NUMERICAL ALGORITHMS FOR THE NON-LINEAR SCALAR WAVE EQUATION

M. J. BAINES

NUMERICAL ANALYSIS REPORT 1/83

ABSTRACT

Roe's approach to the numerical solution of the one-dimensional scalar wave equation is surveyed and extended to two and three dimensions with up to third order accuracy.

CONTENTS

		Page
1.	One-dimensional Second Order Algorithms	1
2.	A One-dimensional Third Order Algorithm	9
3.	Statement of the One-Dimensional Algorithms	11
4.	Two-dimensional Second Order Algorithms	13
5.	A Third-order Two-dimensional Algorithm	22
6.	Statement of the Two-dimensional Algorithms	23
7.	Three-dimensional Algorithm	27
8.	Conclusion	32
	References	34
	Appendix A	

§1. One-dimensional Second Order Algorithms

In recent years Roe [1] has evolved algorithms for systems of hyperbolic conservation laws based on the ideas of 'fluctuation' and 'signal'. The fluctuations measure flux variations in space and the signals are used to update the variables after a time step in such a way as to imitate the physics.

For the scalar equation in one dimension,

$$u_t + f_x \equiv u_t + a(u)u_x = 0, \tag{1.1}$$

the fluctuation in the cell $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ is defined by

$$\phi_j = \int_{\text{cell}} f_x \, dx = f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}, \tag{1.2}$$

equal to the rate of decrease of $\int u dx$ in the cell at time t. The notation f_k denotes $f(u(x_k))$. Note that internal cancellation ensures that

$$\sum \phi_j = 0, \tag{1.3}$$

the sum being taken over all cells.

A first order algorithm is obtained by updating u after a time step Δt by the consistent (in space) addition of the signal Φ_j, where

$$\Phi_j = -(\Delta t/\Delta x)\phi_j \tag{1.4}$$

and $\Delta x = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$. From (1.3) we have

$$\sum_{\substack{\text{new time}\\ \text{level}}} u = \sum_{\substack{\text{old time}\\ \text{level}}} u, \tag{1.5}$$

a discrete conservation law, so the algorithm is in conservation form.

If the signals are sent with the stream, so that they are received by $u_{j+\frac{1}{2}\sigma_j}$ where

$$a_j = \left(\frac{f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}}}{u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}} \right), \qquad \sigma_j = \text{sign}\,(a_j), \tag{1.6}$$

a_j being the stream velocity, then the scheme is simple upwinding. In the case $a_j \geq 0 \ \forall j$, the new value of $u_{j+\frac{1}{2}\sigma_j}$ is then (see Fig. 1).

$$u^{j+\frac{1}{2}\sigma_j} = u^k = u_k + \Phi_{k-\frac{1}{2}}$$

$$= u_k - a_{k-\frac{1}{2}} \frac{\Delta t}{\Delta x} (u_k - u_{k-1})$$

$$= (1-\nu_{k-\frac{1}{2}})u_k + \nu_{k-\frac{1}{2}} u_{k-1}, \qquad (1.7)$$

where

$$\nu_{k-\frac{1}{2}} = a_{k-\frac{1}{2}} \frac{\Delta t}{\Delta x}, \qquad (1.8)$$

from which it can be seen that, provided that $\nu_{k-\frac{1}{2}} \leq 1 \ \forall k$, u^k satisfies the local bound property (LB)

$$u^k \in [\min (u_{k-\frac{1}{2}}, u_{k+\frac{1}{2}}), \max (u_{k-\frac{1}{2}}, u_{k+\frac{1}{2}})], \qquad (1.9)$$

henceforth abbreviated to

$$u^k \in \{\min, \max\}(u_{k-\frac{1}{2}}, u_{k+\frac{1}{2}}). \qquad (1.10)$$

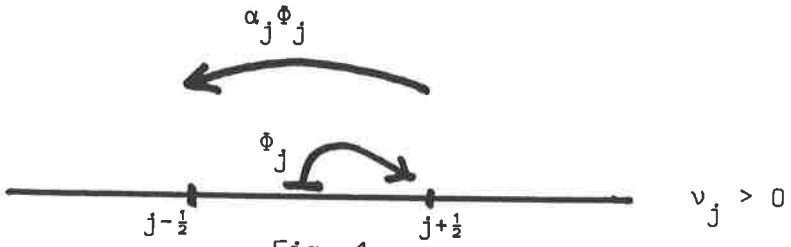

Fig. 1

For the more general case when a_j may change sign $\nu_{k-\frac{1}{2}}$ has to be restricted to $|\nu_{k-\frac{1}{2}}| \leq \frac{1}{2}$, see [2]. Note that the LB property (1.9) or (1.10) implies monotonicity preservation, i.e. monotone data remains monotone after a time step [2].

Higher order algorithms can be achieved by a re-adjustment of the signals but, in accordance with a theorem of Godunov [3], monotonicity preservation must be lost and thus the LB property must also be lost. Re-adjustments in the form of transfers will preserve conservation.

One re-adjustment is to transfer $\alpha_j \Phi_j$, where

$$\alpha_j = \tfrac{1}{2}(1 - |\nu_j|), \qquad (1.11)$$

across the cell $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ against the direction of the stream. First order accuracy is unaffected by such a transfer and the choice of α_j provides

for second order accuracy (see Fig. 1). The result is the Lax-Wendroff scheme [4]. No LB property exists since even when $a_j = a$, a positive constant,

$$u^k = u_k + \Phi_{k-\frac{1}{2}} - \alpha\Phi_{k-\frac{1}{2}} + \alpha\Phi_{k+\frac{1}{2}} \tag{1.12}$$

where $$\alpha = \tfrac{1}{2}(1-\nu) = \tfrac{1}{2}(1-\frac{a\Delta t}{\Delta x}) \tag{1.13}$$

and hence
$$\begin{aligned} u^k &= u_k - (1-\alpha)\nu(u_k-u_{k-1}) - \alpha\nu(u_{k+1}-u_k) \\ &= u_k - \tfrac{1}{2}\nu(1+\nu)(u_k-u_{k-1}) - \tfrac{1}{2}\nu(1-\nu)(u_{k+1}-u_k) \\ &= (1-\nu)u_k + \tfrac{1}{2}\nu(1+\nu)u_{k-1} - \tfrac{1}{2}\nu(1-\nu)u_{k+1} \end{aligned} \tag{1.14}$$

so that $u^k \notin \{\min,\max\}(u_{k-1},u_k,u_{k+1})$ in general. This result is consistent with the spontaneous generation of oscillations by the Lax-Wendroff scheme and with Godunov's theorem (see above).

A local bound does exist under certain circumstances, however. Consider again (1.12) but for variable $a_j \geq 0$ and let

$$\frac{\alpha_{k+\frac{1}{2}}\Phi_{k+\frac{1}{2}}}{\alpha_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}}} = \frac{\alpha_{k+\frac{1}{2}}\phi_{k+\frac{1}{2}}}{\alpha_{k-\frac{1}{2}}\phi_{k-\frac{1}{2}}} = r_k, \text{ say.} \tag{1.15}$$

Then
$$\begin{aligned} u^k &= u_k + \Phi_{k-\frac{1}{2}} - \alpha_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}} + \alpha_{k-\frac{1}{2}}r_k\Phi_{k-\frac{1}{2}} \\ &= \{1-\nu(1-\alpha+\alpha r_k)\}u_k + \nu(1-\alpha+\alpha r_k)u_{k-1}, \end{aligned} \tag{1.16}$$

(dropping suffices on ν and α) from which it can be seen that the LB property holds if

$$0 \leq \nu(1-\alpha+\alpha r_k) \leq 1 \tag{1.17}$$

or, when $0 < \nu < 1$,

$$-1 - \frac{2\nu}{1-\nu} \leq r_k \leq 3 + 2\,\frac{(1-\nu)}{\nu}\,. \tag{1.18}$$

For fluctuation ratios (1.15) satisfying (1.18) the LB property holds.

Another possible readjustment is to transfer $\alpha_j\Phi_j$ across another cell, say that cell adjacent and downstream of the cell $(x_{j-\frac{1}{2}},x_{j+\frac{1}{2}})$, (see Fig. 2). This gives the fully upwinded second order scheme [5] for which again no LB property generally exists. This time, in the $a_j = a > 0$ case,

$$u^{k+1} = u_{k+1} + \Phi_{k+\frac{1}{2}} - \alpha\Phi_{k-\frac{1}{2}} + \alpha\Phi_{k+\frac{1}{2}} \qquad (1.19)$$

$$= u_{k+1} - (1+\alpha)\nu(u_{k+1}-u_k) - \alpha\nu(u_k-u_{k-1})$$

$$= u_{k+1} - \tfrac{1}{2}\nu(3-\nu)(u_{k+1}-u_k) + \tfrac{1}{2}\nu(1-\nu)(u_k-u_{k-1})$$

$$= (1-\tfrac{3}{2}\nu+\tfrac{1}{2}\nu^2)u_{k+1} + \nu(2-\nu)u_k - \tfrac{1}{2}\nu(1-\nu)u_{k-1}, \qquad (1.20)$$

so that $u^{k+1} \notin \{\min, \max\}(u_{k-1}, u_k, u_{k+1})$ in general.

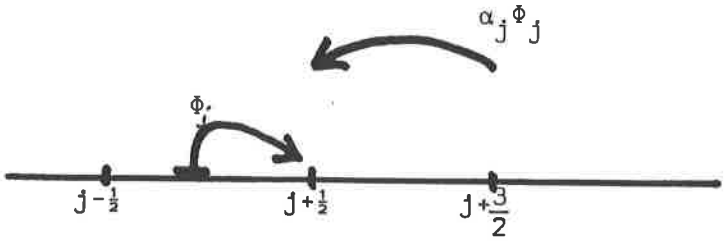

Fig. 2 $a_j > 0$ case

Again, however, the LB property does exist under certain circumstances. From (1.19) in the case $a_j > 0$, we have

$$u^{k+1} = u_{k+1} + \Phi_{k+\frac{1}{2}} - \alpha_{k-\frac{1}{2}}r_k^{-1}\Phi_{k+\frac{1}{2}} + \alpha_{k+\frac{1}{2}}\Phi_{k+\frac{1}{2}}$$

$$= \{1-\nu_{k+\frac{1}{2}}(1+\alpha_{k+\frac{1}{2}}-\alpha_{k+\frac{1}{2}}r_k^{-1})\}u_{k+1} + (1+\alpha_{k+\frac{1}{2}}-\alpha_{k+\frac{1}{2}}r_k^{-1})\nu_{k+\frac{1}{2}}u_k, \qquad (1.21)$$

so that the LB property holds if

$$0 \le \nu_{k+\frac{1}{2}}(1+\alpha_{k+\frac{1}{2}} - \alpha_{k+\frac{1}{2}}r_k^{-1}) \le 1 \qquad (1.22)$$

or, provided that $0 < \nu < 1$, (dropping the suffices)

$$-1 - 2\frac{(1-\nu)}{\nu} \le r_k^{-1} \le 2 + \frac{2\nu}{1-\nu} \quad . \qquad (1.23)$$

For fluctuation ratios $\frac{\phi_{k-\frac{1}{2}}}{\phi_{k+\frac{1}{2}}}$ (c.f. (1.15)) satisfying (1.23) the LB property is satisfied. For more general variable a_j the situation is more complicated (see Sweby & Baines [2]).

