	2.047	2.077	2.073	2.070	2.067	2.065	2.051		
0.7	2.039	2.065	2.062	2.058	2.056	2.053	2.052		
0.8	2.028	2.051	2.047	2.043	2.040	2.038	2.037	2.037	
0.9	2.038	2.059	2.057	2.055	2.053	2.052	2.050	2.049	2.039
1.0	2.033	2.053	2.051	2.049	2.047	2.045	2.044	2.043	2.042

Table C.4: Error estimates*1000 for $y(t)$ and $u(t,x)$ using 3-point interpolation.

	$y(t)$	$u(t,x)$							
$t \setminus x$		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	-1.062	-1.973							
0.2	-1.801	-3.067	-2.156						
0.3	-2.410	-3.871	-3.034	-2.367					
0.4	-2.878	-4.405	-3.626	-2.982	-2.473				
0.5	-3.277	-4.824	-4.093	-3.473	-2.965	-2.564			
0.6	-3.642	-5.191	-4.500	-3.901	-3.396	-2.983	-2.653		
0.7	-3.949	-5.473	-4.815	-4.238	-3.741	-3.325	-2.986		
0.8	-4.237	-5.731	-5.102	-4.543	-4.055	-3.638	-3.289	-3.005	
0.9	-4.497	-5.952	-5.348	-4.806	-4.327	-3.912	-3.557	-3.262	-3.016
1.0	-4.726	-6.132	-5.550	-5.025	-4.556	-4.144	-3.788	-3.485	-3.234

The interpolation will still introduce some disturbance but smaller and in a higher order term and therefore hardly visible. The order ratios of Table C.3 exhibit a much smoother variation than those of Table C.1 although the erratic components in the t-direction can be seen faintly. The error estimates given in Table C.4 agree reasonably well with those from linear interpolation indicating that these were quite adequate.

Bibliography

[1] A. C. Aitken, *On Bernoulli's numerical solution of algebraic equations*, Proc. Roy. Soc. Edinburgh, 46 (1926), pp. 289–305.

[2] V. A. Barker, *Extrapolation*, Hæfte 31, Numerisk Institut, DtH, Lyngby, 1974.

[3] M. J. Brennan and E. S. Schwartz, *A continuous time approach to the pricing of bonds,* Journal of Banking and Finance, 3 (1979), pp. 133–155.

[4] M. J. Brennan and E. S. Schwartz, *The valuation of American put options,* Journal of Finance, 32 (1977), pp. 449–462.

[5] G. G. O'Brien, M. A. Hyman, and S. Kaplan, *A Study of the Numerical Solution of Partial Differential Equations*, J. Math. Phys., 29 (1951), pp. 223–251.

[6] D. Britz and O. Østerby, *Some numerical investigations of the stability of electrochemical digital simulation, particularly as affected by first-order homogeneous reactions*, J. Electroanal. Chem., 368 (1994), pp. 143–147.

[7] E. T. Copson and P. Keast, *On a boundary-value problem for the equation of heat*, J. Inst. Maths. Applics, 2 (1966), pp. 358–363.

[8] J. Crank and P. Nicolson, *A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type*, Proc. Cambridge Philos. Soc., 43 (1947), pp. 50–67. Reprinted in Adv. Comput. Math., 6 (1996), pp. 207–226.

[9] J. Douglas and T. M. Gallie, *On the numerical integration of a parabolic differential equation subject to a moving boundary condition*, Duke Math. J., 22 (1955), pp. 557-571.

[10] J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.

[11] E. C. Du Fort and S. P. Frankel, *Stability Conditions in the Numerical Treatment of Parabolic Differential Equations*, Math. Tables Aid Comput., 7, (1953), pp. 135–152.

[12] A. R. Gourlay and J. Ll. Morris, *The extrapolation of first order methods for parabolic partial differential equations II*, SIAM J. Numer. Anal., 17, (1980), pp. 641–655.

[13] P. M. Gresho and R. L. Lee, *Don't suppress the wiggles – they're telling you something*, Computers and Fluids, 9 (1981), pp. 223–253.

[14] Asbjørn Trolle Hansen, *Martingale Methods in Contingent Claim Pricing and Asymmetric Financial Markets*, Ph.D. thesis, Dept. Oper. Research, Aarhus University, 1998.

[15] Asbjørn Trolle Hansen and Ole Østerby, *Accelerating the Crank-Nicolson method in American option pricing*, Dept. Oper. Research, Aarhus University, 1998.

