On error bounds of finite difference approximations to partial differential equations - temporal behavior and rate of convergence.

Saul Abarbanel
Department of Applied Mathematics
Tel-Aviv University
Tel-Aviv, ISRAEL

Adi Ditkowski
Division of Applied Mathematics
Brown University
Providence, R.I, USA

Bertil Gustafsson

September 26, 2000

Abstract

This paper considers a family of spatially semi-discrete approximations, including boundary treatments, to hyperbolic and parabolic equations. We derive the dependence of the error-bounds on time as well as on mesh size.

1 Introduction

The question of the role of numerically imposed boundary conditions in the solution of parabolic and hyperbolic PDE's has been with us for many years. Many investigators have studied the effect of boundary conditions on the stability of the overall scheme, (i.e. the 'inner algorithm' + boundary conditions), see for example [8], [11], [12], [17], [6], [20], [7], [13], [14], [15], [16], [18],[19] . In this context stability implies convergence of the scheme, at a fixed time t, as the mesh is refined. The question of the temporal behavior of the error was usually not considered.

When constructing higher order schemes, 3^{rd}-order accuracy and above, it turns out that it is difficult to state boundary conditions such that the overall scheme remains stable. The question then arises what happens to the overall accuracy of the numerical solution if the order of accuracy of the inner stencil, (m), is is higher then the order of the boundary conditions, $(m-s)$. This problem has been tackled by Gustafsson, see [9], [10]. His main result, for both parabolic and hyperbolic PDE's, is that if the accuracy of the extra boundary conditions, required for 'numerical closure' of the problem, are one less than that of the inner scheme, then the overall accuracy is not affected. The physical boundary conditions, however, must be approximated to the same order as the inner scheme.

In the present paper we consider a form of differentiation matrices, both hyperbolic and parabolic, which represent a fairly wide family of boundary condition formulation plus central inner schemes. We investigate the dependence of the error on time as well as on mesh size. The main results are as follows:

- In the hyperbolic case, the overall convergence rate is of order $\min(m, m-s+1)$ for all s, in agreement with the results given by Gustafsson [9], [10]. For $s = 0, 1$, the temporal bound of the error behaves as $\sqrt{t}$ for $t \ll 1$, and bends over smoothly to a linear bound as t increases. For $s \geq 2$, the temporal behavior is $\sim \sqrt{t}$ for all t, $t \geq 0$.

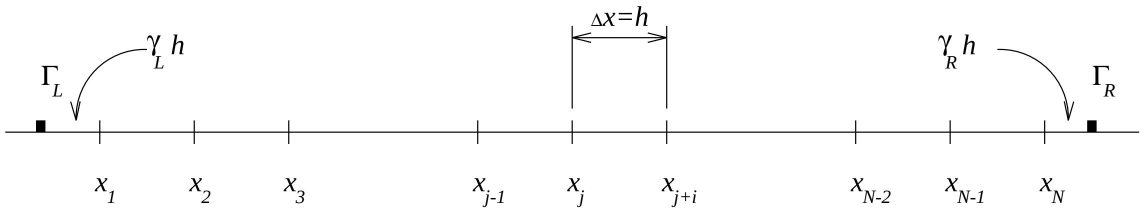

Figure 1: One dimensional grid.

- In the parabolic case, the overall convergence rate is of order m if $s = 0, 1$, and $m - s + 3/2$ if $s \geq 2$. The error is uniformly bounded independent of t for all $t,\ t \geq 0$.

The parabolic results are derived in section 2. The hyperbolic results are derived in sections 3, 4 and appendices A and B. Numerical experiments, given in section 5, demonstrate the validity of these bounds. In fact, in the parabolic case, the numerical convergence rate is $(m - s + 2)$ $s \geq 2$, exceeding the prediction which is only an *upper bound* on the error.

2 The diffusion equation

We consider the following problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t); \qquad \Gamma_L \leq x \leq \Gamma_R,\ \ t \geq 0\ , \tag{2.1a}$$

$$u(x,0) = u_0(x)\ , \tag{2.1b}$$

$$u(\Gamma_L, t) = g_L(t)\ , \tag{2.1c}$$

$$u(\Gamma_R, t) = g_R(t) \tag{2.1d}$$

and $f(x,t) \in\ C^1$.

Let us spatially discretize (2.1a) on the uniform grid presented in Figure 1:

Note that the boundary points do not necessarily coincide with x_1 and x_N. Set $x_{j+1} - x_j = h$, $1 \leq j \leq N-1$; $x_1 - \Gamma_L = \gamma_L h$, $0 \leq \gamma_L < 1$; $\Gamma_R - x_N = \gamma_R h$, $0 \leq \gamma_R < 1$.

The projection onto the above grid of the exact solution $u(x,t)$ to (2.1), is $u_j(t) = u(x_j,t) \overset{\triangle}{=} \mathbf{u}(t)$; $1 \leq j \leq N$. Let M be a matrix representing the second partial derivative at internal points without specifying yet how it is being constructed. Then we may write

$$\frac{d}{dt}\mathbf{u}(t) = [M\mathbf{u}(t) + \mathbf{B} + \mathbf{T}] + \mathbf{f}(t) \ , \tag{2.2}$$

where $\mathbf{T}$ is the truncation error due to the numerical differentiation and $\mathbf{f}(t) = f(x_j,t)$, $1 \leq j \leq N$. The boundary vector $\mathbf{B}$ has entries whose values depend on $g_L, g_R, \gamma_L, \gamma_R$ in such a way that $M\mathbf{u} + \mathbf{B}$ represents u_{xx} everywhere to the desired accuracy. The standard way of finding a numerical approximate solution to (2.1) is to omit $\mathbf{T}$ from (2.2) and solve

$$\frac{d}{dt}\mathbf{v}(t) = [M\mathbf{v}(t) + \mathbf{B}] + \mathbf{f}(t) \ , \tag{2.3}$$

where $\mathbf{v}(t)$ is the numerical approximation to the projection $\mathbf{u}(t)$. An equation for the solution error vector, $\boldsymbol{\epsilon}(t) = \mathbf{u}(t) - \mathbf{v}(t)$, can be found by subtracting (2.3) from (2.2):

$$\frac{d}{dt}\boldsymbol{\epsilon} = M\boldsymbol{\epsilon}(t) + \mathbf{T}(t)^{\dagger} \ , \tag{2.4}$$

with an homogeneous initial and boundary conditions.

We now form the scalar product of (2.4) with $H\boldsymbol{\epsilon}$ where H is a symmetric positive definite matrix of dimensions $N \times N$. We denote $(\cdot, H\cdot)$ by $(\cdot,\cdot)_H$. Equation (2.4) then becomes:

$$\begin{array}{rcl}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H & = & (\boldsymbol{\epsilon}, M\boldsymbol{\epsilon})_H + (\boldsymbol{\epsilon}, \mathbf{T})_H \\ & = & (\boldsymbol{\epsilon}, \frac{H\ M + M^T H}{2}\boldsymbol{\epsilon}) + (\boldsymbol{\epsilon}, H\mathbf{T}) \\ & = & (\boldsymbol{\epsilon}, M_s\boldsymbol{\epsilon}) + (\boldsymbol{\epsilon}, H\mathbf{T}) \ . \end{array} \tag{2.5}$$

If one uses a symmetric stencil, $-\frac{1}{h^2}(\alpha_k u_{j-k} + \ldots + \alpha_0 u_j + \ldots + \alpha_k u_{j+k})$ to represent $\frac{\partial^2 u_j}{\partial x^2}$, $k+1 \leq j \leq N-k$ and uses a non-symmetric stencil near the boundaries, $k+1 > j$, and $j > N-k$ then the structure of $M_s = \frac{1}{2}(H\ M + M^T H)$ is [‡]:

[†] If one uses penalty methods to represent the boundary values ("SAT", see [1], [2], [3]) then (2.2) and (2.3) are modified but (2.4) remains the same.

[‡] In the variable coefficient case M_s has a structure similar to (2.6). However in that case the μ_k's change along the diagonals.

$$M_s = -\frac{1}{h^2}\begin{pmatrix}
 & & & 0 & & & & & & & \\
 & M_{BL} & & \mu_k & \ddots & & & & & & \\
 & & & \vdots & \mu_k & & & & & & \\
0 & \mu_k & \dots & \mu_0 & & \ddots & & & & & \\
 & \ddots & \ddots & & \ddots & & \ddots & & & & \\
 & & & & & & & & & & \\
 & & & & \ddots & & \ddots & & \ddots & \ddots & \\
 & & & & & \ddots & & \mu_0 & \dots & \mu_k & 0 \\
 & & & & & & \mu_k & \vdots & & & \\
 & & & & & & \ddots & \mu_k & & M_{BR} & \\
 & & & & & & & 0 & & &
\end{pmatrix}, \qquad (2.6)$$

which is an $N \times N$ matrix with a $(2k+1)$-diagonal 'core'. The blocks M_{BL} and M_{BR} are each a matrix block of dimensions n_L and n_R respectively. n_L and n_R are determined by the stencil used near the boundaries and are independent of the matrix dimension N. At this point we would like to find a matrix, $\hat{M}$, such that the scalar product $(\boldsymbol{\epsilon}, M_s\boldsymbol{\epsilon})$ is majorized by $(\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon})$. The structure of $\hat{M}$ is:

$$\hat{M} = -\begin{pmatrix}
\lambda/h^2 & & & & & & & & & & \\
 & \ddots & & & & & & & & & \\
 & & \lambda/h^2 & & & & & & & & \\
 & & & c_0 & & & & & & & \\
 & & & & c_0 & & & & & & \\
 & & & & & \ddots & & & & & \\
 & & & & & & c_0 & & & & \\
 & & & & & & & c_0 & & & \\
 & & & & & & & & \lambda/h^2 & & \\
 & & & & & & & & & \ddots & \\
 & & & & & & & & & & \lambda/h^2
\end{pmatrix}, \qquad (2.7)$$

i.e. $\hat{M}$ is a diagonal matrix, with the first n_L and the last n_R entries being λ/h^2. $\lambda > 0$ is a constant and all the other terms, c_0, are positive. It can be shown that for many time-stable schemes (ones that do not admit exponential temporal growth of the error when they are GKS stable) such an $\hat{M}$ can be found. One way to show this is to decompose M_s in the following way:

$$M_s = \frac{1}{h^2}\left[\alpha\, M_1 + M_2 + M_3 + M_4\right]$$

where $\alpha > 0$ is a constant,

$$M_1 = \begin{bmatrix} -2 & 1 & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \\ & & & & & 1 & -2 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} M_L & 0 \\ 0 & \mathbf{0} \end{bmatrix}, \quad M_3 = \begin{bmatrix} \mathbf{0} & 0 \\ 0 & M_R \end{bmatrix},$$

where M_L and M_R are each a matrix-block of dimensions n_L and n_R respectively, having only negative eigenvalues and entries of order unity. M_4 is a non-positive definite matrix. When such a decomposition is possible then it is shown in [1],[2] and [3] that $c_0 = \alpha\pi^2$ and λ is the minimal eigenvalue of $-M_2$ and $-M_3$.

Using the form (2.7) of $\hat{M}$, one gets from (2.5):

$$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \le (\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon}) + (\boldsymbol{\epsilon}, H\mathbf{T}) \ . \tag{2.8}$$

It is reasonable to assume that $H\mathbf{T}$ has the following structure:

$$H\mathbf{T} = \left[T_{BL_1}, T_{BL_2}, \ldots, T_{BL_{n_L}}; T_{M_{N-(N-n_L-1)}}, \ldots, T_{M_{N-n_R}}; T_{BR_{N-n_R+1}}, \ldots, T_{BR_{N-1}}, T_{BR_N}\right]^T \ ,$$

where

$$\begin{array}{lcll} |T_{BL_j}| & < & \alpha_j h^{m-s} & 1 \le j \le n_L \ , \\ |T_{M_k}| & < & \alpha_k h^{m} & n_L+1 \le k \le N-n_R \ , \\ |T_{BR_l}| & < & \alpha_l h^{m-s} & N-n_R+1 \le l \le N \ , \end{array}$$

with the natural numbers s, m satisfying $m \ge 2,\ m \ge s \ge 0$ [§]; i.e. the basic finite differencing is at least second order accurate. Note that the entries in the positive definite matrix H are absorbed into the α's.

