TEL-AVIV UNIVERSITY
The Raymond and Beverly Sackler Faculty of Exact Sciences
School of Mathematical Sciences

Strict Stability of High-Order Compact Implicit Schemes - The Role of Boundary Conditions for Hyperbolic PDEs

Thesis submitted for the degree "Doctor of Philosophy"

by

Alina Chertock

Submitted to the Senate of Tel-Aviv University
November 1998

This work was carried out under the supervision of

Professor Saul S. Abarbanel

This work is dedicated to the memory of my late daughter

ANAT

I would like to express my deepest thanks and sincere appreciation to my supervisor Professor Saul S. Abarbanel for his valuable guidance, continuous encouragement and for his understanding.

I wish to express my appreciation to Dr. S. Tsynkov for many useful suggestions and complete readiness to help.

I would like to thank my colleagues at Tel-Aviv University. In particular to thank Amir Yefet for hours of constructive discussions and for his friendship.

A very special thanks goes to my husband Boris and my son Daniel whose love, help and encouragement have played an essential role in my ability to perform this work.

I would also like to thank my parents and my parents-in-law for standing by me and for their firm belief in my success.

Finally, I would like to thank the Josef Buchmann Doctoral Scholarship Fund for partial support of this research.

Contents

Introduction

This thesis is concerned with the construction of implicit high-order finite-difference numerical schemes, for hyperbolic initial boundary value problems (IBVPs). Examples of computational problems in which low order finite difference methods (second order or less) are not accurate enough include acoustic, the propagation and scattering of electromagnetic waves and fluid dynamics. The advantage of high-order finite difference methods is two-fold: they allow either to increase the accuracy while keeping number of mesh points fixed or reduce the computational cost by decreasing the grid dimension while preserving the accuracy, see [17], [29].

However, in practical computing the schemes most widely used nowadays are still of the first or second order of accuracy. The fact that high-order methods are not used routinely is due to the fact that these schemes demand a more complex treatment of the statement of the "physical" and numerical boundary conditions. To retain the formal accuracy of a high-order scheme, boundary closures must be accomplished with accuracy at most one order less than the interior scheme [9], [10]. On a Cartesian mesh, it is always possible to derive non-symmetrical boundary operators that fulfill the boundary conditions and maintain the overall accuracy of the scheme. The difficulty is to derive high accurate and *stable* operators. In some cases high-order interior schemes have been used in combination with lower-order boundary conditions, as no high-order stable boundary formulation were available (see for example [22]). In so doing, the overall order of accuracy of the scheme can exceed the order of accuracy of the boundary conditions by no more than one. This makes it somewhat hard to justify the additional computation expenses that are associated with the use of high-order methods.

Studying the numerical stability of a fully discrete approximation for a linear hyperbolic partial differential equation is a difficult task even in one space dimension. Using the Laplace transform technique, Godunov and Ryabenkii [8] obtained necessary conditions for stability, analogous to the Von Neumann necessary condition for pure initial value problem. Kreiss established the sufficient condition for stability [15] and developed the complete stability theory for dissipative schemes [16]. His results, including some implicit and non-dissipative

methods were expanded by Osher [27], who derived Kreiss' sufficient conditions for stability as a corollary to a more general result. Gustafsson, Kreiss and Sundström [12] presented a general stability theory, for hyperbolic IBVPs, based on normal mode analysis, for the fully discrete case. According to their theory (known as G-K-S stability theory), in order to ensure stability of the finite domain problem it is sufficient to show that the inner scheme is Cauchy stable on $(-\infty, \infty)$, and that each of the two semi-line problems is stable using the normal mode analysis. For each semi-line problem, a necessary and sufficient condition for stability of the IBVP is that there is no eigensolution. Later Strikwerda generalized this theory to the semi-discrete case [34]. He showed that by using method of line approach in which hyperbolic systems are discretized in space but the time is left continuous, the necessary and sufficient conditions for stability are analogous to those obtained for finite-difference equations by Gustafsson, Kreiss and Sundström in the fully discrete case. The stability for the fully discrete approximation then follows, under mild assumptions (see Kreiss and Wu [20] or Levy and Tadmor [23]), from the stability for the semi-discrete approximation, if Runge-Kutta or other multi step time marching are used. Examples of several high-order schemes that are G-K-S stable can be found in works by Gary [7] and Sjögreen [30]. In these schemes, one-sided difference operators at the points close to the boundary are used to approximate the space derivatives of the differential equation.

An important concept in the analysis is the notion of strong stability. An approximation is strongly stable if the solution, including the values at the boundary points, can be estimated and bounded in terms of all data in the problem: the forcing function, initial data and boundary data. Stability analysis based on the Laplace transform method leads to strong stability if the Kreiss condition is satisfied as shown in [11]. In this book, there is also a complete analysis of a semi-discrete explicit fourth-order approximation based on the standard five-point scheme, which satisfies the Kreiss condition with various boundary conditions. This scheme was later generalized to general order of accuracy $2r$ by Strand, see [32], [33]. Stability analysis of an implicit difference operator for the IBVP was presented in [13], where Gustafsson and Olsson proved strong stability for a fourth-order scheme.

Another type of stability is strict stability, which means that the energy dissipation introduced by the boundaries is essentially preserved by the numerical scheme. In the case of semi-discrete approximations, strict stability implies that for a fixed mesh size h , all eigenvalues of the coefficient matrix of the correspondent system of ordinary differential equations have non-positive real part. For calculations over long time intervals, strict stability is especially important because it prevents exponential growth in time of the error for a fixed mesh size h.

In [24], [33] strictly stable high-order explicit finite difference approximations for IBVPs,

which satisfy an energy estimate, are computed based on the work by Kreiss and Scherer [28], [18]. In these schemes, the growth rates of the analytic and numerical solution are identical. Strict stability is obtained by constructing discrete operators that satisfy summation by parts formula, which imitates the integration by parts formula in the continuous case.

A series of compact high-order strictly stable as well as G-K-S stable schemes was constructed by Carpenter, Gottlieb and Abarbanel [4]. Compact implicit difference schemes are built by using an approximation $\frac{\partial}{\partial x} \rightarrow P^{-1}Q$, where P and Q are non-diagonal difference operators. In spite of additional work required for solving the banded systems, the advantage of these schemes in comparison with explicit difference schemes is that they have a much smaller "error constant". The stability characteristic of different compact fourth- and sixth-order spatial operators were appraised in this work, using the theory of Gustafsson, Kreiss and Sundström (G-K-S theory) for the semi-discrete IBVP. It was shown that many of high-order scalar schemes, which are G-K-S stable, were not strictly stable. Moreover, it was recently found that many high-order schemes, which are strictly stable in the scalar case, exhibit time divergence when they are applied to systems of equations. The underlying reason for the error growth in time is improper imposition of numerical boundary conditions.

In [19], Kreiss and Scherer presented a way to impose analytic boundary conditions by adding a projection to the semi-discrete system. They derived stability results for various explicit difference operators, approximating hyperbolic partial differential equations in several space dimensions, which satisfy a summation by parts rule. Generalizing this technique, Olson [25], [26] proved strict stability for a larger class of finite difference operators than those considered in [19]. Strand [33] obtained the stability results for explicit high-order finite difference approximations using the G-K-S stability theory for the semi-discrete IBVPs. To close the scheme near the boundary he obtained extra boundary conditions by extrapolating the outgoing characteristic variables and by differentiating the analytic boundary conditions and using the partial differential equation for the incoming characteristic variables. However, in some cases the approximation with such boundary conditions had eigenvalues with positive real part and and in order to assure the time stability of the scheme the numerical boundary conditions were modified by adding dissipative terms into the inflow part of the boundary conditions.

For compact high-order difference schemes for one-dimensional hyperbolic systems Carpenter, Gottlieb and Abarbanel [5] introduced a new procedure for imposing boundary conditions so as to ensure strict stability, when using difference operators satisfying a generalized summation by parts property. This methodology is based on solving the differential equation everywhere, including the boundary points. In so doing, the semi-discrete scheme is modified by adding a so called Simultaneous Approximation Term, SAT for short, that takes the

boundary information into account and does not destroy the overall accuracy of the scheme. Using this technique a time stable as well as G-K-S stable fourth-order implicit scheme, which is tridiagonal both on the implicit and on the explicit side, was constructed. This procedure is reminiscent of the usage of penalty terms.

Before proceeding further to the description of our own methodology let us, give here a brief account of the SAT method presented in work by Carpenter, Gottlieb and Abarbanel [5].

Consider the scalar hyperbolic equation

$$\frac{\partial u}{\partial t} + \lambda \frac{\partial u}{\partial x} = 0, \qquad 0 \le x \le 1, \quad t \ge 0, \lambda > 0,$$

with the following boundary and initial conditions:

$$\begin{aligned} u(0,t) &= g(t), \qquad t \ge 0, \\ u(x,0) &= f(x) \qquad 0 \le x \le 1. \end{aligned}$$

For a compact spatial operator, the approximation to the first derivative can be written as

$$P\frac{\partial \vec{v}}{\partial x} = Q\vec{v} \tag{0.0.1}$$

where $\vec{v}$ the vector of the unknowns $(v_0, \ldots, v_N)^T$ corresponding to the grid points $x_0, \ldots, x_N$, and P and Q are matrices, which satisfy the conditions:

1. Equation (0.0.1) is accurate to order m.
2. The matrix P has a simple structure, which is easily invertible.

These two assumptions are common to any useful compact scheme. In order to construct difference approximations for $\frac{\partial}{\partial x}$, which satisfy a generalized summation by parts formula, additional conditions have to be satisfied: there exists a matrix H such that

- the matrix $W = HP$ is a symmetric positive definite matrix;
- the matrix $V = HQ$ is almost skew-symmetric, i.e. $(V + V^T)/2$ has only two elements: v_{00} and v_{NN} $(\mathrm{v}_{00} < 0 < \mathrm{v}_{NN})$.

Applying the difference operator $D_x = P^{-1}Q$ on the differential equation gives a semi-discrete ODE system:

$$P\frac{d\vec{v}}{dt} = -\lambda Q\vec{v} \tag{0.0.2}$$

with boundary conditions not yet incorporated. However, as was mentioned above, the imposition the boundary conditions may destroy the summation by parts property, resulting in an unwanted exponential growth. Unlike the standard procedure of satisfying the boundary conditions directly by setting $v_0 = g(t)$, the SAT method involves an indirect statement of the boundary conditions. It is accomplished by adding a term to the derivative operator, which is proportional to the difference between the discrete value v_0 and the boundary term $g(t)$, and rewriting the approximation (0.0.2) as follows:

$$P\frac{d\vec{v}}{dt} = -\lambda Q\vec{v} + \lambda\tau \mathrm{v}_{00}\vec{S}_0(v_0 - g(t)) \tag{0.0.3}$$

where

$$\vec{S}_0 = H^{-1}(1, 0, \ldots, 0)^T . \tag{0.0.4}$$

It should be noted again, that this equation is solved at all points including boundary points. It was shown in [5] that the SAT method of imposing the analytic boundary conditions did not degrade the overall accuracy of the original spatial approximation. It was also proved that if the spatial operator satisfied a generalized summation by parts energy norm, and the SAT boundary procedure was used then the resulting numerical discretization was strictly stable both for scalar case and for one-dimensional hyperbolic systems.

In the present work the method proposed by Carpenter, Gottlieb and Abarbanel in [5], for constructing time stable high-order finite-difference approximations, for hyperbolic initial boundary value problems (IBVPs), is generalized. Fourth- and sixth-order compact implicit finite-difference schemes are constructed, analyzed, and numerical experiments are performed.

This paper is organized in two parts: Part I discusses numerical methods for solving one- and two-dimensional problems in the scalar case and Part II discusses numerical methods for solving one- and two dimensional hyperbolic systems.

In Part I, Chapter 1 starts with a consideration of the scalar hyperbolic IBVP. Approximating $\frac{\partial}{\partial x} \to D_x = P^{-1}Q$, we change some conditions which the matrices P and Q must satisfy. On the one hand, we assume as before that:

1. The approximation is m-order accurate;
2. The matrix P has a simple structure, which is easily invertible.

   On the other hand, we assume additionally that:

3. The matrix P is a symmetric positive definite matrix;

4. The matrix Q is almost skew-symmetric, except in $(n+1) \times (n+1)$ corners.

These properties of the matrices P and Q enable us to choose the matrix H as the identity matrix. This, in turns, (i) simplifies the construction of the approximation of desirable accuracy from the technical point of view, and (ii) allows us to extend this method to the solution of two-dimensional problems. The boundary conditions are imposed using the SAT boundary procedure with the extra SAT term modified accordingly. Most of the effort in constructing the scheme went into insuring that all eigenvalues of the coefficient matrix of the corresponding ODE system have a negative real part. We shall prove that the norm of the solution error vector, $||\epsilon||$, is bounded by a function of the time t, mesh size h, and the exact solution u, i.e. $||\epsilon|| < Kh^m t,\ K = K(u)$, indicating the convergence of the scheme for all $t > 0$ (and at most a linear temporal growth of the error). Numerical experiments performed in this chapter using fourth- and sixth-order schemes show a good agreement with theoretical results. The convergence rate predicted by the theory of Gustafsson [9], [10] is verified by doing a grid refinement study. The time stability of the scheme is illustrated by both computing the error for long time integrations and determining the eigenvalue spectrum for the semi-discrete system. The actual numerical solution had a temporal error bounded by a constant rather by a linear growth.

In Chapter 2, a numerical approximation for solving two-dimensional hyperbolic scalar problems in a rectangular domain is built by analogy with the one-dimensional case. Using the same differentiation matrices as in the one-dimensional case once in the x-direction and once more in the y-direction, we approximate $\frac{\partial}{\partial x} + \frac{\partial}{\partial y}$ by $D_x + D_y$. In order to ensure time stability of the scheme it is sufficient to show that all eigenvalues of the coefficient matrix have a negative real part. From the fact that each of the matrices has eigenvalues with a negative real part does not follow that the sum of these matrices will preserve this virtue. This implies that the matrix $D_x + D_y$ should be checked for stability. This is done by proving that $Re\,(u, (D_x + D_y)\,u)_H < 0 \quad \forall u \in R^{N^2}$ in some norm H.† Numerical experiments are performed on a hyperbolic model problem in two space-dimensions. The fourth- and sixth-order schemes are examined with respect to convergence rate and long time integrations. The results of numerical simulations agree well with the theoretical results.

Part II is devoted to solving one- and two-dimensional hyperbolic systems. In Chapter 3, the methodology presented in Chapter 1 is adopted to accommodate partially reflecting boundary conditions and to solve the one-dimensional hyperbolic system. As was mentioned

†Of course, if D_x, D_y were negative definite matrices, we would not have paid special attention to the sum of these matrices. Solving IBVPs on irregular domain, Abarbanel and Ditkowski [1] constructed a differentiation matrix whose symmetric part is negative definite. Technically, it was a very difficult task as, in particular, confirmed by the fact that they succeeded to construct only second order explicit approximation for $\partial/\partial x$ and fourth-order explicit approximation for $\partial^2/\partial x^2$.

above, the time stability in the scalar case does not imply the time stability for systems. Despite the fact that for hyperbolic systems we succeeded in proving the time stability only for some special cases, numerical examples show that the method is effective and provide time stability even when a theoretical foundation is lacking. As in the scalar case, the fourth- and sixth-order schemes are used for solving model problems. The formal accuracy of each scheme is determined by doing a grid refinement study. The numerical results show that the convergence rate of the schemes used here agrees well with the theory. In order to investigate numerically if the schemes are time stable we compute the error for long time integrations, and additionally determine the eigenvalue spectrum of the semi-discrete system. In all cases, no eigenvalues with positive real part are found which indicate time stability of the schemes.

As an application where high-order accurate approximations are needed we consider in Chapter 4 the two-dimensional Maxwell's equations in free space. The SAT method used for the diagonalized system in 1-D is adopted to solve the two-dimensional system, which can not be digonalized. The problem is solved using both the fourth- and the sixth-order schemes. Numerical results are compared with those obtained by E. Turkel and A. Yefet in [35], [36]. They solved the same problem by using the Ty(2,4) scheme, which is a fourth-order compact implicit difference scheme on staggered meshes.

In Appendixes A and B, a way to construct sixth- and fourth-order compact implicit difference schemes, which satisfy all above mentioned conditions, is described in detail. It should be observed here that the construction of such schemes is technically a very difficult task, especially in the case of the sixth-order scheme.

Part I

The Scalar Case

Chapter 1

1-D Hyperbolic Equations

1.1 Description of the method and proof of the main theorem

We consider the scalar hyperbolic equation

$$\frac{\partial u}{\partial t} + \lambda \frac{\partial u}{\partial x} = 0, \qquad 0 \le x \le 1, \qquad t \ge 0 \tag{1.1.1}$$

with initial conditions prescribed at $t = 0$,

$$u(x,0) = f(x) \qquad 0 \le x \le 1. \tag{1.1.2}$$

For positive λ we have the boundary condition:

$$u(0,t) = g(t), \qquad t \ge 0. \tag{1.1.3}$$

We want to solve the above problem by finite difference approximations. In this work, we will deal with compact schemes for the discretization of the spatial operator $\frac{\partial}{\partial x}$. We therefore introduce the mesh width h, and divide the interval $[0,1]$ into subintervals of length h. We use with $j = 0, \ldots, N$ and $N = 1/h$ the notation

$$x_j = jh, \qquad u_j(t) = u(x_j, t), \tag{1.1.4}$$

where $u_j(t)$ is the projection of the exact solution $u(x,t)$ unto the grid. We denote by $\vec{u}$ the vector $(u_0(t), \ldots, u_N(t))^T$ and by $\vec{v}$ the numerical approximation to the projection $\vec{u}$.

The implicit approximation for the first derivative can be written as

$$P \frac{\partial \vec{v}}{\partial x} = Q\vec{v} \tag{1.1.5}$$

where $P = (p_{ij})$ and $Q = (q_{ij})$ are $(N+1) \times (N+1)$ Teoplitz matrices with small perturbations at the corners due to the boundary conditions (a detailed discussion regarding

the construction of these matrices is given in Appendixes A and B). Using (1.1.5), we may write the following approximation for (1.1.1)

$$P\frac{d\vec{v}}{dt} = -\lambda Q\vec{v} \tag{1.1.6}$$

In order to satisfy the analytic boundary conditions, (1.1.3), we use the SAT methodology introduced in [5] that involves an indirect treatment of the boundary conditions. Using this method we do not satisfy the boundary conditions directly by imposing $v_0 = g(t)$, but add to the derivative operator a term, which is proportional to the difference between the discrete value v_0 and the boundary term $g(t)$ and solve a derivative equation everywhere, including the boundary points. This approach will be elaborated later.

Throughout this work we make three main assumptions:
1. Equation (1.1.5) is accurate to order m, i.e

$$P\frac{\partial\vec{u}}{\partial x} = Q\vec{u} + P\vec{T}, \tag{1.1.7}$$

where $\vec{T}$ is the truncation error due to the numerical differentiation and

$$\|\vec{T}\| = O(h^m) \tag{1.1.8}$$

2. The matrix P is a symmetric positive definite matrix with simple structure which is easily invertible and there exist positive constants c_0, c_1 independent of N such that

$$c_0\| \vec{u} \|^2 \leq (P\vec{u}, \vec{u}) \leq c_1\| \vec{u} \|^2 \tag{1.1.9}$$

where $\| \vec{u} \|^2 = (\vec{u}, \vec{u})$, and c_1 is the largest eigenvalue of P, i.e. $c_1 = \| P \|$ because P is positive definite symmetric matrix.
3. The matrix Q is almost skew-symmetric except in $(n+1) \times (n+1)$ corners. It means

that

$$\frac{Q+Q^T}{2}=\begin{pmatrix}
\mathbf{q}_{00} & \dots & \mathbf{q}_{0n-1} & \mathbf{q}_{0n} & 0 & & & & & \\
\vdots & & \vdots & \vdots & \vdots & & & 0 & & \\
\mathbf{q}_{0n-1} & \dots & \mathbf{q}_{n-1n-1} & \mathbf{q}_{n-1n} & 0 & & & & & \\
\mathbf{q}_{0n} & \dots & \mathbf{q}_{nn-1} & \mathbf{q}_{nn} & 0 & & & & & \\
0 & \dots & 0 & 0 & 0 & & & & & \\
 & & & & & \ddots & & & & \\
 & & & & & & 0 & 0 & 0 & \dots & 0 \\
 & & & & & & 0 & \mathbf{q}_{ll} & \mathbf{q}_{ll+1} & \dots & \mathbf{q}_{lN} \\
 & & & & & & 0 & \mathbf{q}_{ll+1} & \mathbf{q}_{l+1l+1} & \dots & \mathbf{q}_{l+1N} \\
 & & 0 & & & & \vdots & \vdots & \vdots & & \vdots \\
 & & & & & & 0 & \mathbf{q}_{lN} & \mathbf{q}_{l+1N} & \dots & \mathbf{q}_{NN}
\end{pmatrix} \tag{1.1.10}$$

Actually we shall show in Appendixes A and B that the matrix $Q=(q_{ij})$ can be constructed in such a way that $n=1$, that is

$$\frac{Q+Q^T}{2}=\begin{pmatrix}
\mathbf{q}_{00} & \mathbf{q}_{01} & 0 & & & & & \\
\mathbf{q}_{01} & \mathbf{q}_{11} & 0 & & & 0 & & \\
0 & 0 & 0 & & & & & \\
 & & & \ddots & & & & \\
 & & & & 0 & 0 & 0 & \\
0 & & & & 0 & \mathbf{q}_{N-1N-1} & \mathbf{q}_{N-1N} & \\
 & & & & 0 & \mathbf{q}_{N-1N} & \mathbf{q}_{NN} &
\end{pmatrix} \tag{1.1.11}$$

$$= \begin{pmatrix} q_{00} & \frac{1}{2}(q_{01}+q_{10}) & 0 & & & \\ \frac{1}{2}(q_{01}+q_{10}) & q_{11} & 0 & & 0 & \\ 0 & 0 & 0 & & & \\ & & & \ddots & & \\ & & & 0 & 0 & 0 \\ & 0 & & 0 & q_{N-1N-1} & \frac{1}{2}(q_{NN-1}+q_{N-1N}) \\ & & & 0 & \frac{1}{2}(q_{NN-1}+q_{N-1N}) & q_{NN} \end{pmatrix}$$

We now rewrite the semi-discrete problem for $\vec{v}$ in the following form:

$$P\frac{d\vec{v}}{dt} = -\lambda Q\vec{v} + \lambda\vec{S}_0(v_0 - g(t)) \tag{1.1.12}$$

where

$$\vec{S}_0 = \begin{pmatrix} \tau q_{00} \\ q_{01}+q_{10} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{1.1.13}$$

Theorem 1.1.1 *The approximation (1.1.12),(1.1.13) preserves the order of accuracy - m of the spatial operator and is strictly stable under the following conditions on τ and the corner entries of the matrix Q:*

$$\begin{aligned} &(1-\tau)q_{00} \geq 0 \quad , \qquad q_{11} \geq 0, \\ &q_{NN}u_N^2 + (q_{N-1N}+q_{NN-1})u_N u_{N-1} + q_{N-1N-1}u_{N-1}^2 \geq 0, \quad \forall u_N, u_{N-1} \in \mathbf{R}. \end{aligned} \tag{1.1.14}$$

Proof. Denote as before by $\vec{u} = (u_0(t), \ldots, u_N(t))^T$, i.e the values of the true solution at the grid points and by $\vec{v}$ its numerical approximation . Combining the accuracy condition

found in assumption **1** , with equation (1.1.12) we may write

$$P\frac{d\vec{u}}{dt} = -\lambda Q\vec{u} + \lambda\vec{S}_0(u_0(t) - g(t)) + P\vec{T} \tag{1.1.15}$$

Note that $u_0(t) - g(t) = u(0,t) - g(t) = 0$. To get the equation for the solution error vector, $\vec{\epsilon}(t) = \vec{u}(t) - \vec{v}(t)$, we subtract (1.1.12) from (1.1.15):

$$P\frac{d\vec{\epsilon}}{dt} = -\lambda Q\vec{\epsilon} + \lambda\vec{S}_0\epsilon_0 + P\vec{T} \tag{1.1.16}$$

where $\epsilon_0 = v_0 - g(t) = v_0 - u_0$.
Taking the scalar product of $\vec{\epsilon}$ with (1.1.16) one gets:

$$\frac{1}{2}\frac{d}{dt}(P\vec{\epsilon},\vec{\epsilon}) = -\lambda(Q\vec{\epsilon},\vec{\epsilon}) + \lambda(\vec{S}_0\epsilon_0,\vec{\epsilon}) + (P\vec{T},\vec{\epsilon}) \tag{1.1.17}$$

We notice that $(Q\vec{\epsilon},\vec{\epsilon}) = ((Q+Q^T)\vec{\epsilon}/2,\vec{\epsilon})$ and that means that

$$\begin{aligned}(Q\vec{\epsilon},\vec{\epsilon}) &= q_{00}\epsilon_0^2 + (q_{01}+q_{10})\epsilon_0\epsilon_1 + q_{11}\epsilon_1^2 \\ &\quad + q_{NN}\epsilon_N^2 + (q_{N-1N}+q_{NN-1})\epsilon_{N-1}\epsilon_N + q_{N-1N-1}\epsilon_{N-1}^2\end{aligned} \tag{1.1.18}$$

From (1.1.13) follows that

$$(\vec{S}_0\epsilon_0,\vec{\epsilon}) = \tau q_{00}\epsilon_0^2 + (q_{01}+q_{10})\epsilon_0\epsilon_1 \tag{1.1.19}$$

Using (1.1.18),(1.1.19) in (1.1.17) one gets:

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}(P\vec{\epsilon},\vec{\epsilon}) &= -\lambda(1-\tau)q_{00}\epsilon_0^2 - \lambda q_{11}\epsilon_1^2 \\ &\quad - \lambda\Big[q_{NN}\epsilon_N^2 + (q_{N-1N}+q_{NN-1})\epsilon_{N-1}\epsilon_N + q_{N-1N-1}\epsilon_{N-1}^2\Big] + (P\vec{T},\vec{\epsilon})\end{aligned} \tag{1.1.20}$$

If we require (and achieve by construction) that

$$q_{NN}\epsilon_N^2 + (q_{N-1N}+q_{NN-1})\epsilon_{N-1}\epsilon_N + q_{N-1N-1}\epsilon_{N-1}^2 \geq 0$$

for all $\epsilon_N, \epsilon_{N-1} \in \mathbf{R}$, then for $(1-\tau)q_{00} \geq 0$ and $q_{11} \geq 0$, and define $\vec{T}_1 = 2\vec{T}$, the equation

(1.1.20) leads to the inequality

$$\frac{d}{dt}(P\vec{\epsilon},\vec{\epsilon}) \le (P\vec{T}_1,\vec{\epsilon}) \tag{1.1.21}$$

We now use the inequality

$$\left(P\vec{T}_1,\vec{\epsilon}\right) \le \sqrt{\left(P\vec{T}_1,\vec{T}_1\right)}\sqrt{(P\vec{\epsilon},\vec{\epsilon})} \tag{1.1.22}$$

to obtain

$$2\frac{d}{dt}\sqrt{(P\vec{\epsilon},\vec{\epsilon})} \le \sqrt{(P\vec{T}_1,\vec{T}_1)} \tag{1.1.23}$$

After integrating (1.1.23) and using (1.1.9) we get

$$\| \vec{\epsilon}\,\| \le \frac{1}{2}\sqrt{\frac{c_1}{c_0}} \sup_{0\le\tau\le t} \| \vec{T}_1(\tau)\,\|\, t \tag{1.1.24}$$

which proves the convergence of the scheme for all $t < \infty$ (and at most a linear temporal growth of the error)[‡]. The linear temporal bound on $\| \vec{\epsilon}\,\|$ is given by (1.1.24), shows that the scheme is not only Lax stable but also strictly stable. □

Remarks.

1. The construction of the matrices P and Q will be described in detail in Appendixes A and B. We note that if we succeeded in constructing the matrices P and Q then we know exactly the value of $q_{00},\ q_{11},\ q_{N-1N-1},\ q_{N-1N}+q_{NN-1},\ q_{NN}$. This implies that actually stability of the scheme (1.1.12), (1.1.13) depends only on τ. For example, for our sixth-order implicit scheme with five-order boundary closure the matrices P, Q were constructed in such a way that $q_{00}=-\frac{2}{3},\ q_{11}=\frac{1}{6},\ q_{N-1N-1}=\frac{1}{6},\ q_{N-1N}+q_{NN-1}=\frac{1}{3},\ q_{NN}=\frac{1}{3}$ and therefore the expression

$$q_{NN}\epsilon_N^2+(q_{N-1N}+q_{NN-1})\epsilon_{N-1}\epsilon_N+q_{N-1N-1}\epsilon_{N-1}^2$$

$$=\frac{1}{6}\epsilon_N^2+\frac{1}{3}\epsilon_{N-1}\epsilon_N+\frac{1}{6}\epsilon_{N-1}^2=\frac{1}{6}\left(\epsilon_{N-1}+\epsilon_N\right)^2+\frac{1}{6}\epsilon_N^2$$

[‡]Note that the behavior of ϵ with h depends on the smoothness of the solution. To maintain the order of the approximation we need $u(x,t)\in C^m$, where m is the order of accuracy. If, for example, the initial data contains only a first derivative this will degrade the behavior of $\|\vec{T}_1\|$ with h.

is positive for all $\epsilon_N, \epsilon_{N-1} \in \mathbf{R}$ and the scheme is strictly stable for $\tau \geq 1$, see (1.1.20).

2. For negative λ we have the boundary condition at $x = 1$:

$$u(1,t) = g(t), \qquad t \geq 0 \tag{1.1.25}$$

In this case we will write the following semi-discrete approximation for $\vec{v}$:

$$\widetilde{P}\frac{d\vec{v}}{dt} = -\lambda\widetilde{Q}\vec{v} + \lambda\vec{S}_N(v_N - g(t)) \tag{1.1.26}$$

where $(\widetilde{P})_{ij} = (P)_{N-i,N-j}$, $\quad (\widetilde{Q})_{ij} = -(Q)_{N-i,N-j} \quad$ for all $\quad 0 \leq i,j \leq N \quad$ and

$$\vec{S}_N = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -(q_{01} + q_{10}) \\ -\tau q_{00} \end{pmatrix} \tag{1.1.27}$$

Because of the Teoplitz structure of matrices P and Q it means that the matrices $\widetilde{P}$ and $\dfrac{\widetilde{Q} + \widetilde{Q}^T}{2}$ are almost identical to the matrices P and $\dfrac{Q + Q^T}{2}$. They differ only in the corners which are transformed in such a way that the matrix $\widetilde{P}$ still satisfies the conditions of assumption **2** with the same constants c_0, c_1 and the matrix $\dfrac{\widetilde{Q} + \widetilde{Q}^T}{2}$ is of the form

$$\frac{\widetilde{Q} + \widetilde{Q}^T}{2} = \begin{pmatrix} \tilde{q}_{00} & \frac{1}{2}(\tilde{q}_{01} + \tilde{q}_{10}) & 0 & & & \\ \frac{1}{2}(\tilde{q}_{01} + \tilde{q}_{10}) & \tilde{q}_{11} & 0 & & 0 & \\ 0 & 0 & 0 & & & \\ & & & \ddots & & \\ & & & 0 & 0 & 0 \\ & 0 & & 0 & \tilde{q}_{N-1N-1} & \frac{1}{2}(\tilde{q}_{NN-1} + \tilde{q}_{N-1N}) \\ & & & 0 & \frac{1}{2}(\tilde{q}_{NN-1} + \tilde{q}_{N-1N}) & \tilde{q}_{NN} \end{pmatrix}$$

$$= \begin{pmatrix} -q_{NN} & -\frac{1}{2}(q_{N-1N}+q_{NN-1}) & & \\ -\frac{1}{2}(q_{N-1N}+q_{NN-1}) & -q_{N-1N-1} & 0 & \\ & \ddots & & \\ & 0 & -q_{11} & -\frac{1}{2}(q_{01}+q_{10}) \\ & & -\frac{1}{2}(q_{01}+q_{10}) & -q_{00} \end{pmatrix}$$

and we recall that $(1-\tau)q_{00} \geq 0,\ q_{11} \geq 0$ and the expression $\tilde{q}_{00}u_0^2+(\tilde{q}_{01}+\tilde{q}_{10})u_0u_1+\tilde{q}_{11}u_1^2 = q_{NN}u_0^2+(q_{N-1N}+q_{NN-1})u_0u_1+q_{NN}u_1^2$ is positive for all $u_0, u_1 \in \mathbf{R}$.

