38 Springer Series in Computational Mathematics

High Order Difference Methods for Time Dependent PDE

Bertil Gustafsson

Springer

Springer Series in
Computational
Mathematics

38

Bertil Gustafsson

High Order Difference Methods for Time Dependent PDE

With 94 Figures and 12 Tables

Springer

Bertil Gustafsson
Ledungsvägen 28
75440 Uppsala
Sweden
bertil@stanford.edu

ISBN 978-3-540-74992-9 e-ISBN 978-3-540-74993-6

DOI 10.1007/978-3-540-74993-6

Springer Series in Computational Mathematics ISSN 0179-3632

Library of Congress Control Number: 2007940500

Mathematics Subject Classification (2000): 65M06

Cover design: WMX Design GmbH, Heidelberg
Typesetting: by the author using a Springer LaTeX macro package
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

Many books have been written on finite difference methods (FDM), but there are good reasons to write still another one. The main reason is that even if higher order methods have been known for a long time, the analysis of stability, accuracy and effectiveness is missing to a large extent. For example, the definition of the formal high order accuracy is based on the assumption that the true solution is smooth, or expressed differently, that the grid is fine enough such that all variations in the solution are well resolved. In many applications, this assumption is not fulfilled, and then it is interesting to know if a high order method is still effective. Another problem that needs thorough analysis is the construction of boundary conditions such that both accuracy and stability is upheld. And finally, there has been quite a strong development during the last years, in particular when it comes to very general and stable difference operators for application on initial–boundary value problems.

The content of the book is not purely theoretical, neither is it a set of recipes for various types of applications. The idea is to give an overview of the basic theory and construction principles for difference methods without going into all details. For example, certain theorems are presented, but the proofs are in most cases left out. The explanation and application of the theory is illustrated by using simple model examples. Among engineers, one is often talking about "toy problems" with a certain scepticism, under the assumption that the results have no significance for real world problems. When looking at the scientific production over the years, there is a certain truth to this claim. A method may be working very well, and better than other well known methods, for a certain model problem in one space dimension, but it is not even clear how the method can be generalized to several space dimensions. In this book we try to avoid falling into this trap. The generalization should of course always be possible in the sense that an algorithm based on the same principle can be constructed for the full problem, and hopefully the essential properties of the numerical method have been caught by the analysis of the simpler problem. Sometimes, the theoretical considerations carry over without difficulty, but sometimes we have to rely upon intuition and numerical experiments to be confident about the performance on the large problem. On the other hand, there are many cases where the model problem tells it all. For example, for a hyperbolic system of PDE in one

space dimension, one can transform the system to a diagonal one and limit the analysis to a scalar PDE with constant coefficients. By applying the transformation back, and generalizing to variable coefficients, the results hold for the original problem as well.

When discussing stability and accuracy, the model examples in this book are often of low order accuracy, since they have a simpler structure. The goal is simply to illustrate how the theory works, and it is easier to see the basic mechanisms for simpler schemes. Once the application of the theory is well understood, it should be clear how to apply it for more complicated and powerful high order methods. However, there are two main application areas where we choose to go into more detail about the full implementation of the suggested methods. One application is wave propagation with relevance for acoustics and electromagnetics. In this case the most interesting problems are so large, that there is no realistic low order computational alternative. The other one is incompressible flow governed by the Navier-Stokes equations. Even if low order methods have been used for many applications, the real challenge is turbulent flow, where there is no realistic alternative to high order methods if the smallest scales are to be represented well. In both cases, we present a fairly detailed description of the methods. One good reason for this is to illustrate how the analysis and construction described in earlier chapters is carried out for a more technically complicated problem.

There is a pervading theme in the book, and that is compactness of the computational stencils. A comparison between two approximations of the same order always comes out in favor of the more compact one, i.e., as few grid points as possible should be involved, both in space and time. The smaller error constants often make quite a dramatic change. Padé approximations and staggered grids are examples of this, as well as the box scheme described in 8.

The outline of the book is as follows.

In Chapter 1 there is an analysis of the effectiveness of higher order methods. It is based on Fourier analysis, and the necessary number of points per wave length is estimated for different types of PDE and different orders of accuracy.

Chapter 2 contains a survey of the theory for well-posedness and stability, and the different tools for analysis are described. These tools are based on Fourier analysis for problems with periodic solutions and the energy method and Laplace transform method for initial–boundary value problems. Different kinds of stability definitions are necessary in this case, and we discuss the implications of each one.

In Chapter 3 we discuss how the order of accuracy is connected to the convergence rate, i.e., how fast the numerical solution approaches the true solution when the grid is refined. This is a straightforward analysis for periodic problems, but less obvious when boundaries are involved.

Chapters 2 and 3 can be read independently of the rest of the book as an introduction and a survey of the stability theory for difference approximations in general.

The next three chapters contain a systematic presentation of the different ways of constructing high order approximations. There are the standard centered difference operators, but also Padé type operators as well as schemes based on staggered grids. When constructing the difference schemes, one principle is to approximate the dif-

ferential equation in space first, and then approximate the resulting ODE system in time. But there are other ways to obtain effective difference schemes where the space and time discretizations are not separated, and this is described in Chapter 6.

In Chapter 7 we bring up the special problem of constructing proper approximations of the boundary conditions, and of the modifications of the scheme near the boundaries. A major part of this chapter is devoted to the so called SBP operators, which are based on summation by parts leading to stability in the sense of the energy method.

Chapter 8 is a little different from the other ones, since the box scheme discussed there is only second order accurate. However, because of the compact nature, it is sometimes more accurate than higher order ones, and some recent results regarding this property are presented. Another advantage is that it can easily be generalized to nonuniform grids.

The next two chapters are devoted to applications. The purpose is not to give a survey of problems and numerical methods, but rather to illustrate the application of certain high order methods in some detail for a few problems of high interest in the engineering and scientific communities. In this way, we hope to give an idea of how to handle the technical details. In Chapter 9 we discuss wave propagation problems described by first order hyperbolic PDE with application in acoustics and electromagnetics. Here we concentrate on a class of high order methods based on staggered grids, and demonstrate the effectiveness, even if the solutions are not smooth. We also present a new method of embedding the boundary with a certain definition of the coefficients in the PDE such that the true boundary conditions are well represented. In Chapter 10 we discuss incompressible fluid dynamics. The flow is governed by the Navier–Stokes equations, but we give a special presentation of the Stokes equations, since they play a special role in the Navier–Stokes solver. The method is semi-implicit and fourth order accurate in space. Large linear systems of equations must be solved for each time step, and we present the iterative solver in some detail as well.

A big challenge, particularly in gas dynamics, is computation of solutions to nonlinear problems containing discontinuities or shocks. This requires some special theory and specialized methods. Techniques that work well for nonlinear problems with smooth solutions, like the incompressible Navier–Stokes equations in Chapter 10, are not well suited. We discuss the theory and methods for these problems in Chapter 11. This is an area where low order methods dominate even more than for linear problems, but we give some emphasis on those methods that can be generalized to high order.

Each chapter has a summary section at the end. These sections contain a brief summary of the theory and results of the chapter, and also some historical remarks and comments on available literature.

When electronic computers came in use in the forties, the field of Numerical Analysis expanded very quickly. Already from the beginning, the solution of ordinary and partial differential equations was in focus, and the development of FDM set full speed. Hardly no other methods were considered, and it was not until the late sixties, that finite element methods (FEM) started emerging for solving PDE

problems. In the beginning, FEM were used mostly for steady state problems, and for approximation of the space part of time dependent problems. For approximation in time, finite difference methods were used, and this is still the case for many applications. The great advantage with FEM is their flexibility when it comes to approximation of irregular domains. This flexibility is shared by finite volume methods (FVM), but construction of high order FVM is not that easy.

The geometric flexibility of FEM is not shared by spectral and pseudo-spectral methods, which was the next class of numerical methods that emerged for PDE. Their strength is the very high accuracy relative to the required work. However, these methods have at least as many restrictions as FDM on the type of computational domain, in particular those that are based on approximation by trigonometric polynomials (Fourier methods). Later this difficulty was partly overcome by the use of spectral element methods, where the domain is partitioned into many subdomains with orthogonal polynomials used for approximation on each one of them.