We now majorize $H\ \mathbf{T}$ entry by entry, by $\hat{\mathbf{T}}$,

$$\begin{aligned} \hat{\mathbf{T}} &= \left[\alpha_B h^{m-s}, \ldots, \alpha_B h^{m-s}; \alpha_M h^m, \ldots, \alpha_M h^m; \alpha_B h^{m-s}, \ldots, \alpha_B h^{m-s}\right]^T \\ &= [T_B, \ldots, T_B; T_M, \ldots, T_M; T_B, \ldots, T_B]^T \ , \end{aligned} \tag{2.9}$$

where $\alpha_B = \max_{1\le j\le n_L,\, N-n_R+1\le l\le N}\{\alpha_j, \alpha_l\}$ and $\alpha_M = \max_{n_L+1\le k\le N-n_R}\{\alpha_k\}$.

Using (2.9), $(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H$ is further majorized and (2.8) becomes:

$$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \le (\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon}) + (\boldsymbol{\epsilon}, H\mathbf{T}) \le (\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon}) + (|\boldsymbol{\epsilon}|, \hat{\mathbf{T}}) \ , \tag{2.10}$$

where

$$|\boldsymbol{\epsilon}| = [|\epsilon_1|, \ldots, |\epsilon_j|, \ldots, |\epsilon_N|]^T \ . \tag{2.11}$$

The component-wise version of (2.10) is:

$$\begin{aligned} (\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \ \le\ & \frac{1}{N}\left[\sum_{j=1}^{n_L} \epsilon_j \left(-\frac{\lambda}{h^2}\epsilon_j\right) + \sum_{k=n_L+1}^{N-n_R} \epsilon_k\left(-c_0\epsilon_k\right) + \sum_{l=N-n_R+1}^{N} \epsilon_l \left(-\frac{\lambda}{h^2}\epsilon_l\right)\right] + \\ & \frac{1}{N}\left[\sum_{j=1}^{n_L} |\epsilon_j| T_B + \sum_{k=n_L+1}^{N-n_R} |\epsilon_k| T_M + \sum_{l=N-n_R+1}^{N} |\epsilon_l| T_B\right] \ . \end{aligned} \tag{2.12}$$

[§] The case $s = 0$ is trivial. One can see from the definition of $H\mathbf{T}$ that the convergence rate is h^m.

Next we use the Schwartz inequality of the form

$$|fg| \leq af^2 + \frac{1}{4a}g^2 \ , \qquad a > 0 \ . \tag{2.13}$$

Then (2.12)becomes:

$$\begin{aligned}
(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \ \leq \ & h\sum_{j=1}^{n_L}\left[-\frac{\lambda}{h^2}\epsilon_j^2 + \left(a_B|\epsilon_j|^2 + \frac{1}{4a_B}T_B^2\right)\right] + \\
& h\sum_{k=n_L+1}^{N-n_R}\left[-c_0\epsilon_k^2 + \left(a_M|\epsilon_k|^2 + \frac{1}{4a_M}T_M^2\right)\right] + \\
& h\sum_{l=N-n_R+1}^{N}\left[-\frac{\lambda}{h^2}\epsilon_l^2 + \left(a_B|\epsilon_l|^2 + \frac{1}{4a_B}T_B^2\right)\right] \ .
\end{aligned} \tag{2.14}$$

Let us choose $a_M = \beta c_0 \ \ (0 < \beta < 1)$, and $a_B = \beta\frac{\lambda}{h^2}$, to get

$$\begin{aligned}
(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \ \leq \ & h\sum_{j=1}^{n_L} -(1-\beta)\frac{\lambda}{h^2}\epsilon_j^2 + h\sum_{k=n_L+1}^{N-n_R} -(1-\beta)c_0\epsilon_k^2 + h\sum_{l=N-n_R+1}^{N} -(1-\beta)\frac{\lambda}{h^2}\epsilon_l^2 + \\
& h\sum_{j=1}^{n_L}\frac{1}{4}\frac{h^2}{\lambda\beta}\alpha_B^2 h^{2m-2s} + h\sum_{k=n_L+1}^{N-n_R}\frac{1}{4\beta c_0}\alpha_M^2 h^{2m} + h\sum_{l=N-n_R+1}^{N}\frac{1}{4}\frac{h^2}{\lambda\beta}\alpha_B^2 h^{2m-2s} \ .
\end{aligned} \tag{2.15}$$

Next we note that because $\lambda = O(1)$, $\lambda/h^2 \gg c_0$. We use this fact to simplify (2.15)

$$\begin{aligned}
(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \ \leq \ & -c_0\,h\,(1-\beta)\sum_{j=1}^{N}\epsilon_j^2 + \frac{\alpha_B^2}{4\lambda\beta}\sum_{j=1}^{n_L+n_R} h^{2m+3-2s} + \frac{\alpha_M^2}{4\beta c_0}\sum_{j=1}^{N-n_L-n_R} h^{2m+1} \\
\leq \ & -c_0(1-\beta)(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) + \frac{\alpha_B^2}{4\lambda\beta}(n_L + n_R)h^{2m+3-2s} + \frac{\alpha_M^2}{4\beta c_0}\frac{N-(n_L+n_R)}{N}h^{2m} \\
\leq \ & -c_1(1-\beta)(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H + \frac{\alpha_B^2}{4\lambda\beta}(n_L + n_R)h^{2m+3-2s} + \frac{\alpha_M^2}{4\beta c_0}\left(1 - \frac{n_L+n_R}{N}\right)h^{2m} \ ,
\end{aligned} \tag{2.16}$$

where we introduced $c_1 > 0$, taking into account the equivalence of the L_2 and H norms.

We now distinguish between two cases:

i $s = 0,\, 1$

ii $s \geq 2$

In the first case ($s = 0,\, 1$) the middle term in (2.16) is negligible compared to the last term and we have:

$$\frac{1}{2}\frac{\partial}{\partial t}(\boldsymbol{\epsilon},\boldsymbol{\epsilon})_H = (\boldsymbol{\epsilon},\boldsymbol{\epsilon}_t)_H \leq -\frac{c_2}{2}(\boldsymbol{\epsilon},\boldsymbol{\epsilon})_H + \frac{\alpha_M^2}{4\beta c_0}h^{2m} + O(h^{2m+1})\ ; \quad (s = 0,\, 1)\ , \tag{2.17}$$

where $c_2 = 2c_1(1-\beta)$. From (2.17) it is easily shown that

$$(\boldsymbol{\epsilon},\boldsymbol{\epsilon})_H \leq \frac{\alpha_M^2\ (1 + O(h)\,)}{2\beta c_0 c_2}\left(1 - e^{-c_2 t}\right)h^{2m}\ . \tag{2.18}$$

Note that for all t, the estimate (2.18) has the same dependence on the mesh size, h, as predicted of by Gustafsson [10], namely h^m. In addition, (2.18) describes how the bound on the error evolves in time.

Next we consider the second case, $s \geq 2$. Now the middle term in (2.16) dominates the last one and we can write,

$$(\boldsymbol{\epsilon},\boldsymbol{\epsilon}_t)_H \leq -c_2(\boldsymbol{\epsilon},\boldsymbol{\epsilon})_H + \frac{\alpha_B^2}{2\lambda\beta}(n_L + n_R)h^{2m+3-2s} + O(h^{2m})\ . \tag{2.19}$$

This leads to

$$(\boldsymbol{\epsilon},\boldsymbol{\epsilon})_H \leq \frac{\alpha_B^2(n_L + n_R)}{2\,\lambda\,\beta\,c_2}\left(1 + O(h^{2s-3})\right)h^{2m+3-2s}\left(1 - e^{-c_2 t}\right). \tag{2.20}$$

The rate of convergence is $h^{m-s+3/2}$. The temporal behavior is analogous to the cases of $s = 0,\, 1$, see (2.18). For all s and fixed h, the error is uniformly bounded independent of t

In practice, however, a better rate of convergence of h^{m-s+2} is often obtained; see [10] and the numerical example in Section 5. This only reaffirms the status of the result (2.20) as an *upper* bound on the error.

3 The hyperbolic case, a classical approach.

We consider the following problem

$$\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x} + f(x,t); \qquad \Gamma_L \le x \le \Gamma_R, \;\; t \ge 0, \;\; a > 0 \;\;, \tag{3.1a}$$

$$u(x,0) = u_0(x) \;\;, \tag{3.1b}$$

$$u(\Gamma_R,t) = g_R(t) \tag{3.1c}$$

and $f(x,t) \in \; C^1$.

Let us discretize (3.1) spatially on the same grid as in section 2, and use the same notation for the numerical approximation and for the error vector. The equation for the solution error vector, $\boldsymbol{\epsilon}(t) = \mathbf{u}(t) - \mathbf{v}(t)$, is formally the same as (2.4). After taking the scalar product with $H\boldsymbol{\epsilon}$ we get an equation for $(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H$ which is is formally the same as (2.5). The only difference is that in this case, if one uses a centered-difference scheme, the structure of $M_s = \frac{1}{2}(H\ M + M^T H)$ will be:

$$-\frac{1}{h}\begin{pmatrix} M_{BL} & & \\ & 0 & \\ & & M_{BR} \end{pmatrix} \;\;. \tag{3.2}$$

The difference operator is an $N \times N$ matrix. The blocks M_{BL} and M_{BR} are each matrix blocks of dimensions n_L and n_R respectively, possessing only negative eigenvalues and having entries of order unity. As N increases ($h = 1/N$ decreases), n_L and n_R remain constants. We require that $(n_L + n_R)/N \ll 1$. With this in mind, the scalar product $(\boldsymbol{\epsilon}, M_s\boldsymbol{\epsilon})$ is majorized by $(\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon})$ where

$$\hat{M} = -\frac{1}{h}\begin{pmatrix} \lambda &&&&&&&&& \\ & \ddots &&&&&&&& \\ && \lambda &&&&&&& \\ &&& 0 &&&&&& \\ &&&& 0 &&&&& \\ &&&&& \ddots &&&& \\ &&&&&& 0 &&& \\ &&&&&&& 0 && \\ &&&&&&&& \lambda & \\ &&&&&&&&& \ddots & \\ &&&&&&&&&& \lambda \end{pmatrix} \tag{3.3}$$

and $\lambda > 0$ is the smallest eigenvalue of M_{BL} and M_{BR}. The dimensions of the corner blocks are the same as those of M_{BL} and M_{BR}. Therefore

$$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \leq (\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon}) + (\boldsymbol{\epsilon}, H\mathbf{T}) \leq (\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon}) + (|\boldsymbol{\epsilon}|, \hat{\mathbf{T}}) \ \ , \tag{3.4}$$

where the definitions for $\hat{\mathbf{T}}$ and for $|\boldsymbol{\epsilon}|$ are the same as the ones in section 2, see (2.9) and (2.11).

At this point we remark that one may obtain different estimates on the error-bound using three different approaches to analyze (3.4):

1. The first approach takes advantage of the fact that $(\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon}) \leq 0$, i.e.:

   $$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \leq (\boldsymbol{\epsilon}, \hat{M}\boldsymbol{\epsilon}) + (|\boldsymbol{\epsilon}|, \hat{\mathbf{T}}) \leq ||\boldsymbol{\epsilon}|| \ ||\hat{\mathbf{T}}|| \ \ . \tag{3.5}$$

   Then using (2.9) it follows that

   $$||\boldsymbol{\epsilon}||_H \leq \begin{cases} A\, h^m\, t & A = \max\{\alpha_B, \alpha_M\} = O(1) \quad ; \ s = 0 \\ \alpha_B \sqrt{n_l + n_r}\, h^{m-s+\frac{1}{2}}\, t & \quad ; \ s \geq 1 \end{cases} \tag{3.6}$$

   Note that for $s = 0$ we get what might be considered an optimal temporal bound, i.e. linear in t, and convergence rate, h^m. For $s \geq 1$, however, the convergence rate can be improved on.

2. In the second approach we analyze (3.4) in the same manner as (2.10), taking into account that the c_0 that existed in (2.7) vanishes here.

   We find that for $s = 1$, see Appendix A,

$$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H \leq \frac{1}{Kc_M} \left[\frac{\alpha_B^2}{4\lambda\beta}(n_L + n_R) + \frac{\alpha_M^2}{c_M} \right] h^{2m} \left(e^{2Kc_M t} - 1 \right) \quad , \tag{3.7}$$

   where, $0 < \beta < 1$, $K = O(1)$, $0 < c_M < \infty$.

