Journal of Engineering Mathematics, Vol. 7, No. 1, January 1973
Noordhoff International Publishing - Leyden
Printed in The Netherlands

On the construction of accurate difference schemes for hyperbolic partial differential equations

P. WESSELING

National Aerospace Laboratory NLR, Amsterdam, The Netherlands ★

(Received January 19, 1972 and in revised form June 2, 1972)

SUMMARY

Methods are developed for increasing the fidelity of difference approximations to hyperbolic partial differential equations. A relation between the truncation error and the exact and approximate amplification factors is derived. Based upon this relation, quantitative criteria for the minimization of dissipation and dispersion are derived, and difference schemes which satisfy these criteria are constructed. Completely new schemes, one of them promising, are obtained, together with several well-known schemes. One of these is the Fromm scheme, for which previously only a heuristic derivation could be given. It is shown that in general the accuracy of the Rusanov–Burstein–Mirin scheme is disappointing. A simple modification was found to remedy this deficiency.

1. Introduction

Although we have come a long way towards the realization of L. F. Richardson's dream of accurate long term "weather forecasting by numerical process" [14], it has become increasingly evident, that application of the computer to the solution of complicated initial-boundary-value problems, such as occur in fluid mechanics, can fulfill its promise of bringing a larger class of problems within our reach only if the numerical methods to be employed are designed very carefully. The discovery of the Courant–Friedrichs–Lewy stability condition [3] eliminated one of the causes of failure of Richardson's pioneering calculations. But stability is only one of the properties which a good finite difference scheme should have. Several authors have pointed out, that numerical dissipation and dispersion should be kept small, and have constructed difference schemes that more or less fulfill this requirement [5], [6], [7], [9], [10], [16]. It has also been shown that it may be advantageous to employ so-called conservative difference schemes, i.e. schemes which conserve some quantity that is an approximation to a quantity, like mass or momentum, which is conserved by the exact equations [1], [5], [6], [8], [13]. Finite difference methods are never completely free from dissipation and dispersion. Perhaps Galerkin methods, which may be made free of dispersion, are to be preferred above finite difference methods for many applications; see e.g. [12]. In order to obtain insight into this matter it is necessary to investigate to what extent dissipation and dispersion can be suppressed in finite difference methods, and here the present paper attempts to contribute.

The design of difference schemes with little numerical dissipation and dispersion is hampered by the fact, that one does not know exactly in what way dissipation and dispersion should be minimized. Dissipation and dispersion must be minimized simultaneously, but it is not known quantitatively what weights should be attached to the dissipation and the dispersion, respectively, in a minimization process. The aforementioned authors mainly give qualitative discussions of the dissipation and dispersion properties of various difference schemes. In the following, a constructive principle, which has also been used by Miranker [10], is used in the design of difference schemes with little dispersion and dissipation.

Before proceeding, it is perhaps useful to point out what is new in this paper as compared to [10]. As in [10], several well-known schemes are obtained, but also new ones, one of them perhaps being an attractive alternative to the much used Lax–Wendroff scheme. Furthermore, the Fromm scheme, for which until now only a heuristic derivation was available, is shown to follow from an application of the constructive principle just mentioned. It is shown to what extent and why the accuracy of the only third order predictor-corrector scheme in existence for

★ Present address: Technological University Twente, Enschede, The Netherlands.

hyperbolic systems, the Rusanov–Burstein–Mirin scheme [2], [17] is disappointing. The accuracy of this scheme is greatly improved by a simple modification, which preserves the predictor-corrector form.

2. Dissipation, dispersion and truncation error

The discussion will be limited to the following initial-value problem:

$$\frac{\partial \varphi}{\partial t} + \frac{\partial (u\varphi)}{\partial x} = 0\,, \qquad -\infty < x < \infty\,,\ t \geqq 0$$
$$u = u(x)\,,\ \varphi(0, x) \text{ given}\,. \tag{2.1}$$

This conservation equation is a realistic model for the advection-operator in fluid mechanics, and has been used as such by many authors in search of accurate difference methods for fluid mechanical problems. Moreover, Eq. (2.1) furnishes a particularly severe test for the dissipation and dispersion properties of difference schemes, because dissipation and, when u is constant, dispersion are absent from (2.1).

When discussing the dissipation and dispersion properties of difference schemes it seems best to eliminate the troublesome difficulties that are connected with the application of boundary conditions to partial difference equations, especially the appearance of non-harmonic normal modes. Therefore in (2.1) the range of x is chosen to be $(-\infty, \infty)$.

For simplicity, the number of unknowns is restricted to one, and only explicit two time-level difference schemes are considered. However, the reasoning which follows is more generally applicable, and a difference scheme with an arbitrary number of unknowns is considered later in this paper.

A general expression for explicit two time-level difference schemes which approximate (2.1) is:

$$\varphi_j^{n+1} = \sum_{k_1}^{k_2} a_k \varphi_{j+k}^n\,, \tag{2.2}$$

where $\varphi_j^n = \varphi(n\Delta t, j\Delta x)$. This scheme has a difference molecule consisting of $(k_2 - k_1 + 2)$ points, and will be referred to as a $(k_2 - k_1 + 2)$-point scheme. Its solution is denoted by $\varphi_h(t, x)$.

With u assumed constant, the amplification factor g of Eq. (2.1) is defined in the customary way as the ratio of the amplitude of a harmonic wave $\varphi = a(t)\exp(ivx)$ at time $t + \Delta t$ and time t. Thus, one finds:

$$g(\theta) = \exp(-ic\theta)\,, \tag{2.3}$$

where $c = u\Delta t/\Delta x$ is the Courant number, and $\theta = v \cdot \Delta x$. The amplification factor g_h of the difference scheme (2.2) is found to be:

$$g_h(\theta) = \sum_{k_1}^{k_2} a_k \exp(ik\theta)\,. \tag{2.4}$$

The dissipation $\varepsilon(\theta)$ and dispersion $\alpha(\theta)$ of a wave with wave-number $\theta/\Delta x$ are defined, respectively, as:

$$\varepsilon(\theta) = |g| - |g_h|\,, \quad \alpha(\theta) = \arg g - \arg g_h\,. \tag{2.5}$$