During the last decade, discontinuous Galerkin methods have arisen as a new interesting class of methods. They can be seen as a generalization of FEM, and have the potential of leading to faster algorithms.

In the final chapter, we give a brief introduction to all of the methods mentioned here.

The available commercial and public software is today dominated by algorithms based on finite element methods (or finite volume methods for problems in fluid problems). One reason for this is that the development of unstructured grid generators has made enormous progress during the last decades, and this is important for many applications. On the other hand, for problems where structured grids provide an acceptable representation of the computational domain, it is hard to beat a good high order difference method when it comes to implementation, speed and effectiveness.

We have limited the presentation in this book almost exclusively to uniform grids. The use of Cartesian grids is for convenience, we could of course use any other of the classic coordinate systems like for example cylindrical coordinates. In general, the Cartesian uniform grid can be seen as the model for all cases where a smooth mapping takes the physical domain to a rectangle in 2-D or hyper-rectangle in higher dimensions.

If a smooth mapping cannot be used for transformation of the whole domain, one can use Cartesian coordinates together with some sort of interpolation procedure near the boundary, and construct the finite differences locally on unstructured grids. Another way is to construct one local structured grid that fits the irregular boundary, and another grid that is used in the main part of the domain. These two (or more) grids are then connected via interpolation. A third way is to couple a finite difference method with a finite element or finite volume method near the boundary, and in this way arriving at a hybrid method. Still another way is to embed the irregular boundary in a larger domain with regular boundaries, and to enforce the boundary conditions by some modification of the PDE.

If the need for a nonstructured grid arises from the fact that the solution varies on very different scales, the most general finite difference technique is based on piecewise uniform grids, that are coupled by an interpolation procedure.

There is also the possibility to construct FDM directly on unstructured grids in the whole computational domain. However, the stability is then a harder issue, and furthermore, the effectiveness will not be much better than with FEM.

Acknowledgment
The main part of this book was written after my retirement from the chair at the Division of Scientific Computing at Uppsala University. However, the university has provided all the necessary infrastructure, such as libraries and computers during the whole project. Furthermore, the staff at the Department of Information Technology have been very helpful when needed. During the writing period I have also spent some time as a visitor at Stanford University, and I want to express my gratitude to the Center for Turbulence Research (CTR) and the Institute for Computational Mathematics in Engineering (ICME) for providing very good working conditions.

Finally I would like to thank my wife Margareta for enduring still another major undertaking from my side after formal retirement.

Uppsala, Sweden, November 2007 *Bertil Gustafsson*

Contents

1 When are High Order Methods Effective? 1
1.1 Preliminaries 1
1.2 Wave Propagation Problems 2
1.3 Parabolic Equations 8
1.4 Schrödinger Type Equations 11
1.5 Summary 12

2 Well-posedness and Stability 13
2.1 Well Posed Problems 13
2.2 Periodic Problems and Fourier Analysis 16
2.2.1 The PDE Problem 17
2.2.2 Difference Approximations 21
2.3 Initial–Boundary Value Problems and the Energy Method 29
2.3.1 The PDE Problem 29
2.3.2 Semidiscrete Approximations 33
2.3.3 Fully Discrete Approximations 38
2.4 Initial–Boundary Value Problems and Normal Mode Analysis for Hyperbolic Systems 41
2.4.1 Semidiscrete Approximations 41
2.4.2 Fully Discrete Approximations 59
2.5 Summary 66

3 Order of Accuracy and the Convergence Rate 69
3.1 Periodic Solutions 69
3.2 Initial–Boundary Value Problems 72
3.3 Summary 79

4 Approximation in Space 81
4.1 High Order Formulas on Standard Grids 81
4.2 High Order Formulas on Staggered Grids 85
4.3 Compact Padé Type Difference Operators 87

4.4 Optimized Difference Operators 91
4.5 Summary 93

5 Approximation in Time 95
5.1 Stability and the Test Equation 95
5.2 Runge–Kutta Methods 97
5.3 Linear Multistep Methods 102
5.4 Deferred Correction 108
5.5 Richardson Extrapolation 111
5.6 Summary 113

6 Coupled Space-Time Approximations 115
6.1 Taylor Expansions and the Lax–Wendroff Principle 115
6.2 Implicit Fourth Order Methods 117
6.3 Summary 124

7 Boundary Treatment 127
7.1 Numerical Boundary Conditions 127
7.2 Summation by Parts (SBP) Difference Operators 130
7.3 SBP Operators and Projection Methods 140
7.4 SBP Operators and Simultaneous Approximation Term (SAT) Methods 147
7.5 Summary 155

8 The Box Scheme 157
8.1 The Original Box Scheme 157
8.2 The Shifted Box Scheme 161
8.3 Two Space Dimensions 165
8.4 Nonuniform Grids 169
8.5 Summary 176

9 Wave Propagation 177
9.1 The Wave Equation 177
9.1.1 One Space Dimension 178
9.1.2 Two Space Dimensions 185
9.2 Discontinuous Coefficients 192
9.2.1 The Original One Step Scheme 193
9.2.2 Modified Coefficients 201
9.2.3 An Example with Discontinuous Solution 206
9.3 Boundary Treatment 209
9.3.1 High Order Boundary Conditions 209
9.3.2 Embedded Boundaries 210
9.4 Summary 216

10 A Problem in Fluid Dynamics 219
10.1 Large Scale Fluid Problems and Turbulent Flow 219
10.2 Stokes Equations for Incompressible Flow 220
10.3 A Fourth Order Method for Stokes Equations 223
10.4 Navier–Stokes Equations for Incompressible Flow 228
10.5 A Fourth Order Method for Navier–Stokes Equations 231
10.6 Summary 242

11 Nonlinear Problems with Shocks 245
11.1 Difference Methods and Nonlinear Equations 245
11.2 Conservation Laws 246
11.3 Shock Fitting 251
11.4 Artificial Viscosity 252
11.5 Upwind Methods 257
11.6 ENO and WENO Schemes 261
11.7 Summary 265

12 Introduction to Other Numerical Methods 267
12.1 Finite Element Methods 267
12.1.1 Galerkin FEM 267
12.1.2 Petrov–Galerkin FEM 281
12.2 Discontinuous Galerkin Methods 283
12.3 Spectral Methods 289
12.3.1 Fourier Methods 290
12.3.2 Polynomial Methods 295
12.4 Finite Volume Methods 300

A Solution of Difference Equations 307

B The Form of SBP Operators 311
B.1 Diagonal H-norm 311
B.2 Full H_0-norm 317
B.3 A Padé Type Operator 323

References 325

Index 331

Acronyms

BDF	Backward Differentiation Formulas
CFL	Courant–Friedrichs–Levy
CG	Conjugate Gradient (methods)
CGSTAB	Conjugate Gradient Stabilized (method)
DIRK	Diagonal Implicit Runge–Kutta (methods)
DFT	Discrete Fourier Transform
ENO	Essentially Nonoscillating
ERK	Explicit Runge–Kutta (methods)
FDM	Finite Difference Method(s)
FEM	Finite Element Method(s)
FFT	Fast discrete Fourier Transform
FVM	Finite Volume Method(s)
GKS	Gustafsson–Kreiss–Sundström
IBVP	Initial–Boundary Value Problems
IRK	Implicit Runge–Kutta (methods)
n-D	n space Dimensions
ODE	Ordinary Differential Equations
PDE	Partial Differential Equations
RK	Runge–Kutta
SAT	Simultaneous Approximation Term
SBP	Summation By Parts
SSP	Strong Stability Preserving
TV	Total Variation
TVD	Total Variation Diminishing
WENO	Weighted Essentially Nonoscillating

Chapter 1
When are High Order Methods Effective?

In the modern era of computational mathematics beginning in the forties, most methods in practical use were first or second order accurate. Actually, that is the case even today, and the reason for this low accuracy is probably the simpler implementation. However, from an efficiency point of view, it is most likely that they should be substituted by higher order methods. These require more programming work, and the computer has to carry out more arithmetic operations per grid point. However, for a given error tolerance, the number of grid points can be reduced substantially, and in several dimensions, one may well reduce the computing time and memory requirement by orders of magnitude.

In this chapter, we shall investigate how the order of accuracy affects the performance of the method. We shall use simple model problems to get an idea of what we can expect for different types of PDE.

