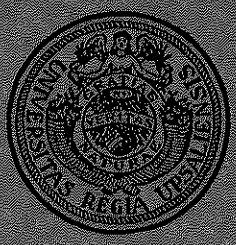
On energy estimates for difference approximations to hyperbolic partial differential equations

by

Gödela Scherer

UPPSALA UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCES



On the existence of energy estimates for difference approximations for hyperbolic systems

by

H-O Kreiss and G Scherer

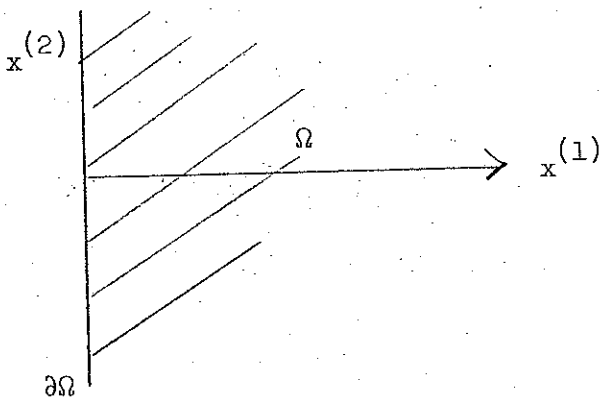
On the existence of energy estimates for difference approximations for hyperbolic systems

by

H.O. Kreiss and G. Scherer

1. Introduction.

Let $x = (x^{(1)},\ldots,x^{(m)})$ denote a point in the real Euclidean space R^m and denote by Ω the half space $x^{(1)} \geq 0$, $-\infty < x^{(i)} < \infty$, $i=2,3,\ldots,m$.



and by $\partial\Omega$ its boundary: $x^{(1)} = 0$, $-\infty < x^{(i)} < \infty$, $i=2,3,\ldots,m$. We consider a symmetric system of partial differential equations

$$\partial u/\partial t = P(x,t,\partial/\partial x)u = \sum_{j=1}^{m} A_j(x,t)\, \partial u/\partial x^{(j)} \quad \text{for} \quad t \geq 0, \tag{1.1}$$

$x \in \Omega$. Here $A_j = A'_j$ are (real) symmetric $n\times n$ matrices which depend Lipschitz continously on x and t and $u = (u^{(1)},u^{(2)},\ldots,u^{(n)})'$ [1] is a real vector function. For $t = 0$ initial values

$$u(x,0) = f(x), \quad x \in \Omega, \quad t = 0 \tag{1.2}$$

and for $x^{(1)} = 0$ boundary conditions

---

[1] if u is a vector then u' denotes its transposed. The corresponding notation holds for matrices.

(1.3) $\quad - Lu(x,t) = 0, \quad t \geq 0, \quad x^{(1)} = 0, \quad -\infty < x^{(i)} < +\infty \ , \ i=2,3,\ldots,m$

are given. Let q denote the number of negative eigenvalues of A_1. Then (1.3) represents q linearly independent relations, i.e. L can be represented by a $(q \times n)$ matrix of rank q:

(1.4) $$\begin{pmatrix} b_{11} & \cdots & b_{1n} \\ & & \\ b_{q1} & \cdots & b_{qn} \end{pmatrix} .$$

Here, the $b_{ij} = b_{ij}(x,t)$ are Lipschitz continuous functions of $x^{(2)},\ldots,x^{(m)}$, t.

Let $\langle u,v \rangle = \sum_{i=1}^{m} u^{(i)} v^{(i)}$, $|u|^2 = \langle u,u \rangle$, u,v real denote the usual vector product and norm of vectors.

For the real functions the L_2 space $L_2(x)$ defined by the scalar product and norm:

$$(u,v) = \int_{\Omega} (u(x,t),v(x,t))dx \ , \quad ||u||_x^2 = (u,u) \quad \text{will be used.}$$

It will be assumed that f and u belong to $L_2(x)$.

The wellposedness of the problem defined by (1.1)-(1.3) is easily obtained if an energy inequality holds.

The differential equation (1.1) can also be written in the form:

$$\partial u/\partial t = \frac{1}{2} \sum_{i=1}^{m} \left[\frac{\partial}{\partial x_j} (A_j u) + A_j \frac{\partial u}{\partial x_j} \right] + B(x,t)u$$

where $B = -\frac{1}{2} \sum_{j=1}^{m} \frac{\partial A}{\partial x_j}$.

The following theorem holds:

Theorem 1.1. The solution of (1.1)-(1.3) satisfy an energy estimate

(1.5) $$\frac{\partial}{\partial t} ||u||^2 \leq 2\alpha ||u||^2 \quad \text{i.e.} \quad ||u(t)|| \leq e^{\alpha t} ||u(0)|| \ ,$$

if the boundary conditions are such that

(1.6) $\qquad (y, A_1 y) \leq 0 \quad$ for every vector v for which Lv=0.

The estimate follows directly by integration by parts.(Compare [4].)

In this paper we want to derive energy estimates for difference approximations to the problem (1.1) - (1.3). In a first step the time derivative is kept and a discretisation introduced in the space direction.
In the next section by defining a difference scheme for the operator d/dx in the quarter plane $0 \leq x < \infty$ a discrete version of the halfboundness inequality [1] in a given norm is designed.
Then the time derivative is approximated by a modified leap frog and the total scheme is proved to be stable in the norm above introduced, by the energy method. It is established that through the outlined method, stable schemes of any order of accuracy can be obtained.
Moreover, given a scheme with a certain accuracy in the interior of the integration domain, one can define a modified scheme at the boundary of one order less, therefore from [1] one can imply that an overall accuracy of the same order as in the interior is attainable.

(1)The operator defined by $P(x,t,\frac{\partial}{\partial x})$ is halfbounded if an inequality $2\,\mathrm{Re}(u,Pu) \leq K \|u\|^2$ holds for all infinitely differentiable u satisfying the boundary conditions (1.3)

2. Halfbounded difference approximations for the operator d/dx

Consider the half-line $0 \leq x < \infty$ and divide it into intervals of length $h > 0$. Let $x_\nu = \nu h$, $\nu = 0,1,2,\ldots;$ denote the gridpoints and $v_\nu = v(x_\nu)$ real scalar gridfunctions with $\sum_{\nu=0}^{\infty} |v_\nu|^2 h < \infty$. We define a discrete scalar product and norm by

$$(u,v)_h = \langle u^I, Hv^I \rangle_h + \sum_{\nu=r}^{\infty} u_\nu v_\nu h =$$

$$= \sum_{i,j=0}^{r-1} h_{ij} u_i u_j h + \sum_{\nu=r}^{\infty} u_\nu u_\nu h, \tag{2.1}$$

and $\quad ||u||_h^2 = (u,u)_h$.

Here $u^I = (u_0,u_1,\ldots,u_{r-1})'$ denotes the vector formed with the first r [1] values of u and $H = H' > 0$ is a positive definite symmetric $r \times r$ matrix. We want to construct difference approximations Q to d/dx such that

$$(u,Qv)_h = -(Qu,v)_h - u_0 v_0 \quad \text{for all } u,v. \tag{2.2}$$

This is a discretisation of the integration by parts formula. An equivalent formulation is given in the following lemma.

Lemma 2.1: The relation (2.2) is equivalent with

$$(u,Qu)_h = -\frac{1}{2} u_0^2 \quad \text{for all } u. \tag{2.3}$$

Proof. $(u,v)_h = (v,u)_h$ shows that (2.2) implies (2.3). If (2.3) holds then

$$(u+v,\, Q(u+v))_h = -(Q(u+v),\, u+v)_h - (u_0+v_0)^2$$

and (2.2) follows easily using (2.3) for u and v.

We can consider u as an infinite column vector $u = (u_0,u_1,\ldots,)'$ and therefore we represent Q as an infinite matrix. We assume that Q has the form

$$hQ = \begin{pmatrix} Q_{11} & Q_{12} \\ -C' & D \end{pmatrix}. \tag{2.4}$$

[1] They correspond to boundary points for which the interior scheme does not apply

Here

$$Q_{11} = \begin{pmatrix} q_{00} & q_{01} \cdots\cdots q_{0\,r-1} \\ & \\ q_{r-1\,0} & \cdots \quad q_{r-1\,r-1} \end{pmatrix} , \quad Q_{12} = \begin{pmatrix} q_{0r} & \cdots & q_{0m} & 0 \cdots \\ & & & \\ q_{r-1\,r} & & q_{r-1\,m} & 0 \cdots \end{pmatrix}$$

$$-C' = \begin{pmatrix} 0 \cdots & -\alpha_s - \alpha_{s-1} & \cdots -\alpha_1 \\ 0 & 0 - \alpha_s & -\alpha_2 \\ \cdots\cdots & & \\ \cdots\cdots & & -\alpha_s \\ 0 & & 0 \\ \cdots\cdots & & \end{pmatrix} \qquad (2.5)$$

and D is the antisymmetric band matrix

$$D = \begin{pmatrix} 0 & \alpha_1 & \cdots & \alpha_s & & \\ -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0 \cdots \\ -\alpha_2 & -\alpha_1 & 0 & \alpha_1 & \cdots & \\ & & & & & \\ \cdots\cdots & & & & & \end{pmatrix}$$

$h^{-1}(-C'\ D)$ represents the operator away from the boundary.

The following approximations of increasing order of accuracy 2s to the derivative are used :

$$(2.6) \quad \frac{\partial}{\partial x} \sim D^{[2s]}(h) \equiv \sum_{\nu=1}^{s} \lambda_\nu D_0(\nu h) , \quad \lambda_\nu = \frac{-2\,(-1)^\nu\,(s!)^2}{(s+\nu)!\ (s-\nu)!} \qquad \nu = 1, 2, \ldots$$

These formulae use the ordinates of 2s+1 symmetric centered points with antisymmetric coefficients and so obtain the highest accuracy possible with this number of equidistant points.

The coefficient of the ν-th right hand term is

$$\alpha_\nu = \frac{-2\,(-1)^\nu\ s!}{\nu\ (s+\nu)!\,(s-\nu)!}$$

The corresponding left hand term is $\alpha_{-\nu} = -\alpha_\nu$

Let h=1 and denote by

$$(2.7)\qquad w_j = \begin{pmatrix} e_j \\ f_j \end{pmatrix} \quad , \quad e_j=(-1)^j \begin{pmatrix} r^j \\ (r-1)^j \\ \vdots \end{pmatrix} \quad , \quad f_j = \begin{pmatrix} 0^j \\ 1^j \\ \vdots \end{pmatrix}, \quad j=0,1,2\ldots$$

the discretisation of $(x-r)^j$.

With the conventions $0^0 = 1$, $e_{-1} = 0$.

The following lemma characterizes the accuracy of Q in the interior (1)

Lemma 2.2. The operator h^{-1} (-C' , D) approximates d/dx of order $2s > 0$ if and only if

$$\sum_{\nu=1}^{s} \alpha_\nu \nu^{2n+1} = \begin{cases} 1/2 & \text{for } n=0 \\ 0 & \text{for } n = 1,2,\ldots s-1 \end{cases} .$$

$h^{-1}(Q_{11} , Q_{12})$ represents the modification of Q at the boundary points.

The accuracy conditions which Q has to fulfill are:

Lemma 2.3 . The operator h^{-1} (Q_{11} , Q_{12}) approximates d/dx with order τ at the points x_ν , $\nu = 0,1,2,\ldots r-1$ if and only if

$$(2.8)\qquad j\, e_{j-1} = Q_{11}\, e_j + Q_{12}\, f_j \quad , \quad j=0,1,2\ldots \tau.$$

We shall now derive necessary and sufficient conditions such that the operator Q satisfies the condition (2.3).

Theorem 2.1 . The operator Q satisfies the relation (2.3) if and only if it can be written as

$$h\,Q = \begin{pmatrix} H^{-1} B & H^{-1} C \\ -C' & D \end{pmatrix}$$

where B is a rxr matrix of the form

$$(2.9)\qquad B = B_1 + B_2 \qquad \text{with}$$

(1) An operator S is an accurate approximation to d/dx of order s at the point x <=> $S\,x^m - \frac{d}{dx} x^m = 0$, m=0,1..s.

$$B_1 = \begin{pmatrix} -\frac{1}{2} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 & b_{01} & \cdots & b_{0r-1} \\ -b_{01} & 0 & b_{12} & \cdots & b_{1r-1} \\ -b_{02} & -b_{12} & 0 & b_{23} & \cdots & b_{2r-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -b_{0r-1} & \cdots & -b_{r-2r-1} & & 0 \end{pmatrix} = -B_2'$$

Proof. We can write (2.3) in the form

$$-\frac{1}{2} u_0^2 = \langle u^I, HQ_{11} u^I \rangle + \langle u^I, HQ_{12} u^{II} \rangle - \langle u^{II}, C'u^I \rangle + \langle u^{II}, Du^{II} \rangle$$

where $u^I = (u_0, \ldots, u_{r-1})'$, $u^{II} = (u_r, u_{r+1}, \ldots)'$.

D is an antisymmetric matrix and therefore $\langle u^{II}, Du^{II} \rangle \equiv 0$. Furthermore, if $u^{II} \equiv 0$ then the above relation becomes

$$-\frac{1}{2} u_0^2 = \langle u^I, HQ_{11} u^I \rangle$$

Therefore $HQ_{11} = B$ must have the form (2.9). Then the above relation is equivalent with

$$0 = \langle u^I, HQ_{12} u^{II} \rangle - \langle u^{II}, C'u^I \rangle = \langle u^I, (HQ_{12} - C)u^{II} \rangle$$

for all vectors u^I, u^{II}. This is only possible if $Q_{12} = H^{-1}C$ and the theorem is proved.

We want to calculate now the coefficients of Q at the boundary points. The elements of B_2 are obtained from (2.8) which can be written as

(2.10) $\quad B_2 e_j = g_j$, $g_j = jHe_{j-1} - B_1 e_j - Cf_j$, $j = 0,1,2,\ldots,\tau$.

By assumption B_2 is antisymmetric. Therefore the following compatibility conditions for the system have to be satisfied

(2.11) $\quad \langle e_i, g_j \rangle + \langle e_j, g_i \rangle = \langle e_i, B_2 e_j \rangle + \langle e_j, B_2 e_i \rangle = 0 \quad$ for $0 \leq i,j \leq \tau$

If these conditions hold the system (2.10) can be resolved as it is expressed in

Lemma 2.4 . Assume that $r \geq \sigma + 1$ and that the relations (2.11) hold. Then we can always find an antisymmetric matrix B_2 such that (2.10) is valid.

We will obtain the elements of the norm matrix H from the relations (2.11) By (2.10) this can also be written as

$$(2.12) \qquad j \langle e_i , H e_{j-1} \rangle + i \langle e_j , H e_{i-1} \rangle = M_{i,j} \quad , \quad 0 \leq i,j \leq \sigma$$

where

$$M_{i,j} = 2 \langle e_i , B_1 e_j \rangle + \langle e_i , C f_j \rangle + \langle e_j , C f_i \rangle \ .$$

A calculation performed in [1] gives

$$(2.13) \qquad M_{i,j} = -(-1)^{i+j} r^{i+j} + \gamma_{i,i+j} \ , \quad \gamma_{i,\sigma} = \sum_{\nu=1}^{s} \alpha_\nu \Big(\sum_{\mu=0}^{\nu-1} \mu^{\sigma-i} (\mu-\nu)^i + \mu^i (\mu-\nu)^{\sigma-i} \Big), \quad \sigma \geq i+j$$

If one introduces the notation $\rho_{i,j} = \langle e_i , H e_j \rangle$ then (2.12) can be written

$$(2.14) \qquad j \rho_{i,j-1} + i \rho_{j,i-1} = M_{i,j} \qquad \text{for } 0 \leq i,j \leq \sigma$$

Here $\rho_{i,-1} = \rho_{-1,i} \equiv 0$ by the convention for e_{-1} and $\rho_{i,j} = \rho_{j,i}$ by the symmetry of H. Note that this system in the $\rho_{i,j}$ is very simple to resolve. We will resolve it and obtain from the solution the elements of H. The matrix H has to be positive definite in order to be used as norm matrix . Lemma 2.9 below states the equivalence between this condition and the positive definiteness of the matrix defined by the $\rho_{i,j}$.

We will now derive conditions for $M_{i,j}$ such that the system (2.14) has a solution (These conditions are resumed in Lemma 2.5) and therefore for the compatibility of (2.10).

For i = j = 0 we obtain from (2.14) and (2.13)

$$(2.15) \qquad 0 = 0 \cdot \rho_{0,-1} + 0 \cdot \rho_{0,-1} = M_{0,0} = -1 + 2 \sum_{\nu=1}^{s} \alpha_\nu \nu \quad \text{ie} \quad \sum_{\nu=1}^{s} \alpha_\nu \nu = \frac{1}{2}$$

Therefore, by Lemma 2.2 the approximation at the interior points has to be at least second order accurate.

$\rho_{i,j} = \rho_{j,i}$ implies $M_{i,j} = j \rho_{i,j-1} + i \rho_{j,i-1} = M_{j,i}$

This condition is satisfied by (2.13). Therefore we need to consider (2.14) only for $i \leq j$.