Assume that at time $t = \bar{t}$ $\varphi(\bar{t}, x)$ possesses a Fourier-transform, and that $\varphi_h(\bar{t}, x) = \varphi(\bar{t}, x)$. Denoting Fourier-transforms by the symbol $\hat{\ }$ we have:

$$\varphi(\bar{t} + \Delta t, x) = \frac{1}{\Delta x}\int_{-\infty}^{\infty} \hat{\varphi}(\theta/\Delta x, \bar{t}) g(\theta) \exp(i\theta x/\Delta x) d\theta\,, \tag{2.6}$$

$$\varphi_h(\bar{t} + \Delta t, x) = \frac{1}{\Delta x}\int_{-\infty}^{\infty} \hat{\varphi}(\theta/\Delta x, \bar{t}) g_h(\theta) \exp(i\theta x/\Delta x) d\theta\,. \tag{2.7}$$

According to Parseval's equality, the truncation error $\varphi_h - \varphi$ satisfies the following equation:

$$\int_{-\infty}^{\infty} \{\varphi_h(\bar{t}+\Delta t, x)-\varphi(\bar{t}+\Delta t, x)\}^2\, dx =$$

$$(2\pi/(\Delta x)^2) \int_{-\infty}^{\infty} |\hat{\varphi}(\theta/\Delta x, \bar{t})|^2 |g(\theta)-g_h(\theta)|^2\, d\theta \tag{2.8}$$

Using the L_2-norm as a measure for the truncation error, Eq. (2.8) shows, that for minimum error one should minimize $\|g-g_h\|$, defined as

$$\|g-g_h\|^2 = \int_{-\infty}^{\infty} \rho(\theta/\Delta x)\, |g(\theta)-g_h(\theta)|^2 d\theta\,, \tag{2.9}$$

where the weight function ρ should be equal to the square of the modulus of the Fourier-transform of the exact solution at time t. This defines the optimum way in which to diminish dissipation and dispersion.

The above procedure is not of any practical use, because at every instant $t = n\Delta t$ the Fourier-transform of the exact solution must be determined, and the minimization process carried out. Obviously, it is much easier to obtain the solution analytically. However, it will be shown that interesting and useful results may be obtained with a fixed weight function, chosen *a priori*. Other applications of this idea are given in [10].

3. The choice of the weighting function

The obvious way to improve the fidelity of a difference scheme, apart from decreasing the mesh-size of the computational grid, is to increase the order of consistency. It is of interest to inquire to what weight function this corresponds. For schemes of type (2.2), the highest possible order of consistency is, in general, $k_2 - k_1$, or higher in exceptional cases. The relation between the order of consistency and the choice of the weight function $\rho(v)$ for schemes of type (2.2) is given by the following theorem.

Theorem. *For stable difference schemes, a necessary condition* for $\|g-g_h\|$ *to be a minimum with weight function* $\rho(\theta/\Delta x) \equiv \delta(\theta/\Delta x)$ *is that the difference scheme has the highest possible order of consistency.*

Proof. Because g and g_h consist of exponentials, $g-g_h$ can be represented by a power series in θ which is uniformly convergent in an arbitrary finite domain $|\theta| \leqq \Theta$. Consider two difference schemes, with orders of consistency M and N, $M < N$, and let $g-g_h$ be given by, respectively,

$$(g-g_h)_M = \sum_0^{\infty} b_m \theta^m\,, \quad (g-g_h)_N = \sum_0^{\infty} c_m \theta^m\,. \tag{3.1}$$

By definition, $b_m = 0$ for $0, 1, 2, \ldots, M$, $b_{M+1} \neq 0$, $c_m = 0$ for $m = 0, 1, 2, \ldots, N$, $c_{N+1} \neq 0$. The following sequence of weighting functions is chosen:

$$\rho_n(\theta/\Delta x) = (n/\Delta x \pi^{\frac{1}{2}}) \exp\left[-(n\theta/\Delta x)^2\right]. \tag{3.2}$$

This sequence belongs to the class of sequences by means of which the delta-function may be defined. One may write:

$$\|g-g_h\|_N^2 - \|g-g_h\|_M^2 = I_\Theta + I_\infty\,, \tag{3.3}$$

with

$$I_\Theta = (n/\Delta x \pi^{\frac{1}{2}}) \int_{-\Theta}^{\Theta} \exp(-(n\theta/\Delta x)^2) \left(\left| \sum_{N+1}^{\infty} c_m \theta^m \right|^2 - \left| \sum_{M+1}^{\infty} b_m \theta^m \right|^2 \right) d\theta\,, \tag{3.4}$$

and

$$I_\infty = (n/\Delta x \pi^{\frac{1}{2}}) \left\{ \int_{-\infty}^{\Theta} + \int_{\Theta}^{\infty} \right\} \exp(-(n\theta/\Delta x)^2)(|g-g_h|_N^2 - |g-g_h|_M^2)\, d\theta\,. \tag{3.5}$$

It will be shown that $I_\Theta + I_\infty$ becomes negative for n large enough. One may write:

$$\left| \sum_{N+1}^{\infty} c_m \theta^m \right|^2 - \left| \sum_{M+1}^{\infty} b_m \theta^m \right|^2 = \sum_{2M+3}^{\infty} d_m \theta^m - |b_{M+1}|^2 \, \theta^{2M+2} . \tag{3.6}$$

The series $\Sigma_{2M+3}^{\infty} d_m \theta^m$ is again uniformly convergent for $\theta \leqq \Theta$. One has

$$(n/\Delta x \pi^{\frac{1}{2}}) \int_{-\Theta}^{\Theta} \exp(-(n\theta/\Delta x)^2) \sum_{2M+3}^{\infty} d_m \theta^m d\theta \leqq$$

$$\leqq (n/\Delta x \pi^{\frac{1}{2}}) D \int_{-\infty}^{\infty} \exp(-(n\theta/\Delta x)^2) \theta^{2M+4} d\theta =$$

$$= 1 \cdot 3 \cdot 5 \ldots (2M+3) 2^{-M-2} (\Delta x/n)^{2M+4} D , \tag{3.7}$$

where $D = \sum_{M+2}^{\infty} |d_{2m}| \Theta^{2m-2M-4}$. Furthermore,

$$(n/\Delta x \pi^{\frac{1}{2}}) \int_{-\Theta}^{\Theta} \exp(-(n\theta/\Delta x)^2) |b_{M+1}|^2 \theta^{2M+2} d\theta \geqq$$

$$\geqq 1 \cdot 3 \cdot 5 \ldots (2M+1) 2^{-M-1} (\Delta x/n)^{2M+2} [1 - \pi^{-\frac{1}{2}} (\Delta x/n\Theta) \exp(-(n\Theta/\Delta x)^2)] |b_{M+1}|^2 . \tag{3.8}$$

With the aid of (3.7) and (3.8) an upper bound for I_Θ is obtained. An upper bound for I_∞ can be derived as follows. Because stable difference schemes are considered, $|g_h| < 1 + K_1 \Delta t$, where K_1 is some positive constant. Furthermore, $|g| = 1$. Hence, $|g - g_h|_N^2 - |g - g_h|_M^2 < K_2 \pi^{\frac{1}{2}}$, where K_2 is some other constant. It follows that

$$I_\infty < (2K_2 n/\Delta x) \int_{\Theta}^{\infty} \exp(-(n\theta/\Delta x)^2) d\theta < (K_2 \Delta x/n\Theta) \exp(-(n\Theta/\Delta x)^2) \tag{3.9}$$

for n large enough. By substitution of Eqs. (3.7), (3.8) and (3.9) in (3.3) one finds that

$$\|g - g_h\|_N^2 - \|g - g_h\|_M^2 < 0 \tag{3.10}$$

for n large enough, which completes the proof.