The above transfers can be regarded as special cases of a general transfer function

$$B(\alpha_j \Phi_j, \alpha_{j-\sigma_j} \Phi_{j-\sigma_j}) \tag{1.24}$$

which has the form $B(b_1,b_2) = b_1$ in the Lax-Wendroff case and $B(b_1,b_2) = b_2$ in the fully upwinded case (see Roe & Baines [6]). The function

$$B(b_1,b_2) = \tfrac{1}{2}(b_1 + b_2) \tag{1.25}$$

corresponds to Fromm's algorithm.

In one version of his second order algorithm Roe [7] uses

$$B(b_1,b_2) = \begin{cases} b_1 & b_1 \le b_2 \\ b_2 & b_2 < b_1 \\ 0 & \end{cases} \quad \left.\begin{array}{l} \\ \\ \end{array}\right\} \begin{array}{l} b_1 b_2 > 0 \\ \\ b_1 b_2 \le 0 \end{array} \tag{1.26}$$

while Sweby [8] has carried out a thorough study of the related function

$$B(b_1,b_2) = \text{minmod}\,(b_1,b_2) \tag{1.27}$$

which selects the argument with minimum modulus. Because of their non-linearity both of these functions escape the limitations of Godunov's theorem (see above) and lead to second order algorithms with the LB property. Note that b_1 is selected in (1.26) if $0 \le r_k \le 1$ and in (1.27) if $-1 \le r_k \le 1$, both consistent with (1.18), while b_2 is selected in (1.26) if $0 \le r_k^{-1} \le 1$, and in (1.27) if $-1 \le r_k^{-1} \le 1$, both consistent with (1.23). Figs. 3 and 4 show the B-functions in the two cases.

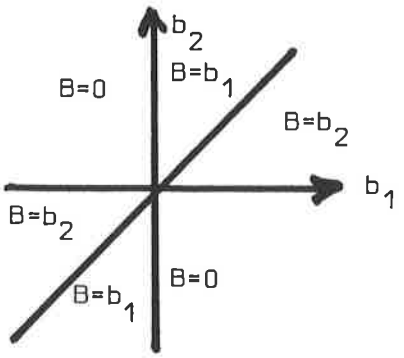

Fig. 3 : B-function (1.26)

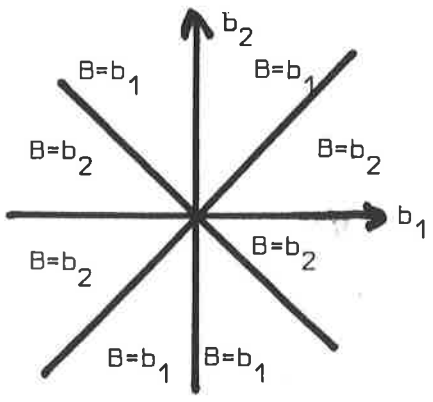

Fig. 4 : B-function (1.27)

More generally for any B function, transferring B_j, say, we have (in the case a_j variable but positive)

$$u^k = u_k + \Phi_{k-\frac{1}{2}} - B_{k-\frac{1}{2}} + B_{k+\frac{1}{2}} \,. \tag{1.28}$$

Let
$$\frac{B(b_1,b_2)}{b_1} = \beta, \qquad \frac{B(b_1,b_2)}{b_2} = \gamma. \tag{1.29}$$

Then, with β, γ referring to the cells $k-\frac{1}{2}$, $k+\frac{1}{2}$, we can write (1.28) as

$$u^k = u_k + \Phi_{k-\frac{1}{2}} - \beta\,\alpha_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}} + \gamma\,\alpha_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}} \tag{1.30}$$

$$= u_k - \{1-(\beta-\gamma)\alpha_{k-\frac{1}{2}}\}\nu_{k+\frac{1}{2}}(u_k-u_{k-1})$$

$$= [1-\nu\{1-(\beta-\gamma)\alpha\}]u_k + \{1-(\beta-\gamma)\alpha\}\nu u_{k-1} \tag{1.31}$$

(dropping suffices on ν and α). Hence the LB property is satisfied when $0 < \nu < 1$ if

$$0 \le \nu\{1-(\beta-\gamma)\alpha\} \le 1, \tag{1.32}$$

which requires

$$-2 - \frac{2(1-\nu)}{\nu} \le \beta - \gamma \le 2 + \frac{2\nu}{1-\nu} \,. \tag{1.33}$$

Conditions (1.18) and (1.23) are reproduced when $\beta = 1$ (the case $B(b_1,b_2) = b_1$) and when $\gamma = 1$ (the case $B(b_1,b_2) = b_2$), respectively. Condition (1.33) is in fact always satisfied by the B-functions in (1.26) and (1.27) since $-1 \le \beta \le 1$ and $-1 \le \gamma \le 1$ for these functions.

An important effect in second order algorithms is the reduction of the numerical diffusion, or spreading, evident in first order schemes. The choices (1.26) and (1.27) generally achieve second order accuracy and so reduce spreading, and in addition they possess the LB property and so do not introduce spurious oscillations. However, these B-functions provide somewhat safe choices and larger transfers are consistent with the LB property (further inhibiting diffusion) in restricted regions of the b_1, b_2 plane. Thus the choice (Roe [7])

$$B(b_1,b_2) = \left\{ \begin{array}{ll} b_1 & b_1 \geq b_2 \\ b_2 & b_2 > b_1 \\ 0 & \end{array} \right\} \begin{array}{l} b_1 b_2 \geq 0 \\ \\ b_1 b_2 \leq 0 \end{array} \qquad (1.34)$$

(c.f. (1.26)) will satisfy the LB property provided that

$$\left. \begin{array}{ll} \dfrac{b_1}{b_2} \leq 2 + \dfrac{2\nu}{1-\nu}\,, & \dfrac{b_2}{b_1} \leq 2 + \dfrac{2(1-\nu)}{\nu} \\ \\ = \dfrac{2}{1-\nu} & = \dfrac{2}{\nu} \end{array} \right\} \qquad (1.35)$$

(see (1.18) and (1.23). The region of the b_1, b_2 plane given by (1.35) is shown shaded in Fig. 5.

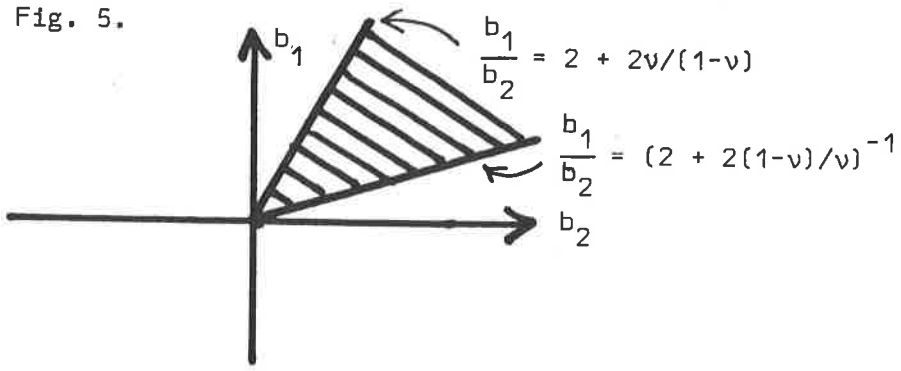

Fig. 5

In smooth regions we might expect the ratio $\dfrac{b_1}{b_2}$ to be not too far from 1 and in that case second order accuracy with minimal spreading and the LB property are all achieved. As solutions become less smooth $\dfrac{b_1}{b_2}$ will move out of the shaded region in Fig. 5. But in that case second order accuracy may be of less interest than the LB property, and so we maintain the latter property (and continuity of the B-functions) by setting

$$B(b_1,b_2) = \left\{ \begin{array}{ll} (2 + \dfrac{2\nu}{1-\nu})b_2 & (b_1 \geq b_2) \\ \{2 + \dfrac{2(1-\nu)}{\nu}\}b_1 & (b_2 < b_1) \end{array} \right\} \qquad (1.36)$$

in the unshaded regions of the first quadrant in Fig. 5. We end up with the B-function

$$B(b_1,b_2) = \left\{ \begin{array}{ll} b_1 & 1 \le \frac{b_1}{b_2} \le 2 + \frac{2\nu}{1-\nu} \\ b_2 & 1 \le \frac{b_2}{b_1} \le 2 + \frac{2(1-\nu)}{\nu} \\ 2 + \frac{2\nu}{1-\nu}\, b_2 & \frac{b_1}{b_2} \ge 2 + \frac{2\nu}{1-\nu} \\ 2 + \frac{2(1-\nu)}{\nu} b_1 & \frac{b_2}{b_1} \ge 2 + \frac{2(1-\nu)}{\nu} \end{array} \right\} \quad (1.37)$$

for $b_1b_2 > 0$ with $B(b_1,b_2) = 0$ for $b_1b_2 \le 0$. This choice satisfies the LB property in non-smooth regions and has the maximum second order anti-diffusion effect in smooth regions.

Roe [7] has advocated a simpler form of B-function contained within (1.37), namely,

$$B(b_1,b_2) = \left\{ \begin{array}{ll} b_1 & 1 \le \frac{b_1}{b_2} \le 2 \\ b_2 & 1 \le \frac{b_2}{b_1} \le 2 \\ 2b_2 & \frac{b_1}{b_2} \ge 2 \\ 2b_1 & \frac{b_2}{b_1} \ge 2 \end{array} \right. \quad (1.38)$$

for $b_1b_2 > 0$ with $B(b_1,b_2) = 0$ for $b_1b_2 < 0$ and has demonstrated its non-diffusive properties.

Note that the LB property forces peaks and troughs in the data to be trimmed undesirably and that the Roe schemes above (although not (1.27)) revert to first order accuracy at extrema, causing extra diffusion. A suggestion for combatting this effect is to add the rule that, if $b_1b_2 < 0$,

$$B_j = \left\{ \begin{array}{ll} B_{j+\sigma_j} & \left|\frac{b_1}{b_2}\right| < 1 \\ B_{j-\sigma_j} & \left|\frac{b_1}{b_2}\right| > 1 . \end{array} \right. \quad (1.39)$$

This ensures that second order accuracy is maintained at the most important point, closest to the peak or trough: the unwanted LB property is not preserved,

of course.

§2. A One-dimensional Third Order Algorithm

An attractive feature of the second order algorithms discussed above is the compactness of their support, everything depending on the conditions in two adjacent cells. A third order algorithm can be achieved using the same support but the LB property will hold only for a further restricted set of values of b_1/b_2.

To obtain the third order algorithm we carry out a further readjustment of the original signals, transferring this time a proportion of the difference of fluctuations across a cell. This leaves first and second order accuracy unaffected and, with an appropriate choice of weight, third order accuracy is met.

Thus, in the $a_j > 0$ case, transfer

$$\tau_j \Phi_j \text{ - } \tau_{j-\sigma_j} \Phi_{j-\sigma_j}, \tag{2.1}$$

where

$$\tau_j = \frac{1}{6}(1 - |\nu_j|^2), \tag{2.2}$$

across the cell $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ in the direction of the stream.

An alternative process, which achieves the same result, is to carry out the transfers as shown in Fig. 6 (for the $a_j > 0$ case) for each Φ_j.