For $i = 0$ we get from (2.14)

$$\rho_{0,j-1} = \frac{1}{j} M_{0,j} \qquad j=1,2\ldots .$$

If $i > 0$ then $\rho_{i,j-1}$ can be explicitly calculated.

$$(2.14') \qquad \rho_{i,j-1} = \frac{1}{j} M_{i,j} - \frac{i}{j} \rho_{i-1,j} .$$

If $i-1 > 0$ and $j < \zeta$ then (2.14') can be used to replace $\rho_{i-1,j}$ by $\rho_{i-2,j+1}$ and obtain $\rho_{i,j-1} = \frac{1}{j} M_{i,j} - \frac{i}{j(j+1)} M_{i-1,j+1} + \frac{i(i-1)}{j(j+1)} \rho_{i-2,j+1}$

Threfore by recursion (2.14) can thenbe written in the form

$$\rho_{i,j} = \rho_{j,i} \quad , \quad \rho_{i,-1} = \rho_{-1,i} = 0$$

$$(2.16) \qquad \rho_{i,j-1} = \frac{1}{j} M_{i,j} - \frac{i}{j(j+1)} M_{i-1,j+1} + \cdots + (-1)^{\alpha} \frac{(i-1)\ldots(i-\alpha+1)}{j(j+1)\ldots(j+\alpha)} M_{i-\alpha,j+\alpha}$$

$$+ (-1)^{\alpha+1} \frac{i(i-1)\ldots(i-\alpha)}{j(j+1)\ldots(j+\alpha)} \rho_{i-\alpha-1,j+\alpha} \qquad 0 < i \le j \le \zeta$$

where $\alpha = \min(i-1, \zeta - j)$.

If $i-1 < \zeta - j$, ie $i+j \le \zeta$, then $\alpha = i-1$ and we obtain

$$(2.17) \qquad \rho_{i-\alpha-1,j+\alpha} = \rho_{0,i+j-1} = \frac{1}{i+j} M_{0,i+j}$$

Thus (2.15) and (2.16) imply that $\rho_{i,j-1}$ is compltely determined by the $M_{i,j}$ provided $i + j \le \zeta$.

If $i + j > \zeta$ then $\alpha = \zeta - j$ and

$$(2.18) \qquad \rho_{i-\alpha-1,j+\alpha} = \rho_{i+j-1-\zeta,\zeta} = \rho_{\nu,\zeta} \quad , \quad \nu = i+j-1-\zeta$$

There are no further relations which $\rho_{\nu,\zeta}$ need to satisfy.

If we use the equation (2.16) with $i = n$, $j = n$ and $i = n-1$, $j = n+1$, for $n < \zeta$ we obtain representations for $\rho_{n,n-1}$ and $\rho_{n-1,n}$ but by (2.13) $\rho_{n,n-1} = \rho_{n-1,n}$
If $i + j = 2n > \zeta$ then by (2.18)' these two relations determine $\rho_{\nu,\zeta}$, $\nu = 2n-1-\zeta$ and no conditions for the $M_{i,j}$ result.
If $i + j = 2n \le \zeta$ we obtain from (2.16) , (2.17)

$$\frac{1}{n} M_{n,n} - \frac{n}{n(n+1)} M_{n-1,n+1} + \frac{n(n-1)}{n(n+1)(n+2)} M_{n-2,n+2} + \cdots$$

$$+ (-1)^n \frac{n(n-1) \cdots 1}{n(n+1) \cdots 2n} M_{0,2n} = \rho_{n,n-1} = \rho_{n-1,n} = \frac{1}{n+1} M_{n-1,n+1} -$$

$$\frac{(n-1)}{(n+1)(n+2)} M_{n-2,n+2} + \cdots + (-1)^{n-1} \frac{(n-1) \cdots 1}{(n+1)(n+2) \cdots 2n} M_{0,2n} .$$

This relation can also be written as

$$(2.19) \quad \frac{1}{2n} M_{n,n} - \frac{1}{n+1} M_{n-1,n+1} + \frac{(n-1)}{(n+1)(n+2)} M_{n-2,n+2} - \cdots +$$

$$+ (-1)^n \frac{(n-1) \cdots 1}{(n+2) \cdots 2n} M_{0,2n} = 0 \qquad \text{for } n=1,2,.. \text{ with } 2n \leq \tau$$

We obtain

Lemma 2.5 . The system (2.14) has a solution if and only if the difference approximation in the interior is at least second order accurate and if the $M_{i,j}$ satisfy the relations (2.19).

We have also a special result for the diagonal norm schemes discussed in [2].

Lemma 2.6. If the $M_{i,j}$ depend only on $i + j$, i.e.

$$M_{i,j} = R_{i+j} \quad \text{and} \quad R_0 = 0$$

then the system (2.14) has the solution

$$\rho_{i,-1} = \rho_{-1,j} = 0, \quad \rho_{i,j} = \frac{R_{i+j+1}}{i+j+1} , \quad 0 \leq i,j \leq \tau, \; i + j < 2\tau.$$

Proof. Introduce the above expression for $\rho_{i,j}$ into (2.14).

A consequence of this lemma is that if

$$M_{n,n} = M_{n-1,n+1} = \cdots = M_{0,2n} = R_{2n}$$

then the relation (2.19) must hold i.e.

$$(2.20) \quad \frac{1}{2n} - \frac{1}{n+1} + \frac{n-1}{(n+1)(n+2)} - \cdots + (-1)^n \frac{(n-1) \cdots 1}{(n+1) \cdots 2n} = 0 \qquad n = 1,2,.. 2n \leq \tau$$

The $M_{i,j}$ are expressed in terms of $J_{i,c}$ in formula (2.13). These in turn in [1, pp 205-206] were written in function of sums of type $\sum_{\nu=1}^{s} \alpha_\nu \nu^m$. (Compare accuracy characterisation lemma 2.2).

Therefore we will obtain the sought conditions of compatibility for the system (2.14) in terms of accuracy conditions which the chosen scheme in the interior has to fulfill.

In fact, by introducing (2.13) into (2.19) one obtains:

$$\frac{1}{2n} J_{n,2n} - \frac{1}{n+1} J_{n-1,2n} + \frac{n-1}{(n+1)(n+2)} J_{n-2,2n} \cdots (-1)^n \frac{(n-1)\cdots 1}{(n+2)\cdots 2n} J_{0,2n}$$

$$- (-1)^{2n} r^{2n} \left[\frac{1}{2n} - \frac{1}{n+1} + \cdots (-1)^n \frac{(n-1)\cdots 1}{(n+2)\cdots 2n} \right] = 0$$

The last term is equal to zero by (2.20).

From [1, formula (2.20) for c even] one obtains :

$$(2.21) \qquad J_{i,2n} = \sum_{\nu=1}^{s} \alpha_\nu N_{i,2n}(\nu) .$$

where
$$N_{i,2n}(\nu) = \sum_{\mu=0}^{\nu-1} \mu^{2n-i} (\mu-\nu)^i + \mu^i (\mu-\nu)^{2n-i}$$

$$= \gamma_1^{(i,2n)} \nu^{2n+1} + \gamma_3^{(i,2n)} \nu^{2n-1} + \cdots \gamma_{2n+1}^{(i,2n)} \nu$$

$$(2.22) \quad \text{ie} \quad J_{i,2n} = \gamma_1^{(i,2n)} \sum_{\nu=1}^{s} \alpha_\nu \nu^{2n+1} + \cdots \gamma_{2n+1}^{(i,2n)} \sum_{\nu=1}^{s} \alpha_\nu \nu$$

The $\gamma_{2n+1}^{(i,2n)}$ is independent of i (this is proved by expanding the $N_{i,2n}(\nu)$). Therefore using again (2.20) it also drops out and the relations (2.19) are reduced to

$$\beta_{n,0} \sum_{\nu=1}^{s} \alpha_\nu \nu^{2n+1} + \beta_{n,2} \sum_{\nu=1}^{s} \alpha_\nu \nu^{2n-1} + \cdots \beta_{n,2n-2} \sum_{\nu=1}^{s} \alpha_\nu \nu^3$$

where $\beta_{n,0} = \frac{\gamma_1^{(n,2n)}}{2n} - \frac{\gamma_1^{(n-1,2n)}}{n-1} + \cdots \neq 0$ because the $\gamma_1^{(i,2n)}$ are different from zero and have alternate signs. (see [1]).

Then induction gives us

$$\sum_{\nu=1}^{s} \alpha_\nu \nu^{2n+1} = 0 \quad , \quad \text{for} \quad n = 1, 2, \ldots ; \quad 2n \leq c$$

Therefore we obtain from lemma 2.5 and 2.2

Lemma 2.7 . The system (2.14) has a solution if and only if

$$\sum_{\nu=1}^{s} \alpha_\nu \nu^{2n+1} = \begin{cases} 1/2 & \text{for } n = 0 \\ 0 & \text{for } n = 1, 2, \ldots ; \ 2n \leq c \end{cases}$$

ie if and only if the approximation in the interior is accurate of order $c+1$ if c is odd.

If the system has a solution then the $\rho_{i,j}$ are determined for $0 \le i,j \le \zeta$ if one specifies those $\rho_{\nu,\zeta}$, $\nu = 0,1,\dots \zeta$ which are not defined by the system They can be used to define a symmetric matrix

$$P_\zeta = \begin{pmatrix} \rho_{0,0} & \rho_{0,1} & \dots & \rho_{0,\zeta} \\ \rho_{0,1} & \rho_{1,1} & \dots & \rho_{1,\zeta} \\ \vdots & & & \\ \rho_{0,\zeta} & \rho_{1,\zeta} & \dots & \rho_{\zeta,\zeta} \end{pmatrix} = P'_\zeta$$

The parameters $\rho_{\nu,\zeta}$ should be chosen so as to obtain a positive definite P_ζ . The following lemma can be used

Lemma 2.8 . If

$$P_{\zeta-1} = \begin{pmatrix} \rho_{0,0} & \rho_{0,1} & \dots & \rho_{0,\zeta-1} \\ \vdots & & & \\ \rho_{0,\zeta-1} & \dots & & \rho_{\zeta-1,\zeta-1} \end{pmatrix} = P'_{\zeta-1}$$

is positive definite then one can choose $\rho_{\zeta,\zeta}$ such that also $P_\zeta > 0$ independently of the values of $\rho_{\nu,\zeta}$, $\nu = 0,1,\dots \zeta-1$

Proof The proof is obtained by developing the determinant of P_ζ by the elements of the last row and balancing with the value of $\rho_{\zeta,\zeta}$. Notice that the system (2.14) does not depend on $\rho_{\zeta,\zeta}$.

Now a positive definite rxr matrix H has to be determined such that

(2.23) $\langle e_i, H e_j \rangle = \rho_{i,j} \qquad 0 \le i,j \le \zeta$.

In fact the following lemma holds:

Lemma 2.9 . If $r \ge \zeta + 1$ and the matrix P_ζ is positive definite then thre are $H = H' > 0$ such that (2.23) holds.
In particular if $r = \zeta + 1$ then H is uniquely defined by

$$E' H E = P_\zeta \ , \ E = (e_0, \dots . e_{r-1}).$$

In the following it will be essential that the matrix has the form

(2.24) $$H = \begin{pmatrix} \lambda_0 & 0 & \dots & 0 \\ 0 & h_{11} & \dots & h_{1,r-1} \\ \vdots & \vdots & & \vdots \\ 0 & h_{1,r-1} & \dots & h_{r-1,r-1} \end{pmatrix} = \begin{pmatrix} \lambda_0 & 0 \\ 0 & H_1 \end{pmatrix}$$

We have

Lemma 2.10. If P_ζ is positive definite then one can choose H in the form (2.24) such that (2.23) holds. In general one has to take $r \ge \zeta + 2$.

Proof Let $e_j = \tilde{e}_j + \tilde{\tilde{e}}_j$ where

$$\tilde{e}_j = (-1)^j \begin{pmatrix} 0 \\ (r-1)^j \\ \vdots \\ 1 \end{pmatrix}, \quad \tilde{\tilde{e}}_j = (-1)^j \begin{pmatrix} r^j \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

(2.23) is equivalent with

$$(2.25) \quad \tilde{\rho}_{ij} = \langle \tilde{e}_i, H_1 \tilde{e}_j \rangle = \rho_{ij} - \lambda_0 (-1)^{i+j} r^{i+j}$$

If P_τ is positive definite then the matrix $\tilde{P}_\tau$ formed by the $\tilde{\rho}_{ij}$ is also positive definite provided we choose λ_0 sufficiently small.

We can now prove the main result of this section.

Theorem 2.2 For every $\zeta+1 = 2s$, $s = 1,2,\ldots$ there is an H of the form (2.24) and a scalar product of the form (2.1) and an approximation Q of d/dx which is accurate of order τ for $x = x_\nu$, $\nu = 0,1,2,\ldots,r-1$ and accurate of order $\zeta+1$ for $x = x_\nu$, $\nu \geq r$ such that (2.2) holds.

Proof. We need only to show that we can make the matrix $P_\tau > 0$ if we choose r sufficiently large. The ρ_{ij} are the solutions of the system (2.14). We can split $\rho_{ij} = \tilde{\rho}_{ij} + \tilde{\tilde{\rho}}_{ij}$ into two parts. $\tilde{\rho}_{ij}$ is the solution of

$$(2.26) \quad j\,\tilde{\rho}_{i,j-1} + i\,\tilde{\rho}_{j,i-1} = -(-1)^{i+j} r^{i+j} \quad 0 \leq i,j \leq \tau, \quad i+j > 0$$

$$\tilde{\rho}_{i,-1} = \tilde{\rho}_{-1,j} = 0, \quad \tilde{\rho}_{i,j} = \tilde{\rho}_{ji}$$

and

$$(2.27) \quad j\tilde{\tilde{\rho}}_{i,j-1} + i\tilde{\tilde{\rho}}_{j,i-1} = J_{i,i+j}, \quad 0 \leq i,j \leq \tau, \; i+j > 0,$$

$$\tilde{\tilde{\rho}}_{i,-1} = \tilde{\tilde{\rho}}_{-1,j} = 0, \quad \tilde{\tilde{\rho}}_{i,j} = \tilde{\tilde{\rho}}_{ji}.$$

By Lemma 2.6 the system (2.26) has a solution

$$\tilde{\rho}_{i,j} = \frac{(-1)^{i+j} r^{i+j+1}}{i + j + 1} \quad , \quad 0 \leq i,j \leq \tau \ .$$

The matrix formed by the $\tilde{\rho}_{i,j}$ can be written as

$$\begin{pmatrix} \frac{r}{1} & -\frac{r^2}{2} & \cdots\cdots\cdots & (-1)^{\tau} \frac{r^{\tau+1}}{\tau+1} \\ \frac{-r^2}{2} & \frac{r^3}{3} & \cdots\cdots\cdots & (-1)^{\tau+1} \frac{r^{\tau+2}}{\tau+2} \\ & & \cdots\cdots\cdots & \\ (-1)^{\tau} \frac{r^{\tau+1}}{\tau+1} & & \cdots\cdots\cdots & \frac{r^{2\tau+1}}{2\tau+1} \end{pmatrix} = D\,G\,D$$

where

$$D = r^{1/2} \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & -r & 0 & \cdots & 0 \\ 0 & 0 & r^2 & \ddots & \\ & & & & \\ 0 & \cdots & \cdots & 0 & (-r)^{\tau} \end{pmatrix}$$

and

$$G = \begin{pmatrix} 1 & \frac{1}{2} & \cdots\cdots & \frac{1}{\tau+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots\cdots & \frac{1}{\tau+2} \\ & & \cdots\cdots & \\ \frac{1}{\tau+1} & \frac{1}{\tau+2} & \cdots\cdots & \frac{1}{2\tau+1} \end{pmatrix}$$

is the well known Hilbert matrix which is positive definite. Also, the eigenvalues of DGD are of the order r and therefore the matrix DGD will dominate the matrix formed by the solution of $\tilde{\rho}_{ij}$ of the system (2.27). Thus, for r sufficiently large, the matrix $P_\tau > 0$

This proves the theorem.

We consider now two examples.

Example 1. The case $\tau = 1$ has already been considered in [2].

We choose

$$hQ = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

with the scalar product

$$(u,v)_h = \frac{1}{2} u_0 v_0 h + \sum_{\nu=1}^{\infty} u_\nu v_\nu h$$

i.e. $r = 1$ and $H = 1/2$.

It is a diagonal norm scheme.

Example 2. For $\tau = 3$ the approximation must be fourth order accurate in the interior. We choose $s = 2$, $\alpha_1 = 2/3$, $\alpha_2 = -1/12$. A simple calculation shows that the ρ_{ij} with $0 \le i \le j$ are given by

$$\rho_{00} = M_{01}, \quad \rho_{01} = \frac{1}{2} M_{02}, \quad \rho_{02} = \frac{1}{3} M_{03}, \quad \rho_{03} = M_{13} - \frac{3}{4} M_{22}$$

$$\rho_{11} = \frac{1}{2} M_{12} - \frac{1}{6} M_{03}, \quad \rho_{12} = \frac{1}{4} M_{22}, \quad \rho_{13} = \frac{1}{2} M_{23} - \frac{3}{2} \rho_{22}$$

$$\rho_{23} = \frac{1}{6} M_{33} .$$

No further relations have to be satisfied. Therefore we can choose ρ_{22}, ρ_{33} arbitrary and can make the matrix $P_\tau > 0$ provided

$$\begin{pmatrix} \rho_{00} & \rho_{01} \\ \rho_{01} & \rho_{11} \end{pmatrix} = \begin{pmatrix} M_{01} & \frac{1}{2} M_{02} \\ \frac{1}{2} M_{02} & \frac{1}{2} M_{12} - \frac{1}{6} M_{03} \end{pmatrix} \quad > 0.$$

From (2.13) we obtain

$$\begin{pmatrix} M_{00} & M_{01} & M_{02} & M_{03} \\ & M_{11} & M_{12} & M_{13} \\ & & M_{22} & M_{23} \\ & & & M_{33} \end{pmatrix} = \begin{pmatrix} 0 & r-\frac{1}{2} & r^2+\frac{1}{6} & r^3 \\ & -r^2+\frac{1}{6} & r^3 & -r^4+\frac{1}{6} \\ & & -r^4-\frac{1}{6} & r^5 \\ & & & -r^6+\frac{1}{6} \end{pmatrix}.$$

Therefore

$$\begin{pmatrix} M_{01} & \frac{1}{2} M_{02} \\ \frac{1}{2} M_{02} & \frac{1}{2} M_{12} - \frac{1}{6} M_{03} \end{pmatrix} = \begin{pmatrix} r - \frac{1}{2} & \frac{-r^2}{2} + \frac{1}{12} \\ \frac{-r^2}{2} + \frac{1}{12} & \frac{1}{3} r^3 \end{pmatrix}$$

which is positive definite for $r \geq 2$.