When Δx is so small that $\hat{\varphi}(\theta/\Delta x, \bar{t})$ is appreciably different from zero only in a small neighbourhood of $\theta = 0$ one may conclude heuristically from the preceding theorem that difference schemes with maximum consistency will be optimal. However, in practice one wishes for reasons of efficiency to make calculations with as large a step-size as possible. For larger step-sizes, $\hat{\varphi}(\theta/\Delta x, \bar{t})$ will differ from zero in a relatively large neighbourhood of $\theta = 0$. Accordingly, $\|g - g_h\|$ should be minimized with a weight function different from a delta-function.

A discussion will be given of results obtained with the following weight functions:

(i) $\rho_1(v) = \delta(v)$. This weight function results in "classical" schemes with maximum consistency.

(ii) $\rho_2(v) = 1/v^2$. This weight function equals minus the square of the Fourier-transform of the step-function. Difference schemes for which $\|g - g_h\|$ is minimal with this weight function are optimal for the calculation over one time-step, with a step-function as initial condition. This weight function is less heavily concentrated around $v = 0$ than the delta-function.

(iii) $\rho_3(v) = \delta(v - \pi/\Delta x)$. First or second order consistency is imposed as a constraint on the minimization of $\|g - g_h\|$ for four- or five-point schemes, respectively, in order to ensure convergence as $\Delta x \to 0$. Schemes obtained with this weight function give a better resolution of the shortest waves (with $v = \pi/\Delta x$) that can be resolved by the difference scheme, than schemes with maximum consistency.

(iv) $\rho_{n+2}(v) = \delta(v - \pi/n\Delta x)$, $n = 2, 3$. As a constraint on the minimization process, $O(n-1)$ consistency is imposed for $(n+2)$-point difference schemes. The reason behind this choice of weight function is as follows. With a polynomial of degree $(n-1)$ as initial condition, schemes with consistency of order $(n-1)$ give the solution with error zero. Through the function-values at the $(n+1)$ points from which the solution in a given point on the next time-level is calculated

one can always fit a function $f(x)$ consisting of a linear combination of a polynomial of degree $(n-1)$ and a sine or cosine with wavelength $n\Delta x$. It turns out that minimization with weight function ρ_n results in difference schemes that propagate $f(x)$ without error. With a "classical" difference scheme, with consistency $O(n)$, polynomials of $O(n)$ are propagated without error. Because the spectrum of $f(x)$ contains more short wavelength components than the polynomial of $O(n)$ through the given points, one expects that with schemes designed to propagate $f(x)$ without error a more accurate representation of short wavelengths is obtained, though less accurate than schemes obtained with ρ_3 as weight function.

In the next section, 3-, 4- and 5-point difference schemes, obtained by minimization of $\|g-g_h\|$ with the foregoing weight functions will be given, together with results of trial calculations.

4. Difference schemes and trial calculations

Only schemes of the following form are considered:

$$\varphi_j^{n+1} = \sum_{k=-2}^{1} a_k \varphi_{j+k}^n .$$

The difference schemes are of what will be called "characteristic interpolation" type. That is, when u is constant they reduce to

$$\varphi_j^{n+1} = \varphi(n\Delta t, j\Delta x - u\Delta t) , \tag{4.2}$$

and $\varphi(n\Delta t, j\Delta t - u\Delta t)$ is evaluated by interpolation between the points of the difference molecule at the $n\Delta t$ time level. For all schemes obtained in this way, the Courant–Friedrichts–Lewy condition

$$c \leqq 1 \tag{4.3}$$

was found to be sufficient for stability.

For the 4-point schemes ($a_{-2}=0$) the requirement of first order consistency leads to the following relations:

$$a_{-1} = (1+c-a_0)/2 , \quad a_1 = (1-c-a_0)/2 . \tag{4.4}$$

Minimization of $\|g-g_h\|$ with ρ_1, ρ_2, ρ_3 and ρ_4 as weight functions under the restrictions given by (4.4) leads to difference schemes 1 to 4, for which a_0 is given by:

$$\text{scheme 1: } a_0 = (1-c^2) , \tag{4.5}$$

$$\text{scheme 2: } a_0 = (1-|c|) , \tag{4.6}$$

$$\text{scheme 3: } a_0 = \cos^2 \pi c/2 , \tag{4.7}$$

$$\text{scheme 4: } a_0 = \cos \pi c/2 . \tag{4.8}$$

For the minimization with weight factor ρ_1 use is made of the theorem given in section 3. According to this theorem, $\|g-g_h\|$ is minimized by the scheme with the highest possible order of consistency.

With u constant, schemes 1 and 2 are the well-known Lax–Wendroff [8] and Courant–Isaacson–Rees [4] schemes, respectively. It is interesting to note that with error-norm (2.8) the Courant–Isaacson–Rees scheme rather than the Lax–Wendroff scheme is the optimal 4-point scheme for the propagation of a step-function over one time-step.

For the 5-point difference schemes c will be assumed to be positive. The case with c negative may be treated by replacing c by $|c|$ in the following formulae and by interchanging a_1 and a_{-1}, and a_2 and a_{-2}.