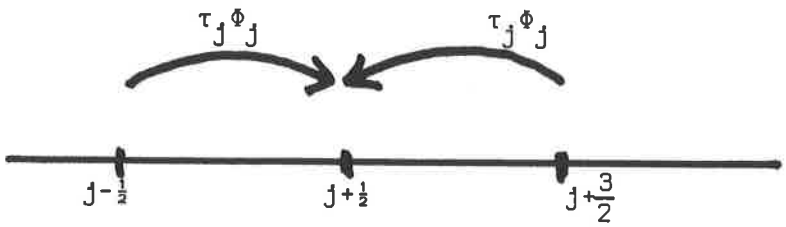

Fig. 6

Then the updated value of u_k is

$$u^k = u_k + \Phi_{k-\frac{1}{2}} - \alpha_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}} + \alpha_{k+\frac{1}{2}}\Phi_{k+\frac{1}{2}} + \left(\tau_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}} - \tau_{k-\frac{3}{2}}\Phi_{k-\frac{3}{2}}\right) - \left(\tau_{k+\frac{1}{2}}\Phi_{k+\frac{1}{2}} - \tau_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}}\right) \tag{2.3}$$

$$= u_k + \Phi_{k-\frac{1}{2}} - \alpha_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}} + \alpha_{k-\frac{1}{2}} r_k \Phi_{k-\frac{1}{2}} + \tau_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}} - \tau_{k-\frac{3}{2}} \frac{\alpha_{k-\frac{1}{2}}}{\alpha_{k-\frac{3}{2}}} r_{k-1}^{-1}\Phi_{k-\frac{1}{2}}$$

$$- \tau_{k+\frac{1}{2}} \frac{\alpha_{k-\frac{1}{2}}}{\alpha_{k+\frac{1}{2}}} r_k \Phi_{k-\frac{1}{2}} + \tau_{k-\frac{1}{2}}\Phi_{k-\frac{1}{2}}$$

$$= u_k - \left[\left(1-\alpha_{k-\frac{1}{2}}+2\tau_{k-\frac{1}{2}}\right) + \left(1 - \frac{\tau_{k+\frac{1}{2}}}{\alpha_{k+\frac{1}{2}}}\right)\alpha_{k-\frac{1}{2}} r_k - \tau_{k-\frac{3}{2}} \frac{\alpha_{k-\frac{1}{2}}}{\alpha_{k-\frac{3}{2}}} r_{k-1}^{-1}\right]\nu_{k-\frac{1}{2}}(u_k - u_{k-1}), \tag{2.4}$$

and $u^k \in \{\min, \max\}(u_k, u_{k-1})$ if

$$0 \le \left[\left(1-\alpha_{k-\frac{1}{2}} + 2\tau_{k-\frac{1}{2}}\right) + \left(1 - \frac{\tau_{k+\frac{1}{2}}}{\alpha_{k+\frac{1}{2}}}\right)\alpha_{k-\frac{1}{2}} r_k - \tau_{k-\frac{3}{2}} \frac{\alpha_{k-\frac{1}{2}}}{\alpha_{k-\frac{3}{2}}} r_{k-1}^{-1}\right]\nu_{k-\frac{1}{2}} \le 1 \tag{2.5}$$

or $$0 \le A_k + B_k r_k - C_k r_{k-1}^{-1} \le 1, \tag{2.6}$$

where $$A_k = \left(1-\alpha_{k-\frac{1}{2}}+2\tau_{k-\frac{1}{2}}\right)\nu_{k-\frac{1}{2}},\ B_k = \left(1 - \frac{\tau_{k+\frac{1}{2}}}{\alpha_{k+\frac{1}{2}}}\right)\alpha_{k-\frac{1}{2}}\nu_{k-\frac{1}{2}},\ C_k = \frac{\tau_{k-\frac{3}{2}}\,\alpha_{k-\frac{1}{2}}\nu_{k-\frac{1}{2}}}{\alpha_{k-\frac{3}{2}}}. \tag{2.7}$$

Since $B_k > 0$ and $C_k > 0$ it is sufficient that r_k, r_{k-1} should satisfy

$$r \le r_k \le R, \qquad r \le r_{k-1} \le R, \tag{2.8}$$

where $r, R \ge 0$ and

$$0 = A_k + B_k r - C_k r^{-1}, \qquad , A_k + B_k R - C_k R^{-1} = 1. \tag{2.9}$$

Now $$A_k + B_k \rho - C_k \rho^{-1} \tag{2.10}$$

is monotone for $\rho > 0$ and lies between 0 and 1 if ρ lies between r and R.

Since

$$A_k = (1-\alpha_{k-\frac{1}{2}}+2\tau_{k-\frac{1}{2}})\nu_{k-\frac{1}{2}} = [\tfrac{1}{2}(1+\nu_{k-\frac{1}{2}}) + \tfrac{1}{3}(1-\nu_{k-\frac{1}{2}}^2)]\nu_{k-\frac{1}{2}}$$

$$= \tfrac{1}{6}\nu_{k-\frac{1}{2}}(1+\nu_{k-\frac{1}{2}})(5-2\nu_{k-\frac{1}{2}}), \tag{2.11}$$

its slope is $5 - 6\nu_{k-\frac{1}{2}}(1-\nu_{k-\frac{1}{2}})$ which is positive for $0 < \nu_{k-\frac{1}{2}} < 1$ and $A_k = 0$ when $\nu_{k-\frac{1}{2}} = 0$, $A_k = 1$ when $\nu_{k-\frac{1}{2}} = 1$. Hence

$$0 < A_k < 1 \tag{2.12}$$

provided that $0 < \nu_{k-\frac{1}{2}} < 1$. Thus we may take r and R to be the roots

$$r = -\tfrac{1}{2}\frac{A_k}{B_k} + \tfrac{1}{2}\left(\frac{A_k^2}{B_k^2} + 4\frac{C_k}{B_k}\right)^{\frac{1}{2}} \tag{2.13}$$

and

$$R = \frac{\tfrac{1}{2}(1-A_k)}{B_k} + \tfrac{1}{2}\left(\frac{(1-A_k)^2}{B_k^2} + \frac{4C_k}{B_k}\right)^{\frac{1}{2}}$$

of (2.9).

For values of r_k or r_{k-1} outside the limits in (2.13) the LB property is lost. However one can then use one of the LB preserving second order schemes above.

Note that for second order accuracy $C_k = 0$ and r, R reduce to

$$r = 0, \qquad R = \frac{1-A_k}{B_k} = \frac{1-\nu_{k-\frac{1}{2}}(1-\alpha_{k-\frac{1}{2}})}{\alpha_{k-\frac{1}{2}}\nu_{k-\frac{1}{2}}}, \tag{2.14}$$

in accordance with (1.17) (considering only positive values of r_k).

§3. Statement of the One-dimensional Algorithms

Summarising the one-dimensional schemes:-

For each cell $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$:

1. evaluate the fluctuation

$$\phi_j = f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}} \tag{3.1}$$

and the signal $\Phi_j = -\frac{\Delta t}{\Delta x}\phi_j = -\frac{\Delta t}{\Delta x}(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}})$; (3.2)

2. evaluate the CFL number

$$\nu_j = \frac{\Delta t}{\Delta x}\frac{(f_{j+\frac{1}{2}} - f_{j-\frac{1}{2}})}{(u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}})} \tag{3.3}$$

and a direction given by

$$\sigma_j = \text{sign}\,(\nu_j)\ ; \tag{3.4}$$

3. add Φ_j to $u_{j+\frac{1}{2}}$ or $u_{j-\frac{1}{2}}$ according as $\sigma_j = 1$ or $\sigma_j = -1$.

The resulting scheme is first order accurate with the LB property

$$u_j^{n+1} \in [\text{min.}, \text{max.}]\,(u_j^n,\ u_{j+\sigma}^n)\ ; \tag{3.5}$$

4. transfer $\alpha_j \Phi_j$

where $$\alpha_j = \tfrac{1}{2}(1-|\nu_j|) \tag{3.6}$$

from $u_{j+\frac{1}{2}\sigma_j}$ to $u_{j-\frac{1}{2}\sigma_j}$.

This scheme is second order accurate but has no LB property.

5. Transfer $\tau_j \Phi_j$

where $$\tau_j = \frac{1}{6}(1-|\nu_j|^2) \tag{3.7}$$

to $u_{j+\frac{1}{2}\sigma_j}$ from both adjacent nodes.

The resulting scheme is third order accurate with no LB property.

For a second order LB preserving scheme, replace steps 4 and 5 by 4a.

Transfer $B(\alpha_j\Phi_j,\ \alpha_{j-\sigma_j}\Phi_{j-\sigma_j})$ to $u_{j-\frac{1}{2}\sigma_j}$ from $u_{j+\frac{1}{2}\sigma_j}$, where the B-function is one of those mentioned above:

- if the B-function is as in (1.26) or (1.27), the scheme is second order (except at extrema of the data in the case of (1.26)) and possesses the LB property at all points (other than extrema).
- if the B-function is as in (1.37) or (1.38) the scheme is second order accurate only if the ratio b_1/b_2 of the arguments b_1, b_2 of the B-function lies within certain limits, but the

LB property is maintained and the scheme has much less numerical diffusion.

The above algorithms all refer to the scalar non-linear equation, and the demonstrations of the LB property refer to the case when the wave speed has one sign throughout. The case of a general wave speed has been discussed fully by Sweby & Baines [2], and Sweby [8]. The extension to systems of non-linear equations has been the subject of a special investigation of Roe [9] who, using a certain linearised form of the Jacobian matrix, diagonalises the system prior to applying one of the algorithms above.

§4. Two-dimensional Second Order Alorithms

Algorithms such as those described above have been applied to two-dimensional conservation laws, for which the scalar equation is

$$u_t + f_x + g_y \equiv u_t + a(u)u_x + b(u)u_y = 0, \tag{4.1}$$

via a form of operator splitting in which one-dimensional steps are taken alternately in the x and y directions [10]. Here however we generalise the ideas of the one-dimensional method directly into two dimensions seeking to avoid such a splitting technique (see [11]).

First we generalise the fluctuation ϕ by identifying it with the rate of decrease of $\int u d\Omega$ over a cell at time t (c.f. (1.2)). For the quadrilateral in Fig. 7 this is

$$\iint_{ABCD} (f_x+g_y)\,d\Omega = \oint_{ABCD} (f,g)\cdot d\underline{S}. \tag{4.2}$$

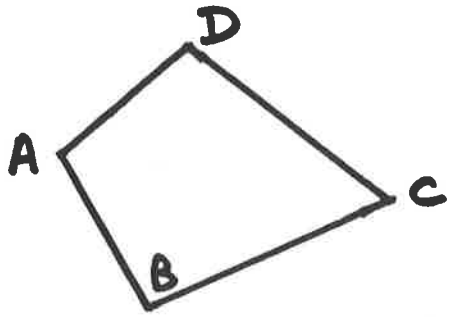

Fig. 7

Using trapezium rule integration we obtain the approximation

$$\sum_{\text{sides like CD}} \{\tfrac{1}{2}(f_C+f_D)(y_D-y_C) + \tfrac{1}{2}(g_C+g_D)(x_C-x_D)\} \tag{4.3}$$

When the quantities (4.3) are added over all cells there is total internal cancellation, so that signals constructed from (4.3) in the manner of earlier work will lead to

$$\sum_{\substack{\text{new}\\ \text{time level}}} u \;=\; \sum_{\substack{\text{old}\\ \text{time level}}} u \;, \tag{4.4}$$

the discrete conservation law. We shall define ϕ to be (4.3).