We choose H in the form (2.24) and set $r = \tau+2$. Then for every choice of λ_0, ρ_{22}, ρ_{33} H is uniquely determined. (The only restriction on λ_0, ρ_{22}, ρ_{33} is that $\tilde{P}_\tau$ is positive definite).

After having determined H one can obtain B_2 from the equations (2.10). B_2 is a $(r+1)\times(r+1)$ matrix and there are only 4 vector relations connecting e_0, e_1, e_2, e_3 with g_0, g_1, g_2, g_3. Let e_4 be the unit vector $e_4 = (0,0,0,0,1)'$. One can define

$$g_4 = B_2 e_4$$

from the compatability conditions.

$$\langle e_4, g_4 \rangle = 0, \quad \langle e_j, g_4 \rangle = -\langle e_4, g_j \rangle \quad j = 0,1,2,3.$$

Then

$$B_2 = GE^{-1}, \quad G = (g_0,\ldots,g_4), \quad E = (e_0,\ldots,e_4).$$

Thus the scalar product and the approximation depends on the three parameters, λ_0, ρ_{22}, ρ_{33}.

3. Stability of difference approximations.

In this section we consider difference approximations for the problem (1.1)-(1.3). To begin with assume that the boundary conditions have the special form

$$(3.1) \qquad u^I(x,t) = (u^{(1)}(x,t),\ldots,u^{(q)}(x,t))' = 0,\ x \in \partial\Omega\ ,\ t \geq 0.$$

We introduce gridpoints $x_\nu = x_{\nu_1,\ldots,\nu_s} = (\nu_1 h,\ldots,\nu_s h) \in \Omega$, and while keeping the time derivative, approximate (1.1) inthe interior of Ω by

$$(3.2) \qquad dv(x,t)/dt = \sum_{\ell=1}^{m} A_\ell(x,t)\ Q_\ell v(x,t),\ x \in \Omega\ ,x^{(1)} > 0,\ t \geq 0.$$

Here x represents gridpoints and Q_ℓ are consistent difference approximations for $\partial/\partial x_\ell$. The operator Q_1 is of the form discussed in the preceding section and $Q_2,\ldots,Q_m$ are centered approximations of the form (2.6)

Assume that the v satisfies the boundary conditions (3.1)

$$(3.4) \qquad v^I(x,t) = 0, \quad x \in \partial\Omega,\ t \geq 0$$

and that the other components $v^{II}(x,t) = (v^{(q+1)}(x,t),\ldots,v^{(n)}(x,t))'$ satisfy (3.2), i.e.

$$(3.5) \qquad dv^{II}(x,t)/dt = \sum_{\ell=1}^{m} (A_\ell(x,t)\ Q_\ell v(x,t))^{II},\ x \in \partial\Omega\ ,\ t \geq 0.$$

The initial conditions

$$(3.6) \qquad v(x,0) = f(x),\ \ x \in \Omega\ ,\ t = 0\ ,\ \text{have to be compatible with (3.4)}$$

Now a discrete scalar Product for these gridfunctions $v(x_\nu, t)$ hasto be defined scalar product $(u,v)_h$ of type (2.1) in the first variable x_1.

Let $\phi(x_)$ denote a function of $x_ = (x^{(2)},\ldots,x^{(m)})$.

An operator S by is introduced

$$(3.7)\qquad S\phi(x_) = \sum_{\nu_2=-\infty}^{+\infty} \cdots\cdots \sum_{\nu_m=-\infty}^{+\infty} \phi(\nu_2 h,\ldots,\nu_m h)\, h^{m-1}$$

Let $u = u_j(x_)$, $v = v_j(x_)$ be scalar functions of $x = (jh,x_)$. Then we can compute $(u,v)_h$ for every fixed $x_$ and we define the m-dimensional scalar product and norm by

$$(3.8)\qquad ((u,v))_h = S(u,v)_h,\quad |||u|||_h^2 = ((u,u))_h.$$

For $u = (u^{(1)},\ldots,u^{(n)})'$, $v = (v^{(1)},\ldots,v^{(n)})'$ vector functions, the scalar product $((u,v))_h$ is defined in the natural way by

$$(3.9)\qquad ((u,v))_h = \sum_{i=1}^{n} ((u^{(i)},v^{(i)}))_h = S\Big(\sum_{i,j=0}^{r-1} h_{ij} < u_i(x_), v_j(x_) > h + \sum_{j=r}^{\infty} < u_j(x_), v_j(x_) > h\Big).$$

It is clear that $((u,v))_h$ is a bilinear form and defines a norm. Therefore the usual inequalities

$$(3.10a)\qquad |((u,v))_h| \le |||u|||_h\, |||v|||_h,\quad |||u+v|||_h \le |||u|||_h + |||v|||_h$$

are valid. Also

$$(3.10b)\qquad ((u,Av))_h = ((A'u,v))_h \text{ if } A \text{ is a matrix.}$$

Next, an estimate for $\frac{d}{dt}|||v|||_h^2$ will be derived. For this we need the following:

Lemma 3.1. Let $u = u_j(x_)$, $v = v_j(x_)$ be vector functions of $x = (jh,x_)$. Assume that $|||u|||_h < \infty$, $|||v|||_h < \infty$.

Let $A(x,t)$ be a uniformly Lipschitz continuous nxn matrix then

$$(3.10)\qquad ((u,Q_\ell Av))_h = ((u,AQ_\ell v))_h + C_\ell[u,v],\quad \ell = 1,2,\ldots,m,$$

where

$$|C_\ell[u,v]| \le \beta_\ell\, |||u|||_h\, |||v|||_h.$$

and β_ℓ is a constant which does not depend on u, v nor h.

Furthermore

$$(3.11) \qquad ((u,Q_\ell v))_h = -((Q_\ell u,v))_h \quad \text{for } \ell=2,3,\dots,m$$

and

$$(3.12) \qquad ((u,Q_1 v))_h = -((Q_1 u,v))_h - S<u_0(x_-), v_0(x_-)>$$

provided Q_1 and H are such that (2.2) holds.

Proof. (3.1o) is derived from the fact that for a scheme $Q_\ell = \sum_{j=-p}^{q} \gamma_j E_\ell^j$

$Q_\ell A(x)v(x) = A(x)Q_\ell v(x) + \sum_{j=-p}^{q} B_j(x) v(x+jhe_\ell)$ where e_ℓ is the unit vector in the x_ℓ direction and $E_\ell^j v(x) = v(x+jhe_\ell)$. Here $B_j(x)$ are uniformly bounded matrices (the bound is depending on the Lipschitz constant)

For $\ell \geq 2$

$$((u,Q_\ell v))_h = \sum_{i,j=0}^{r-1} h_{ij} S<u_i(x_-), Q_\ell v_j(x_-)>_h + \sum_{j=r}^{\infty} S<u_j(x_-), Q_\ell v_j(x_-)>_h$$

For difference operators of the form (2.6)

$$S<u_i(x_-), Q_\ell v_j(x_-)> = -S<Q_\ell u_i(x_-), v_j(x_-)>$$

and (3.11) follows.

(3.12) is obtained by applying the operator S to the identity (2.2) which is valid for all components $u^{(i)}$, $v^{(i)}$, and sum.

We can now prove

Theorem 3.1. Assume that a onedimensional scalar product (2.1) is given, that the matrix H has the form (2.24) and that Q_1, H are such that (2.2) is valid. Then the solutions of the approximation (3.2)-(3.6) satisfy an energy estimate

$$(3.13) \qquad \frac{d}{dt} |||v|||_h^2 \leq \text{const.} \; |||v|||_h^2$$

if the boundary conditions satisfy

(3.14) $\langle v, A_1 v \rangle \leq 0$ for all vectors v with $v^T = 0$.

Proof. Let P denote the projection

$$Pu(x) = \begin{cases} u(x) & \text{if } x^{(1)} > 0 \\ \begin{pmatrix} 0 \\ u^{II}(x) \end{pmatrix} & \text{if } x^{(1)} = 0 \end{cases}.$$

Then the problem can be rewritten

$$(3.15) \qquad dv/dt = Rv, \quad v = Pv, \quad x \in \Omega, \ t \geq 0,$$
$$v(x,0) = Pf(x) = f(x),$$

where

$$R = P \sum_{\ell=1}^{m} A_\ell(x,t)\, Q_\ell .$$

The functions v fulfill the boundary conditions now and

$$\frac{d}{dt} |||v|||_h^2 = ((v,dv/dt))_h + ((dv/dt,v))_h =$$

$$= \sum_{\ell=1}^{m} ((v,PA_\ell Q_\ell v))_h + ((PA_\ell Q_\ell v,v))_h .$$

By assumption the scalar product is of the form

$$((u,v))_h = \lambda_0 \langle u_0, v_0 \rangle_h + \text{terms which are independent of } u_0, v_0.$$

Therefore

$$((u,Pv))_h = ((Pu,v))_h \quad \text{and} \quad ((Pu,Pv))_h = ((Pu,v))_h .$$

Thus by lemma (3.1) and (3.10b)

$$((v,PA_\ell Q_\ell v))_h + ((PA_\ell Q_\ell v,v))_h =$$

$$((v,PA_\ell Q_\ell Pv))_h + ((PA_\ell Q_\ell Pv,v))_h =$$

$$((v,P(A_\ell Q_\ell - Q_\ell A_\ell)Pv))_h + B_\ell \leq$$

$$\leq \text{const.} \, |||v|||_h^2 + B_\ell ,$$

where B_ℓ appears from the fact that Q_ℓ is not antisymmetric in x_1.

$$B_\ell = \begin{cases} 0 & \text{if } \ell \geq 2 \\ S\,\langle Pv_0(x_-,t),\ A_1(0,x_-,t)\,Pv_0(x_-,t)\rangle \leq 0 & \text{if } \ell = 1 \end{cases}$$

For the time discretisation an analogous theorem to the one proved in [3] is used. The proof for the present case follows along the same lines.

Theorem 3.2. Consider a two-step method

$$(3.16)\qquad (I-kR_0)v(t+k) = (I+kR_0)v(t-k) + 2k\,R_1 v(t).$$

It is stable if for all gridfunctions u which satisfy the boundary conditions

$$((u,R_0u))_h \leq \text{const.}\ |||u|||_h^2\ ,\ ((u,R_1u))_h = 0,\quad h|||R_1|||_h \leq 1-\eta\ ,\ \eta > 0$$

Several choices of R_0, R_1 can be tried.

If $R_1 = 0$, $R_0 = R$ then it follows immediately that the approximation (3.16) is stable for the problem (3.15). However then

completely implicit.

A method which is essentially explicit in all directions but x_1 can be obtained in the following way. Write the differential equations in the form

$$\partial u/\partial t = \frac{1}{2} \sum_{\ell=1}^{m} (A_\ell\ \partial u/\partial x_\ell + \partial(A_\ell u)/\partial x_\ell) + Cu$$

where

$$C = -\frac{1}{2} \sum_{\ell=1}^{m} \partial A_\ell/\partial x_\ell$$

et

$$R_0 = \frac{1}{2} P(A_1Q_1+Q_1A_1) + PC$$

$$R_1 = R - R_0 = \frac{1}{2} P \sum_{\ell=2}^{m} (A_\ell\ Q_\ell + Q_\ell\ A_\ell).$$

Then $(v,R_1v) = 0$ and the approximation is stable if k/h is sufficiently small, i.e. $k|||R_1|||_h \leq 1 - \eta < 1$.

One can split $R_0 = R_{00} + R_{01}$ with $(v,R_{01}v)_h = 0$ such that the whole approximation is essentially explicit, by replacing R_0 by R_{00} and R_1 by

$R_1 + R_{01}$. Denote by P_1 the projection

$$P_1 u = \begin{cases} 0 & \text{for } x^{(1)} = 0 , \text{ i.e. } x \in \partial\Omega \\ u & \text{for } x^{(1)} > 0 . \end{cases}$$

and set

$$R_{01} = \frac{1}{2} PQ - \frac{1}{2}(I-P_1)\, PQ(I-P_1), \quad Q = A_1 Q_1 + Q_1 A_1$$

$$R_{00} = PC + \frac{1}{2}(I-P_1)\, PQ(I-P_1) .$$

Then by (3.12)

$$((u,Qv))_h = -((Qu,v))_h - S < u_0(x_-), A_1(0,x_-)v_0(x_-) >$$

$$- S < A_1(0,x_-)u_0(x_-), v_0(x_-) >.$$

for all functions u,v. Therefore for all u with $u = Pu$

$$2((u,R_{01}u))_h = ((u,P_1PQu))_h + ((u,PQP_1u))_h -$$

$$((u,P_1PQP_1u))_h = ((u,P_1PQPu))_h + ((u,PQPP_1u))_h - ((u,P_1PQPP_1u))_h =$$

$$- ((PQPP_1u,u))_h - ((P_1PQPu,u))_h + ((P_1PQP_1Pu,u))_h =$$

$$- 2\,((u,R_{01}u))_h$$

Thus R_{01} is antisymmetric.

We turn now to the general boundary conditions (1.3). We construct a orthogonal matrix U such that (1.3) is equivalent with

$$(Uu)^I = \tilde{u}^I = 0 .$$

and the projection $\tilde{P}$ which projects the gridfunctions into gridfunctions which satisfy the boundary conditions is given by

$$\tilde{P}u = \begin{cases} u(x) & \text{for } x^{(1)} > 0 , \\ U'PUu & \text{for } x^{(1)} = 0 . \end{cases}$$

For $x^{(1)} > 0$ we use again (3.2) but replace (3.4) and (3.5) by

(3.4a) $(Uv)^I = 0$ i.e. $\tilde{P}v = v,$

$$(3.5a) \qquad (U\, dv/dt)^{II} = \sum_{\ell=1}^{m} (U\, A_\ell(x,t)\, Q_\ell\, v(x,t))^{II},$$

for $x \in \partial\Omega$, $t \geq 0$.

Assume now that dU/dt exists and is uniformly bounded. Then we can differentiate (3.4a) with respect to t and obtain

$$(U\, dv/dt)^{I} = -\,(dU/dt\; v)^{I},$$

and (3.4a), (3.5a) can be written as

$$dv/dt = \tilde{P} \sum_{\ell=1}^{m} A_\ell(x,t)\, Q_\ell\, v(x,t) + B_1 v, \quad \tilde{P}v = v$$

where

$$B_1 v = -\,U' \begin{pmatrix} (dU/dt\; v)^{I} \\ 0 \end{pmatrix}$$

is a bounded operator. Therefore we can essentially also write the problem (3.2), (3.4a), (3.5a) in the form (3.15)

$$dv/dt = \tilde{R} v_1 + P_1 B_1 v, \quad \tilde{P}v = v$$

with

$$\tilde{R} = \tilde{P} \sum_{\ell=1}^{m} A_\ell(x,t)\, Q_\ell\, v(x,t).$$

Now

$$< u_0, \tilde{P} v_0 > \; = < \tilde{P} u_0, v_0 >$$

and therefore also

$$((u, \tilde{P}v))_h = ((\tilde{P}u, v))_h$$

if H is of the form (2.24). Then all the estimates are valid again and theorem 3.1 and the construction of the difference approximations is also valid for the general boundary conditions 1.3.

References

[1] Gustafsson B. " The convergence rate for difference approximations to mixed initial boundary value problems ". Math. of Comp. vol 29, 130, (1975)

[2] Kreiss H-O. and Scherer G. "Finite elements and finite difference methods for hyperbolic partial differential equations". Proc. of Symp. on Mathematical Aspects of Finite Elements in Partial Differential Equations. Academic Press (1974).

[3] Kreiss H-O. and Widlund O. "Difference approximations for initial value problems for partial differential equations " Rep.#7, Dept. of Comp.Sc. Uppsala University (1967).

[4] Richtmyer R. and Morton K.W. "Difference Methods for Initial Value Problems". J,Wiley (1967)

Numerical computations with energy estimates schemes

by

Godela Scherer

April 77

Contents

1 Introduction

In [6] and [7] stable difference schemes for a class of hyperbolic partial differential equations were developed.

We will describe now the numerical properties of these schemes and the experiments done with the third and fourth order approximations. The test results are compared with the ones obtained by J.Oliger in [8] who did extensive computations with some of the test problems.

Notation :

For the difference methods , the domain of integration is covered by a grid of uniform meshsize. In all space directions the net is of size $h = 1/N$, $N \in \mathbb{N}$. The time step is taken equal to k. As usual the rate k/h is called λ . The gridpoints are denoted by $x_\nu(t) = x(\nu h, t)$ and the gridfunction values: $v_\nu(t) = v(x_\nu(t))$ $\nu = 0,1,2\ldots$

2 Diagonal norm schemes

2.1 Existence theorem

The simplest form which the energy conserving schemes considered in [7] can adopt is the one corresponding to a diagonal norm matrix. In order to obtain such a boundary modification for a 2s order scheme in the interior of the right halfplane $0 \leq x < \infty$, 2s boundary points are approximated by an s-th order scheme. The resulting schemes are therefore of accuracy order at most s + 1. (see Gustafsson [4]).

It has been suggested without proof in [6] that it is possible to construct such schemes for $s = 1, 2, 3, 4$.

Theorem. 2.1

Given an approximation of $\partial / \partial x$ of order 2s in the interior, one can define a difference approximation of order s for 2s boundary points such that the total scheme given by

$$\frac{1}{h}\begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} B & C \\ -C^{*} & D \end{pmatrix}$$ is halfbounded in norm $H = \begin{pmatrix} \Lambda & \\ & I \end{pmatrix}$, provided that a certain Vandermonde system of equations in the λ's of $\Lambda = \operatorname{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{2s-1})$ has positive solution.