3. Sometimes it is useful to rewrite the approximation (1.1.12) in the following matrix form

$$P\frac{d\vec{v}}{dt} = -\lambda \mathbf{Q}\vec{v} - \lambda \vec{S}_0 g(t), \tag{1.1.28}$$

where $\mathbf{Q}$ and S are $(N+1)\times(N+1)$ matrices defined by

$$\mathbf{Q} = Q - S, \qquad S = \begin{pmatrix} \tau q_{00} & 0 & \dots & 0 \\ q_{01}+q_{10} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & 0 & \\ 0 & & & \end{pmatrix} \tag{1.1.29}$$

and the vector $\vec{S}_0$ is defined by (1.1.13).

Note that all the boundary information is incorporated into the matrix $\mathbf{Q}$ and that the time-stability of the numerical scheme (1.1.28) depends directly on the properties of this matrix.

It should be also observed that if the inequalities (1.1.14) hold then the matrix $\mathbf{Q}$ is positive definite, that is

$$(\vec{v}, \mathbf{Q}\vec{v}) = \frac{1}{2}\left(\vec{v}, \left(\mathbf{Q}+\mathbf{Q}^T\right)\vec{v}\right) \geq 0 \quad \forall \vec{v} \in \mathbf{R}^N.$$

And it follows that real part of each eigenvalue of the matrix $P^{-1}\mathbf{Q}$ is positive. One can verify this by writing:

$$P^{-1}\mathbf{Q} = P^{-\frac{1}{2}}\left(P^{-\frac{1}{2}}\mathbf{Q}P^{-\frac{1}{2}}\right)P^{\frac{1}{2}}$$

which means that the matrices $P^{-1}\mathbf{Q}$ and $P^{-\frac{1}{2}}\mathbf{Q}P^{-\frac{1}{2}}$ are similar and therefore have the same eigenvalues. And since P is a positive definite symmetric matrix (and therefore also the matrices $P^{\frac{1}{2}}$ and $P^{-\frac{1}{2}}$ are positive definite matrices), the matrix $P^{-\frac{1}{2}}\mathbf{Q}P^{-\frac{1}{2}}$ satisfies

$$\left(\vec{v}, \left(P^{-\frac{1}{2}}\mathbf{Q}P^{-\frac{1}{2}}\right)\vec{v}\right) = \left(P^{-\frac{1}{2}}\vec{v}, \mathbf{Q}\left(P^{-\frac{1}{2}}\vec{v}\right)\right) \geq 0 \quad \forall \vec{v} \in \mathbf{R}^N.$$

The last inequality implies that the real part of each eigenvalue of $P^{-\frac{1}{2}}\mathbf{Q}P^{-\frac{1}{2}}$ and therefore of $P^{-1}\mathbf{Q}$ is positive.

In the next section we show a graphical representation of this fact. Figure (1.5), (1.6) show the eigenvalue spectrum of $-P^{-1}\mathbf{Q}$ for fourth-order and sixth-order approximation respectively for various grids. All eigenvalues of these matrices ($N = 20, 40, 60, 80$) are distinct and no eigenvalues with positive real part exist.

4. In a similar fashion, if we define

(1.1.30) $$\widetilde{\mathbf{Q}} = \tilde{Q} - \tilde{S},\ \tilde{S} = \begin{pmatrix} & & 0 \\ & 0 & \vdots \\ & & 0 \\ 0 & \dots\ 0 & -(q_{01}+q_{10}) \\ 0 & \dots\ 0 & -\tau q_{00} \end{pmatrix},\ \vec{S}_N = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -(q_{01}+q_{10}) \\ -\tau q_{00} \end{pmatrix}.$$

we can rewrite the approximation (1.1.26) for $\lambda < 0$ as follows:

(1.1.31) $$\tilde{P}\frac{d\vec{v}}{dt} = -\lambda\widetilde{\mathbf{Q}}\vec{v} - \lambda\vec{S}_N g(t),$$

In this case it can be shown that the matrix $\widetilde{\mathbf{Q}}$ is negative definite and all eigenvalues of the matrix $\tilde{P}^{-1}\widetilde{\mathbf{Q}}$ have negative real part.

1.2 Numerical results

In this section we consider the scalar model problem

$$u_t(x,t) + u_x(x,t) = 0, \qquad 0 \le x \le 1, \qquad t \ge 0 \tag{1.2.1}$$

$$u(x,0) = f(x), \qquad 0 \le x \le 1, \tag{1.2.2}$$

$$u(0,t) = g(t), \qquad t \ge 0 \tag{1.2.3}$$

with $f(x) = \sin \omega x, \quad g(t) = -\sin \omega t.$

The exact solution is

$$u(x,t) = \sin \omega (x-t) \qquad 0 \le x \le 1, \qquad t \ge 0. \tag{1.2.4}$$

In order to highlight the difference in the quality of results obtained using standard and SAT-type boundary conditions, we solve the scalar model equation using both types of boundary conditions.

To solve the model problem (1.2.1), (1.2.2), (1.2.3) we use two different difference operators: fourth-order compact and six-order compact (see Appendixes A and B for details). Here the order of the difference operator refers to the order of the global accuracy that the theory of Gustafsson [9], [10] predicts. There it is proved that in our case boundary conditions of at least order $m-1$ must be imposed to retain mth-order global accuracy. Therefore we use a fourth-order difference operator which is of order three at the boundary and order four in the interior, and a sixth-order difference operator of order five at the boundary and order six in the interior. The standard fourth-order Runge-Kutta method is used for time integration in the case of the fourth-order difference operator and a sixth-order Runge-Kutta method (developed by Butcher [2], [3]) is used in case of the sixth-order difference operator. The time step is chosen small enough to ensure the local stability of the Runge-Kutta method. In the case of conventional implementation of boundary conditions we overwrite the value of the solution at the boundary point with the analytic boundary condition at the end of each Runge-Kutta stage.

Conventional boundary conditions. Table (1.1) shows a grid convergence study for both spatial discratizations. The absolute error $\log_{10}(L_2)$ at a fixed time $t = \mathrm{T}$ and the convergence rate between two grids are plotted. The convergence rate is computed as

$$\log_{10}\left(\frac{\| u - v^{h_1}\|_2}{\| u - v^{h_2}\|_2}\right) / \log_{10}\left(\frac{h_1}{h_2}\right), \tag{1.2.5}$$

where $u = (u(x_0,t), u(x_2,t), \ldots, u(x_N,t))^T$ is the projection of the exact solution, v^h is the numerical solution with mesh width h, and $\| u - v^h\|_2$ is the discrete L_2 norm of the absolute error.

We see in this table that for relative short time integration ($T = 0.5$) the convergence rate of sixth-order scheme is approximately 6. The convergence rate of fourth-order scheme asymptotes to the theoretical value of 4. For the schemes to be strictly stable no eigenvalues with positive real part are allowed to exist. Therefore we investigated numerically if the schemes are strictly stable by both measuring the error for long time integration and computing eigenvalues of the ODE system obtained after semi-discretization.

In this case of imposing conventional boundary conditions a system of ODE's results having the form

$$P\frac{dv_j^h}{dt} = -Qv_j^h, \quad j = 0, \ldots, N \tag{1.2.6}$$

Noting that the physical boundary condition $g(t)$ will be imposed at the grid point $j = 0$, the last equation can be rewritten as

$$\widehat{P}\frac{dv_j^h}{dt} = -\widehat{Q}v_j^h + B_j g(t), \quad j = 1, \ldots, N \tag{1.2.7}$$

where $\widehat{P}, \widehat{Q}$ are $N \times N$ matrices and B_j is the "j^{th}" term of the vector $\vec{B} = (q_{10}, q_{20}, \ldots, q_{N0})^T$ which gives the dependence of the "j^{th}" scheme on the boundary data.

Figures (1.1),(1.2) show the error as a function of time for the fourth-order compact scheme and the sixth-order compact scheme respectively for different grids. Clearly there is an exponential growth in time for the sixth-order scheme, but not for the fourth-order one. Figures (1.4)-(1.3) show the semi-discrete eigenvalues spectrum of the ODE system, i.e. the eigenvalues of the matrix $-\widehat{P}^{-1}\widehat{Q}$ defined above. In figure (1.4) we see that for the fourth-order scheme there are no eigenvalues with positive real part. It will be seen later that this fortuitous situation fails when one consider the case of a system of equations rather then the scalar partial differential equation. In figure (1.3) we see that the eigenvalue spectra of the ODE system for the sixth-order scheme stretches into the right half plane and since the exponential growth is caused by the eigenvalues having positive real part we get the unwanted growth. The time divergence seen in the sixth-order scheme is a result of imposing the conventional boundary conditions.

	Fourth-order compact		Sixth-order compact	
Grid	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate
21	-2.798		-3.510	
31	-3.431	3.60	-4.535	5.82
41	-3.901	3.76	-5.331	6.37
61	-4.580	3.86	-6.408	6.12
81	-5.069	3.91	-7.169	6.09

Table 1.1: Grid convergence of two high-order schemes on $u_t + u_x = 0$, using conventional implementation of boundary conditions. Here $\omega = 2\pi$, CFL = 0.1, T = 10 for the fourth-order scheme and $T = 0.5$ for the sixth-order scheme.

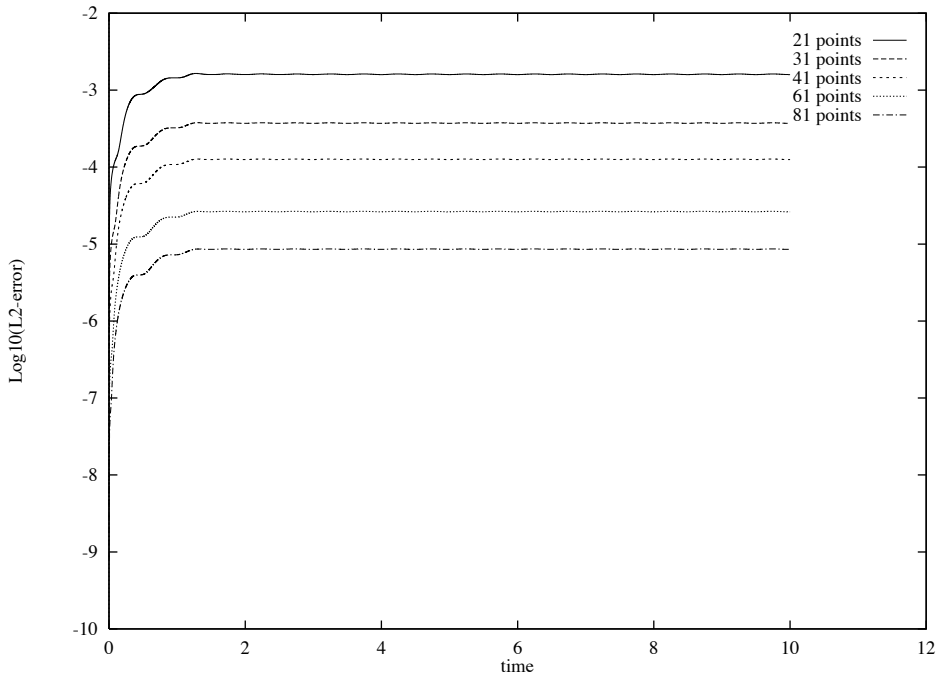

Figure 1.1: The L_2-error as a function of time for the fourth-order approximation using conventional implementation of boundary conditions with CFL $= 0.1$, $\omega = 2\pi$.

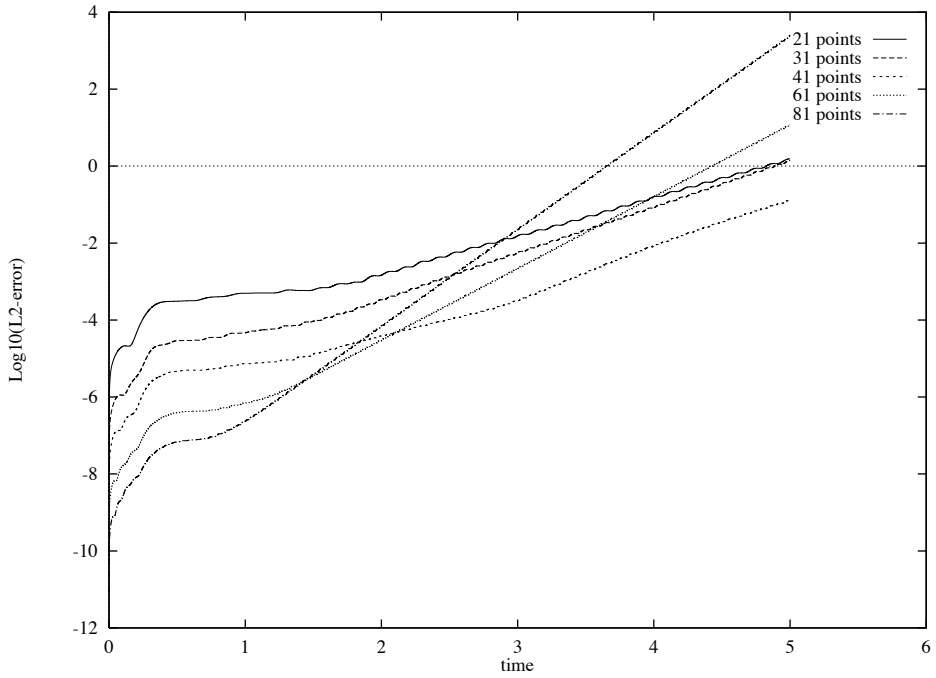

Figure 1.2: The L_2-error as a function of time for the sixth-order approximation using conventional implementation of boundary conditions with CFL $= 0.1$, $\omega = 2\pi$.

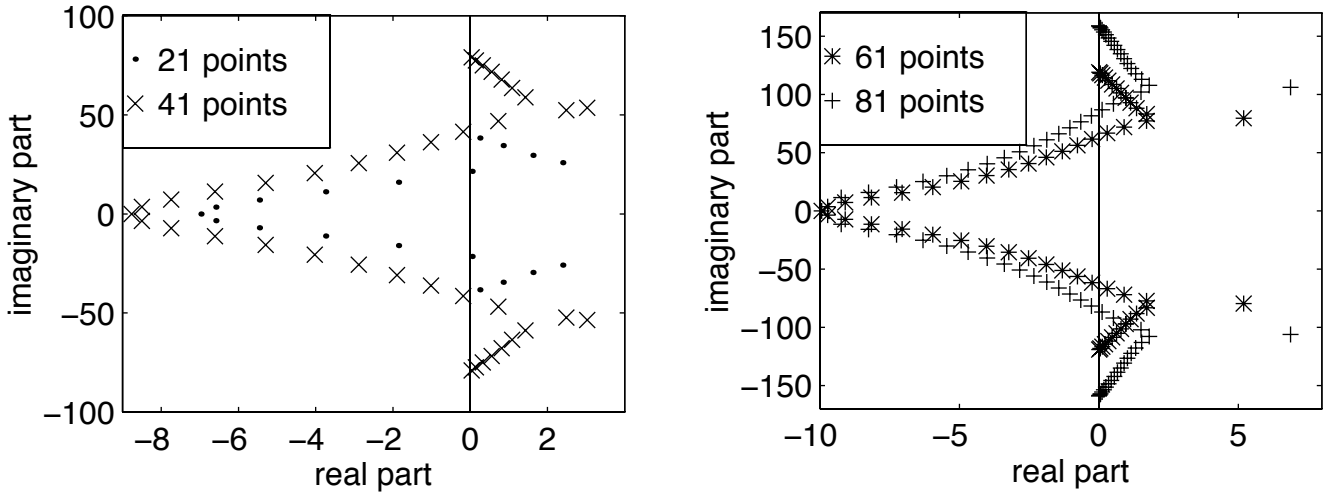

Figure 1.3: Magnification of semi-discrete eigenvalue spectrum close to imaginary axis for the sixth-order approximation using conventional implementation of boundary conditions.

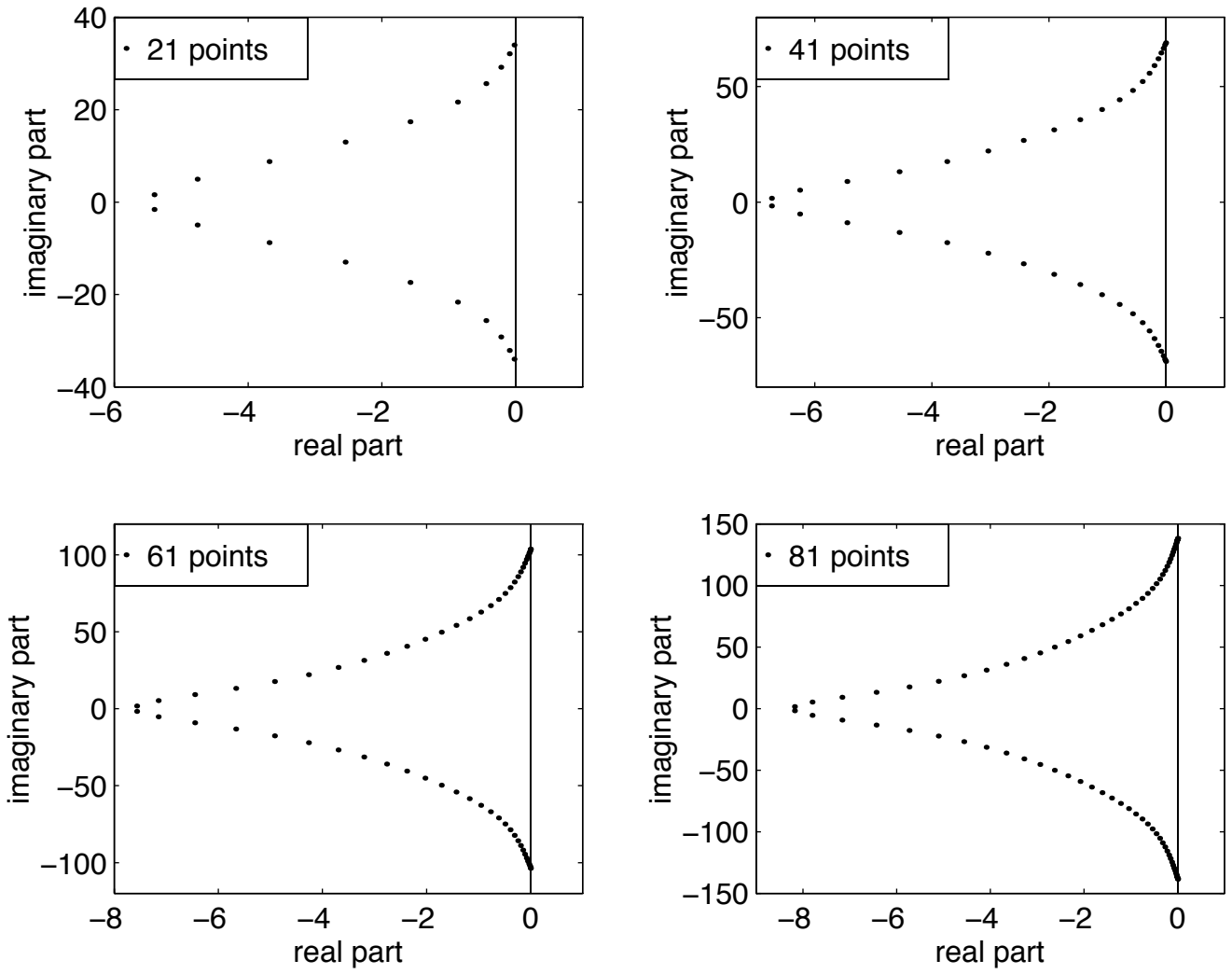

Figure 1.4: Semi-discrete eigenvalue spectrum for the fourth-order approximation using conventional implementation of boundary conditions.

SAT boundary conditions. We now solve the model problem (1.2.1), (1.2.2), (1.2.3) using SAT method for treating the boundary conditions.

Tables (1.2), (1.3) show a grid refinement study for the fourth-order and the sixth-order compact difference operators with different SAT parameters τ. As in the case of conventional boundary conditions we plot the absolute error $\log_{10}(L_2)$ at the time $t = T = 10$ (extracted from computations run to $T = 100$) and the convergence rate computed as in (1.2.5). We see that the SAT procedure for boundary treatment does not destroy the formal accuracy of spatial discretization. The numerical results agree well with the theory of Gustafsson [9], [10] and give the predicted accuracy. We can also see that the magnitude of the error is dependent on the value of the parameter τ.

It was proven in section 1.1 that semi-discrete approximation (1.1.12) and (1.1.13) obtained with the SAT method is strictly stable. From results of Kreiss and Wu [20] and Levi and Tadmor [23] follow that the fully discrete approximation is stable if a locally stable Runge-Kutta method is used for time integration. Again, the standard fourth-order Runge-Kutta method is used for time integration in the case of the fourth-order spatial difference operator and the sixth-order Runge-Kutta method (developed by Butcher [2], [3]) is used in case of the sixth-order spatial difference operator. The time step is chosen small enough to ensure the local stability of the Runge-Kutta method. Figures (1.7) - (1.8) show the error as a function of time for different SAT parameters τ , for different grids and CFL numbers. In all cases the error remained bounded for all grids and CFLs for time as large as $T = 100$. No exponential growth was found for the SAT method, indicating time stability. Figures (1.5)-(1.6) show semi-discrete eigenvalues spectrum for this method, i.e. the eigenvalues of the matrix $-P^{-1}\mathbf{Q}$ defined by (1.1.28), (1.1.29) (see remarks for section 1.1). As we can see in these figures no eigenvalues with positive real part exist.

We also solved the problem (1.2.1), (1.2.2), (1.2.3) for different values of ω. In figure (1.9), (1.10) we show the approximate solution of the problem computed at the time $t = 10$ using the sixth-order compact scheme with $\tau = 2$, CFL $= 0.1$, $\omega = 30\pi$ and the number of grid points $N = 80$.

	Fourth-order compact		Sixth-order compact	
Grid	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate
21	-3.480		-4.944	
31	-4.132	3.7	-6.050	6.28
41	-4.612	3.84	-6.805	6.04
61	-5.319	4.01	-7.859	5.986
81	-5.841	4.17	-8.608	5.995

Table 1.2: Grid convergence of two high-order schemes on $u_t + u_x = 0$, using SAT implementation of boundary conditions with the SAT parameter $\tau = 1$. Here $\omega = 2\pi$, CFL $= 0.1$, T $= 10$.

	Fourth-order compact		Sixth-order compact	
Grid	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate
21	-3.632		-5.012	
31	-4.315	3.99	-6.203	6.75
41	-4.816	4.00	-7.044	6.73
61	-5.541	4.12	-8.170	6.39
81	-6.061	4.16	-8.949	6.23

Table 1.3: Grid convergence of two high-order schemes on $u_t + u_x = 0$, using SAT implementation of boundary conditions with the SAT parameter $\tau = 2$. Here $\omega = 2\pi$, CFL $= 0.1$, T $= 10$.

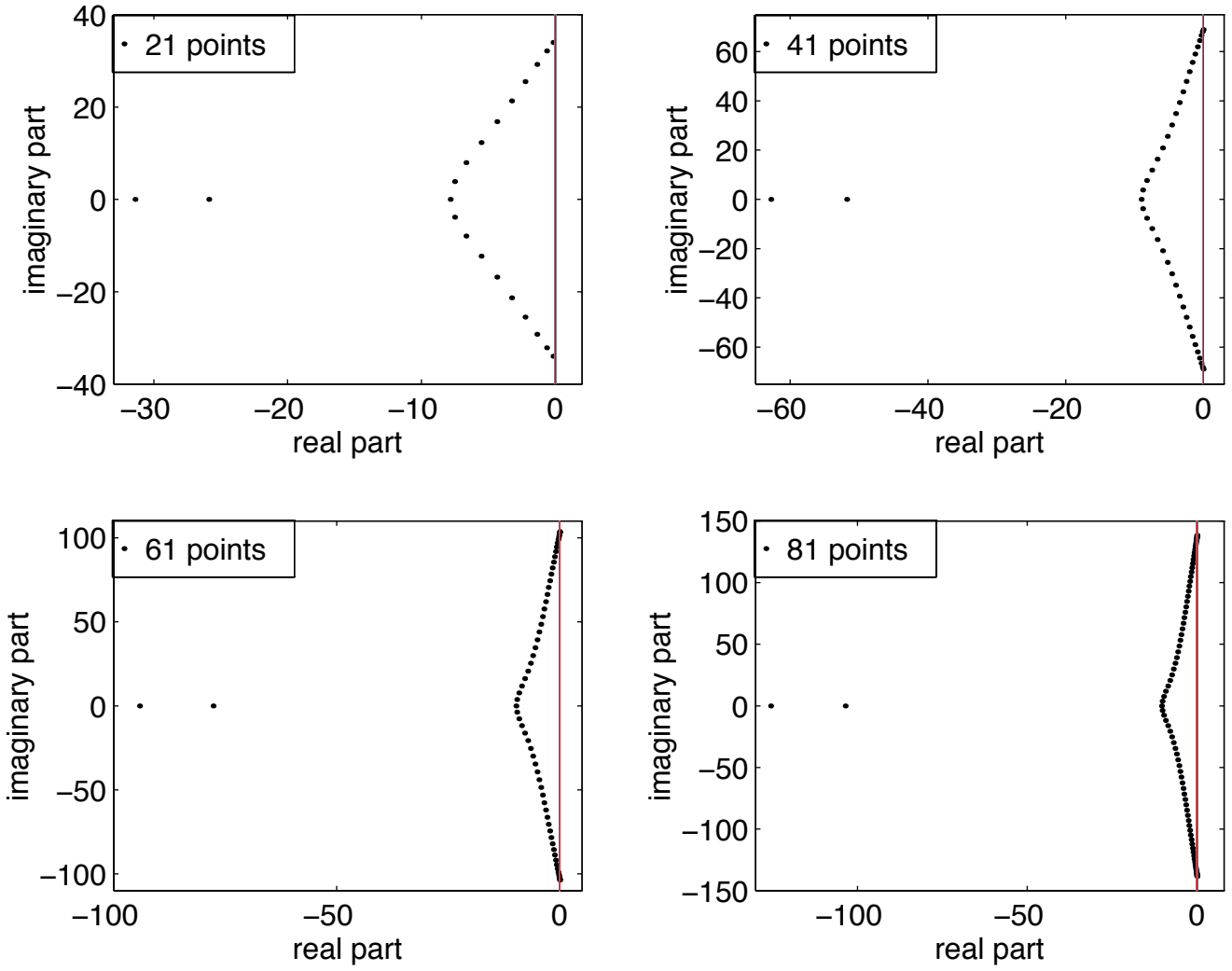

Figure 1.5: Semi-discrete eigenvalue spectrum for the fourth-order approximation using SAT implementation of boundary conditions with $\tau = 2$.

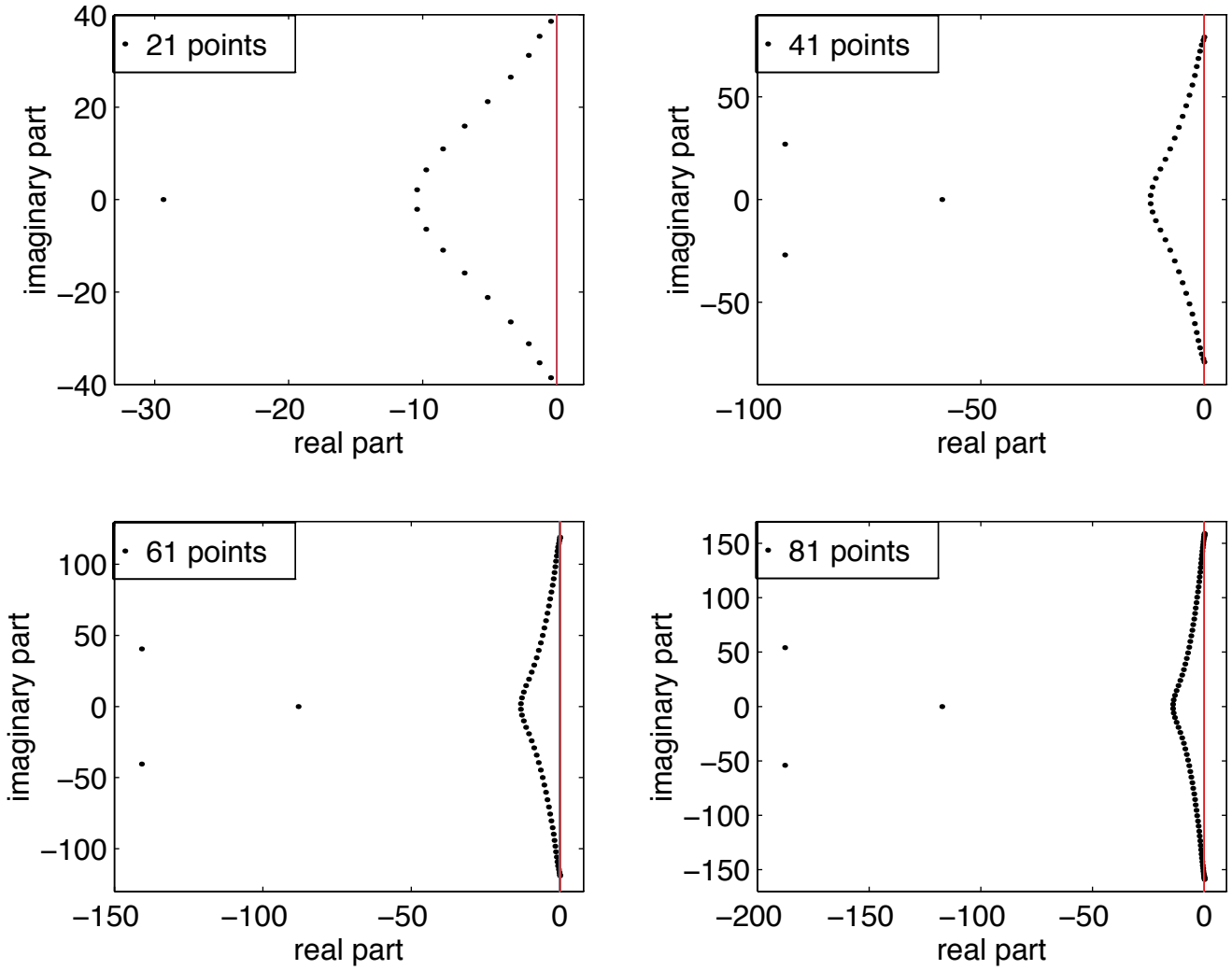

Figure 1.6: Semi-discrete eigenvalue spectrum for the sixth-order approximation using SAT implementation of boundary conditions with $\tau = 2$.

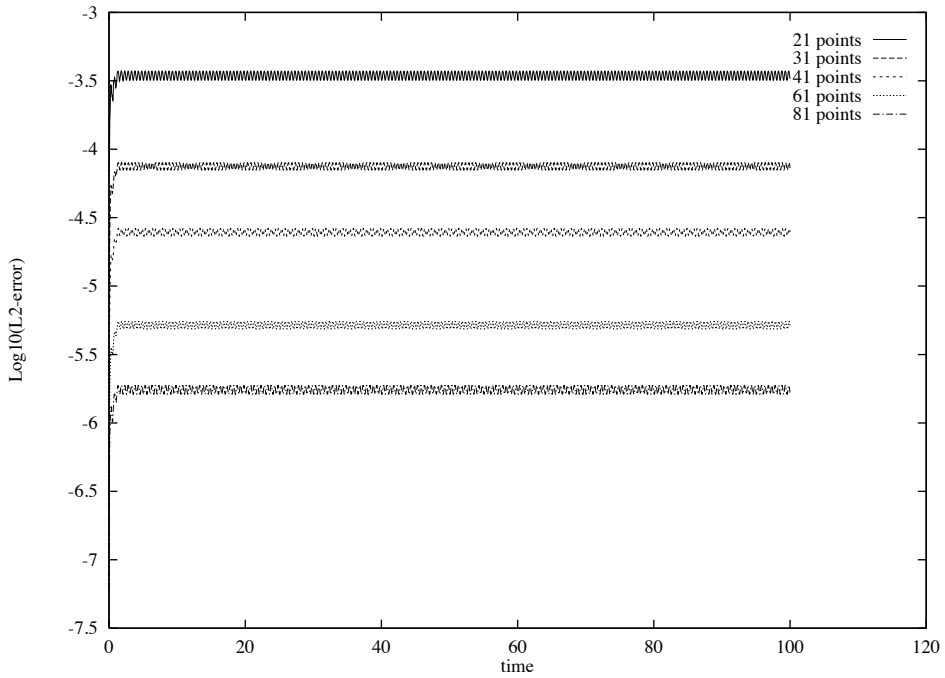

Figure 1.7: The L_2-error as a function of time for the fourth-order approximation using SAT implementation of boundary conditions with $\tau = 1,\ \mathrm{CFL} = 0.5,\ \omega = 2\pi$.