1.1 Preliminaries

Every numerical method for solution of differential equations is based on some sort of discretization, such that the computer can handle it in finite time. The most common discretization parameter is the step size h, which denotes the typical distance between points in a grid where the solution can be computed. If the true solution can formally be expressed as an infinite sum, the discretization parameter is N, which denotes the finite number of terms in the approximating sum. For difference methods, the approximation is related to the differential equation by the *truncation error* τ, and the *order of accuracy* is defined as p if $\tau \sim h^p$. Under certain conditions that will be described in Chapter 3, this leads to error estimates of the same order, i.e., the *error in the solution* is also proportional to h^p. Those methods that have $p > 2$ are usually called higher order methods.

The difference approximations throughout this book will be built by the basic centered, forward and backward difference operators on a uniform grid with step size h

$$x_j = jh, \quad j = 0, \pm 1, \pm 2, \ldots .$$

Grid functions in space are defined by $u(x_j) \to u_j$, and the difference operators are

$$\begin{aligned} D_0 u_j &= (u_{j+1} - u_{j-1})/(2h), \\ D_+ u_j &= (u_{j+1} - u_j)/h, \\ D_- u_j &= (u_j - u_{j-1})/h. \end{aligned}$$

We also define the shift operator by

$$E u_j = u_{j+1}.$$

All the difference operators commute, such that for example

$$D_0 D_+ D_- = D_+ D_- D_0 = D_0 D_- D_+ .$$

For any difference operator Q we use the simplified notation Qu_j, which is to be interpreted as $(Qu)_j$. This notation is used even when j is fixed, i.e., Qu_0 means $(Qu)_{j=0}$.

The time discretization is done on a uniform grid

$$t_n = nk, \quad n = 0, 1, \ldots ,$$

where k is the time step. The approximation of a function $u(x_j, t_n)$ is denoted by u_j^n.

1.2 Wave Propagation Problems

In this section we shall consider wave propagation problems represented by the simple model equation

$$u_t = u_x$$

satisfied by the simple wave $e^{i\omega(x+t)}$, where ω is the wave number . It may seem as a complication to consider complex solutions in the analysis, also when we know that the solutions are real, but actually it is a simplification. The reason is that the algebraic operations become easier in this way when Fourier analysis is used.

The most straightforward way of finding the order of accuracy of a certain difference approximation is Taylor expansion . It is easily shown that for any sufficiently smooth function $u(x)$, we have

$$D_0 u(x) = u_x + \frac{h^2}{6} u_{xxx} + \mathscr{O}(h^4),$$

i.e., D_0 is a second order approximation of $\partial/\partial x$. The leading part of the truncation error can now be eliminated by including a difference approximation of it. Again, it is easily shown by Taylor expansion that

$$D_0 D_+ D_- u(x) = u_{xxx} + \mathcal{O}(h^2)\,,$$

which gives us the fourth order approximation

$$Q_4 u(x) = u_x + \mathcal{O}(h^4)\,,$$

where

$$Q_4 = D_0(I - \frac{h^2}{6} D_+ D_-)\,.$$

After a few more Taylor expansions, one obtains the sixth order approximation

$$Q_6 = D_0(I - \frac{h^2}{6} D_+ D_- + \frac{h^4}{30} D_+^2 D_-^2)\,.$$

Since the solution of our model problem is periodic, we consider the computational domain $0 \le x \le 2\pi\,,\ 0 \le t$, and the grid

$$x_j = jh\,,\quad j = 0, 1, \ldots, N\,,\quad (N+1)h = 2\pi\,.$$

With the notation $Q_2 = D_0$, we have the three alternative approximations

$$\frac{du_j}{dt} = Q_p u_j\,,\quad p = 2, 4, 6\,. \tag{1.1}$$

With the ansatz

$$u_j(t) = \hat{u}(t) e^{i\omega x_j}\,,$$

we get the *Fourier transform* of the equation (1.1)

$$\frac{d\hat{u}}{dt} = \hat{Q}_p \hat{u}\,,$$

where $\hat{Q}_p$ is the Fourier transform of Q_p. Since

$$D_0 e^{i\omega x} = \frac{i}{h} \sin\xi\, e^{i\omega x}\,,$$
$$D_+ D_- e^{i\omega x} = -\frac{4}{h^2} \sin^2 \frac{\xi}{2}\, e^{i\omega x}\,,$$

where $\xi = \omega h$, we get

$$\begin{aligned}
\hat{Q}_2 &= \frac{i}{h} \sin\xi\,,\\
\hat{Q}_4 &= \frac{i}{h} \sin\xi\,(1 + \frac{2}{3} \sin^2 \frac{\xi}{2})\,,\\
\hat{Q}_6 &= \frac{i}{h} \sin\xi\,(1 + \frac{2}{3} \sin^2 \frac{\xi}{2} + \frac{8}{15} \sin^4 \frac{\xi}{2})\,.
\end{aligned} \tag{1.2}$$

The solution in Fourier space is $\hat{u}(t) = \exp(\hat{Q}_p t)$, which gives the solution in physical space

$$u_j(t) = e^{i\omega x_j + \hat{Q}_p t}\,.$$

The approximation changes $i\omega$ in the exponent to $\hat{Q}_p = \hat{Q}_p(\xi)$, and it makes sense to compare these quantities. After normalization by the factor h/i, we compare ξ with $h\hat{Q}_p(\xi)/i$. Assuming for convenience that N is an even number, the highest wave number that can be represented on the grid is $\omega = N/2 = (\pi - h/2)/h$. Since h is arbitrarily small, the range of ξ is $0 \le |\xi| \le \pi$. Figure 1.1 shows how $h\hat{Q}_p(\xi)/i$ approaches ξ for increasing p. The true wave speed for our problem is -1, and for the approximation it is $-\hat{Q}_p/(i\omega)$, which is always less than 1 in magnitude. The waves will be slowed down by the approximation, more for higher frequencies, and this error is called the *dispersion error*.

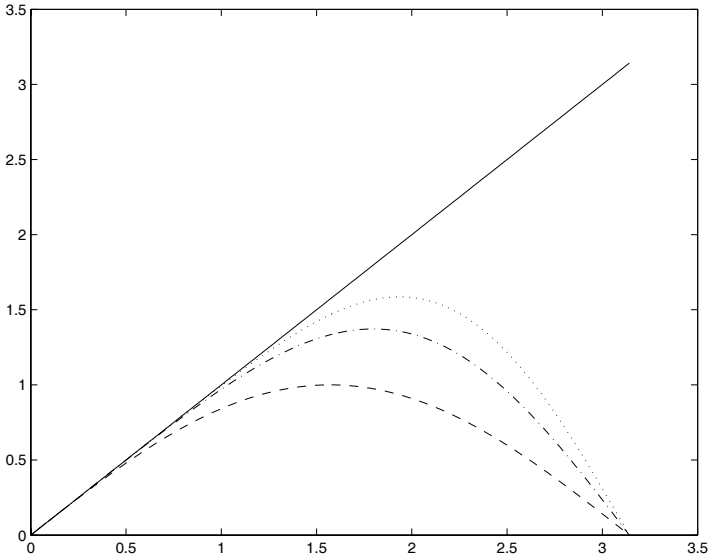


Fig. 1.1 $\xi(-)$, $\hat{Q}_p(\xi)h/i$, $p = 2(--)$, $4(-\cdot)$, $6(\cdot\cdot)$

Next we will estimate the number of grid points $N_p = N + 1$ that are necessary for achieving a certain accuracy, and for convenience, we assume that ω is positive. The solution is periodic in both time and space, and the length of one period in time is $2\pi/\omega$. If we want to compute q periods, the time interval is $0 \le t \le 2\pi q/\omega$. We also introduce the number of grid points per wavelength

$$M_p = N_p/\omega = 2\pi/\xi\,, \quad p = 2,4,6\,.$$

We assume that $\xi << 1$, and investigate the error defined as

$$v^{(p)}(t) = \max_j |e^{i\omega(x_j+t)} - e^{i\omega x_j + \hat{Q}_p t}| = |1 - e^{(-i\omega+\hat{Q}_p)t}|.$$

By Taylor expansion in terms of ξ, we obtain

$$v^{(2)}(t) \approx \frac{\omega t \xi^2}{6} \leq \frac{\pi q \xi^2}{3} = \frac{4\pi^3 q}{3M_2^2}. \tag{1.3}$$

By prescribing the maximum error ε for $v^{(2)}$, we obtain an expression for M_2 in terms of q and ε. By applying the same procedure for $p = 4$ and $p = 6$, we get the complete list

$$\begin{aligned} M_2 &\approx 2\pi \left(\frac{\pi q}{3\varepsilon}\right)^{1/2}, \\ M_4 &\approx 2\pi \left(\frac{\pi q}{15\varepsilon}\right)^{1/4}, \\ M_6 &\approx 2\pi \left(\frac{\pi q}{70\varepsilon}\right)^{1/6}. \end{aligned} \tag{1.4}$$

Since the work per grid point increases by a constant factor (independent of q and ε) for each level of increased accuracy, we note that a higher order method always wins if the accepted error level is low enough, and/or the time interval is large enough. Furthermore, the gain is more pronounced in several space dimensions. If an explicit time integrator is used, there is a limit on the time step for stability reasons. If the number of grid points in each space direction can be reduced by a factor $\alpha > 1$ by using a higher order method, the total reduction of the number of grid points for a problem with three space dimensions, is a factor α^4.