This last condition is valid for $s < 5$.

In fact, it can be proved numerically that for $5 \leq s \leq 14$ the solution of the Vandermonde system has negative elements and therefore the above described schemes cannot be defined for larger values of s.

Proof

As was pointed out in [6 , pp 206], the elements of the norm matrix Λ : λ_i , i = 0 , ... 2s-1, are determined from formula 2.24

$$\sigma \sum_{v=0}^{2s-1} \lambda_v (2s-v)^{\sigma-1} = r^\sigma - (-1)^\sigma g_\sigma \qquad \sigma = 1, 2, \ldots 2s$$

with $g_\sigma = \begin{cases} -1/2 & \sigma = 1 \\ 0 & = 3,5\ldots \\ B_6 & = 2,4,\ldots \\ 2 \sum_{\nu=1}^{s} \alpha_\nu \sum_{\mu=0}^{\nu-1} \mu^2 (\mu-\nu)^2 & = 2s \end{cases}$

If one takes σ to the right hand side, the system has a Vandermonde coefficient matrix with integer elements of the form

$$\begin{pmatrix} 1 & \ldots & & 1 \\ 2s & \ldots & 3\ \ 2 & 1 \\ \vdots & & & \vdots \\ 2s^{2s-1} & \ldots & & 2^{2s-1}\ 1 \end{pmatrix} \begin{matrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_{2s-1} \end{matrix} = \begin{pmatrix} 1\,(2s + (-1/2)) \\ 1/2\,(2s^2 + B_2) \\ 1/3\,(2s^3 + 0) \\ \vdots \\ 1/2s\,(2s^{2s} - g_{2s}) \end{pmatrix}$$

We denote the matrix by V.

These systems are resolved using an adaptation of the Vandermonde system solver described in [9] .

The algorithm is based on the fact that there is a particularly simple factorization of the inverse of the coefficient matrix. In fact in [9] it was proved that $V^{-1} = UL$, where U and L are upper and lower triangular matrices respectively, L having unit diagonal. Both L and U can be decomposed in a product of bidiagonal matrices and in the case of U bidiagonal and diagonal matrices.

The elements of $(\lambda_0, \ldots \lambda_{2s-1})$ are obtained from the right hand side by recursive premultiplication with these matrices.

The computation of the elements of Λ were not continued further than s=14. It was noted that for each s, the minimum λ was smaller than the minimum of the λ' s associated with s-1.

Although it is possible to prove that an approximation with diagonal norm always exists if one takes enough boundary points, it seems pointless to approximate more abcissae with a lower accuracy scheme.

2.2 Analysis of the 3^d order scheme

We will use in the interior the following difference approximations of increasing order of accuracy 2s

$$\frac{\partial}{\partial x} \sim D^{[2s]}(h) = \sum_{\nu=1}^{s} \lambda_\nu \, D_o(\nu h) \quad , \quad \lambda_\nu = \frac{-2(-1)^\nu (s!)^2}{(s+\nu)\,!\,(s-\nu)!} \qquad s=1,2\ldots$$

These formulae give the highest possible accuracy for that number of points. They use the ordinates of 2s+1 symmetric centered points with antisymmetric coefficients.

The coefficient of the ν-th right hand term (not counting the central term whose coefficient is denoted by α_o and is 0) is

$\alpha_\nu = \dfrac{-2\,(-1)^\nu\, s!}{\nu\,(s+\nu)!\,(s-\nu)!}$. The corresponding left hand term is

$\alpha_{-\nu} = -\alpha_\nu$

Compare [3] for a table of coefficients from 2s=2 to 26.

The above defined approximations can be obtained by differentiating the Gaussian polynomials of order 2s which interpolate the function values at 2s+1 symmetrically situated equidistant points.

The truncation error for them is

$$\frac{h^{2s}}{(2s+1)!} \left(\sum_{\nu=1}^{s} \lambda_\nu \nu^{2s} \right) \frac{\partial^{2s+1}}{\partial x^{2s+1}}(x) + O\left(h^{2s+1}\right) .$$

The 2s boundary points are approximated by a scheme of order s, as was described in [6] and again in the previous section.

Using the results in [4] the final scheme in the halfplane can attain order of accuracy s+1.

Extensive numerical experiments with the above described schemes were only done for the third order method, ie fourth order approximation with second order modification at the boundary points. The choice seems reasonable: in the first place it was demonstrated in [5] that the 4^{th} order schemes when applied to the Cauchy problem for the scalar equation $u_t = c\, u_x$ were only slightly less efficient than for example the 6^{th} order schemes, as long as the allowed error is 1% and the integration is not over very long time intervals. Moreover, for the mixed initial-boundary value problem, where the boundary points are treated implicitly, there is probably no gain in using higher order approximations.

In order to define the 3^{rd} order difference approximation at the boundary points, the elements of the norm matrix Λ and of the matrix B of Q have to be known. The elements of the norm matrix had already been computed in [6]. For the elements of B, the 12 equations resulting from the accuracy conditions (3.1.1) (defined below) were used. By Householder transformations, a linearly independent set of 6 equations was determined for the 6 unknown elements of B. This system was solved by Crout reduction; the approximate solution was subsequently refined by iteration.

Let $C = \frac{1}{h} \sum_{j=-p}^{q} a_j E^j$ denote the difference approximation for $\partial/\partial x$ at a point x in the quarter plane, with E the translation operator. Then by a Taylor expansion one obtains for a sufficiently smooth function $\phi(x)$

$$C\,\phi(x) = \frac{1}{h}\left[\sum_{j=-p}^{q} a_j \phi(x,jh)\right] = \frac{1}{h}\Big[\sum_j a_j\, \phi(x) + h\Big(\sum_j a_j\, j\Big)\phi'(x) + \frac{h^i}{i!}\Big(\sum_j a_j\, j^i\Big)\,\phi^{(i)}(x) \ldots\Big] \tag{2.2.1}$$

In particular $C\,x^{q_1+1}$ gives the truncation error constants at the point x, x^{q_1+1} being the first monomial such that $C\,x^{q_1+1} - (q_1+1)x^{q_1} \neq 0$

One gets a better idea of the error incurred in calculating the coefficients if one looks at the error of the operator $(Q - \frac{\partial}{\partial x})$ when applied to the monomials 1, x, x^2. It is at most $\sigma(10^{-5})$.

The truncation error constants can be obtained from the relation (2.2.1).

In fact, $Q\,\bar{x}^3 - \frac{\partial}{\partial x}\bar{x}^3 = 3!\,\bar{c}$ $(\bar{x} = x_0, x_1, \ldots)$.

Here c_i is the coefficient of the $\sigma(h^2)$ term in the truncation error.

The values of c_i for the boundary points are :

-0.42161303 D+00
0.16669321 D+00
-0.19370812 D-01
-0.37403517 D-01 .

In time, a modified leapfrog scheme will be used so that the total scheme is essentially explicit except for the 2s boundary points.

We like to remark that the schemes will be stable in norm $H = \begin{bmatrix} \Lambda & \\ & I \end{bmatrix}$

as was proved in [6] . It is therefore of interest to know the equivalence constants between the L_2- norm and the H-norm.

The discrete L_2-norm of a vector in one space dimension: $\bar{v} = (v_0, v_1, \ldots)$

is $L_2(\bar{v}) = \left(\sum_{i=0}^{\infty} |v_i|^2 \right)^{1/2} = (\bar{v},\bar{v})^{1/2}$

and the H-norm is $H(\bar{v}) \equiv \|\bar{v}\|_H \equiv \sqrt{\left(\sum_{i=0}^{2s-1} |v_i|^2 \lambda_{ii} + \sum_{i=2s}^{\infty} |v_i|^2 \right)^{1/2}} = (\bar{v}, H\bar{v})^{1/2}$

One seeks constants m,M , positive , and such that

$\forall\, \bar{v}$, $mL_2(\bar{v}) \leq H(\bar{v}) \leq ML_2(\bar{v})$; equivalently, $[mL_2(\bar{v})]^2 \leq [H(v)]^2 \leq [ML_2(v)]^2$.

Ie, find m^2 / $0 \leq (\bar{v}, (H - m^2 I)\bar{v})$. The eigenvalues of $(H - m^2 I)$ are $\lambda(h) - m^2$

(λ (h) : eigenvalue of H).

Therefore choosing $m^2 \leq \lambda_{min}(H)$ one obtains the desired inequality.

Similarly , M^2 has to be chosen such that $M^2 \geq \lambda_{max}(H)$ = largest eigenvalue of H.

One "closest" equivalence relation is then

$$\sqrt{\lambda}_{min}(H)\ L_2(\bar{v}) \leq H(v) \leq \sqrt{\lambda}_{max}(H)\ L_2(\bar{v})$$

For s=1,..4 ; the norm matrix elements are listed in [6] .

With increasing s, the values of λ_{min} , λ_{max} decrease/increase respectively ie the bounds for $H(\bar{v})$ tend further apart.

The elements of the diagonal norm matrix for the fourth order scheme in the interior are :

$\lambda_0 = 0.3541666667$
$\lambda_1 = 1.229166667$
$\lambda_2 = 0.8958333333$
$\lambda_3 = 1.020833333$. (These values have been rounded to 10 decimals)

An equivalence between the L_2 -norm and the H-norm is:

$$0.59511903\ L_2(\bar{v}) \leq H(v) \leq 1.1086779\ L_2(\bar{v})$$

Columns 1-3

-0.14117546D+01	0.17353036D+01	-0.23529389D+00
-0.50000265D+00	0.0	0.50000121D+00
0.93023173D-01	-0.68604836D+00	0.0
0.30611506D-01	-0.47671514D-06	-0.60204189D+00

Columns 4-6

-0.88233133D-01	0.0	0.0
0.39591574D-06	0.0	0.0
0.68604754D+00	-0.93023258D-01	0.0
0.0	0.65306143D+00	-0.81632678D-01

Coefficients of the third order schemes for the 4 boundary points.

2.3. Numerical experiments

Scalar equations.

As first test equations, the scalar equations $u_t = \pm u_x$ were considered. The integration interval was $[0,1]$.

The initial value funcions $f(x)$ were elected 1-periodic. The necessary boundary value conditions i.e. $u(1,t)$ for $u_t = u_x$ and $u(0,t)$ for $u_t = -u_x$ were defined as the analytic solution for the corresponding Cauchy Problem with periodic initial values.

The solution is then $f(x \pm t)$ respectively and is 1-periodic in space and time. The problems are wellposed with the estimate $\| u(.,t) \|_2 = \| u(.,0) \|_2 = \| f(.) \|_2$. Here $\| v(.) \|_2$ denotes the usual L_2 norm in $[0,1]$

The order of the scheme in the interior is $O(h^4) + O(k^2)$.

We give computational results for 2 values of λ :

1) λ was selected as $\lambda = h$ which implies $k = h^2$.
2) By choosing $\lambda = 1/4$ the h^4 and k^2 terms are approximately of the same magnitude for the problems with initial value function $f(x) = \sin 4\pi x$.

We analyze first the results of the case

$$u_t = u_x \qquad x \in [0,1]$$

$$u(x,0) = \sin 4\pi x \quad , \quad u(1,t) = \sin 4\pi t = g(t)$$

The integration interval was divided into subintervals of length $h = \frac{1}{20}$.

Let $v_\nu(t)$ denote the calculated solution for time t at $x = \nu h$, and

$\bar{v}(t) = (v_0(t) \ldots v_N(t))'$.

The additional time step required to start leap-frog, $v_\nu(k)$, was defined by the analytic solution ie $v_\nu(k) = \sin 4\pi(\nu h + k)$. $\nu = 0,1,..N$.

The scheme can be written in matrix notation :

$$(2.3.1) \qquad (I - kQ_2)\,\tilde{v}(t+k) = (I + kQ_2)\,\tilde{v}(t-k) + 2k\left[Q_1\,\tilde{v}(t) + G(t)\right]$$

$$v_N(t+k) = g(t+k) \quad .$$

Here $\tilde{v}(t) = (v_0(t), v_1(t) \ldots v_{N-1}(t))'$

and $G(t) = \frac{1}{h}\,(0, \ldots 0, -q_{41}\,v_N(t) \ldots -q_{21}\,v_N(t))'$

$$Q_2 = \frac{1}{h} \begin{pmatrix} q_{11} & & \\ & 0 & \\ & & \end{pmatrix}_{N\times N}$$

$$Q_1 = \frac{1}{h} \begin{pmatrix} 0 & q_{12} \cdots & q_{16} & & & \\ q_{21} & & q_{26} & & & \\ q_{41} & & q_{46} & & & \\ 0 & 0 & \alpha_1 & \alpha_2 & & \\ & & \ddots & & & \\ & & & & -q_{46} \cdots & -q_{42} \\ & & & & -q_{26} \cdots & -q_{22} \end{pmatrix}_{N\times N} .$$

For the stability analysis we have to prove that the scheme is stable for the two related halfplane problems obtained by removing the right or left boundary and extending the integration domain to $\pm\ \infty$ respectively. For simplicity the left halfplane problem is transformed into a right halplane one through a change of variables : $x' = -x$

The operator $Q = \frac{1}{h} \begin{pmatrix} q_{11} & \cdots & q_{16} \\ & & \\ q_{41} & \cdots & q_{46} \end{pmatrix}$ was constructed as an approximation

to $\partial/\partial x$ in $0 \leq x < \infty$ ie for the first halfplane problem.

In analogy with leapfrog for an ODE, when discretizing in time, one has to separate Q into its symmetric and antisymmetric parts in norm

$$\Lambda \equiv \operatorname{diag}(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$$

The symmetric part is $1/h \cdot \Lambda^{-1} \begin{pmatrix} -1/2 & & \\ & 0 & \end{pmatrix} = 1/h \cdot \begin{pmatrix} q_{11} & & \\ & 0 & \end{pmatrix}$

and the antisymmetric $Q - Q_2$.

For the second halfplane one approximates only the 2^d, 3^{rd}, and 4^{th} boundary points (x_1, x_2, x_3) with the scheme defined by

$$-\frac{1}{h} \begin{pmatrix} q_{22} & \cdots & q_{26} \\ & & \\ q_{42} & \cdots & q_{46} \end{pmatrix} \begin{pmatrix} v_1(t) \\ \vdots \\ v_5(t) \end{pmatrix} + \frac{1}{h} \begin{pmatrix} -q_{21} & v_0(t) \\ & \\ -q_{41} & v_0(t) \end{pmatrix}$$ for the operator $\partial/\partial x$.

Here the coefficient matrix of $v_1(t) \ldots v_5(t)$ is already antisymmetric in norm $\operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$.

The term in $v_0(t)$ can be considered as an inhomogeneous term which does not affect stability.

The above described scheme (2 .3.1) is stable in norm defined by

$$\begin{pmatrix} \Lambda & & & \\ & I & & \\ & & \lambda_3 \; \lambda_2 & \lambda_1 \end{pmatrix}_{N\times N} .$$

In order to prove stability, the well known theorem in [5 , pp 36] is applied, modified for the present case , in the theorem presented below.

Theorem 2.2

Consider a difference approximation of the form $(I-kQ_2)\bar{v}(t+k) = (I+kQ_2)\bar{v}(t-k) + 2kQ_1\bar{v}(t)$ for which the difference operators Q_1, Q_2 fulfill the conditions:

1) $\mathrm{Re}(\bar{v},Q_1\bar{v})_H = 0$, $\mathrm{Re}(\bar{v},Q_2\bar{v})_H \leq 0$, $k\|Q_1\|_H \leq 1-\eta$ for η real constant $0 < \eta < 1$

2) for each fixed h the norm $\|Q_2\|_H$ is bounded.

The $\bar{v}(.,t)$ considered are those that fulfill the boundary conditions of the continuous problem and for which $\|\bar{v}(.,t)\|_H$ exists.

Then for every $\bar{v}(0)$, $\bar{v}(k)$, such that $\|\bar{v}(0)\|_H$, $\|\bar{v}(k)\|_H$ exists, there is a solution $\bar{v}(x,t)$ with

$$\eta\,(\|\bar{v}(.,t+k)\|_H^2 + \|\bar{v}(.,t)\|_H^2) \leq (2-\eta)\,(\|\bar{v}(.,k)\|_H^2 + \|\bar{v}(.,0)\|_H^2),$$
$$t = 0,\ k,\ \ldots$$

Note that as leapfrog involves 3 time levels, the scheme has to be reduced to a 2-step method in the usual way, introducing an auxiliary vector $\bar{\bar{v}}(x,t) = (\bar{v}(x,t),\ \bar{v}(x,t+k))^t$, with norm

$$\|\bar{\bar{v}}(.,t)\|_H^2 = \sum_{j=0}^{1} \|\bar{v}(.,t+jk)\|_H^2 .$$

In the condition 1), the sought estimate $k\|Q_1\|_H \leq 1 - \eta$ gives a bound for the maximum permitted value of the rate k/h.

The spectral norm for HQ_1 was calculated for the problem $u_t = u_x$, $x \in [0,1]$, $h = 1/20$.

Then $\frac{1}{h}\|HQ_1\|_2 = \frac{1}{h}\sqrt{1,822}$, and $\frac{k}{h}\|Q_1\|_H = \lambda \times 1,349815$.
This implies $\lambda_{max} < 7.408424 \times 10^{-1}$.

For simplicity, the discrete L_2-norm was used. This is defined as

$$\|\bar{v}(t)\|_2 = \left(h \sum_{\nu=0}^{N-1} |v_\nu(t)|^2\right)^{1/2}.$$

The error function is denoted by $e_\nu(t) = v_\nu(t) - u(x_\nu, t)$.
The initial function and the solution has for any time t, the L_2-norm = 0.70711.
For $\lambda = 0.25$, and $\lambda = h$ the values of $\|\bar{e}(t)\|_2$ and $\|\bar{v}(t)\|_2$ are listed for several values of t.