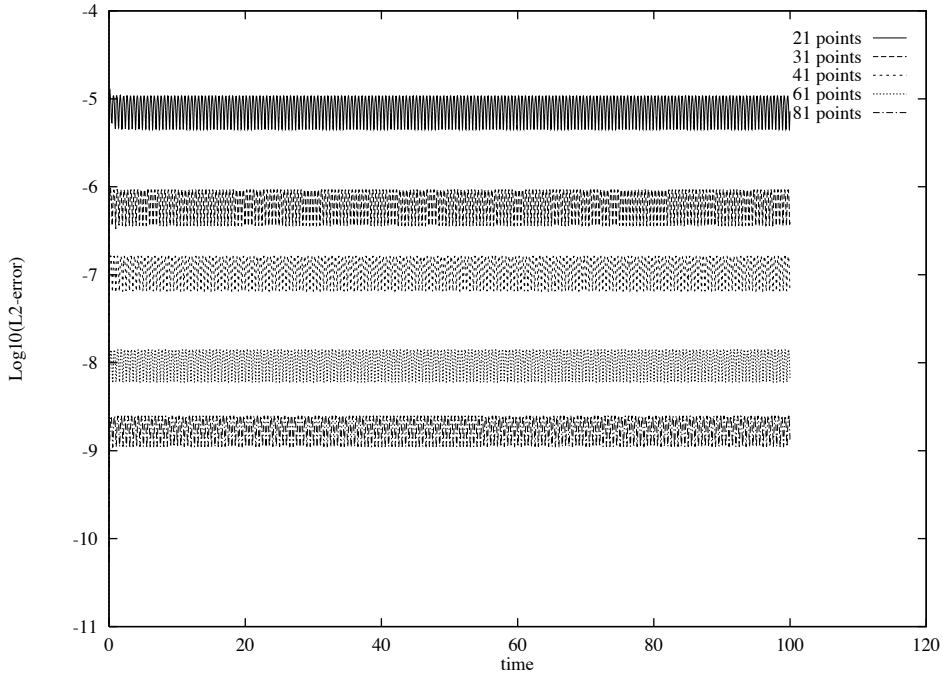

Figure 1.8: The L_2-error as a function of time for the sixth-order approximation using SAT implementation of boundary conditions with $\tau = 2,\ \mathrm{CFL} = 0.1,\ \omega = 2\pi$.

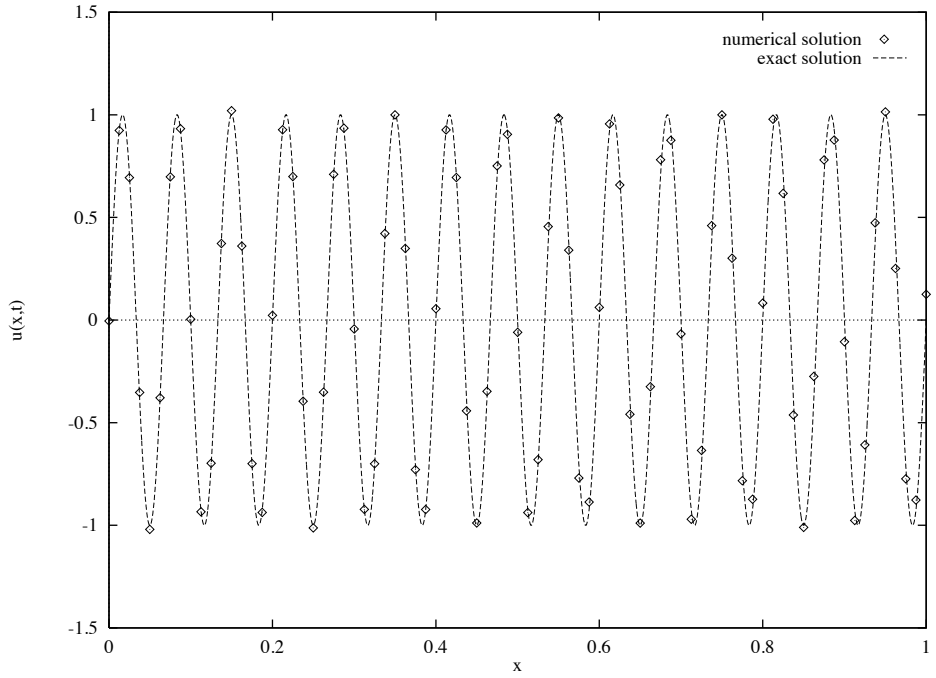

Figure 1.9: The numerical solution at the time $t = 10$ obtained with the sixth-order scheme using SAT implementation of boundary conditions with $\tau = 2$, CFL $= 0.1$, $\omega = 30\pi$, N $= 81$.

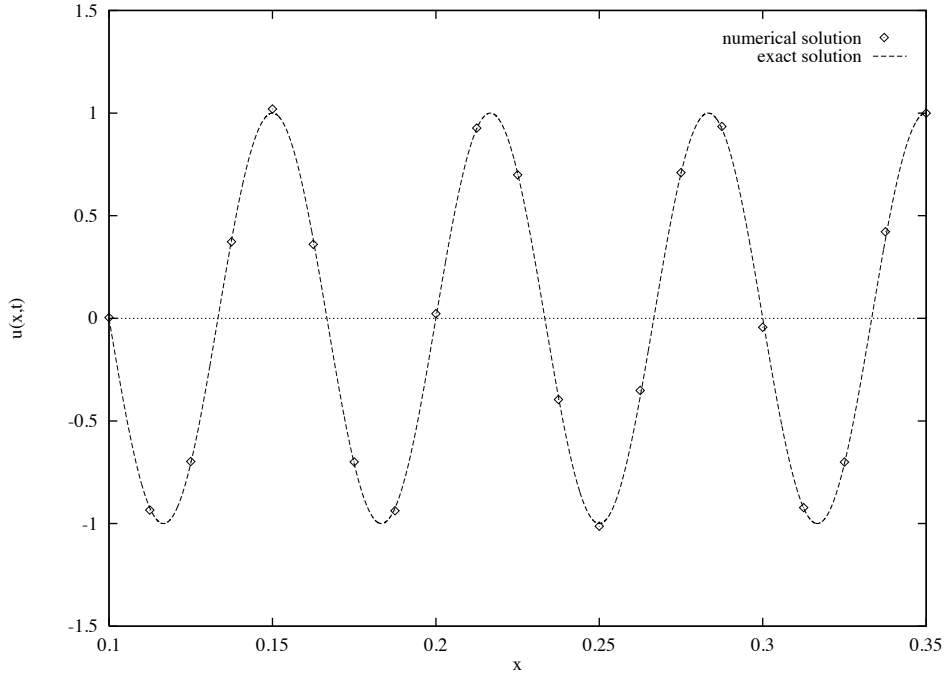

Figure 1.10: Magnification of the numerical solution at the time $t = 10$ obtained with the sixth-order scheme using SAT implementation of boundary conditions with $\tau = 2$, CFL $= 0.1$, $\omega = 30\pi$, N $= 81$.

Chapter 2

2-D Hyperbolic Equations

2.1 Description of the method and proof of main results

In this section we show how to use the one-dimensional scheme, whose properties were described in the previous chapter, for the two-dimensional case. We consider the following linear differential equation, with constant coefficients, in a rectangular domain Ω with boundary curve $\partial\Omega$:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} = 0, \qquad x, y \in \Omega, \quad t \geq 0, \tag{2.1.1}$$

with initial condition prescribed at $t = 0$:

$$u(x, y, 0) = f(x, y), \qquad x, y \in \Omega, \tag{2.1.2}$$

and the boundary condition:

$$u(x, y, t)|_{\partial\Omega} = g_B(t), \qquad t \geq 0. \tag{2.1.3}$$

Without lost of generality we assume that Ω is a square

$$\Omega = \left\{(x, y) \in \mathbf{R}^2 \mid 0 \leq x \leq 1,\ 0 \leq y \leq 1\right\},$$

and if, for example, $a > 0,\ b < 0$ then we have the following boundary conditions

$$u(0, y, t) = g^{(1)}(y, t) \tag{2.1.4}$$

$$u(x, 1, t) = g^{(2)}(x, t), \qquad t \geq 0. \tag{2.1.5}$$

We begin with dividing the continuous domain Ω into N^2 uniform intervals of width h where $h = \Delta x = \Delta y = 1/N$. We use for $i = 0, \ldots, N$ and $j = 0, \ldots, N$ the notation

$$x_i = ih, \quad y_j = jh, \quad u_{ij}(t) = u(x_i, y_j, t), \tag{2.1.6}$$

where $u_{ij}(t)$ is the projection of the exact solution $u(x,y,t)$ unto the grid. We arrange the solution projection array in vectors according to rows, starting from the bottom of the domain Ω and denote

$$\begin{aligned} \vec{U}(t) &= (u_{00}, u_{10}, \ldots, u_{N0}; \ldots; u_{0k}, u_{1k}, \ldots, u_{Nk}; \ldots; u_{0N}, u_{1N}, \ldots, u_{NN})^T \\ &= (\vec{u}_0, \ldots, \vec{u}_k, \ldots, \vec{u}_N)^T . \end{aligned} \tag{2.1.7}$$

If we arrange this array by columns (instead of rows) we will have the following structure

$$\begin{aligned} \vec{U}^c(t) &= (u_{00}, u_{01}, \ldots, u_{0N}; \ldots; u_{k0}, u_{k1}, \ldots, u_{kN}; \ldots; u_{N0}, u_{N1}, \ldots, u_{NN})^T \\ &= (\vec{u}^c_0, \ldots, \vec{u}^c_k, \ldots, \vec{u}^c_N)^T . \end{aligned} \tag{2.1.8}$$

As one can see, the vector $\vec{U}^c(t)$ is a specific permutation of $\vec{U}(t)$,

$$\vec{U}^c(t) = R\vec{U}(t), \tag{2.1.9}$$

where $R = R^T = R^{-1}$ is an $(N+1)^2 \times (N+1)^2$ orthogonal matrix whose each row contains $(N+1)^2 - 1$ zeros and a single 1 somewhere. If the domain is not a square then $R \neq R^T$, but still $RR^T = I$.

The continuous derivative $\dfrac{\partial \vec{u}_k}{\partial x}$ $(k = 0, \ldots, N)$ is then replaced with a finite-difference representation

$$P\frac{\partial \vec{u}_k}{\partial x} = Q\vec{u}_k + P\vec{T}^{(x)}_k \tag{2.1.10}$$

and the continuous derivative $\dfrac{\partial \vec{u}_k}{\partial y}$ $(k = 0, \ldots, N)$ is replaced by

$$\widetilde{P}\frac{\partial \vec{u}^c_k}{\partial y} = \widetilde{Q}\vec{u}^c_k + \widetilde{P}\vec{T}^{(y)}_k \tag{2.1.11}$$

where P, $\widetilde{P}$ and Q, $\widetilde{Q}$ are $(N+1) \times (N+1)$ matrices which have exactly the same structure as in one-dimensional case and vectors $\vec{T}^{(x)}_k$, $\vec{T}^{(y)}_k$ are the truncation error due to the numerical differentiation. Recall that the superscript "$\sim$" is used when the "inflow" boundary is on the right side of the one-dimensional domain.

Using (2.1.9), (2.1.10) and (2.1.11) we can write:

$$\begin{aligned} \left(a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y}\right) u_{ij}(t) &= \left[aD\vec{U} + b\widetilde{D}\vec{U}^c + \vec{T}^{(x)} + \vec{T}^{(y)}\right]_{ij} \\ &= \left[aD\vec{U} + bR\widetilde{D}R\vec{U} + \vec{T}^{(x)} + R\vec{T}^{(y)}\right]_{ij} \end{aligned} \tag{2.1.12}$$

where D and $\widetilde{D}$ are the following $(N+1)^2 \times (N+1)^2$ block-diagonal matrices:

$$(2.1.13)\quad D = \begin{pmatrix} P^{-1}Q & & & \\ & P^{-1}Q & & \\ & & \ddots & \\ & & & P^{-1}Q \end{pmatrix}, \quad \widetilde{D} = \begin{pmatrix} \tilde{P}^{-1}\tilde{Q} & & & \\ & \tilde{P}^{-1}\tilde{Q} & & \\ & & \ddots & \\ & & & \tilde{P}^{-1}\tilde{Q} \end{pmatrix}.$$

and $\vec{T}^{(x)} = \left(\vec{T}_1^{(x)}, \ldots, \vec{T}_N^{(x)}\right)^T$ and $\vec{T}^{(y)} = \left(\vec{T}_1^{(y)}, \ldots, \vec{T}_N^{(y)}\right)^T$ are truncation errors.

Before proceeding to the semi-discrete problem let us define by analogy to the one-dimensional case the following $(N+1)\times(N+1)$ matrices $\mathbf{Q}$, $\widetilde{\mathbf{Q}}$ and S, $\widetilde{S}$ and $(N+1)$ long vectors $\vec{S}_0$, S_N :

$$(2.1.14)\qquad \mathbf{Q} = Q - S, \qquad S = \begin{pmatrix} \tau q_{00} & 0 & \ldots & 0 \\ q_{01}+q_{10} & 0 & \ldots & 0 \\ 0 & & & \\ \vdots & & 0 & \\ 0 & & & \end{pmatrix}, \qquad \vec{S}_0 = \begin{pmatrix} \tau q_{00} \\ q_{01}+q_{10} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$(2.1.15)\quad \widetilde{\mathbf{Q}} = \tilde{Q} - \tilde{S}, \quad \tilde{S} = \begin{pmatrix} & & & 0 \\ & 0 & & \vdots \\ & & & 0 \\ 0 & \ldots & 0 & -(q_{01}+q_{10}) \\ 0 & \ldots & 0 & -\tau q_{00} \end{pmatrix}, \quad \vec{S}_N = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -(q_{01}+q_{10}) \\ -\tau q_{00} \end{pmatrix}.$$

We now define the $(N+1)^2$ vectors, $\vec{V} = (\vec{v}_0, \ldots, \vec{v}_k, \ldots, \vec{v}_N)^T$ and $\vec{V}^c = (\vec{v}_0^c, \ldots, \vec{v}_k^c, \ldots, \vec{v}_N^c)^T$ where $\vec{v}_k$ and $\vec{v}_k^c$ the numerical approximation to the projection $\vec{u}_k$ and $\vec{u}_k^c$, $(k = 0, \ldots, N)$ respectively and write the semi-discrete problem in the following way:

$$(2.1.16)\qquad \frac{d\vec{V}}{dt} = -\left[a\mathbf{D} + bR\widetilde{\mathbf{D}}R\right]\vec{V} - a\vec{G}^{(x)} - bR\vec{G}^{(y)},$$

where

$$\mathbf{D} = \begin{pmatrix} P^{-1}\mathbf{Q} & & & \\ & P^{-1}\mathbf{Q} & & \\ & & \ddots & \\ & & & P^{-1}\mathbf{Q} \end{pmatrix}, \quad \widetilde{\mathbf{D}} = \begin{pmatrix} \tilde{P}^{-1}\widetilde{\mathbf{Q}} & & & \\ & \tilde{P}^{-1}\widetilde{\mathbf{Q}} & & \\ & & \ddots & \\ & & & \tilde{P}^{-1}\widetilde{\mathbf{Q}} \end{pmatrix},$$

$$\vec{G}^{(x)} = \begin{pmatrix} P^{-1}\vec{S}_0 g_0^{(1)}(t) \\ P^{-1}\vec{S}_0 g_1^{(1)}(t) \\ \vdots \\ P^{-1}\vec{S}_0 g_N^{(1)}(t) \end{pmatrix}, \quad \vec{G}^{(y)} = \begin{pmatrix} \tilde{P}^{-1}\vec{S}_N g_0^{(2)}(t) \\ \tilde{P}^{-1}\vec{S}_N g_1^{(2)}(t) \\ \vdots \\ \tilde{P}^{-1}\vec{S}_N g_N^{(2)}(t) \end{pmatrix}. \tag{2.1.17}$$

Since $S\vec{u}_k - \vec{S}_0 g_k^{(1)}(t) = 0$ and also $S\vec{u}_k^c - \vec{S}_N g_k^{(2)}(t) = 0 \quad (k = 0, \ldots, N)$, we may write for the vector $\vec{U}$:

$$\frac{d\vec{U}}{dt} = \left[-a\mathbf{D} - bR\widetilde{\mathbf{D}}R\right]\vec{U} - a\vec{G}^{(x)} - bR\vec{G}^{(y)} + \vec{T}^{(x)} + R\vec{T}^{(y)}. \tag{2.1.18}$$

Subtracting (2.1.16) from (2.1.18) we get:

$$\frac{d\vec{E}}{dt} = \left[-a\mathbf{D} - bR\widetilde{\mathbf{D}}R\right]\vec{E} + \vec{\mathbf{T}}, \tag{2.1.19}$$

where $\vec{E} = \vec{U} - \vec{V}$ is the two dimensional array of the errors arranged by rows as a vector and $\vec{\mathbf{T}}$ is proportional to the truncation error.

We recall that in section 1.1 it was proven that if the inequalities (1.1.14) hold then the real part of each eigenvalue of the matrix $P^{-1}\mathbf{Q}$ is positive and the real part of each eigenvalue of the matrix $\tilde{P}^{-1}\widetilde{\mathbf{Q}}$ is negative. Therefore all eigenvalues of $\mathbf{D}$ have a positive real part and all eigenvalues of $R\widetilde{\mathbf{D}}R$ have a negative real part. To prove the time stability of the scheme (2.1.16) it is sufficient to show that $H\left(-a\mathbf{D} - bR\widetilde{\mathbf{D}}R\right) + \left[H\left(-a\mathbf{D} - bR\widetilde{\mathbf{D}}R\right)\right]^T \leq 0$ for any symmetric positive definite matrix H.

To show this, we define now a symmetric positive definite matrix, $H = P^{1/2}(R\tilde{P}R)P^{1/2}$ and consider the following scalar product:

$$\left(\left[H\left(-a\mathbf{D}-bR\widetilde{\mathbf{D}}R\right)+\left(-a\mathbf{D}-bR\widetilde{\mathbf{D}}R\right)^{T}H\right]\vec{E},\vec{E}\right)$$

(2.1.20)

$$=-a\left(\left[H\mathbf{D}+\mathbf{D}^{T}H\right]\vec{E},\vec{E}\right)-b\left(\left[HR\widetilde{\mathbf{D}}R+R\widetilde{\mathbf{D}}^{T}RH\right]\vec{E},\vec{E}\right)$$

It can be verified by direct multiplication and using the properties of block-diagonal matrices and of the permutation matrix R, that any block-diagonal matrix M is commutative with the matrix of the form $R\widetilde{\mathbf{D}}R$, i.e., for example, $MR\widetilde{\mathbf{D}}R=R\widetilde{\mathbf{D}}RM$. Using this information, and the fact that $RR=I$ we can write:

$$H\mathbf{D}+\mathbf{D}^{T}H=P^{1/2}(R\tilde{P}R)P^{1/2}P^{-1}\mathbf{Q}+\mathbf{Q}^{T}P^{-1}P^{1/2}(R\tilde{P}R)P^{1/2}$$

$$=(R\tilde{P}R)\mathbf{Q}+(R\tilde{P}R)\mathbf{Q}^{T}=R\tilde{P}R\left(\mathbf{Q}+\mathbf{Q}^{T}\right),$$

(2.1.21)

$$HR\widetilde{\mathbf{D}}R+R\widetilde{\mathbf{D}}^{T}RH=P^{1/2}(R\tilde{P}R)P^{1/2}R\tilde{P}^{-1}\widetilde{\mathbf{Q}}R+R\widetilde{\mathbf{Q}}^{T}\tilde{P}^{-1}RP^{1/2}(R\tilde{P}R)P^{1/2}$$

$$=PR\widetilde{\mathbf{Q}}R+PR\widetilde{\mathbf{Q}}^{T}R=PR\left(\widetilde{\mathbf{Q}}+\widetilde{\mathbf{Q}}^{T}\right)R$$

Denoting $\vec{\varphi}=(R\tilde{P}^{1/2}R)\vec{E}$ and using again the fact that $(R\tilde{P}^{1/2}R)M=M(R\tilde{P}^{1/2}R)$ we obtain:

$$\left(R\tilde{P}R\left(\mathbf{Q}+\mathbf{Q}^{T}\right)\vec{E},\vec{E}\right)=\left(R\tilde{P}^{1/2}RR\tilde{P}^{1/2}R\left(\mathbf{Q}+\mathbf{Q}^{T}\right)\vec{E},\vec{E}\right)$$

(2.1.22)

$$=\left(R\tilde{P}^{1/2}R\left(\mathbf{Q}+\mathbf{Q}^{T}\right)\vec{E},R\tilde{P}^{1/2}R\vec{E}\right)=\left(\left(\mathbf{Q}+\mathbf{Q}^{T}\right)\vec{\varphi},\vec{\varphi}\right),$$

In a similar fashion, denoting $\vec{\eta}=(RP^{1/2})\vec{E}$ we get

$$\left(PR\left(\widetilde{\mathbf{Q}}+\widetilde{\mathbf{Q}}^{T}\right)R\vec{E},\vec{E}\right)=\left(P^{1/2}R\left(\widetilde{\mathbf{Q}}+\widetilde{\mathbf{Q}}^{T}\right)R\vec{E},P^{1/2}\vec{E}\right)$$

(2.1.23)

$$=\left(R\left(\widetilde{\mathbf{Q}}+\widetilde{\mathbf{Q}}^{T}\right)RP^{1/2}\vec{E},P^{1/2}\vec{E}\right)=\left(\left(\widetilde{\mathbf{Q}}+\widetilde{\mathbf{Q}}^{T}\right)\vec{\eta},\vec{\eta}\right).$$

Taking into account (2.1.21), (2.1.22), (2.1.23), the fact that $a > 0,\ b < 0$ and

$$((\mathbf{Q} + \mathbf{Q}^T)\,\vec{\varphi}, \vec{\varphi}) \geq 0, \quad \left(\left(\widetilde{\mathbf{Q}} + \widetilde{\mathbf{Q}}^T\right)\vec{\eta}, \vec{\eta}\right) \leq 0 \qquad \forall \vec{\varphi}, \vec{\eta} \in \mathbf{R}^{(N+1)^2}$$

we can conclude that if the one-dimensional inequalities (1.1.14) hold, then

$$\left(\left[H\left(-a\mathbf{D} - bR\widetilde{\mathbf{D}}R\right) + \left(-a\mathbf{D} - bR\widetilde{\mathbf{D}}R\right)^T H\right]\vec{E}, \vec{E}\right)$$

$$= -a\,((\mathbf{Q} + \mathbf{Q}^T)\,\vec{\varphi}, \vec{\varphi}) - b\left(\left(\widetilde{\mathbf{Q}} + \widetilde{\mathbf{Q}}^T\right)\vec{\eta}, \vec{\eta}\right) \leq 0$$

for all $\vec{E} \in \mathbf{R}^{(N+1)^2}$.

Remark. It should be observed that when the scheme (2.1.16) was used in practice, the one-dimensional algorithm was implemented on each row to compute the numerical approximation to u_x , and on each column to compute the numerical approximation to u_y . It means that in practice we solved the following equation:

$$\frac{d}{dt}[V] = -\left[a\mathbf{D}[V] + b[V]\widetilde{\mathbf{D}}^T + a\vec{G}^{(1)}(t)\vec{S}_0^T P^{-1} + b\tilde{P}^{-1}\vec{S}_N\left(\vec{G}^{(2)}(t)\right)^T\right],$$

where $[V]$ is $(N+1)\times(N+1)$ matrix with the elements v_{ij} and

$$\vec{G}^{(1)}(t) = \begin{pmatrix} g_0^{(1)}(t) \\ g_1^{(1)}(t) \\ \vdots \\ g_N^{(1)}(t) \end{pmatrix}, \qquad \vec{G}^{(2)}(t) = \begin{pmatrix} g_0^{(2)}(t) \\ g_1^{(2)}(t) \\ \vdots \\ g_N^{(2)}(t) \end{pmatrix},$$

and the matrices $\mathbf{D}$, $\widetilde{\mathbf{D}}$ and P, $\tilde{P}$ and the vectors $\vec{S}_0$, S_N were defined earlier. Note that in practice P^{-1}, $\tilde{P}^{-1}$ are never evaluated. Rather the decompositions $P = LU$ and $\tilde{P} = \tilde{L}\tilde{U}$ are calculated. L and U ($\tilde{L}$ and $\tilde{U}$) are bidiagonal matrices with one of them having "ones" along the diagonal. Hence, the inversion of L and U ($\tilde{L}$ and $\tilde{U}$) is very cheap.

2.2 Numerical results

Here we consider the problem

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}+\frac{\partial u}{\partial y} = 0, \qquad 0\le x\le 1, \quad 0\le y\le 1, \quad t\ge 0, \tag{2.2.1}$$

$$u(0,y,t) = \sin\omega(y-2t), \tag{2.2.2}$$

$$u(x,0,t) = \sin\omega(x-2t), \tag{2.2.3}$$

$$u(x,y,0) = \sin\omega(x+y). \tag{2.2.4}$$

The analytic solution of this problem is:

$$u(x,y,t) = \sin\omega(x+y-2t) \tag{2.2.5}$$

We shall now use the SAT method to solve the problem (2.2.1)-(2.2.4) , as well as using the conventional implementation of boundary conditions. Two difference operators were used: fourth-order with third-order boundary closure and sixth-order with fifth-order boundary closure. The temporal discretization was accomplished with the standard fourth-order Runge-Kutta algorithm in the case of fourth-order difference operator and with a sixth-order Runge-Kutta algorithm developed by Butcher [2], [3] in the case of sixth-order difference operator. In the case of conventional implementation of boundary conditions the value of the solution at the boundary point was overridden with the analytic boundary condition at the end of each Runge-Kutta stage.

Conventional boundary conditions. To check on the order of accuracy, the runs were repeated for $\Delta x=\Delta y=1/20,\ 1/30,\ 1/40,\ 1/60$ and $1/80$. Doubling the grid at constant CFL, should decrease the error at time $t=T$ by a factor $(\frac{1}{2})^p$ where $p=4,6$ is order of the method. The formal accuracy of each scheme was determined in this manner. Table (2.1) shows the results of this study. The $\log_{10}$ of the L_2 error at time $t=T$ and the convergence rate are the entries. $T=0.4$ in the case of the sixth-order scheme and $T=10$ in the case of the fourth-order scheme. As one can see for relative short-time integration the convergence rate of the sixth-order scheme is approximately 6 and the convergence rate of the fourth-order scheme asymptotes to the theoretical value of 4.

The error as a function of time for the fourth-order and the sixth-order schemes is shown in Figures (2.2) and (2.3) respectively for different grids. $\mathrm{CFL}=0.5$ were chosen for the fourth-order scheme and $\mathrm{CFL}=0.1$ were chosen for the sixth-order scheme. As in the one

dimensional case, the runs are time stable in the case of the fourth-order scheme. The results obtained by using the sixth-order scheme diverge exponentially from the analytic solution.

SAT boundary conditions. To check on the order of accuracy, the runs were repeated for $\Delta x = \Delta y = 1/20,\ 1/30,\ 1/40,\ 1/60$ and $1/80$. Table (2.2) shows a grid convergence study for both spatial operators. The absolute error $\log_{10}(L_2)$ at a fixed time $\mathrm{T} = 10$ and the convergence rate are plotted. As one can see the formal accuracy of the spatial operator is unaffected by SAT boundary treatment.

The simulations were all run to equivalent times $T = 100$ for both the fourth- and the sixth-order schemes and different grids. $\mathrm{CFL} = 0.25,\ \tau = 1$ were chosen for the fourth-order scheme and $\mathrm{CFL} = 0.1,\ \tau = 2$ were chosen for the sixth-order scheme. Figure (2.4), (2.5) show a plot of the error of the solution to the problem (2.2.1)-(2.2.4) for the fourth-order and the sixth-order respectively. The $\log_{10}$ of the L_2 error is plotted as a function of time for five grid densities: 21, 31, 41, 61 and 81 points, respectively. It is clear that both schemes give good results, no exponential growth exists, indicating time stability of the schemes.

Figure (2.1) shows the 3-D plot of the numerical solution at the time $T = 2$ obtained using the sixth-order scheme with $N = 60,\ \mathrm{CFL} = 0.1,\ \tau = 2,\ \omega = 2\pi$.

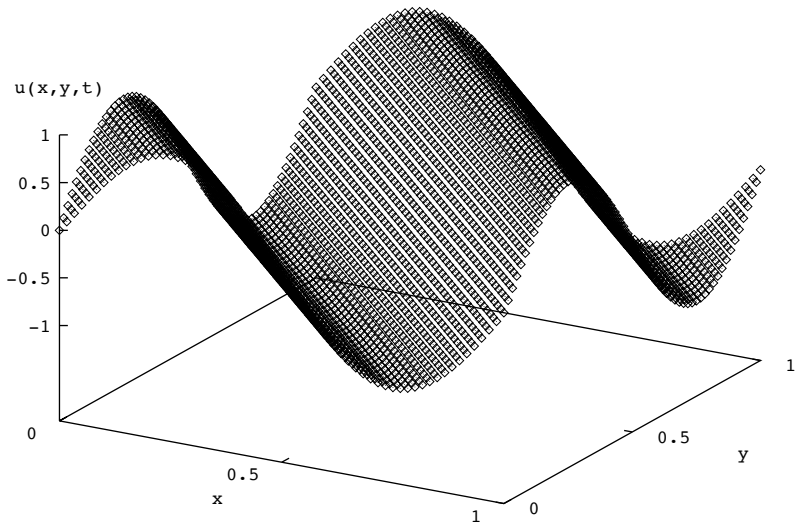

Figure 2.1: Numerical solution at the time $T = 2$ obtained with the sixth-order approximation using SAT implementation of boundary conditions with $N = 60,\ \mathrm{CFL} = 0.1,\ \tau = 2,\ \omega = 2\pi$.

	Fourth-order compact		Sixth-order compact	
Grid	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate
21	-2.786		-3.536	
31	-3.461	3.83	-4.619	6.15
41	-3.947	3.89	-5.378	6.07
61	-4.638	3.92	-6.420	5.92
81	-5.131	3.95	-7.143	5.83

Table 2.1: Grid convergence of two high-order schemes on $u_t+u_x+u_y = 0$, using conventional implementation of boundary conditions with CFL $= 0.1$ and $T = 0.4$ for the sixth-order scheme, and CFL $= 0.25$ and $T = 10$ for the fourth-order scheme. $\omega = 2\pi$.

	Fourth-order compact		Sixth-order compact	
Grid	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate
21	-3.389		-4.909	
31	-4.100	4.04	-5.991	6.14
41	-4.599	4.00	-6.757	6.14
61	-5.310	4.04	-7.835	6.06
81	-5.813	4.03	-8.575	6.00

Table 2.2: Grid convergence of two high-order schemes on $u_t + u_x + u_y = 0$, using SAT implementation of boundary conditions with the SAT parameter $\tau = 2$ and CFL $= 0.1$ for the sixth-order scheme and the SAT parameter $\tau = 1$ and CFL $= 0.25$ for the fourth-order scheme. $T = 10$, $\omega = 2\pi$.

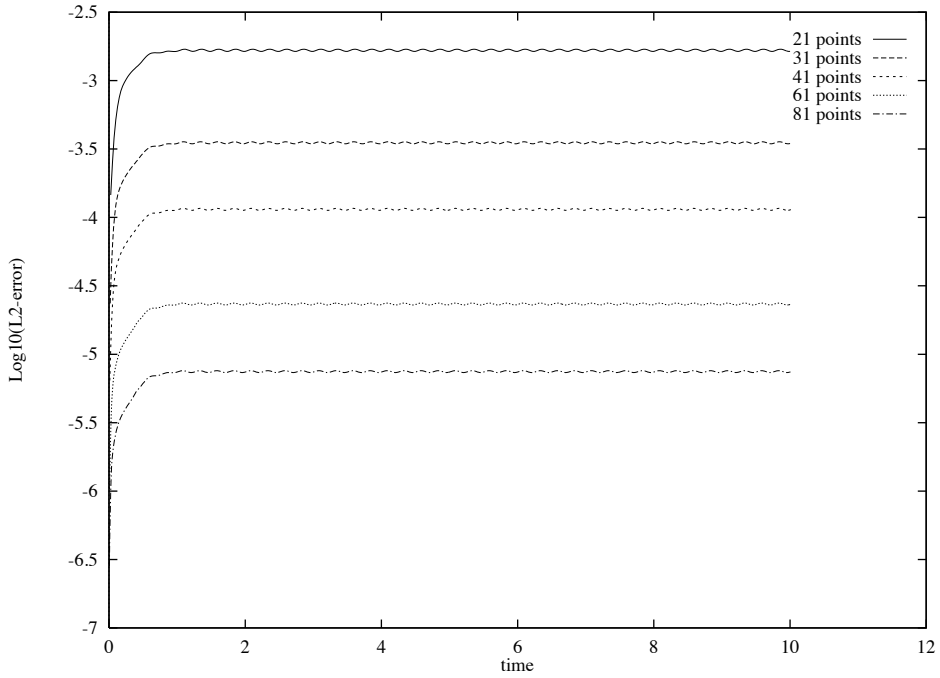

Figure 2.2: The L_2-error as a function of time for the fourth-order approximation using conventional implementation of boundary conditions with CFL $= 0.25$, $\omega = 2\pi$.