Requiring second order consistency, one finds:

$$\left.\begin{aligned} a_{-2} &= (a_0 - 1 + c^2)/3 , \\ a_{-1} &= -a_0 + 1 + c(1-c)/2 , \\ a_1 &= (-2a_0 + 2 - 3c + c^2)/6 . \end{aligned}\right\} \tag{4.9}$$

With ρ_1, ρ_2, ρ_3 and ρ_5 as weight functions, minimization of $\|g-g_h\|$ under the restrictions imposed by (4.9) results in schemes 5 to 8, with a_0 given by, respectively,

$$\text{scheme 5}: a_0 = (1-c/2-c^2+c^3/2)\,, \tag{4.10}$$

$$\text{scheme 6}: a_0 = 1-(3c+c^2)/4\,, \tag{4.11}$$

$$\text{scheme 7}: a_0 = (5-2c^2+3\cos\pi c)/8\,, \tag{4.12}$$

$$\text{scheme 8}: a_0 = (-2+9c-c^2-12\cos(2\pi/3-\pi c/3))/4\,. \tag{4.13}$$

By intuitive reasoning, Fromm [7] has constructed a 5-point difference scheme with reduced dispersion, which he calls a "zero average phase error difference scheme". The phase error is identical to what is called dispersion in the present paper. The expression "zero average phase error" refers qualitatively to the fact, that as a function of Courant number the phase-error has positive as well as negative values. For most difference schemes, the phase error is negative. Scheme 6 turns out to be identical to Fromm's scheme. Hence we have the following result:

Fromm's "zero average phase error" difference scheme is optimal for the calculation over one time-step of the solution with a step-function as initial condition. The qualitative property of small dispersion corresponds to the fact that among all 5-point "characteristic-interpolation" difference schemes Fromm's scheme has the smallest possible value of $\|g-g_h\|$ if the weight function ρ_2 is used.

The 8 schemes defined above are an approximation of Eq. (4.2), and therefore also of

$$\frac{\partial\varphi}{\partial t} + u\frac{\partial\varphi}{\partial x} = 0 \tag{4.14}$$

instead of Eq. (2.1). Therefore, when $u(x)$ is not constant the schemes just given must be modified. A simple modification may be derived as follows. Eq. (2.1) is equivalent to

$$(1+u^2)^{\frac{1}{2}}\frac{\partial\varphi}{\partial s} + \varphi\frac{du}{dx} = 0\,,$$

or

$$\frac{\partial \ln\varphi}{\partial s} = -(1+u^2)^{-\frac{1}{2}}\frac{du}{dx}\,, \tag{4.15}$$

where $\partial/\partial s$ denotes differentiation along the characteristic of Eq. (2.1). From Eq. (4.15) it follows, that

$$\varphi_j^{n+1} = \varphi(n\Delta t\,, j\Delta x - u\Delta t)\exp\left(-\Delta t\frac{du}{dx}\right) + O((\Delta t)^2)\,,$$

or

$$\varphi_j^{n+1} = \Gamma\varphi(n\Delta t\,, j\Delta x - u\Delta t) + O((\Delta t)^2)\,, \tag{4.16}$$

with $\Gamma = 1-\Delta t\,(du/dx)$. Because the schemes derived above approximate Eq. (4.2), they may be made to approximate Eq. (4.16) simply by multiplying the coefficients a_k by Γ. The quantities Γ and c are evaluated at $x=j\Delta x$. In the trial calculations to be described shortly du/dx was evaluated analytically.

When $u(x)$ is variable, with the coefficients a_k multiplied by Γ schemes 1 to 8 are first order consistent in t, whereas the order of consistency in x is 1 for schemes 2, 3 and 4, 2 for schemes 1, 6, 7 and 8, and 3 for scheme 5. For difference schemes with but one unknown it is a simple matter to increase the order of consistency in t and make it equal to that in x, and write the scheme in conservation form, see e.g. Refs. [1], [5], [6], [7]. With more than one unknown, increasing the order of consistency in t, especially beyond 2, usually involves much labour and results in complicated difference schemes. Nevertheless, schemes of higher order consistency in t are used in practice, because of the more favourable balance between accuracy and step-size that one may hope to obtain with these schemes. The Lax–Wendroff scheme is usually employed in its second-order form, conveniently written as a predictor-corrector scheme. A third order conservative predictor-corrector scheme for hyperbolic systems of conservation laws has been

constructed by Rusanov [17]. The same scheme has been derived by Burstein and Mirin [2]. It will be formulated in the next section. It corresponds to minimization with weight function ρ_1.

Trial calculations have been made with the Rusanov–Burstein–Mirin (RBM) scheme, the Lax–Wendroff scheme in the predictor-corrector form given in [15], p. 303, and schemes 2 to 8. The damping coefficient ω in the RBM scheme was assigned in two ways: $\omega=3$ and $\omega=c^2(4-c^2)$. For these trial calculations $u(x)$ was chosen as:

$$u(x)=(a+b\cos^2\pi x)^{-1}\,. \tag{4.17}$$

If initially φ is periodic in x with period 1 the solution is periodic in t with period $p=a+b/2$. This property is used to evaluate the errors in the numerical solution.

The following test-cases were studied:

$$\begin{aligned}
&\text{case 1: } a=1\,, \quad b=1\,, \quad \varphi_0(x)=H(x-\tfrac{1}{2})\,,\\
&\text{case 2: } a=1\,, \quad b=1\,, \quad \varphi_0(x)=\sin^2\pi x\,,\\
&\text{case 3: } a=1.05\,, b=1.9\,, \varphi_0(x)=H(x-\tfrac{1}{2})\,,\\
&\text{case 4: } a=1.05\,, b=1.9\,, \varphi_0(x)=\sin^2\pi x\,,
\end{aligned}$$

where $H(x)$ is the periodic step-function with wavelength 1. For the test-cases listed above, table 1 gives for the various schemes the error norm ε, defined as:

$$\varepsilon=\frac{1}{m}\sum_{k=0}^{m}|\varphi(t,k\Delta x)-\varphi_h(t,k\Delta x)|\,, \qquad (m=1/\Delta x)\,. \tag{4.18}$$

For all calculations, $\Delta t=\Delta x$.

TABLE 1

Average error for test-cases 1 *to* 4

n = number of points in difference molecule.