$\lambda = 0.25$ h = 0.5 D-01, k = 0.125 D-01

t	$\|\bar{v}(t)\|_2$	$\|\bar{e}(t)\|_2$
0.25 D-01 (1)	0.70725	0.82146 D-2
5 (2)	0.63769	0.94110 D-1

(1) First time step calculated with the scheme t = 2k

(2) 400 time steps

$\lambda = h$ $\quad h = 0.5$ D-01 , $\quad k = 0.25$ D-2

t	$\|\bar{v}(t)\|_2$	$\|\bar{e}(t)\|_2$
0.5 D-2 (1)	0.70712	0.20230 D-2
1	0.66929	0.88047 D-1
4 (2)	0.63383	0.10885

(1) $t = 2k$

(2) 1600 time step

For comparison, one test example with $f(x) = \sin 2\pi x$ was run with $\lambda = 0.25$, $h = 0.5$ D-1 $\quad k = 0.125$ D-1 . Again the L_2-norm of the solution is 0.70711.

t	$\|\bar{v}(t)\|_2$	$\|\bar{e}(t)\|_2$
0.25 D-1	0.70712	0.12943 D-2
5	0.70980	0.59235 D-2

Although the truncation error is smaller with $\lambda = h$, the number of time steps needed to reach a certain t is 5 times larger . In figure 1 the solution at t=4 is plotted against the analytic solution.

The representation of the lower frequency is considerably better. In Appendix A fig. 2 and 3 give a clear comparison of the two representations.

We analyse now

$u_t = - u_x$, $x \in [0,1]$ with initial values and boundary conditions $u(x,0) = \sin 4\pi x$ and $u(0,t) = \sin 4\pi(-t) = g(t)$. The solution is then $u(x,t) = \sin 4\pi(x-t)$.

For this problem, Oliger in [8] gives extensive comparisons of different schemes. For some computations, the integration interval was divided in 20 and again the two values of λ used were $\lambda = 0.25$ and h. But for 7 of the 21 gridpoints the order of approximation is only $0(h^2)$. Therefore , in order to show better the accuracy of the method, another test run with $h = \frac{1}{40}$ was done.

The scheme is $(I-kQ_2)\ \tilde{v}(t+k) = (I+kQ_2)\ \tilde{v}(t-k) + 2k\left[Q_1\ \tilde{v}(t) + G(t)\right]$

$$v_0(t+k) = g(t+k)$$

where $\tilde{v}(t) = (v_1(t),\ldots v_N(t))'$ and

$$G(t) = (-q_{21}\ v_0(t),\ -q_{31}\ v_0(t), -q_{41}\ v_0(t), 0\ldots 0)'.$$

Q_2 and Q_1 can be derived from the corresponding matrices for the positive problem by changing the sign of the elements and then by rotating the columns so that the order is N, N-1, ...1, followed by a similar change of the rows.

The norm matrix for $u_t = u_x$ changed in the same way is a norm for the present case.

The time step $\bar{v}(k)$ is defined by the analytic solution .

We will denote this scheme with C1. The numerical results for the 20-point net and $\lambda = 0.25$ are compared with the methods C2, C4 and C5 of [8].

The C4-scheme is 4^{th} order in the interior with a $0(h+k^2)$ approximation for $\nu = N$: $v_\nu(t+k) = v_\nu(t-k) - 2\lambda\left\{\left[v_\nu(t+k) + v_\nu(t-k)\right]/2 - v_{\nu-1}(t)\right\}$ and leapfrog in space as well as in time for $\nu = 1$, and $\nu = N-1$, ie an $0(h^2+k^2)$ approximation . The C5-method is second order in the interior with the above $0(h+k^2)$ extrapolation for $\nu = N$. It corresponds to the diagonal norm scheme for $\beta = 1$ developed in [6] .

The C 2 scheme is 4^{th} order in the interior with a $O(h^3)$ extrapolation for

ν = 1,N-1,N.

The following are the results of the computations.

For a grid of 21 points i.e. h = 1/20

λ = 0.25 h = 0.5 D-1 k = 0.125 D-1

t	$\|\bar{v}(t)\|_2$	$\|\bar{e}(t)\|_2$	C4: $\|\bar{e}(t)\|_2$	C5: $\|\bar{e}(t)\|_2$	C2 : $\|\bar{e}(t)\|_2$
0.025	0.70657	0.48907 D-2			
1	0.67054	0.72194 D-1	1.28 D-1	3.87 D-1	.134 D-1
2	0.6337	0.97992 D-1	1.88 D-1	4.44 D-1	.13 D-1
4	0.6375	0.92474 D-1	1.75 D-1	4.8 D-1	.125 D-1

The error for our 3^d order scheme is approximately 7 times as large as for Oligers 4^{th} order scheme.

For λ = h h = 0.5 D-1 k = 0.25 D-2

t	$\|\bar{v}(t)\|_2$	$\|\bar{e}(t)\|_2$
0.005	0.70708	0.10639 D-2
1	0.669	0.85785 D-1
2	0.62932	0.11187
4	0.6333	0.10571 (1)

(1) 1600 time steps

For 41 grid points i.e.

λ = 0.25 h = 0.25 D-1 k = 0.00625

t	$\|\bar{v}(t)\|_2$	$\|\bar{e}(t)\|_2$
0.0125	0.70709	0.53558 D-3
1	0.70362	0.73863 D-2
2	0.69987	0.97573 D-2
4	0.69992	0.96818 D-2

The error is approximately 1/10 after duplicating the number of gridpoints. Fig. 4 and 5 compare the approximations when using a grid of 21 or 41 points respectively. Notice that in this last case the calculated solution is still a fairly good approximation after 4 time periods.

System of equations in one space dimension

As a second example a system of two scalar equations in one space dimension was considered. The equations were coupled through the reflective boundary conditions.

If $w(x,t) = [u(x,t), v(x,t)]'$ then

$$w_t = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} w_x \qquad x \in [0,1]$$

The initial values were $w(x,0) = [f(x),0]'$ and

the boundary conditions $u(0,t) = v(0,t)$ and $v(1,t) = u(1,t)$

Two values for $f(x)$ were tried:

a) $f(x) = \sin \pi x$

b) $f(x) = \begin{cases} 1 - 20/3\, d & d < 0.15 \\ 0 & d \geq 0.15 \end{cases}$ $\qquad d \equiv |x - 1/2|$

i.e. the roof function.

Well-posedness of the problem is easily proved via an energy estimate:

$$\frac{d}{dt}(w,w) = 2\left(w, \frac{\partial}{\partial t} w\right) = 2\left[-\int_0^1 uu_x dx + \int_0^1 vv_x dx\right] =$$

$-[u^2(1) - u^2(0)] + [v^2(1) - v^2(0)] = 0$ for the given boundary conditions.

For the extra time step the analytic solution $w(k) = \begin{bmatrix} f(x_v - k) \\ 0 \end{bmatrix}$ was taken.

The grid solutions of $u(t)$ and $v(t)$ are denoted by $U(t)$ and $V(t)$ respectively. The absolute value of the error function is defined at point x_ν by

$$|e_\nu(t)|^2 = |u(x_\nu,t) - U_\nu(t)|^2 + |v(x_\nu,t) - V_\nu(t)|^2$$

The L_2-norm of the calculated solution $W_\nu(t) = [U_\nu(t), V_\nu(t)]'$ and the error function defined by $\|.\|^2 = h \sum_{\nu=0}^{N} |.|^2$

In the numerical tests the integration interval was divided in 50 sub-intervals.

For the Initial value function a) λ was defined as 0.2, h= 0.02, k = 0.004

t	$\|\bar{e}(t)\|_2$	$\|\bar{W}(t)\|_2$	$\|\bar{w}(t)\|_2$
0.08	0.36454 D-3	0.70712	0.70712
		0.71975	
1.0	0.84558 D-2	0.70712	0.70711
3.6	0.9701 D-2	0.71995	0.71978

For f(x) : The roof function and

$\lambda = 0.25$, h = 1/50 , k = 0.005

t	$\|\bar{e}(t)\|_2$	$\|\bar{W}(t)\|_2$	$\|\bar{w}(t)\|_2$	$\|\bar{e}(t)\|_2$ Oliger
0.01	0.84437 D-2	0.3176	0.31552	
1	0.2618 D-1	0.3163	0.31763	0.3 4 D- 1
2	0.2219 D-1	0.31593	0.31763	0.3 31 D- 1
3	0.24667 D-1	0.31564	0.31763	0.3 61 D- 1

We compare with Oligers 4^{th} order approximation with $O(h^3)$ extrapolation at the boundary described in [8]

Notice that the error of the scheme is smaller than the one for Oliger's.

A graphical display of the solution at t = 3 is given in Fig.7 of Appendix A. The error is of the order 10^{-2}.

Shallow water equations

These equations describe the divergent motion of water or other non-viscuous, incompressible, hydrostatic fluid with a free top surface if the height of the fluid column is small compared with for example the maximum radius of curvature of the top surface.

The fluid is contained in a rectangular area, the flow being periodic in the west-east direction with a period equal to the length of the rectangle. For the southern and northern boundaries two sets of homogeneous boundary conditions were given. On the top surface, the excess pressure above atmospheric pressure is taken as zero.

The coordinate system is cartesian, centered on Earth, with the x-axis in the west-east direction and the y-axis in the south-north direction.

Then, using the Euler equations of motion, one can obtain a system involving the top surface elevation h and the velocities u, v at the same top surface, in x and y directions.

It is

$$\begin{pmatrix} u \\ v \\ \phi \end{pmatrix}_t = \begin{pmatrix} -u & -1 & \\ & -u & \\ -\phi & & -u \end{pmatrix} \begin{pmatrix} u \\ v \\ \phi \end{pmatrix}_x + \begin{pmatrix} -v & & \\ & -v & -1 \\ & -\phi & -v \end{pmatrix} \begin{pmatrix} u \\ v \\ \phi \end{pmatrix}_y + \begin{pmatrix} & f & \\ -f & & \\ & & \end{pmatrix} \begin{matrix} u \\ v \\ \phi \end{matrix}$$

Here $\phi = gh$ is the geopotential and f is the Coriolis parameter

The integration area is $0 \leq x \leq L$, $0 \leq y \leq D$, $t \geq 0$

This system of equations, after fixing the coefficient matrices at values $u = u_o$, $v = v_o$, $\phi = \phi_o$, can be transformed to a system with symmetric matrix for the x derivative and diagonal for the y, by the change of variables:

$$\bar{\Psi} = \mathcal{R}\,\bar{\omega} \quad , \quad \bar{\omega} = \begin{pmatrix} u \\ v \\ \phi \end{pmatrix} \quad , \quad \mathcal{R} = \begin{pmatrix} \sqrt{\Phi_0} & & \\ & \sqrt{\Phi_0}/\sqrt{2} & 1/\sqrt{2} \\ & \sqrt{\Phi_0}/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \quad \mathcal{R}^{-1} = \begin{pmatrix} 1/\sqrt{\Phi_0} & & \\ & 1/\sqrt{\Phi_0}\sqrt{2} & 1/\sqrt{\Phi_0}\sqrt{2} \\ & 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

The system then looks like

$$\bar{\Psi}_t = \begin{pmatrix} -u_0 & \sqrt{\Phi_0}/\sqrt{2} & \sqrt{\Phi_0}/\sqrt{2} \\ -\sqrt{\Phi_0}/\sqrt{2} & -u_0 & \\ \sqrt{\Phi_0}/\sqrt{2} & & -u_0 \end{pmatrix} \bar{\Psi}_x + \begin{pmatrix} -v_0 & & \\ & -v_0 - \Phi_0 & \\ & & -v_0+\Phi_0 \end{pmatrix} \bar{\Psi}_y + \begin{pmatrix} 0 & f/\sqrt{2} & f/\sqrt{2} \\ -f/\sqrt{2} & & \\ f/\sqrt{2} & & \end{pmatrix} \bar{\Psi}$$

We will denote the matrices by: A, B, C respectively and $\sqrt{\Phi}_0 = c$

The boundary conditions will be chosen so as to obtain a halfbounded differential operator i.e. one for which an estimate of the type

$\frac{d}{dt} \| \bar{\Psi}(t) \|^2 \leq K \| \bar{\Psi}(t) \|^2$ holds, where the norm is defined by

$$(\bar{\Psi}, \bar{\Psi}) = \| \bar{\Psi}(t) \|^2 \equiv \int_0^D \int_0^L \left(|\Psi_1(x,y,z)|^2 + |\Psi_2(x,y,z)|^2 + |\Psi_3(x,y,z)|^2 \right) dx\,dy$$

In the x-direction, $\bar{\Psi}$ will be periodic i.e. $\bar{\Psi}(x,y,t) = \bar{\Psi}(x+L,y,t)$.

Then

$$\frac{d}{dt}(\bar{\Psi},\bar{\Psi}) = (\bar{\Psi}, \frac{\partial}{\partial t}\bar{\Psi}) + (\frac{\partial}{\partial t}\bar{\Psi}, \bar{\Psi}) = 2\left(\bar{\Psi}, A\frac{\partial \bar{\Psi}}{\partial x}\right) + 2\left(\bar{\Psi}, B\frac{\partial \bar{\Psi}}{\partial y}\right) + (\bar{\Psi}, C\bar{\Psi}) + (C\bar{\Psi}, \bar{\Psi})$$

The term involving A is zero because of the periodic boundary condition in the x direction, and the term in C disappears also because the matrix is antisymmetric.

Therefore

$$\frac{d}{dt}(\bar{\Psi},\bar{\Psi}) = 2(\bar{\Psi}, B\frac{\partial \bar{\Psi}}{\partial y}) = (\bar{\Psi}, B\bar{\Psi})\Big|_0^D \quad \text{by integration by parts}$$

$$= \left(\int_0^L \bar{\Psi}^* B \bar{\Psi}\, dx \right)\Big|_0^D$$

We will consider problems for which ϕ is always positive and also $\phi >> u^2 + v^2$. Then for ψ_2 a boundary condition at y=0 has to be given, and for ψ_3 one at y=D. An additional one has to be imposed for ψ_1 at y=0 or y=D depending on the sign of the parameter v_0. Now

$$\frac{d}{dt}(\bar{\psi},\bar{\psi}) = \int_0^L v_0\left[-|\psi_1(x,D,t)|^2 dx + |\psi_1(x,0,t)|^2\right] dx +$$

$$\int_0^L (-v_0-c)\left[|\psi_2(x,D,t)|^2 - |\psi_2(x,0,t)|^2\right] dx +$$

$$\int_0^L (-v_0+c)\left[|\psi_3(x,D,t)|^2 - |\psi_3(x,0,t)|^2\right] dx$$

Numerical Results.-

The numerical experiments were made on a modification of the example considered by Elvius and Sundström in [2]. In order to obtain a constant coefficient problem, u_0, v_0, ϕ_0 were taken equal to the corresponding parameter values in [2].

The initial values suggested "describe a single wave on a constant velocity field with the geopotential chosen so that the nonlinear balance equation is initially satisfied":

$$u(x,y,0) = u_0 - u_1 \sin\frac{2\pi x}{L} \cos\frac{2\pi y}{L}$$

$$v(x,y,0) = v_0 + u_1 \cos\frac{2\pi x}{L} \sin\frac{2\pi y}{L}$$

$$\phi(x,y,0) = \phi_0 + fv_0\left(x - \frac{1}{2}L\right) - fu_0\left(y - \frac{1}{4}L\right) + \frac{Lf}{2} u_1 \sin\frac{2\pi x}{L}\sin\frac{2\pi y}{L}$$

$$+ \left(\frac{1}{2}u_1\right)^2 \left(\cos\frac{4\pi x}{L} + \cos\frac{4\pi y}{L}\right)$$

The area of integration was $0 \le x \le L$, $0 \le y \le \frac{L}{2} = D$.

The Coriolis parameter is $f = 10^{-4}\ s^{-1}$

The parameter values used were

$u_0 = 5$ m/s , $v_0 = 0$ m/s , $u_1 = 10$ m/s , $\phi_0 = 10^4 m^2/s^2$,

$L = 4 \times 10^6$ m.

The integration region was divided into a 20 x 10 point grid with space interval $h = 2 \times 10^5$ m.

In [2] the time step was of one hour. This gives a value for λ of 0.18×10^{-1}.

Since v_0 is 0 there is no need of a boundary condition for ψ_1. And for this choice of the coefficient matrices, the following set of homogeneous boundary conditions give an energy estimate of the solution :

a) $\psi_2(x,0,t) = 0$

$\psi_3(x,D,t) = 0 \qquad \forall x, \forall t$ It gives the estimate $\|\bar{\psi}(t)\| \leq \|\bar{\psi}(0)\|$

For continuity reasons, the set of values from the initial conditions were assigned:

$\psi_2(x,0,t) = \psi_2(x,0,0)$

$\psi_3(x,D,t) = \psi_3(x,D,0) \qquad \forall x, \forall t.$

After some time an homogeneous altitude h is attained.

b) Boundary conditions of type $L\bar{\psi} = 0$ for example :

$\psi_2(x,0,t) + \psi_3(x,0,t) = 0$

$\psi_3(x,D,t) + \psi_2(x,D,t) = 0 \qquad \forall x, \forall t$ It gives $\|\bar{\psi}(t)\| = \|\bar{\psi}(0)\|$

It corresponds to a generalized reflection.

Difference scheme

The integration area is covered by a grid of width h in both x and y directions.

The gridpoints are denoted by the pair (x_i, y_j) $0 \le i \le n$, $0 \le j \le m$.