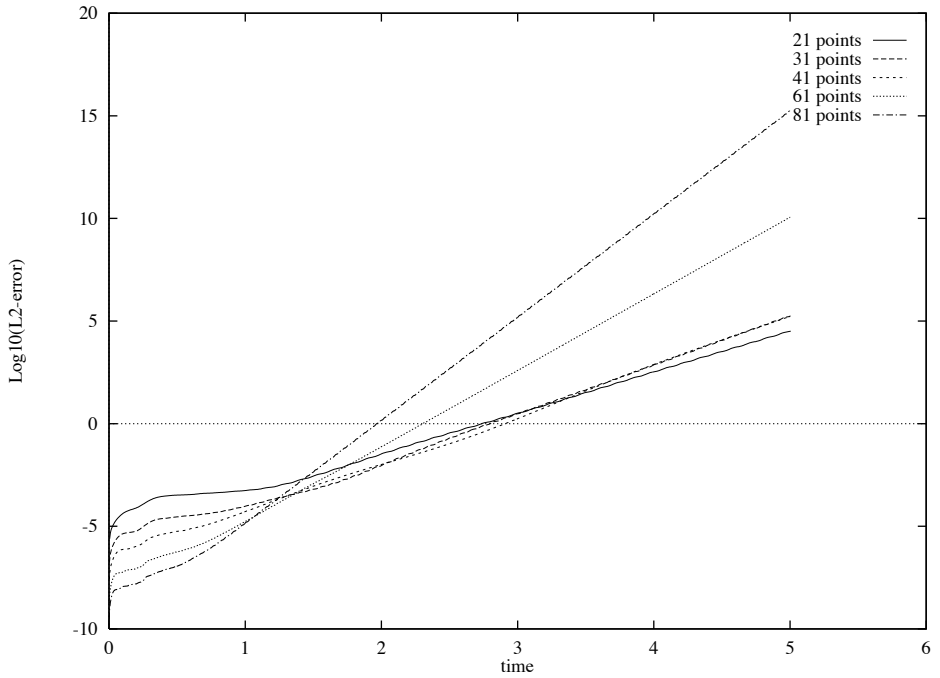

Figure 2.3: The L_2-error as a function of time for the sixth-order approximation using conventional implementation of boundary conditions with CFL $= 0.1$, $\omega = 2\pi$.

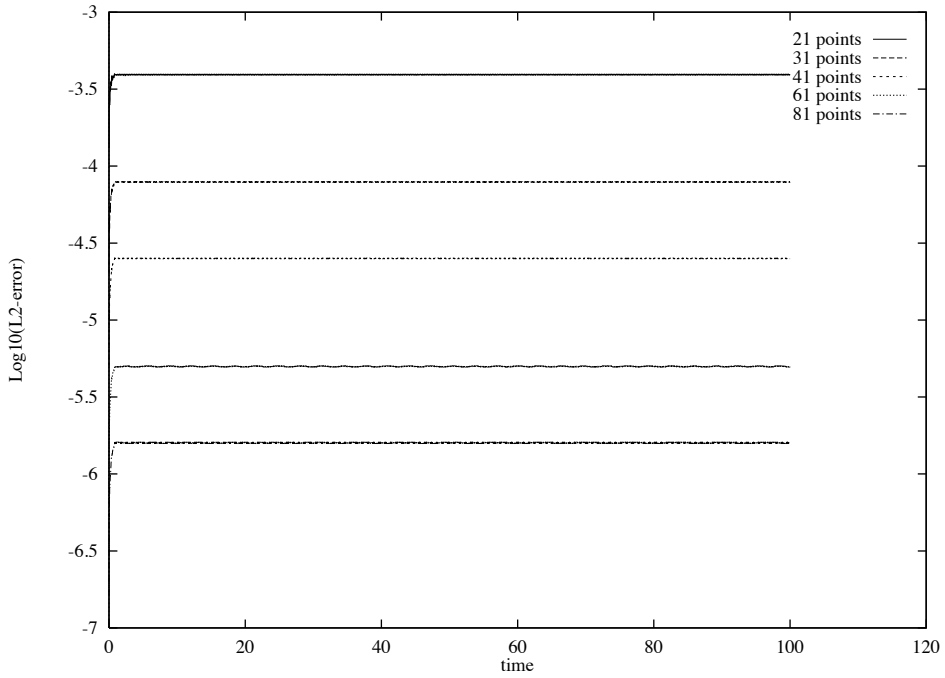

Figure 2.4: The L_2-error as a function of time for the fourth-order approximation using SAT implementation of boundary conditions with $\text{CFL} = 0.25,\ \tau = 1,\ \omega = 2\pi$.

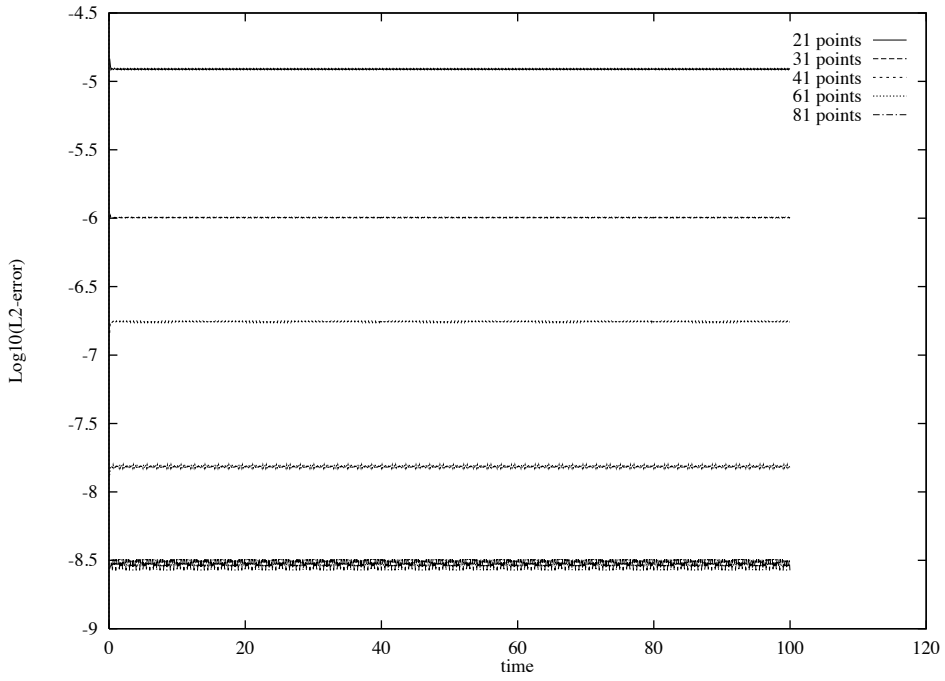

Figure 2.5: The L_2-error as a function of time for the sixth-order approximation using SAT implementation of boundary conditions with $\text{CFL} = 0.1,\ \tau = 2,\ \omega = 2\pi$.

Part II

The Hyperbolic System

Chapter 3

1-D Hyperbolic Systems

3.1 General theory and description of the method

Consider a first order hyperbolic system of partial differential equations

$$\frac{\partial \mathbf{u}}{\partial t} + A\frac{\partial \mathbf{u}}{\partial x} = 0, \quad 0 \le x \le 1, \quad t \ge 0 \tag{3.1.1}$$

where without loss of generality $\mathbf{u}(x,t) = (u^1(x,t), \ldots, u^r(x,t))^T$ and A is a diagonal matrix with constant entries:

$$A = \begin{pmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_k & & & & \\ & & & \lambda_{k+1} & & & \\ & & & & \ddots & & \\ & & & & & \lambda_r & \end{pmatrix}, \qquad \begin{aligned} &\lambda_1 > \lambda_2 > \ldots, \lambda_k > 0, \\ &\lambda_r < \ldots < \lambda_{k+2} < \lambda_{k+1} < 0. \end{aligned}$$

The solution of (3.1.1) is uniquely determined if we prescribe initial values

$$\mathbf{u}(x,0) = \mathbf{f}(x), \qquad 0 \le x \le 1, \tag{3.1.2}$$

and boundary conditions

$$\begin{aligned} \mathbf{u}^{\mathrm{I}}(0,t) &= L\mathbf{u}^{\mathrm{II}}(0,t) + g^{\mathrm{I}}(t) \\ \mathbf{u}^{\mathrm{II}}(1,t) &= R\mathbf{u}^{\mathrm{I}}(1,t) + g^{\mathrm{II}}(t), \qquad t \ge 0, \end{aligned} \tag{3.1.3}$$

where L and R are fixed matrices of orders $k \times (r-k)$ and $(r-k) \times k$, respectively, $g^{\mathrm{I}}(t)$ a given k-vector, $g^{\mathrm{II}}(t)$ a given $(r-k)$-vector, and

$$\mathbf{u}^{\mathrm{I}} = (u^1, \ldots, u^k)^T, \quad \mathbf{u}^{\mathrm{II}} = (u^{k+1}, \ldots, u^r)^T \tag{3.1.4}$$

is a partition of $\mathbf{u}$ into its outflow and inflow components, respectively, corresponding to the partition of A.

It is well known that (3.1.3) is well posed for any L and R, but in order to assure that the solution of (3.1.1) is bounded in time (under the condition that $g^{\mathrm{I}}(t)$ and $g^{\mathrm{II}}(t)$ are bounded in time), it is sufficient to assume that

$$\| L \| \cdot \| R \| \leq 1 \tag{3.1.5}$$

where the non-square matrix norm is defined by

$$\| L \| = \rho(L^T L)^{\frac{1}{2}} \tag{3.1.6}$$

and $\rho(L^T L)$ is the spectral radius of $L^T L$.

In order to solve the initial-boundary value problem (3.1.1) by a finite difference approximation, we introduce, as in scalar case, a mesh size h and denote by $\mathbf{u}^i = (u_0^i, u_1^i, \ldots, u_N^i)^T$, $i = 1, \ldots, r$, vectors of unknowns corresponding to the grid points $x_0, \ldots, x_N \quad (N = 1/h)$ and by $\mathbf{v}^i$ the numerical approximation to $\mathbf{u}^i$. Assuming that we have matrices $P, Q, \tilde{P}, \tilde{Q}$ as in the scalar case and the vectors $\vec{S}_0, \vec{S}_N$ are

$$\vec{S}_0 = \begin{pmatrix} \tau q_{00} \\ q_{01} + q_{10} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \vec{S}_N = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -(q_{01} + q_{10}) \\ -\tau q_{00} \end{pmatrix} \tag{3.1.7}$$

we approximate the (3.1.1) by the following scheme:

$$\begin{aligned} P\frac{d\mathbf{v}^i}{dt} &= -\lambda_i Q\mathbf{v}^i + \lambda_i \vec{S}_0(\mathbf{v}_0^i - (L\mathbf{v}^{\mathrm{II}} + g^{\mathrm{I}})_0^i), \qquad 1 \leq i \leq k \\ \tilde{P}\frac{d\mathbf{v}^i}{dt} &= -\lambda_i \tilde{Q}\mathbf{v}^i + \lambda_i \vec{S}_N(\mathbf{v}_N^i - (R\mathbf{v}^{\mathrm{I}} + g^{\mathrm{II}})_N^i), \qquad k+1 \leq i \leq r \end{aligned} \tag{3.1.8}$$

To prove the convergence of the scheme (3.1.8) we will derive an equation for the error function $\mathcal{E}$ and will show that its discrete norm (to be defined later) is bounded by a function $F(t, h, u)$, where t, h and u are the time, the mesh size, and the exact solution respectively.

Since $\mathbf{u}_0^i - (L\mathbf{u}^{\mathrm{II}} + g^{\mathrm{I}})_0^i = 0$ for $1 \leq i \leq k$ and $\mathbf{u}_N^i - (R\mathbf{u}^{\mathrm{I}} + g^{\mathrm{II}})_N^i = 0$ for $k+1 \leq i \leq r$, we may write for the vectors $\mathbf{u}^i$

$$\begin{aligned} P\frac{d\mathbf{u}^i}{dt} &= -\lambda_i Q\mathbf{u}^i + \lambda_i \vec{S}_0(\mathbf{u}_0^i - (L\mathbf{u}^{\mathrm{II}} + g^{\mathrm{I}})_0^i) + P\mathbf{T}^i, \qquad 1 \leq i \leq k \\ \tilde{P}\frac{d\mathbf{u}^i}{dt} &= -\lambda_i \tilde{Q}\mathbf{u}^i + \lambda_i \vec{S}_N(\mathbf{u}_N^i - (R\mathbf{u}^{\mathrm{I}} + g^{\mathrm{II}})_N^i) + \tilde{P}\mathbf{T}^i, \qquad k+1 \leq i \leq r \end{aligned} \tag{3.1.9}$$

where $\mathbf{T} = (\mathbf{T^0}, \ldots, \mathbf{T^k}, \mathbf{T^{k+1}}, \ldots, \mathbf{T^r})$ is the r long vector of the truncation error due to numerical differencing.

Denote by $\varepsilon^i = \mathbf{u}^i - \mathbf{v}^i \quad (1 \leq i \leq r)$ the solution error vectors and substructing (3.1.8) from (3.1.9) to get

$$
\begin{aligned}
P\frac{d\varepsilon^i}{dt} &= -\lambda_i Q\varepsilon^i + \lambda_i \vec{S}_0(\varepsilon_0^i - (L\varepsilon^{\mathrm{II}})_0^i) + P\mathbf{T}^i, \qquad 1 \leq i \leq k \\
\tilde{P}\frac{d\varepsilon^i}{dt} &= -\lambda_i \tilde{Q}\varepsilon^i + \lambda_i \vec{S}_N(\varepsilon_N^i - (R\varepsilon^{\mathrm{I}})_N^i) + \tilde{P}\mathbf{T}^i, \qquad k+1 \leq i \leq r
\end{aligned}
\tag{3.1.10}
$$

We define now the scalar product

$$
(\varepsilon^i, \varepsilon^j) = \sum_{m=0}^{N} \varepsilon_m^i \varepsilon_m^j \tag{3.1.11}
$$

and the discrete norms

$$
\|\varepsilon^{\mathrm{I}}\|^2 = \sum_{i=1}^{k} \frac{\|R\|}{\lambda_i}(\varepsilon^i, \varepsilon^i), \qquad \|\varepsilon^{\mathrm{II}}\|^2 = \sum_{i=k+1}^{r} \frac{\|L\|}{|\lambda_i|}(\varepsilon^i, \varepsilon^i) \tag{3.1.12}
$$

and

$$
\|\mathcal{E}\|^2 = \|\varepsilon^{\mathrm{I}}\|^2 + \|\varepsilon^{\mathrm{II}}\|^2, \tag{3.1.13}
$$

where $\mathcal{E}$ is the $r \times N$ long error vector whose first $k \times N$ entries are the entries of ε^{I} and the other $(r-k) \times N$ entries are the ones of $\varepsilon^{\mathrm{II}}$.

Differentiating the scalar products $(P\varepsilon^i, \varepsilon^i)$ and $(\tilde{P}\varepsilon^i, \varepsilon^i)$ and using the equation (3.1.10) yields

$$
\begin{aligned}
\frac{d}{dt}(P\varepsilon^i, \varepsilon^i) &= -\lambda_i (Q\varepsilon^i, \varepsilon^i) + \lambda_i (\vec{S}_0, \varepsilon^i)(\varepsilon_0^i - (L\varepsilon^{\mathrm{II}})_0^i) + (P\mathbf{T}^i, \varepsilon^i), \quad 1 \leq i \leq k \\
\frac{d}{dt}(\tilde{P}\varepsilon^i, \varepsilon^i) &= -\lambda_i (\tilde{Q}\varepsilon^i, \varepsilon^i) + \lambda_i (\vec{S}_N, \varepsilon^i)(\varepsilon_N^i - (R\varepsilon^{\mathrm{I}})_N^i) + (\tilde{P}\mathbf{T}^i, \varepsilon^i), \quad k+1 \leq i \leq r
\end{aligned}
\tag{3.1.14}
$$

We now use the definitions of $\vec{S}_0$ and $\vec{S}_N$, the properties of Q and $\tilde{Q}$ (from assumption **3** and remarks from chapter 1) and the fact that the λ_i are positive for $1 \leq i \leq k$ and are negative for $k+1 \leq i \leq r$ to get

$$\frac{d}{dt}(P\varepsilon^i,\varepsilon^i)=\lambda_i(\tau-1)q_{00}(\varepsilon_0^i)^2-\lambda_i q_{11}(\varepsilon_1^i)^2-\lambda_i\tau q_{00}(L\varepsilon^{\rm II})_0^i\varepsilon_0^i-\lambda_i(q_{01}+q_{10})\varepsilon_1^i(L\varepsilon^{\rm II})_0^i$$

$$-\lambda_i\left[q_{NN}(\varepsilon_N^i)^2+(q_{N-1N}+q_{NN-1})\varepsilon_{N-1}^i\varepsilon_N^i+q_{N-1N-1}(\varepsilon_{N-1}^i)^2\right]+(P\mathbf{T}^i,\varepsilon^i),\qquad 1\le i\le k$$

(3.1.15)

$$\frac{d}{dt}(\widetilde{P}\varepsilon^i,\varepsilon^i)=|\lambda_i|(\tau-1)q_{00}(\varepsilon_N^i)^2-|\lambda_i|q_{11}(\varepsilon_{N-1}^i)^2-|\lambda_i|\tau q_{00}(R\varepsilon^{\rm I})_N^i\varepsilon_N^i-|\lambda_i|(q_{01}+q_{10})\varepsilon_{N-1}^i(R\varepsilon^{\rm I})_N^i$$

$$-|\lambda_i|\left[q_{NN}(\varepsilon_0^i)^2+(q_{N-1N}+q_{NN-1})\varepsilon_0^i\varepsilon_1^i+q_{N-1N-1}(\varepsilon_1^i)^2\right]+(\widetilde{P}\mathbf{T}^i,\varepsilon^i),\qquad k+1\le i\le r$$

We multiply the first equation of (3.1.15) by $\dfrac{\|R\|}{\lambda_i}$ and sum up from $i=0$ to k and we multiply the second one by $\dfrac{\|L\|}{|\lambda_i|}$ and sum up from $i=k+1$ to r. Assuming that q_{N-1N-1} is positive we can rewrite these equations as follows:

$$\frac{d}{dt}\sum_{i=0}^{k}\frac{\|R\|}{\lambda_i}(P\varepsilon^i,\varepsilon^i)+\frac{d}{dt}\sum_{i=k+1}^{r}\frac{\|L\|}{|\lambda_i|}(\widetilde{P}\varepsilon^i,\varepsilon^i)=$$

$$\sum_{i=0}^{k}\Big[\,\|R\|\,(\tau-1)q_{00}(\varepsilon_0^i)^2-\|R\|\,q_{11}(\varepsilon_1^i)^2-\|R\|\,\tau q_{00}(L\varepsilon^{\rm II})_0^i\varepsilon_0^i$$

$$-\|R\|\,(q_{01}+q_{10})\varepsilon_1^i(L\varepsilon^{\rm II})_0^i-\|R\|\left(\frac{q_{N-1N}+q_{NN-1}}{2\sqrt{q_{N-1N-1}}}\varepsilon_N^i+\sqrt{q_{N-1N-1}}\varepsilon_{N-1}^i\right)^2$$

$$-\|R\|\left(q_{NN}-\frac{(q_{N-1N}+q_{NN-1})^2}{4q_{N-1N-1}}\right)(\varepsilon_N^i)^2\Big]+\sum_{i=0}^{k}\frac{\|R\|}{\lambda_i}(P\mathbf{T}^i,\varepsilon^i)$$

$$+\sum_{i=k+1}^{r}\Big[\,\|L\|\,(\tau-1)q_{00}(\varepsilon_N^i)^2-\|L\|\,q_{11}(\varepsilon_{N-1}^i)^2-\|L\|\,\tau q_{00}(R\varepsilon^{\rm I})_N^i\varepsilon_N^i$$

$$-\|L\|\,(q_{01}+q_{10})\varepsilon_{N-1}^i(R\varepsilon^{\rm I})_N^i-\|L\|\left(\frac{q_{N-1N}+q_{NN-1}}{2\sqrt{q_{N-1N-1}}}\varepsilon_0^i+\sqrt{q_{N-1N-1}}\varepsilon_1^i\right)^2$$

$$-\|L\|\left(q_{NN}-\frac{(q_{N-1N}+q_{NN-1})^2}{4q_{N-1N-1}}\right)(\varepsilon_0^i)^2\Big]+\sum_{i=k+1}^{r}\frac{\|L\|}{|\lambda_i|}(\widetilde{P}\mathbf{T}^i,\varepsilon^i)$$

Again we require the expression $q_{NN}\varepsilon_0^2+(q_{N-1N}+q_{NN-1})\varepsilon_0\varepsilon_1+q_{NN}\varepsilon_1^2$ be positive for all

$\varepsilon_0, \varepsilon_1 \in \mathbf{R}$. It implies

$$q_{N-1N-1} > 0, \quad q_{NN} - \frac{(q_{N-1N} + q_{NN-1})^2}{4q_{N-1N-1}} > 0. \tag{3.1.16}$$

We next define a new discrete scalar products:

$$[\varepsilon^{\mathrm{I}}, \varepsilon^{\mathrm{I}}]_m = \sum_{i=1}^{k} \varepsilon_m^i \varepsilon_m^i$$

$$[\varepsilon^{\mathrm{II}}, \varepsilon^{\mathrm{II}}]_m = \sum_{i=k+1}^{r} \varepsilon_m^i \varepsilon_m^i \tag{3.1.17}$$

Replacing the sums in the last equation with these vector operations and using the properties of the matrices P and $\widetilde{P}$ we get an estimate for the discrete norm $\|\mathcal{E}\|$:

$$\begin{aligned} c_0 \frac{d}{dt} \|\mathcal{E}\|^2 \leq\ & (\tau - 1) q_{00} \|R\| [\varepsilon^{\mathrm{I}}, \varepsilon^{\mathrm{I}}]_0 - \|R\| q_{11} [\varepsilon^{\mathrm{I}}, \varepsilon^{\mathrm{I}}]_1 - \|R\| \tau q_{00} [L\varepsilon^{\mathrm{II}}, \varepsilon^{\mathrm{I}}]_0 - \beta \|R\| [\varepsilon^{\mathrm{I}}, \varepsilon^{\mathrm{I}}]_N \\ & + (\tau - 1) q_{00} \|L\| [\varepsilon^{\mathrm{II}}, \varepsilon^{\mathrm{II}}]_N - \|L\| q_{11} [\varepsilon^{\mathrm{II}}, \varepsilon^{\mathrm{II}}]_{N-1} - \|L\| \tau q_{00} [R\varepsilon^{\mathrm{I}}, \varepsilon^{\mathrm{II}}]_N - \beta \|L\| [\varepsilon^{\mathrm{II}}, \varepsilon^{\mathrm{II}}]_0 \\ & - 2 \|R\| \mathbf{q}_{01} \sum_{i=1}^{k} (\varepsilon^{\mathrm{I}})_1^i (L\varepsilon^{\mathrm{II}})_0^i - 2 \|L\| \mathbf{q}_{01} \sum_{i=k+1}^{r} (\varepsilon^{\mathrm{II}})_{N-1}^i (R\varepsilon^{\mathrm{I}})_N^i \\ & + \sum_{i=0}^{k} \frac{\|R\|}{\lambda_i} (P\mathbf{T}^i, \varepsilon^i) + \sum_{i=k+1}^{r} \frac{\|L\|}{|\lambda_i|} (\widetilde{P}\mathbf{T}^i, \varepsilon^i) \end{aligned}$$

where

$$\mathbf{q}_{01} = \frac{1}{2}(q_{01} + q_{10}), \qquad \beta = q_{NN} - \frac{(q_{N-1N} + q_{NN-1})^2}{4q_{N-1N-1}} > 0.$$

Substituting the estimates

$$[L\varepsilon^{\mathrm{II}}, \varepsilon^{\mathrm{I}}]_0 \ \leq\ \|L\| \cdot \|\varepsilon^{\mathrm{II}}\|_0 \cdot \|\varepsilon^{\mathrm{I}}\|_0,$$

$$[R\varepsilon^{\mathrm{I}}, \varepsilon^{\mathrm{II}}]_N \ \leq\ \|R\| \cdot \|\varepsilon^{\mathrm{I}}\|_N \cdot \|\varepsilon^{\mathrm{II}}\|_N,$$

$$\sum_{i=1}^{k} (\varepsilon^{\mathrm{I}})_1^i (L\varepsilon^{\mathrm{II}})_0^i \ \leq\ \sqrt{\sum_{i=1}^{k} [(\varepsilon^{\mathrm{I}})_1^i]^2 \cdot \sum_{i=1}^{k} [(L\varepsilon^{\mathrm{II}})_0^i]^2} \ \leq\ \|L\| \cdot \|\varepsilon^{\mathrm{I}}\|_1 \cdot \|\varepsilon^{\mathrm{II}}\|_0,$$

$$\sum_{i=1}^{k} (\varepsilon^{\mathrm{I}})_1^i (L\varepsilon^{\mathrm{II}})_0^i \ \leq\ \sqrt{\sum_{i=k+1}^{r} \left[(\varepsilon^{\mathrm{II}})_{N-1}^i\right]^2 \cdot \sum_{i=k+1}^{r} [(R\varepsilon^{\mathrm{I}})_N^i]^2} \ \leq\ \|R\| \cdot \|\varepsilon^{\mathrm{I}}\|_N \cdot \|\varepsilon^{\mathrm{II}}\|_{N-1}$$

where

$$\|\varepsilon^{\rm I}\|_m \;=\; \sqrt{[\varepsilon^{\rm I},\varepsilon^{\rm I}]_m},$$

$$\|\varepsilon^{\rm II}\|_m \;=\; \sqrt{[\varepsilon^{\rm II},\varepsilon^{\rm II}]_m}, \qquad m=0,1,N-1,N$$

into the last inequality for $\|\mathcal{E}\|$, and collecting like terms yields

$$\begin{aligned} c_0\frac{d}{dt}\|\mathcal{E}\|^2 \;\le\; & \Big\{(\tau-1)q_{00}\cdot\|R\|\cdot\|\varepsilon^{\rm I}\|_0^2 - \|R\|\,q_{11}\|\varepsilon^{\rm I}\|_1^2 + |\tau q_{00}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_0\cdot\|\varepsilon^{\rm II}\|_0 \\ & +2|\mathbf{q}_{01}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_1\cdot\|\varepsilon^{\rm II}\|_0 - \beta\,\|L\|\cdot\|\varepsilon^{\rm II}\|_0^2\Big\} \\ & +\Big\{(\tau-1)q_{00}\cdot\|R\|\cdot\|\varepsilon^{\rm II}\|_N^2 - \|R\|\,q_{11}\|\varepsilon^{\rm I}\|_{N-1}^2 + |\tau q_{00}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_N\cdot\|\varepsilon^{\rm II}\|_N \\ & +2|\mathbf{q}_{01}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_N\cdot\|\varepsilon^{\rm II}\|_{N-1} - \beta\,\|L\|\cdot\|\varepsilon^{\rm I}\|_N^2\Big\} \\ & +\sum_{i=0}^{k}\frac{\|R\|}{\lambda_i}(P\mathbf{T}^i,\varepsilon^i) + \sum_{i=k+1}^{r}\frac{\|L\|}{|\lambda_i|}(\tilde{P}\mathbf{T}^i,\varepsilon^i) \end{aligned} \tag{3.1.18}$$

We require now each curly bracket to be negative. Thus we need

$$\begin{aligned} & (\tau-1)q_{00}\cdot\|R\|\cdot\|\varepsilon^{\rm I}\|_0^2 - \|R\|\,q_{11}\|\varepsilon^{\rm I}\|_1^2 + |\tau q_{00}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_0\cdot\|\varepsilon^{\rm II}\|_0 \\ & +2|\mathbf{q}_{01}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_1\cdot\|\varepsilon^{\rm II}\|_0 - \beta\,\|L\|\cdot\|\varepsilon^{\rm II}\|_0^2 \;\le 0 \end{aligned}$$

and also

$$\begin{aligned} & (\tau-1)q_{00}\cdot\|R\|\cdot\|\varepsilon^{\rm II}\|_N^2 - \|R\|\,q_{11}\|\varepsilon^{\rm I}\|_{N-1}^2 + |\tau q_{00}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_N\cdot\|\varepsilon^{\rm II}\|_N \\ & +2|\mathbf{q}_{01}|\cdot\|R\|\cdot\|L\|\cdot\|\varepsilon^{\rm I}\|_N\cdot\|\varepsilon^{\rm II}\|_{N-1} - \beta\,\|L\|\cdot\|\varepsilon^{\rm I}\|_N^2 \;\le 0 \end{aligned}$$

for all $\varepsilon^{\rm I},\varepsilon^{\rm II}\in\mathbf{R}$.

It is possible to show that both inequalities are satisfied (and hence the algorithm is time stable) if

$$q_{11} > 0, \quad q_{N-1N-1} > 0, \quad (\tau - 1)q_{00} < 0,$$

$$\beta = q_{NN} - \frac{(q_{N-1N} + q_{NN-1})^2}{4q_{N-1N-1}} > 0, \tag{3.1.19}$$

$$\frac{1}{4}\tau^2 q_{11} q_{00}^2 \,\|R\|\cdot\|L\| + (\tau - 1)q_{00}\left(\beta q_{11} - \mathbf{q}_{01}^2 \,\|R\|\cdot\|L\|\right) < 0.$$

Assuming for the moment that these inequalities hold we can write

$$c_0 \frac{d}{dt}\|\mathcal{E}\|^2 \le \sum_{i=0}^{k} \frac{\|R\|}{\lambda_i}(P\mathbf{T}^i, \varepsilon^i) + \sum_{i=k+1}^{r} \frac{\|L\|}{|\lambda_i|}(\tilde{P}\mathbf{T}^i, \varepsilon^i) \tag{3.1.20}$$

Using (1.1.9) and the definition (3.1.12) of the discrete norms we get

$$c_0 \frac{d}{dt}\|\mathcal{E}\|^2 \le c_1 \,\|\mathbf{T}\|\|\mathcal{E}\| \tag{3.1.21}$$

and after dividing by $\|\mathcal{E}\|$

$$\frac{d}{dt}\|\mathcal{E}\| \le \frac{c_1}{c_0}\,\|\mathbf{T}\| \tag{3.1.22}$$

leading to

$$\|\mathcal{E}\| \le \frac{c_1}{c_0}\,\|\mathbf{T}\|\, t. \tag{3.1.23}$$

We are ready now to formulate the theorem:

Theorem 3.1.1 *Let the method defined by equation (3.1.8) satisfies (3.1.19), for the discretization of the hyperbolic system (3.1.1) with initial and boundary conditions (3.1.2),(3.1.3). Then it is stable and leads to an error whose norm is growing linearly in time.*

Remark.

We recall that in order to solve the hyperbolic system numerically we use the same matrices $P, Q, \tilde{P}, \tilde{Q}$ and the same vectors $\vec{S}_0, \vec{S}_N$ as in the scalar case, i.e.

$$q_{00} = -\frac{2}{3}, \quad q_{11} = \frac{1}{6} > 0, \quad \mathbf{q}_{01} = \frac{1}{3},$$

$$\beta = q_{NN} - \frac{(q_{N-1N} + q_{NN-1})^2}{4q_{N-1N-1}} = \frac{1}{6} > 0.$$

With this choice of the matrix Q the inequalities (3.1.19) hold if

$$\frac{1- \| R \|\cdot\| L \| -\sqrt{D}}{2 \| R \|\cdot\| L \|} \leq \tau \leq \frac{1- \| R \|\cdot\| L \| +\sqrt{D}}{2 \| R \|\cdot\| L \|} \tag{3.1.24}$$

where

$$D = (1- \| R \|\cdot\| L \|)(1 - 5 \| R \|\cdot\| L \|).$$

We can choose τ , which satisfies (3.1.24), if $D \geq 0$. This happens if $\| R \|\cdot\| L \| \leq 1/5$. But this is only a sufficient condition, because numerical experiments (see discussion in the next section) show that the numerical solution converges to the analytical solution for all $t < \infty$ even if $1/5 < \| R \|\cdot\| L \| \leq 1$.