The contribution to the sum (4.3) from the points C and D is

$$\tfrac{1}{2}(f_C+f_D)(y_D-y_C) + \tfrac{1}{2}(g_C+g_D)(x_C-x_D) + \tfrac{1}{2} f_D(y_A-y_D) + \tfrac{1}{2}g_D(x_D-x_A)$$

$$+ \tfrac{1}{2}f_C(y_C-y_B) + \tfrac{1}{2}g_C(x_B-x_C) \tag{4.5}$$

$$= \tfrac{1}{2}f_C(y_D-y_B) + \tfrac{1}{2}f_D(y_A-y_C) + \tfrac{1}{2}g_C(x_B-x_D) + \tfrac{1}{2}g_D(x_C-x_A) \tag{4.6}$$

+Note: (see next page) Rather than splitting the fluctuations (4.3) into arrow fluctuations it would be more in line with the non-splitting philosophy here to use (4.3) or (4.8) in the form of a single signal sent wholly to a target or targets. However it has been found (Baines[11]; that the LB properties (§5) cannot be preserved unless biassed forms of (4.8) used which lead to conservation difficulties in the non-linear case.

If the quadrilateral is a rectangle on the xy grid (see Fig. 8),

$$y_D - y_B = +(y_A - y_C) = \Delta y, \quad \text{say},$$
$$x_C - x_A = -(x_B - x_D) = \Delta x, \quad \text{say}. \tag{4.7}$$

Then (4.6) becomes

$$\tfrac{1}{2}(f_D + f_C)\Delta y + \tfrac{1}{2}(g_D - g_C)\Delta x \quad . \tag{4.8}$$

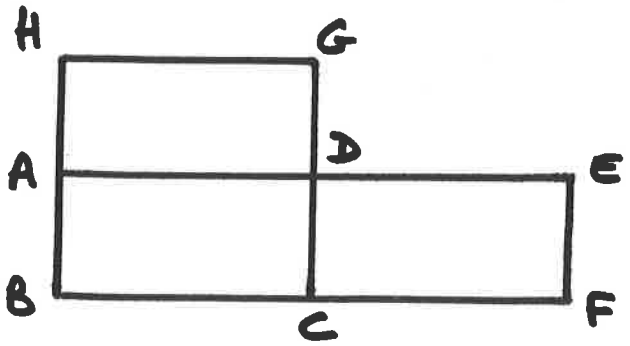

Fig. 8

Now, associating the term $\frac{1}{2}(g_D - g_C)$ with CD, leaving $\frac{1}{2}f_D\Delta y$, $\frac{1}{2}f_C\Delta y$ to be associated with AD, BC, respectively, the total contribution of CD from the cells ABCD, CDEF (See Fig. 8) to the sum in (4.3) is

$$\psi_g = (g_D - g_C)\Delta x. \tag{4.9}$$

Similarly the contribution from DA is

$$\psi_f = (f_D - f_A)\Delta y. \tag{4.10}$$

Now $\sum\phi$, as defined by (4.3), may equally be taken to be the sum of the quantities ψ_f, ψ_g taken over all <u>arrows</u> CD, AD in the grid of rectangles. Comparing ψ_f, ψ_g in (4.9), (4.10) with ϕ in (1.2) of §1, we see that there is a close correspondence, each component differing only by a length factor. Since we need to divide the arrow-fluctuations ψ_f, ψ_g by an <u>area</u> to obtain the correct dimensions for u (as compared with a length in one-dimension), we can construct a two-dimensional algorithm[+]

(+ see previous page)

based on that in one dimension by the simultaneous updating of u by the addition of signals

$$\Phi_f = -\frac{\Delta t}{\Delta x \Delta y}\psi_f \qquad \Phi_g = -\frac{\Delta t}{\Delta x \Delta y}\psi_g$$
$$= -\frac{\Delta t}{\Delta x}(f_D - f_A) \qquad = -\frac{\Delta t}{\Delta y}(g_D - g_C), \tag{4.11}$$

in line with the definition of Φ_j in (1.2) and (1.4) of §1. The advantage of this approach is that many of the results for the algorithms in one dimension go over into two dimensions. Although akin to splitting the method operates simultaneously in the x and y directions.

In a first order scheme the signals (4.11) may be sent with the stream components (indicated by the signs of $\frac{\partial f}{\partial u}$, $\frac{\partial g}{\partial u}$ or their approximations) to update u at the appropriate end of the arrows, i.e. using the signs of

$$a_{AD} = \frac{f_D - f_A}{u_D - u_A} \qquad \text{and} \qquad b_{CD} = \frac{g_D - g_C}{u_D - u_C} \tag{4.12}$$

for all arrows DA, CD. For this scheme, which is first order accurate, we can show taking the case $a,b > 0$, a two-dimensional LB property, from which monotonicity preservation in a range of directions can be deduced. The value of u_D at the new time level will be (see Fig. 9)

Fig. 9

$$u^D = u_D + \Phi_f + \Phi_g \tag{4.13}$$
$$= u_D - \nu_1(u_D - u_A) - \nu_2(u_D - u_C), \tag{4.14}$$

where $\nu_1 = a_{AD}\frac{\Delta t}{\Delta x}$, $\nu_2 = b_{CD}\frac{\Delta t}{\Delta y}$. (4.15)

Hence $$u^D = \{1-(\nu_1+\nu_2)\}u_D + \nu_1 u_A + \nu_2 u_C, \tag{4.16}$$

showing that, for $\nu_1 \geq 0$, $\nu_2 \geq 0$, $\nu_1 + \nu_2 \leq 1$, u^D satisfies the LB property

$$u^D \in \{\min,\max\}\ (u_D, u_A, u_C). \tag{4.17}$$

As in §1, for the general non-linear case it can be shown that the restriction needs to be tightened to $|\nu_1 + \nu_2| \leq \frac{1}{2}$.

We can use the results of §1 to develop a number of algorithms which are second order in the sense that terms in u_{xx} and u_{yy} in the truncation error are correctly matched. For example, a two-dimensional Lax-Wendroff type algorithm can be constructed by redistributing the signals (4.11) using transfers of $\alpha_1\Phi_f$ and $\alpha_2\Phi_g$ from D to A and C respectively, where

$$\begin{aligned} \alpha_1 &= \tfrac{1}{2}(1-|\nu_1|)\ , \\ \alpha_2 &= \tfrac{1}{2}(1-|\nu_2|)\ , \end{aligned} \tag{4.18}$$

(see Fig. 10). The result will not have the LB property, nor will it match the u_{xy} term in the truncation error.

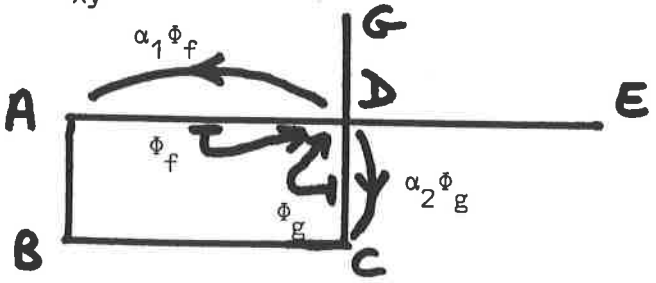

Fig. 10

Taking the latter point first, we introduce lateral transfers $\alpha_3\Phi_f$ from D to G and $\alpha_4\Phi_g$ from D to E. Existing first and second order terms in the truncation error are unaffected by this process and it is easily shown that the cross-derivative term in u_{xy} is matched if

$$\nu_1\alpha_3 + \nu_2\alpha_4 = \nu_1\nu_2\ , \tag{4.19}$$

which has a parametric solution

$$\alpha_3 = \nu_2 \cos^2\theta\ , \qquad \alpha_4 = \nu_1 \sin^2\theta\ . \tag{4.20}$$

To re-establish the LB property we use the B-function of (1.24). Then, corresponding to (1.30), we have for the updated value of u_D

$$\begin{aligned} u^D = u_D &+ \Phi_f - \beta\alpha_1\Phi_f + \gamma\alpha_1\Phi_f \\ &+ \Phi_g - \beta\alpha_2\Phi_g + \gamma\alpha_2\Phi_g \\ &- \beta\alpha_3\Phi_f + \gamma\alpha_3\Phi_f \\ &- \beta\alpha_4\Phi_g + \gamma\alpha_4\Phi_g, \end{aligned} \tag{4.21}$$

where β,γ are defined by (1.29) and in fact vary from term to term. Thus

$$\begin{aligned} u^D = u_D &+ (1-\beta\alpha_1+\gamma\alpha_1-\beta\alpha_3+\gamma\alpha_3)\nu_1(u_D-u_A) \\ &+ (1-\beta\alpha_2+\gamma\alpha_2-\beta\alpha_4+\gamma\alpha_4)\nu_2(u_D-u_C) \end{aligned} \tag{4.22}$$

and the LB property depends on the inequalities

$$\left.\begin{aligned} 1 - \beta\alpha_1 + \gamma\alpha_1 - \beta\alpha_3 + \gamma\alpha_3 &\geq 0 \\ 1 - \beta\alpha_2 + \gamma\alpha_2 - \beta\alpha_4 + \gamma\alpha_4 &\geq 0 \\ \nu_1(1-\beta\alpha_1+\gamma\alpha_1-\beta\alpha_3+\gamma\alpha_3) + \nu_2(1-\beta\alpha_2+\gamma\alpha_2-\beta\alpha_4+\gamma\alpha_4) &\leq 1. \end{aligned}\right\} \tag{4.23}$$

If we take the B-function to be that in (1.26) we have

$$0 \leq \beta \leq 1, \qquad 0 \leq \gamma \leq 1. \tag{4.24}$$

Thus

$$\begin{aligned} 1 - \beta\alpha_1 + \gamma\alpha_1 - \beta\alpha_3 + \gamma\alpha_3 &\geq 1 - \alpha_1 - \alpha_3 \\ &= \tfrac{1}{2}(1+\nu_1) - \nu_2\cos^2\theta, \end{aligned} \tag{4.25}$$

from (4.18) and (4.20). This is non-negative provided that

$$\cos^2\theta \leq \frac{(1+\nu_1)}{2\nu_2} . \tag{4.26}$$

Similarly $1 - \beta\alpha_2 + \gamma\alpha_2 - \beta\alpha_3 + \gamma\alpha_3 \geq 0$ if $\sin^2\theta \leq \frac{(1+\nu_2)}{2\nu_1}$.