The vector $\bar{\psi}(t)$ denotes the difference approximation to the solution and is

$$\bar{\psi}(t) = (\psi_1(x_0,y_0,t), \psi_1(x_0,y_1,t), \dots \psi_1(x_0,y_m t), \psi_1(x_1,y_0,t) \dots, \dots$$
$$\psi_2(x_0,y_0,t), \dots \quad \psi_3(x_0,y_0,t) \dots).$$

The second order centered leapfrog method is used for the time derivative. The problem can be reduced for every t after fixing x, to the solution of the following system of equations:

$$\begin{pmatrix} I & & \\ & I-ckQ_2^- & \\ & & I-ckQ_2^+ \end{pmatrix} \begin{pmatrix} \psi_1(t+k) \\ \psi_2(t+k) \\ \psi_3(t+k) \end{pmatrix} = \begin{pmatrix} I & & \\ & I+ckQ_2^- & \\ & & I+ckQ_2^+ \end{pmatrix} \begin{pmatrix} \psi_1(t-k) \\ \psi_2(t-k) \\ \psi_3(t-k) \end{pmatrix} +$$

$$\frac{2k}{} \begin{pmatrix} -u_0 D & -c/\sqrt{2}D & c/\sqrt{2}D \\ -c/\sqrt{2}D & -u_0 D & \\ c/\sqrt{2}D & & -u_0 D \end{pmatrix} \begin{pmatrix} \psi_1(t) \\ \psi_2(t) \\ \psi_3(t) \end{pmatrix} +$$

$$\frac{2k}{} \begin{pmatrix} 0 & & \\ & cQ_1^- & \\ & & cQ_1^+ \end{pmatrix} \begin{pmatrix} \psi_1(t) \\ \psi_2(t) \\ \psi_3(t) \end{pmatrix} +$$

$$\frac{2k}{} \begin{pmatrix} & f/\sqrt{2} & f/\sqrt{2} \\ -f/\sqrt{2} & & \\ -f/\sqrt{2} & & \end{pmatrix} \begin{pmatrix} \psi_1(t) \\ \psi_2(t) \\ \psi_3(t) \end{pmatrix}$$

Here Q_2^- and Q_2^+ are the symmetric parts of the 4^{th} order difference approximation Q when applied respectively to a "negative" scalar equation : $c\frac{\partial}{\partial y}$ with $c < 0$, and to $c\frac{\partial}{\partial y}$ with $c > 0$.

Similarly Q_1^- and Q_1^+ are the antisymmetric parts.

D is the 4^{th} order antisymmetric difference approximation :

$$\frac{4}{3} D_0(h) - \frac{1}{3} D_0(2h) .$$

The additional time step $\bar{\psi}(k)$ was defined by

$$\bar{\psi}(k) \equiv \bar{\psi}(0) + k \left.\frac{\partial \bar{\psi}}{\partial t}\right|_{t=0} = \bar{\psi}(0) + k \left(A \left.\frac{\partial \bar{\psi}}{\partial x}\right|_{t=0} + B \left.\frac{\partial \bar{\psi}}{\partial y}\right|_{t=0} + C \bar{\psi}|_{t=0}\right)$$

The scheme is conditionally stable in norm defined by the Krönecker product of matrices :

$$\mathcal{H} \equiv I_3 \otimes I_n \otimes H_m$$

The subindices indicate the corresponding dimensions

The stability is proved via an analog of theorem 2.2

In order to get an estimate of the values of λ which give stability for our scheme, the maximum value of λ for the scheme of 4^{th} order approximation D in both x and y directions and leap frog in time was calculated for the associated Cauchy Problem with periodic boundary conditions in both space directions.

The condition for stability is:

$k \|Q_1\| \leq 1 - \eta$ with η constant $0 < \eta < 1$

When applying this to our problem Q_1 is $A D_x \bar{\psi}(t) + B D_y \bar{\psi}(t)$

One can calculate the norm of $\|Q_1\|$ by calculating $\|\hat{Q}_1\|$ i.e. the norm of the Fourier Transform:

$$k \hat{Q}_1 = i \frac{k}{h} A \underbrace{\left(\frac{4}{3} \sin\zeta - \frac{1}{6} \sin 2\zeta\right)}_{\alpha} + i \frac{k}{h} B \underbrace{\left(\frac{4}{3} \sin\eta - \frac{1}{6} \sin 2\eta\right)}_{\beta}$$

here ζ, η are $\zeta \equiv \omega_x h$, $\eta \equiv \omega_y h$

Then the eigenvalues of $k\hat{Q}_1$ are z such that

$\det(z\bar{I} - \hat{Q}_1) = 0$ where $k\hat{Q}_1 =$

$$i\frac{k}{h} \begin{vmatrix} -u_0\alpha - v_0\beta & -\sqrt{\phi_0}/\sqrt{2} & \sqrt{\phi_0}/\sqrt{2} \\ -\sqrt{\phi_0}/\sqrt{2} & -u_0\alpha + (-v_0 - \sqrt{\phi_0})\beta & \\ \sqrt{\phi_0}/\sqrt{2} & & -u_0\alpha + (-v_0 + \sqrt{\phi_0})\beta \end{vmatrix}$$

The eigenvalues of $k\hat{Q}_1$ are

$$z_{1,2} = \frac{ik}{h}\left(u_0\alpha + v_0\beta \pm \sqrt{\phi_0(\alpha^2 + \beta^2)}\right)$$

$$z_3 = \frac{ik}{h}(u_0\alpha + v_0\beta)$$

The approximation is stable if $\lambda |u_0 \alpha + v_0 \beta| < 1$

$$\lambda |u_0 \alpha + v_0 \beta \pm \sqrt{\phi_0 (\alpha^2 + \beta^2)}| < 1$$

Here $|u_0 \alpha + v_0 \beta| \leq \frac{5}{3} (|u_0| + |v_0|)$

and $|u_0 \alpha + v_0 \beta \pm \sqrt{\phi_0 (\alpha^2+\beta^2)}| \leq \frac{5}{3} (|u_0| + |v_0|) + \sqrt{\phi_0} \frac{1}{3} \sqrt{50}$

Using the Elvius-Sundström values of the parameters this implies that

for $\lambda \leq 4.13 \times 10^{-3}$ there is stability.

For the boundary conditions a) the following figures give the equipotential curves of the geopotential ϕ at time t=0, t=24 hours and t=48 hours. At each t the scale with which ϕ is represented was recalculated according to the maximum and minimum value of ϕ.

The L_2-norm of the solution $\bar{\psi}$ is:

t=0 0.286828 E 11

=24 h. 0.284961 E 11

=48 h. 0.284957 E 11

Minimum and Maximum of geopotential at gridpoints

0.7949 E 4 , 0.1195 E 5

Equipotential curves scale : 0.8013 E 3
contour lines

```
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 3
2 3 3 4 4 4 4 4 3 3 2 2 2 1 1 1 1 1 2 2 2
2 3 3 4 4 4 4 4 3 3 2 2 1 1 0 0 0 1 1 2 2
2 3 4 4 4 4 4 4 4 3 2 1 1 0 0 0 0 0 1 1 2
2 3 4 4 4 4 4 4 4 3 2 1 1 0 0 0 0 0 1 1 2
2 3 3 4 4 4 4 4 3 3 2 1 1 0 0 0 0 0 1 1 2
2 2 3 3 4 4 4 3 3 2 2 1 1 0 0 0 0 0 1 1 2
2 2 3 3 3 3 3 3 3 2 2 1 1 0 0 0 0 0 1 1 2
2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 2
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

FIG 1

Minimum and Maximum of geopotential at gridpoints

0.9988 E 4 , 0.1006 E 5

Equipotential curves scale : 0.1582 E 2
Contour lines

```
3 2 2 1 1 0 0 0 0 0 1 1 2 2 3 3 4 4 4 4 3
3 3 2 1 1 0 0 0 0 0 0 1 2 2 3 4 4 4 4 4 3
3 3 2 1 0 0 0 0 0 0 0 1 2 2 3 4 4 4 4 4 3
2 2 2 1 1 1 1 1 1 1 1 1 2 2 2 3 3 3 3 3 2
2 2 2 1 1 1 1 1 1 1 1 2 2 2 2 2 3 3 3 2 2
3 2 2 1 1 0 0 0 0 0 0 1 2 2 3 3 4 4 4 4 3
2 2 2 1 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 2 2
3 2 2 1 1 1 1 1 1 1 1 1 2 2 2 3 3 3 3 3 3
3 3 2 1 0 0 0 0 0 0 0 1 2 2 3 4 4 4 4 4 3
3 2 2 1 0 0 0 0 0 0 0 1 1 2 3 3 4 4 4 4 3
3 2 2 1 1 0 0 0 0 0 1 1 1 2 3 3 4 4 4 3 3
```

Fig 2

Minimum and Maximum of geopotential at gridpoints

0.1001 E 5 , 0.1003 E 5

Equipotential curves scale : 0.4690 E 1
Contour lines

```
2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 1 1 1 1 1 1 2 2 2 2 3 3 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
0 0 0 1 1 2 3 3 4 4 4 4 3 3 2 1 1 1 0 0 0
1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1
4 4 3 3 2 1 0 0 0 0 0 0 0 1 1 2 3 4 4 4 4
2 2 2 2 2 2 2 2 3 3 3 2 2 2 2 2 2 2 2 2 2
0 0 0 1 1 2 3 3 4 4 4 4 3 2 2 1 1 0 0 0 0
2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 2
2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 2 2 2 3 3 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

Fig 3

For the boundary conditions b) Fig. 4 and 5 give equipotential curves at t=24 hours and 48 hours.

The L_2-norm of the solution is:

t= 24 hours 0.286038 E 11

= 48 hours 0.284997 E 11

Minimum and Maximum of geopotential at gridpoints
0.9059 E 4 , 0.1096 E 5

Equipotential curves scale : 0.3815 E 3
Contour lines

2 2 2 3 3 3 4 4 4 4 3 3 3 2 2 2 1 1 1 2 2
2 2 2 3 3 4 4 4 4 4 3 3 3 2 2 2 1 1 1 2 2
1 2 2 3 3 4 4 4 4 4 3 3 3 2 2 1 1 1 1 1 1
1 2 2 3 4 4 4 4 4 4 4 3 2 2 1 1 0 0 0 1 1
1 1 2 3 3 4 4 4 4 4 4 3 2 2 1 0 0 0 0 0 1
0 1 2 3 3 4 4 4 4 4 4 3 2 1 1 0 0 0 0 0 0
0 1 2 2 3 4 4 4 4 4 3 3 2 1 1 0 0 0 0 0 0
0 1 1 2 3 3 4 4 4 3 3 2 2 1 0 0 0 0 0 0 0
0 1 1 2 2 3 3 3 3 3 2 2 2 1 1 0 0 0 0 0 0
1 1 1 2 2 2 3 3 3 2 2 2 1 1 1 0 0 0 0 0 1
1 1 1 2 2 2 2 3 2 2 2 2 1 1 1 0 0 0 0 0 1