   For $s \geq 2$ we get

$$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H \leq \frac{1}{Kc_M} \left[\frac{\alpha_B^2}{4\lambda\beta}(n_L + n_R) \right] h^{2(m-s+1)} \left(e^{2Kc_M t} - 1 \right) \quad . \tag{3.8}$$

   The temporal bound (3.7) is effectively exponential, unlike the (much better) linear bound given in (3.6). The convergence rate here is h^m, as found by Gustafsson [9], [10] and is an improvement over (3.6) with $s = 1$.

   In the case of $s \geq 2$, equation (3.8), it can be shown that the best temporal bound is for $c_M \ll 1$ (note from the Schwartz's inequality, eq. A.2 , that c_M can be chosen arbitrary to lie on $(0, \infty)$), and this yields $||\boldsymbol{\epsilon}||_H \sim O(1)\, h^{m-s+1} \sqrt{t}$.This estimate is valid for all $0 < t \ll \frac{2}{K c_M}$. Thus by choosing very small c_M we get a very good bound, i.e. $\sim \sqrt{t}$ for very large t, with a convergence rate of h^{m-s+1}.

   Note that in the case of $s = 1$, equation (3.7), choosing $c_M \ll 1$ will not really improve the exponential growth, since the coefficient of $\left(e^{2Kc_M t} - 1\right)$ contains a $\left(\frac{1}{c_M^2}\right)$ term.

To summarize, the second approach gives, for $s \geq 2$, the better temporal bound, ($\sim \sqrt{t}$), and convergence rate, h^{m-s+1}. For $s = 1$ the first approach gives a better temporal bound ($\sim t$), while the second approach predicts a better convergence rate (h^m). For $s = 0$ the first approach gives a better prediction.

We have developed a third, and completely different, approach that predicts "optimal" convergence rates and temporal bounds for all $s \geq 1$. This theory is delineated in the next section.

4 The hyperbolic case, an optimization approach.

In this section we would like to derive a temporal bound on the error, more benign than the one found in section 3, for the problem (3.1). The approach we take here is to define a new problem, which is somewhat similar to the original one. We shall show that the solution of this auxiliary problem bounds, in some sense (to be described shortly), the error of the original problem.

In this section we use the same notations and assumptions as in section 3, equations (3.2)-(3.4).

Let $\boldsymbol{\psi}(t)$, $\boldsymbol{\epsilon}(t) \in C^1$ and

$$\begin{aligned} \frac{d}{dt} q(\boldsymbol{\epsilon}(t)) &\leq g(\boldsymbol{\epsilon}(t)) \ , \qquad (4.1) \\ \frac{d}{dt} p(\boldsymbol{\psi}(t)) &= f(\boldsymbol{\psi}(t)) \ , \qquad (4.2) \end{aligned}$$

where $f, g, p, q : R^N \longrightarrow R$, $f, g \in C$ and $0 \leq p, q \in C^1$. Suppose that

$$\forall \boldsymbol{\phi}; \ \ g(\boldsymbol{\phi}) \leq f(\boldsymbol{\phi}) \qquad (4.3)$$

and that there are constants $c_2 \geq c_1 \geq 0$ such that

$$\forall \boldsymbol{\phi}; \quad c_1\, q(\boldsymbol{\phi}) \leq p(\boldsymbol{\phi}) \leq c_2\, q(\boldsymbol{\phi}) \ . \qquad (4.4)$$

Let

$$\Phi_\rho^{(q)} = \{\boldsymbol{\phi} \,|\, q(\boldsymbol{\phi}) = \rho\} \qquad (4.5)$$

and

$$\Phi_\rho^{(p)} = \{\boldsymbol{\phi} \,| c_1\, \rho \leq p(\boldsymbol{\phi}) \leq c_2\, \rho\} \ . \qquad (4.6)$$

Note that $\Phi_\rho^{(q)} \subset \Phi_\rho^{(p)}$. Suppose that there exist a vector, $\boldsymbol{\phi}_{\max}$ such that

$$f(\boldsymbol{\phi}_{\max}) = \max_{\boldsymbol{\phi} \in \Phi_\rho^{(p)}} f(\boldsymbol{\phi}) \ . \qquad (4.7)$$

We shall denote:

$$f_{\max}(\rho) \;\equiv\; f(\boldsymbol{\phi}_{\max}) \quad . \tag{4.8}$$

Now we would like to present the following lemma:

Lemma 1 *Let the function $\rho(t)$ be defined by the following differential equation:*

$$\frac{d}{dt}(\rho) \;=\; \hat{f}(\rho) \;\equiv\; f_{\max}(\rho) + \delta(t) \; ; \qquad \delta(t) \;>\; 0 \quad , \tag{4.9}$$

with the initial condition

$$\rho(t = t_0) \;=\; \rho_0 \quad . \tag{4.10}$$

Then, if at $t \;=\; t_0$, $\rho_0 \geq q(\,\boldsymbol{\epsilon}(t_0)\,)$, then

$$\rho(t) \geq q(\,\boldsymbol{\epsilon}(t)\,) \quad \forall t \geq t_0 \quad . \tag{4.11}$$

Proof: Note that for each time t_1 where $\rho(t_1) = q(\,\boldsymbol{\epsilon}(t_1)\,)$, by the definitions of $\Phi^{(p)}_{\rho}$ and $\Phi^{(q)}_{\rho}$, we have $\boldsymbol{\epsilon}(t_1) \;\in\; \Phi^{(q)}_{\rho} \;\subset\; \Phi^{(p)}_{\rho}$. Therefore $f(\,\boldsymbol{\epsilon}(t_1)\,) \;\leq\; f(\,\boldsymbol{\phi}_{\max}(t_1)\,) \;=\; f_{\max}(\,\rho(t_1)\,)$. Using this observation we can write the following chain of inequalities:

$$\begin{aligned} \frac{d}{dt}q(\,\boldsymbol{\epsilon}(t_1)\,) \;&\leq\; g(\,\boldsymbol{\epsilon}(t_1)\,) \\ &\leq\; f(\,\boldsymbol{\epsilon}(t_1)\,) \\ &\leq\; f_{\max}(\,\rho(t_1)\,) \\ &<\; \hat{f}(\,\rho(t_1)\,) \\ &=\; \frac{d}{dt}\rho(t_1) \; . \end{aligned} \tag{4.12}$$

Therefore there is a Δt s.t. $\forall\, t_1 \;<\; t \;<\; t_1 + \Delta t$

$$\rho(t) \;>\; q(\,\boldsymbol{\epsilon}(t)\,) \quad .$$

In particular this is true if at $t = t_0$, $\rho_0 = q(\,\boldsymbol{\epsilon}(t_0)\,)$. Since $\rho(t) \in C^1$, then if at $t = t_0$, $\rho(t_0) > q(\,\boldsymbol{\epsilon}(t_0)\,)$, there is a Δt s.t. $\forall\, t_0 < t < t_0 + \Delta t$, $\rho(t) > q(\,\boldsymbol{\epsilon}(t)\,)$. There might be a Δt^*, s.t. at $t_1 = t_0 + \Delta t^*$, $\rho(t_1) > q(\,\boldsymbol{\epsilon}(t_1)\,)$; then by (4) the process repeats itself. This completes the proof of the Lemma.

Remarks:

1 If $f \in C^1$ then the technique of Lagrange multipliers can be used to find $\boldsymbol{\phi}_{\max}$ and $f_{\max}$.

2 If $f \in C^1$ and $\nabla f(\boldsymbol{\phi}) \neq \mathbf{0}$, $\forall\ \boldsymbol{\phi} \in \mathrm{Int}\,\Phi_\rho^{(p)}$, then $f_{\max}(\rho) = \max_{\boldsymbol{\phi}\in\partial\Phi_\rho^{(p)}} f(\boldsymbol{\phi})$, where $\partial\Phi_\rho^{(p)}$ is the boundary of $\Phi_\rho^{(p)}$ and $\mathrm{Int}\,\Phi_\rho^{(p)}$ is its interior.

3 Since p and q are non-negative quantities, c_1 can be taken as 0.

In the rest of this section we shall use Lemma 1 in order to derive the convergence rate and a temporal bound for the problem (3.1).

Let us introduce a family of vector functions, $\boldsymbol{\psi}$, as follows:

$$\frac{\partial}{\partial t}\psi_j^B = -\frac{\lambda}{h}\psi_j^B + \hat{T}_j^B \qquad 1 \leq j \leq n_L, \quad N - n_R + 1 \leq j \leq N \tag{4.13}$$

$$\frac{\partial}{\partial t}\psi_j^M = \hat{T}_j^M \qquad n_L + 1 \leq j \leq N - n_R \ . \tag{4.14}$$

Note that these equations can be written as:

$$\frac{\partial}{\partial t}\boldsymbol{\psi} = \hat{M}\boldsymbol{\psi} + \hat{\mathbf{T}} \ , \tag{4.15}$$

where $\hat{M}$ is defined in (3.3), and $\boldsymbol{\psi}|_{t=0} = 0$.

We now identify the different terms in (4.1) and (4.2) by:

$$\begin{aligned} q &= (\boldsymbol{\epsilon},\boldsymbol{\epsilon})_H = ||\boldsymbol{\epsilon}||_H^2 \ , \\ g &= (\boldsymbol{\epsilon},\hat{M}\boldsymbol{\epsilon}) + (|\boldsymbol{\epsilon}|,\hat{\mathbf{T}}) \ , \\ p &= (\boldsymbol{\psi},\boldsymbol{\psi}) = ||\boldsymbol{\psi}||^2 \ , \\ f &= (\boldsymbol{\psi},\hat{M}\boldsymbol{\psi}) + (\boldsymbol{\psi},\hat{\mathbf{T}}) \ , \end{aligned} \tag{4.16}$$

where the definition of $|\boldsymbol{\epsilon}|$ is given in (2.11). The functions f, g, p and q satisfy the conditions in the Lemma. Since p and q are square of norms we take $c_1 = 0$, following Remark 3. Also since $f(\mathbf{0}) = 0$, $f(\boldsymbol{\phi}_{max}) > 0 \ \forall ||\boldsymbol{\phi}_{max}|| > 0$ and $\nabla \hat{f} \neq \mathbf{0}$, then $\boldsymbol{\phi}_{\max} \in \{\boldsymbol{\phi}\,|p(\boldsymbol{\phi}) = c_2\,\rho\}$. In the rest of the section we shall use $c_2\,\rho = R^2$. We now propose to find $\boldsymbol{\phi}_{max}$, and thus $R^2(t)$, by resorting to the technique of Lagrange multipliers. In particular, we write:

$$\nabla \hat{f} \; = \; L \nabla p \ \ , \tag{4.17}$$

where L is the Lagrange multiplier. Component-wise we have, see (4.16)

$$\left(\nabla \hat{f}(\boldsymbol{\phi}_{max})\right)_j = \frac{1}{N}\left\{\begin{array}{lc} -\frac{2\lambda}{h}(\phi_{max})_j^B + \hat{T}_j^B & 1 \le j \le n_L \ , \quad N - n_R + 1 \le j \le N \\ \\ \qquad \hat{T}_j^M & n_L + 1 \le j \le N - n_R \end{array}\right. \ , \tag{4.18}$$

$$(\nabla p(\boldsymbol{\phi}_{max}))_j = \frac{2}{N}\left\{\begin{array}{lc} (\phi_{max})_j^B & 1 \le j \le n_L \ , \quad N - n_R + 1 \le j \le N \\ \\ {\phi_{max}}_j^M & n_L + 1 \le j \le N - n_R \end{array}\right. \ . \tag{4.19}$$