Indeed, there is a substantial gain already in one space dimension, and quite modest error levels. Table 1.1 shows M_p for a 1 % error level and $q = 20$ and $q = 200$ respectively.

Table 1.1 M_p for 1 % error level

q	M_2	M_4	M_6
20	287	28	13
200	909	51	20

For several space dimensions, the total number of grid points for a second order method becomes totally unrealistic for long time integration. Clearly it pays to use a fourth order method in these cases, even if the computing time per grid point is longer. Going to sixth order is more doubtful in one or two space dimensions.

The formal order of accuracy, as well as the estimates above, are derived under the assumption that the solution $u(x,t)$ is smooth. In practice this is seldom the case, and one might wonder how the higher order methods behave for less smooth solutions. Consider for example the problem

$$\begin{aligned} u_t &= u_x\,, \quad -1 \le x \le 1\,,\ 0 \le t\,, \\ u(x,0) &= |\sin(\pi x/2)|^r\,, \end{aligned} \tag{1.5}$$

where r is an odd number. The solution $|\sin(\pi(x+t)/2)|^r$ is 2-periodic in both time and space. The derivative of order r is discontinuous, i.e., the solution becomes smoother for higher r. Figures 1.2, 1.3, 1.4 show the solutions for $r = 1, 3, 5$ and its approximations $u^{(p)}$ obtained by a formally pth order accurate method. The figures show the solution at $t = 6$ when the true solution is back at its initial state for the third time. Even for $r = 1$, the higher order methods give better solutions. The dramatic change between the 2nd and 4th order methods is clearly visible.

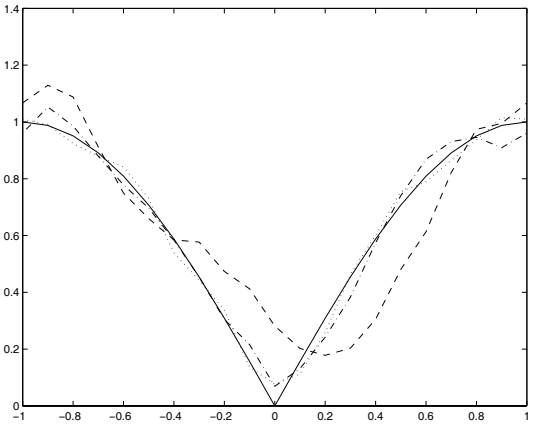


Fig. 1.2 $u(x,6),\ r = 1\ (-),\ u^{(p)},\ p = 2(--),\ p = 4(-\cdot),\ p = 6(\cdot\cdot),\ N = 20$

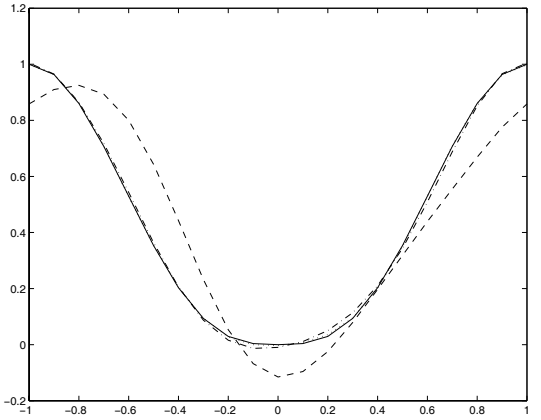


Fig. 1.3 $u(x,6),\ r = 3\ (-),\ u^{(p)},\ p = 2(--),\ p = 4(-\cdot),\ p = 6(\cdot\cdot),\ N = 20$

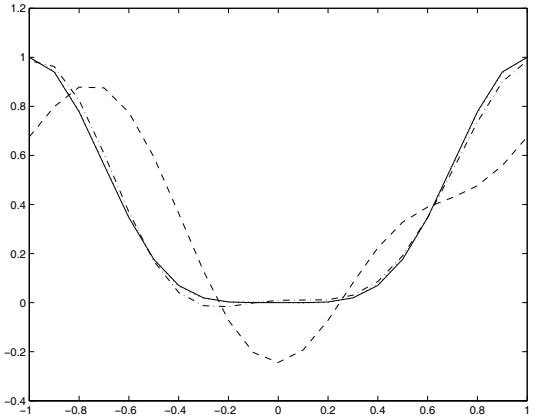


Fig. 1.4 $u(x,6),\ r=5\ (-),\ u^{(p)},\ p=2(--),\ p=4(-\cdot),\ p=6(\cdot\cdot),\ N=20$

The next three figures show the l_2-error $\sum_j |v_j(t)|^2 h$ for $r=1,3,5$ as a function of time for $N=20$ and $N=40$. For the case $r=1$, the convergence rate is roughly linear ($\sim h$) for all three methods, but the error is significantly smaller for the higher order ones. For the smoother cases $r=3$ and $r=5$, the convergence rate goes up considerably as expected.

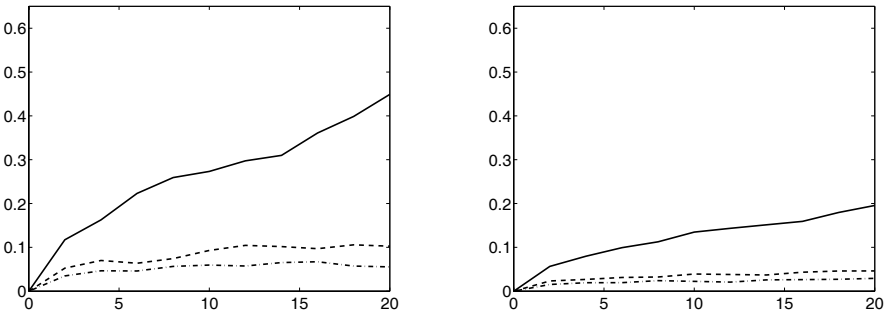
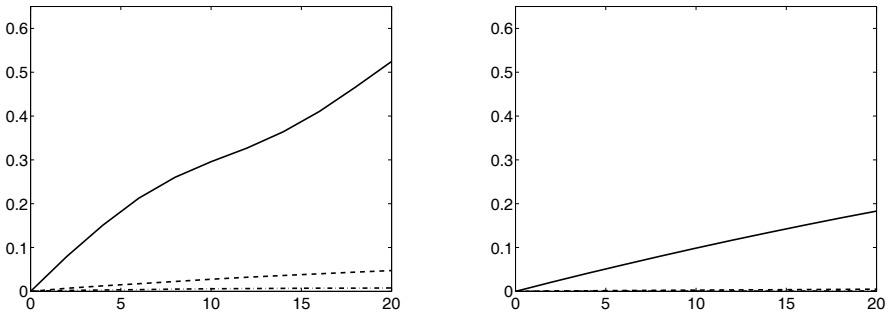
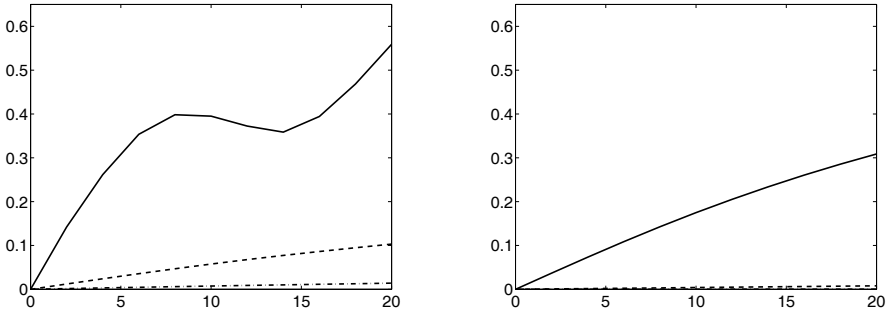


