Positive Schemes for Solving Multi-dimensional Hyperbolic Systems of Conservation Laws

Xu-Dong Liu[*] Peter D. Lax[†]

January 18, 1995

To Tony Jameson, master of computational fluid dynamics,
with friendship and admiration

Abstract

This paper introduces a new positivity principle for numerical schemes for solving hyperbolic systems of conservation laws in many space variables. A family of positive schemes has been designed that are second order accurate in the space variables. Second order accuracy in time is achieved by a second order accurate energy conserving Runge-Kutta method due to Shu, [17] and [18].

[0] submitted to CFD Journal

[*] Courant Institute of Mathematical Sciences, 251 Mercer street, New York, NY 10012, xliu@cims.nyu.edu. The author was supported by National Science Foundation grants DMS-9112654 and DMS-9406467, and Air Force grant AFOSR F49620-94-1-0132

[†] Courant Institute of Mathematical Sciences, 251 Mercer street, New York, NY 10012. The author was supported by Applied Mathematical Sciences Program of the U.S. Department of Energy under contract DE-FG02-88ER25053

Our positive schemes have a very simple structure using characteristic decomposition, treating each space variable separately. Only about 150 (180) FORTRAN 77 lines are needed for solving the two(three)-dimensional Euler equations. Positive schemes can be constructed for all hyperbolic conservation laws where Roe averages are known, and where the eigenvectors and eigenvalues of the Roe matrix can be evaluated explicitly.

Excellent numerical results have been obtained. Comparison with successful high order accurate schemes shows that the positive schemes are competitive with the best schemes available today.

1 Introduction

Over the last decade a great deal of effort has been devoted to designing total variation diminishing (TVD) numerical schemes for solving hyperbolic conservation laws, see Harten's basic paper [4]. Strictly speaking, TVD schemes exist only for scalar conservation laws, and for linear hyperbolic systems in one space variable. Hyperbolic systems of conservation laws in one space variable can be treated in a formal manner. No TVD schemes can exist for non-scalar hyperbolic systems in more than one space variable, linear or non-linear. The possibility of focusing shows that total variation norms are not bounded. The only functional known to be bounded for solutions of linear hyperbolic equations is **energy**. This is particularly easy to show, as Friedrichs has done in [3], for symmetric hyperbolic linear systems

$$U_t + \sum_{s=1}^{d} A_s U_{x_s} = 0, \tag{1}$$

A_s real symmetric matrices, whose dependence on x is Lipschitz continuous. It is easy to show in this case that the L^2 norm of a solution is of bounded growth

$$\|U(t)\| \le e^{ct}\|U(0)\|, \tag{2}$$

where c is related to the Lipschitz constant. Friedrichs [3] has shown that solutions of such equations can be approximated by solutions of difference equations of form

$$U_J^{m+1} = \sum_K C_K U_{J+K}^m, \tag{3}$$

J a lattice point, U_J^m an approximation to the value of the solution $U(x,t)$ of (1) at the point $x = (j_1\Delta x_1, \cdots, j_d\Delta x_d)$ at time $t = m\Delta t$. The coefficient matrices C_K are required to have the following properties:

i) C_K is symmetric,

ii) C_K is nonnegative,

iii) $\sum_K C_K = I$, where I is the identity matrix,

iv) $C_K = 0$ except for a finite set of K,

v) C_K depends Lipschitz continuously on x.

We show now that the l^2 norm of solutions of difference schemes satisfying these properties has bounded growth:

$$\|U^{m+1}\| \leq (1 + const\Delta)\,\|U^m\|, \tag{4}$$

where the discrete l^2 norm is defined as

$$\|U\|^2 = \sum \mid U_J \mid^2 .$$

The value of the constant in (4) depends on the Lipschitz constant.