Fig. 4

Minimum and Maximum of geopotential at gridpoints

0.8939 E 4 , 0.1113 E 5

Equipotential curves scale : 0.4397 E 3

Contour lines

```
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
2 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2 2 2 2 3 2
3 3 3 2 2 1 1 1 0 0 0 0 1 1 2 2 3 3 3 3 3
4 4 3 3 2 1 1 0 0 0 0 0 0 1 2 2 3 3 4 4 4
4 4 4 3 2 1 1 0 0 0 0 0 0 1 2 2 3 4 4 4 4
4 4 3 3 2 2 1 0 0 0 0 0 1 1 2 2 3 4 4 4 4
4 4 3 3 2 2 1 1 1 1 1 1 1 1 2 2 3 3 4 4 4
3 3 3 2 2 2 2 2 1 1 1 1 2 2 2 2 3 3 3 3 3
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

Fig 5

2.4 Experiments with a refined grid

It was proved in [6] that the highest overall accuracy attainable for a fourth order difference scheme in the interior , when using a diagonal norm matrix is third order.

In order to obtain a more uniform error, a grid refinement at the boundaries was tried out.

The model problem examined is the scalar equation :

$u_t = - u_x, \quad x \in [0,1]$

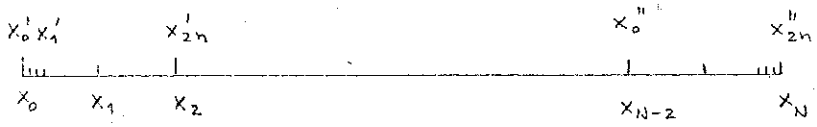
$u(x,0) = \sin 4\pi x \quad , \quad u(0,t) = \sin 4\pi (-t) \equiv g(t)$

On the interval $[0,1]$ two grids are defined :

1) an uniform grid of interval size h_1 . The gridpoints are denoted by x_ν.

2) an uniform grid refinement for the intervals : $[0,2h_1]$ and $[1-2h_1,1]$.

The distance between meshpoints is $h_1/n = h_2$. The gridpoints are denoted x'_ν and x''_ν respectively. (see following fig.)



The corresponding gridfunctions are denoted by : $v(x_\nu,t)$, $v'(x'_\nu,t)$ and $v''(x''_\nu,t)$

One seeks to obtain gridfunction values $v(x_\nu,t)$ which approximate the analytic solution $u(\nu h_1,t)$ with order h_1^3 at the boundary points . Then according to [4] overall accuracy of order h_1^4 can be reached.

In time, second order leapfrog with stepsize k is used for all grids.

Three sets of calculations are performed : for v, v', v'' .

The two initial time steps required for the 3 functions are obtained from the initial conditions and the analytic solution at time t=k.

The gridfunction values $v'(x'_\nu, t)$ are calculated as an independent set like the solution of the problem $u_t = -u_x$ extensively discussed in section 2.3.

From them, $v(x_1,t)$ and $v(x_2,t)$ are defined by $v(x_1,t) = v'(x'_n, t)$, $v(x_2,t) = v'(x'_{2n}, t)$.

Next, the gridfunction v is approximated by the 4^{th} order scheme :

$4/3\ D_0(h) - 1/3\ D_0(2h)$ at the points x_ν, $\nu = 3, \ldots N-2$.

Last, the function values $v''(x''_\nu, t)$ can be obtained considering them as an approximation to the partial problem:

$u_t = -u_x \quad x \in [1-2h_1, 1]$ with boundary conditions : $u(1-2h_1, t) = v(x_{N-2}, t)$.

With these values the definition of $v(x, t)$ is completed by assigning $v(x_{N-1}, t) = v''(x''_n, t)$ and $v(x_N, t) = v''(x''_{2n}, t)$.

In the interior the gridsize was taken as $h_1 = 1/20$.

The value $v(x_N, t)$ is obtained as a boundary point and therefore approximated with order $O(h_2^2)$. Threfore we want to choose n such that $h_2^2 = h_1^2/n^2 \sim h_1^3$. For example take $n = 1/\sqrt{h_1} \sim 4.47$.

In the numerical experiments we took n=5.

For simplicity the same time step k was taken for the refinement and the original grid.

Note that if $\lambda_1 = k/h_1$, $\lambda_2 = k/h_2$ then $\lambda_2 = n\,\lambda_1$, therefore care must be taken that both λ fulfill the stability requirement. λ_2 was defined as $\lambda_2 = 0.25$, then $\lambda_1 = 0.05$.

With the chosen values of n, each subinterval $[0, x_2]$ and $[x_{N-2}, x_N]$ consists of 11 gridpoints, 8 of them being boundary points ie function values approximating the solution with order $O(h_2^2)$. Because the proportion of these points is so large, $v(x_1,t)$, $v(x_2,t)$, $v(x_{N-1}, t)$ and $v(x_N, t)$ are probably only approximated with order h_2^2. But this is enough to assure $O(h_1^3)$ for them.

The following are the results of the computations.

t	$\|v(t)\|_2$	$\|e(t)\|_2$	$\|v(t)\|_2^{(1)}$	$\|e(t)\|_2^{(1)}$
0.005	0.7071	0.186E-3	0.7071	0.3519E-3
2	0.7045	0.2 E-1	0.7019	0.3109E-1
4	0.7045	0.2E-1	0.7020	0.3035E-1

A detailed stability analysis will follow.

(1) these values correspond to method C2 described in [8].

3 Full matrix norm schemes

3.1 Computation of the coefficients of the 4^{th} order scheme

From the full matrix norm schemes whose existence was proved in ([7], theorem 2.2) for every order of accuracy in the interior, we choose the fourth order scheme which corresponds to the already considered fourth order difference approximation to $\frac{\partial}{\partial x}$, $\frac{4}{3} D_0(h) - \frac{1}{3} D_0(2h)$, with a third order modification for the boundary points $x_0, x_1, \ldots x_4$.

For the calculation of the coefficients, there are two possible ways:

1) Follow the steps of the constructive proof of theorem 2.2. The idea is basically the following:

The norm matrix H has the form $H = \begin{pmatrix} \lambda_0 & & \\ & h_{11} \cdots h_{14} & \\ & \vdots \quad \vdots & \\ & h_{14} \cdots h_{44} & \end{pmatrix}$

Instead of calculating the elements of H from the accuracy conditions:

(3.1.1) $B_2 e_j = j\, H\, e_{j-1} \triangleq B_1\, e_j - C f_j \qquad j = 0, 1, \ldots 3$;

one calculates from them the elements of a matrix (p_{ij}), $0 \leq i, j \leq s$, $\langle e_i, H e_j \rangle \equiv p_{ij}$.

In fact, for the present example these values have already been evaluated in [7].

If the matrix (p_{ij}) is positive definite, and one elects a small enough value of the parameter λ_0, then it is easy to proof that one obtains a positive definite matrix H.

There are two degrees of freedom in the equations for the p_{ij}, and p_{22} p_{33} can be chosen so that $(p_{ij}) > 0$.

Using the fact that for a positive definite matrix, each principal submatrix determinant has to be positive, the selected values for

p_{22} and p_{33} were the ones that would define a singular submatrix perturbed at the third decimal digit i.e. if d_i , i = 2,3 defines a singular matrix, then $p_{ii} = d_i\ (1 + 10^{-3})$. The perturbation is arbitrary but it was found by trying other values for it that it did not affect significantly the values of H.

The chosen values of p_{22} and p_{33} give the smallest eigenvalues of (p_{ij}). Even so they range from 0,35 D-02 to 0.11 D+5.

They were calculated using the QR factorization with Householder transformation.

The elements of (p_{ij}) are so organized that maximum accuracy can be obtained with this method.

The parameter λ_0 was chosen equal to 10^{-2}. It is not possible to select a larger value (closer to 1) which would define a better norm matrix.

The elements of B_2 are computed following the formulae in [7] .

In the next page the values of the calculated elements of B and H are listed.

NORM MATRIX

0.10000D-01	0.0	0.0	0.0	0.0
0.0	0.81447D+01	-0.11908D+02	0.10481D+02	-0.41676D+01