A consistent value of θ is $\frac{\pi}{4}$, giving

$$\alpha_3 = \tfrac{1}{2}\nu_2, \qquad \alpha_4 = \tfrac{1}{2}\nu_1. \tag{4.27}$$

For the final condition in (4.23), we require

$$\begin{aligned}
&\nu_1(1-\beta\alpha_1+\gamma\alpha_1-\beta\alpha_3+\gamma\alpha_3) + \nu_2(1-\beta\alpha_2+\gamma\alpha_2-\beta\alpha_4+\gamma\alpha_4) \\
&\leq \nu_1(1+\alpha_1+\alpha_3) + \nu_2(1+\alpha_2+\alpha_4) \\
&= \nu_1\{\tfrac{1}{2}(3-\nu_1) + \nu_2\cos^2\theta\} + \nu_2\{\tfrac{1}{2}(3-\nu_2) + \nu_1\sin^2\theta\} \\
&= \tfrac{1}{2}\nu_1(3-\nu_1) + \tfrac{1}{2}\nu_2(3-\nu_2) + \nu_1\nu_2 \\
&= \frac{3}{2}(\nu_1+\nu_2) - \tfrac{1}{2}(\nu_1-\nu)^2
\end{aligned} \tag{4.28}$$

to be not greater than 1, which is the case if, for example

$$|\nu_1 + \nu_2| \leq \frac{2}{3}. \tag{4.29}$$

(the worst case is when $\nu_1 = \nu_2$).

If the B-function is taken as that in (1.27) we have only the weaker condition

$$|\beta| \leq 1, \quad |\gamma| \leq 1, \tag{4.30}$$

c.f. (4.24), and the corresponding conditions on ν_1,ν_2 and θ are then

$$\cos^2\theta \leq \frac{\nu_1}{2\nu_2}, \qquad \sin^2\theta \leq \frac{\nu_2}{2\nu_1}, \tag{4.31}$$

and

$$2(\nu_1+\nu_2) - (\nu_1^2+\nu_2^2) + \nu_1\nu_2 \leq 1.$$

These can be satisfied, in particular, if $\theta = \tan^{-1}(b/a)$ and (ν_1,ν_2) belongs to a rather small neighbourhood of $(0,0)$.

Under these conditions we have the LB property

$$u^D \in \{\min,\max\}(u_D, u_A, u_C), \tag{4.32}$$

Seeking now maximum anti-diffusion, as in (1.34), we set $\beta = 1$ in (4.22) (corresponding to $B(b_1,b_2) = b_1$) obtaining

$$u^D = u_D + (1-\alpha_1-\alpha_3+\gamma_1\alpha_1+\gamma_3\gamma_3)\nu_1(u_D-u_A) + (1-\alpha_2-\alpha_4+\gamma_2\alpha_2+\gamma_4\alpha_4)\nu_2(u_D-u_C), \tag{4.33}$$

(see Fig. 10), where

$$\left.\begin{aligned} \gamma_1 &= \frac{(\alpha_1\Phi_f)_{DE}}{(\alpha_1\Phi_f)_{AD}} \quad , \quad \gamma_2 = \frac{(\alpha_2\Phi_g)_{DG}}{(\alpha_2\Phi_g)_{CD}} \\ \gamma_3 &= \frac{(\alpha_3\Phi_f)_{BC}}{(\alpha_3\Phi_f)_{AD}} \quad , \quad \gamma_4 = \frac{(\alpha_4\Phi_g)_{AB}}{(\alpha_4\Phi_g)_{CD}} \quad . \end{aligned}\right\} \tag{4.34}$$

Suppose that $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are bounded, i.e.

$$g \le \gamma_i \le G, \quad (i = 1,2,3,4). \tag{4.35}$$

Then the coefficient of u_A in (4.33)

$$\begin{aligned} 1 - \alpha_1 - \alpha_3 + \gamma_1\alpha_1 + \gamma_3\alpha_3 &\ge 1 + (\alpha_1+\alpha_3)(g-1) \\ &\ge 0 \end{aligned} \tag{4.36}$$

if

$$g \ge 1 - (\alpha_1+\alpha_3)^{-1}. \tag{4.37a}$$

Similarly for the coefficient of u_C we require $g \ge 1 - (\alpha_2+\alpha_4)^{-1}$. (4.37b)

Finally, for the coefficient of u_D in (4.33) to be non-negative we require

$$\begin{aligned} &\nu_1(1-\alpha_1-\alpha_3+\gamma_1\alpha_1+\gamma_3\alpha_3) + \nu_2(1-\alpha_2-\alpha_4+\gamma_2\alpha_2+\gamma_4\alpha_4) \\ &\le \nu_1(1 + (\alpha_1+\alpha_3)(G-1)) + \nu_2(1+(\alpha_2+\alpha_4)(G-1)) \\ &\le 1 \end{aligned} \tag{4.38}$$

which holds if

$$G \le 1 + \{1 - (\nu_1+\nu_2)\}/\{\nu_1(\alpha_1+\alpha_3) + \nu_2(\alpha_2+\alpha_4)\}. \tag{4.39}$$

Therefore, provided that the four ratios $\gamma_1,\gamma_2,\gamma_3,\gamma_4$ lie in the range

$$(\max.[1-(\alpha_1+\alpha_3)^{-1},\ 1-(\alpha_2+\alpha_4)^{-1}],\ 1+\Gamma), \tag{4.40}$$

where $$\Gamma = \{1-(\nu_1+\nu_2)\}/\{(\alpha_1+\alpha_3)\nu_1 + (\alpha_2+\alpha_4)\nu_2\}, \tag{4.41}$$

then the coefficients of u_A, u_C, u_D in (4.33) are non-negative and the LB property holds. This is the range for which the Lax-Wendroff-type two-dimensional scheme referred to above has the LB property.

If we now set $\gamma = 1$ in (4.22) (corresponding to $B(b_1,b_2) = b_2$) we obtain

$$u^D = u_D + (1-\beta_1\alpha_1-\beta_3\alpha_3+\alpha_1+\alpha_3)\nu_1(u_D-u_A)$$
$$+ (1-\beta_2\alpha_2-\beta_4\alpha_4+\alpha_2+\alpha_4)\nu_2(u_D-u_C), \tag{4.42}$$

where $$\beta_1 = \gamma_1^{-1},\quad \beta_2 = \gamma_2^{-1},\ \beta_3 = \gamma_3^{-1},\ \beta_4 = \gamma_4^{-1} \tag{4.43}$$

(see (4.34)). For the LB property it is then sufficient that the β_i $(i = 1,2,3,4)$ lie in the range

$$(1-\Gamma,\ \min.[1+(\alpha_1+\alpha_3)^{-1},\ 1+(\alpha_2+\alpha_4)^{-1}]), \tag{4.44}$$

i.e. that the coefficients γ_i lie in the range

$$\left(\max.[\{1+(\alpha_1+\alpha_3)^{-1}\}^{-1},\ \{1+(\alpha_2+\alpha_4)^{-1}\}^{-1}],\ (1-\Gamma)^{-1}\right). \tag{4.45}$$

Taking (4.40) and (4.45) together, and noting that all ν's and α's are positive, we see that for the LB property to hold it is sufficient that $\gamma_1,\gamma_2,\gamma_3,\gamma_4$ lie in the range

$$\left(\max.[1+(\alpha_1+\alpha_3)^{-1},\ 1+(\alpha_2+\alpha_4)^{-1}],[\min.\ (1+\Gamma),(1-\Gamma)^{-1}]\right). \tag{4.46}$$

With α_3,α_4 given by (4.27), this gives the range

$$\left(\max.\left[1-\frac{2}{3-\nu_1+\nu_2},\ 1-\frac{2}{3+\nu_1-\nu_2} \right],\ \min.\left[1+\frac{\{1-(\nu_1+\nu_2)\}}{\frac{1}{2}(\nu_1+\nu_2)-\frac{1}{2}(\nu_1-\nu_2)^2},\ \left\{ 1-\frac{\{1-(\nu_1+\nu_2)\}}{\frac{1}{2}(\nu_1+\nu_2)-\frac{1}{2}(\nu_1-\nu_2)^2} \right\}^{-1} \right] \right) \tag{4.47}$$

If, further $\nu_1 \le 1$, $\nu_2 \le 1$ and $\nu_1 + \nu_2 \le \frac{2}{3}$ c.f. (4.29), the range for γ_i, and therefore for $\frac{b_1}{b_2}$, can be replaced by

$$\tfrac{1}{2} \le \frac{b_1}{b_2} \le 2, \tag{4.48}$$

(i = 1,2,3,4) as in the one-dimensional case (1.38), with the same choice of B-function as (1.38) when (4.48) is not satisfied. More generally we can define, as in (1.37),

$$B(b_1,b_2) = \left\{ \begin{array}{ll} b_1 & 1 \le \frac{b_1}{b_2} \le k \\ b_2 & 1 \le \frac{b_2}{b_1} \le K \\ kb_2 & \frac{b_1}{b_2} \ge k \\ Kb_1 & \frac{b_1}{b_2} \le K \end{array} \right\}, \tag{4.49}$$

where

$$\left. \begin{array}{l} k = \max.\left[1-\frac{2}{3-\nu_1+\nu_2},\ 1-\frac{2}{3+\nu_1-\nu_2} \right] \\ K = \min.\left[1+\frac{\{1-(\nu_1+\nu_2)\}}{\frac{1}{2}(\nu_1+\nu_2)-\frac{1}{2}(\nu_1-\nu_2)^2},\ \left\{ 1-\frac{\{1-(\nu_1+\nu_2)\}}{\frac{1}{2}(\nu_1+\nu_2)-\frac{1}{2}(\nu_1-\nu_2)^2} \right\}^{-1} \right]. \end{array} \right\} \tag{4.50}$$

§5. A Third-order Two-dimensional Algorithm

We now construct a two-dimensional third order scheme along the lines of §2.

Using the transfers in §2 on the signals of (4.11) we ensure that terms in the truncation error proportional to u_x, u_{xx}, u_{xxx} and u_y, u_{yy}, u_{yyy}

are matched and, with the lateral transfers of §4 included, the term in u_{xy} is also matched. It remains to deal with the terms u_{xxy} and u_{xyy} in the truncation error. These can be matched by further lateral transfers (in the $a_j, b_j > 0$ case) of the form

$$\tfrac{1}{2}\nu_2(1-\nu_2)\Phi_f \tag{5.1}$$

from G to D and from C to D (see Figs. 10 and 11) and

$$\tfrac{1}{2}\nu_1(1-\nu_1)\Phi_g \tag{5.2}$$

from E to D and from A to D.

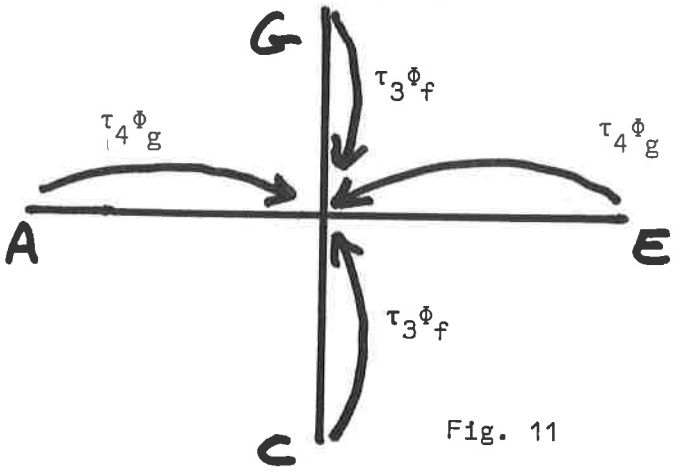

Fig. 11

transfers for u_{xxy} and u_{xyy}:

$\tau_3 = \tfrac{1}{2}\nu_2(1-\nu_2)$

$\tau_4 = \tfrac{1}{2}\nu_1(1-\nu_1)$

It is possible to analyse this scheme and to determine that it is stable in a neighbourhood of $(\nu_1,\nu_2) = (0,0)$ [see Appendix A by P.L. Roe]. It is also possible (but intricate) to find regions of the (b_1,b_2) plane for which the scheme possesses the LB property.