	n	*Case* 1, $t=1.5$			*Case* 2, $t=1.5$		
Δx, Δt	—	0.025	0.05	0.0625	0.025	0.05	0.0625
Scheme 2	3	0.1851	0.2251	0.2843	0.0979	0.1666	0.1942
Scheme 3	4	0.1951	0.2690	0.2992	0.1096	0.1851	0.2145
Lax–Wendroff	4	0.1595	0.2425	0.2613	0.0268	0.0858	0.1179
Scheme 4	4	0.1411	0.2179	0.2401	0.0336	0.0816	0.1097
Scheme 5	5	0.0963	0.1624	0.1929	0.0055	0.0277	0.0446
Scheme 6	5	0.0965	0.1626	0.1921	0.0057	0.0278	0.0443
Scheme 7	5	0.0973	0.1651	0.1960	0.0059	0.0286	0.0456
Scheme 8	5	0.0963	0.1624	0.1930	0.0055	0.0277	0.0446
RBM $\omega=3$	6	0.1248	0.2118	0.2331	0.0146	0.0577	0.0819
RBM, $\omega=c^2(4-c^2)$	6	0.1019	0.1631	0.1972	0.0029	0.0225	0.0380
		Case 3, $t=2$			*Case* 4, $t=2$		
		0.025	0.05	0.0625	0.025	0.05	0.0625
Scheme 2	3	0.2333	0.2897	0.3174	0.1780	0.2715	0.3112
Scheme 3	4	0.2191	0.2923	0.3267	0.1556	0.2440	0.2832
Lax–Wendroff	4	0.2154	0.3022	0.3135	0.0906	0.1894	0.2532
Scheme 4	4	0.1869	0.2716	0.2892	0.0805	0.1843	0.2338
Scheme 5	5	0.1256	0.2038	0.2278	0.0254	0.0795	0.1123
Scheme 6	5	0.1252	0.2029	0.2257	0.0255	0.0778	0.1113
Scheme 7	5	0.1320	0.2111	0.2367	0.0298	0.0900	0.1235
Scheme 8	5	0.1257	0.2040	0.2282	0.0255	0.0799	0.1125
RBM $\omega=3$	6	0.1820	0.2488	0.2466	0.0609	0.1550	0.1851
RBM, $\omega=c^2(4-c^2)$	6	0.1411	0.2208	0.2618	0.0215	0.0875	0.1183

Several interesting conclusions can be drawn from table 1. In the first place, among the 4-point schemes considered scheme 4 is clearly the most accurate by a fair margin, except when the solution is smooth and the mesh is fine. For the test-cases considered here it is significantly more accurate than the Lax–Wendroff scheme. In figure 1 dissipation and dispersion of scheme

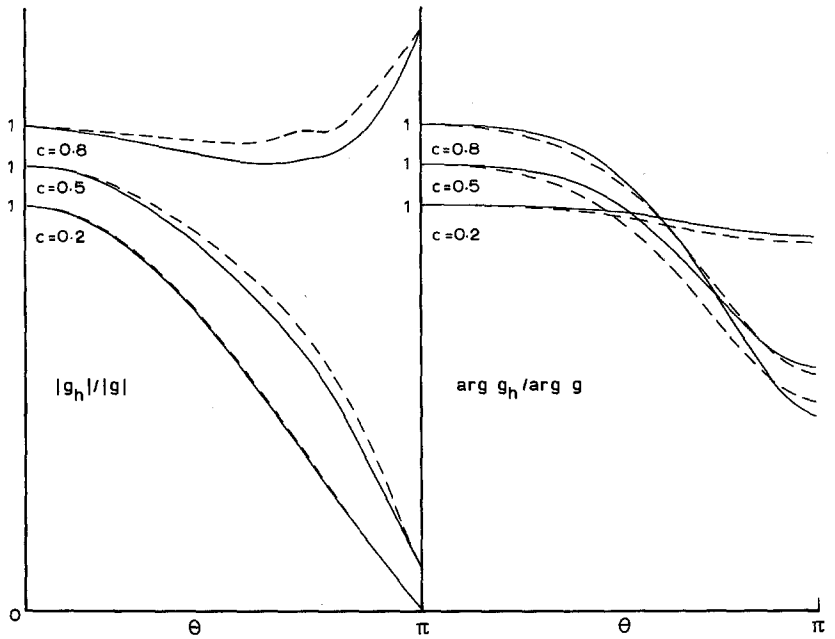

Figure 1. Dispersion and dissipation; ——, Law–Wendroff; – – –, scheme 4.

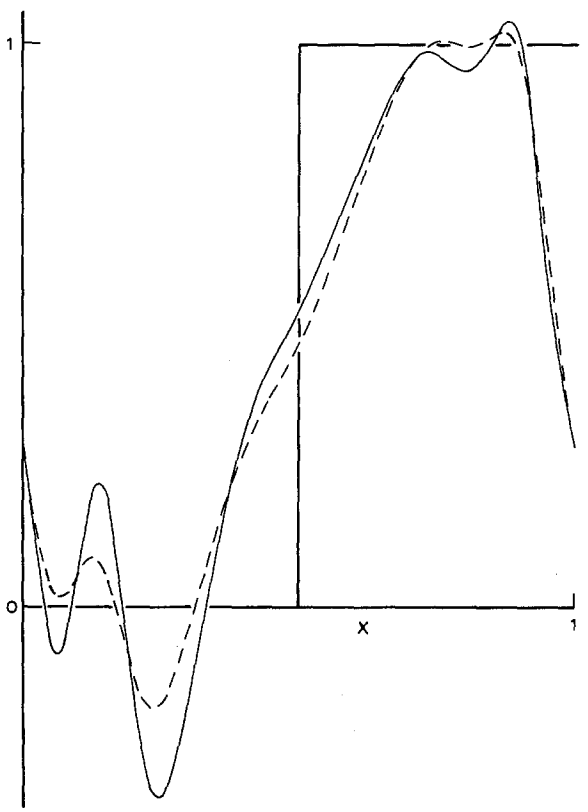

Figure 2. Results for test-case 3. ——, exact solution; ——, Lax–Wendroff; – – –, scheme 4.

4 and the Lax–Wendroff scheme are compared. The figure shows that scheme 4 has more dissipation but less dispersion than the Lax–Wendroff scheme. In figure 2 the solutions for test-case 3 are displayed; the results are qualitatively similar to test-case 1. Oscillations similar to those that are typical for the Lax–Wendroff scheme when the solution is discontinuous are also displayed by scheme 4, but with a smaller amplitude, as is to be expected from the fact that scheme 4 has more dissipation. A variety of stratagems, all embodying the introduction of additional dissipation, have been proposed in the literature for the removal of the oscillations exhibited by the Lax–Weddroff scheme. But in regions where the solution is smooth this additional dissipation usually causes the difference scheme to be less accurate then the original Lax–Wendroff scheme. Scheme 4, however, seems to be more accurate than the Lax–Wendroff scheme for smooth solutions also, as may be seen from the results of test-cases 2 and 4. This is probably due to the fact that the adverse effect of the extra dissipation of scheme 4 is compensated by the better dispersion properties. Figure 3 gives the solutions for test-case 4; the results are similar to those of test-case 2.