0.0	-0.11908D+02	0.20261D+02	-0.16983D+02	0.76273D+01
0.0	0.10481D+02	-0.16983D+02	0.14767D+02	-0.60338D+01
0.0	-0.41676D+01	0.76273D+01	-0.60338D+01	0.32855D+01

MATRIX B

-0.50000000D+00	0.19566667D+01	-0.29050000D+01	0.19300000D+01	-0.48166667D+00
-0.19566667D+01	0.77803139D-12	0.19156944D+01	0.14450000D+01	-0.14040278D+01
0.29050000D+01	-0.19156944D+01	0.10451350D-11	-0.48704167D+01	0.38811111D+01
-0.19300000D+01	-0.14450000D+01	0.48704167D+01	0.17361636D-12	-0.14120833D+01
0.48166667D+00	0.14040278D+01	-0.38811111D+01	0.14120833D+01	0.11460740D-16

4th order scheme

The eigenvalues of the norm matrix are :

0.1 D-1

0.59493 D-1

0.38777

0.11984 D 1

0.44813 D 2 .

The equivalence relation between the L_2-norm and the H-norm is then

$$1.0\ D\text{-}1\ \ L_2(\bar{v}) \leq H(\bar{v}) \leq 6.694251\ \ L_2(\bar{v})$$

The error committed while calculating the coefficients is illustrated by the result of applying $(Q - \frac{\partial}{\partial x})$ to the monomials 1, x, x^2, x^3 . It is largest at the point x_0 and is of order $0(10^{-12})$.
The truncation error constants result from $Q\bar{x}^4 - \frac{\partial}{\partial x}\bar{x}^4 = 4!\ \bar{c}$ $(\bar{x} = x_0, x_1 \ldots)$, $\bar{c} = (c_0, c_1, \ldots c_4)$. The values of c_i are:

-47.916667
-0.83904338
4.6561708
4.5625596
-4.1831888 .

These values compare very unfavorably with for example the truncation error constants for the 3rd order scheme studied in the last section. In the next section, results of computations of the solution of the scalar equations $u_t = \pm u_x$ with the above defined method are shown.

2) Another possibility of defining the third order modification for the boundary points is to deduce the elements of the norm matrix H and

of B_2 directly from the accuracy conditions. Again there are 3 degrees of freedom in the resulting system. λ_0 was chosen and two of the unknowns were predetermined. For none of the sets of values tried, the matrix H resulted positive definite. When we defined the accuracy conditions with a discretization of x^j at different points, for example with $e^j = (0, 1, \ldots)^i$, then for the values of the parameters tried there were less negative eigenvalues but still we were not able to find a set of values defining a positive definite norm matrix H.

3.2 Numerical Results.

Scalar equations.

We consider first the equation $u_t = -u_x \quad x \in [0,1]$, with initial value function $\sin 4\pi x$. In a manner already discussed in the third order scheme case, for the right hand boundary points $x_{N-4}, \ldots x_N$, the operator Q approximation to $\frac{\partial}{\partial x}$ has to be separated in its symmetric and antisymmetric parts in norm H. The solution at these points is then given implicitly, whereas at the left hand boundary points x_1, x_2, x_3, x_4, the solution is explicitly calculated.

In any case, the system for the gridfunction values of $X_{N-4}, \ldots x_N$ is simple, involving only the values $u(x_\nu)$ and $u(x_N)$ in each equation,

$\nu = N-4, \ldots N.$

An estimate for the maximum stable value of λ for the above problem with $h = \frac{1}{20}$ was obtained from the L_2-norm of $\| HQ_1 \|$. It is

$$\lambda_{max} = \frac{1}{8.04} .$$

The results of the test runs were:

$\lambda = 0.25$ D-1 , $h = 0.5$ D-1 , $k = 0.00125$

t	$\|\bar{e}(t)\|_2$	$\|\bar{v}(t)\|_2$	$\|\bar{v}(t)\|_H$
0.0025	0.37106 D-1	0.70791	0.68682
0.25	0.78022 D-1	0.68625	0.68185
0.5	0.12351	0.67065	0.67127
1	0.1173	0.633	0.6239
2	0.15692	0.59104	0.59226

Figure 8 in Appendix A shows a plot of the solution after 1 time period.

The equation $u_t = u_x$ was resolved for $x \in [0,1]$. Here the grid-function at left hand boundary points was calculated with the third order approximation to $\frac{\partial}{\partial x}$.

The function values at the right hand boundary were calculated using the 4^{th} order interior scheme and the necessary values outside of the interval were defined by the analytic solution.

The results obtained using the initial value function $\sin 2\pi x$ are

$\lambda = 0.25$ D-1 $\quad h = 0.5$ D-1 $\quad k = 0.125$ D-2

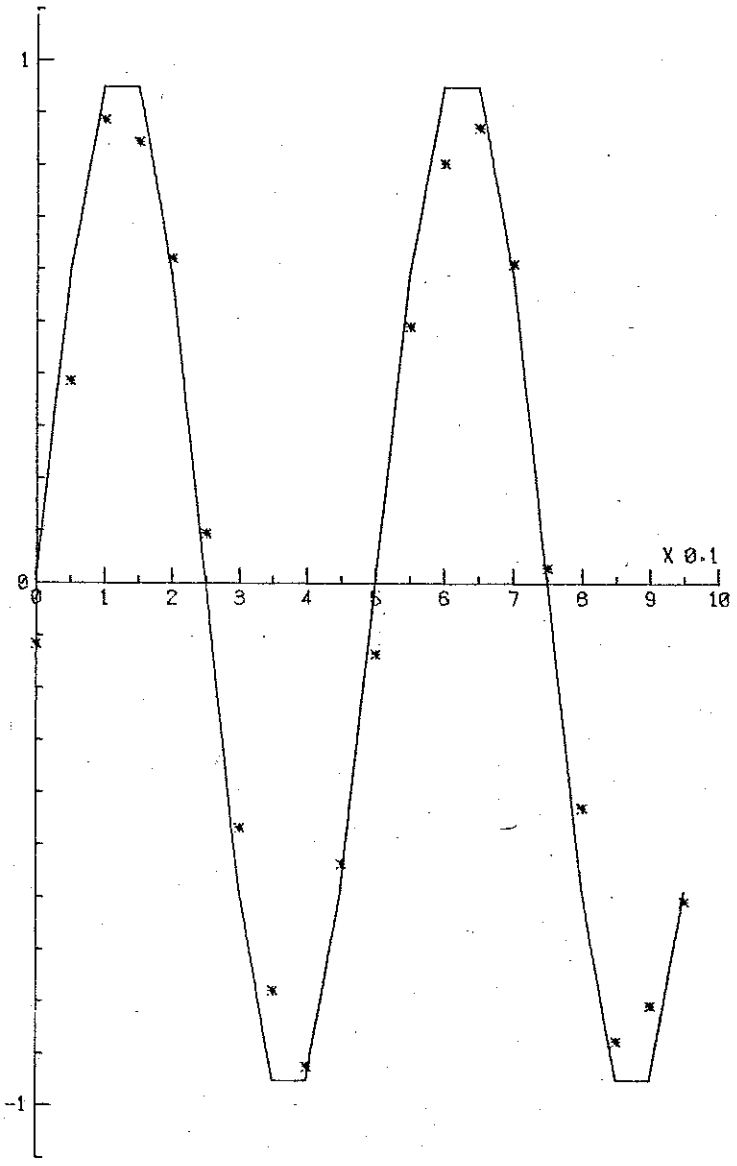
t	$\|\bar{e}(t)\|_2$	$\|\bar{v}(t)\|_2$
0.0025	0.1427 D-2	0.70714
0.25	0.57978 D-2	0.70584
0.5	0.51152 D-2	0.70678
1	0.70014 D-2	0.71007
2	0.83883 D-2	0.71364

Fig.9 shows a graphic display of the solution after 1 period in time.

Appendix A

The analytic solution (represented by the continuous line) versus the calculated (asteriks) of some of the test examples described are plotted.

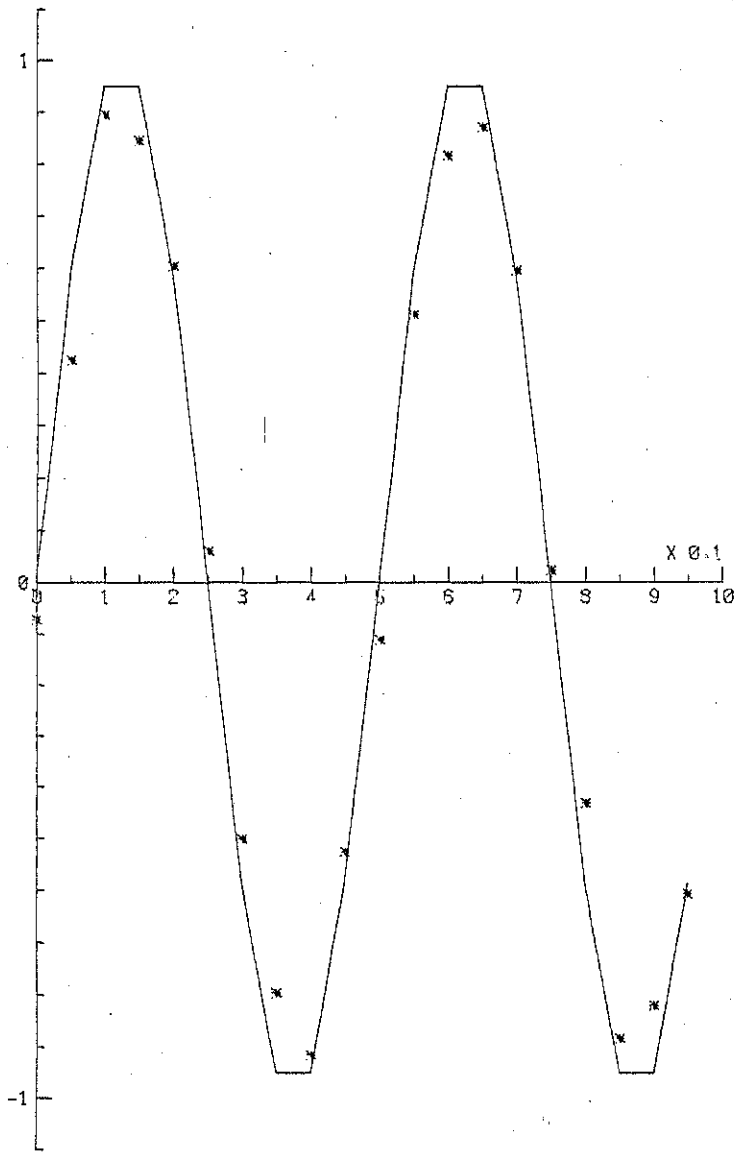
All solutions are bounded by 1. The ordinate axis is subdivided into 10 subintervals to simplify the estimate of the error.



- SOLUTION AT T = 4.000

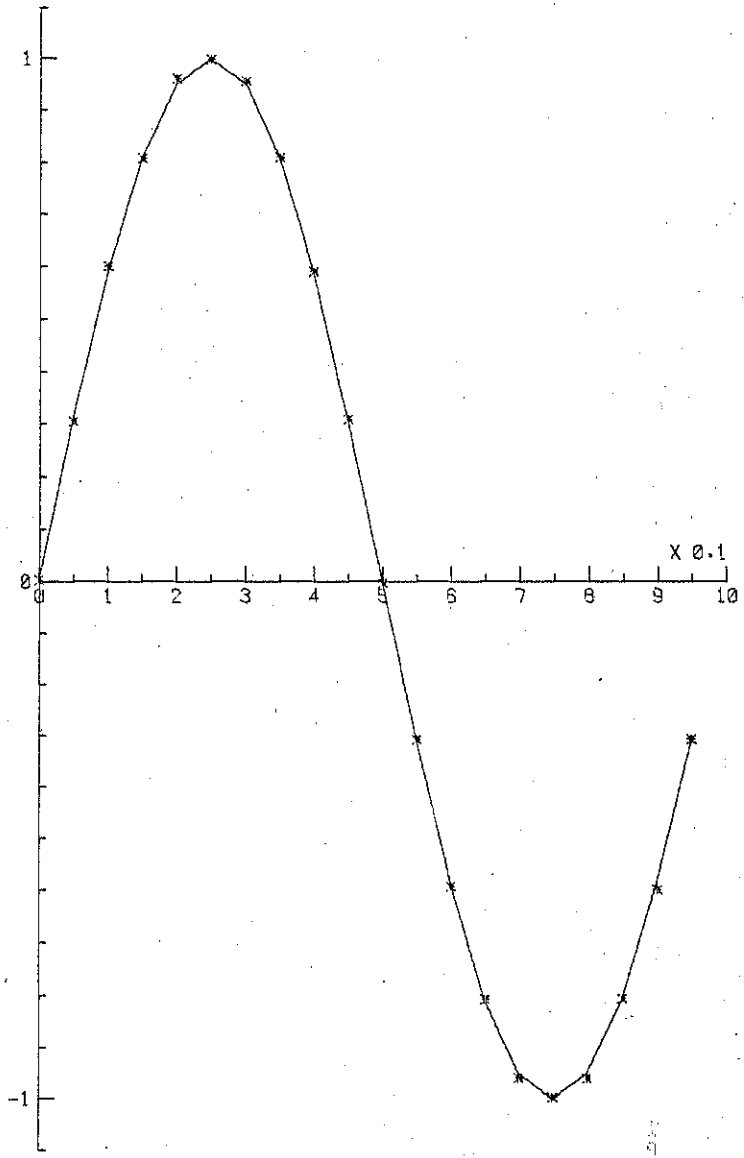
FIG. 1

A2



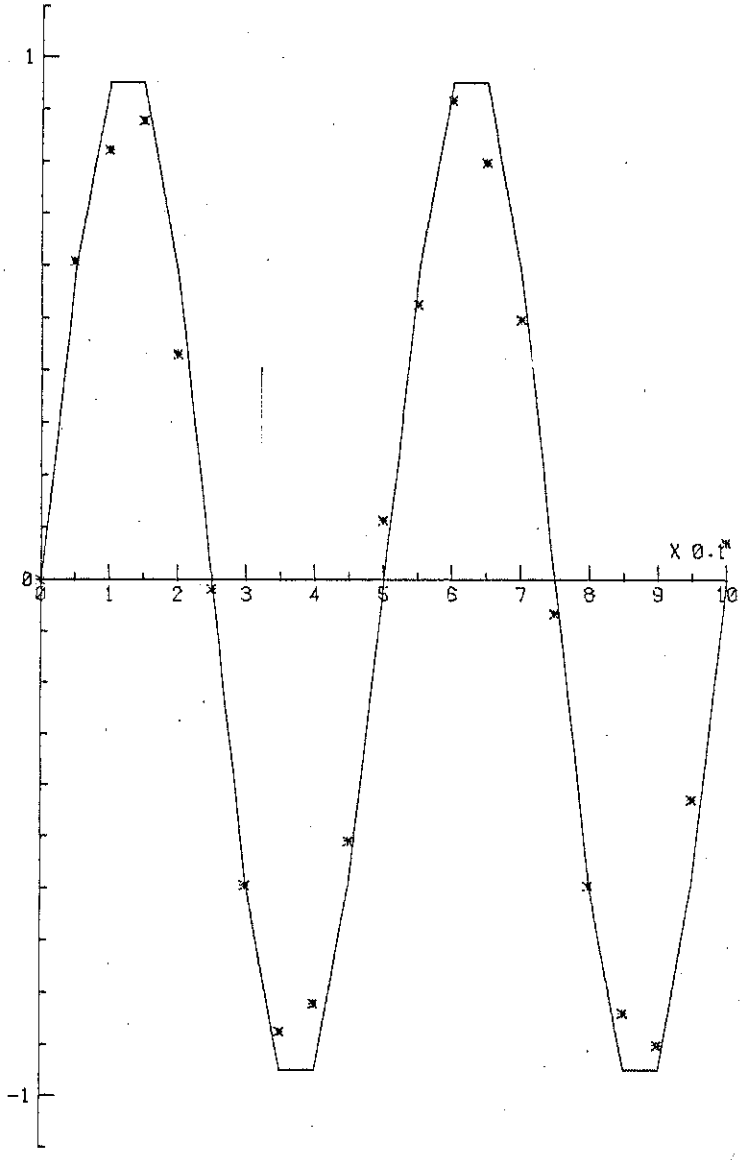
SOLUTION AT T = 5.000

FIG. 2



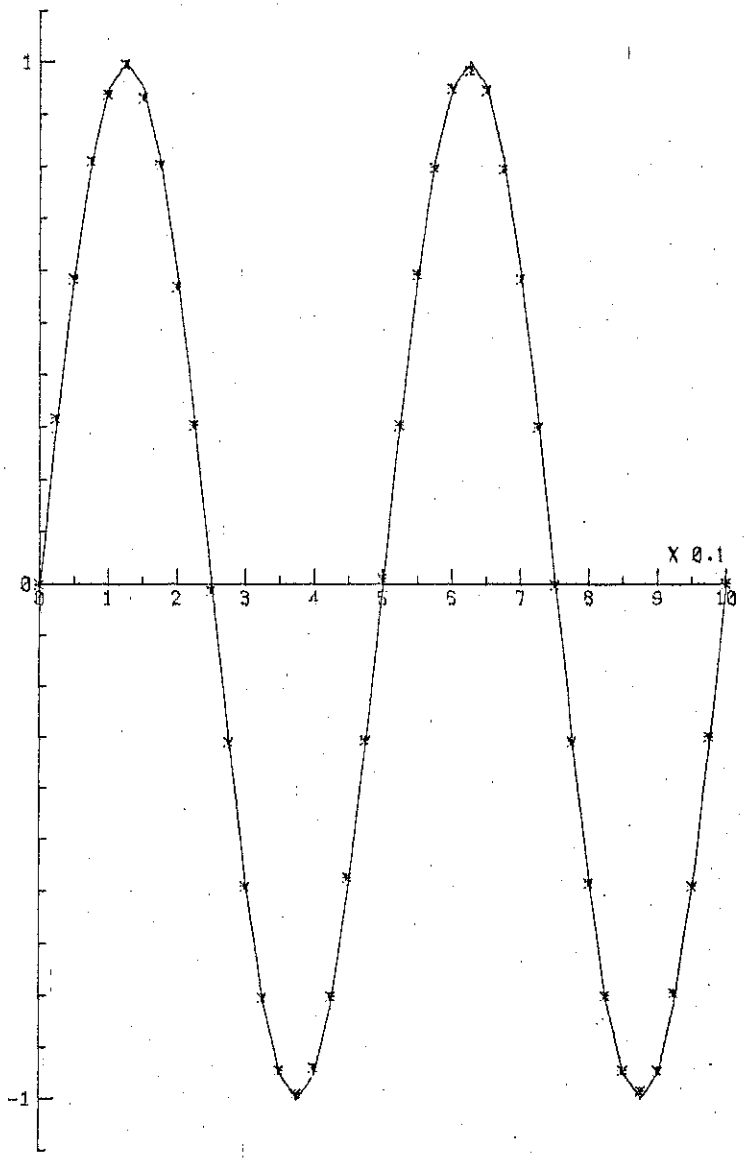
SOLUTION AT T = 5.000

FIG. 3



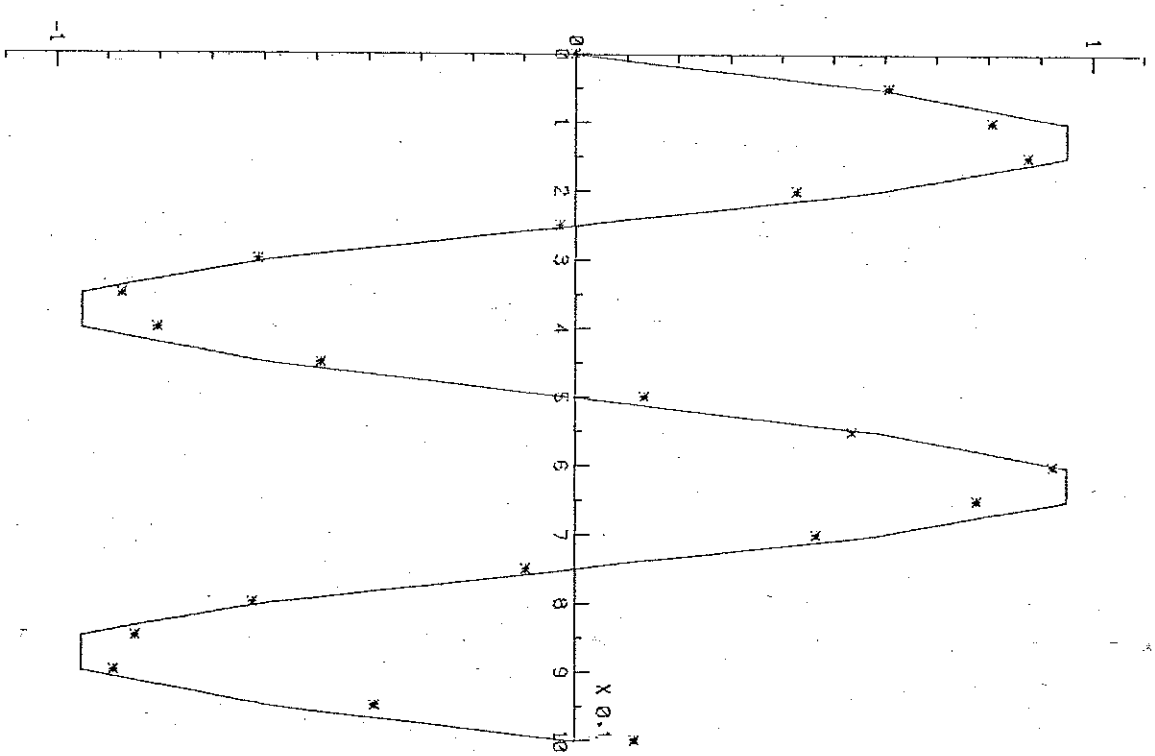
SOLUTION AT T = 4.000

FIG. 4



SOLUTION AT T = 4.000

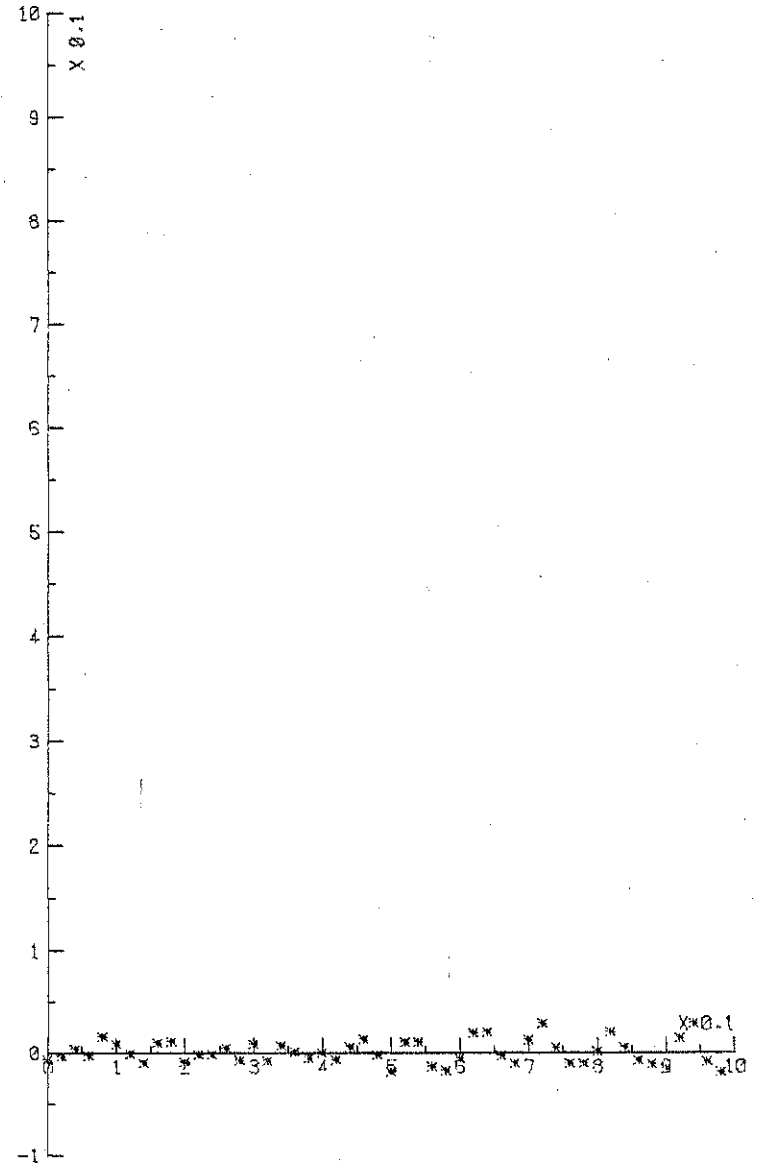
FIG. 5



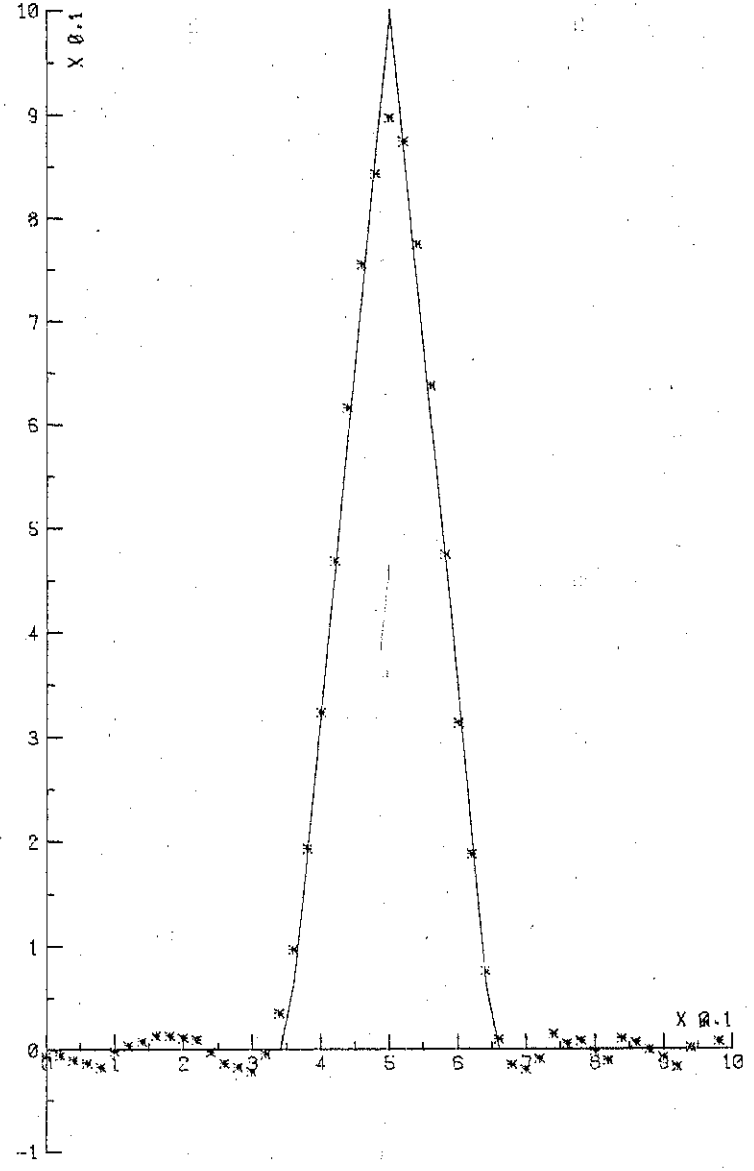
SOLUTION AT T = 4.000

Fig. 6

A5



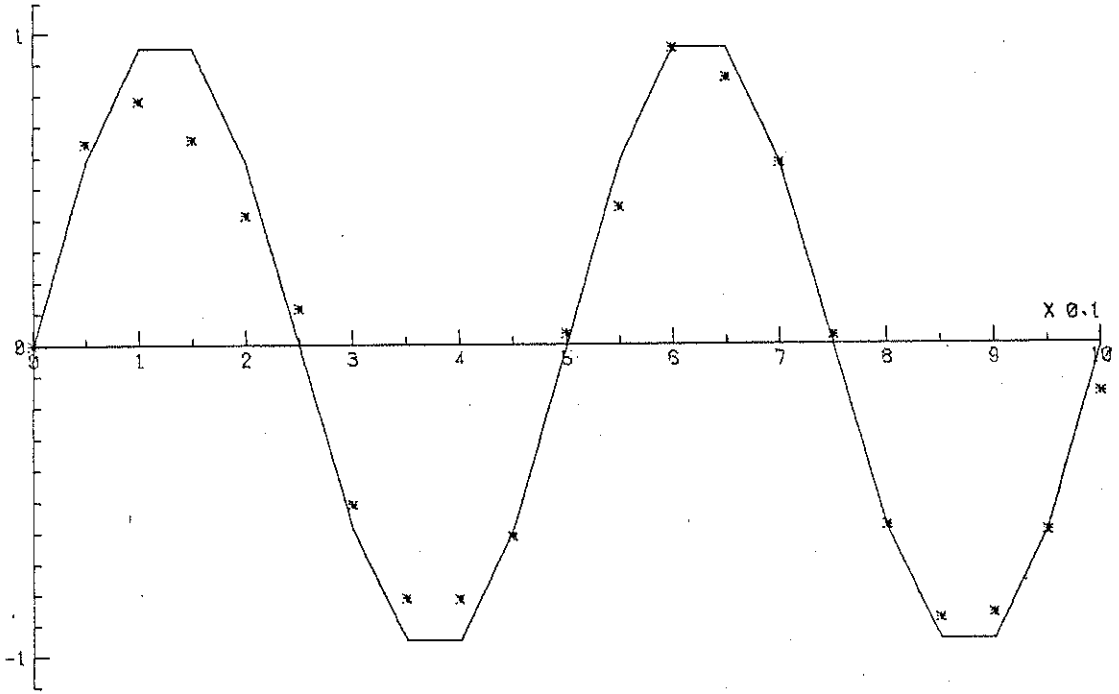
SOLUTION AT T = 3.000



SOLUTION AT T = 3.000

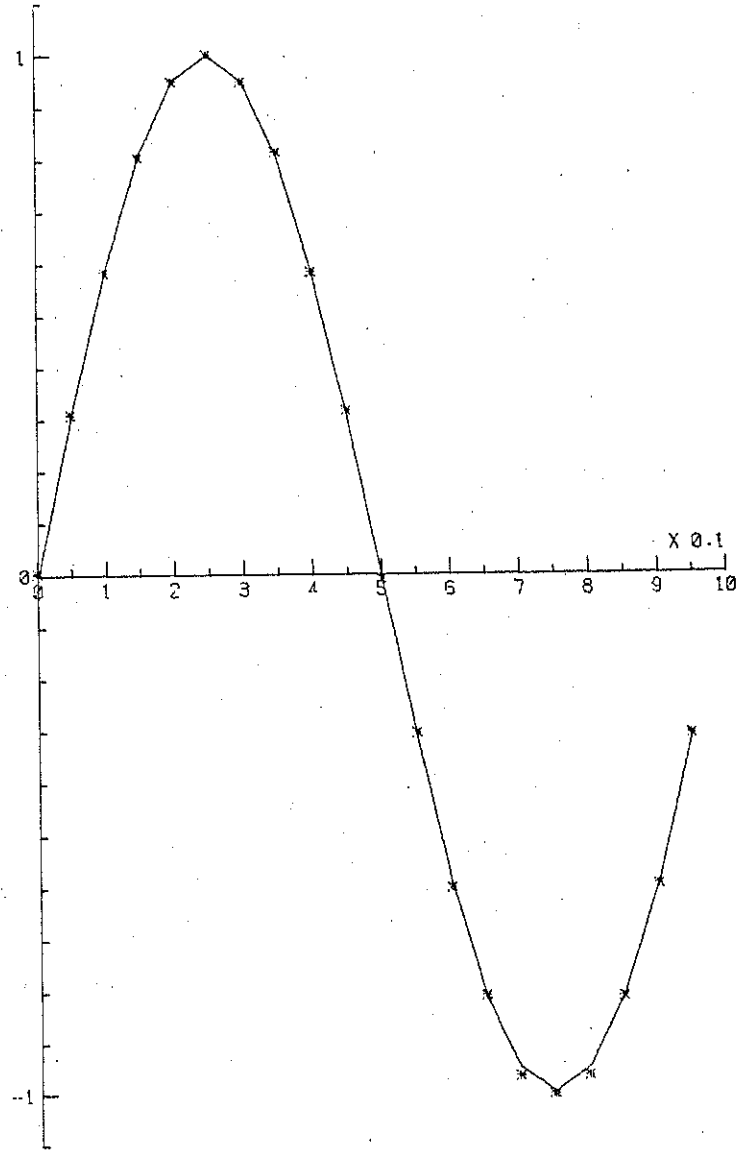
FIG 7

A6



SOLUTION AT T = 1.000

Fig 8



SOLUTION AT T = 1.000

Fig. 9

Acknowledgements

I wish to express my deep gratitude to Professor H-O Kreiss for many hours of inspiring and encouraging discussions.

Bibliography

(1) "Mesh refinement". Browning G., Kreiss H-O, and Oliger J. Math of Comp. (73) v. 27

(2) "Computationally efficient schemes and boundary conditions for a fine-mesh barotropic model based on the shallow water equations." Elvius T. and Sundström A. Tellus XXV 1973,2.

(3) "On high order approximations of hyperbolic partial differential equations by a Fourier Method" Fornberg B. Rep #39 Dept. of Computer Sc. Uppsala University 1972

(4) "The convergence rate for difference approximations to mixed initial boundary value problems" Gustafsson B. Math. of Comp. vol. 29, #130, April 1975

(5) "Methods for the approximate solution of time dependent problems" Kreiss H-O and Oliger J. Garp Publ. Series #10, Feb. 73

(6) "Finite elements and finite difference methods for hyperbolic PDE". Kreiss H-O and Scherer G. Math. Aspects of finite elements in PDE. Academic Press 74.

(7) "On the existence of enery estimates for difference approximations for hyperbolic systems." Kreiss H-O and Scherer G.

(8) "Fourth order difference methods for the initial boundary value problem for hyperbolic equations." Oliger J. Math. of Comp. vol. 28, 1974, #125

(9) "Solution of Vandermonde systems of equations". Björck A. and Pereyra V. Publ. 70.02. Dept. of Computer Sc. UCV,1970.