Proof: Take the scalar product of (3) with U_J^{m+1}

$$\mid U_J^{m+1} \mid^2 = \sum (U_J^{m+1}, C_K U_{K+J}^m). \tag{5}$$

Since C_K is symmetric and nonnegative, $(U, C_K V)$ can be regarded as an inner product, to which the Schwarz inequality applies; combined with the inequality between arithmetic and geometric mean we get

$$(U, C_K V) \leq \sqrt{(U, C_K U)}\sqrt{(V, C_K V)} \leq \frac{1}{2}(U, C_K U) + \frac{1}{2}(V, C_K V). \tag{6}$$

Using this on the right side of (5) gives

$$\mid U_J^{m+1} \mid^2 \leq \frac{1}{2}\sum (U_J^{m+1}, C_K U_J^{m+1}) + \frac{1}{2}\sum (U_{J+K}^m, C_K U_{J+K}^m). \tag{7}$$

Carrying out the summation with respect to K, using condition iii) and multiplying by 2 gives

$$\mid U_J^{m+1} \mid^2 \leq \sum \left(U_{J+K}^m, C_K U_{J+K}^m\right). \tag{8}$$

Now sum with respect to J, and introduce $K + J = N$ as new index of summation:

$$\sum_J \mid U_J^{m+1} \mid^2 \leq \sum_{N,K} \left(U_N^m, C_K(N-K) U_N^m\right). \tag{9}$$

Replace on the right $C_K(N-K)$ by $C_K(N)$; using Lipschitz continuity, and the fact that K ranges over a finite stencil, we get that the right side of (9) is

$$\leq \sum_{N,K} (U_N^m, C_K(N)U_N^m) + O(\Delta) \sum \mid U_N^m \mid^2,$$

which, using iii), is equal to

$$\sum \mid U_N^m \mid^2 (1 + O(\Delta)) .$$

Setting this into (9) gives (4). Q.e.d..

In the usual variables the Euler equations are not symmetric but symmetrizable, i.e. can be written in the form

$$B_0 U_t + \sum B_s U_{x_s} = 0, \tag{10}$$

where all the B_s are symmetric and B_0 is positive. Multiplying (10) by B_0^{-1} gives

$$U_t + \sum A_s U_{x_s} = 0, \tag{11}$$

where

$$A_s = B_0^{-1} B_s. \tag{12}$$

Since B_0 is positive, it has a symmetric square root S:

$$B_0 = S^2, \qquad S^T = S. \tag{13}$$

Multiplying (12) by S on the left, S^{-1} on the right gives

$$S A_s S^{-1} = S^{-1} B_s S^{-1}. \tag{14}$$

The matrix on the right is symmetric; therefore so is the matrix on the left. In words: the matrices A_s can be symmetrized by the same similarity transformation.

In section 2 we show how to construct difference schemes that can be put in form (3), where the $C_K(J)$ have the following two properties:

a) $\sum\limits_K C_K(J) = I$,
b) C_K is of the form,

$$C_K(J) = \sum \tilde{A}_s(J \pm \frac{1}{2}e_s),$$

where the matrix $\tilde{A}_s(J \pm \frac{1}{2}e_s)$ commutes with $A_s(J \pm \frac{1}{2}e_s)$, and has positive eigenvalues; here e_s is the unit vector in the x_s direction. It follows from a) and b) that

$$S\tilde{A}_s S^{-1},$$

is a symmetric and nonnegative matrix, where $S = S(J \pm \frac{1}{2}e_s)$.

Assume now that the matrices B_s depend Lipschitz continuously on x; then so does $S = B_0^{1/2}$ and C_K. We define D_K by

$$D_K(J) = S(J)C_K(J)S^{-1}(J). \tag{15}$$

It follows that the $D_K(J)$ have the following properties:

i) D_K differs by $O(\Delta)$ from a symmetric matrix,

ii) The symmetric part of D_K has eigenvalues that are $\geq -O(\Delta)$,

iii) $\sum\limits_K D_K = I$,

iv) $D_K = 0$ except for a finite number of K,

v) D_K depends Lipschitz continuously on x.

Under these conditions we can deduce an analogue of inequality (4). We introduce $V_J = S(J)U_J$ as new variable and multiply (3) by $S(J)$; we obtain

$$V_J^{m+1} = S(J)U_J^{m+1} = \sum S(J)C_K U_{J+K}^m = \sum S(J)C_K S^{-1}(J)V_{J+K}^m = \sum D_K(J)V_{J+K}^m.$$

We can now proceed as before and deduce that

$$\|V^{m+1}\| \leq (1 + O(\Delta))\|V^m\|.$$

In terms of U, this can be expressed as

$$\|U^{m+1}\|_S^2 \leq (1 + O(\Delta))\|U^m\|_S^2, \tag{16}$$

where $\|U\|_S^2$ is defined as

$$\|U\|_S^2 = \sum (V_J, V_J) = \sum (SU_J, SU_J) = \sum (U_J, S^2 U_J) = \sum (U_J, B_0 U_J).$$

If B_0 depends on t as well, as it does in our case, then also the norm $\|U\|_S$ depends on t. Since B_0 is assumed to depend Lipschitz continuously on t, it follows that

$$\|U\|_{S(t_{m+1})} \leq (1 + O(\Delta))\|U\|_{S(t_m)}.$$

Combined with (16) we deduce recursively that

$$\|U\|_{S(t_m)} \leq (1 + mO(\Delta))\|U\|_{S(0)}.$$

Q.e.d..