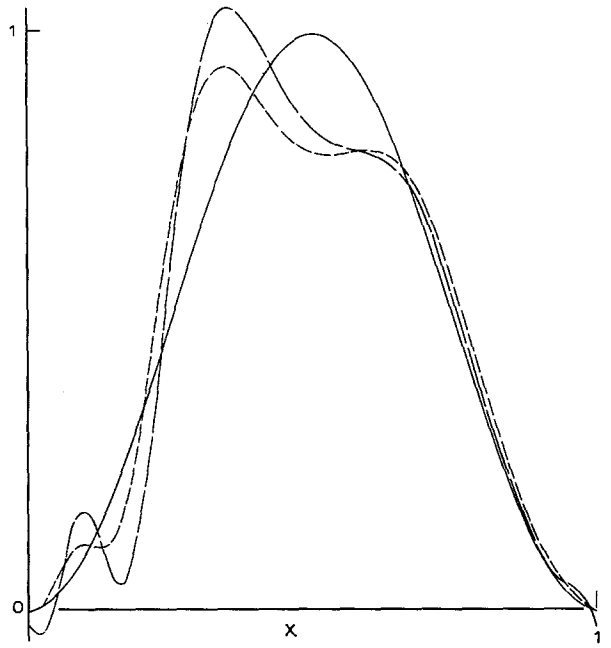

Figure 3. Results for test-case 4. ——, exact solution; — — —, Lax–Wendroff; – – – –, scheme 4.

The model-equation (2.1) with initial condition as in test-cases 1 and 3 gives an accurate representation of the development of a contact-discontinuity in gasdynamics. However, equation (2.1) does not contain the non-linear "steepening" behaviour of the advection-operator in fluid mechanics, and therefore the trial calculations reported in this paper do not accurately simulate the development of a shock-wave. An investigation whether, in the case of the true gasdynamical equations, scheme 4 is more accurate than the Lax–Wendroff scheme, not only for contact-discontinuities but also for shock-waves, would be of interest.

Obviously, because of the appearance of a trigonometric function, scheme 4 requires more computer-time than the Lax–Wendroff scheme. However, it seems likely that the cosine could be replaced by a low-order polynomial without appreciable changes in the results. Furthermore, the fact that scheme 4 consists of a predictor only and does not contain a corrector, is time-saving.

As is to be expected, the results of the 5-point schemes are more accurate than the results of the 4-point schemes. The differences in accuracy between the 5-point schemes are small. Increasing the order of consistency in t gave only a slight improvement of the results. The agreement between the results of these schemes corresponds to the fact, that their amplification

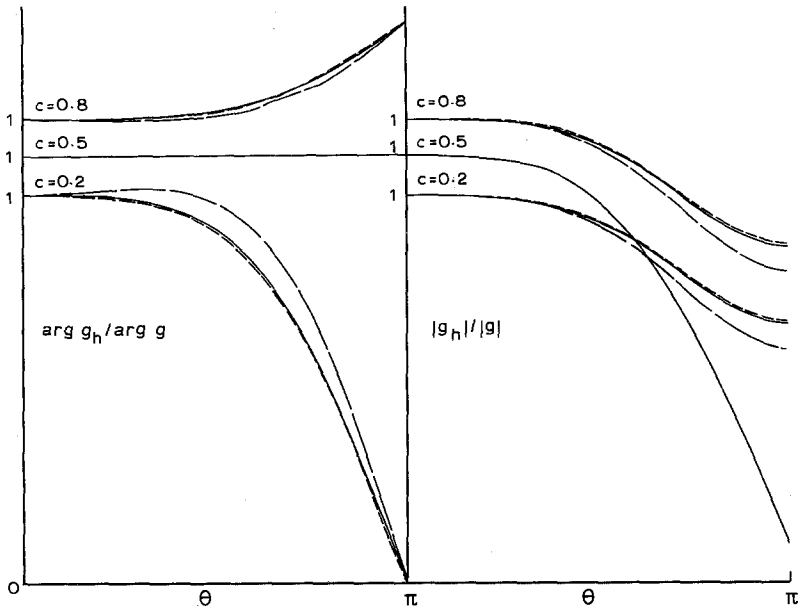

Figure 4. Dispersion and dissipation; ——, scheme 5; — — —, scheme 6 (Fromm); – – –, scheme 8.

factors do not differ much, as shown in figure 4. The results of scheme 7 are slightly worse than the results of the other 5-point schemes, and the results of scheme 3 are worse than the results of the other 4-point schemes. Apparently, the weight function ρ_3 overemphasizes large wave-numbers at the expense of small wave-numbers.

The accuracy of the RBM-scheme is seen to depend strongly on the damping coefficient ω especially when the solution is smooth. It seems that this fact has not yet received sufficient attention. In calculations reported in [2] and [17] more or less arbitrary choices for ω are made. Furthermore, in the case with more than one unknown it can be shown that regardless of the value of ω the accuracy will always be appreciably less than what can be obtained in the scalar case with a proper choice of ω. However, this deficiency can be remedied by a simple modification of the RBM scheme. It seems worthwhile to devote a special section to this matter, in view of the practical importance of the RBM scheme: it is the only predictor-corrector scheme in existence with higher order of accuracy than the Lax–Wendroff scheme.

In the next section it will be argued, that the best choice for ω is $\omega = c^2(4-c^2)$. Table 1 shows, that with this choice of ω the RBM scheme is only slightly more accurate than the 5-point schemes. With $\omega = 3$ the accuracy is significantly worse. With $\Delta x = 0.05$ the results of the 5-point schemes are appreciably better than the results of the RBM scheme with the slightly larger step-size $\Delta x = 0.0625$. This suggests that the most economical way to increase the accuracy may be not to use the RBM scheme, but to develop a predictor-corrector form of one of the 5-point schemes, and use this with slightly diminished step-size.

5. Modification of the Rusanov–Burstein–Mirin-scheme

The RBM scheme ([2], [17]) is an approximation to hyperbolic systems of the following type:

$$\frac{\partial \varphi}{\partial t} = \frac{\partial f(\varphi, x, t)}{\partial x} + g(\varphi. x, t), \tag{5.1}$$

where φ, f and g are vectors. For simplicity, g will be assumed to be zero. The hyperbolicity implies that the eigenvalues of the matrix $F = \partial f/\partial \varphi$ are real and distinct.