Fig. 1.5 l_2-error, $r=1,\ p=2(--),\ p=4(-\cdot),\ p=6(\cdot\cdot)$



(a) N=20 (b) N=40

Fig. 1.6 l_2-error, $r = 3,\ p = 2(--),\ p = 4(-\cdot),\ p = 6(\cdot\cdot)$



(a) N=20 (b) N=40

Fig. 1.7 l_2-error, $r = 5,\ p = 2(--),\ p = 4(-\cdot),\ p = 6(\cdot\cdot)$

1.3 Parabolic Equations

For *parabolic* equations, the situation is a little different. Consider the heat equation in its simplest form and the model problem

$$\begin{aligned} u_t &= u_{xx}\,, \\ u(x,0) &= e^{i\omega x} \end{aligned}$$

with the solution $u(x,t) = e^{-\omega^2 t + i\omega x}$. The standard second order approximation is

$$\begin{aligned} \frac{du_j}{dt} &= D_+ D_- u_j\,, \\ u_j(0) &= e^{i\omega x_j} \end{aligned}$$

with the solution $u_j(t) = e^{\hat{P}_2(\xi)t+i\omega x_j}$, where

$$\hat{P}_2(\xi) = -\frac{4}{h^2}\sin^2\frac{\xi}{2}$$

is the Fourier transform of D_+D_-. Apparently, the accuracy is determined by the ability of $\hat{P}_2(\xi)$ to approximate $-\omega^2$.

The fourth and sixth order approximations are

$$\frac{du_j}{dt} = D_+D_-(I - \frac{h^2}{12}D_+D_-)u_j\,,$$
$$\frac{du_j}{dt} = D_+D_-(I - \frac{h^2}{12}D_+D_- + \frac{h^4}{90}(D_+D_-)^2)u_j$$

with the solution $u_j(t) = e^{\hat{P}_p(\xi)t+i\omega x_j}\,,\ p = 4, 6$, where

$$\hat{P}_4(\xi) = -\frac{4}{h^2}\sin^2\frac{\xi}{2}(1 + \frac{1}{3}\sin^2\frac{\xi}{2})\,.$$
$$\hat{P}_6(\xi) = -\frac{4}{h^2}\sin^2\frac{\xi}{2}(1 + \frac{1}{3}\sin^2\frac{\xi}{2} + \frac{8}{45}\sin^4\frac{\xi}{2})\,.$$

Figure 1.8 shows a comparison between $\xi^2 = \omega^2h^2$ and $-\hat{P}_p(\xi)h^2,\ p = 2, 4, 6$. All three approximations give a stronger damping with time than the true solution has.

In order to estimate the necessary number of grid points, we consider the error

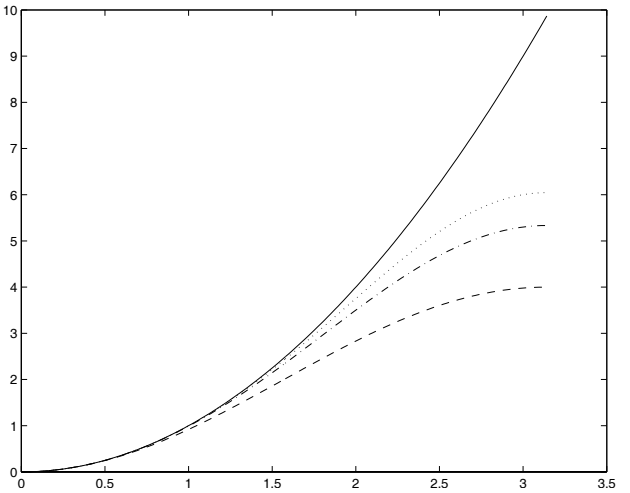


Fig. 1.8 $\xi^2(-)\,,\ -\hat{P}_p(\xi)h^2\,,\ p = 2(--)\,,\ 4(-\cdot)\,,\ 6(\cdot\cdot)$

$$v^{(p)}(t) = \max_j |e^{i\omega x_j - \omega^2 t} - e^{i\omega x_j + \hat{P}_p(\xi)t}| = e^{-\omega^2 t}|1 - e^{(\omega^2 + \hat{P}_p(\xi))t}|, \; p = 2, 4, 6 .$$

For small $|\xi|$, we have by Taylor expansion

$$\hat{P}_2(\xi) \approx -\omega^2\big(1 - \frac{\xi^2}{12} + \mathcal{O}(\xi^4)\big) ,$$

giving the approximative error

$$v^{(2)}(t) \approx \frac{1}{12}\omega^2\xi^2 t e^{-\omega^2 t} . \tag{1.6}$$

In contrast to the hyperbolic case, the error is maximal at a point in time which is independent of the number of computed periods:

$$\max_t(v^{(2)}(t)) \approx \frac{\xi^2}{12e} \quad \text{for } t = \frac{1}{\omega^2} .$$

For the prescribed error level ε, the necessary number of points per wave length is

$$M_2 \approx \frac{2\pi}{(12e\varepsilon)^{1/2}} .$$

A similar calculation for the fourth and sixth order cases gives the number of points per wave length

$$M_4 \approx \frac{2\pi}{(90e\varepsilon)^{1/4}} ,$$
$$M_6 \approx \frac{2\pi}{(560e\varepsilon)^{1/6}} .$$

Table 1.2 shows M_p for the error levels $\varepsilon = 0.01, 0.001, 0.0001$.

Table 1.2 M_p for different error levels, parabolic equations

ε	M_2	M_4	M_6
0.01	11	5	4
0.001	35	9	6
0.0001	110	16	9

Since M_p is independent of the length of the time integration (except for the first short part), it takes stronger accuracy requirements to get a real advantage of higher order accurate methods for second order parabolic problems. It can be shown that this conclusion holds also for higher order parabolic equations, since they all have strong damping with time, which causes the error to peak after a short time.

1.4 Schrödinger Type Equations

The *Schrödinger equation* in its simplest form is

$$u_t = iu_{xx},$$

with complex solutions u. With the usual Fourier component as initial function, the solution is $u(x,t) = e^{i(\omega x - \omega^2 t)}$. The only difference from the parabolic case, is the extra constant i multiplying the space derivative. However, this difference is quite significant when it comes to choosing the best order of approximation, since we don't have any damping of the amplitudes any longer. The behavior is more like the hyperbolic case, with increasing error all the time until it reaches a level $\mathcal{O}(1)$.

The approximations of the Schrödinger equation are obtained precisely as for the parabolic model problem in the previous section, except for an extra factor i multiplying them. Therefore, we don't have to go through the whole analysis again, but rather substitute $e^{-\omega^2 t}$ by $e^{-i\omega^2 t}$. The error derived in (1.6) is now obtained as

$$v^{(2)}(t) \approx \frac{1}{12}\omega^2 \xi^2 t,$$

and similarly for the $v^{(4)}$ and $v^{(6)}$. Therefore, we have the same situation as for the hyperbolic case. The only difference is that the length of a period in time is now $2\pi/\omega^2$, but for q periods it leads to the same type of inequality as (1.3):

$$v^{(2)} \approx \frac{\omega^2 t \xi^2}{12} \leq \frac{\pi q \xi^2}{6} = \frac{2\pi^3 q}{3M_2^2}.$$

With a prescribed error level ε, we get in analogy with (1.4)

$$\begin{aligned} M_2 &\approx 2\pi \left(\frac{\pi q}{6\varepsilon}\right)^{1/2}, \\ M_4 &\approx 2\pi \left(\frac{\pi q}{45\varepsilon}\right)^{1/4}, \\ M_6 &\approx 2\pi \left(\frac{\pi q}{270\varepsilon}\right)^{1/6}. \end{aligned} \tag{1.7}$$

Compared to the parabolic case, there is now an extra factor q involved. The error will grow with increasing time intervals for integration, but the influence becomes weaker with increasing order of accuracy. For the 1 % error level, we get Table 1.3. The results are very similar to the hyperbolic case, and again the most dramatic advantage is obtained by going from second to fourth order. Note however, that one period in time is now $2\pi/\omega^2$, which is shorter than in the hyperbolic case. So if the total time interval for integration is $[0,T]$ for both cases, the comparison between lower and higher order methods comes out even more favorable for the higher order ones in the case of Schrödinger type equations.