Using (4.18) and (4.19) in (4.17) we can write L as one of the following two ratios:

$$L = \frac{-\frac{2\lambda}{h}(\phi_{max})_j^B + \hat{T}_j^B}{2\,(\phi_{max})_j^B} \ \ , \tag{4.20}$$

and

$$L = \frac{\hat{T}_j^M}{2\,(\phi_{max})_j^M} \ \ . \tag{4.21}$$

Recall that all the $\hat{T}_j^B$'s are equal to each other, ($T_B \overset{\triangle}{=} \hat{T}_j^B$), and the same is true for the $\hat{T}_j^M$'s ($T_M \overset{\triangle}{=} \hat{T}_j^M$). Therefore it follows from (4.20) that all the $(\phi_{max})_j^B$'s are equal to each other; and from (4.21) we have that all the $(\phi_{max})_j^M$'s are the same. Thus, using (4.18) and (4.19), problem (4.17) instead of having $(N+1)$ unknowns, has only 3, namely ϕ_B, ϕ_M and L, where

$$\phi_B = (\phi_{max})_j^B \quad , \tag{4.22}$$

$$\phi_M = (\phi_{max})_j^M \quad . \tag{4.23}$$

In terms of ϕ_B, ϕ_M and L, the expressions for $\hat{f}$, $p(\boldsymbol{\phi}_{max})$ and equations (4.20) and (4.21) become:

$$\begin{aligned}
p(\boldsymbol{\phi}_{max}) &= n\,h\,{\phi_B}^2 \,+\, (1 - n\,h){\phi_M}^2 \,; \qquad (\,n \,=\, n_l \,+\, n_r\,) \;\;, && (4.24)\\
\hat{f} &= -2nh\frac{\lambda}{h}{\phi_B}^2 + 2nhT_B\phi_B + 2(1-nh)T_M\phi_M + \delta \;\;, && (4.25)\\
L &= \frac{-\frac{2\lambda}{h}\phi_B + T_B}{2\,\phi_B} \quad , && (4.26)\\
L &= \frac{T_M}{2\,\phi_M} \quad . && (4.27)
\end{aligned}$$

Eliminating L between (4.26) and (4.27) we have a relation between ϕ_B and ϕ_M:

$$\phi_M = \frac{T_M\phi_B}{T_B - \frac{2\lambda}{h}\phi_B} \quad . \tag{4.28}$$

We can now write $p(\boldsymbol{\phi}_{max})$ and $\hat{f}$ as a function of ϕ_B only:

$$\begin{aligned}
p(\boldsymbol{\phi}_{max}) &= {\phi_B}^2\left[nh + (1-nh)\left(\frac{T_M}{T_B - \frac{2\lambda}{h}\phi_B}\right)^2\right] \quad , && (4.29)\\
\hat{f} &= 2\phi_B\left[-n\lambda\phi_B + nhT_B + (1-nh)\frac{T_M^2}{T_B - \frac{2\lambda}{h}\phi_B}\right] + \delta \;\;. && (4.30)
\end{aligned}$$

Remark: The classical approach is to solve (4.29) for $\phi_B = \phi_B(\rho)$ where, see paragraph after (4.16), $c_2\,\rho \overset{\triangle}{=} p(\boldsymbol{\phi}_{max})$. Then substitute this $\phi_B(\rho)$ into (4.30) to find $\hat{f}(\rho)$. One then goes to (4.9), and solves for $\rho(t)$. From Lemma 1, $\sqrt{\rho}$ is the bound on the error-norm. However, this method involves solving a quartic equation, and therefore we use a different approach.

In order to simplify the calculations we introduce a new function, σ, see (4.32) below and then we solve (4.9). Finally we will show that the solution we get is indeed $\boldsymbol{\phi}_{\max}$. Using $R^2 = c_2\,\rho$ we rewrite equation (4.9) as follows:

$$c_2\,\hat{f} = \frac{d}{dt}(R^2) = \left[\frac{d}{d\sigma}(R^2)\frac{d\sigma}{dt}\right] \quad , \tag{4.31}$$

where

$$\sigma = \frac{(T_B - \frac{2\lambda}{h}\phi_B)}{T_M} \ . \tag{4.32}$$

Then from (4.31)

$$\frac{1}{c_2}\frac{d}{dt}\sigma = \frac{\hat{f}(\sigma)}{\frac{d}{d\sigma}R^2(\sigma)} \quad . \tag{4.33}$$

Note that c_2 is a "norm-equivalence" constant of order unity, and in order to simplify the notation we absorb it into the time t. After some manipulation (4.33) becomes,

$$\frac{d\sigma}{dt} = -n\lambda\sigma^2 \frac{\sigma^2 + \left(\frac{T_B}{T_M}\right)\sigma + \frac{2(1-nh)}{nh} + \delta_0}{nh\sigma^3 + (1-nh)\left(\frac{T_B}{T_M}\right)} \quad , \tag{4.34}$$

where we choose

$$\delta(t) = \delta_0 \frac{(T_B - T_M\sigma)nh^2 T_M}{2\lambda\sigma} \ . \tag{4.35}$$

Note that at $t = 0$, $\phi = 0$ and from (4.32) $\sigma = T_B/T_M$, see initial conditions for (4.15), and that $d\sigma/dt < 0 \quad \forall\sigma > 0$. Before solving (4.34) for $\sigma(t)$, we rewrite (4.29) ($p(\boldsymbol{\phi}_{max}) = c_2\,\rho = R^2$) in terms of σ, using (4.32):

$$R^2 = \frac{nh^3 T_M^2}{4\lambda^2}\left(\frac{T_B}{T_M} - \sigma\right)^2 \left[1 + \frac{1-nh}{nh}\frac{1}{\sigma^2}\right] \tag{4.36}$$

We now use (4.34) to write an expression for $dt/d\sigma$:

$$\begin{aligned} \frac{dt}{d\sigma} &= -\frac{1}{n\lambda\sigma^2}\frac{nh\sigma^3 + (1-nh)\left(\frac{T_B}{T_M}\right)}{\sigma^2 + \left(\frac{T_B}{T_M}\right)\sigma + \frac{2(1-nh)}{nh} + \delta_0} \\ &= \frac{k_1}{(\sigma-\sigma_1)} + \frac{k_2}{(\sigma-\sigma_2)} + \frac{k_3\sigma + k_4}{\sigma^2} \quad , \end{aligned} \tag{4.37}$$

where:

$$\begin{aligned}
\sigma_1 &= \frac{1}{2}\left(\frac{T_B}{T_M}\right)\left[-1+\sqrt{1-4\frac{1}{\left(\frac{T_B}{T_M}\right)^2}\left[\frac{1-nh}{nh}+\delta_0\right]}\;\right] \quad , \\
\sigma_2 &= \frac{1}{2}\left(\frac{T_B}{T_M}\right)\left[-1-\sqrt{1-4\frac{1}{\left(\frac{T_B}{T_M}\right)^2}\left[\frac{1-nh}{nh}+\delta_0\right]}\;\right] \quad , \\
k_1 &= \frac{h}{\lambda}\,\frac{\sigma_1^3+\frac{1-nh}{nh}\left(\frac{T_B}{T_M}\right)}{\sigma_1^2(\sigma_2-\sigma_1)} \quad , \\
k_2 &= -\frac{h}{\lambda}\,\frac{\sigma_2^3+\frac{1-nh}{nh}\left(\frac{T_B}{T_M}\right)}{\sigma_2^2(\sigma_2-\sigma_1)} \quad , \\
k_3 &= -\frac{(1-nh)(\sigma_1+\sigma_2)\left(\frac{T_B}{T_M}\right)}{\lambda n\sigma_1^2\sigma_2^2} \quad , \\
k_4 &= -\frac{(1-nh)\left(\frac{T_B}{T_M}\right)}{\lambda n\sigma_1\sigma_2} \quad .
\end{aligned} \tag{4.38}$$

Using the fact that

$$\sigma_1+\sigma_2 = -\left(\frac{T_B}{T_M}\right) \quad , \tag{4.39}$$

$$\sigma_1\sigma_2 = 2\frac{1-nh}{nh}+\delta_0 \tag{4.40}$$

and that

$$\frac{T_B}{T_M} = O(h^{-s}) = Kh^{-s} \qquad s \geq 1 \quad , \tag{4.41}$$

we find that

$$\begin{aligned}
\sigma_1 &= -\frac{2}{\left(\frac{T_B}{T_M}\right)}\left(\frac{1-nh}{nh}+\frac{\delta_0}{2}\right)-\frac{4}{\left(\frac{T_B}{T_M}\right)^3}\left(\frac{1-nh}{nh}+\frac{\delta_0}{2}\right)^2+O\left(\frac{1}{h^3\left(\frac{T_B}{T_M}\right)^5}\right) \quad , \qquad (4.42) \\
\sigma_2 &= -\left(\left(\frac{T_B}{T_M}\right)+\sigma_1\right)
\end{aligned}$$

$$= -\left(\frac{T_B}{T_M}\right) + \frac{2}{\left(\frac{T_B}{T_M}\right)}\left(\frac{1-nh}{nh}+\frac{\delta_0}{2}\right) + \frac{4}{\left(\frac{T_B}{T_M}\right)^3}\left(\frac{1-nh}{nh}+\frac{\delta_0}{2}\right)^2 + O\left(\frac{1}{h^3\left(\frac{T_B}{T_M}\right)^5}\right) \quad , \tag{4.43}$$

$$k_1 = -k_3 + \frac{h}{2\lambda}\,\frac{\sigma_1}{\sigma_2} + + \frac{h}{2\lambda}\left(\frac{\sigma_1+\sigma_2}{\sigma_1^2(\sigma_1-\sigma_2)} - \frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\right)\delta_0 \quad , \tag{4.44}$$

$$k_2 = -\frac{h}{\lambda}\left[1+\frac{1}{2}\frac{\sigma_1}{\sigma_2}+\frac{1}{2}\frac{\sigma_1+\sigma_2}{\sigma_2^2(\sigma_2-\sigma_1)}\delta_0\right] \quad , \tag{4.45}$$

$$k_3 = \frac{h}{2\lambda}\left[\frac{(\sigma_1+\sigma_2)^2}{\sigma_1\sigma_2} - \left(\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\right)^2\delta_0\right] \quad , \tag{4.46}$$

$$k_4 = \frac{h}{2\lambda}\left[(\sigma_1+\sigma_2) - \left(\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\right)\delta_0\right] \quad , \tag{4.47}$$

$$\frac{k_4}{|\sigma_1|} = -k_3 + \frac{h}{2\lambda}\left(1+\frac{\sigma_1}{\sigma_2}\right) - \frac{h}{2\lambda}\frac{\sigma_1+\sigma_2}{(\sigma_1\sigma_2)^2}(\sigma_1+2\sigma_2)\delta_0 \quad . \tag{4.48}$$

Integrating (4.37) (noting, again, that the initial condition $\phi_B(t=0)=0$ implies $\sigma(t=0)=T_B/T_M$) we have:

$$\begin{aligned} t &= \left[k_1\ln(\sigma-\sigma_1)+k_2\ln(\sigma-\sigma_2)+k_3\ln(\sigma)-\frac{k_4}{\sigma}\right] - \\ &\quad \left[k_1\ln\left(\frac{T_B}{T_M}-\sigma_1\right)+k_2\ln\left(\frac{T_B}{T_M}-\sigma_2\right)+k_3\ln\left(\frac{T_B}{T_M}\right)-\frac{k_4}{\left(\frac{T_B}{T_M}\right)}\right] \quad . \end{aligned} \tag{4.49}$$

Recalling that $T_B/T_M = O(h^{-s})$, one can show from (4.18)-(4.48) that if

$$\delta_0 = o(h^{2(s-1)}) \quad , \tag{4.50}$$

then

$$k_1\ln\left(\frac{T_B}{T_M}-\sigma_1\right)+k_3\ln\left(\frac{T_B}{T_M}\right)-\frac{k_4}{\left(\frac{T_B}{T_M}\right)} = O(h) \tag{4.51}$$

and also

$$k_2\left[\ln(\sigma-\sigma_2)-\ln\left(\frac{T_B}{T_M}-\sigma_2\right)\right] = O(h) \qquad \forall\left(\frac{T_B}{T_M}\right)\geq\sigma\geq 0 \quad . \tag{4.52}$$

The implication of (4.51) and (4.52) is that (4.49) is equivalent to

$$t = k_1 \ln(\sigma - \sigma_1) + k_3 \ln(\sigma) - \frac{k_4}{\sigma} + O(h) \ . \tag{4.53}$$