The fraction $\tau_1 \Delta t$ of the time-step Δt at which the first predictor is evaluated is an arbitrary parameter in the RBM scheme. However, the amplification factor is found not to depend on τ_1,

so that the choice of this parameter is irrelevant in the present context, and has very little influence on the accuracy. In [2] and [17] calculations are made with $\tau_1 = \frac{1}{3}$. With this choice the RBM scheme is defined as follows:

$$\varphi^{(1)}_{j+\frac{1}{2}} = (\varphi^n_{j+1} + \varphi^n_j)/2 + \sigma(f^n_{j+1} - f^n_j)/3\,, \tag{5.2}$$

$$\varphi^{(2)}_j = \varphi^n_j + 2\sigma(f^{(1)}_{j+\frac{1}{2}} - f^{(1)}_{j-\frac{1}{2}})/3\,, \tag{5.3}$$

$$\varphi^{n+1}_j = \varphi^n_j + \sigma(-2f^n_{j+2} + 7f^n_{j+1} - 7f^n_{j-1} + 2f^n_{j-2})/24 + 3\sigma(f^{(2)}_{j+1} - f^{(2)}_{j-1})/8 - \frac{\omega}{24}\delta^4\varphi^n_j\,, \tag{5.4}$$

where $\sigma = \Delta t/\Delta x$, and $\delta^4\varphi^n$ denotes an undivided fourth difference. The difference scheme is of third order accuracy in x and t. Equations (5.2) to (5.4) may be termed the first predictor, second predictor and corrector, respectively. The advantage of the predictor-corrector formulation is, that the matrix F and its time-derivatives do not enter, thus eliminating several matrix multiplications, which are very time-consuming. The term multiplied by ω in (5.4) is called the damping term; it is necessary to stabilize the scheme. Stability requires ([2], [17]):

$$c_m^2(4 - c_m^2) \leqq 3\,, \tag{5.5}$$

where $c_m = \sigma\lambda_m$, λ_m the absolutely largest eigenvalue of F.

In order to investigate how the damping coefficient ω should be chosen, the dissipation and dispersion properties of the RBM scheme are studied. To this end the linearizing assumption that F is constant is made. There exists a matrix U with the property $UFU^{-1} = \Lambda$, with Λ a diagonal matrix. The RBM scheme can be written in the following form:

$$\begin{aligned}\psi^{n+1}_j = \psi^n_j &+ C(-\psi^n_{j+2} + 8\psi^n_{j+1} - 8\psi^n_{j-1} + \psi^n_{j-2})/12 + \\ &+ C^2(\psi^n_{j+2} - 2\psi^n_j + \psi^n_{j-2})/8 \\ &+ C^3(\psi^n_{j+2} - 2\psi^n_{j+1} + 2\psi^n_{j-1} - \psi^n_{j-2})/12 - \frac{\omega}{24}\delta^4\psi^n_j\,,\end{aligned} \tag{5.6}$$

where $\psi = U\varphi$ and $C = \sigma\Lambda$.

First the case with only unknown (with $C = c$) is considered. Figure 5 gives the dissipation

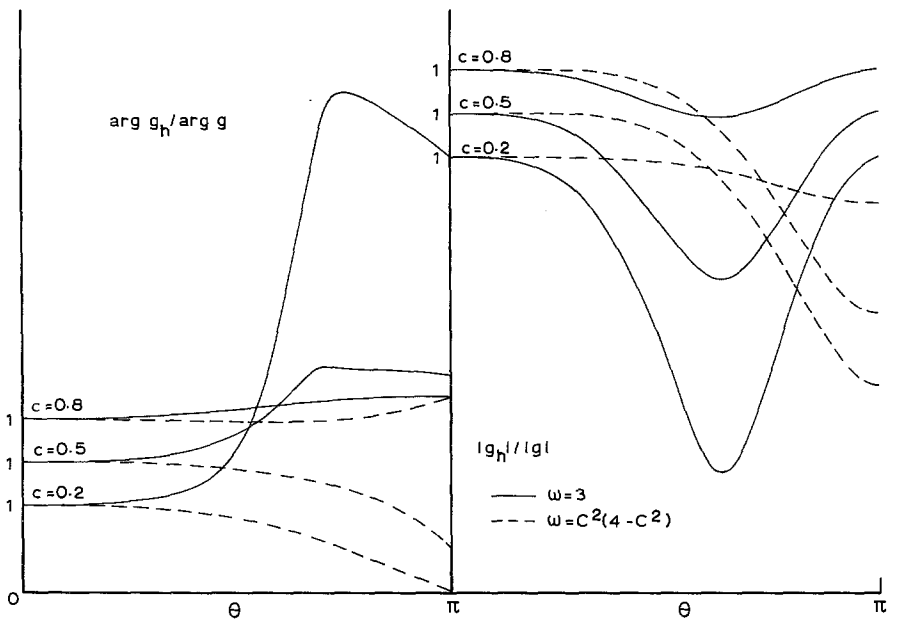

Figure 5. Dispersion and dissipation of the Rusanov–Burstein–Mirin scheme.

and dispersion properties of (5.6) for several values of c with $\omega=3$ and $\omega=c^2(4-c^2)$, the two extremes of the stability condition (5.5). From this figure the reason why the accuracy with $\omega=3$ is so much less than with $\omega=c^2(4-c^2)$ (see table 1) is immediately apparent. With $\omega=3$ the damping for moderate wavenumbers is more, for large wavenumbers less than for $\omega=c^2(4-c^2)$. In view of the fact that the dispersion increases as the wavenumber increases, this is undesirable; wavenumbers with much dispersion should be damped out more quickly than wavenumbers with little dispersion. For $\omega=3$ this is not the case; for $\theta=\pi$, where the dispersion is greatest, damping is even completely absent.

Hence it is clear why the choice of ω has such a large influence on the accuracy. The question arises, whether there is a better choice for ω than $c^2(4-c^2)$. This does not appear to be so. Assigning $\omega=c^2(4-c^2)$ is found to correspond to minimization of (2.9) with a delta-function at zero wavenumber as weighting function (finding out to what weighting function $\omega=3$ corresponds does not seem to be easy). Spreading of the weighting function to non-zero wavenumbers may be expected to have roughly the same influence on a third order 6-point scheme, such as the RBM-scheme, as on the second and third order 5-point schemes discussed in the preceding section. Here it was found that spreading of the weighting-function does not have much influence on the dissipation and dispersion (figure 4), unless high wavenumbers are emphasized very strongly, in which case the scheme becomes inaccurate, see for example the results obtained with scheme 7 (table 1). Furthermore, with $\omega=c^2(4-c^2)$, the RBM scheme is fourth order accurate in x, although still third order in t.