Table 1.3 M_p for 1 % error level

q	M_2	M_4	M_6
20	203	22	11
200	642	38	16

1.5 Summary

The first theoretical analysis regarding optimal order of accuracy for first order hyperbolic equations was done by Kreiss and Oliger 1972 [Kreiss and Oliger, 1972]. This type of analysis has been presented in this chapter, also for higher order differential equations. The first rule of thumb is that the advantage of high order methods is more pronounced for problems where small errors in the solution are required. Secondly, for real equations $\partial u/\partial t = a\partial^q u/\partial t^q$, a real and q odd, there is an extra advantage with high order methods for long time integrations. This extra advantage is there also for complex equations with $a = i$ and even q, giving the solutions a wave propagation character. For problems in several space dimensions, the advantage with high order methods is even more pronounced.

For parabolic problems with real a and even q, the advantage with higher order methods is less. The reason for this is that there is an inherent damping in the equation, which means that the errors are not allowed to accumulate with time as for hyperbolic problems. Even if the integration is carried out over a long time interval, the error behaves more like short time integration.

The analysis presented here is based on the behavior of the approximation when applied to a single wave with a fixed wave number ω. If, for a whole wave package, the highest wave number ω_0 that is of interest to us is determined a priori, then the guidelines derived in this chapter tell us what method should be used to obtain a certain accuracy for the whole solution. One could of course have more involved criteria, where for example less accuracy is required for higher wave numbers, and then the conclusions would be modified. One can also discuss in terms of group velocity as Trefethen did in [Trefethen, 1983], see also [Strikwerda, 1989].

The time discretization can of course also be included in the analysis, as was done in [Swartz and Wendroff, 1974] for hyperbolic problems. The results are in line with the ones summarized above. A more detailed comparison between different schemes for wave propagation problems was carried out by Zingg in [Zingg, 2000]. Further investigations are carried out in Chapters 6 and 9.

Chapter 2
Well-posedness and Stability

Stability is a fundamental concept for any type of PDE approximation. A stable approximation is such that small perturbations in the given data cause only small perturbations in the solutions. Furthermore, the solutions converge to the true solution of the PDE as the step size h tends to zero. The extra condition required for this property is that the PDE problem is well posed. In this chapter we shall present a survey of the basic theory for the well-posedness and stability. The theory can be divided into three different techniques: Fourier analysis for Cauchy and periodic problems, the energy method and Laplace analysis (also called normal mode analysis) for initial–boundary value problems. In order to emphasize the similarities between the continuous and discrete case, we treat the application of each technique to both the PDE and the finite difference approximations in the same section (the Laplace technique for PDE is omitted).

2.1 Well Posed Problems

We consider a general initial–boundary value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= Pu + F\,, \quad 0 \le t\,, \\ Bu &= g\,, \\ u &= f\,, \quad t = 0\,. \end{aligned} \tag{2.1}$$

Here P is a differential operator in space, and B is a boundary operator acting on the solution at the spacial boundary. (Throughout this book, we will refer to t as the time coordinate, and to the remaining independent variables as the space variables, even if the physical meaning may be different.) There are three types of data that are fed into the problem: F is a given forcing function , g is a boundary function and f is an initial function. (By "function" we mean here the more general concept "vector function", i.e., we are considering systems of PDE.) A well posed problem

has a unique solution u, and there is an estimate

$$||u||_I \le K(||f||_{II} + ||F||_{III} + ||g||_{IV}), \tag{2.2}$$

where K is a constant independent of the data. In general, there are four different norms involved, but $||\cdot||_I$ and $||\cdot||_{II}$ are often identical.

Let v be the solution of the perturbed problem

$$\begin{aligned} \frac{\partial v}{\partial t} &= Pv + F + \delta F\ ,\ \ 0 \le t\,, \\ Bv &= g + \delta g\,, \\ v &= f + \delta f\ ,\ \ t = 0\,. \end{aligned} \tag{2.3}$$

Assuming that P and B are linear operators, we subtract (2.1) from (2.3), and obtain for the perturbation $w = v - u$ of the solution

$$\begin{aligned} \frac{\partial w}{\partial t} &= Pw + \delta F\ ,\ \ 0 \le t\,, \\ Bw &= \delta g\,, \\ w &= \delta f\ ,\ \ t = 0\,. \end{aligned}$$

The estimate (2.2) can now be applied to w:

$$||w||_I \le K(||\delta f||_{II} + ||\delta F||_{III} + ||\delta g||_{IV})\,.$$

Hence, if K has a moderate size, small perturbations δf, δF, δg in the data cause a small perturbation w in the solution.

As an example, consider a scalar problem in one space dimension and $0 \le x \le 1$. Then $u = u(x,t)$, $F = F(x,t)$, $g = g(t)$, $f = f(x)$, and we choose

$$||u||_I^2 = ||u(\cdot,t)||^2 = \int_0^1 |u(x,t)|^2 dx\,.$$

If the boundary conditions are

$$\begin{aligned} u(0,t) &= g_0(t)\,, \\ u(1,t) &= g_1(t)\,, \end{aligned}$$

then a typical estimate has the form

$$||u(\cdot,t)||^2 \le K\Big(||f(\cdot)||^2 + \int_0^t ||F(\cdot,\tau)||^2 d\tau + \int_0^t \left(|g_0(\tau)|^2 + |g_1(\tau)|^2\right) d\tau\Big)\,,$$

where K may depend on t, but not on f, F, g_0, g_1. We could of course reformulate this estimate as in (2.2), but it is more convenient to keep the squared norms.

If the domain in space doesn't have any boundaries, there is of course no boundary condition. However, for a difference approximation, there has to be boundaries

for computational reasons. A special, but frequent case, is that the solutions are periodic in space. In that case the computation is done in a bounded domain, with the requirement that the solution and all its derivatives are equal at the both ends of the interval. This is often used as a model problem, since Fourier analysis can be used for investigating stability.

For periodic problems, the solution can be written as a Fourier series, and the behavior of the coefficients is the key issue. Let $v(x) = [v^{(1)}\, v^{(2)} \ldots v^{(m)}]^T$ be a 2π-periodic vector function. The following lemma connects the size of these coefficients with the L_2-norm of $v(x)$, which is defined by

$$||v(\cdot)||^2 = \int_0^{2\pi} |v(x)|^2 dx, \quad |v|^2 = \sum_{\nu=1}^{m} |v^{(\nu)}|^2 .$$

Lemma 2.1. (Parseval's relation) *) Let $v(x)$ be represented by its Fourier series*

$$v(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{v}(\omega) e^{i\omega x} .$$

Then

$$||v(\cdot)||^2 = \sum_{\omega=-\infty}^{\infty} |\hat{v}(\omega)|^2 .$$

□

As an example, consider the heat equation in its simplest form

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \le x \le 2\pi, \ \ 0 \le t, \\ u(x,0) &= f(x). \end{aligned} \tag{2.4}$$

The solution can be written as a Fourier series with time dependent coefficients

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{u}(\omega,t) e^{i\omega x}, \tag{2.5}$$

and the coefficients satisfy

$$\begin{aligned} \hat{u}_t &= -\omega^2 \hat{u}, \\ \hat{u}(\omega,0) &= \hat{f}(\omega), \end{aligned}$$

where $\hat{f}(\omega)$ are the Fourier coefficients of the initial data. The solution is

$$\hat{u}(\omega,t) = e^{-\omega^2 t} \hat{f}(\omega),$$

and by Parseval's relation

$$||u(\cdot,t)||^2 = \sum_{\omega=-\infty}^{\infty} |\hat{u}(\omega,t)|^2 \le \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = ||f(\cdot)||^2 .$$

Here we have proven well-posedness by simply finding the explicit form of the solution.