The difference schemes studied in this paper are nonlinear; when we write them in linear form (3), the coefficient matrices C_K depend on the solution being computed. In the applications of interest these solutions contain shocks and contact discontinuities; therefore the coefficient matrices $C_K(J)$ not only fail to be Lipschitz continuous, they are not even continuous. So our analysis of the boundedness of the l^2 norm of the solution is not applicable. A possible way of salvaging our argument is to note that in inequality (6) the left side is substantially smaller than the right side, unless the vector U and V are nearly equal. For unless the vectors U and V are nearly proportional, the Schwarz inequality is a strict inequality. Similarly, unless (U, CU) and (V, CV) are nearly equal, their geometric mean is substantially less than their arithmetic mean. This shows that at a discontinuity the left side of (8) is substantially less than the right side. We don't at this moment see how to show that this gain is enough to counterbalance what we may lose when we replace on the right in inequality (9) the matrix $C_K(N-K)$ by $C_K(N)$, but at least we have found a plausible reason why our scheme is as stable as it appears to be in numerical experiments.

In this paper we construct finite difference approximation in conservation form to symmetric and symmetrizable hyperbolic systems of conservation laws that are second order accurate and yet can be written in positive form (3). This is achieved by combining a second order accurate scheme with an upwind scheme, using an appropriate flux limiter, see LeVeque [13] for a general discussion.

In verifying positivity we use the fact that the sum of positive matrices is positive, and that the product of commuting positive matrices is positive. We use a combination of several such schemes. One of them, which we call the least dissipative, is, in one space dimension, similar to one proposed by Helen Yee in [23]. In [9], Jameson discusses this scheme in the scalar case; he observes that it has the local extremum diminishing (LED) property. Fluxes in the several coordinate directions are handled independently of each other. Second order accuracy in time is achieved by using an energy preserving second order time step.

Our positive schemes have the following advantages.

Great simplicity and very low cost. No Riemann solvers or MUSCL-type reconstruction is needed. By using dimension-by-dimension split and the method of lines, cost for each time step is proportional to dimension. Thanks to Roe averages only simple algebraic operations are involved, programs are short and run fast. For solving the two-dimensional Euler equations, the positive schemes need less than 150 lines in FORTRAN 77 (excluding input/output, initial, boundary condition subroutines). For three-dimension, the positive schemes need less than 180 lines. Positive schemes are at least as fast as MUSCL schemes in multi-dimension. One-dimensional positive schemes are 40% faster than MUSCL schemes.

Wide range of applicability. Positive schemes can be applied to all hyperbolic conservation laws where Roe averages are known, and where the eigenvalues and eigenvectors of the Roe matrix can be evaluated explicitly.

The disadvantage of our positive schemes is that they are limited to first order accuracy at extrema, although second order elsewhere. ENO schemes [7], [18], [14] and [10] have no such limitation.

High resolution numerical experiments show that positive schemes are capable of resolving multi-dimensional hyperbolic systems of conservation laws. The numerical results are comparable to the numerical results using ENO [7] [18] and PPM [22].

The organization of this paper is as follows. Section 2.1 describes the notion of a positive scheme for solving symmetric nonlinear equations. In section 2.2 we construct two positive schemes, both second order accurate. The scheme used in practice is an appropriate combination of them. Another family of one-dimensional positive schemes based on Lax-Wendroff are constructed in the section. The Runge-Kutta method to achieve second order accuracy in time is described in section 2.2. In section 2.3 we explain how to combine the fluxes in all space direction. In section 2.4 we explain why the scheme constructed in section 2.3 work also for symmetrizable systems.