Similarly, in the case of the fourth-order scheme,

$$q_{00} = -\frac{5}{8}, \quad q_{11} = \frac{1}{8} > 0, \quad \mathbf{q}_{01} = \frac{1}{4},$$

$$\beta = q_{NN} - \frac{(q_{N-1N} + q_{NN-1})^2}{4q_{N-1N-1}} = \frac{1}{4} > 0.$$

leading to

$$\frac{4 - 2 \| R \|\cdot\| L \| -2\sqrt{D}}{5 \| R \|\cdot\| L \|} \leq \tau \leq \frac{4 - 2 \| R \|\cdot\| L \| +2\sqrt{D}}{5 \| R \|\cdot\| L \|} \tag{3.1.25}$$

where

$$D = (2- \| R \|\cdot\| L \|)(2 - 6 \| R \|\cdot\| L \|).$$

We can find τ , which satisfies (3.1.25), if $\| R \| \cdot \| L \| \leq 1/3$. Numerical experiments performed in the next section show that the fourth-order scheme is time stable even if $1/3 < \| R \|\cdot\| L \| \leq 1$.

3.2 Numerical experiments

Consider the hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + A\frac{\partial \mathbf{u}}{\partial x} = 0, \quad 0 \le x \le 1, \quad t \ge 0 \tag{3.2.1}$$

where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix} \tag{3.2.2}$$

with initial data

$$u(x,0) = \sin 2\pi x, \qquad v(x,0) = -\sin 2\pi x, \qquad 0 \le x \le 1, \tag{3.2.3}$$

and boundary conditions

$$u(0,t) = v(0,t), \qquad v(1,t) = u(1,t), \qquad t \ge 0. \tag{3.2.4}$$

The exact solution is

$$\begin{aligned} u(x,t) &= \sin 2\pi(x-t), \\ v(x,t) &= -\sin 2\pi(x+t), \qquad 0 \le x \le 1, \quad t \ge 0. \end{aligned} \tag{3.2.5}$$

Note that due to (3.2.4), $\| R \| \cdot \| L \| = 1$ and thus we test the most severe reflection case.

As in the scalar case we solved the problem (3.2.1) - (3.2.4) numerically using two different schemes: fourth-order compact with third-order boundary closure and sixth-order compact with fifth-order boundary closure. And again we compare two methods for implementation of the boundary conditions: (i) conventional - which implies the overwriting of the value of the solution at the boundary point with the analytic boundary condition after each Runge-Kutta stage and (ii) the SAT method described in the previous section. In all cases, the standard fourth-order Runge-Kutta method is used for time integration, with a suitable Δt such that the desired overall accuracy is maintained.

Conventional boundary conditions. In Chapter 1 (Part I) it was shown that for the scalar case the fourth-order scheme is time-stable while the sixth-order scheme is not when using conventional implementation of boundary conditions. Using these schemes for solving the test problem (3.2.1) - (3.2.4) we found that both scheme failed to be time-stable when applied to system of equations. Figure (3.1), (3.2) show L_2-error as function of time for the fourth-order compact scheme and sixth-order compact scheme respectively for different grids. As one can see results diverge exponentially from the analytic solution.

On the other hand, we shall show that SAT procedure ensures time-stability (only a sub-linear temporal growth) for the hyperbolic system, for both the fourth- and the sixth-order schemes.

SAT boundary conditions. First of all we verify that SAT implementation of boundary conditions retains the formal accuracy of the spatial operator. Results of the grid convergence study of the spatial operators with SAT parameter $\tau = 2$ for both orders of accuracy are presented in table (3.1). The absolute error $\log_{10}(L_2)$ at a fixed time $t = \mathrm{T}$ and the convergence rate are the entries.. The convergence rate is computed as

$$\log_{10}\left(\frac{\| \mathbf{u} - \mathbf{u}^{h_1}\|_2}{\| \mathbf{u} - \mathbf{u}^{h_2}\|_2}\right) / \log_{10}\left(\frac{h_1}{h_2}\right), \tag{3.2.6}$$

where $\mathbf{u} = (\mathbf{u}(x_0,t), \mathbf{u}(x_2,t), \ldots, \mathbf{u}(x_N,t))^T$ is the exact solution, $\mathbf{u}^h$ is the numerical solution with mesh width h, and $\| \mathbf{u} - \mathbf{u}^h\|_2$ is the discrete L_2 norm of the absolute error. The data in this table indicate that the convergence rate asymptotically approaches to the theoretical value of 4 for the fourth-order operator and to 6 for the sixth-order operator. Figure (3.3), (3.4) show the error as a function of time for long time integration using the fourth-order and the sixth-order difference operators respectively for different grids. No exponential growth exists and both schemes are found to be strictly stable. In figure (3.5), (3.6) the eigenvalue spectrum for both schemes for different grids are shown. One can see that there are no eigenvalues with positive real part.

	Fourth-order compact		Sixth-order compact	
Grid	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate
21	-2.657		-4.371	
31	-3.332	3.83	-5.462	6.19
41	-3.817	3.89	-6.231	6.15
61	-4.506	3.91	-7.299	6.07
81	-4.998	3.94	-8.041	5.97

Table 3.1: Grid convergence of two high-order schemes for $u_t + Au_x = 0$, using SAT implementation of boundary conditions with the SAT parameter $\tau = 2$ and CFL $= 0.5$ for the fourth-order scheme, CFL $= 0.1$ for the sixth-order scheme. $T = 10$.

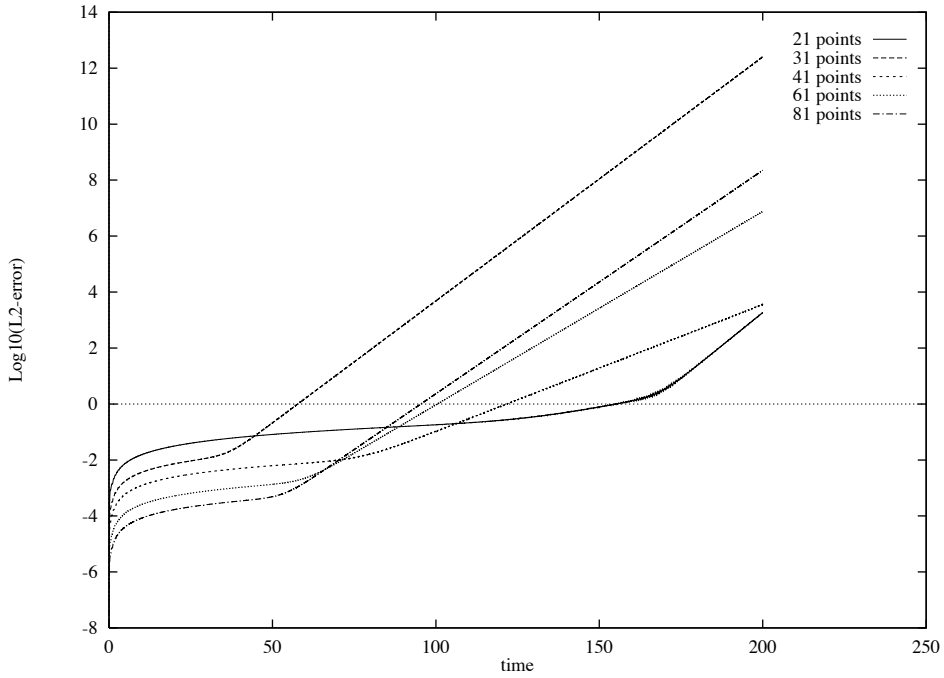

Figure 3.1: The L_2-error as a function of time for the fourth-order approximation using conventional implementation of boundary conditions with $\mathrm{CFL} = 0.5$.

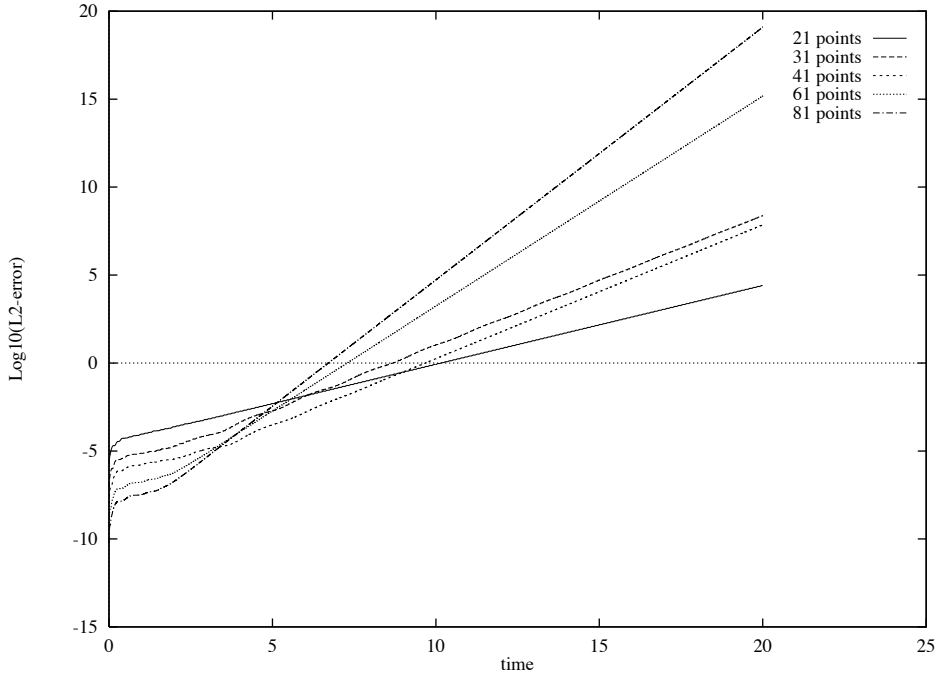

Figure 3.2: The L_2-error as a function of time for the sixth-order approximation using conventional implementation of boundary conditions with $\mathrm{CFL} = 0.1$.

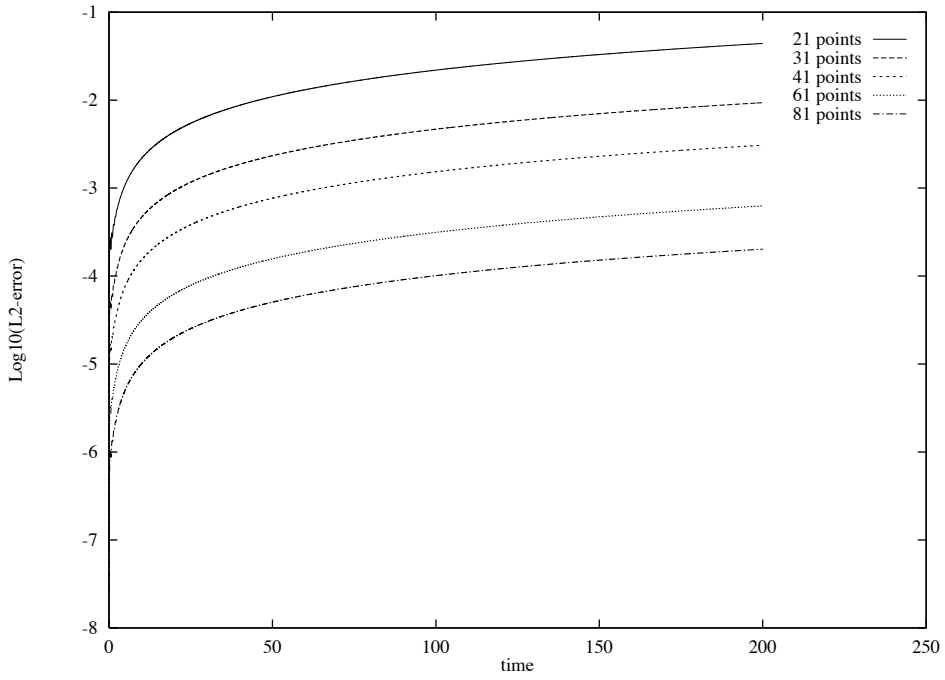

Figure 3.3: The L_2-error as a function of time for the fourth-order approximation using SAT method for implementation of boundary conditions with $\tau = 2,\ \mathrm{CFL} = 0.5$.

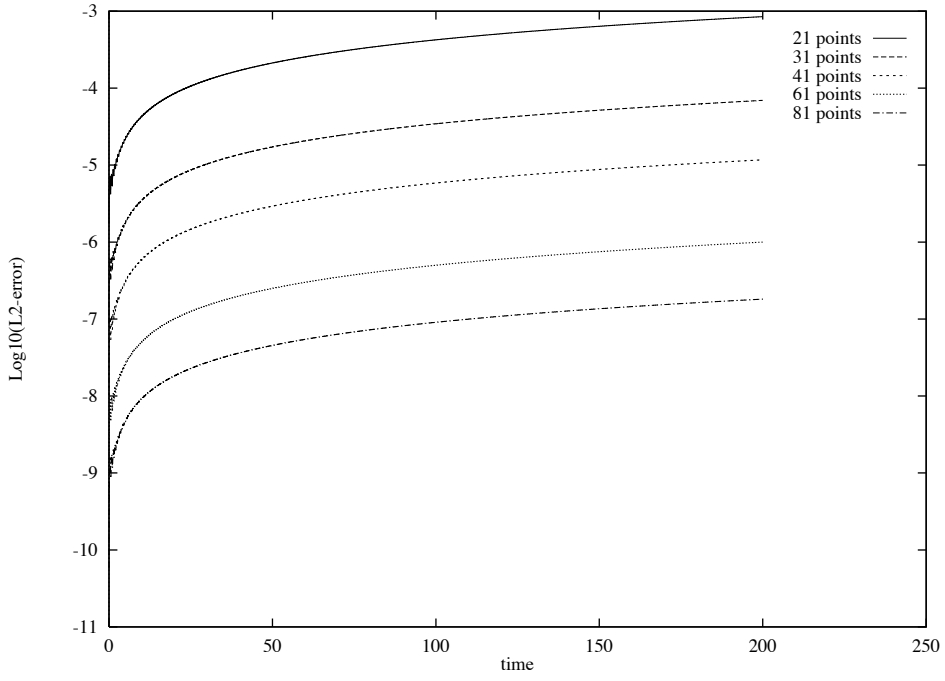

Figure 3.4: The L_2-error as a function of time for the sixth-order approximation using SAT method for implementation of boundary conditions with $\tau = 2,\ \mathrm{CFL} = 0.1$.

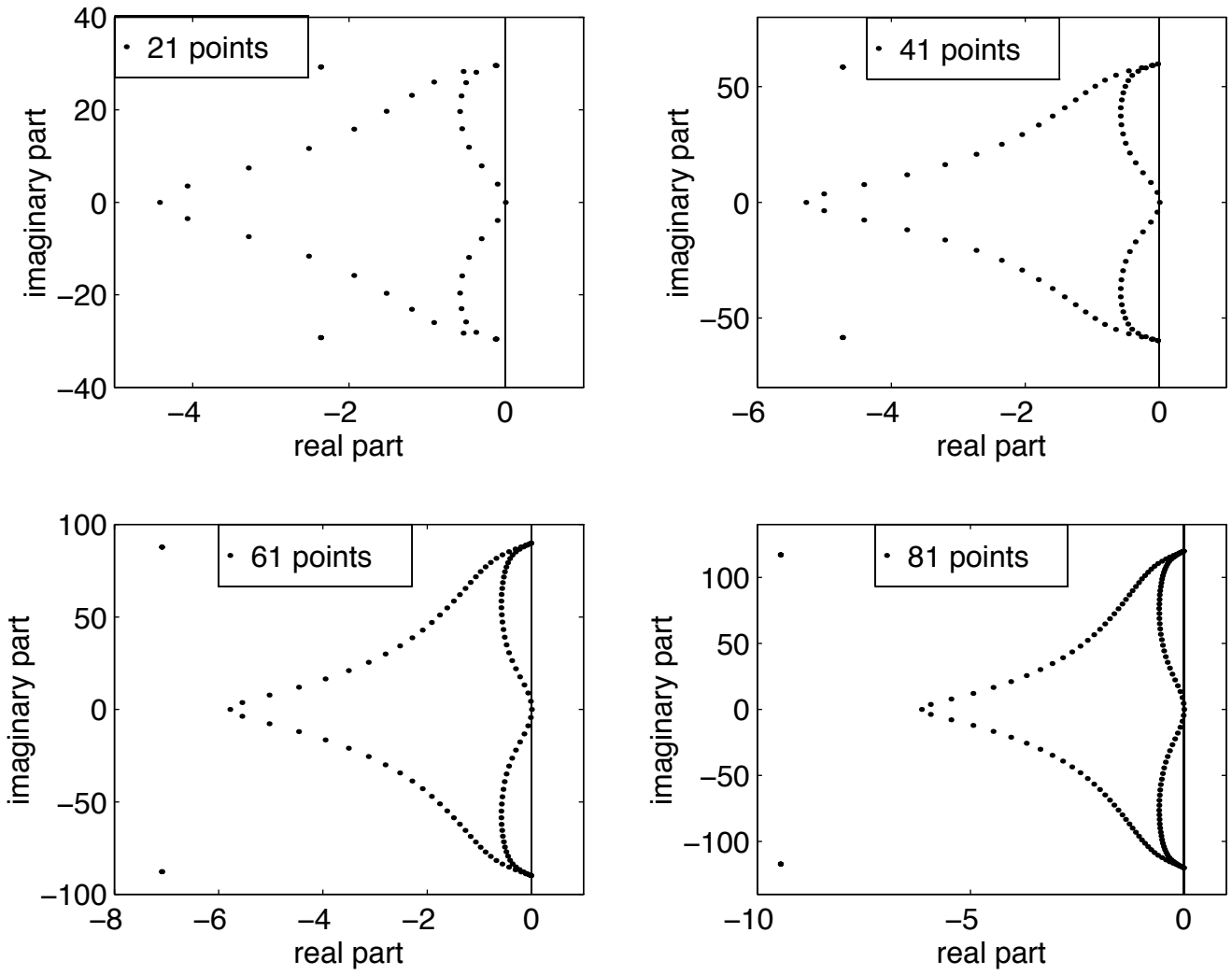

Figure 3.5: Semi-discrete eigenvalue spectrum for the fourth-order approximation using SAT method for implementation of boundary conditions with $\tau = 2$.

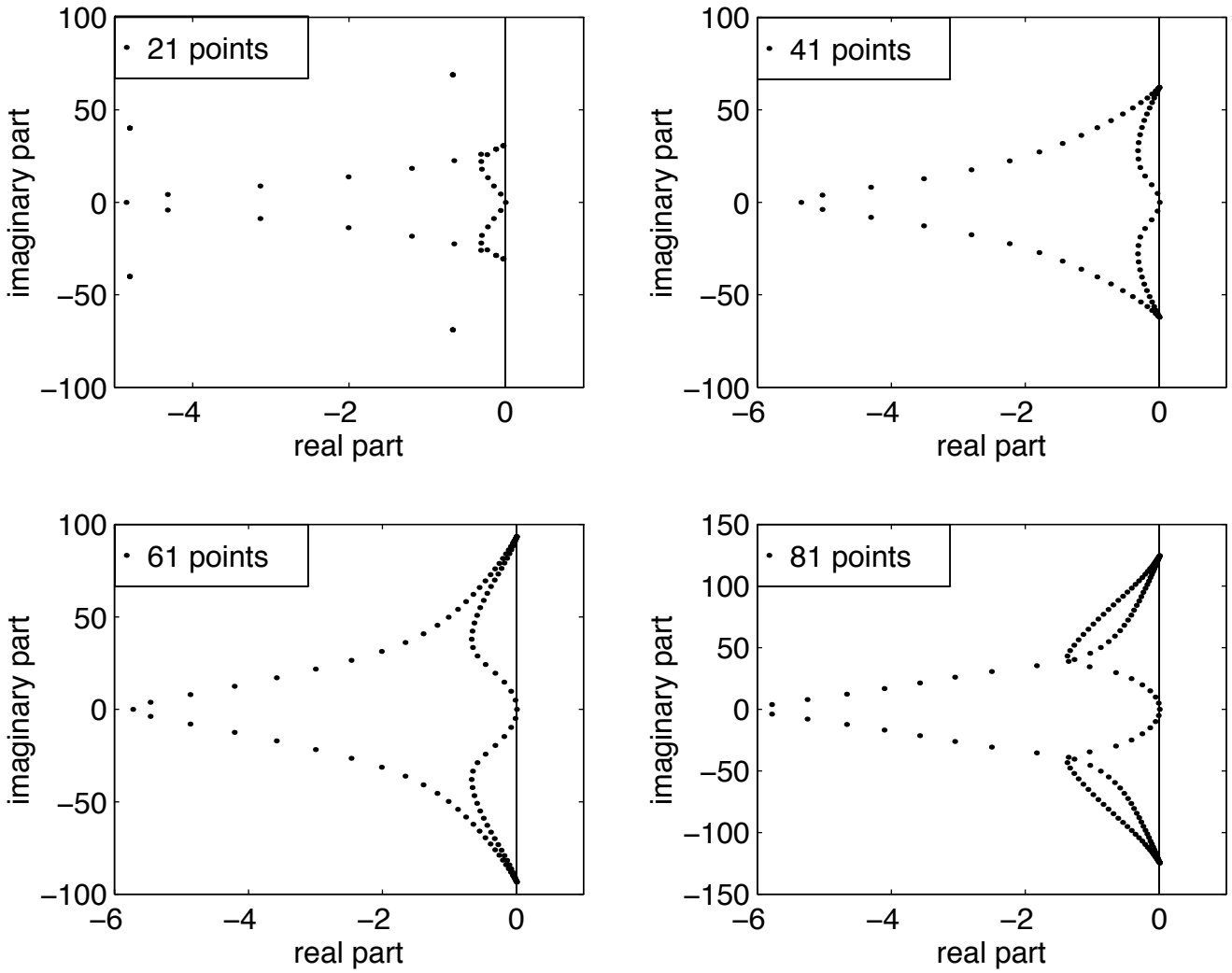

Figure 3.6: Magnification of semi-discrete eigenvalue spectrum close to imaginary axis for the sixth-order approximation using SAT method for implementation of boundary conditions with $\tau = 2$.

Chapter 4

2-D Hyperbolic Systems

4.1 Application to Maxwell's equations

As an application where high-order accurate approximation are needed we consider Maxwell's equations. In free space they are given by:

$$
\begin{aligned}
\frac{\partial \mathrm{B}}{\partial \tilde{t}} + \nabla \times \mathrm{E} &= 0, \qquad \text{(Faraday's law)} \\
\frac{\partial \mathrm{D}}{\partial \tilde{t}} - \nabla \times \mathrm{H} &= -\mathrm{J}, \qquad \text{(Ampere's law)} \\
\mathrm{B} &= \mu \mathrm{H}, \\
\mathrm{D} &= \epsilon \mathrm{E},
\end{aligned}
\tag{4.1.1}
$$

coupled with Gauss's law

$$
\begin{aligned}
\nabla \cdot \mathrm{B} &= 0, \\
\nabla \cdot \mathrm{D} &= 0.
\end{aligned}
\tag{4.1.2}
$$

If we assume perfectly conducting conditions at the outer edge of the domain then the boundary conditions are:

$$
\begin{aligned}
\vec{n} \times \mathrm{E} &= 0, \\
\vec{n} \cdot \mathrm{H} &= 0
\end{aligned}
\tag{4.1.3}
$$

where $\vec{n}$ is a normal vector to the surface of the domain.

To simplify the notation we shall consider the two dimensional case with ϵ, μ constants and $J = 0$. We nondimensionalize the variables: $t = c\tilde{t}/L$, $x = \tilde{x}/L$, $y =$

$\tilde{y}/L$, $E = \mathrm{E}$, $H = \sqrt{\frac{\epsilon}{\mu}}\mathrm{H}$, where ϵ and μ are the permittivity and permeability coefficients, in free space, respectively, c is the speed of light and L is a length of the domain. The 2-D version of system (4.1.1), (4.1.2) decouples into two independent sets of equations. We shall consider the TM (Transverse magnetic) system in a square domain $\Omega = \{(x,y) \in \mathbf{R}^2 \mid 0 \le x \le 1,\ 0 \le y \le 1\}$. The TM equations then become:

$$
\begin{aligned}
\frac{\partial E_z}{\partial t} &= \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \qquad (x,y) \in \Omega, \quad t \ge 0 \\
\frac{\partial H_x}{\partial t} &= -\frac{\partial E_z}{\partial y} \qquad (4.1.4) \\
\frac{\partial H_y}{\partial t} &= \frac{\partial E_z}{\partial x}
\end{aligned}
$$

with the boundary conditions

$$
\begin{aligned}
E_z(0,y,t) &= E_z(1,y,t) = 0, \quad t \ge 0, \\
E_z(x,0,t) &= E_z(x,1,t) = 0. \qquad (4.1.5)
\end{aligned}
$$

We take as initial conditions,

$$
\begin{aligned}
E_z(x,y,0) &= \sin(\omega_1 x)\sin(\omega_2 y), \qquad (x,y) \in \Omega, \\
H_x(x,y,0) &= 0, \qquad (4.1.6) \\
H_y(x,y,0) &= 0,
\end{aligned}
$$

where $\omega_1 = \pi n$ and $\omega_2 = \pi m$ ($n, m = \pm 1, \pm 2, \pm 3, \ldots$).

The exact solution is

$$
\begin{aligned}
E_z(x,y,t) &= \sin(\omega_1 x)\sin(\omega_2 y)\cos(\omega t), \\
H_x(x,y,t) &= -\frac{\omega_2}{\omega}\sin(\omega_1 x)\cos(\omega_2 y)\sin(\omega t), \qquad (4.1.7) \\
H_y(x,y,t) &= \frac{\omega_1}{\omega}\cos(\omega_1 x)\sin(\omega_2 y)\sin(\omega t),
\end{aligned}
$$

where $\omega = \sqrt{\omega_1^2 + \omega_2^2}$.

Matrix form of the equations (4.1.4) is:

$$\frac{\partial}{\partial t}\begin{pmatrix} E_z \\ H_x \\ H_y \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} E_z \\ H_x \\ H_y \end{pmatrix} + \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\frac{\partial}{\partial y}\begin{pmatrix} E_z \\ H_x \\ H_y \end{pmatrix}$$

$$= A_1\frac{\partial}{\partial x}\begin{pmatrix} E_z \\ H_x \\ H_y \end{pmatrix} + A_2\frac{\partial}{\partial y}\begin{pmatrix} E_z \\ H_x \\ H_y \end{pmatrix}, \tag{4.1.8}$$

where

$$A_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The SAT method for implementation of boundary conditions is used for diagonalized systems in one dimension. We encounter a problem when dealing with the two dimensional problem, because it is impossible to diagonalize the two matrices A_1 and A_2 simultaneously. To overcome this problem we consider the two dimensional Maxwell's equations (4.1.4) in each space dimension independently. We decompose (4.1.8) into the following one dimensional Maxwell's equations†:

$$\frac{\partial}{\partial t}\begin{pmatrix} E_z \\ H_y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} E_z \\ H_y \end{pmatrix}, \tag{4.1.9}$$

$$\frac{\partial}{\partial t}\begin{pmatrix} E_z \\ H_x \end{pmatrix} = -\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\frac{\partial}{\partial y}\begin{pmatrix} E_z \\ H_x \end{pmatrix}, \tag{4.1.10}$$

with $E_z = 0$ at the boundaries (see (4.1.5)), and we denote

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

We shall limit our detailed discussion only to equation (4.1.9). The treatment of the equation (4.1.10) is similar.

We diagonalize the matrix A and change the variables. Let M be a diagonalizing matrix of A and Λ a diagonal matrix having the eigenvalues of A, i.e.

$$M^{-1}AM = \Lambda = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.1.11}$$

†This decomposition is not, of course, equivalent to the original system (4.1.8). It is done for lack of a 2-D characteristic theory. This practice follows what has been done previously in the context of 2-D gas dynamics, see [11].

and

$$M = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad M^{-1} = \frac{1}{2}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{4.1.12}$$

(4.1.9) is transformed to

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix}, \tag{4.1.13}$$

where

$$\begin{pmatrix} u \\ v \end{pmatrix} = M^{-1}\begin{pmatrix} E_z \\ H_x \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -E_z + H_y \\ E_z + H_y \end{pmatrix}$$

The boundary conditions can be written as

$$u(0,y,t) = v(0,y,t), \qquad v(1,y,t) = u(1,y,t) \tag{4.1.14}$$

This is equivalent to the requirement of $E_z = 0$ on the boundaries. Note also that (4.1.14) is in the form (3.1.3) with $g^{\mathrm{I}}(t) = g^{\mathrm{II}}(t) = 0$ and $R = L = 1$.

We add to the system (4.1.13) an artificial zero term which is similar to the SAT term for one dimensional hyperbolic system and rewrite it as follows:

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \alpha\,[u(0,y,t) - v(0,y,t)] \\ \beta\,[v(1,y,t) - u(1,y,t)] \end{pmatrix}. \tag{4.1.15}$$

where α and β are some constants.

When we return to the original variables, i.e E_z, H_y , we get:

$$\begin{aligned} \frac{\partial}{\partial t}\begin{pmatrix} E_z \\ H_y \end{pmatrix} &= A\frac{\partial}{\partial x}\begin{pmatrix} E_z \\ H_y \end{pmatrix} + M\begin{pmatrix} \alpha\,[u(0,y,t) - v(0,y,t)] \\ \beta\,[v(1,y,t) - u(1,y,t)] \end{pmatrix} \\ &= A\frac{\partial}{\partial x}\begin{pmatrix} E_z \\ H_y \end{pmatrix} + \begin{pmatrix} -\alpha\,[u(0,y,t) - v(0,y,t)] + \beta\,[v(1,y,t) - u(1,y,t)] \\ \alpha\,[u(0,y,t) - v(0,y,t)] + \beta\,[v(1,y,t) - u(1,y,t)] \end{pmatrix} \end{aligned} \tag{4.1.16}$$

Using the fact that

$$\begin{pmatrix} E_z \\ H_y \end{pmatrix} = M\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -u + v \\ u + v \end{pmatrix}$$

we replace the boundary terms $u(0,y,t) - v(0,y,t),\ v(1,y,t) - u(1,y,t)$ in (4.1.16) by the original variables $E_z(0,y,t),\ E_z(1,y,t)$.

Thus (4.1.16) becomes

$$\frac{\partial}{\partial t}\begin{pmatrix} E_z \\ H_y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} E_z \\ H_y \end{pmatrix} + \begin{pmatrix} \alpha E_z(0,y,t) + \beta E_z(1,y,t) \\ -\alpha E_z(0,y,t) + \beta E_z(1,y,t) \end{pmatrix} \tag{4.1.17}$$

We now call the attention to the fact that the systems (4.1.9) and (4.1.17) are equivalent (see (4.1.5)).
In a similar fashion we get for E_z, H_x a system which is equivalent to (4.1.10):

$$\frac{\partial}{\partial t}\begin{pmatrix} E_z \\ H_x \end{pmatrix} = -\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\frac{\partial}{\partial y}\begin{pmatrix} E_z \\ H_x \end{pmatrix} + \begin{pmatrix} +\alpha E_z(x,0,t) + \beta E_z(x,1,t) \\ \alpha E_z(x,0,t) - \beta E_z(x,1,t) \end{pmatrix} \tag{4.1.18}$$

When we approximate the non-diagonalized equations (4.1.17) and (4.1.18) numerically by using the SAT method for implementation of boundary conditions we shall add a SAT boundary terms for both directions which resemble the artificial zero terms which appear in the equations (4.1.17), (4.1.18). Let Δx and Δy be mesh widths in the $x-$ and $y-$directions, and divide the axes into sub-intervals of length Δx and Δy respectively. For $i = 0, \ldots, N_1$ and $j = 0, \ldots, N_2$ we use the notation:

$$E_{z_{ij}}(t) = E_z(x_i, y_j, t), \quad H_{x_{ij}}(t) = H_x(x_i, y_j, t), \quad H_{y_{ij}}(t) = H_y(x_i, y_j, t),$$

$$x_i = i\Delta x, \quad y_j = j\Delta y,$$

$$N_1\Delta x = 1, \quad N_2\Delta y = 1,$$

where $E_{z_{ij}}(t)$, $H_{x_{ij}}(t)$ and $H_{y_{ij}}(t)$ are vector grid functions. We denote by $e_{z_{ij}}$, $h_{x_{ij}}$ and $h_{y_{ij}}$ the numerical approximations to the projections $E_{z_{ij}}(t)$, $H_{x_{ij}}(t)$ and $H_{y_{ij}}(t)$ respectively. Without lost of generality we take $N = N_1 = N_2$, i.e. $\Delta x = \Delta y$.