§6. Statement of the Two-dimensional Algorithms

Summarising the two-dimensional schemes :-

For each x-arrow (like AD in Fig. 12):

1. evaluate a fluctuation

$$\psi_f = (f_D - f_A)\Delta y \tag{6.1}$$

and a signal

$$\Phi_f = -\frac{\Delta t}{\Delta x \Delta y}\psi_f = -\frac{\Delta t}{\Delta x}(f_D - f_A)\ , \tag{6.2}$$

2. evaluate a CFL number

$$\nu_1 = \frac{\Delta t}{\Delta x}\frac{(f_D - f_A)}{(u_D - u_A)} \tag{6.3}$$

and a direction $\sigma_1 = \text{sign}\,(\nu_1)$; (6.4)

3. add Φ_f to u_D or u_A according as $\sigma_1 = 1$ or $\sigma_1 = -1$.

- repeat steps 1-3 for each y-arrow (like CD in Fig. 13), defining corresponding quantities $\psi_g, \Phi_g, \nu_2, \sigma_2$.

The resulting scheme is first order with the LB property

$$u^{n+1}_{i,j} \in [\min,\max](u^n_{i,j},\ u^n_{i-\sigma_1,j},\ u_{i,j-\sigma_2}) \tag{6.5}$$

Further, for each x-arrow:

4. transfer $\alpha_1\Phi_f$,

where $$\alpha_1 = \tfrac{1}{2}(1-|\nu_1|), \tag{6.6}$$

from u_D to u_A or from u_A to u_D according as $\sigma_1 = 1$ or $\sigma_1 = -1$;

5. transfer $\alpha_3\Phi_f$,

where $$\alpha_3 = |\nu_2|\cos^2\theta, \tag{6.7}$$

from u_D to u_G $(\sigma_1=1,\ \sigma_2=1)$
u_D to u_C $(\sigma_1=1,\ \sigma_2=-1)$
u_A to u_H $(\sigma_1=-1,\ \sigma_2=1)$
u_A to u_B $(\sigma_1=-1,\ \sigma_2=-1)$ (6.8)

- repeat steps 4-5 for each y-arrow, defining

$$\alpha_2 = \tfrac{1}{2}(1-|\nu_2|),\quad \alpha_4 = |\nu_1|\sin^2\theta. \tag{6.9}$$

The scheme so far is second order accurate but has no LB property.

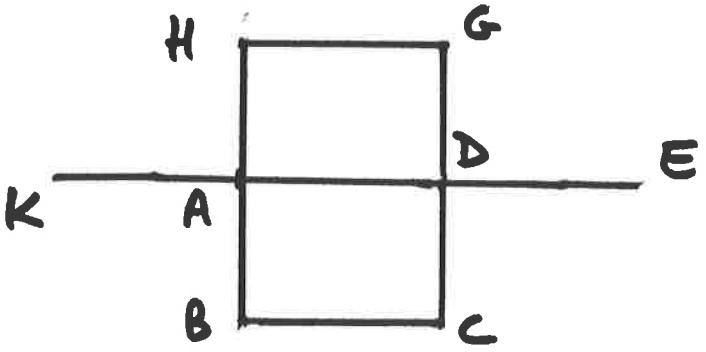

Fig. 12

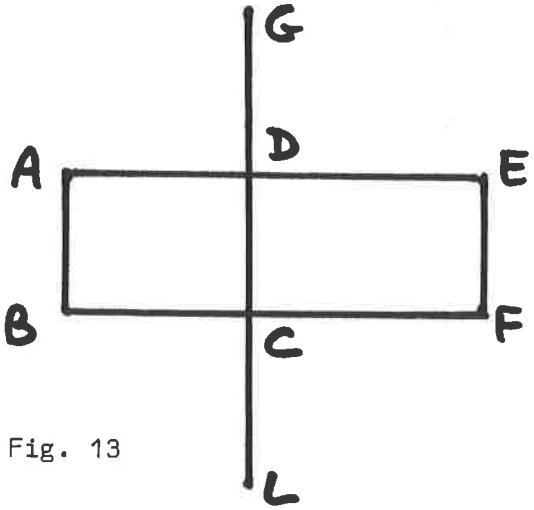

Fig. 13

Finally, for each x-arrow:

6. transfer $\tau_1\Phi_f$,

   where $$\tau_1 = \frac{1}{6}(1-|\nu_1|^2), \tag{6.10}$$

   from u_A and u_E to u_D (if $\sigma_1 = 1$) or from u_D and u_K to u_A (if $\sigma_1 = -1$);

7. transfer $\tau_3\Phi_f$,

   where $$\tau_3 = \tfrac{1}{2}|\nu_2|(1-|\nu_2|) \tag{6.11}$$

   from u_G and u_C to u_D ($\sigma_1 = 1$), or from u_H and u_B to u_A ($\sigma_1 = -1$);

- repeat steps 6-7 for each y-arrow, defining

$$\tau_2 = \frac{1}{6}(1-|\nu_2|^2), \qquad \tau_4 = \tfrac{1}{2}|\nu_1|(1-|\nu_1|). \tag{6.12}$$

The resulting scheme is third order accurate but has no LB property.

For a second order LB preserving scheme, replace steps 4-7 by

4a. transfer $B(\alpha_1\Phi_f, \alpha_1^u \Phi_f^u)$ from u_D to u_A or from u_A to u_D according as $\sigma_1 = 1$ or $\sigma_1 = -1$, where α_1^u and Φ_f^u are the

values of α_1 and Φ_f for the adjacent upwind arrow, and the B-function is as discussed below.

5a. transfer B $(\alpha_3\Phi_f,\ \alpha_3^p\Phi_f^p)$ as in step 5 above, where α_3^p and α_f^p are the values of α_3 and Φ_f for adjacent parallel arrow upwind w.r.t. ν_2.

Taking the B-function to be as in (1.26) yields a scheme which is second order except at flat points of the data and possesses the LB property (6.5) at all points (other than flat points), provided that

$$\left.\begin{aligned} &\cos^2\theta \le \frac{1+|\nu_1|}{2|\nu_2|}\ , \quad \sin^2\theta \le \frac{1+|\nu_2|}{2|\nu_1|} \\ &\text{(satisfied by } \cos^2\theta = \sin^2\theta = \tfrac{1}{2}\text{)} \\ &\text{and} \qquad |\nu_1 + \nu_2| \le \frac{2}{3}\ . \end{aligned}\right\} \tag{6.13}$$

Taking the B-function to be as in (1.27) yields a scheme which is everywhere second order and possesses the LB property provided that

$$\left.\begin{aligned} &\cos^2\theta \le \frac{|\nu_1|}{2|\nu_2|}\ , \quad \sin^2\theta \le \frac{|\nu_2|}{2|\nu_1|} \\ &\text{(satisfied by } \tan\theta = \nu_2/\nu_1\text{)} \\ &\text{and} \qquad 2(\nu_1+\nu_2) - \nu_1\nu_2 - (\nu_1^2+\nu_2^2) \le 1. \end{aligned}\right\} \tag{6.14}$$

Taking the B-function to be as in (4.49) or (1.38) yields a scheme which is second order provided that the ratio $\frac{b_1}{b_2}$ of the arguments b_1, b_2 of the B-functions used is sufficiently close to 1, but it possesses the LB property and also has strong anti-diffusion properties.

§7. Three-dimensional Algorithm

We now consider the extension to three dimensions. If the cell is the rectangular box ABCD A'B'C'D', orientated as in Fig. 14, the fluctuation in the box is

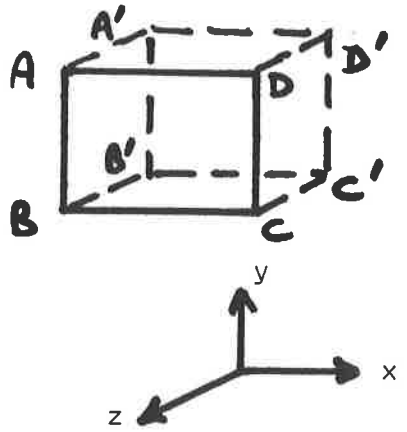

Fig. 14

$$\iiint_{\text{box}} (f_x+g_y+h_z)\,d\Omega = \oint (f,g,h)\cdot d\underline{S} \qquad (7.1)$$

which, using trapezium rule integration, gives

$$\phi = \sum_{\substack{\text{sides like}\\ ABCD}} \tfrac{1}{4}(h_A+h_B+h_C+h_D)\Delta x\Delta y, \qquad (7.2)$$

and we note that $\sum_{\text{boxes}} \phi = 0$ and updating the u's by the ϕ's will give an algorithm in conservation form. (7.3)

Considering the contribution of the points A,B,C,D to the whole sum in (7.2) we obtain

$$\tfrac{1}{4}(-f_A-f_B+f_C+f_D)\Delta y\Delta z + \tfrac{1}{4}(g_A+g_D-g_B-g_C)\Delta z\Delta x + \tfrac{1}{4}(h_A+h_B+h_C+h_D)\Delta x\Delta y. \qquad (7.4)$$

Then the contribution from the points C,D to the sum of the four ϕ's from boxes which have CD as an edge is

$$\psi_g = (g_D-g_C)\Delta z\Delta x. \qquad (7.5)$$

Hence $\sum\phi$, as defined by (7.2), may be broken down into the sum of terms like (7.5) and

$$\psi_f = (f_D-f_A)\Delta y\Delta z, \quad \psi_h = (h_D-h_D)\Delta x\Delta y \qquad (7.6)$$

which, as in §4, we may call arrow-fluctuations. We can then construct a three-dimensional algorithm by updating u values with signals

$$\left.\begin{aligned} \Phi_f &= \frac{-\Delta t}{\Delta x \Delta y \Delta z}\,\psi_f = -\frac{\Delta t}{\Delta x}\,(f_D - f_A) \\ \Phi_g &= \frac{-\Delta t}{\Delta x \Delta y \Delta z}\,\psi_g = -\frac{\Delta t}{\Delta y}\,(g_D - g_C) \\ \Phi_h &= \frac{-\Delta t}{\Delta x \Delta y \Delta z}\,\psi_h = -\frac{\Delta t}{\Delta z}\,(h_D - h_{D'}). \end{aligned}\right\} \qquad (7.7)$$

Taking the approximate wave speeds

$$a = \frac{f_D - f_A}{u_D - u_A}\,, \qquad b = \frac{g_D - g_C}{u_D - u_C}\,, \qquad c = \frac{h_D - h_{D'}}{u_D - u_{D'}} \qquad (7.8)$$

as positive, the corresponding first order scheme is to update u_D by the three signals (7.7). The value of u_D at the new time level will then be

$$u^D = u_D + \Phi_f + \Phi_g + \Phi_h \qquad (7.9)$$

$$= u_D - \nu_1(u_D - u_A) - \nu_2(u_D - u_C) - \nu_3(u_D - u_{D'}), \qquad (7.10)$$

where $$\nu_1 = a_{DA}\frac{\Delta t}{\Delta x}\,, \quad \nu_2 = b_{DC}\frac{\Delta t}{\Delta y}\,, \quad \nu_3 = c_{DD'}\frac{\Delta t}{\Delta z}\,. \qquad (7.11)$$

Hence $$u^D = \{1-(\nu_1+\nu_2+\nu_3)\}u_D + \nu_1 u_A + \nu_2 u_C + \nu_3 u_{D'}, \qquad (7.12)$$

showing that, if $\nu_1, \nu_2, \nu_3 \geq 0$ and $|\nu_1 + \nu_2 + \nu_3| \leq 1$, u satisfies the LB property

$$u^D \in [\min,\max](u_A, u_C, u_D) \qquad (7.13)$$

(see Fig. 14).