We now recall again that in order to evaluate R, we need to solve for $\sigma = \sigma(t;h)$. We assume that the following (possibly asymptotic) expansion of σ is valid:

$$\sigma(t;h) = |\sigma_1| h^{\nu} \sum_{j=0}^{\infty} h^{\gamma_j} \alpha_j(t) \ \ , \tag{4.54}$$

where $\gamma_0 = 0$, $\gamma_{j+1} > \gamma_j$. In Appendix B we carry out the analysis involving the asymptotic expansion (4.54) and we obtain there the following results:

i For $s = 1$ one gets $\nu = 0$, and

$$t = \frac{nK^2}{4\lambda}\left[\frac{1}{\alpha_0(t)} - \ln\left(1 + \frac{1}{\alpha_0(t)}\right)\right] + O(h) \ \ , \tag{4.55}$$

from which one deduces the following behavior of $\alpha_0(t)$:

$$\alpha_0(t) \cong \frac{\sqrt{n}K}{\sqrt{8\lambda}} \frac{1}{\sqrt{t}}; \qquad \frac{t}{h} \gg 1\, ,\, t \ll 1 \tag{4.56}$$

$$\alpha_0(t) \cong \frac{nK^2}{4\lambda} \frac{1}{t}; \qquad 1 \ll t \tag{4.57}$$

and $\forall t \geq 0$,

$$R = \left[\frac{nK^2}{4\lambda} + O(h)\right] \frac{T_M}{\alpha_0(t)} \ \ . \tag{4.58}$$

Thus the error-bound, R, has a temporal rise like $\sqrt{t}$ for short times, and 'bends-over' to a linear growth. We also note that the bound on the error, R, has the same convergence rate, T_M, as the inner scheme.

ii $s \geq 2$: One gets

$$\nu = s - 1 \tag{4.59}$$

and

$$t = \frac{nK^2}{8\lambda}\frac{1}{\alpha_0^2(t)} + O(h), \qquad h \ll t \ . \tag{4.60}$$

Then, using (B.12),

$$\begin{aligned} R &= \left[\frac{nK^2}{4\lambda} + O(h)\right]\frac{h^{1-s}}{\alpha(t)}T_M \\ &\approx \frac{\sqrt{n}K}{\sqrt{2\lambda}}h^{1-s}T_M\sqrt{t} \ . \end{aligned} \tag{4.61}$$

Remark : Although the temporal growth indicated in (4.61) ($s \geq 2$) is slower than the linear one ($t \gg 1$) for the case $s = 1$, the actual R, predicted by (4.61), is larger than the error bound in the case $s = 1$, (4.58), because $h^{1-s}\sqrt{t} \gg t$ for practical values of h and t. Thus it is still worthwhile to use boundary stencils of higher accuracy, namely $s = 1$.

Finally we prove that the R^2 which we got is indeed the maximal one. Note that equations (4.24) and (4.25) can be rewritten, for a given R^2, as:

$$R^2 = nh{\phi_B}^2 + (1-nh){\phi_M}^2 \ , \tag{4.62}$$

$$\phi_M = \frac{1}{2(1-nh)T_M}\left[2nh\frac{\lambda}{h}{\phi_B}^2 - 2nhT_B\phi_B + \hat{f}\right] \ , \tag{4.63}$$

i.e. the contour of (4.62) is a canonical ellipse with axes $2\frac{R}{\sqrt{n\,h}}$ and $2\frac{R}{\sqrt{(1-n\,h)}}$ in ϕ_B and ϕ_M respectively. The contours of (4.63), for a fixed $\hat{f}$, are paraboli with a minimum at $\phi_B = \frac{h\,T_B}{\lambda} > 0$. As $\hat{f}$ increases the paraboli 'climb up' the ϕ_M axis. The geometrical interpretation of Lagrange multipliers is to find the $\hat{f}$ for which the corresponding paraboli are tangential to the ellipse. The maximal $\hat{f}$ is the desired solution; see Figure 2. Clearly, this maximal $\hat{f}$ corresponds to the top parabola and in this case ϕ_B, $\phi_M > 0$. Thus all we have to show is that ϕ_B and ϕ_M are positive.

As can be seen from (4.28) and (4.32),

$$\phi_B = \frac{h}{2\,\lambda}\left(T_B - T_M\,\sigma\right) \ , \tag{4.64}$$

$$\phi_M = \frac{\phi_B}{\sigma} \ . \tag{4.65}$$

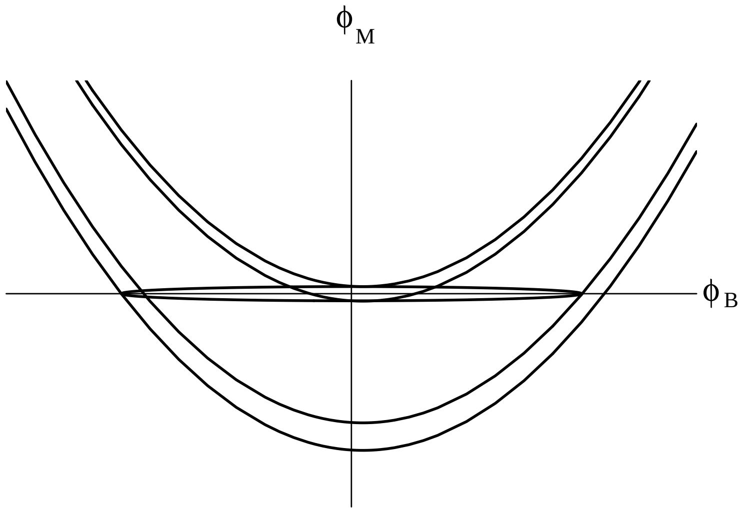

Figure 2:

Since $\sigma = T_B/T_M$ at $t = 0$ and is a monotonically decreasing positive function of t, $\forall t > 0$, then ϕ_B and ϕ_M are indeed positive.

5 Numerical examples

In this section two examples are given: a variable coefficient diffusion equation and a simple hyperbolic problem.

5.1 Diffusion equation

We solve the following problem:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left((1+x)^2 \, \frac{\partial u}{\partial x} \right) \; + \; 100\,(1+x)\; \cos(10\,x); \qquad 0 \le x \le 1, \tag{5.1a}$$

$$u(x,0) = \frac{\sin(\frac{\log(1+x)}{\log\, 2}\,\pi}{\sqrt{1+x}} + \frac{\cos(10\,x)}{1+x} \quad , \tag{5.1b}$$

$$u(0,t) = 1 \quad , \tag{5.1c}$$

$$u(1,t) = \frac{\cos(10)}{2} \approx -0.419536 \quad . \tag{5.1d}$$

The exact solution to this problem is:

$$u(x,t) \;=\; \exp\left[-\left(\left(\frac{\pi}{\log 2}\right)^2+\frac{1}{4}\right)\,t\,\right]\,\frac{\sin\left(\frac{\log(1+x)}{\log 2}\,\pi\right)}{\sqrt{1+x}}\;+\;\frac{\cos(10\,x)}{1+x}\quad .$$

For solving the problem (5.1) we used a 4th order scheme for the middle points and 3rd, 2nd and first order stencils near the boundaries. The details of the schemes are given in Appendix C. Though variable-coefficients problems were not considered explicitly in Section 2, it can be shown (see footnote regarding to eq. (2.6)) that the schemes used satisfy all the required conditions, (2.7) - (2.9), given in Section 2, see [4].

local order		$\|\|\epsilon(t=2)\|\|_{L_2}$ for $1/h=$						convergence rate	
middle	boundary	32	64	128	256	512	1024	theory	measured
4	3	1.04e-4	3.95e-6	2.31e-7	1.42e-8	8.88e-10	5.26e-11	4	4.04
4	2	1.05e-3	7.59e-5	5.12e-6	3.32e-7	2.11e-8	1.34e-9	3.5	3.95
4	1	4.50e-3	5.56e-4	6.76e-5	8.29e-6	1.03e-6	1.27e-7	2.5	3.02

Table 1: The errors for the diffusion problem.

The result are presented in figure 3 and table 1. As one can see, the time behavior of the solution is bounded, as predicted by the theory. The computed convergence rate agrees with the theory when the difference between the order of the scheme in the middle and at the boundary is 1, i.e. $s=1$, and is better then the theoretical prediction by half ($m-s+2$ rather then $m-s+3/2$) for all other cases, $s>1$. This state of affairs is not unusual when one derives an *upper* bound on the error, as we have done here.

5.2 Hyperbolic equation

We solved the following problem:

$$\frac{\partial u}{\partial t}=\frac{\partial u}{\partial x}\ , \tag{5.3a}$$

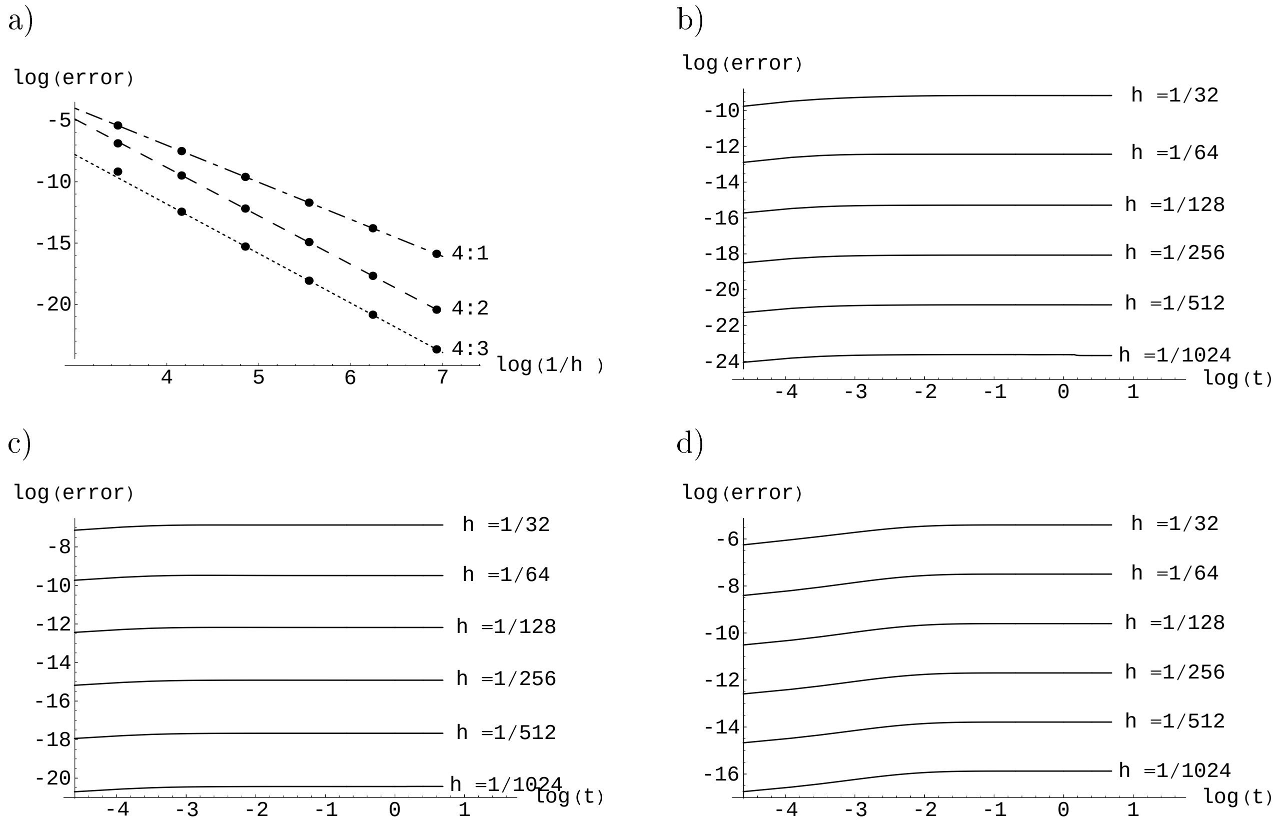

Figure 3: Diffusion equation: (a) Plots of $\log(\epsilon)$ vs. $\log(1/h)$ at $t = 2$ for the different schemes. (b), (c) and (d), plots of $\log(\epsilon)$ vs. $\log(t)$ for the 4:3, 4:2 and 4:1 schemes respectively.

$$u(x,0) = \cos(10\,x) \quad , \tag{5.3b}$$

$$u(1,t) = \cos(10(1+t)) \quad . \tag{5.3c}$$

The exact solution to this problem is:

$$u(x,t) \;=\; \cos(10(x+t)) \quad .$$

For solving this problem, we used a 4th order scheme for the middle points and 3rd, 2nd and first order stencils near the boundaries that were derived by Strand, [19]. Note that it is not clear whether or not these schemes satisfy the conditions on the differentiate matrix given is Sections 3 and 4.