Before proceeding to the semi discrete problem let us define:

$$D_x = \widetilde{P}^{-1}\widetilde{Q}, \qquad D_y = P^{-1}Q, \tag{4.1.19}$$

where $(N+1) \times (N+1)$ matrices P, Q and $\widetilde{P}, \widetilde{Q}$ are the same matrices were used for solving hyperbolic system in one dimensional case and described in detail in Chapter 1 (Part I) and in Appendixes A and B. We note that in practice P^{-1} and $\widetilde{P}^{-1}$ are never evaluated. Rather the decomposition $P = LU$ and $\widetilde{P} = \widetilde{L}\widetilde{U}$ is calculated once for each matrix. L and U ($\widetilde{L}$ and $\widetilde{U}$) are bidiagonal matrices with one of them having "ones" along the diagonal. Hence, the inverse of L and U ($\widetilde{L}$ and $\widetilde{U}$) is very cheap (two additions and three multiples per point).

Let $[e_z]$, $[h_x]$ and $[h_y]$ be the $(N+1)\times(N+1)$ matrices with the elements $e_{z_{ij}}$, $h_{x_{ij}}$ and $h_{y_{ij}}$ respectively and denote by $[e_z]_j^R$, $[h_x]_j^R$ and $[h_y]_j^R$ the j^{th} row of each of these matrices and by $[e_z]_i^C$, $[h_x]_i^C$ and $[h_y]_i^C$ the i^{th} column of each of these matrices.

We now write the semi-discrete approximation to (4.1.17) as:

$$
\begin{aligned}
\frac{d}{dt}[e_z]_j^R &= D_x[h_y]_j^R - \tilde{P}^{-1}\left(\vec{S}_0 e_{z_{0j}} + \vec{S}_N e_{z_{Nj}}\right), \\
\frac{d}{dt}[h_y]_j^R &= D_x[e_z]_j^R - \tilde{P}^{-1}\left(-\vec{S}_0 e_{z_{0j}} + \vec{S}_N e_{z_{Nj}}\right),
\end{aligned}
\tag{4.1.20}
$$

and the semi-discrete approximation to (4.1.18) as:

$$
\begin{aligned}
\frac{d}{dt}[e_z]_i^C &= -[h_x]_i^C D_y^T + P^{-1}\left(\vec{S}_0 e_{z_{i0}} + \vec{S}_N e_{z_{iN}}\right), \\
\frac{d}{dt}[h_x]_i^C &= -[e_z]_i^C D_y^T + P^{-1}\left(\vec{S}_0 e_{z_{i0}} - \vec{S}_N e_{z_{iN}}\right),
\end{aligned}
\tag{4.1.21}
$$

where the $(N+1)$ long vectors $\vec{S}_N$ and $\vec{S}_0$ are exactly the same vectors like in one-dimensional case, i.e.

$$
\vec{S}_0 = \begin{pmatrix} \tau q_{00} \\ (q_{01}+q_{10}) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \vec{S}_N = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -(q_{01}+q_{10}) \\ -\tau q_{00} \end{pmatrix}
\tag{4.1.22}
$$

We now compose the two one dimensional systems into the two dimensional set and approximate the equations (4.1.8) in the following way:

$$
\begin{aligned}
\frac{d}{dt}[e_z] &= D_x[h_y] - \left([e_z]_0^C \vec{S}_0^T + [e_z]_N^C \vec{S}_N^T\right)\tilde{P}^{-1} \\
&\quad -[h_x]D_y^T + P^{-1}\left(\vec{S}_0[e_z]_0^R + \vec{S}_N[e_z]_N^R\right) \\
\frac{d}{dt}[h_x] &= -[e_z]D_y^T + P^{-1}\left(\vec{S}_0[e_z]_0^R - \vec{S}_N[e_z]_N^R\right) \\
\frac{d}{dt}[h_y] &= D_x[e_z] - \left(-[e_z]_0^C \vec{S}_0^T + [e_z]_N^C \vec{S}_N^T\right)\tilde{P}^{-1}
\end{aligned}
\tag{4.1.23}
$$

4.2 Maxwell's equations: Numerical simulations

The problem (4.1.4), (4.1.5), (4.1.7) was solved using both the fourth-order scheme and the sixth-order scheme. The boundary conditions are imposed using the SAT algorithm described above. In all cases, the temporal advance is via the standard fourth-order Runge-Kutta method. The time step is chosen small enough to ensure the local stability of the Runge-Kutta method and retain the desired overall accuracy. The simulations were all run to equivalent times $T = 100$ for both the fourth- and the sixth-order schemes and different grids $(N = N_1 = N_2 = 20, 40, 80)$. We choose CFL $= 1/10$, $\tau = 2$ for the fourth-order scheme, and CFL $= 1/15$, $\tau = 2$ for the sixth-order scheme. In figures (4.1)-(4.3) the log_{10} of the L_2 error is computed for both schemes and different grids. As one can see the error grows linearly in time; no exponential growth exists, indicating temporal stability of the schemes. Figure (4.5) shows the e_z component of the numerical solution at time $T = 2$ obtained by using the sixth-order scheme with $N = 80$, $\tau = 2$.

Unlike previous chapters, where we compared two procedures for imposing of boundary conditions (the conventional procedure and the SAT procedure), in this section we shall compare our results with the results obtained by E. Turkel and A. Yefet, see [35], [36]. They solved the same problem by using the Ty(2,4) scheme, which is a fourth-order compact implicit difference scheme on staggered meshes. For time integration they used the staggered leap-frog method. The Ty(2,4) algorithm was run for: $N = 20$, CFL $= 1/18$; $N = 40, 80$, CFL $= 1/44$. The log_{10} of the L_2 error, obtained by using the Ty(2,4) fourth-order scheme, is plotted in figure (4.4). Note that the "Ty" algorithm was run with a time step, Δt, almost twice smaller for $N = 20$ and almost four and a half times smaller for $N = 40, 80$ than one used for the fourth-order "SAT" scheme. It should be also observed that the results obtained by using the "SAT" schemes and presented in figures (4.1)-(4.3) are printed every Δt step while the results obtained by using the Ty(2,4) scheme and presented in figure (4.4) are printed every $1/(10\Delta t)$ steps (i.e. only a 1000 points are printed, in contrast to about 20,000-80,000 points for our printout graphs).

In order to check on the order of accuracy, the runs were repeated for $(N = N_1 = N_2 = 20, 40, 80)$. Table (4.1) shows a grid refinement study for all three spatial operators. The absolute error $\log_{10}(L_2)$ at a fixed time $t = \mathrm{T}$ and the convergence rate between two grids are plotted. The results in this table agree very well with the predicted ones for fourth- and sixth-order. We note that the error obtained by using the Ty(2,4) fourth-order scheme is smaller that the error in "SAT" fourth-order scheme, but the "SAT" sixth-order scheme outperforms both.

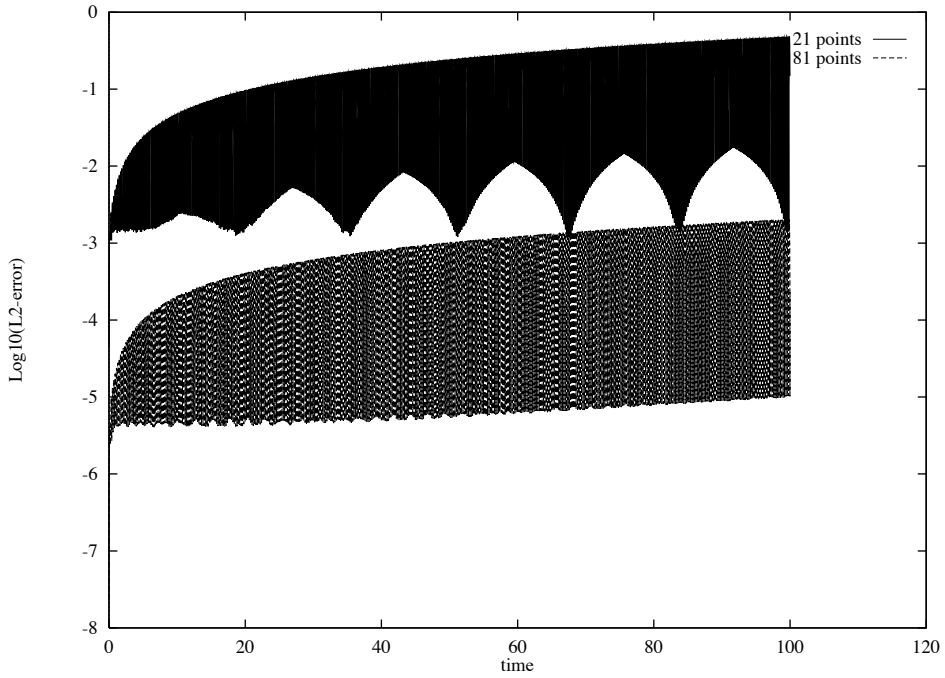

Figure 4.1: The L_2-error as a function of time for the "SAT" fourth-order approximation with CFL $= 0.1, \tau = 2,\ \omega_1 = 3\pi,\ \omega_2 = 4\pi,\ \omega = 5\pi$. N $= 20, 80$.

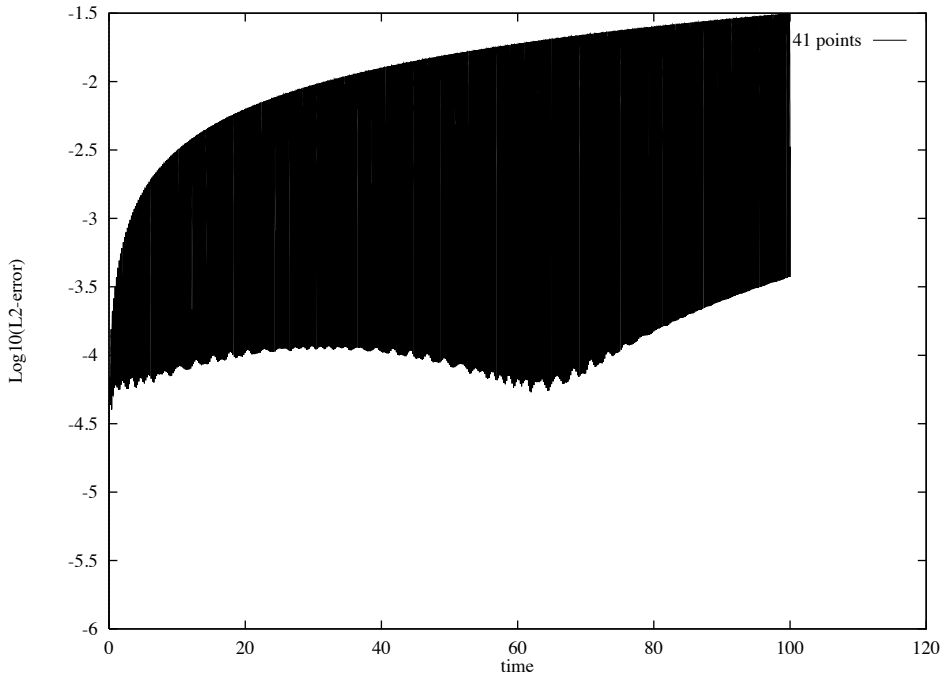

Figure 4.2: The L_2-error as a function of time for the "SAT" fourth-order approximation with CFL $= 0.1,\ \tau = 2,\ \omega_1 = 3\pi,\ \omega_2 = 4\pi,\ \omega = 5\pi$. N $= 40$.

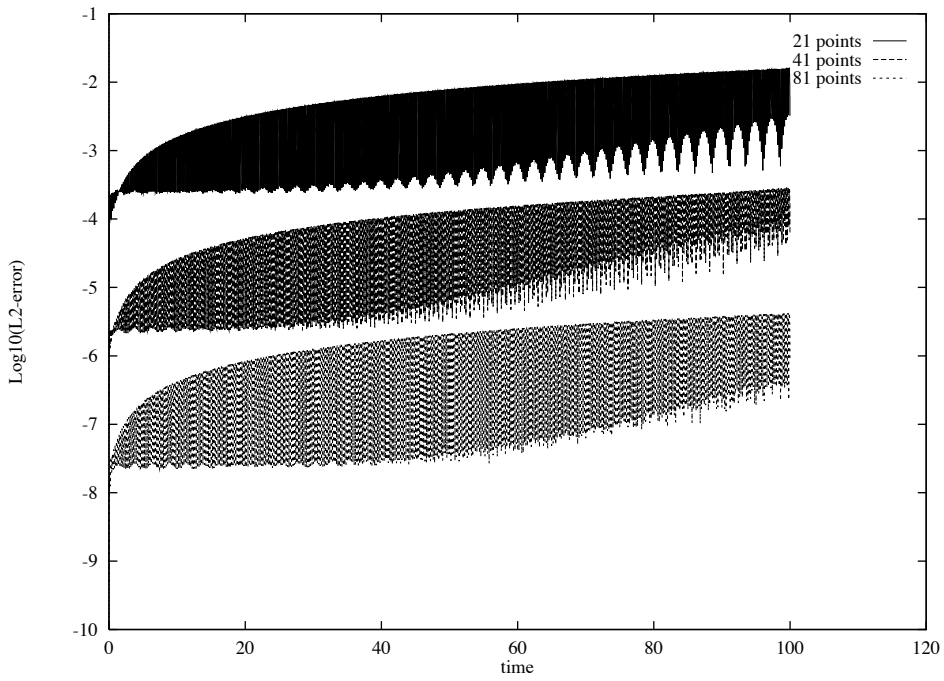

Figure 4.3: The L_2-error as a function of time for the "SAT" sixth-order approximation with CFL $= 1/15,\ \tau = 2,\ \omega_1 = 3\pi,\ \omega_2 = 4\pi,\ \omega = 5\pi$. N $= 20, 40, 80$.

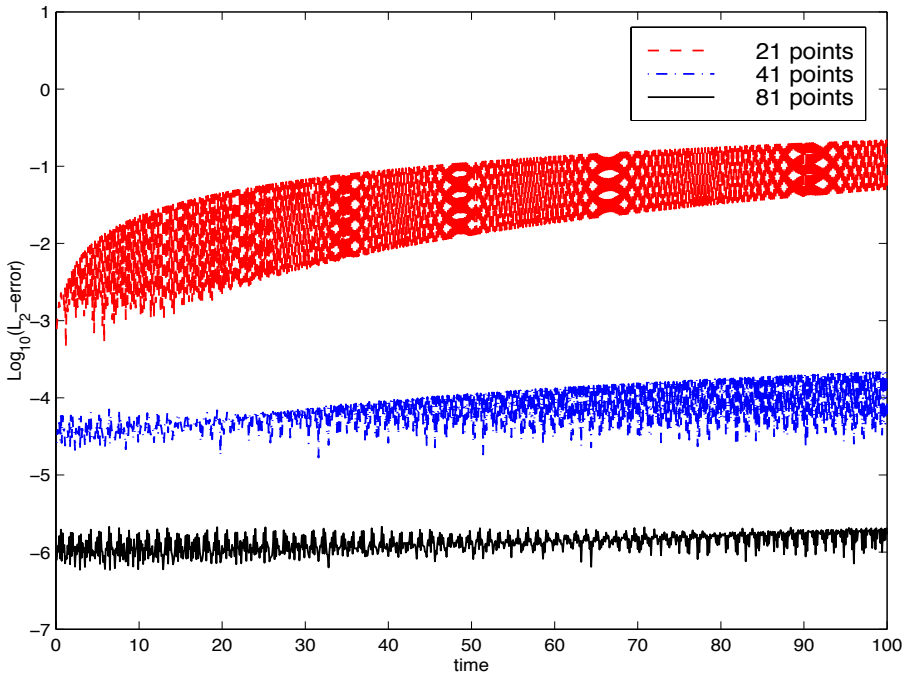

Figure 4.4: The L_2-error as a function of time for the Ty(2,4) fourth-order approximation for $N = 20$: CFL $= 1/18$; for $N = 40, 80$: CFL $= 1/44$. $\omega_1 = 3\pi,\ \omega_2 = 4\pi,\ \omega = 5\pi$.

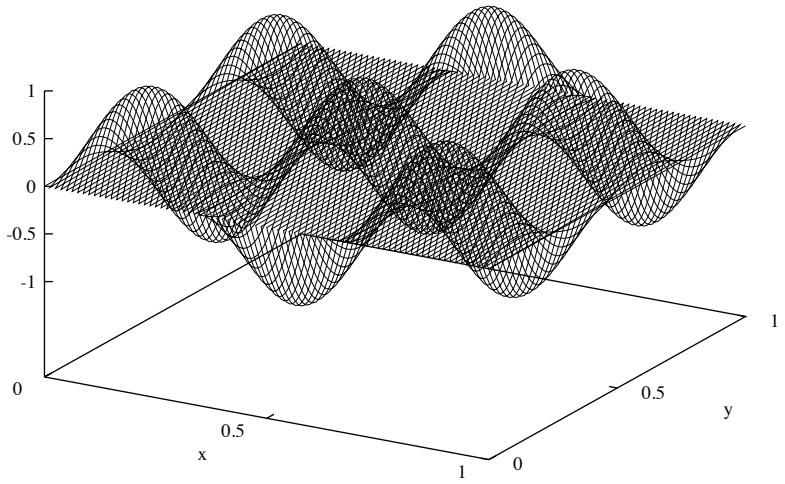

Figure 4.5: e_z component of the numerical solution at $T = 2$ obtained using "SAT" sixth-order approximation with $N = 80,\ \text{CFL} = 1/15,\ \tau = 2,\ \omega_1 = 3\pi,\ \omega_2 = 4\pi,\ \omega = 5\pi$.

	Ty(2,4) fourth-order		"SAT" fourth-order		"SAT" sixth-order	
Grid	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate	$\log_{10}(L_2)$	Rate
21	-2.677		-2.644		-3.580	
41	-4.234	5.17	-4.089	4.80	-5.416	6.10
81	-5.751	5.03	-5.326	4.11	-7.261	6.13

Table 4.1: Grid convergence of schemes for the two-dimensional Maxwell equations. $T = 10,\ \omega_1 = 3\pi,\ \omega_2 = 4\pi,\ \omega = 5\pi$. Here $\text{CFL} = 1/10$ for the "SAT" fourth-order scheme and $\text{CFL} = 1/15$ for the "SAT" sixth-order scheme. For Ty(2,4): $N = 20,\ \text{CFL} = 1/18;\ N = 40, 80,\ \text{CFL} = 1/44$.

Appendix A

Construction of the Sixth-Order Compact Scheme

Here we derive an implicit scheme for (1.1.1) which is formally sixth-order accurate. We begin with approximation of the first derivative $\frac{\partial v}{\partial x}$ at inner points in the form

$$\left(\frac{\partial v}{\partial x}\right)_j = \frac{1}{h}\left[\frac{a_1\mu\delta + a_2\mu\delta^3}{1 + b_1\delta^2 + b_2\delta^4}\right] v_j \tag{A.0.1}$$

where

$$\delta v_j = v_{j+1/2} - v_{j-1/2}, \qquad \mu v_j = \frac{v_{j+1/2} + v_{j-1/2}}{2};$$

or in the equivalent form

$$b_2\frac{\partial v_{j+2}}{\partial x} + (b_1 - 4b_2)\frac{\partial v_{j+1}}{\partial x} + (1 - 2b_1 + 6b_2)\frac{\partial v_j}{\partial x} + (b_1 - 4b_2)\frac{\partial v_{j-1}}{\partial x} + b_2\frac{\partial v_{j-2}}{\partial x} =$$
$$\frac{1}{2h}\left[\, a_2 v_{j+2} + (a_1 - 2a_2)\, v_{j+1} - (a_1 - 2a_2) v_{j-1} - a_2 v_{j-2}\, \right] \tag{A.0.2}$$

The relations between the coefficients a_1, a_2, b_1, b_2 are derived by matching the Taylor series coefficients of various orders. The first unmatched coefficient determines the formal truncation error of the approximation (A.0.1). In order to generate a sixth-order scheme we require that

$$\begin{cases} a_1 = 1 \\ \\ b_1 = \frac{1}{6}a_1 - a_2 \\ \\ \frac{1}{12}b_1 + b_2 = \frac{1}{120}a_1 + \frac{1}{4}a_2 \end{cases} \tag{A.0.3}$$

(A.0.3) will in general yield a pentadiagonal left-hand side. However the choice of $b_2 = 0$ gives a more convenient tridiagonal operator. In this case

$$b_2 = 0 \quad \Rightarrow \quad a_1 = \frac{1}{2}, \quad a_2 = \frac{1}{60}, \quad b_1 = \frac{1}{5} \tag{A.0.4}$$

It follows from this and (A.0.1) that we have a sixth-order compact scheme in interior points defined implicitly as

$$\frac{1}{3}\frac{\partial v_{j+1}}{\partial x}+\frac{\partial v_j}{\partial x}+\frac{1}{3}\frac{\partial v_{j-1}}{\partial x}=\frac{1}{36h}\left[\,-v_{j-2}-28v_{j-1}+28v_{j-1}+v_{j+1}\,\right] \tag{A.0.5}$$

Using this results we will write the approximation for the first derivative in the form

$$\left(P\frac{\partial v}{\partial x}\right)_j=\frac{1}{h}\left(Qv\right)_j,\quad j=0,\ldots,N \tag{A.0.6}$$

where P and Q are $(N+1)\times(N+1)$ matrices with boundary closures of arbitrary size n such that (we denote $l=N-n$):

$$Q=\begin{pmatrix}
q_{00} & \cdots & q_{0n-1} & q_{0n} & & & & & & & \\
\vdots & & \vdots & \vdots & & & & & 0 & & \\
q_{n-10} & \cdots & q_{n-1n-1} & q_{n-1n} & \frac{1}{36} & & & & & & \\
q_{n0} & \cdots & q_{n-1n} & q_{nn} & \frac{7}{9} & \frac{1}{36} & & & & & \\
0 & \cdots & -\frac{1}{36} & -\frac{7}{9} & 0 & \frac{7}{9} & \frac{1}{36} & & & & \\
 & & & & & \ddots & & & & & \\
 & & & & -\frac{1}{36} & -\frac{7}{9} & 0 & \frac{7}{9} & \frac{1}{36} & \cdots & 0 \\
 & & & & & -\frac{1}{36} & -\frac{7}{9} & q_{ll} & q_{ll+1} & \cdots & q_{lN} \\
 & & & & & & -\frac{1}{36} & q_{l+1l} & q_{l+1l+1} & \cdots & q_{l+1N} \\
 & & 0 & & & & & \vdots & \vdots & & \vdots \\
 & & & & & & & q_{Nl} & q_{Nl+1} & \cdots & q_{NN}
\end{pmatrix}$$

$$P = \begin{pmatrix} p_{00} & \dots & p_{0n} & & & & & \\ \vdots & & \vdots & & & 0 & & \\ p_{n0} & \dots & p_{nn} & \frac{1}{3} & & & & \\ 0 & \dots & \frac{1}{3} & 1 & \frac{1}{3} & & & \\ & & & & \ddots & & & \\ & & & & \frac{1}{3} & 1 & \frac{1}{3} & \\ & & & & & \frac{1}{3} & p_{ll} & \dots & p_{lN} \\ & & 0 & & & & \vdots & & \vdots \\ & & & & & & p_{Nl} & \dots & p_{NN} \end{pmatrix}$$

Denote by $\widehat{P}_1, \widehat{P}_2, \widehat{Q}_1, \widehat{Q}_2$ the corners of the matrices P and Q respectively, i.e.

$$\widehat{P}_1 = \begin{pmatrix} p_{00} & \dots & p_{0n} \\ \vdots & & \vdots \\ p_{n0} & \dots & p_{nn} \end{pmatrix}, \qquad \widehat{P}_2 = \begin{pmatrix} p_{ll} & \dots & p_{lN} \\ \vdots & & \vdots \\ p_{Nl} & \dots & p_{NN} \end{pmatrix}$$

$$\widehat{Q}_1 = \begin{pmatrix} q_{00} & \dots & q_{0n} \\ \vdots & & \vdots \\ q_{n0} & \dots & q_{nn} \end{pmatrix}, \qquad \widehat{Q}_2 = \begin{pmatrix} q_{ll} & \dots & q_{lN} \\ \vdots & & \vdots \\ q_{Nl} & \dots & q_{NN} \end{pmatrix}$$

Recall that we are looking for a matrix P which is symmetric positive definite and for a matrix Q which satisfies the Assumption **3** in Chapter 1 (Part I). Note that the matrix P is automatically symmetric in interior area (without the corners) and the matrix Q is automatically skew-symmetric in interior area (without the corners). For P to be symmetric we must have, therefore

$$(\widehat{P}_1)^T = \widehat{P}_1, \quad (\widehat{P}_2)^T = \widehat{P}_2 \tag{A.0.7}$$

We will show separately that the matrix P is positive definite. For Q to satisfy the Assumption **3**, it must satisfy:

$$\frac{\widehat{Q}_1^T + \widehat{Q}_1}{2} = \begin{pmatrix} q_{00} & \frac{1}{2}(q_{01}+q_{10}) & 0 & \dots \\ \frac{1}{2}(q_{01}+q_{10}) & q_{11} & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{A.0.8}$$

$$\frac{\widehat{Q}_2^T + \widehat{Q}_2}{2} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots \\ \dots & 0 & 0 & 0 \\ \dots & 0 & q_{N-1N-1} & \frac{1}{2}(q_{NN-1}+q_{N-1N}) \\ \dots & 0 & \frac{1}{2}(q_{NN-1}+q_{N-1N}) & q_{NN} \end{pmatrix} \tag{A.0.9}$$

In addition to these properties the matrices P and Q must satisfy the order properties of the boundary. To retain the formal six-order accuracy of the interior scheme, boundary closure must be accomplished to at least fifth-order accuracy (see [9], [10]). The demand that the boundary closure be fifth-order accurate is equivalent to its ability to annihilate a fifth-order polynomial in x exactly. Our problem is linear, which means that it is sufficient to check the basis elements $v_j = j^r \quad (r = 0, \dots, 5)$. The derivatives of the basis functions are $rj^{r-1} \quad (r = 0, \dots, 5)$. Substituting the test functions and their derivatives into the matrices $\widehat{P}_1, \widehat{Q}_1$ respectively yields the expression

$$r\sum_{j=0}^{n} p_{kj} j^{r-1} + \frac{r}{3}\delta_{kn}(n+1)^{r-1} = \sum_{j=0}^{n} q_{kj} j^r + \frac{1}{36}\delta_{kn-1}(n+1)^r + \frac{7}{9}\delta_{kn}(n+1)^r + \frac{1}{36}\delta_{kn}(n+2)^r, \quad k = 0, \dots, n \tag{A.0.10}$$

$$\delta_{kn} = \begin{cases} 0 & \text{if } n \neq k \\ 1 & \text{if } n = k \end{cases}.$$

Substituting into the matrices $\widehat{P}_2, \widehat{Q}_2$ yields the expression (we recall that $l = N - n$)

$$r\sum_{j=l}^{N} p_{kj} j^{r-1} + \frac{r}{3}\delta_{kl}(l-1)^{r-1} =$$

$$\sum_{j=l}^{N} q_{kj} j^{r} - \frac{1}{36}\delta_{kl+1}(l-1)^{r} - \frac{7}{9}\delta_{kl}(l-1)^{r} - \frac{1}{36}\delta_{kl}(l-2)^{r}, \quad k = \overline{l, N} \tag{A.0.11}$$

The block size n of the boundary matrices $\widehat{P}_1, \widehat{P}_2$ and $\widehat{Q}_1, \widehat{Q}_2$ could been any value and so far it has not been specified. As it has been mentioned above, we demand that the boundary closure be at least fifth-order accurate, therefore $n \geq 5$. For $n = 5$, equation (A.0.10) can be written concisely in matrix notation as

$$\widehat{P}_1 \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 4 & 12 & 32 & 80 \\ 0 & 1 & 6 & 27 & 108 & 405 \\ 0 & 1 & 8 & 48 & 256 & 1280 \\ 0 & 1 & 10 & 75 & 500 & 3125 \end{pmatrix} = \widehat{Q}_1 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 & 32 \\ 1 & 3 & 9 & 27 & 81 & 243 \\ 1 & 4 & 16 & 64 & 256 & 1024 \\ 1 & 5 & 25 & 125 & 625 & 3125 \end{pmatrix}$$

$$+ \frac{1}{36} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 6 & 36 & 216 & 1296 & 7776 \\ 29 & 163 & 913 & 5095 & 28321 & 156775 \end{pmatrix} \tag{A.0.12}$$

and equation (A.0.11) can be written as

$$\widehat{P}_2 \begin{pmatrix} 0 & 1 & 2(N-5) & 3(N-5)^2 & 4(N-5)^3 & 5(N-5)^4 \\ 0 & 1 & 2(N-4) & 3(N-4)^2 & 4(N-4)^3 & 5(N-4)^4 \\ 0 & 1 & 2(N-3) & 3(N-3)^2 & 4(N-3)^3 & 5(N-3)^4 \\ 0 & 1 & 2(N-2) & 3(N-2)^2 & 4(N-2)^3 & 5(N-2)^4 \\ 0 & 1 & 2(N-1) & 3(N-1)^2 & 4(N-1)^3 & 5(N-1)^4 \\ 0 & 1 & 2N & 3N^2 & 4N^3 & 5N^4 \end{pmatrix}$$

$$= \widehat{Q}_2 \begin{pmatrix} 1 & N-5 & (N-5)^2 & (N-5)^3 & (N-5)^4 & (N-5)^5 \\ 1 & N-4 & (N-4)^2 & (N-4)^3 & (N-4)^4 & (N-4)^5 \\ 1 & N-3 & (N-3)^2 & (N-3)^3 & (N-3)^4 & (N-3)^5 \\ 1 & N-2 & (N-2)^2 & (N-2)^3 & (N-2)^4 & (N-2)^5 \\ 1 & N-1 & (N-1)^2 & (N-1)^3 & (N-1)^4 & (N-1)^5 \\ 1 & N & N^2 & N^3 & N^4 & N^5 \end{pmatrix} \tag{A.0.13}$$

$$-\frac{1}{36}\begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where

$$\begin{aligned} \alpha_0 &= 29, \\ \alpha_1 &= -163 + 29N, \\ \alpha_2 &= 913 - 326N + 29N^2, \\ \alpha_3 &= -5095 - 2739N - 489N^2 + 29N^3, \\ \alpha_4 &= (N-7)^4 + 48(N-6)^3 + 28(N-6)^4, \\ \alpha_5 &= (N-7)^5 + 60(N-6)^4 + 28(N-6)^6, \\ \beta_0 &= 1, \\ \beta_1 &= N-6, \\ \beta_2 &= (N-6)^2, \\ \beta_3 &= (N-6)^3, \\ \beta_4 &= (N-6)^4, \\ \beta_5 &= (N-6)^5. \end{aligned}$$