Analogously to the two-dimensional case we can construct a second order scheme by transferring quantities

$$\alpha_{11}\Phi_f, \quad \alpha_{22}\Phi_g, \quad \alpha_{33}\Phi_h, \qquad (7.14)$$

where

$$\alpha_{ii} = \tfrac{1}{2}(1-|\nu_i|) \quad (i = 1,2,3), \qquad (7.15)$$

as in the one-dimensional case, and adding lateral transfers.

Introducing lateral transfers $\alpha_{1i}\Phi_f$ (i = 1,2,3) (as indicated in Fig. 15), (c.f. Fig. 14), together with $\alpha_{2i}\Phi_g$, $\alpha_{3i}\Phi_h$ (i = 1,2,3), we find that in order to match the u_{yz}, u_{zx} and u_{xy} terms in the truncation error we need

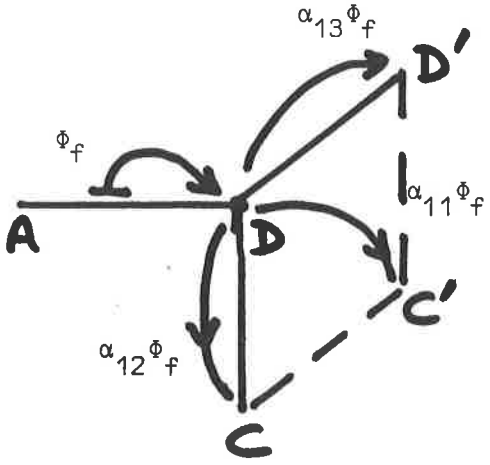

Fig. 14

$$\left.\begin{aligned}
(\alpha_{23}+\alpha_{22})\nu_2 + (\alpha_{32}+\alpha_{33})\nu_3 &= \nu_2\nu_3 \\
(\alpha_{31}+\alpha_{33})\nu_3 + (\alpha_{13}+\alpha_{11})\nu_1 &= \nu_3\nu_1 \\
(\alpha_{12}+\alpha_{11})\nu_1 + (\alpha_{21}+\alpha_{22})\nu_2 &= \nu_1\nu_2 \\
\alpha_{11}\nu_1 + \alpha_{22}\nu_2 + \alpha_{33}\nu_3 &= \nu_1\nu_2\nu_3 ,
\end{aligned}\right\} \qquad (7.16)$$

which has the particular solution

$$\alpha_{11} = \tfrac{1}{3}\nu_2\nu_3,\ \alpha_{22} = \tfrac{1}{3}\nu_3\nu_1,\ \alpha_{33} = \tfrac{1}{3}\nu_1\nu_2. \qquad (7.17)$$

$$\left.\begin{aligned}
\alpha_{23} + \alpha_{22} &= \tfrac{1}{2}\nu_3, \quad \alpha_{32} + \alpha_{33} = \tfrac{1}{2}\nu_2 \\
\alpha_{31} + \alpha_{33} &= \tfrac{1}{2}\nu_1, \quad \alpha_{13} + \alpha_{11} = \tfrac{1}{2}\nu_3 \\
\alpha_{12} + \alpha_{11} &= \tfrac{1}{2}\nu_2, \quad \alpha_{21} + \alpha_{22} = \tfrac{1}{2}\nu_1,
\end{aligned}\right\} \qquad (7.18)$$

which leads to

$$\left.\begin{aligned}
\alpha_{23} &= \nu_3(\tfrac{1}{2} - \tfrac{1}{3}\nu_1), \quad \alpha_{32} = \nu_2(\tfrac{1}{2} - \tfrac{1}{3}\nu_1) \\
\alpha_{31} &= \nu_1(\tfrac{1}{2} - \tfrac{1}{3}\nu_2) \quad \alpha_{13} = \nu_3(\tfrac{1}{2} - \tfrac{1}{3}\nu_2) \\
\alpha_{12} &= \nu_2(\tfrac{1}{2} - \tfrac{1}{3}\nu_3) \quad \alpha_{21} = \nu_1(\tfrac{1}{2} - \tfrac{1}{3}\nu_3).
\end{aligned}\right\} \qquad (7.19)$$

The scheme can be made third order by adding further transfers: the u_{xxx}, u_{yyy}, u_{zzz} terms in the truncation error are matched by double transfers of

$$\tau_{11}\Phi_f, \qquad \tau_{12}\Phi_g, \qquad \tau_{33}\Phi_h,$$

as in the one and two-dimensional cases, where

$$\tau_{ii} = \tfrac{1}{6}(1-|\nu_i|^2) \qquad (i = 1,2,3). \qquad (7.19)$$

For the u_{xxy}, u_{yyx} etc. terms we can use double lateral transfers, as in the two-dimensional case: i.e. we transfer $\tau_{12}\Phi_f$, $\tau_{13}\Phi_f$ etc.

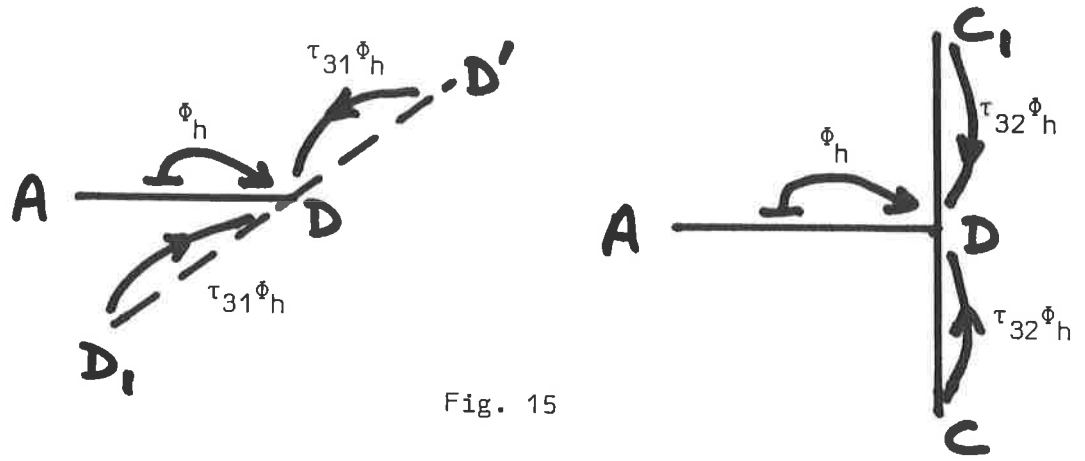

Fig. 15

as illustrated partially in Fig. 15. We then find that the

$$\left.\begin{aligned} \tau_{12} &= \tfrac{1}{2}\nu_2(1-\nu_2) & \tau_{21} &= \tfrac{1}{2}\nu_1(1-\nu_1) \\ \tau_{23} &= \tfrac{1}{2}\nu_3(1-\nu_3) & \tau_{32} &= \tfrac{1}{2}\nu_2(1-\nu_2) \\ \tau_{31} &= \tfrac{1}{2}\nu_1(1-\nu_1) & \tau_{13} &= \tfrac{1}{2}\nu_3(1-\nu_3)\ , \end{aligned}\right\} \qquad (7.20)$$

as in the two-dimensional case.

We conclude the discussion of three-dimensional schemes by consideration of the LB property in a second order scheme using B-functions. For second order we omit transfers involving τ_{ij} but, because it is the only genuinely three-dimensional term, retain the transfers which match u_{xyz} in the truncation error, namely, those involving α_{11}, α_{22}, α_{33}.

For this second order scheme, with $\nu_1 \geq 0$, $\nu_2 \geq 0$, $\nu_3 \geq 0$, the new value of u_D, without using B-functions (or with $B(b_1,b_2) = b_1$) is (see Fig. 13)

$$u^D = u_D + \sum\Big[\Phi_f - \alpha_1\Phi_f + \alpha_1\Phi_f^{DE} - \alpha_{12}\Phi_f + \alpha_{12}\Phi_f^{BC} - \alpha_{13}\Phi_f + \alpha_{13}\Phi_f^{A'D'}$$

$$- \alpha_{11}\Phi_f + \alpha_{11}\Phi_f^{B'C'}\Big], \qquad (7.21)$$

where the sum is over f,g,h, with appropriate 1,2,3 and superfix permutations. Using B-functions with β,γ defined by (1.29) we obtain

$$u^D = u_D + \sum_{\substack{f,g,h \\ 1,2,3}}\Big[1 - \beta\alpha_1 + \alpha\gamma_1 - \beta\alpha_{12} + \gamma\alpha_{12} - \beta\alpha_{13} + \gamma\alpha_{13} - \beta\alpha_{11} + \gamma\alpha_{11}\Big]\Phi_f \qquad (7.22)$$

$$= u_D + q_f\Phi_f + q_g\Phi_g + q_h\Phi_h, \qquad (7.23)$$

where $q_f = 1 - \beta\alpha_1 + \gamma\alpha_1 - \beta\alpha_{12} + \gamma\alpha_{12} - \beta\alpha_{13} + \gamma\alpha_{13} - \beta\alpha_{11} + \gamma\alpha_{11}$ (7.24)

with similar expressions for q_g, q_h. Thus

$$u^D = u_D - \nu_1 q_f(u_D - u_A) - \nu_2 q_g(u_D - u_C) - \nu_3 q_h(u_D - u_{D'}) \tag{7.25}$$

and the LB property depends on the inequalities

$$\nu_1 q_f \geq 0, \quad \nu_2 q_g \geq 0, \quad \nu_3 q_h \geq 0 \tag{7.26}$$

$$1 - \nu_1 q_f - \nu_2 q_g - \nu_3 q_h \geq 0. \tag{7.27}$$

For (7.26) we require for example $q_f \geq 0$, i.e. from (7.24)

$$\beta\alpha_1 - \gamma\alpha_1 + \beta\alpha_{12} - \gamma\alpha_{12} + \beta\alpha_{13} - \gamma\alpha_{13} + \beta\alpha_{11} - \gamma\alpha_{11} \leq 1 \ . \tag{7.28}$$

If we choose the B-function as in (1.26) then β, γ satisfy the conditions (4.24), namely,

$$0 \leq \beta \leq 1, \qquad 0 \leq \gamma \leq 1 \tag{7.29}$$

and hence, from (7.28), we need

$$\alpha_1 + \alpha_{12} + \alpha_{13} + \alpha_{11} \leq 1,$$

i.e.
$$\tfrac{1}{2}(1-\nu_1) + \nu_2(\tfrac{1}{2} - \tfrac{1}{3}\nu_3) + \nu_3(\tfrac{1}{2} - \tfrac{1}{3}\nu_2) + \tfrac{1}{3}\nu_2\nu_3 \leq 1 \tag{7.30}$$

or

$$\nu_1 - \tfrac{1}{3}(\nu_2+\nu_3) + \tfrac{2}{3}(1-\nu_2)(1-\nu_3) \leq \tfrac{5}{3}, \tag{7.31}$$

which is satisfied if $\nu_1 \leq 1$, $\nu_2 \leq 1$, $\nu_3 \leq 1$.