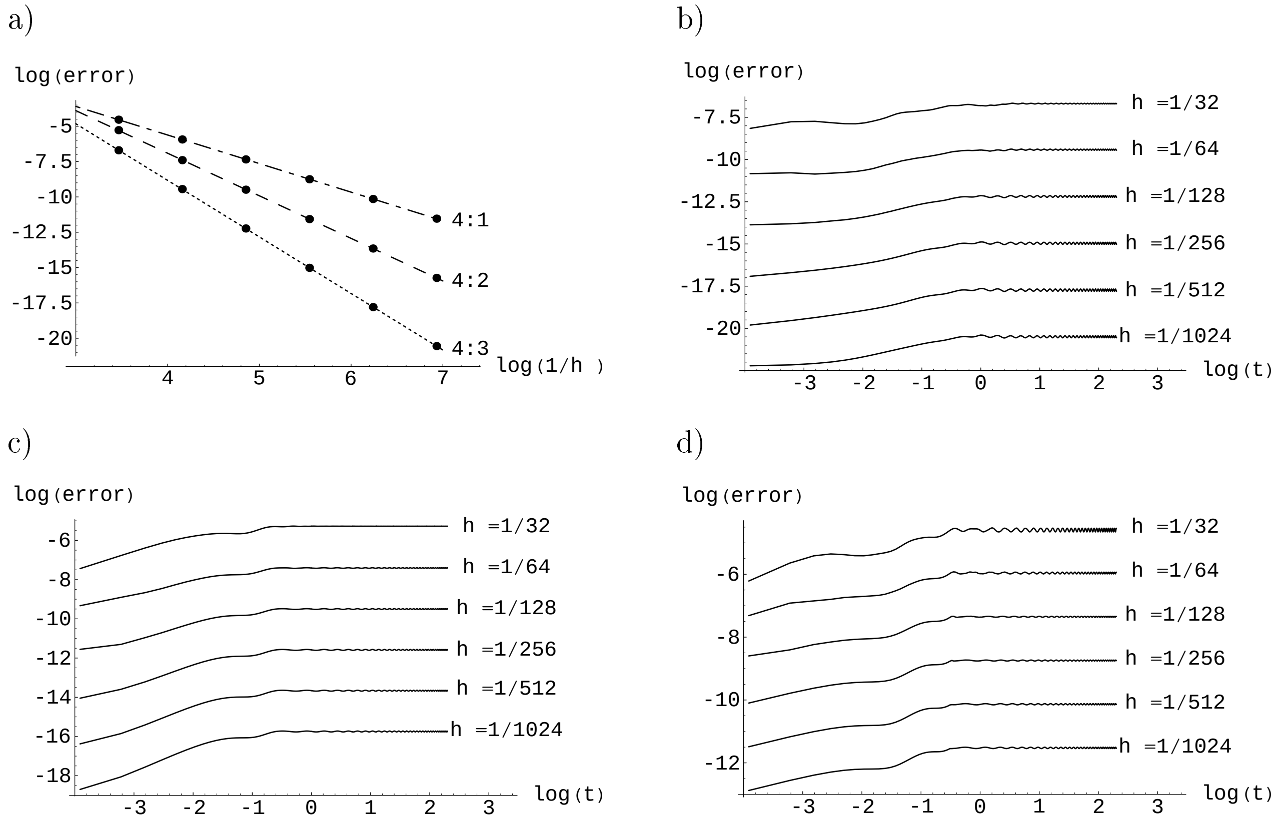

Figure 4: Hyperbolic equation: (a) Plots of $\log(\epsilon)$ vs. $\log(1/h)$ at $t = 10$ for the different schemes. (b), (c) and (d), plots of $\log(\epsilon)$ vs. $\log(t)$ for the 4:3, 4:2 and 4:1 schemes respectively.

The result are presented in figure 4 and table 2. As can be seen, here, the actual convergence rate agrees with the theory for all cases. The theoretical time behavior, which predicts a linear growth in time, is too conservative, as can be seen in figure 4. It is often observed that the error is bounded in time for hyperbolic problems in which the wave 'stays in the

local order		$\|\|\epsilon(t=10)\|\|_{L_2}$ for $1/h =$						convergence rate	
middle	boundary	32	64	128	256	512	1024	theory	measured
4	3	1.23e-3	7.91e-5	4.87e-6	3.01e-7	1.87e-8	1.19e-9	4	4.00
4	2	5.11e-3	6.12e-4	7.58e-5	9.46e-6	1.18e-6	1.48e-7	3	3.01
4	1	1.07e-2	2.65e-3	6.46e-4	1.59e-4	3.92e-5	9.76e-6	2	2.02

Table 2: The errors for the hyperbolic problem.

computational domain for a limited time'. Some examples for this case are scalar advection equation, such as the one given here, and wave scattering problems. In the cases when the wave is 'restricted' to the domain, as in the case of periodic boundary conditions or perfectly conducting cavities, the numerical solution often propagates with a slightly different velocity than the exact solution. This phenomenon manifests itself as a 'linear' growth in time, for very large times. For examples see [5]

6 Conclusions

1. The dependence of the error bounds on mesh size and time has been determined for a family of spatially semi-discrete approximations, including boundary treatments, to hyperbolic and parabolic partial differential equations.

2. The form of the theory presented herein is extendible to multi-dimensional problems. For instance, the case of $m = 4$ for the 2-D diffusion equation was investigated in [1], and fits into the present framework. For the hyperbolic case, with $m = 2$, see [1] and [3].

3. The present framework can also accommodate systems of PDE's. The case of parabolic systems, with $m = 2$, has been analyzed, see [3].

4. An analysis similar to the one presented in section 4 can also be carried out for the case of the diffusion equation. However, since the bounds are similar to the ones derived in section 2, this analysis is not included in this paper.

5. The present results point to the importance of understanding the dependence of the error on both time and mesh size. Thus, for a given grid, the temporal behavior will tell us when the numerical results exceeds a given error threshold. Conversely for a given time, say when the error grows linearly with t, we can decide what is the necessary mesh size that is needed on order to stay below that threshold.

Appendix A

In this Appendix an analysis, similar to the one done for the parabolic case in section 2, will be carried out for the hyperbolic case.

The component-wise version of (3.4) is:

$$\begin{aligned}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \quad \leq \quad & \frac{1}{N}\left[\sum_{j=1}^{n_L} \epsilon_j\left(-\frac{\lambda}{h}\epsilon_j\right) + \sum_{l=N-n_R+1}^{N} \epsilon_l\left(-\frac{\lambda}{h}\epsilon_l\right)\right] + \\ & \frac{1}{N}\left[\sum_{j=1}^{n_L} |\epsilon_j| T_B + \sum_{k=n_L+1}^{N-n_R} |\epsilon_k| T_M + \sum_{l=N-n_R+1}^{N} |\epsilon_l| T_B\right] \quad . \end{aligned} \tag{A.1}$$

Next we use the Schwartz inequality of the form

$$|fg| \leq cf^2 + \frac{1}{4c}g^2 \; , \quad c > 0 \; . \tag{A.2}$$

Then (A.1)becomes:

$$\begin{aligned}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \quad \leq \quad & h\sum_{j=1}^{n_L}\left[-\frac{\lambda}{h}\epsilon_j^2 + \left(c_B|\epsilon_j|^2 + \frac{1}{4c_B}T_B^2\right)\right] + \\ & h\sum_{k=n_L+1}^{N-n_R}\left[c_M|\epsilon_k|^2 + \frac{1}{4c_M}T_M^2\right] + \end{aligned}$$

$$h \sum_{l=N-n_R+1}^{N} \left[-\frac{\lambda}{h}\epsilon_l^2 + \left(c_B|\epsilon_l|^2 + \frac{1}{4c_B}T_B^2 \right) \right] \quad . \tag{A.3}$$

Let us choose $c_B = \beta\frac{\lambda}{h}$ $(0 < \beta < 1)$, to get

$$\begin{aligned} (\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \ \leq \ & h\sum_{j=1}^{n_L} -(1-\beta)\frac{\lambda}{h}\epsilon_j^2 + h\sum_{k=n_L+1}^{N-n_R} c_M\epsilon_k^2 + h\sum_{l=N-n_R+1}^{N} -(1-\beta)\frac{\lambda}{h}\epsilon_l^2 + \\ & h\sum_{j=1}^{n_L} \frac{1}{4}\frac{h}{\lambda\beta}\alpha_B^2 h^{2m-2s} + h\sum_{k=n_L+1}^{N-n_R} \frac{1}{4c_M}\alpha_M^2 h^{2m} + h\sum_{l=N-n_R+1}^{N} \frac{1}{4}\frac{h}{\lambda\beta}\alpha_B^2 h^{2m-2s} \quad . \end{aligned} \tag{A.4}$$

Next we note that $-(1-\beta)\lambda/h < 0$ while $c_M > 0$. We use this fact to simplify (A.4)

$$\begin{aligned} (\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_t)_H \ & \leq \ c_M\, h\sum_{j=1}^{N}\epsilon_j^2 + \frac{\alpha_B^2}{4\lambda\beta}\sum_{j=1}^{n_L+n_R} h^{2m+2-2s} + \frac{\alpha_M^2}{4c_M}\sum_{j=1}^{N-n_L-n_R} h^{2m+1} \\ & \leq \ c_M(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) + \frac{\alpha_B^2}{4\lambda\beta}(n_L+n_R)h^{2m+2-2s} + \frac{\alpha_M^2}{c_M}\frac{N-(n_L+n_R)}{N}h^{2m} \\ & \leq \ Kc_M(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H + \frac{\alpha_B^2}{4\lambda\beta}(n_L+n_R)h^{2m+2-2s} + \frac{\alpha_M^2}{c_M}h^{2m} \quad , \end{aligned} \tag{A.5}$$

where we introduced K, taking account of the equivalence of the L_2 and H norms.

We now distinguish between the two cases:

i $s = 1$

ii $s \geq 2$

In the first case ($s = 1$) the last two terms term in (A.5) are of the same order, and we have

$$\frac{\partial}{\partial t}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H \leq 2Kc_M(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H + 2\left[\frac{\alpha_B^2}{4\lambda\beta}(n_L+n_R) + \frac{\alpha_M^2}{c_M}\right]h^{2m} \quad , \tag{A.6}$$

then, it can be shown that

$$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H \leq \frac{1}{Kc_M} \left[\frac{\alpha_B^2}{4\lambda\beta}(n_L + n_R) + \frac{\alpha_M^2}{c_M} \right] h^{2m} \left(e^{2Kc_M t} - 1 \right) \quad . \tag{A.7}$$

Next we consider the third case, $s \geq 2$. Now the middle term in (A.5) dominates the last one and we can write,

$$\frac{\partial}{\partial t}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H \leq 2Kc_M(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H + 2 \left[\frac{\alpha_B^2}{4\lambda\beta}(n_L + n_R) \right] h^{2(m-s+1)} \quad , \tag{A.8}$$

and

$$(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})_H \leq \frac{1}{Kc_M} \left[\frac{\alpha_B^2}{4\lambda\beta}(n_L + n_R) \right] h^{2(m-s+1)} \left(e^{2Kc_M t} - 1 \right) \quad . \tag{A.9}$$

Appendix B

In this Appendix the asymptotic analysis for $\sigma(t; h)$ (and thus $R(t; h)$) is carried out.

The asymptotic expansion of $\sigma(t; h)$, see equation (4.54) is:

$$\sigma(t; h) = |\sigma_1| h^{\nu} \sum_{j=0}^{\infty} h^{\gamma_j} \alpha_j(t) \quad . \tag{B.1}$$

One can show from (4.42), with $\delta_0 = o(h^{2(s-1)})$, that

$$|\sigma_1| = \frac{2}{nK} h^{s-1} \left[1 - \frac{nK}{2} Q_s h + o(h) \right] \quad , \tag{B.2}$$

where

$$Q_s = \begin{cases} \frac{2}{K}\left(1 + \frac{2}{n^2K^2}\right) & s = 1 \\ \frac{2}{K} & s \geq 2 \end{cases} \quad . \tag{B.3}$$

Using (B.2) and (B.3), equation (B.1) can be written as:

$$\begin{aligned} \sigma(t; h) &= \frac{2}{nK} h^{s-1+\nu} \alpha_0(t) \left[1 - \frac{nK}{2} Q_s h + o(h) \right] \left[1 + \frac{\alpha_1(t)}{\alpha_0(t)} h^{\gamma_1} + \ldots + o(h) \right] \\ &= \frac{2}{nK} h^{s-1+\nu} \alpha_0(t) \left[1 - \frac{nK}{2} Q_s h + \frac{\alpha_1(t)}{\alpha_0(t)} h^{\gamma_1} + \frac{\alpha_2(t)}{\alpha_0(t)} h^{\gamma_2} + \ldots + o(h) \right] \quad , \end{aligned} \tag{B.4}$$

(As long as γ_2 has not been determined we really don't know how many terms to carry.)