The numerical experiments are described in section 3. In section 3.1 several one-dimensional problems are solved, a Riemann problem, an interaction of two waves, and low density and internal energy Riemann problem. In section 3.2 we present a calculation of a plane shock diffracted by a wedge; the flow is rather complicated, containing double Mach reflection. The second example is flow in a wind tunnel obstructed by a step; this flow, too, is rather complicated, containing a strong rarefaction wave, a Mach reflection and many simple reflections. Most details are well resolved by our positive scheme.

We thank Tony Jameson for his valuable comments, and Stanley Osher and Chi-Wang Shu for their valuable comments on ENO. We thank Bjorn Sjogreen for calling our attention to references [23] and [9]. We also thank Smadar Karni for helpful discussions.

2 Positivity Principle and Positive Schemes

2.1 Positivity Principle

We consider multi-dimensional hyperbolic systems of conservation laws

$$U_t + \sum_{s=1}^{d} F_s(U)_{x_s} = 0, \tag{17}$$

where $U = (u_1, \cdots, u_n)^T \in \mathbf{R^n}$ and $x = (x_1, x_2, \cdots, x_d) \in \mathbf{R^d}$. We assume that all Jacobian matrices $A_s = \bigtriangledown F_s$ are symmetric or simultaneously symmetrizable with the same similarity transformation. We construct a uniform Cartesian grid $\{\Omega_J\}$ in $\mathbf{R^d}$, where $J = (j_1, j_2, \cdots, j_d)$ is a lattice point in which all j_s are integers. In this uniform grid we denote cell-averages as $U_J = \frac{1}{|\Omega|} \int\limits_{\Omega_J} U(x,t)\,dx$, where $\mid \Omega \mid$ is the volume of the cell Ω_J.

Conservative schemes are of the form

$$U_J^* = U_J - \sum_{s=1}^{d} \frac{\Delta t}{\Delta x_s}[F_{J+1/2e_s} - F_{J-1/2e_s}] \tag{18}$$

where Δt is the time step and Δx_s is the spatial step in the x_s dimension.

We call the schemes (18) **positive**, if we can write the right side as

$$U_J^* = \sum_K C_K U_{J+K}, \tag{19}$$

so that the coefficient matrices C_K, which themselves depend on all the U_{J+K} that occur in (18), have the following properties

(i) Each C_K is symmetric positive definite i.e. $C_K \geq 0$;

(ii) $\sum\limits_K C_K = I$.

Note that since the right side of (18) is a non-linear function of the U_{J+K}, there are many ways of writing it in form (19).

In section 2.4 we shall extend the notion of positive scheme to symmetrizable systems.

2.2 Positive Scheme in One-Dimension

We consider one-dimensional hyperbolic systems of conservation laws

$$U_t + F(U)_x = 0. \tag{20}$$

We consider schemes in conservation form, i.e.

$$U_j^* = U_j - \frac{\Delta t}{\Delta x}[F_{j+1/2} - F_{j-1/2}]. \tag{21}$$

The numerical flux F is obtained by mixing a second order accurate scheme with numerical flux F^{acc}, with another dissipative scheme with numerical flux F^{diss}. We combine them using the flux limiting philosophy [1] and [8] of having a numerical flux of the form

$$F = F^{diss} + L(F^{acc} - F^{diss}), \tag{22}$$

where the flux limiter L is near the identity when the flow is smooth, and near zero otherwise. For the accurate scheme we take not Lax-Wendroff but central differencing as in [23] and [9]. For the dissipative scheme we have two candidates of upwind type; one is the least dissipative, the other more dissipative. Different flux limiters are chosen in the two cases. This way we come up with two distinct positive schemes; an appropriate combination of the two positive schemes is again positive. Note that although centered differencing leads to an unconditionally unstable scheme, the combined scheme is stable.

The schemes are second order in the smooth regions with respect to space discretization. To achieve second order accuracy in time we employ the second order energy preserving Runge-Kutta of Shu [17] and [18].

Both dissipative schemes use the Roe average $\hat{U} = \hat{U}(U, V)$, which has the property

$$F(U) - F(V) = A(\hat{U})(U - V), \tag{23}$$

where $A = \bigtriangledown F$. Roe [16] has shown how to construct $\hat{U}$ for the Euler flux in all dimensions. We write $F(U_{j+1}) - F(U_j) = A_{j+1/2}(U_{j+1} - U_j)$ and then drop the subscript $j + 1/2$ to avoid clutter. The spectral resolution of A is $A = R\Lambda R^{-1}$, Λ diagonal. The entries λ^i of Λ are the eigenvalues of A, the columns of R the right eigenvectors, the rows of R^{-1} the left eigenvectors.