[16] D. R. Hartree and J. R. Womersley, *A Method for the Numerical or Mechanical Solution of Certain Types of Partial Differential Equations*, Proc. Royal Soc. London, Ser. A, 161 (1937), pp. 353–366.

[17] D. C. Joyce, *Survey of extrapolation processes in numerical analysis*, SIAM Review, 13 (1971), pp. 435–490.

[18] P. Keast and A. R. Mitchell, *Finite difference solution of the third boundary problem in elliptic and parabolic equations*, Numer. Math., 10, (1967), pp. 67–75.

[19] P. Laasonen, *Über eine Methode zur Lösung der Wärmeleitungsgleichung*, Acta Math., 81 (1949), pp. 309–317.

[20] J. D. Lawson and J. Ll. Morris, *The extrapolation of first order methods for parabolic partial differential equations I*, SIAM J. Numer. Anal., 15, (1978), pp. 1212–1224.

[21] P. D. Lax and R. D. Richtmyer, *Survey of the Stability of Linear Finite Difference Equations*, Comm. Pure Appl. Math., 9, (1956), pp. 267–293.

[22] P. D. Lax and B. Wendroff, *Systems of conservation laws*, Comm. Pure Appl. Math., 13, (1960), pp. 217–237.

[23] B. Lindberg, *On smoothing and extrapolation for the trapezoidal rule*, BIT, 11 (1971), pp. 29–52.

[24] F. A. Longstaff and E. S. Schwartz, *Interest rate volatility and the term structure: A two-factor general equilibrium model*,
J. Finance, 47 (1992), pp. 1259–1282.
See also Journal of Fixed Income (1993), pp. 7–14.

[25] A. R. Mitchell and D. F. Griffiths, *The Finite Difference Method in Partial Differential Equations*, John Wiley, Chichester, 1980.

[26] P. L. J. van Moerbeke, *On optimal stopping and free boundary problems*, Arch. Rational Mech. Anal., 60 (1976), pp. 101–148.

[27] D. W. Peaceman and H. H. Rachford, *The numerical solution of parabolic and elliptic differential equations*, J. SIAM, 3 (1955), pp. 28–41.

[28] C. E. Pearson, *Impulsive end condition for diffusion equation*,
Math. Comp., 19 (1965), pp. 570–576.

[29] Lewis F. Richardson, *The Approximate Numerical Solution by Finite Differences of Physical Problems Involving Differential Equations*,
Phil. Trans. Roy. Soc. London, Series A, 210 (1910), pp. 307–357.

[30] Lewis F. Richardson and J. Arthur Gaunt,
The Deferred Approach to the Limit I - II,
Phil. Trans. Roy. Soc. London, Series A, 226 (1927), pp. 299–361.

[31] Robert D. Richtmyer, *Difference Methods for Initial-Value Problems*,
Interscience, New York, 1957.

[32] Werner Romberg, *Vereinfachte Numerische Integration*,
Norske Vid. Selsk. Forh., Trondheim, 28 (1955), pp. 30–36.

[33] J. Stefan, *Über die Theorie der Eisbildung, insbesondere über die Eisbildung im Polarmeere*, Akad. Wiss. Wien, Mat. Nat. Classe, Sitzungsberichte, 98 (1889), pp. 965–983.

[34] G. W. Stewart, *Introduction to Matrix Computations*,
Academic Press, New York, 1973.

[35] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989.

[36] Øystein Tødenes, *On the numerical solution of the diffusion equation*,
Math. Comp., 24, (1970), pp. 621–627.

[37] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*,
HMSO, London, 1963.

[38] W. L. Wood and R. W. Lewis, *A comparison of time marching schemes for the transient heat conduction equation*,
Int. J. Num. Meth. Engrg., 9, (1975), 679–689.

[39] Y. G. D'Yakonov, *On the application of disintegrating difference operators*,
USSR Comp. Math., 3 (1963), 511–515.
See also vol. 2 (1962), pp. 55–77 and pp. 581–607.

[40] O. Østerby, *Stability of finite difference formulas for linear parabolic equations*, 2nd Int. Coll. on Numerical Analysis, Plovdiv, D. Bainov and V. Covachev (eds.), VSP Press, Utrecht (1994), pp. 165–176.

[41] O. Østerby, *Five ways of reducing the Crank-Nicolson oscillations*,
BIT, 43 (2003), pp. 811–822.