Next we consider, see (4.53),

$$t = k_1 \ln(\sigma - \sigma_1) + k_3 \ln(\sigma) - \frac{k_4}{\sigma} + O(h) \quad . \tag{B.5}$$

Using (4.44) and (4.48), eq. (B.5) becomes

$$\begin{aligned} t &= k_3 \ln \frac{\sigma}{\sigma - \sigma_1} + k_3 \frac{|\sigma_1|}{\sigma} + O(h) \\ &= k_3 \left[- \ln \frac{\sigma - \sigma_1}{\sigma} + \frac{|\sigma_1|}{\sigma} \right] + O(h) \\ &= k_3 \left[- \ln \left(1 + \frac{|\sigma_1|}{\sigma} \right) + \frac{|\sigma_1|}{\sigma} \right] + O(h) \\ &= k_3 \left\{ - \ln \left[1 + \frac{1}{h^\nu \alpha_0(t) \left[1 + \frac{\alpha_1(t)}{\alpha_0(t)} h^{\gamma_1} + \ldots \right]} \right] + \frac{1}{h^\nu \alpha_0(t) \left[1 + \frac{\alpha_1(t)}{\alpha_0(t)} h^{\gamma_1} + \ldots \right]} \right\} + O(h) \quad , \end{aligned} \tag{B.6}$$

or

$$t = \frac{nK^2}{4\lambda} (1 + O(h)) \left[- \ln (1 + B) + B \right] + O(h) \quad , \tag{B.7}$$

where

$$B = \frac{1}{h^\nu \alpha_0(t) \left[1 + \frac{\alpha_1(t)}{\alpha_0(t)} h^{\gamma_1} + \ldots \right]} \quad . \tag{B.8}$$

The case of $s = 1$

Since the LHS of (B.7) is t, the RHS must also be a function of t only (up to $O(h)$), and it follows immediately, see (B.8), that $\nu = 0$. To the lowest order of approximation we then have

$$\begin{aligned} t &= \frac{nK^2}{4\lambda} (1 + O(h)) \left[- \ln \left(1 + \frac{1}{\alpha_0(t)} \right) + \frac{1}{\alpha_0(t)} \right] + O(h) \\ &= \frac{nK^2}{4\lambda} \left[- \ln \left(1 + \frac{1}{\alpha_0(t)} \right) + \frac{1}{\alpha_0(t)} \right] + O(h) \quad . \end{aligned} \tag{B.9}$$

It is not easy to describe analytically the behavior of $\alpha_0(t) \ \ \forall t$, therefore we limit the discussion to the following two cases,

i $\alpha_0(t) \gg 1$

ii $\alpha_0(t) \ll 1$

and we ask what are the ranges of t for which each case, (i) or (ii), is valid.

i $\alpha_0(t) \gg 1$, i.e.

$$\ln\left(1+\frac{1}{\alpha_0}\right) = \frac{1}{\alpha_0} - \frac{1}{2}\frac{1}{\alpha_0^2} + \ldots \quad .$$

Then eq. (B.9) becomes

$$t = \frac{nK^2}{4\lambda}\left[\frac{1}{\alpha_0} - \left(\frac{1}{\alpha_0} - \frac{1}{2\alpha_0^2} + \ldots\right)\right] = \frac{nK^2}{8\lambda}\frac{1}{\alpha_0^2} + O(h) + O(\frac{1}{\alpha_0^3})$$

and

$$\alpha_0 = \frac{\sqrt{n}k}{\sqrt{8\lambda}}\frac{1}{\sqrt{t}}\left[1 + O(h) + O(\frac{1}{\alpha_0})\right] \quad ,$$

and since $\alpha_0(t) \gg 1$ this expression is valid for

$$\frac{t}{h} \gg 1 \quad \text{and} \quad t \ll 1 \quad .$$

ii $\alpha_0(t) \ll 1$ then eq. (B.9) may be written as

$$t = \frac{nK^2}{4\lambda}\left[\frac{1}{\alpha_0} - \ln\left(1 + \frac{1}{\alpha_0(t)}\right)\right] = \frac{nK^2}{4\lambda}\left[\frac{1}{\alpha_0} - \ln\left(\frac{1}{\alpha_0(t)}\right)\right] + O(h) \quad ,$$

so, asymptotically

$$\alpha_0 \cong \frac{nK^2}{4\lambda t} \quad ,$$

and this is valid for

$$1 \ll t \quad .$$

The case of $s \geq 2$

From (B.7), using the formula of expanding $\ln(1+x)$, $x \ll 1$

$$t = \frac{nK^2}{8\lambda} h^{2(1-s)} \frac{(1+O(h)}{h^{2\nu}\alpha_0^2} \frac{1}{\left(1+\frac{\alpha_1(t)}{\alpha_0(t)}h^{\gamma_1}+\frac{\alpha_2(t)}{\alpha_0(t)}h^{\gamma_2}\right)^2} \left[1-\frac{(\frac{2}{3})h^{s-1}}{\alpha_0\left(1+\frac{\alpha_1(t)}{\alpha_0(t)}h^{\gamma_1}+\frac{\alpha_2(t)}{\alpha_0(t)}h^{\gamma_2}\right)}\right] + O(h) \ . \tag{B.10}$$

It is clear that $\nu = 1-s$ for the RHS to be function of t only, to order h, and so

$$t = \frac{nK^2}{8\lambda\alpha_0^2(t)}\left[1-2\frac{\alpha_1(t)}{\alpha_0(t)}h^{\gamma_1}+O(h^{\gamma_2})\right]+O(h) \ . \tag{B.11}$$

Using the lowest order term of the approximation we get

$$\alpha_0 = \frac{\sqrt{n}k}{\sqrt{8\lambda}\sqrt{t}} \ . \tag{B.12}$$

To find γ_1 we substitute (B.11) in (B.10) and we find

$$0 = -2\frac{\alpha_1(t)}{\alpha_0(t)}h^{\gamma_1}+O(h^{\gamma_2})+O(h) \ . \tag{B.13}$$

Since by assumption $\gamma_1 < \gamma_2$, we must have

$$\gamma_1 = 1 \ . \tag{B.14}$$

It is clear from (B.13) that

$$\alpha_1 \sim \alpha_0 \sim \frac{1}{\sqrt{t}} \ . \tag{B.15}$$

Next we obtain the expression for R, using (4.46). R depends on σ and σ is expressed by (4.54). Again, to the lowest order approximation

$$\begin{aligned} \sigma &= |\sigma_1| h^{\nu}\alpha_0(t) \\ &= \frac{2}{nK}h^{s-1}h^{1-s}\alpha_0(t) = \frac{2}{nK}\alpha_0(t) \ . \end{aligned} \tag{B.16}$$

Figure 5: $\frac{1}{\alpha_0(t)}$ vs. t.

To summarize, when $s = 1$ we have, from (4.36), again to the lowest order approximation,

$$R = \frac{nK^2}{4\lambda}\frac{T_M}{\alpha_0(t)} \quad . \tag{B.17}$$

Thus R has a temporal rise like $\sqrt{t}$ for short times, and 'bends-over' to a linear growth. A graph of $\frac{1}{\alpha_0(t)}$ vs. t, using formula (B.9) with $h \to 0$, illustrates the temporal behavior of $R(t)$, see figure 5.

For the case, $s \geq 2$, we get,

$$\begin{aligned} R &= \frac{nK^2}{4\lambda}\frac{T_M h^{1-s}}{\alpha_0(t)} \\ &= \frac{\sqrt{n}K}{\sqrt{2\lambda}} T_M h^{1-s}\sqrt{t} \quad . \end{aligned} \tag{B.18}$$

Appendix C

In this Appendix the scheme used in the diffusion equation example is given. We used the grid presented in Figure 1. with $\gamma_L = \gamma_R = 1$.

The matrix M, see eq. (2.2) and (2.3), which represents the operator $\frac{\partial}{\partial x}\left(a(x)\,\frac{\partial u}{\partial x}\right)$ with a 4th order accuracy in the middle and 3rd order near the boundary is:

$$M = \frac{1}{12\,h^2}\begin{bmatrix}
M_{1,1} & M_{1,2} & M_{1,3} & M_{1,4} & M_{1,5} & M_{1,6} & 0 & & & \\
M_{2,1} & M_{2,2} & M_{2,3} & M_{2,4} & M_{2,5} & M_{2,6} & 0 & & & \\
M_{3,1} & M_{3,2} & M_{3,3} & M_{3,4} & M_{3,5} & 0 & & & & \\
0 & M_{3,2} & M_{3,3} & M_{3,4} & M_{3,5} & M_{3,6} & 0 & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\
 & & & M_{k,k-2} & M_{k,k-1} & M_{k,k} & M_{k,k+1} & M_{k,k+2} & & \\
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & & & & M_{N-2,N-4} & M_{N-2,N-3} & M_{N-2,N-2} & M_{N-2,N-1} & M_{N-2,N} \\
 & & & & M_{N-1,N-5} & M_{N-1,N-4} & M_{N-1,N-3} & M_{N-1,N-2} & M_{N-1,N-1} & M_{N-1,N} \\
 & & & & M_{N,N-5} & M_{N,N-4} & M_{N,N-3} & M_{N,N-2} & M_{N,N-1} & M_{N,N}
\end{bmatrix}, \tag{C.1}$$

where

$$\begin{aligned}
M_{1,1} &= 110\,a(x_1) - 96\,a(x_1+\frac{h}{2}) + 11\,a(x_1+h) + 16\,a(x_1+\frac{3\,h}{2}) - 6\,a(x_1+2\,h) + \mu_{1,1}, \\
M_{1,2} &= -200\,a(x_1) + 16\,a(x_1+\frac{h}{2}) + 200\,a(x_1+h) - 160\,a(x_1+\frac{3\,h}{2}) + 40\,a(x_1+2\,h) + \mu_{1,2}, \\
M_{1,3} &= 125\,a(x_1) + 200\,a(x_1+\frac{h}{2}) - 426\,a(x_1+h) + 280\,a(x_1+\frac{3\,h}{2}) - 65\,a(x_1+2\,h) + \mu_{1,3}, \\
M_{1,4} &= -40\,a(x_1) - 160\,a(x_1+\frac{h}{2}) + 280\,a(x_1+h) - 176\,a(x_1+\frac{3\,h}{2}) + 40\,a(x_1+2\,h) + \mu_{1,4},
\end{aligned}$$

$$
\begin{aligned}
M_{1,5} &= 5\,a(x_1)+40\,a(x_1+\frac{h}{2})-65\,a(x_1+h)+40\,a(x_1+\frac{3\,h}{2})-9\,a(x_1+2\,h)+\mu_{1,5},\\
M_{1,6} &= 0,
\end{aligned}
$$

$$
\begin{aligned}
M_{2,1} &= -6\,a(x_2-h)+16\,a(x_2-\frac{h}{2}),\\
M_{2,2} &= 16\,a(x_2-h)-16\,a(x_2-\frac{h}{2})+a(x_2+h)-16\,a(x_2+\frac{h}{2}),\\
M_{2,3} &= -20\,a(x_2-h)+16\,a(x_2+\frac{h}{2}),\\
M_{2,4} &= 15\,a(x_2-h)-a(x_2+h),\\
M_{2,5} &= -6\,a(x_2-h),\\
M_{2,6} &= a(x_2-h),
\end{aligned}
$$

$$
\begin{aligned}
M_{k,k-2} &= -a(x_k-h),\\
M_{k,k-1} &= 16\,a(x_k-\frac{h}{2}),\\
M_{k,k} &= a(x_k-h)-16\,a(x_k-\frac{h}{2})-16\,a(x_k+\frac{h}{2})+a(x_k+h),\\
M_{k,k+1} &= 16\,a(x_k+\frac{h}{2}),\\
M_{k,k+2} &= -a(x_k+h),
\end{aligned}
$$