For the least dissipative scheme we define the "absolute value" of A as $\mid \mathring{A} \mid = R \mid \Lambda \mid R^{-1}$ where $\mid \Lambda \mid = diag\,(\mid \lambda^i \mid)$. We construct the numerical flux $F_{j+1/2}$ in (21) as a mixture of a 2nd order accurate centered difference flux

$$F^c_{j+1/2} = \frac{F(U_j) + F(U_{j+1})}{2} \tag{24}$$

and a first order accurate upwind flux

$$F^{up}_{j+1/2} = \frac{F(U_j) + F(U_{j+1})}{2} - \frac{1}{2} \mid \mathring{A} \mid (U_{j+1} - U_j)\,. \tag{25}$$

The mixture is accomplished with the aid of a limiter L^0:

$$F^0_{j+1/2} = \frac{F(U_j) + F(U_{j+1})}{2} - \frac{1}{2} \mid \mathring{A} \mid (U_{j+1} - U_j) + \frac{1}{2} L^0 \mid \mathring{A} \mid (U_{j+1} - U_j) \tag{26}$$

where L^0 is a limiter i.e. matrix that is near I in a smooth region of the solution, and near 0 otherwise. This makes $F^0_{j+1/2}$ the 2nd order accurate central difference flux in the smooth regions and the least dissipative flux otherwise.

We choose L^0 to commute with A, specially we take

$$L^0 = R\Phi^0 R^{-1}, \quad \text{where} \quad \Phi^0 = diag\left(\phi^0(\theta^i)\right) \quad \text{is a diagonal matrix.}$$

We take $\phi^0(\theta)$ as some limiter function such that, see Sweby [20],

$$0 \leq \phi^0\left(\theta\right), \frac{\phi^0\left(\theta\right)}{\theta} \leq 2, \quad \phi^0(1) = 1.$$

Here $\phi^0(\theta^i)$ could be different for each i.

We define θ^i in terms of the i-th component of differences of the characteristic variables:

$$\theta^i = \frac{l^i(U_{j'+1} - U_{j'})}{l^i(U_{j+1} - U_j)} \qquad 1 \leq i \leq n, \qquad \text{where} \qquad j^{'} = \begin{cases} j-1 & \text{if} \quad \lambda^i \geq 0 \\ j+1 & \text{otherwise.} \end{cases} \tag{27}$$

The l^i are the left eigenvectors of A associated with λ^i, This completes the construction of the least dissipative scheme.

Our choice in (27) is somewhat different from Helen Yee's scheme in [23], where she define θ^i as

$$\theta^i = \frac{l^i_{j'+1/2}(U_{j'+1} - U_{j'})}{l^i_{j+1/2}(U_{j+1} - U_j)}.$$

We exploit the precise form of (27) in proving positivity.

We now prove that scheme (26) is positive. We define

$$\mathring{A}^+ = R\Lambda^+ R^{-1}, \qquad \mathring{A}^- = R\Lambda^- R^{-1}, \tag{28}$$

where two diagonal matrices are

$$\Lambda^+ = \begin{pmatrix} \max(\lambda^1, 0) & & \\ & \ddots & \\ & & \max(\lambda^n, 0) \end{pmatrix}, \qquad \Lambda^- = \begin{pmatrix} \min(\lambda^1, 0) & & \\ & \ddots & \\ & & \min(\lambda^n, 0) \end{pmatrix}.$$

Obviously

$$\begin{aligned} A &= \mathring{A}^+ + \mathring{A}^-, \qquad & \mid \mathring{A} \mid &= \mathring{A}^+ - \mathring{A}^-, \\ \mathring{A}^+ &\geq 0 \qquad & \mathring{A}^- &\leq 0. \end{aligned} \tag{29}$$

By definition (27)

$$\theta^i = \frac{l^i(U_j - U_{j-1})}{l^i(U_{j+1} - U_j)} \qquad \text{when} \quad \lambda^i \geq 0. \tag{30}$$