Assume next, that we want to solve the heat equation backward

$$\begin{aligned} u_t &= u_{xx}\,, \quad 0 \le x \le 2\pi\,, \;\; 0 \le t \le T\,, \\ u(x,T) &= \phi(x)\,. \end{aligned}$$

By the variable transformation $\tau = T - t$, $v(x,\tau) = u(x,t) = u(x,T-\tau)$, we get

$$\begin{aligned} v_\tau &= -v_{xx}\,, \quad 0 \le x \le 2\pi\,, \;\; 0 \le \tau \le T\,, \\ v(x,0) &= \phi(x)\,. \end{aligned}$$

By the same procedure as above, we obtain

$$\hat{v}(\omega,\tau) = e^{\omega^2\tau}\hat{\phi}(\omega)\,.$$

It is now impossible to obtain an estimate of the type

$$||v(\cdot,\tau)|| \le K||\phi(\cdot)||\,,$$

where K is a constant, since

$$||v(\cdot,\tau)||^2 = \sum_{\omega=-\infty}^{\infty} e^{2\omega^2\tau}|\hat{\phi}(\omega)|^2\,.$$

This shows that, given a measured heat distribution at a given time T, it is in theory impossible to find the true heat distribution at an earlier time, except for very smooth ϕ with fast decaying Fourier coefficients. The problem is ill posed . In practice it means, that it is extremely difficult to get any reasonable accuracy at $t = 0$, since small errors in the measurements give rise to large errors in the solution. (It should be said that there are numerical methods for ill posed problems, but a discussion of those is outside the scope of this book.)

2.2 Periodic Problems and Fourier Analysis

In this section we shall discuss the so called *Cauchy problem* , i.e., the domain in space is the whole real line. However, for convenience we will assume that the solutions are 2π-periodic in space, i.e., $u(x,t) = u(x+2\pi,t)$, such that we can deal with a finite interval $[0,2\pi]$. When the concept Cauchy problem is used a few times in the text, it refers either to the periodic case or to the case where $u(x,t) = 0$ outside some finite interval in space.

We begin by considering the PDE problem before discretization.

2.2.1 The PDE Problem

Consider the general problem in one space dimension

$$\begin{aligned} \frac{\partial u}{\partial t} &= P(\partial/\partial x)u + F(x,t)\ , \ \ 0 \le t\,, \\ u(x,0) &= f(x)\,, \end{aligned} \tag{2.6}$$

where $P(\partial/\partial x)$ is a linear differential operator , i.e.,

$$P(\partial/\partial x)(\alpha u + v) = \alpha P(\partial/\partial x)u + P(\partial/\partial x)v\,,$$

if α is a constant. Before defining well-posedness, we consider the example

$$u_t = Au_x\,, \quad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}\,, \quad A = \begin{bmatrix} 0 & 4 \\ 1 & 0 \end{bmatrix}\,.$$

The matrix A can be diagonalized, i.e., there is a *similarity transformation* such that

$$T^{-1}AT = \Lambda = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}\,,$$

where

$$T = \begin{bmatrix} 1 & 1 \\ 1/2 & -1/2 \end{bmatrix}\,, \quad T^{-1} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}\,.$$

By the substitution $v = T^{-1}u,\ g = T^{-1}f$, we get the new system

$$\begin{aligned} v_t &= \Lambda v_x\,, \\ v(x,0) &= g(x)\,, \end{aligned}$$

with the solution

$$\begin{aligned} v^{(1)}(x,t) &= g^{(1)}(x+2t)\,, \\ v^{(2)}(x,t) &= g^{(2)}(x-2t)\,. \end{aligned}$$

The norm is defined by

$$||v(\cdot,t)||^2 = \int_0^{2\pi} |v(x,t)|^2 dx\,, \quad |v(x,t)|^2 = |v^{(1)}(x,t)|^2 + |v^{(2)}(x,t)|^2\,,$$

and we have the inequality

$$||Tv|| = \Big(\int_0^{2\pi} |Tv|^2 dx\Big)^{1/2} \le \Big(\int_0^{2\pi} |T|^2 |v|^2 dx\Big)^{1/2} = |T| \cdot ||v||^2\,,$$

where $|T|$ is the matrix norm defined by

$$|T| = \max_{|v|=1} |Tv| .$$

By periodicity it follows that

$$\int_0^{2\pi} |g^{(\nu)}(x \pm 2t)|^2 dx = \int_0^{2\pi} |g^{(\nu)}(x)|^2 dx ,$$

and we get

$$||u(\cdot,t)|| \le |T| \cdot ||v(\cdot,t)|| = |T| \cdot ||g(\cdot)|| = |T| \cdot ||T^{-1} f(\cdot)|| \le |T| \cdot |T^{-1}| \cdot ||f(\cdot)|| .$$

Since the matrix A is not symmetric, the *condition number* $K = |T| \cdot |T^{-1}|$ is greater than 1, but the estimate

$$||u(\cdot,t)|| \le K||f(\cdot)||$$

is the best one we can get.

Next, consider the trivial example $u_t = \alpha u$, where α is a positive constant. Obviously the solution satisfies

$$||u(\cdot,t)|| = e^{\alpha t}||f(\cdot)|| .$$

These two examples indicate that the following definition of well-posedness is appropriate:

Definition 2.1. The problem (2.6) is well posed if for $F(x,t) = 0$ there is a unique solution satisfying

$$||u(\cdot,t)|| \le Ke^{\alpha t}||f(\cdot)|| , \tag{2.7}$$

where K and α are constants independent of $f(x)$. □

If the forcing function F is nonzero, one can show that for a well posed problem, the estimate

$$||u(\cdot,t)|| \le Ke^{\alpha t}\left(||f(\cdot)|| + \int_0^t ||F(\cdot,\tau)|| d\tau\right) \tag{2.8}$$

holds. This is a useful estimate, and it means that the forcing function can be disregarded in the analysis. The norm $||\cdot||_{III}$ in (2.2) is defined by

$$||F||_{III} = \int_0^t ||F(\cdot,\tau)|| d\tau .$$

For the simple examples treated so far, the existence of solutions is a trivial matter, in fact we have constructed them. However, questions concerning existence of solutions in the general case is beyond the scope of this book. Uniqueness, on the other hand, follows immediately from the condition (2.7). Assume that there is another solution v of the problem (2.6). Then by linearity, the difference $w = u - v$ satisfies the initial value problem

$$\frac{\partial w}{\partial t} = P(\partial/\partial x)w\,,\quad 0 \le t\,,$$
$$w(x,0) = 0\,.$$

The condition (2.7) then implies that $w = 0$, i.e., $v = u$.

Next we shall discuss how to verify that the estimate (2.7) holds. In the previous section we saw how the problem (2.4) was converted into a simple set of ordinary differential equations by the Fourier transform, i.e., after writing the solution as a Fourier series. The operator $\partial/\partial x^2$ becomes $-\omega^2$ acting on the Fourier coefficients $\hat{u}$. Let us now apply this technique to general problems in one space dimension. Consider the problem (2.6), where $P(\partial/\partial x)$ is a differential operator with *constant coefficients* . This means that it has the form

$$P(\partial/\partial x) = \sum_{\nu=0}^{q} A_\nu \frac{\partial^\nu}{\partial x^\nu}\,,$$

where the matrices A_ν are independent of x and t. By writing the solution as a Fourier series of the form (2.5), the vector coefficients are obtained as

$$\hat{u}(\omega,t) = e^{\hat{P}(i\omega)t}\hat{f}(\omega)\,,$$

where

$$\hat{P}(i\omega) = \sum_{\nu=0}^{q} A_\nu (i\omega)^\nu\,.$$

Note that if u is a vector with m components, i.e., there are m differential equations in (2.6), then $\hat{P}(i\omega)$ is an $m \times m$ matrix, and it is called the *symbol* or *Fourier transform* of $P(\partial/\partial x)$.