Using Mathematica to solve these equations for the matrices $\widehat{Q}_1, \widehat{Q}_2$ results in the expressions

$$\begin{aligned} \widehat{Q}_1 &= \widehat{P}_1 \begin{pmatrix} -137/60 & 5 & -5 & 10/3 & -5/4 & 1/5 \\ -1/5 & -13/12 & 2 & -1 & 1/3 & -1/20 \\ 1/20 & -1/2 & -1/3 & 1 & -1/4 & 1/30 \\ -1/30 & 1/4 & -1 & 1/3 & 1/2 & -1/20 \\ 1/20 & -1/3 & 1 & -2 & 13/12 & 1/5 \\ -1/5 & 5/4 & -10/3 & 5 & -5 & 137/60 \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1/36 & 1/6 & -5/12 & 5/9 & -5/12 & 1/6 \\ -11/60 & 41/36 & -3 & 157/36 & -139/36 & 47/20 \end{pmatrix}, \end{aligned} \tag{A.0.14}$$

$$\widehat{Q}_2 = \widehat{P}_2 \begin{pmatrix} -137/60 & 5 & -5 & 10/3 & -5/4 & 1/5 \\ -1/5 & -13/12 & 2 & -1 & 1/3 & -1/20 \\ 1/20 & -1/2 & -1/3 & 1 & -1/4 & 1/30 \\ -1/30 & 1/4 & -1 & 1/3 & 1/2 & -1/20 \\ 1/20 & -1/3 & 1 & -2 & 13/12 & 1/5 \\ -1/5 & 5/4 & -10/3 & 5 & -5 & 137/60 \end{pmatrix}$$

$$- \begin{pmatrix} 47/20 & -139/36 & 157/36 & -3 & 41/36 & -11/60 \\ 1/6 & -5/12 & 5/9 & -5/12 & 1/6 & -1/36 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{A.0.15}$$

which relate the matrix $\widehat{Q}_1$ to the matrix $\widehat{P}_1$ and the matrix $\widehat{Q}_2$ to the matrix $\widehat{P}_2$ through fifth-order accuracy constraints. Solving for $\widehat{Q}_1$ and $\widehat{P}_1$ such that such that $\widehat{P}_1^T = \widehat{P}_1$ and equation (A.0.8) is satisfied yields :

$$p_{00} = -\frac{63410865331}{4945536000} - \frac{491453}{21200}\, p_{25} - \frac{163111}{63600}\, p_{34} - \frac{163111}{10600}\, p_{35} + \frac{163111}{3975}\, p_{45},$$

$$p_{01} = -\frac{39314269}{2289600} - \frac{8064}{265}\, p_{25} - \frac{896}{265}\, p_{34} - \frac{5376}{265}\, p_{35} + \frac{14601}{265}\, p_{45},$$

$$p_{02} = \frac{68626721}{5151600} + \frac{6336}{265}\, p_{25} + \frac{704}{265}\, p_{34} + \frac{4489}{265}\, p_{35} - \frac{11264}{265}\, p_{45},$$

$$p_{03} = -\frac{548278393}{61819200} - \frac{3959}{265}\, p_{25} - \frac{1408}{795}\, p_{34} - \frac{2816}{265}\, p_{35} + \frac{22528}{795}\, p_{45},$$

$$p_{04} = \frac{243243557}{65940480} + \frac{4353}{848}\, p_{25} + \frac{625}{848}\, p_{34} + \frac{1451}{424}\, p_{35} - \frac{625}{53}\, p_{45},$$

$$p_{05} = -\frac{188258177}{412128000} - \frac{4353}{5300}\, p_{25} - \frac{837}{5300}\, p_{34} - \frac{2511}{2650}\, p_{35} + \frac{2023}{1325}\, p_{45},$$

$$p_{11} = -\frac{69122837699}{989107200} - \frac{553237}{4240}\, p_{25} - \frac{189359}{12720}\, p_{34} - \frac{189359}{2120}\, p_{35} + \frac{179819}{795}\, p_{45},$$

$$p_{12} = \frac{149692169}{4121280} + \frac{3456}{53}\, p_{25} + \frac{437}{53}\, p_{34} + \frac{2304}{53}\, p_{35} - \frac{6144}{53}\, p_{45},$$

$$p_{13} = -\frac{1479095389}{82425600} - \frac{40521}{1060}\, p_{25} - \frac{5209}{1060}\, p_{34} - \frac{15097}{530}\, p_{35} + \frac{15536}{265}\, p_{45},$$

$$
\begin{aligned}
p_{14} &= \frac{723727793}{247276800} + \frac{16159}{1060}\, p_{25} + \frac{5033}{3180}\, p_{34} + \frac{8213}{530}\, p_{35} - \frac{8207}{795}\, p_{45},\\
p_{15} &= -\frac{179173199}{329702400} - \frac{10491}{4240}\, p_{25} - \frac{459}{4240}\, p_{34} - \frac{3497}{2120}\, p_{35} + \frac{459}{265}\, p_{45},\\
p_{22} &= -\frac{229036553}{9158400} - \frac{29653}{1060}\, p_{25} - \frac{6357}{1060}\, p_{34} - \frac{3171}{530}\, p_{35} + \frac{22248}{265}\, p_{45},\\
p_{23} &= \frac{120185699}{8242560} + \frac{1191}{106}\, p_{25} + \frac{415}{106}\, p_{34} - \frac{133}{53}\, p_{35} - \frac{2472}{53}\, p_{45},\\
p_{24} &= \frac{17198251}{82425600} - \frac{5961}{1060}\, p_{25} - \frac{1369}{1060}\, p_{34} - \frac{3577}{530}\, p_{35} + \frac{176}{265}\, p_{45},\\
p_{25} &= p_{25},\\
p_{33} &= -\frac{2197435691}{247276800} + \frac{1067}{1060}\, p_{25} - \frac{8831}{3180}\, p_{34} + \frac{7069}{530}\, p_{35} + \frac{25784}{795}\, p_{45},\\
p_{34} &= p_{34},\\
p_{35} &= p_{35},\\
p_{44} &= \frac{2986442821}{989107200} - \frac{277}{4240}\, p_{25} - \frac{5039}{12720}\, p_{34} - \frac{5039}{2120}\, p_{35} - \frac{4501}{795}\, p_{45},\\
p_{45} &= p_{45},\\
p_{55} &= \frac{580995101}{549504000} + \frac{67}{21200}\, p_{25} + \frac{243}{21200}\, p_{34} + \frac{729}{10600}\, p_{35} - \frac{243}{1325}\, p_{45},\\
q_{00} &= \frac{388639}{11448} + \frac{3240}{53}\, p_{25} + \frac{360}{53}\, p_{34} + \frac{2160}{53}\, p_{35} - \frac{5760}{53}\, p_{45},\\
q_{01} &= -\frac{18524686181}{329702400} - \frac{429809}{4240}\, p_{25} - \frac{144683}{12720}\, p_{34} - \frac{435109}{6360}\, p_{35} + \frac{142828}{795}\, p_{45},\\
q_{02} &= \frac{3897839983}{98910720} + \frac{29633}{424}\, p_{25} + \frac{3481}{424}\, p_{34} + \frac{30481}{636}\, p_{35} - \frac{20038}{159}\, p_{45},\\
q_{03} &= -\frac{2049530987}{82425600} - \frac{44823}{1060}\, p_{25} - \frac{5687}{1060}\, p_{34} - \frac{15471}{530}\, p_{35} + \frac{21158}{265}\, p_{45},
\end{aligned}
$$

$$
\begin{aligned}
q_{04} &= \frac{32307251}{3663360} + \frac{6383}{424}\, p_{25} + \frac{2693}{1272}\, p_{34} + \frac{7231}{636}\, p_{35} - \frac{4538}{159}\, p_{45},\\
q_{05} &= -\frac{1111336813}{989107200} - \frac{10259}{4240}\, p_{25} - \frac{1611}{4240}\, p_{34} - \frac{15559}{6360}\, p_{35} + \frac{2978}{795}\, p_{45},\\
q_{10} &= \frac{18414785381}{329702400} + \frac{429809}{4240}\, p_{25} + \frac{144683}{12720}\, p_{34} + \frac{435109}{6360}\, p_{35} - \frac{142828}{795}\, p_{45},\\
q_{11} &= -\frac{394363}{11448} - \frac{3240}{53}\, p_{25} - \frac{360}{53}\, p_{34} - \frac{2160}{53}\, p_{35} + \frac{5760}{53}\, p_{45},\\
q_{12} &= -\frac{2679152503}{61819200} - \frac{18239}{265}\, p_{25} - \frac{2321}{265}\, p_{34} - \frac{33563}{795}\, p_{35} + \frac{111938}{795}\, p_{45},\\
q_{13} &= \frac{5673768169}{164851200} + \frac{81941}{2120}\, p_{25} + \frac{41447}{6360}\, p_{34} + \frac{52261}{3180}\, p_{35} - \frac{87134}{795}\, p_{45},\\
q_{14} &= -\frac{4614530453}{329702400} - \frac{50817}{4240}\, p_{25} - \frac{12713}{4240}\, p_{34} - \frac{9519}{2120}\, p_{35}\frac{11918}{265}\, p_{45},\\
q_{15} &= \frac{29920589}{19782144} + \frac{815}{424}\, p_{25} + \frac{279}{424}\, p_{34} + \frac{1663}{636}\, p_{35} - \frac{826}{159}\, p_{45},\\
q_{22} &= 0\\
q_{23} &= -\frac{3086574227}{247276800} + \frac{6899}{1060}\, p_{25} - \frac{1589}{1060}\, p_{34} + \frac{31279}{1590}\, p_{35} + \frac{32848}{795}\, p_{45},\\
q_{24} &= \frac{4561833307}{494553600} - \frac{14059}{2120}\, p_{25} + \frac{3149}{2120}\, p_{34} - \frac{43739}{3180}\, p_{35} - \frac{23134}{795}\, p_{45},\\
q_{25} &= -\frac{6161203}{9158400} + \frac{1257}{1060}\, p_{25} - \frac{567}{1060}\, p_{34} - \frac{111}{530}\, p_{35} + \frac{678}{265}\, p_{45},\\
q_{33} &= 0\\
q_{34} &= -\frac{61484501}{20606400} + \frac{961}{265}\, p_{25} - \frac{563}{795}\, p_{34} + \frac{4837}{795}\, p_{35} + \frac{9538}{795}\, p_{45},\\
q_{35} &= \frac{5319631}{98910720} - \frac{319}{424}\, p_{25} + \frac{153}{424}\, p_{34} + \frac{529}{636}\, p_{35} - \frac{70}{159}\, p_{45},
\end{aligned}
$$

$$
\begin{aligned}
q_{44} &= 0 \\
q_{45} &= \frac{1024301777}{989107200} + \frac{271}{4240}\, p_{25} - \frac{441}{4240}\, p_{34} - \frac{5029}{6360}\, p_{35} - \frac{532}{795}\, p_{45}, \\
q_{55} &= 0.
\end{aligned}
$$

Solving for $\widehat{Q}_2$ and $\widehat{P}_2$ such that $\widehat{P}_2^T = \widehat{P}_2$ and equation (A.0.9) is satisfied yields :

$$
\begin{aligned}
p_{N,N} &= -\frac{63410865331}{4945536000} - \frac{491453}{21200}\, p_{N-2,N-5} - \frac{163111}{63600}\, p_{N-3,N-4} - \frac{163111}{10600}\, p_{N-3,N-5} + \\
&\quad + \frac{163111}{3975}\, p_{N-4,N-5}, \\
p_{N,N-1} &= -\frac{39314269}{2289600} - \frac{8064}{265}\, p_{N-2,N-5} - \frac{896}{265}\, p_{N-3,N-4} - \frac{5376}{265}\, p_{N-3,N-5} \\
&\quad + \frac{14601}{265}\, p_{N-4,N-5}, \\
p_{N,N-2} &= \frac{68626721}{5151600} + \frac{6336}{265}\, p_{N-2,N-5} + \frac{704}{265}\, p_{N-3,N-4} + \frac{4489}{265}\, p_{N-3,N-5} - \frac{11264}{265}\, p_{N-4,N-5}, \\
p_{N,N-3} &= -\frac{548278393}{61819200} - \frac{3959}{265}\, p_{N-2,N-5} - \frac{1408}{795}\, p_{N-3,N-4} - \frac{2816}{265}\, p_{N-3,N-5} \\
&\quad + \frac{22528}{795}\, p_{N-4,N-5}, \\
p_{N,N-4} &= \frac{243243557}{65940480} + \frac{4353}{848}\, p_{N-2,N-5} + \frac{625}{848}\, p_{N-3,N-4} + \frac{1451}{424}\, p_{N-3,N-5} - \frac{625}{53}\, p_{N-4,N-5}, \\
p_{N,N-5} &= -\frac{188258177}{412128000} - \frac{4353}{5300}\, p_{N-2,N-5} - \frac{837}{5300}\, p_{N-3,N-4} - \frac{2511}{2650}\, p_{N-3,N-5} \\
&\quad + \frac{2023}{1325}\, p_{N-4,N-5}, \\
p_{N-1,N-1} &= -\frac{69122837699}{989107200} - \frac{553237}{4240}\, p_{N-2,N-5} - \frac{189359}{12720}\, p_{N-3,N-4} - \frac{189359}{2120}\, p_{N-3,N-5} \\
&\quad + \frac{179819}{795}\, p_{N-4,N-5},
\end{aligned}
$$

$$p_{N-1,N-2} = \frac{149692169}{4121280} + \frac{3456}{53}\, p_{N-2,N-5} + \frac{437}{53}\, p_{N-3,N-4} + \frac{2304}{53}\, p_{N-3,N-5} - \frac{6144}{53}\, p_{N-4,N-5},$$

$$p_{N-1,N-3} = -\frac{1479095389}{82425600} - \frac{40521}{1060}\, p_{N-2,N-5} - \frac{5209}{1060}\, p_{N-3,N-4} - \frac{15097}{530}\, p_{N-3,N-5} +$$

$$+ \frac{15536}{265}\, p_{N-4,N-5},$$

$$p_{N-1,N-4} = \frac{723727793}{247276800} + \frac{16159}{1060}\, p_{N-2,N-5} + \frac{5033}{3180}\, p_{N-3,N-4} + \frac{8213}{530}\, p_{N-3,N-5}$$

$$-\frac{8207}{795}\, p_{N-4,N-5},$$

$$p_{N-1,N-5} = -\frac{179173199}{329702400} - \frac{10491}{4240}\, p_{N-2,N-5} - \frac{459}{4240}\, p_{N-3,N-4} - \frac{3497}{2120}\, p_{N-3,N-5}$$

$$+\frac{459}{265}\, p_{N-4,N-5},$$

$$p_{N-2,N-2} = -\frac{229036553}{9158400} - \frac{29653}{1060}\, p_{N-2,N-5} - \frac{6357}{1060}\, p_{N-3,N-4} - \frac{3171}{530}\, p_{N-3,N-5}$$

$$+ \frac{22248}{265}\, p_{N-4,N-5},$$

$$p_{N-2,N-3} = \frac{120185699}{8242560} + \frac{1191}{106}\, p_{N-2,N-5} + \frac{415}{106}\, p_{N-3,N-4} - \frac{133}{53}\, p_{N-3,N-5} - \frac{2472}{53}\, p_{N-4,N-5},$$

$$p_{N-2,N-4} = \frac{17198251}{82425600} - \frac{5961}{1060}\, p_{N-2,N-5} - \frac{1369}{1060}\, p_{N-3,N-4} - \frac{3577}{530}\, p_{N-3,N-5} + \frac{176}{265}\, p_{N-4,N-5},$$

$$p_{N-2,N-5} = p_{N-2,N-5},$$

$$p_{N-3,N-3} = -\frac{2197435691}{247276800} + \frac{1067}{1060}\, p_{N-2,N-5} - \frac{8831}{3180}\, p_{N-3,N-4} + \frac{7069}{530}\, p_{N-3,N-5}$$

$$+ \frac{25784}{795}\, p_{N-4,N-5},$$

$$p_{N-3,N-4} = p_{N-3,N-4},$$

$$
\begin{aligned}
p_{N-3,N-5} &= p_{N-3,N-5}, \\
p_{N-4,N-4} &= \frac{2986442821}{989107200} - \frac{277}{4240}\, p_{N-2,N-5} - \frac{5039}{12720}\, p_{N-3,N-4} - \frac{5039}{2120}\, p_{N-3,N-5} \\
&\quad - \frac{4501}{795}\, p_{N-4,N-5}, \\
p_{N-4,N-5} &= p_{N-4,N-5}, \\
p_{N-5,N-5} &= \frac{580995101}{549504000} + \frac{67}{21200}\, p_{N-2,N-5} + \frac{243}{21200}\, p_{N-3,N-4} + \frac{729}{10600}\, p_{N-3,N-5} \\
&\quad - \frac{243}{1325}\, p_{N-4,N-5}, \\
q_{N,N} &= \frac{388639}{11448} + \frac{3240}{53}\, p_{N-2,N-5} + \frac{360}{53}\, p_{N-3,N-4} + \frac{2160}{53}\, p_{N-3,N-5} - \frac{5760}{53}\, p_{N-4,N-5}, \\
q_{N,N-1} &= -\frac{18524686181}{329702400} - \frac{429809}{4240}\, p_{N-2,N-5} - \frac{144683}{12720}\, p_{N-3,N-4} - \frac{435109}{6360}\, p_{N-3,N-5} \\
&\quad + \frac{142828}{795}\, p_{N-4,N-5}, \\
q_{N,N-2} &= \frac{3897839983}{98910720} + \frac{29633}{424}\, p_{N-2,N-5} + \frac{3481}{424}\, p_{N-3,N-4} + \frac{30481}{636}\, p_{N-3,N-5} \\
&\quad - \frac{20038}{159}\, p_{N-4,N-5}, \\
q_{N,N-3} &= -\frac{2049530987}{82425600} - \frac{44823}{1060}\, p_{N-2,N-5} - \frac{5687}{1060}\, p_{N-3,N-4} - \frac{15471}{530}\, p_{N-3,N-5} \\
&\quad + \frac{21158}{265}\, p_{N-4,N-5}, \\
q_{N,N-4} &= \frac{32307251}{3663360} + \frac{6383}{424}\, p_{N-2,N-5} + \frac{2693}{1272}\, p_{N-3,N-4} + \frac{7231}{636}\, p_{N-3,N-5} \\
&\quad - \frac{4538}{159}\, p_{N-4,N-5}, \\
q_{N,N-5} &= -\frac{1111336813}{989107200} - \frac{10259}{4240}\, p_{N-2,N-5} - \frac{1611}{4240}\, p_{N-3,N-4} - \frac{15559}{6360}\, p_{N-3,N-5} +
\end{aligned}
$$

$$+\frac{2978}{795}\,p_{N-4,N-5},$$

$$q_{N-1,N} = \frac{18414785381}{329702400}+\frac{429809}{4240}\,p_{N-2,N-5}+\frac{144683}{12720}\,p_{N-3,N-4}+\frac{435109}{6360}\,p_{N-3,N-5}-$$

$$-\frac{142828}{795}\,p_{N-4,N-5},$$

$$q_{N-1,N-1} = -\frac{394363}{11448}-\frac{3240}{53}\,p_{N-2,N-5}-\frac{360}{53}\,p_{N-3,N-4}-\frac{2160}{53}\,p_{N-3,N-5}+\frac{5760}{53}\,p_{N-4,N-5},$$

$$q_{N-1,N-2} = -\frac{2679152503}{61819200}-\frac{18239}{265}\,p_{N-2,N-5}-\frac{2321}{265}\,p_{N-3,N-4}-\frac{33563}{795}\,p_{N-3,N-5}$$

$$+\frac{111938}{795}\,p_{N-4,N-5},$$

$$q_{N-1,N-3} = \frac{5673768169}{164851200}+\frac{81941}{2120}\,p_{N-2,N-5}+\frac{41447}{6360}\,p_{N-3,N-4}+\frac{52261}{3180}\,p_{N-3,N-5}$$

$$-\frac{87134}{795}\,p_{N-4,N-5},$$

$$q_{N-1,N-4} = -\frac{4614530453}{329702400}-\frac{50817}{4240}\,p_{N-2,N-5}-\frac{12713}{4240}\,p_{N-3,N-4}-\frac{9519}{2120}\,p_{N-3,N-5}$$

$$+\frac{11918}{265}\,p_{N-4,N-5},$$

$$q_{N-1,N-5} = \frac{29920589}{19782144}+\frac{815}{424}\,p_{N-2,N-5}+\frac{279}{424}\,p_{N-3,N-4}+\frac{1663}{636}\,p_{N-3,N-5}-\frac{826}{159}\,p_{N-4,N-5},$$

$$q_{N-2,N-2} = 0$$

$$q_{N-2,N-3} = -\frac{3086574227}{247276800}+\frac{6899}{1060}\,p_{N-2,N-5}-\frac{1589}{1060}\,p_{N-3,N-4}+\frac{31279}{1590}\,p_{N-3,N-5}$$

$$+\frac{32848}{795}\,p_{N-4,N-5},$$

$$q_{N-2,N-4} = \frac{4561833307}{494553600}-\frac{14059}{2120}\,p_{N-2,N-5}+\frac{3149}{2120}\,p_{N-3,N-4}-\frac{43739}{3180}\,p_{N-3,N-5}$$

$$- \frac{23134}{795}\, p_{N-4,N-5},$$

$$q_{N-2,N-5} = -\frac{6161203}{9158400} + \frac{1257}{1060}\, p_{N-2,N-5} - \frac{567}{1060}\, p_{N-3,N-4} - \frac{111}{530}\, p_{N-3,N-5} + \frac{678}{265}\, p_{N-4,N-5},$$

$$q_{N-3,N-3} = 0$$

$$q_{N-3,N-4} = -\frac{61484501}{20606400} + \frac{961}{265}\, p_{N-2,N-5} - \frac{563}{795}\, p_{N-3,N-4} + \frac{4837}{795}\, p_{N-3,N-5} + \frac{9538}{795}\, p_{N-4,N-5},$$

$$q_{N-3,N-5} = \frac{5319631}{98910720} - \frac{319}{424}\, p_{N-2,N-5} + \frac{153}{424}\, p_{N-3,N-4} + \frac{529}{636}\, p_{N-3,N-5} - \frac{70}{159}\, p_{N-4,N-5},$$

$$q_{N-4,N-4} = 0$$

$$q_{N-4,N-5} = \frac{1024301777}{989107200} + \frac{271}{4240}\, p_{N-2,N-5} - \frac{441}{4240}\, p_{N-3,N-4} - \frac{5029}{6360}\, p_{N-3,N-5} - \frac{532}{795}\, p_{N-4,N-5},$$

$$q_{N-5,N-5} = 0.$$

We have now to choose free parameters p_{25}, p_{34}, p_{35}, p_{45}, $p_{N-2,N-5}$, $p_{N-3,N-4}$, $p_{N-3,N-5}$, $p_{N-4,N-5}$ such that the matrix P will be positive definite and the expression $q_{NN}u_N^2 + (q_{N-1N} + q_{NN-1})u_N u_{N-1} + q_{N-1N-1}u_{N-1}^2$ will be positive for all $u_N, u_{N-1} \in \mathbf{R}$. The last requirement is needed in order to insure the positive definiteness of the low right hand corner of $\frac{Q+Q^T}{2}$. There are many possibilities to choose parameters which satisfy all the above stated criteria. Choosing the specific values of those parameters:

$$p_{25} = -\frac{7471}{699840}, \quad p_{34} = \frac{1}{3}, \quad p_{35} = 0, \quad p_{45} = \frac{1}{3},$$

$$p_{N-2,N-5} = -\frac{731}{139968}, \quad p_{N-3,N-4} = \frac{1}{3}, \quad p_{N-3,N-5} = 0, \quad p_{N-4,N-5} = \frac{1}{3}$$

we get

$$\widehat{P}_1 = \begin{pmatrix} \frac{6965509}{27993600} & \frac{30563}{77760} & \frac{-281}{1296} & \frac{51007}{349920} & \frac{-19075}{373248} & \frac{19273}{2332800} \\ \frac{30563}{77760} & \frac{10875041}{5598720} & \frac{-4159}{15552} & \frac{171491}{466560} & \frac{-209207}{1399680} & \frac{45253}{1866240} \\ \frac{-281}{1296} & \frac{-4159}{15552} & \frac{1786169}{1399680} & \frac{10219}{46656} & \frac{27791}{466560} & \frac{-7471}{699840} \\ \frac{51007}{349920} & \frac{171491}{466560} & \frac{10219}{46656} & \frac{1382789}{1399680} & \frac{1}{3} & 0 \\ \frac{-19075}{373248} & \frac{-209207}{1399680} & \frac{27791}{466560} & \frac{1}{3} & \frac{5603021}{5598720} & \frac{1}{3} \\ \frac{19273}{2332800} & \frac{45253}{1866240} & \frac{-7471}{699840} & 0 & \frac{1}{3} & \frac{27992569}{27993600} \end{pmatrix}, \tag{A.0.16}$$

$$\widehat{P}_2 = \begin{pmatrix} \frac{139965257}{139968000} & \frac{1}{3} & 0 & \frac{-731}{139968} & \frac{100373}{9331200} & \frac{44129}{11664000} \\ \frac{1}{3} & \frac{28005133}{27993600} & \frac{1}{3} & \frac{67423}{1332800} & \frac{-464311}{6998400} & \frac{-43139}{1866240} \\ 0 & \frac{1}{3} & \frac{6952357}{6998400} & \frac{65387}{233280} & \frac{371203}{2332800} & \frac{112511}{1749600} \\ \frac{-731}{139968} & \frac{67423}{1332800} & \frac{65387}{233280} & \frac{7863337}{6998400} & \frac{6853}{77760} & \frac{-2801}{32400} \\ \frac{100373}{9331200} & \frac{-464311}{6998400} & \frac{371203}{2332800} & \frac{6853}{77760} & \frac{34458673}{27993600} & \frac{88303}{388800} \\ \frac{44129}{11664000} & \frac{-43139}{1866240} & \frac{112511}{1749600} & \frac{-2801}{32400} & \frac{88303}{388800} & \frac{17135237}{139968000} \end{pmatrix}, \tag{A.0.17}$$

$$\widehat{Q}_1 = \begin{pmatrix} \frac{-2}{3} & \frac{554577}{5598720} & \frac{-1708109}{2799360} & \frac{38413}{93312} & \frac{-418787}{2799360} & \frac{27143}{1119744} \\ \frac{-7412017}{5598720} & \frac{1}{6} & \frac{246851}{174960} & \frac{-999373}{2799360} & \frac{230747}{1866240} & \frac{-11383}{559872} \\ \frac{1708109}{2799360} & \frac{-246851}{174960} & 0 & \frac{1009613}{1399680} & \frac{252707}{2799360} & \frac{-5071}{466560} \\ \frac{-38413}{93312} & \frac{999373}{2799360} & \frac{-1009613}{1399680} & 0 & \frac{129581}{174960} & \frac{98947}{2799360} \\ \frac{418787}{2799360} & \frac{-230747}{1866240} & \frac{-252707}{2799360} & \frac{-129581}{174960} & 0 & \frac{4351153}{5598720} \\ \frac{-27143}{1119744} & \frac{11383}{559872} & \frac{5071}{466560} & \frac{-98947}{2799360} & \frac{-4351153}{5598720} & 0 \end{pmatrix}, \tag{A.0.18}$$

$$\widehat{Q}_2 = \begin{pmatrix} 0 & \frac{21765521}{27993600} & \frac{87463}{2799360} & \frac{-10271}{2332800} & \frac{-5515}{559872} & \frac{309251}{27993600} \\ \frac{-21765521}{27993600} & 0 & \frac{665203}{874800} & \frac{757411}{13996800} & \frac{543931}{9331200} & \frac{-188999}{2799360} \\ \frac{-87463}{2799360} & \frac{-665203}{874800} & 0 & \frac{5296429}{6998400} & \frac{-2046989}{13996800} & \frac{422449}{2332800} \\ \frac{10271}{2332800} & \frac{-757411}{13996800} & \frac{-5296429}{6998400} & 0 & \frac{905953}{874800} & \frac{-641321}{2799360} \\ \frac{5515}{559872} & \frac{-543931}{9331200} & \frac{2046989}{13996800} & \frac{-905953}{874800} & \frac{1}{6} & \frac{21586961}{27993600} \\ \frac{-309251}{27993600} & \frac{188999}{2799360} & \frac{-422449}{2332800} & \frac{641321}{2799360} & \frac{-12255761}{27993600} & \frac{1}{3} \end{pmatrix}. \tag{A.0.19}$$

This results in

$$\frac{\widehat{Q}_1 + \widehat{Q}_1^T}{2} = \begin{pmatrix} -\frac{2}{3} & -\frac{1}{6} & 0 & 0 & 0 & 0 \\ -\frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \frac{\widehat{Q}_2 + \widehat{Q}_2^T}{2} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{3} \end{pmatrix}. \tag{A.0.20}$$

It means that matrix Q is such that

$$\frac{Q + Q^T}{2} = \begin{pmatrix} -\frac{2}{3} & -\frac{1}{6} & 0 & \dots & 0 & 0 & 0 \\ -\frac{1}{6} & \frac{1}{6} & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{6} & \frac{1}{3} \end{pmatrix} \tag{A.0.21}$$

The next task is to show that P is positive definite. We write the symmetric matrix P as a sum of three symmetric matrices:

$$P = P^L + P^M + P^R \tag{A.0.22}$$

The matrix P^M is given by

$$P^M = \begin{pmatrix}
\frac{1}{50} & 0 & & & & & & & & & & & & & \\
0 & \frac{1}{50} & 0 & & & & & & & & & & & & \\
 & 0 & \frac{1}{50} & 0 & & & & & & & & & & & \\
 & & 0 & \frac{1}{40} & 0 & & & & & & & & & & \\
 & & & 0 & \frac{1}{20} & 0 & & & & & & & & & \\
 & & & & 0 & \frac{1}{2} & \frac{1}{3} & & & & & & & & \\
 & & & & & \frac{1}{3} & 1 & \frac{1}{3} & & & & & & & \\
 & & & & & & \frac{1}{3} & 1 & \frac{1}{3} & & & & & & \\
 & & & & & & & \ddots & \ddots & \ddots & & & & & \\
 & & & & & & & & \frac{1}{3} & 1 & \frac{1}{3} & & & & \\
 & & & & & & & & & \frac{1}{3} & 1 & \frac{1}{3} & & & \\
 & & & & & & & & & & \frac{1}{3} & \frac{1}{2} & 0 & & \\
 & & & & & & & & & & & 0 & \frac{1}{20} & 0 & \\
 & & & & & & & & & & & & 0 & \frac{1}{40} & 0 \\
 & & & & & & & & & & & & & 0 & \frac{1}{50} \\
 & & & & & & & & & & & & & & 0
\end{pmatrix} \tag{A.0.23}$$

and the matrices P^L, P^R are given by

$$P^L = \begin{pmatrix}
\frac{6405637}{27993600} & \frac{30563}{77760} & \frac{-281}{1296} & \frac{51007}{349920} & \frac{-19075}{373248} & \frac{19273}{2332800} & 0 & & \\
\frac{30563}{77760} & \frac{53815333}{27993600} & \frac{-4159}{15552} & \frac{171491}{466560} & \frac{-209207}{1399680} & \frac{45253}{1866240} & 0 & & \\
\frac{-281}{1296} & \frac{-4159}{15552} & \frac{8790877}{6998400} & \frac{10219}{46656} & \frac{27791}{466560} & \frac{-7471}{699840} & 0 & & \\
\frac{51007}{349920} & \frac{171491}{466560} & \frac{10219}{46656} & \frac{1347797}{1399680} & \frac{1}{3} & 0 & 0 & & 0 \\
\frac{-19075}{373248} & \frac{-209207}{1399680} & \frac{27791}{466560} & \frac{1}{3} & \frac{1064617}{1119744} & \frac{1}{3} & 0 & & \\
\frac{19273}{2332800} & \frac{45253}{1866240} & \frac{-7471}{699840} & 0 & \frac{1}{3} & \frac{13995769}{27993600} & 0 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\
 & & & & & & & & \\
 & & & 0 & & & & & \\
 & & & & & & & & 0
\end{pmatrix}, \tag{A.0.24}$$

$$P^R = \begin{pmatrix} 0 & & & & & & & & & & & & \\ & & & & & & & & 0 & & & & \\ & & & & & & & & & & & & \\ & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & \\ & & & 0 & \frac{69981257}{139968000} & \frac{1}{3} & 0 & \frac{-731}{139968} & \frac{100373}{9331200} & \frac{44129}{11664000} & & & \\ & & & 0 & \frac{1}{3} & \frac{26605453}{27993600} & \frac{1}{3} & \frac{67423}{1332800} & \frac{-464311}{6998400} & \frac{-43139}{1866240} & & & \\ & & 0 & 0 & 0 & \frac{1}{3} & \frac{6777397}{6998400} & \frac{65387}{233280} & \frac{371203}{2332800} & \frac{112511}{1749600} & & & \\ & & & 0 & \frac{-731}{139968} & \frac{67423}{1332800} & \frac{65387}{233280} & \frac{7723369}{6998400} & \frac{6853}{77760} & \frac{-2801}{32400} & & & \\ & & & 0 & \frac{100373}{9331200} & \frac{-464311}{6998400} & \frac{371203}{2332800} & \frac{6853}{77760} & \frac{33898801}{27993600} & \frac{88303}{388800} & & & \\ & & & 0 & \frac{44129}{11664000} & \frac{-43139}{1866240} & \frac{112511}{1749600} & \frac{-2801}{32400} & \frac{88303}{388800} & \frac{14335877}{139968000} & & & \\ & & & & & & & & & & & & \\ & & & & & & & & & & & & \\ & & & & & & & & & & & & \end{pmatrix}. \tag{A.0.25}$$

We shall show that P^M is positive definite and P^L, P^R are non-negative definite. The matrices P^L and P^R are $N \times N$ matrices with zero entries except 6×6 upper-left (lower-right) symmetric blocks and it is sufficient to show that these blocks are positive definite. According to Silvester's criteria a symmetric matrix is positive definite if all the principle minors of this matrix are positive. Using this we denote by M_i the minor of order i of a matrix. For the matrix P^L we have:

$$\begin{aligned} M_1 &= \frac{6405637}{27993600} \\ M_2 &= \frac{22366252104972}{783641640960000} \\ M_3 &= \frac{1630773022037561105317}{5484237660094464000000} \\ M_4 &= \frac{2094606180203467710911012 9}{947676267664323379200000 00} \end{aligned} \tag{A.0.26}$$

$$M_5 = \frac{3561011738804887607629921}{213227160224472760320000000}$$

$$M_6 = \frac{1785383434364225698345266211 99}{3070471107232407748608000000000}$$

and for the matrix P^R the principle minors of the lower-right block are:

$$M_1 = \frac{69981257}{139968000}$$

$$M_2 = \frac{1426526576794421}{3918208204800000}$$

$$M_3 = \frac{8144797301100014982137}{27421188300472320000000}$$

(A.0.27)

$$M_4 = \frac{7147677520659895589786682 17}{2369190669160808448000000000}$$

$$M_5 = \frac{1488481488223238977730788075321}{4264543204489455206400000000000}$$

$$M_6 = \frac{2981729143600744031812016450 9893}{1919044442020254842880000000000000000}$$

Consequently, the matrices P^L and P^R are non-negative definite.