It follows from (30) that for $\lambda^i \geq 0$,

$$e_i R^{-1}(U_{j+1} - U_j) = l^i(U_{j+1} - U_j) = \frac{1}{\theta^i} l^i(U_j - U_{j-1}) = \frac{1}{\theta^i} e_i R^{-1}(U_j - U_{j-1}), \tag{31}$$

where e_i is the i-th unit row vector and $l^i = e_i R^{-1}$. Define the matrix $\tilde{\Phi}^0$ as

$$\tilde{\Phi}^0 = \begin{pmatrix} \frac{\phi^0(\theta^1)}{\theta^1} & & \\ & \ddots & \\ & & \frac{\phi^0(\theta^n)}{\theta^n} \end{pmatrix};$$

we obtain from (31) that for $\lambda^i \geq 0$

$$e_i \Phi^0 R^{-1}(U_{j+1} - U_j) = e_i \tilde{\Phi}^0 R^{-1}(U_j - U_{j-1}),$$

or equivalently

$$\max(\lambda^i, 0) e_i \Phi^0 R^{-1}(U_{j+1} - U_j) = \max(\lambda^i, 0) e_i \tilde{\Phi}^0 R^{-1}(U_j - U_{j-1}).$$

Using (28) we can write above equation in matrix form:

$$\Lambda^+ \Phi^0 R^{-1}(U_{j+1} - U_j) = \Lambda^+ \tilde{\Phi}^0 R^{-1}(U_j - U_{j-1}).$$

Multiply by R on the left and obtain the crucial equation

$$\mathring{A}^+ R \Phi^0 R^{-1}(U_{j+1} - U_j) = \mathring{A}^+ R \tilde{\Phi}^0 R^{-1}(U_j - U_{j-1}). \tag{32}$$

Here and below we denote

$$\tilde{L}^0 = R \tilde{\Phi}^0 R^{-1}.$$

Therefore, from (26), (29) and (32), and since L^0 commutes with $\mid \mathring{A} \mid$,

$$\begin{aligned} F^0_{j+1/2} &= F^{up}_{j+1/2} + \tfrac{1}{2} \mid \mathring{A} \mid L^0(U_{j+1} - U_j) \\ &= F^{up}_{j+1/2} + \tfrac{1}{2}\mathring{A}^+ L^0(U_{j+1} - U_j) - \tfrac{1}{2}\mathring{A}^- L^0(U_{j+1} - U_j) \\ &= F^{up}_{j+1/2} + \tfrac{1}{2}\mathring{A}^+ \tilde{L}^0(U_j - U_{j-1}) - \tfrac{1}{2}\mathring{A}^- L^0(U_{j+1} - U_j). \end{aligned} \tag{33}$$

We put the subscript $j + 1/2$ back in (33):

$$F^0_{j+1/2} = F^{up}_{j+1/2} + \frac{1}{2}\mathring{A}^+_{j+1/2}\tilde{L}^0_{j+1/2}(U_j - U_{j-1}) - \frac{1}{2}\mathring{A}^-_{j+1/2}L^0_{j+1/2}(U_{j+1} - U_j). \tag{34}$$

Similarly we obtain the flux $F_{j-1/2}$ as

$$F^0_{j-1/2} = F^{up}_{j-1/2} + \frac{1}{2}\mathring{A}^+_{j-1/2}L^0_{j-1/2}(U_j - U_{j-1}) - \frac{1}{2}\mathring{A}^-_{j-1/2}\tilde{L}^0_{j-1/2}(U_{j+1} - U_j). \tag{35}$$

Using (23) and (25) we can write

$$F^{up}_{j+1/2} - F^{up}_{j-1/2} = \frac{1}{2}(A_{j+1/2} - \mid \mathring{A}_{j+1/2} \mid)(U_{j+1} - U_j) + \frac{1}{2}(A_{j-1/2} + \mid \mathring{A}_{j-1/2} \mid)(U_j - U_{j-1}). \tag{36}$$

We obtain from (21,34-36)

$$\begin{aligned} U^*_j = U_j - \tfrac{\Delta t}{2\Delta x}[\quad & (A_{j+1/2} - \mid \mathring{A}_{j+1/2} \mid)(U_{j+1} - U_j) + (A_{j-1/2} + \mid \mathring{A}_{j-1/2} \mid)(U_j - U_{j-1}) \\ & + \mathring{A}^+_{j+1/2}\tilde{L}^0_{j+1/2}(U_j - U_{j-1}) - \mathring{A}^-_{j+1/2}L^0_{j+1/2}(U_{j+1} - U_j) \\ & - \mathring{A}^+_{j-1/2}L^0_{j-1/2}(U_j - U_{j-1}) + \mathring{A}^-_{j-1/2}\tilde{L}^0_{j-1/2}(U_{j+1} - U_j)\,]. \end{aligned} \tag{37}$$