Considering now the condition (7.27) we need

$$1 - \textstyle\sum \nu_1(1 + \alpha_1 + \alpha_{12} + \alpha_{13} + \alpha_{11}) \geq 0, \tag{7.32}$$

i.e.
$$1 - \textstyle\sum \nu_1(\tfrac{1}{2}(3-\nu_1) + \nu_2(\tfrac{1}{2} - \tfrac{1}{3}\nu_3) + \nu_3(\tfrac{1}{2} - \tfrac{1}{3}\nu_2) + \tfrac{1}{3}\nu_2\nu_3) \geq 0$$

or
$$1 - \textstyle\sum \tfrac{1}{2}\nu_1(3-\nu_1) + \tfrac{1}{2}\nu_1(\nu_2+\nu_3) - \tfrac{1}{3}\nu_1\nu_2\nu_3) \geq 0. \tag{7.33}$$

This is met if

$$\frac{3}{2}(\nu_1+\nu_2+\nu_3) - \frac{1}{2}(\nu_1^2+\nu_2^2+\nu_3^2) + \frac{1}{2}\nu_1(\nu_1+\nu_3) + \frac{1}{2}\nu_2(\nu_3+\nu_1) + \frac{1}{2}\nu_3(\nu_1+\nu_2)$$

$$-\nu_1\nu_2\nu_3 \leq 1 \qquad (7.34)$$

or

$$\frac{3}{2}(\nu_1+\nu_2+\nu_3) + \frac{1}{2}(\nu_3\nu_1+\nu_1\nu_2+\nu_2\nu_3) - \nu_1\nu_2\nu_3 \leq 1, \qquad (7.35)$$

which is satisfied in a neighbourhood of $(\nu_1,\nu_2,\nu_3) = (0,0,0)$.

There are many other solutions for the coefficients in (7.16) and the conditions stated above should not be regarded as optimal.

Since the coefficients of u_D, u_A, u_C, $u_{D'}$ are non-negative we have the LB property

$$u^D \in [\text{min.}, \text{max}]\ (u_D,u_A,u_C,u_{D'}). \qquad (7.36)$$

It is possible to analyse the conditions under which the LB property holds in much greater detail than this, but there are no new principles and the algebra is very intricate.

§8. Conclusion

The LB principle exhibited above together with the principle of TVD (total variation diminishing)[12] has been used [13], [14] to prove the convergence of difference schemes to weak solutions of the underlying hyperbolic differential equation. More directly, it is a control on the slope of the solution preventing any new extrema occurring (a type of monotonicity preservation). Convergence to the unique physical solution requires a further element [15] which has been developed for such schemes as the above [16] and is easily incorporated into them.

The extension to systems in 2 and 3 dimensions needs further analysis of the Roe matrix in these cases. First ideas are that the fluctuations will need to be decomposed on to the eigenvalues of a Roe matrix which,

for the equation

$$\underset{\sim}{u}_t + A\underset{\sim}{u}_x + B\underset{\sim}{u}_y = 0, \tag{8.1}$$

would be of the form

$$A \cos \delta + B \sin \delta \ . \tag{8.2}$$

For an x-arrow the angle δ would be the inclination of the shock normal (if present) to the x-direction and would need to be estimated from local u values. For example, one such estimation in the case of the x-arrow would be (see Fig. 16) the average (over the components of $\underline{u}$) of δ given by

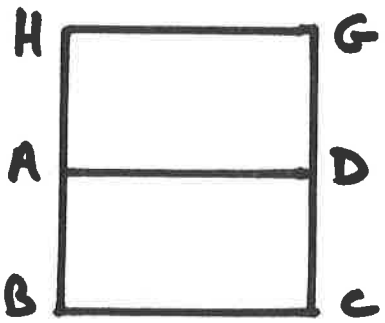

Fig. 16

$$\tan \delta = \frac{u_D - u_A}{\frac{1}{4}(u_G + u_H - u_B - u_C)} \ . \tag{8.3}$$

Finally, distortion of the basic grid used here would not destroy conservation but would reduce accuracy to an extent. The LB principle, however, would continue to hold where appropriate.

Acknowledgements

Thanks are due to P.L. Roe and P.K. Sweby for useful discussions.

References

[1] Roe, P.L. RAE Technical Report TR 81047, (1981).

[2] Sweby, P.K. & Baines, M.J. Num. An. Report 8/81, Univ. of Reading, (1982).

[3] Godunov, S.K., Mat. Sb., 47, p. 271, (1959).

[4] Lax, P.D. & Wendroff, B., Comm. Pure. Appl. Math. 13, p. 217, (1960)

[5] Warming, R.F. & Beam, R., AIAA Journal 14, p. 1241, (1976).

[6] Roe, P.L. & Baines, M.J., Proc. 7th GAMM Conference, Paris, p. 281, (Vieweg), (1981).

[7] Roe, P.L., private communication.

[8] Sweby, P.K., Ph.D. Thesis, Univ. of Reading, (1982).

[9] Roe, P.L., J. Comput. Phys., 43, p. 357, (1981).

[10] Sells, C.C.L., RAE Technical Report TR 80065, (1980).

[11] Baines, M.J., Num. An. Report 4/81, Univ. of Reading, (1981).

[12] Harten, A., NYU Report DOE/ER/030 77-167, (1982).

[13] Sanders, R., Ph.D. Thesis, Univ. of California, Los Angeles, (1981).

[14] Harten, A. & Lax, P.D. SIAM J. Num. Anal. 18, p. 285, (1981).

[15] Harten, A. & Hyman, J.M., Report LA-9105, Los Alamos National Lab., (1982)

[16] Sweby, P.K., Num. An. Report 8/82, Univ. of Reading, (1982).

APPENDIX A - by P.L. Roe

Stability of the Scheme of §5

If the Fourier angles are θ_1, θ_2, the amplification factor λ for the scheme in §5 is (see Fig. A1)

$$\lambda = 1 - \frac{\nu_1}{2}(e^{i\theta_1} - e^{-i\theta_1}) + \frac{\nu_1^2}{2}(e^{i\theta_1} + e^{-i\theta_1} - 2)$$
$$- \frac{\nu_2}{2}(e^{i\theta_2} - e^{-i\theta_2}) + \frac{\nu_2^2}{2}(e^{i\theta_2} + e^{-i\theta_2} - 2)$$
$$+ \nu_1\nu_2(1 - e^{-i\theta_1})(1 - e^{-i\theta_2}). \qquad (A.1)$$

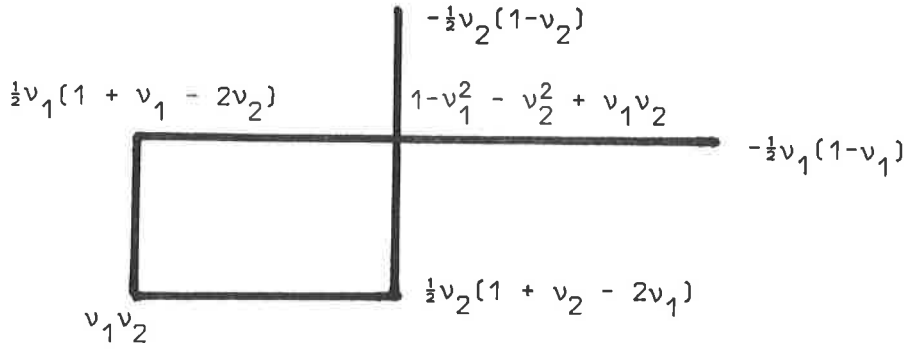

Fig. A.1

Hence

$$\lambda - 1 = -i\nu_1 \sin\theta_1 + \nu_1^2(\cos\theta_1 - 1)$$
$$- i\nu_2 \sin\theta_2 + \nu_2^2(\cos\theta_2 - 1)$$
$$+ \nu_1\nu_2(1 - \cos\theta_1 + i\sin\theta_1)(1 - \cos\theta_2 + i\sin\theta_2)$$
$$= X + iY, \qquad (A.2)$$

where

$$X = \nu_1^2(\cos\theta_1 - 1) + \nu_2^2(\cos\theta_2 - 1)$$
$$+ \nu_1\nu_2[1 - \cos\theta_1 - \cos\theta_2 + \cos(\theta_1+\theta_2)] \qquad (A.3)$$

$$Y = -\nu_1 \sin\theta_1 - \nu_2 \sin\theta_2 + \nu_1\nu_2[\sin\theta_1 + \sin\theta_2 - \sin(\theta_1+\theta_2)] \qquad (A.4)$$

The condition for stability is $\lambda\lambda^* \leq 1$

$$\text{i.e.} \qquad (1 + X + iY)(1 + X - iY) \leq 1,$$

$$1 + 2X + X^2 + Y^2 \leq 1 ,$$

$$\text{or} \qquad X^2 + Y^2 + 2X \leq 0. \tag{A.5}$$

For sufficiently small ν_1, ν_2, we need consider only the quadratic terms in this expression, i.e.

$$Q = (\nu_1 \sin\theta_1 + \nu_2 \sin\theta_2)^2 + 2\nu_1^2(\cos\theta_1 - 1) + 2\nu_2^2(\cos\theta_2 - 1)$$

$$+ 2\nu_1\nu_2 [1 - \cos\theta_1 - \cos\theta_2 + \cos\overline{\theta_1 + \theta_2}] \tag{A.6}$$

$$= \nu_1^2(\sin^2\theta_1 + 2\cos\theta_1 - 2) + \nu_2^2(\sin^2\theta_2 + 2\cos\theta_2 - 2)$$

$$+ 2\nu_1\nu_2 [1 - \cos\theta_1 - \cos\theta_1 + \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 + \sin\theta_1\sin\theta_2]$$

$$= -\nu_1^2(1 - \cos\theta_1)^2 - \nu_2^2(1 - \cos\theta_2)^2$$

$$+ 2\nu_1\nu_2(1 - \cos\theta_1)(1 - \cos\theta_2)$$

$$= -[\nu_1(1 - \cos\theta_1) - \nu_2(1 - \cos\theta_2)]^2 \tag{A.7}$$

which is negative as required, except for the special case $\nu_1 = \nu_2$, $\theta_1 = \theta_2$. In this special case

$$\left.\begin{aligned} X &= \nu^2(\cos 2\theta - 1) = -2\nu^2\sin^2\theta \\ Y &= \nu^2(2\sin\theta - \sin 2\theta) - 2\nu\sin\theta \\ &= 2\nu^2\sin\theta(1 - \cos\theta) - 2\nu\sin\theta. \end{aligned}\right\} \tag{A.8}$$

Hence

$$X^2 + Y^2 + 2X = 4\nu^4\sin^4\theta + 4\nu^4\sin^2\theta(1 - \cos\theta)^2 - 8\nu^3\sin^2\theta(1 - \cos\theta) + 4\nu^2\sin^2\theta$$

$$- 4\nu^2\sin^2\theta$$