The case with an arbitrary number of unknowns will now be considered. It is clear that giving ω the value $c_m^2(4-c_m^2)$ results in good dissipation and dispersion properties for one component of (5.6) only, namely the one corresponding to the largest element c_m of the diagonal-matrix C. The accuracy of the other components of (5.6) may be as bad as the accuracy of the scalar case with $\omega=3$. Fortunately, there is a remedy which leaves the conservative predictor-corrector form of the RBM-scheme untouched: replace the scalar ω by the matrix $\Omega=\sigma^2\Lambda^2(4E-\Lambda^2)$, with E the identity matrix. This results in optimal damping of each component of (5.6). Eq. (5.4) becomes:

$$\varphi_j^{n+1}=\varphi_j^n+\sigma(-2f_{j+2}^n+7f_{j+1}^n-7f_{j-1}^n+2f_{j-2}^n)/24+ \\ +3\sigma(f_{j+1}^{(2)}-f_{j-1}^{(2)})/8-\tfrac{1}{24}\delta^4 d_j^n,\quad d=U^{-1}\Omega U\varphi\,. \tag{5.7}$$

The matrix multiplications necessary for finding d can usually be done analytically beforehand. For this the eigenvalues and the diagonalizing matrix U are needed. These may be considered known, because in practical application they are needed for a good understanding of the phenomenon under study. Replacement of (5.4) by (5.7) makes the RBM scheme slightly more time-consuming, because in addition to the function evaluations needed to determine f^n, $f^{(1)}$ and $f^{(2)}$ an additional function evaluation to determine d is necessary.

The effectiveness of replacing (5.4) by (5.7) is demonstrated by means of the following model problem, which is a version with two unknowns of the model problem discussed in the previous section:

$$\partial\varphi/\partial t+\partial F/\partial x=0 \tag{5.8}$$

where

$$\varphi=\begin{pmatrix}\varphi_1\\ \varphi_2\end{pmatrix},\quad F=A\varphi\,,\quad A=\begin{pmatrix}c_1 & c_2\\ c_2 & c_1\end{pmatrix},\quad c_{1,2}=(\lambda_1\pm\lambda_2)/2\,,$$

$\lambda_{1,2}=(a_{1,2}+b_{1,2}\cos^2\pi x)^{-1}$, $a_{1,2}$ and $b_{1,2}$ are arbitrary constants. If φ is initially periodic in x with period 1 and if $a_1+\frac{1}{2}b_1=p/k$, $a_2+\frac{1}{2}b_2=p$ with k an arbitrary integer, the solution is periodic in t with period p. The four cases that were calculated are listed in table 2.

The average error in φ_1 and φ_2, called ε_1 and ε_2, respectively, is defined as in Eq. (4.15) and listed in table 3. The results confirm, that with the corrector given by (5.7) the scheme is considerably more accurate than with the corrector given by (5.4), especially when the solution is smooth.

TABLE 2

Definition of test-cases 5 *to* 8

	$\varphi_1(0, x)$	$\varphi_2(0, x)$	a_1	b_1	a_2	b_2
Case 5	$H(x-\frac{1}{2})$	$\cos^2\pi(x-\frac{1}{4})$	1.05	0.9	2.5	1
Case 6	$\sin^2\pi x$	$\cos^2\pi(x-\frac{1}{4})$	1.05	0.9	2.5	1
Case 7	$H(x-\frac{1}{2})$	$\cos^2\pi(x-\frac{1}{4})$	1.4	0.2	2.8	0.4
Case 8	$\sin^2\pi x$	$\cos^2\pi(x-\frac{1}{4})$	1.4	0.2	2.8	0.4

TABLE 3

Average error for test-cases 5 *to* 8

Damping term		*Case* 5, $t=3$		*Case* 6, $t=3$		*Case* 7, $t=3$		*Case* 8, $t=3$	
	Δx, Δt	0.025	0.0625	0.025	0.0625	0.025	0.0625	0.025	0.0625
Eq. (5.4)	ε_1	0.1056	0.2241	0.0023	0.0383	0.0913	0.1868	0.00095	0.0118
	ε_2	0.0268	0.0405	0.0038	0.0428	0.0322	0.0390	0.00107	0.0155
Eq. (5.7)	ε_1	0.1032	0.2157	0.0025	0.0394	0.0610	0.1866	0.000092	0.0036
	ε_2	0.0314	0.0663	0.0022	0.0288	0.0150	0.0303	0.000081	0.0027

REFERENCES

[1] A. Arakawa, Computational design for long-term integration of the equations of fluid motion: Two-dimensional incompressible flow, *J. Comp. Phys.*, 1 (1966) 119–143.

[2] S. Z. Burstein and A. A. Mirin, Third order difference methods for hyperbolic equations, *J. Comp. Phys.*, 5 (1970) 547–571.

[3] R. Courant, K. Friedrichs and H. Lewy, Über die Partiellen Differenzengleichungen der Mathematischen Physik, *Math. Ann.*, 100 (1928) 32–74.

[4] R. Courant, E. Isaacson and M. Rees, On the solution of non-linear hyperbolic differential equations by finite differences, *Comm. Pure Appl. Math.*, 5 (1959) 243–255.

[5] W. P. Crowley, Second-order numerical advection, *J. Comp. Phys.*, 1 (1967) 471–484.

[6] W. P. Crowley, Numerical advection experiments, *Monthly Weather Rev.*, 96 (1968) 1–11.

[7] J. E. Fromm, A method for reducing dispersion in convective difference schemes, *J. Comp. Phys.*, 3 (1968) 176–189.

[8] P. D. Lax and B. Wendroff, Systems of conservation laws, *Comm. Pure Appl. Math.*, 13 (1960) 217–237.

[9] H. Lomax, P. Kutler and F. B. Fuller, The numerical solution of partial differential equations governing advection, AGARDograph 146 (1970).

[10] W. L. Miranker, Difference schemes with best possible truncation error, *Num. Math.*, 17 (1971) 124–142.

[11] C. R. Molenkamp, Accuracy of finite difference methods applied to the advection equation, *J. App. Meteor.*, 1 (1968) 160–167.

[12] S. A. Orszag, Numerical simulation of incompressible flows within simple boundaries: accuracy, *J. Fluid Mech.*, 49 (1971) 75–112.

[13] S. A. Piacsek and G. P. Williams, Conservation properties of convection difference schemes, *J. Comp. Phys.*, 6 (1970) 392–405.

[14] L. F. Richardson, *Weather prediction by numerical process*, London: Cambridge University Press 1922; New York: Dover 1965.

[15] R. D. Richtmyer and K. W. Morton, *Difference methods for initial-value problems*, New York: Interscience Publishers 1967.

[16] K. V. Roberts and N. O. Weiss, Convective difference schemes, *Math. Comp.*, 20 (1966) 272–298.

[17] V. V. Rusanov, Difference schemes of the third order of accuracy for continuous computation of discontinuous solutions, *Soviet Math. Dokl.*, 2 (1968) 771–774.