We claim that (37) is the desired positive expression of U^*_j as

$$U^*_j = C_{-1}U_{j-1} + C_0 U_j + C_1 U_{j+1}.$$

We denote $A_{j\pm 1/2}$ as $A_{\pm 1/2}$. From (37) we have the coefficient matrix C_1

$$C_1 = -\frac{\Delta t}{2\Delta x}\left((A_{1/2} - \mid \mathring{A}_{1/2} \mid) - \mathring{A}^-_{1/2}L^0_{1/2} + \mathring{A}^-_{-1/2}\tilde{L}^0_{-1/2}\right). \tag{38}$$

Using (29) we can rewrite C_1 as

$$C_1 = -\frac{\Delta t}{2\Delta x}\left(\mathring{A}^-_{1/2}(2I - L^0_{1/2}) + \mathring{A}^-_{-1/2}\tilde{L}^0_{-1/2}\right) \tag{39}$$

Since ϕ^0 was chosen so that $0 \leq \phi^0(\theta), \frac{\phi^0(\theta)}{\theta} \leq 2$, it follows that $0 \leq \tilde{L}^0$, $L^0 \leq 2I$. By construction $\mathring{A}^- \leq 0$. Since the sum of positive matrices is positive, and the product of commuting positive matrices is positive, it follows that $C_1 \geq 0$.

Similarly the coefficient matrix $C_{-1} \geq 0$.

From (37), the coefficient matrix C_0 is,

$$\begin{aligned} C_0 = I - \frac{\Delta t}{2\Delta x} \Big(& -(A_{1/2} - \mid \mathring{A}_{1/2} \mid) + (A_{-1/2} + \mid \mathring{A}_{-1/2} \mid) \\ & + \mathring{A}^+_{1/2} \tilde{L}^0_{1/2} + \mathring{A}^-_{1/2} L^0_{1/2} - \mathring{A}^+_{-1/2} L^0_{-1/2} - \mathring{A}^-_{-1/2} \tilde{L}^0_{-1/2} \Big) . \end{aligned} \tag{40}$$

Using (29) we can rewrite C_0 as

$$\begin{aligned} C_0 & = I - \frac{\Delta t}{2\Delta x} \Big(-2\mathring{A}^-_{1/2} + \mathring{A}^+_{1/2} \tilde{L}^0_{1/2} + 2\mathring{A}^+_{-1/2} - \mathring{A}^-_{-1/2} \tilde{L}^0_{-1/2} + \mathring{A}^-_{1/2} L^0_{1/2} - \mathring{A}^+_{-1/2} L^0_{-1/2} \Big) \\ & \geq I - \frac{\Delta t}{2\Delta x} [-2\mathring{A}^-_{1/2} + \mathring{A}^+_{1/2} \tilde{L}^0_{1/2}] - \frac{\Delta t}{2\Delta x} [2\mathring{A}^+_{-1/2} - \mathring{A}^-_{-1/2} \tilde{L}^0_{-1/2}]. \end{aligned} \tag{41}$$

The matrices in the brackets are positive. Under the CFL condition

$$\frac{\Delta t}{\Delta x} \max_{1 \leq i \leq n, U} \mid \lambda^i \mid \leq \frac{1}{2} \tag{42}$$

each is $\leq \frac{1}{2} I$; therefore the coefficient matrix $C_0 \geq 0$.

We can read off from (37) that the sum of the three coefficients is the identity matrix I. Therefore the least dissipative scheme (21, 26) is positive.

This scheme is very close to the ones described in [23] from point of view of TVD and in [9] from point of view of local extremum diminishing. The scheme gives sharp shocks but admits entropy violating solutions, because it has zero dissipation in the field of zero eigenvalue. To overcome this difficulty, Harten [4] suggested an entropy fix increasing the magnitude of the eigenvalue up to certain artificial amount.

In the following we shall construct schemes that are more dissipative than the least dissipative positive scheme including those with an entropy fix. Entropy fix destroys the positivity principle for most limiter functions; by choosing the minmod limiter, we obtain dissipative schemes that are positive. However these schemes are smeary because of the minmod limiter and the added dissipation. We then combine the least dissipative positive scheme and the more dissipative positive schemes in an appropriate way to have the best features of both. These combined schemes again are positive.