The matrix P^M is a symmetric diagonally dominant matrix with all entries on the main diagonal positive. Therefore P^M is a positive definite matrix (see [14], theorem 6.1.10). Using this information and also the fact that the matrices P^R and P^L are non-negative definite, we infer that the symmetric matrix

$$P = P^L + P^M + P^R$$

is positive definite.

Appendix B

Construction of the Fourth-Order Compact Scheme

Here we derive an implicit scheme for (1.1.1) which is formally fourth-order accurate. We begin with approximation of the first derivative $\frac{\partial v}{\partial x}$ at inner points in the form

$$\left(\frac{\partial v}{\partial x}\right)_j = \frac{1}{h}\left[\frac{a_1\mu\delta}{1+b_1\delta^2}\right] v_i \tag{B.0.1}$$

where

$$\delta v_j = v_{j+1/2} - v_{j-1/2}, \qquad \mu v_j = \frac{v_{j+1/2} + v_{j-1/2}}{2};$$

or in the equivalent form

$$b_1\frac{\partial v_{j+1}}{\partial x} + (1-2b_1)\frac{\partial v_j}{\partial x} + b_1\frac{\partial v_{j-1}}{\partial x} = \frac{a_1}{2h}[\, v_{j+1} - v_{j-1}\,] \tag{B.0.2}$$

The relations between the coefficients a_1, b_1 are derived by matching the Taylor series coefficients of various orders. The first unmatched coefficient determines the formal truncation error of the approximation (B.0.1). In order to generate a fourth-order scheme we require that

$$a_1 = 1, \qquad b_1 = \frac{1}{6} \tag{B.0.3}$$

It follows from this and (B.0.1) that we have a fourth-order compact scheme in interior points defined implicitly as

$$\frac{1}{4}\frac{\partial v_{j+1}}{\partial x} + \frac{\partial v_j}{\partial x} + \frac{1}{4}\frac{\partial v_{j-1}}{\partial x} = \frac{3}{4h}[\, v_{j-1} - v_{j+1}\,] \tag{B.0.4}$$

Using this results we will write the approximation for the first derivative in the form

$$\left(P\frac{\partial v}{\partial x}\right)_j = \frac{1}{h}(Qv)_j, \quad j = 0, \ldots, N \tag{B.0.5}$$

where P and Q are $(N+1)\times(N+1)$ matrices with boundary closures of arbitrary size n such that (we denote $l = N - n$):

$$Q = \begin{pmatrix} q_{00} & \dots & q_{0n} & & & & & \\ \vdots & & \vdots & & & 0 & & \\ q_{n0} & \dots & q_{nn} & \frac{3}{4} & & & & \\ 0 & \dots & -\frac{3}{4} & 0 & \frac{3}{4} & & & \\ & & & \ddots & & & & \\ & & & -\frac{3}{4} & 0 & \frac{3}{4} & & \\ & & & & -\frac{3}{4} & q_{ll} & \dots & q_{lN} \\ & 0 & & & & \vdots & & \vdots \\ & & & & & q_{Nl} & \dots & q_{NN} \end{pmatrix}$$

$$P = \begin{pmatrix} p_{00} & \dots & p_{0n} & & & & & \\ \vdots & & \vdots & & & 0 & & \\ p_{n0} & \dots & p_{nn} & \frac{1}{4} & & & & \\ 0 & \dots & \frac{1}{4} & 1 & \frac{1}{4} & & & \\ & & & & \ddots & & & \\ & & & \frac{1}{4} & 1 & \frac{1}{4} & & \\ & & & & \frac{1}{4} & p_{ll} & \dots & p_{lN} \\ & 0 & & & & \vdots & & \vdots \\ & & & & & p_{Nl} & \dots & p_{NN} \end{pmatrix}$$

Denote by $\widehat{P}_1, \widehat{P}_2, \widehat{Q}_1, \widehat{Q}_2$ the corners of the matrices P and Q respectively, i.e.

$$\widehat{P}_1 = \begin{pmatrix} p_{00} & \dots & p_{0n} \\ \vdots & & \vdots \\ p_{n0} & \dots & p_{nn} \end{pmatrix}, \qquad \widehat{P}_2 = \begin{pmatrix} p_{ll} & \dots & p_{lN} \\ \vdots & & \vdots \\ p_{Nl} & \dots & p_{NN} \end{pmatrix}$$

$$\widehat{Q}_1 = \begin{pmatrix} q_{00} & \dots & q_{0n} \\ \vdots & & \vdots \\ q_{n0} & \dots & q_{nn} \end{pmatrix}, \qquad \widehat{Q}_2 = \begin{pmatrix} q_{ll} & \dots & q_{lN} \\ \vdots & & \vdots \\ q_{Nl} & \dots & q_{NN} \end{pmatrix}$$

Recall that we are looking for a matrix P which is symmetric positive definite and for a matrix Q which satisfies the Assumption **3** in Chapter 1 (Part I). Note that the matrix P is automatically symmetric in interior area (without the corners) and the matrix Q is automatically skew-symmetric in interior area (without the corners). For P to be symmetric we must have, therefore

$$(\widehat{P}_1)^T = \widehat{P}_1, \quad (\widehat{P}_2)^T = \widehat{P}_2 \tag{B.0.6}$$

We will show separately that the matrix P is positive definite. For Q to satisfy the Assumption **3**, it must satisfy:

$$\frac{\widehat{Q}_1^T + \widehat{Q}_1}{2} = \begin{pmatrix} q_{00} & \frac{1}{2}(q_{01} + q_{10}) & 0 & \dots \\ \frac{1}{2}(q_{01} + q_{10}) & q_{11} & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{B.0.7}$$

$$\text{(B.0.8)} \qquad \frac{\widehat{Q}_2^T + \widehat{Q}_2}{2} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots \\ \dots & 0 & 0 & 0 \\ \dots & 0 & q_{N-1N-1} & \frac{1}{2}(q_{NN-1} + q_{N-1N}) \\ \dots & 0 & \frac{1}{2}(q_{NN-1} + q_{N-1N}) & q_{NN} \end{pmatrix}$$

In addition to these properties the matrices P and Q must satisfy the order properties of the boundary. To retain the formal fourth-order accuracy of the interior scheme, boundary closure must be accomplished to at least third-order accuracy (see [9], [10]). The demand that the boundary closure be third-order accurate is equivalent to its ability to annihilate a third-order polynomial in x exactly. Our problem is linear, which means that it is sufficient to check the basis elements $v_j = j^r \quad (r = 0, \dots, 3)$. The derivatives of the basis functions are $rj^{r-1} \quad (r = 0, \dots, 3)$. Substituting the test functions and their derivatives into the matrices $\widehat{P}_1, \widehat{Q}_1$ respectively yields the expression

$$\text{(B.0.9)} \qquad r\sum_{j=0}^{n} p_{kj} j^{r-1} + \frac{r}{4}\delta_{kn}(n+1)^{r-1} = \sum_{j=0}^{n} q_{kj} j^r + \frac{3}{4}\delta_{kn}(n+1)^r, \quad k = 0, \dots, n$$

$$\delta_{kn} = \begin{cases} 0 & \text{if } n \neq k \\ 1 & \text{if } n = k \end{cases}.$$

Substituting into the matrices $\widehat{P}_2, \widehat{Q}_2$ yields the expression (we recall that $l = N - n$)

$$\text{(B.0.10)} \qquad r\sum_{j=l}^{N} p_{kj} j^{r-1} + \frac{r}{4}\delta_{kl}(l-1)^{r-1} = \sum_{j=l}^{N} q_{kj} j^r - \frac{3}{4}\delta_{kl}(l-1)^r, \quad k = \overline{l, N}$$

The block size n of the boundary matrices $\widehat{P}_1, \widehat{P}_2$ and $\widehat{Q}_1, \widehat{Q}_2$ could been any value and so far it has not been specified. As it has been mentioned above, we demand that the boundary closure be at least fifth-order accurate, therefore $n \geq 3$. For $n = 3$, equation (B.0.9) can be written concisely in matrix notation as

$$\text{(B.0.11)} \qquad \widehat{P}_1 \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & 3 \\ 0 & 1 & 4 & 12 \\ 0 & 1 & 6 & 27 \end{pmatrix} = \widehat{Q}_1 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 3 & 11 & 40 & 144 \end{pmatrix}$$

and equation (B.0.10) can be written as

$$\widehat{P}_2 \begin{pmatrix} 0 & 1 & 2(N-3) & 3(N-3)^2 \\ 0 & 1 & 2(N-2) & 3(N-2)^2 \\ 0 & 1 & 2(N-1) & 3(N-1)^2 \\ 0 & 1 & 2N & 3N^2 \end{pmatrix} = \widehat{Q}_2 \begin{pmatrix} 1 & N-3 & (N-3)^2 & (N-3)^3 \\ 1 & N-2 & (N-2)^2 & (N-2)^3 \\ 1 & N-1 & (N-1)^2 & (N-1)^3 \\ 1 & N & N^2 & N^3 \end{pmatrix}$$

$$- \frac{1}{4} \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{B.0.12}$$

where

$$\begin{aligned} \alpha_0 &= 3, \\ \alpha_1 &= 3N - 11, \\ \alpha_2 &= 3N^2 - 22N + 40, \\ \alpha_3 &= 3(N-3)(N-4)^2. \end{aligned}$$

Using Mathematica to solve these equations for the matrices $\widehat{Q}_1, \widehat{Q}_2$ results in the expressions

$$\widehat{Q}_1 = \widehat{P}_1 \begin{pmatrix} -11/60 & 3 & -3/2 & 1/3 \\ -1/3 & -1/1 & 1 & -1/6 \\ 1/6 & -1 & 1/2 & 1/3 \\ -1/3 & 3/2 & -3 & 11/6 \end{pmatrix}$$

$$+ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 7/24 & -5/4 & 17/8 & -23/12 \end{pmatrix}, \tag{B.0.13}$$

$$\widehat{Q}_2 = \widehat{P}_2 \begin{pmatrix} -11/60 & 3 & -3/2 & 1/3 \\ -1/3 & -1/1 & 1 & -1/6 \\ 1/6 & -1 & 1/2 & 1/3 \\ -1/3 & 3/2 & -3 & 11/6 \end{pmatrix}$$

$$- \begin{pmatrix} 7/24 & -5/4 & 17/8 & -23/12 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{B.0.14}$$

which relate the matrix $\widehat{Q}_1$ to the matrix $\widehat{P}_1$ and the matrix $\widehat{Q}_2$ to the matrix $\widehat{P}_2$ through fifth-order accuracy constraints. Solving for $\widehat{Q}_1$ and $\widehat{P}_1$ such that such that $\widehat{P}_1^T = \widehat{P}_1$ and

equation (B.0.7) is satisfied yields :

$$
\begin{aligned}
p_{00} &= \frac{66527}{288} - 32q_{23} - 64p_{23} - 191p_{33}, \\
p_{01} &= \frac{2086}{9} - 32q_{23} - 63p_{23} - 192p_{33}, \\
p_{02} &= -\frac{25915}{288} + 12q_{23} + 24p_{23} + 75p_{33}, \\
p_{03} &= \frac{75}{4} - 2q_{23} - 5p_{23} - 16p_{33}, \\
p_{11} &= \frac{12845}{18} - 96q_{23} - 200p_{23} - 591p_{33}, \\
p_{12} &= -\frac{5183}{36} + 18q_{23} + 43p_{23} + 120p_{33}, \\
p_{13} &= 26 - 4q_{23} - 8p_{23} - 21p_{33}, \\
p_{22} &= \frac{5183}{288} - 8p_{23} - 15p_{33}, \\
p_{23} &= p_{23}, \\
p_{33} &= p_{33}, \\
q_{00} &= -522 + 71q_{23} + 144p_{23} + 432p_{33}, \\
q_{01} &= \frac{16685}{24} - 95q_{23} - 192p_{23} - 576p_{33}, \\
q_{02} &= -\frac{5183}{24} + 28q_{23} + 60p_{23} + 180p_{33}, \\
q_{03} &= \frac{171}{4} - 5q_{23} - 12p_{23} - 36p_{33}, \\
q_{10} &= -\frac{16691}{24} + 95q_{23} + 192p_{23} + 576p_{33},
\end{aligned}
$$

$$
\begin{aligned}
q_{11} &= \frac{1043}{2} - 72q_{23} - 144p_{23} - 432p_{33}, \\
q_{12} &= \frac{5183}{24} - 27q_{23} - 60p_{23} - 180p_{33}, \\
q_{13} &= -42 + 4q_{23} + 12p_{23} + 36p_{33}, \\
q_{22} &= 0, \\
q_{23} &= q_{23}, \\
q_{33} &= 0.
\end{aligned}
$$

Solving for $\widehat{Q}_2$ and $\widehat{P}_2$ such that $\widehat{P}_2^T = \widehat{P}_2$ and equation (B.0.8) is satisfied yields :

$$
\begin{aligned}
p_{N,N} &= \frac{66527}{288} - 32q_{N-2,N-3} - 64p_{N-2,N-3} - 191p_{N-3,N-3}, \\
p_{N,N-1} &= \frac{2086}{9} - 32q_{N-2,N-3} - 63p_{N-2,N-3} - 192p_{N-3,N-3}, \\
p_{N,N-2} &= -\frac{25915}{288} + 12q_{N-2,N-3} + 24p_{N-2,N-3} + 75p_{N-3,N-3}, \\
p_{N,N-3} &= \frac{75}{4} - 2q_{N-2,N-3} - 5p_{N-2,N-3} - 16p_{N-3,N-3}, \\
p_{N-1,N-1} &= \frac{12845}{18} - 96q_{N-2,N-3} - 200p_{N-2,N-3} - 591p_{N-3,N-3}, \\
p_{N-1,N-2} &= -\frac{5183}{36} + 18q_{N-2,N-3} + 43p_{N-2,N-3} + 120p_{N-3,N-3}, \\
p_{N-1,N-3} &= 26 - 4q_{N-2,N-3} - 8p_{N-2,N-3} - 21p_{N-3,N-3}, \\
p_{N-2,N-2} &= \frac{5183}{288} - 8p_{N-2,N-3} - 15p_{N-3,N-3},
\end{aligned}
$$

$$
\begin{aligned}
p_{N-2,N-3} &= p_{N-2,N-3}, \\
p_{N-3,N-3} &= p_{N-3,N-3}, \\
q_{N,N} &= -522 + 71q_{N-2,N-3} + 144p_{N-2,N-3} + 432p_{N-3,N-3}, \\
q_{N,N-1} &= \frac{16685}{24} - 95q_{N-2,N-3} - 192p_{N-2,N-3} - 576p_{N-3,N-3}, \\
q_{N,N-2} &= -\frac{5183}{24} + 28q_{N-2,N-3} + 60p_{N-2,N-3} + 180p_{N-3,N-3}, \\
q_{N,N-3} &= \frac{171}{4} - 5q_{N-2,N-3} - 12p_{N-2,N-3} - 36p_{N-3,N-3}, \\
q_{N-1,N} &= -\frac{16691}{24} + 95q_{N-2,N-3} + 192p_{N-2,N-3} + 576p_{N-3,N-3}, \\
q_{N-1,N-1} &= \frac{1043}{2} - 72q_{N-2,N-3} - 144p_{N-2,N-3} - 432p_{N-3,N-3}, \\
q_{N-1,N-2} &= \frac{5183}{24} - 27q_{N-2,N-3} - 60p_{N-2,N-3} - 180p_{N-3,N-3}, \\
q_{N-1,N-3} &= -42 + 4q_{N-2,N-3} + 12p_{N-2,N-3} + 36p_{N-3,N-3}, \\
q_{N-2,N-2} &= 0, \\
q_{N-2,N-3} &= q_{N-2,N-3}, \\
q_{N-3,N-3} &= 0.
\end{aligned}
$$

We have now to choose free parameters q_{23}, p_{23}, p_{33}, $q_{N-2,N-3}$, $p_{N-2,N-3}$, $p_{N-3,N-3}$, such that the matrix P will be positive definite and the expression $q_{NN}u_N^2+(q_{N-1N}+q_{NN-1})u_Nu_{N-1}+q_{N-1N-1}u_{N-1}^2$ will be positive for all $u_N, u_{N-1} \in \mathbf{R}$. The last requirement is needed in order to insure the positive definiteness of the low right hand corner of $\frac{Q+Q^T}{2}$. There are many possibilities to choose parameters which satisfy all the above stated criteria. Choosing the specific values of those parameters:

$$q_{23} = \frac{427}{576}, \quad p_{23} = \frac{1}{4}, \quad p_{33} = 1,$$

$$q_{N-2,N-3} = \frac{143}{192}, \quad p_{N-2,N-3} = \frac{1}{4}, \quad p_{N-3,N-3} = 1,$$

we get

$$\text{(B.0.15)} \qquad \widehat{P}_1 = \begin{pmatrix} \frac{79}{288} & \frac{11}{36} & \frac{-25}{288} & \frac{5}{288} \\ \frac{11}{36} & \frac{13}{9} & \frac{35}{288} & \frac{5}{144} \\ \frac{-25}{288} & \frac{35}{288} & \frac{287}{288} & \frac{1}{4} \\ \frac{5}{288} & \frac{5}{144} & \frac{1}{4} & 1 \end{pmatrix}, \qquad \widehat{P}_2 = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{48} & \frac{1}{96} \\ \frac{1}{4} & \frac{287}{288} & \frac{53}{288} & \frac{-13}{288} \\ \frac{1}{48} & \frac{53}{288} & \frac{10}{9} & \frac{7}{36} \\ \frac{1}{96} & \frac{-13}{288} & \frac{7}{36} & \frac{47}{288} \end{pmatrix},$$

$$\text{(B.0.16)} \qquad \widehat{Q}_1 = \begin{pmatrix} \frac{-5}{8} & \frac{451}{576} & \frac{-29}{144} & \frac{25}{576} \\ \frac{-595}{576} & \frac{1}{8} & \frac{181}{192} & \frac{-5}{144} \\ \frac{29}{144} & \frac{-181}{192} & 0 & \frac{427}{576} \\ \frac{-25}{576} & \frac{5}{144} & \frac{-427}{576} & 0 \end{pmatrix}, \qquad \widehat{Q}_2 = \begin{pmatrix} 0 & \frac{143}{192} & \frac{-1}{48} & \frac{5}{192} \\ \frac{-143}{192} & 0 & \frac{163}{192} & \frac{-5}{48} \\ \frac{1}{48} & \frac{-163}{192} & \frac{1}{8} & \frac{45}{64} \\ \frac{-5}{192} & \frac{5}{48} & \frac{-29}{64} & \frac{3}{8} \end{pmatrix}.$$

This results in

$$\text{(B.0.17)} \qquad \frac{\widehat{Q}_1 + \widehat{Q}_1^T}{2} = \begin{pmatrix} -\frac{5}{8} & -\frac{1}{8} & 0 & 0 \\ -\frac{1}{8} & \frac{1}{8} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \frac{\widehat{Q}_2 + \widehat{Q}_2^T}{2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & \frac{1}{8} & \frac{3}{8} \end{pmatrix}.$$

It means that matrix Q is such that

$$\frac{Q+Q^t}{2} = \begin{pmatrix} -\frac{5}{8} & -\frac{1}{8} & 0 & \dots & 0 & 0 & 0 \\ -\frac{1}{8} & \frac{1}{8} & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & 0 & \dots & 0 & \frac{1}{8} & \frac{3}{8} \end{pmatrix} \tag{B.0.18}$$

The next task is to show that P is positive definite. We write the symmetric matrix P as a sum of three symmetric matrices:

$$P = P^L + P^M + P^R \tag{B.0.19}$$

The matrix P^M is given by

$$P^M = \begin{pmatrix} \frac{1}{40} & 0 & & & & & & & & & \\ 0 & \frac{1}{30} & 0 & & & & & & & & \\ & 0 & \frac{1}{20} & 0 & & & & & & & \\ & & 0 & \frac{1}{2} & \frac{1}{4} & & & & & & \\ & & & \frac{1}{4} & 1 & \frac{1}{4} & & & & & \\ & & & & \frac{1}{4} & 1 & \frac{1}{4} & & & & \\ & & & & & \ddots & \ddots & \ddots & & & \\ & & & & & & \frac{1}{4} & 1 & \frac{1}{4} & & \\ & & & & & & & \frac{1}{4} & 1 & \frac{1}{4} & \\ & & 0 & & & & & & \frac{1}{4} & \frac{1}{2} & 0 \\ & & & & & & & & 0 & \frac{1}{20} & 0 \\ & & & & & & & & & 0 & \frac{1}{30} & 0 \\ & & & & & & & & & & 0 & \frac{1}{40} \end{pmatrix} \tag{B.0.20}$$

and the matrices P^L, P^R are given by

(B.0.21)
$$P^L = \begin{pmatrix} \frac{359}{1440} & \frac{11}{36} & \frac{-25}{288} & \frac{5}{288} & 0 & & \\ \frac{11}{36} & \frac{127}{90} & \frac{35}{288} & \frac{5}{144} & 0 & & \\ \frac{-25}{288} & \frac{35}{288} & \frac{1363}{1440} & \frac{1}{4} & 0 & & \\ \frac{5}{288} & \frac{5}{144} & \frac{1}{4} & \frac{1}{2} & 0 & & \\ 0 & 0 & 0 & 0 & 0 & & \\ & & & & & 0 & \\ & & & & & & 0 \end{pmatrix},$$

(B.0.22)
$$P^R = \begin{pmatrix} 0 & & & & & & \\ & & & & & & \\ & & 0 & 0 & 0 & 0 & 0 \\ & & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{48} & \frac{1}{96} \\ 0 & & 0 & \frac{1}{4} & \frac{1363}{1440} & \frac{53}{288} & \frac{-13}{288} \\ & & 0 & \frac{1}{48} & \frac{53}{288} & \frac{97}{90} & \frac{7}{36} \\ & & 0 & \frac{1}{96} & \frac{-13}{288} & \frac{7}{36} & \frac{199}{1440} \end{pmatrix}.$$

We shall show that P^M is positive definite and P^L, P^R are non-negative definite. The matrices P^L and P^R are $N \times N$ matrices with zero entries except 4×4 upper-left (lower-right) symmetric blocks and it is sufficient to show that these blocks are positive definite. According to Silvester's criteria a symmetric matrix is positive definite if all the principle minors of this matrix are positive. Using this we denote by M_i the minor of order i of a matrix. For the matrix P^L we have:

$$M_1 = \frac{359}{1440}$$

$$
\begin{aligned}
M_2 &= \frac{33493}{129600} \\
M_3 &= \frac{668420969}{2985984000} \\
M_4 &= \frac{27250240067}{286654464000}
\end{aligned}
\tag{B.0.23}
$$

and for the matrix P^R the principle minors of the lower-right block are:

$$
\begin{aligned}
M_1 &= \frac{199}{1440} \\
M_2 &= \frac{4801}{43200} \\
M_3 &= \frac{31546961}{331776000} \\
M_4 &= \frac{1287837683}{31850496000}
\end{aligned}
\tag{B.0.24}
$$

Consequently, the matrices P^L and P^R are non-negative definite.

The matrix P^M is a symmetric diagonally dominant matrix with all entries on the main diagonal positive. Therefore P^M is a positive definite matrix (see [14] theorem 6.1.10). Using this information and also the fact that the matrices P^R and P^L are non-negative definite, we infer that the symmetric matrix

$$P = P^L + P^M + P^R$$

is positive definite.

Concluding Remarks

In this work a methodology for constructing compact implicit high-order finite-difference schemes, for hyperbolic initial boundary value problems was presented. The SAT procedure for imposing the analytical boundary conditions proposed by Carpenter, Gottlieb and Abarbanel in [5] was generalized in such a way that (I) it essentially simplified the construction of the approximation of the desirable accuracy from the technical point of view, (II) allowed to apply this technique to the solution of two-dimensional problems. Temporal stability in one space dimension was achieved by constructing such approximations that all eigenvalues of the coefficient matrix of the corresponding system of ordinary differential equations had a negative real part. On the other hand, convergence of the scheme was proved directly by deriving an equation for the error and bounding the error norm. It was shown that L_2 norm of the solution error might at most grow linearly in time with the time coefficient being proportional to h^m where h is the mesh size and m the spatial order of accuracy. In order to solve two-dimensional scalar problems $\partial/\partial x + \partial/\partial y$ was approximated by the sum of two differentiation matrices $D_x + D_y$ where both D_x and D_y had eigenvalues with a negative real part. Since the sum matrix $D_x + D_y$ does not necessarily preserves this property, strict stability of the scheme was proved by showing that $Re\,(u, (D_x + D_y)\,u)_H < 0 \quad \forall u \in R^{N^2}$ in some norm H. Numerical studies on hyperbolic scalar IBVPs in one and two space dimensions have been performed using fourth- and sixth-order compact implicit difference schemes. Boundary conditions were imposed using SAT boundary procedure. Numerical results that support the theoretical analysis have been obtained. It have been shown that the actual numerical solution had a temporal error bounded by a constant rather by a linear growth. These theoretical and numerical results were presented in Chapters 1 and 2 (Part I).

In Chapter 3 (Part II) the methodology presented in Chapter 1 was used to solve one-dimensional hyperbolic systems. Analytical proof of time stability for hyperbolic systems was obtained for a restricted class of problems, namely when $\| L \| \cdot \| R \| \leq 1/5$ for the sixth-order accurate scheme and $\| L \| \cdot \| R \| \leq 1/3$ for the fourth-order scheme. However, it has been numerically verified, by both measuring the error for long time integrations and determining the eigenvalue spectrum of the semi-discrete system, that the method was

effective and provided time stability even when a theoretical foundation was missing, e.g. even for $\|L\| \cdot \|R\| = 1$.

The numerical experiments were concluded by solving in Chapter 4 (Part II) the two-dimensional Maxwell's equations in free space. The SAT method used for solving diagonalized systems in 1-D was adopted to solve the two-dimensional system, which could not be diagonalized simultaneously. Numerical results obtained by using both fourth- and sixth-order "SAT" schemes were compared with the results obtained by using the fourth-order Ty(2,4) scheme derived by E. Turkel and A. Yefet in [35], [36].

Construction of the sixth- and fourth-order compact implicit difference schemes used throughout this work was discussed in detail in Appendix A and B respectively.

Bibliography

[1] S. Abarbanel and A. Ditkowski. Asymptotically stable fourth-order accurate schemes for the diffusion equation on complex shapes. *J. Comput. Phys.* **133,** 279-288 (1997).

[2] J. C. Butcher. On the integration processes of A. Huta. *J. Austral. Math. Soc.* **3,** 203 (1963).

[3] J. C. Butcher. On Runge-Kutta processes of high order. *J. Austral. Math. Soc.* **4,** 179 (1964).

[4] M. H. Carpenter, D. Gottlieb and S. Abarbanel. The stability of numerical boundary treatments for compact high-order finite-difference schemes. *J. Comput. Phys.* **108(2),** 272-295 (1993).

[5] M. H. Carpenter, D. Gottlieb and S. Abarbanel. Time-stable boundary conditions for finite difference schemes solving hyperbolic systems: Methodology and applications to high-order compact schemes. *J. Comput. Phys.* **111(2),** 220-236 (1994).

[6] M. H. Carpenter, D. Gottlieb and S. Abarbanel. The theoretical accuracy of Runge-Kutta time discretization for initial boundary value problem: A careful study of the boundary error. *ICASE Report* **93-83,** (Dec.1993).

[7] J. Gary. On boundary conditions for hyperbolic difference schemes. *J. Comput. Phys.* **26,** 339-351 (1978).

[8] S. K. Godunov and V. S. Ryabenkii. Spectral Criteria for the stability of boundary value problems for non-self-adjoint difference equations. *Uspeki Mat.* **18 VIII,** 3-15 (1963).

[9] B. Gustafsson. The convergence rate for difference approximations to mixed initial boundary value problems. *Math. Comp.* **29(130),** 396-406 (1975).

[10] B. Gustafsson. The convergence rate for difference approximations to general mixed initial boundary value problems. *SIAM J. Numer. Anal.* **18(2),** 179-190 (1981).

[11] B. Gustafsson, H.-O. Kreiss and J. Oliger. *Time Dependent Problems and Difference Methods.* Wiley and Sons, Inc., 1995.

[12] B. Gustafsson, H.-O. Kreiss and A. Sundström. Stability theory of differens approximations for mixed initial boundary value problems. II. *Math. Comp.* **26,** 649-686 (1972).

[13] B. Gustafsson and P. Olsson. Fourth order difference methods for hyperbolic IBVPs. Technical report 94.04, Research Institute Of Advanced Computer Science, NASA Ames Research Center, 1994.

[14] R. A. Horn, C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1989.

[15] H.-0. Kreiss. Difference approximations for the initial boundary value problem for hyperbolic differential equations. *Numerical Solutions of Nonlinear Partial Differential Equations* edited by D. Greenspan. Wiley, New-York, 1966.

[16] H.-O. Kreiss. Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comp.* **22,** 703-714 (1968).

[17] H.-O. Kreiss and J. Oliger. Comparison of accurate methods for the integration of hyperbolic equations. *Tellus*, **3**,1972.

[18] H.-O. Kreiss and G. Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. In *Mathematical Aspects of Finite Elements in Partial Differential Equations.* Academic Press. Inc., 1974.

[19] H.-O. Kreiss and G. Scherer. On the existence of energy estimates for difference approximations for hyperbolic systems. Technical report, Dept. of Scientific Computing, Uppsala University, 1977.

[20] H.-O. Kreiss and L. Wu. On the stability definition of difference approximations for the initial boundary value problems. *Appl. Num. Math.* **12,** 213-227 (1993).

[21] L. Lapidus and J. H. Seinfeld. *Numerical Solution of Ordinary Differential Equations.* Academic Press, New-York and London, 1971.

[22] S. K. Lele. Compact finite difference schemes with spectral-like resolution. *J. Comput. Phys.* **103(1),** 16-42 (1992).

[23] D. Levi and A. Tadmor. From semi-discrete to fully discrete: stability of Runge-Kutta schemes by the energy method. *SIAM Rev.* **40(1),** 40-73 (1998).

[24] P. Olsson. High-order difference methods and data parallel implementation. PhD Thesis, Uppsala University, Department of Scientific Computing, 1992.

[25] P. Olsson. Summation by parts, projections, and stability I. *Math. Comp.* **64(211),** 1035-1065 (1995).

[26] P. Olsson. Summation by parts, projections, and stability II. *Math. Comp.* **64(212),** 1473-1493 (1995).

[27] S. Osher. Systems of difference equations with general homogeneous boundary conditions. *Tran. Amer. Math. Soc.* **137,** 177-201 (1969).

[28] G. Scherer. On energy estimates for difference approximations for hyperbolic partial differential equations. PhD Thesis, Uppsala University, Department of Scientific Computing, 1977.

[29] B. Shwartz and B. Wendroff. The relative efficiency of finite-difference methods. I:Hyperbolic problems and splines. *SIAM J. Numer. Anal.* **11(5),** 979-993 (1974).

[30] B. Sjögreen. High order centered difference methods for the compressible Navier-Stokes equations. *J. Comput. Phys.* **116,** (1995).

[31] B. Strand. Summation by parts for finite difference approximations for d/dx. *J. Comput. Phys.* **110,** 47-67 (1994).

[32] B. Strand. High-order difference method for hyperbolic IBVP. In *International Conference on Spectral and High Order Methods.* Houston Journal of Mathematics, 1995.

[33] B. Strand. Numerical studies of hyperbolic initial boundary value problems. Acta Univ. Ups., *Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology*, 1996.

[34] J. C. Strikwerda. Initial boundary value problem for the method of lines. *J. Comput. Phys.* **34,** 94-110 (1980).

[35] A. Yefet and E. Turkel. Fourth Order Compact Implicit Method for the Maxwell Equations with Discontinuous Coefficients. To appear in *Appl. Num. Math.*

[36] E. Turkel. High-Order Methods. *Advances in Computational Electrodynamics: The Finite-Difference Time-Domain Method*, A. Taflove editor, Artech House, Boston, 63-109, 1998.