$$
\begin{aligned}
M_{N-1,N-5} &= a(x_{N-1}+h),\\
M_{N-1,N-4} &= -6\,a(x_{N-1}+h),\\
M_{N-1,N-3} &= 15\,a(x_{N-1}+h)-a(x_{N-1}-h),\\
M_{N-1,N-2} &= -20\,a(x_{N-1}+h)+16\,a(x_{N-1}-\frac{h}{2}),\\
M_{N-1,N-1} &= 16\,a(x_{N-1}+h)-16\,a(x_{N-1}+\frac{h}{2})-16\,a(x_{N-1}-\frac{h}{2})+a(x_{N-1}-h),
\end{aligned}
$$

$$M_{N-1,N} = -6\,a(x_{N-1}+h) + 16\,a(x_{N-1}+\frac{h}{2}),$$

$$\begin{aligned}
M_{N,N-5} &= 0,\\
M_{N,N-4} &= 5\,a(x_N) + 40\,a(x_N-\frac{h}{2}) - 65\,a(x_N-h) + 40\,a(x_N-\frac{3\,h}{2}) - 9\,a(x_N-2\,h) + \mu_{N,N-4},\\
M_{N,N-3} &= -40\,a(x_N) - 160\,a(x_N-\frac{h}{2}) + 280\,a(x_N-h) - 176\,a(x_N-\frac{3\,h}{2}) + 40\,a(x_N-2\,h) + \mu_{N,N-3},\\
M_{N,N-2} &= 125\,a(x_N) + 200\,a(x_N-\frac{h}{2}) - 426\,a(x_N-h) + 280\,a(x_N-\frac{3\,h}{2}) - 65\,a(x_N-2\,h) + \mu_{N,N-2},\\
M_{N,N-1} &= -200\,a(x_N) + 16\,a(x_N-\frac{h}{2}) + 200\,a(x_N-h) - 160\,a(x_N-\frac{3\,h}{2}) + 40\,a(x_N-2\,h) + \mu_{N,N-1},\\
M_{N,N} &= 110\,a(x_N) - 96\,a(x_N-\frac{h}{2}) + 11\,a(x_N-h) + 16\,a(x_N-\frac{3\,h}{2}) - 6\,a(x_N-2\,h) + \mu_{N,N}.
\end{aligned}$$

The μ's are given by

$$\begin{aligned}
\mu_{1,1} &= 5\,a(x_1)\,\tau_1, & \mu_{N,N} &= 5\,a(x_N)\,\tau_N,\\
\mu_{1,2} &= -10\,a(x_1)\,\tau_1, & \mu_{N,N-1} &= -10\,a(x_N)\,\tau_N,\\
\mu_{1,3} &= 10\,a(x_1)\,\tau_1, & \mu_{N,N-2} &= 10\,a(x_N)\,\tau_N,\\
\mu_{1,4} &= -5\,a(x_1)\,\tau_1, & \mu_{N,N-3} &= -5\,a(x_N)\,\tau_N,\\
\mu_{1,5} &= 1\,a(x_1)\,\tau_1, & \mu_{N,N-4} &= 1\,a(x_N)\,\tau_N
\end{aligned}$$

and

$$\tau_1 = \tau_N = -\frac{890398092}{87407185} \quad .$$

The vector **B**, see eq. (2.2) and (2.3), is :

$$\mathbf{B} = \frac{1}{12\,h^2}\,\left(-\tau_1\,g_L(t), 0, 0, \ldots, 0, -\tau_N\,g_R(t)\right)^T \quad .$$

For the second-order approximation near the boundaries we use the same matrix M but for the first and last rows we take:

$$
\begin{aligned}
M_{1,1} &= 102\,a(x_1) - 124\,a(x_1 + \frac{h}{2}) + 58\,a(x_1 + h) - 12\,a(x_1 + \frac{3\,h}{2}) + \mu_{1,1},\\
M_{1,2} &= -184\,a(x_1) + 192\,a(x_1 + \frac{h}{2}) - 84\,a(x_1 + h) + 16\,a(x_1 + \frac{3\,h}{2}) + \mu_{1,2},\\
M_{1,3} &= 106\,a(x_1) - 84\,a(x_1 + \frac{h}{2}) + 30\,a(x_1 + h) - 4\,a(x_1 + \frac{3\,h}{2}) + \mu_{1,3},\\
M_{1,4} &= -24\,a(x_1) + 16\,a(x_1 + \frac{h}{2}) - 4\,a(x_1 + h) + \mu_{1,4},\\
M_{1,5} &= 0,\\
M_{1,6} &= 0,
\end{aligned}
$$

$$
\begin{aligned}
M_{N,N-5} &= 0,\\
M_{N,N-4} &= 0,\\
M_{N,N-3} &= -24\,a(x_N) + 16\,a(x_N - \frac{h}{2}) - 4\,a(x_N - h) + \mu_{1,4}, + \mu_{N,N-3},\\
M_{N,N-2} &= 106\,a(x_N) - 84\,a(x_N - \frac{h}{2}) + 30\,a(x_N - h) - 4\,a(x_N - \frac{3\,h}{2}) + \mu_{N,N-2},\\
M_{N,N-1} &= -184\,a(x_N) + 192\,a(x_N - \frac{h}{2}) - 84\,a(x_N - h) + 16\,a(x_N - \frac{3\,h}{2}) + \mu_{N,N-1},\\
M_{N,N} &= 102\,a(x_N) - 124\,a(x_N - \frac{h}{2}) + 58\,a(x_N - h) - 12\,a(x_N - \frac{3\,h}{2}) + \mu_{N,N},
\end{aligned}
$$

where

$$
\begin{array}{rclrcl}
\mu_{1,1} &=& 4\,a(x_1)\,\tau_1, & \mu_{N,N} &=& 4\,a(x_N)\,\tau_N,\\
\mu_{1,2} &=& -6\,a(x_1)\,\tau_1, & \mu_{N,N-1} &=& -6\,a(x_N)\,\tau_N,\\
\mu_{1,3} &=& 4\,a(x_1)\,\tau_1, & \mu_{N,N-2} &=& 4\,a(x_N)\,\tau_N,\\
\mu_{1,4} &=& -1\,a(x_1)\,\tau_1, & \mu_{N,N-3} &=& -1\,a(x_N)\,\tau_N,
\end{array}
$$

$$
\tau_1 = \tau_N = -\frac{4214190}{225719} \;\;,
$$

and the vector **B**, as before , is :

$$\mathbf{B} = \frac{1}{12\,h^2}\,\left(-\tau_1\,g_L(t), 0, 0, \ldots, 0, -\tau_N\,g_R(t)\right)^T \quad .$$

For the first-order approximation near the boundaries we take:

$$\begin{aligned}
M_{1,1} &= 66\,a(x_1) - 72\,a(x_1 + \frac{h}{2}) + 18\,a(x_1 + h) + \mu_{1,1}, \\
M_{1,2} &= -96\,a(x_1) + 96\,a(x_1 + \frac{h}{2}) - 24\,a(x_1 + h) + \mu_{1,2}, \\
M_{1,3} &= 30\,a(x_1) - 24\,a(x_1 + \frac{h}{2}) + 6\,a(x_1 + h) + \mu_{1,3}, \\
M_{1,4} &= 0, \\
M_{1,5} &= 0, \\
M_{1,6} &= 0,
\end{aligned}$$

$$\begin{aligned}
M_{N,N-5} &= 0, \\
M_{N,N-4} &= 0, \\
M_{N,N-3} &= 0, \\
M_{N,N-2} &= 66\,a(x_N) - 72\,a(x_N - \frac{h}{2}) + 18\,a(x_N - h) + \mu_{N,N-2}, \\
M_{N,N-1} &= -96\,a(x_N) + 96\,a(x_N - \frac{h}{2}) - 24\,a(x_N - h) + \mu_{N,N-1}, \\
M_{N,N} &= 30\,a(x_N) - 24\,a(x_N - \frac{h}{2}) + 6\,a(x_N - h) + \mu_{N,N},
\end{aligned}$$

where

$$\begin{array}{lcllcl}
\mu_{1,1} &=& 3\,a(x_1)\,\tau_1, & \mu_{N,N} &=& 3\,a(x_N)\,\tau_N, \\
\mu_{1,2} &=& -3\,a(x_1)\,\tau_1, & \mu_{N,N-1} &=& -3\,a(x_N)\,\tau_N, \\
\mu_{1,3} &=& 1\,a(x_1)\,\tau_1, & \mu_{N,N-2} &=& 1\,a(x_N)\,\tau_N,
\end{array}$$

$$\tau_1 = \tau_N = -\frac{2622883}{312498} ,$$

and the vector $\mathbf{B}$, is :

$$\mathbf{B} = \frac{1}{12\,h^2}\,\left(-\tau_1\,g_L(t), 0, 0, \ldots, 0, -\tau_N\,g_R(t)\right)^T \quad .$$

References

[1] S. Abarbanel, A. Ditkowski, Multi-Dimensional Asymptotically Stable 4$^{\text{th}}$-Order Accurate Schemes for the Diffusion Equation. ICASE Report No.96-8, February 1996. Also, Asymptotically Stable Fourth-Order Accurate Schemes for the Diffusion Equation on Complex Shapes. *J. Comput. Phys.*, **133**(2), 1997.

[2] S. Abarbanel, A. Ditkowski, Multi-dimensional asymptotically stable schemes for advection-diffusion equations.ICASE report 47-96. To appear *Computers and Fluids.*

[3] A. Ditkowski, Bounded-Error Finite Difference Schemes for Initial Boundary Value Problems on Complex Domains. Thesis, Department of Applied Mathematics, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel. 1997.

[4] A. Ditkowski, *Fourth-Order Bounded-Error Schemes for the Variable Coefficients Diffusion Equation in Complex Geometries* , In preparation.

[5] A. Ditkowski, K. H. Dridi, and J. S. Hesthaven, *Stable Cartesian Grid Methods for Maxwells Equations in Complex Geometries*, J. Comput. Phys. 1999 – submitted.

[6] B. Gustafsson, H.O. Kreiss, and A. Sundström, Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems. II, *Math. Comp.* **26**, 1972. 649-686.

[7] B. Gustafsson, H.O. Kreiss, and J. Oliger, Time Dependent Problems and Difference Methods. John Wiley & Sons, Inc., 1995.

[8] S.K. Goganov 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**, 1975. 396-406.

[10] B. Gustafsson, The Convergence Rate for Difference Approximations to General Mixed Initial Boundary Value Problems. *SIAM J. Numer. Amal.* **18**, No. 2, 1981. 179-190.

[11] H.O. 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.

[12] H.O. Kreiss. Stability Theory for Difference approximations of Mixed Initial Boundary Value Problem. **I.** *Math. Comp.*, **22**, 703-714, 1968.

[13] H.O. Kreiss, G. Scherer. Finite Element and Finite Difference Methods for Hyperbolic Partial Differential Equations.*Mathematical Aspects of Finite Element in Partial Differential Equations*, Academic Press, Inc., 1974.

[14] H.O. Kreiss, G. Scherer. On the Existence of Energy estimates for Difference Approximations for Hyperbolic Systems. Technical report, Dept. of Scientific Computing, Uppsala University, 1977

[15] P. Olsson, Summation by Parts, Projections and Stability. I. *Math. of Comp.*, **64**(211), 1995. 1035-1065.

[16] P. Olsson, Summation by Parts, Projections and Stability. II. *Math. of Comp.*, **64**(212), 1995. 1473-1493.

[17] S. Osher. Systems of Difference Equations with General Homogeneous Boundary Conditions. *Tran. Amer. Math. Soc.*, **v.137**, 1969, pp. 177-201.

[18] B. Strand, Summation by Parts for Finite Difference Approximations for d/dx. *J. Comput. Phys.*, **110**, 1994. 47-67.

[19] B. Strand, High Order Difference Approximations for Hyperbolic Initial boundary Value Problems. Thesis, Dept of Scientific Computing, Uppsala University, Uppsala, Sweden, 1996

[20] J.C. Strikwerda. Initial Boundary Value Problems for Method of Lines. *J. Comput. Phys.*, **34**, 1980. 94-110.