By Parseval's relation, we get

Theorem 2.1. *The problem (2.6) is well posed if and only if there are constants K and α such that for all ω*

$$|e^{\hat{P}(i\omega)t}| \le Ke^{\alpha t}\,. \tag{2.9}$$

□

It is often easier to study the eigenvalues of a matrix rather than the norm. We have

Definition 2.2. The *Petrovski condition* is satisfied if the eigenvalues $\lambda(\omega)$ of $\hat{P}(i\omega)$ satisfy the inequality

$$Re\,(\lambda(\omega)) \le \alpha\,, \tag{2.10}$$

where α is a constant independent of ω. □

Clearly this condition is necessary for stability. There are many ways of prescribing extra conditions such that it is also sufficient for well-posedness. One such condition is given by

Theorem 2.2. *The Petrovski condition is necessary for well-posedness. It is sufficient if there is a constant K and a matrix $T(\omega)$ such that $T^{-1}(\omega)\hat{P}(i\omega)T(\omega)$ is diagonal and $|T^{-1}(\omega)| \cdot |T(\omega)| \le K$ for all ω.* □

Problems in several space dimensions are treated in the same way. By defining the vectors

$$\mathbf{x} = [x^{(1)}\, x^{(2)} \ldots x^{(d)}]^T\,,$$
$$\boldsymbol{\omega} = [\omega^{(1)}\, \omega^{(2)} \ldots \omega^{(d)}]^T\,,$$

the symbol $\hat{P}(i\boldsymbol{\omega})$ is well defined by the formal transition $\partial/\partial x^{(\nu)} \to i\omega^{(\nu)}$. For example, when using the more common notation $x = x^{(1)}$, $y = x^{(2)}$, the differential operator

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \frac{\partial}{\partial x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{\partial}{\partial y}$$

has the symbol

$$\hat{P}(i\boldsymbol{\omega}) = i \begin{bmatrix} \omega^{(1)} & \omega^{(2)} \\ \omega^{(2)} & \omega^{(1)} \end{bmatrix}.$$

The two theorems above now hold exactly as stated with the generalized definition of $\omega \to \boldsymbol{\omega}$.

The symbol in our example has purely imaginary eigenvalues, which implies that the Petrovski condition is satisfied with $\alpha = 0$. Furthermore, since $\hat{P}(i\boldsymbol{\omega})$ is skew-Hermitian, i.e., $(\hat{P}^*(i\boldsymbol{\omega}) = -\hat{P}(i\boldsymbol{\omega}))$, it can be diagonalized by a unitary matrix. This implies that the conditions of Theorem 2.2 are satisfied (with $K = 1$ and $\alpha = 0$), which makes the Petrovski condition sufficient for well-posedness.

General first order systems have the form

$$\frac{\partial u}{\partial t} = \sum_{\nu=1}^{d} A_\nu \frac{\partial u}{\partial x^{(\nu)}}\,, \tag{2.11}$$

and they are quite common in applications. They are called *hyperbolic* if the symbol

$$\hat{P}(i\omega) = i \sum_{\nu=1}^{d} A_\nu \omega^{(\nu)}$$

has real eigenvalues and can be diagonalized by a matrix $T(\boldsymbol{\omega})$ with bounded condition number. Obviously, the Petrovski condition is satidfied for such systems.

If the PDE system doesn't have constant coefficients A_ν, then the Fourier analysis cannot be applied in a straightforward way as was done above. If $A_\nu = A_\nu(x)$, and the Fourier series is formally inserted, we get equations where the coefficients $\hat{u}(\boldsymbol{\omega})$ depend also on x, and it doesn't lead anywhere. The analysis can still be based on Fourier technique, but the theory becomes much more involved (see [Hörmander, 1985]), and we don't discuss it further here. (For difference approximations we shall briefly indicate what can be done.)

2.2.2 Difference Approximations

The discretization in time is done on a uniform grid $t_n = nk$, $n = 0, 1, \ldots$, where k is the step size. Consider first the classic simple approximation of (2.4)

$$\begin{aligned} u_j^{n+1} &= Qu_j^n\,, \quad j = 0, 1, \ldots, N\,, \\ u_j^0 &= f_j\,, \quad j = 0, 1, \ldots, N\,, \end{aligned} \tag{2.12}$$

where $Q = I + kD_+D_-$. The solution can be expanded in a finite Fourier series

$$u_j^n = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{u}_\omega^n e^{i\omega x_j}\,,$$

where, for convenience, it is assumed that N is even. The coefficients are obtained by

$$\hat{u}_\omega^n = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{N} u_j^n e^{-i\omega x_j} h,$$

which is called the *Discrete Fourier Transform* (DFT), often called the *Fast Fourier Transform* (FFT), which refers to the fast algorithm for computing it. The Fourier series is plugged into (2.12), and since the grid functions $\{e^{i\omega x_j}\}_{\omega=-N/2}^{N/2}$ are linearly independent, we obtain

$$\hat{u}_\omega^{n+1} e^{i\omega x_j} = Q\hat{u}_\omega^n e^{i\omega x_j} = (I + kD_+D_-)\hat{u}_\omega^n e^{i\omega x_j} = (1 - 4\lambda \sin^2 \frac{\xi}{2})\hat{u}_\omega^n e^{i\omega x_j}\,,$$

where $\xi = \omega h$, $\lambda = k/h^2$. The function $\hat{Q}(\xi) = 1 - 4\lambda \sin^2 \frac{\xi}{2}$ is called the (discrete) Fourier transform of the difference operator Q, and we have

$$\hat{u}_\omega^{n+1} = \hat{Q}\hat{u}_\omega^n\,. \tag{2.13}$$

Instead of having a difference operator acting on the whole grid function u_j^n, we have obtained a very simple scalar equation for each Fourier component. The obvious condition for nongrowing solutions is

$$|\hat{Q}(\xi)| \le 1\,, \quad |\xi| \le \pi\,, \tag{2.14}$$

and it is satisfied if and only if $\lambda \le \frac{1}{2}$.

Going back to the physical space, we introduce the discrete norm

$$||u^n||_h^2 = \sum_{j=0}^{N} |u_j^n|^2 h. \tag{2.15}$$

In analogy with Lemma 2.1 we have

Lemma 2.2. (*The* **discrete Parseval's relation**) *Let v_j be represented by its Fourier series*

$$v_j = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{v}_\omega e^{i\omega x_j} .$$

Then

$$||v||_h^2 = \sum_{\omega=-N/2}^{N/2} |\hat{v}_\omega|^2 . \qquad \square$$

By using the discrete Parseval's relation it follows from (2.14) that

$$||u^{n+1}||_h^2 = \sum_{\omega=-N/2}^{N/2} |\hat{u}_\omega^{n+1}|^2 = \sum_{\omega=-N/2}^{N/2} |\hat{Q}\hat{u}_\omega^n|^2 \le \sum_{\omega=-N/2}^{N/2} |\hat{u}_\omega^n|^2 = ||u^n||_h^2 ,$$

and by repeating this inequality for decreasing n, we obtain the final estimate

$$||u^n||_h \le ||f||_h$$

in analogy with the continuous case.

Let us next consider the same problem, but with a lower order term added:

$$u_t = u_{xx} + \alpha u , \quad \alpha > 0 .$$

The difference scheme is (2.12), but now with $Q = I + kD_+D_- + \alpha kI$. By doing the same analysis as above, we arrive at

$$\hat{u}^{n+1} = (1 - 4\lambda \sin^2 \frac{\xi}{2} + \alpha k)\hat{u}^n , \quad |\xi| \le \pi ,$$

and the best estimate we can obtain for all ξ is

$$|\hat{u}^{n+1}| \le (1 + \alpha k)|\hat{u}^n| .$$

This leads to

$$||u^n||_h^2 \le (1 + \alpha k)^{2n} ||f||_h^2 \le e^{2\alpha nk} ||f||_h^2 = e^{2\alpha t_n} ||f||_h^2 .$$

Referring back to the discussion of well-posedness above, this growth corresponds to the growth of the solution to the differential equation itself for the special case $f =$const., i.e., $\hat{f}(\omega) = 0$ for $\omega \neq 0$. The solution to the differential equation is $u = e^{\alpha t}$ if $f \equiv 1$. Hence, we cannot expect any better estimate.

In order to include such lower order terms in the class of problems we want to solve, it is reasonable to generalize the stability definition to

$$||u^n||_h \le e^{\alpha t_n} ||f||_h .$$