The more dissipative schemes are constructed as follows. We define a class of "absolute values" of A as $\mid \dot{A} \mid = Rdiag(\mu^i)R^{-1}$, where the diagonal matrix $diag(\mu^i) \geq \mid \Lambda \mid$; for the least dissipative scheme equality holds. We construct the flux $F_{j+1/2}$ in (21) as

$$F^1_{j+1/2} = \frac{F(U_{j+1}) + F(U_j)}{2} - \frac{1}{2} \mid \dot{A} \mid (U_{j+1} - U_j) + \frac{1}{2} L^1 \mid \dot{A} \mid (U_{j+1} - U_j), \tag{43}$$

where the limiter $L^1 = R\, diag\left(\phi^1(\theta^i)\right) R^{-1}$. Note that L^1 commutes with A. Here $\phi^1(\theta)$ is the minmod limiter function

$$\phi^1(\theta) = \begin{cases} 0 & \theta \leq 0, \\ \theta & 0 < \theta \leq 1, \\ 1 & \theta > 1. \end{cases}$$

This has the properties $0 \leq \phi^1(\theta), \frac{\phi^1(\theta)}{\theta} \leq 1, \phi^1(1) = 1$. Every θ^i is defined as in (27). We shall prove below that also the class of more dissipative schemes is positive under an appropriate CFL condition.

We define

$$\dot{A}^+ = Rdiag\left(sgn_+(\lambda^i)\mu^i\right) R^{-1} \qquad \dot{A}^- = Rdiag\left(sgn_-(\lambda^i)\mu^i\right) R^{-1}, \tag{44}$$

where

$$sgn_+(a) = \begin{cases} 1 & a \geq 0 \\ 0 & a < 0 \end{cases}, \quad sgn_-(a) = \begin{cases} 0 & a \geq 0 \\ -1 & a < 0 \end{cases}.$$

Note that $\dot{A}^+$ and $\dot{A}^-$ commute with A. Clearly

$$\begin{gathered} \mid \dot{A} \mid = \dot{A}^+ - \dot{A}^-, \\ \dot{A}^+ \geq \mathring{A}^+, \qquad \dot{A}^- \leq \mathring{A}^-. \end{gathered} \tag{45}$$

However $A \neq \dot{A}^+ + \dot{A}^-$ in general; this is why ϕ^1 is restricted to the minmod. Following the same analysis as before, the coefficient matrix C_1 of U_{j+1} is

$$C_1 = -\frac{\Delta t}{2\Delta x} \left((A_{1/2} - \mid \dot{A}_{1/2} \mid) - \dot{A}^-_{1/2} L^1_{1/2} + \dot{A}^-_{-1/2} \tilde{L}^1_{-1/2} \right). \tag{46}$$

Using (29,45), we can rewrite C_1 as

$$C_1 = -\frac{\Delta t}{2\Delta x}\left((\mathring{A}^+_{1/2} - \dot{A}^+_{1/2}) + \mathring{A}^-_{1/2} + \dot{A}^-_{1/2}(I - L^1_{1/2}) + \dot{A}^-_{-1/2}\tilde{L}^1_{-1/2}\right). \qquad (47)$$

We claim that all four terms in the parentheses on the right are negative. Since ϕ^1 is the minmod limiter function, $0 \le \phi^1(\theta), \frac{\phi^1(\theta)}{\theta} \le 1$; it follows that $0 \le \tilde{L}^1,\ L^1 \le I$. By construction $\mathring{A}^- \le 0, \dot{A}^- \le 0$, and $\mathring{A}^+_{1/2} - \dot{A}^+_{1/2} \le 0$. Since all matrices with the same subscript commute, this proves that $C_1 \ge 0$.

Similarly the coefficient matrix $C_{-1} \ge 0$.

The coefficient matrix C_0 is

$$\begin{aligned} C_0 = I - \tfrac{\Delta t}{2\Delta x}\Big(&-(A_{1/2} - \mid \dot{A}_{1/2} \mid) + (A_{-1/2} + \mid \dot{A}_{-1/2} \mid) \\ &+\dot{A}^+_{1/2}\tilde{L}^1_{1/2} + \dot{A}^-_{1/2}L^1_{1/2} - \dot{A}^+_{-1/2}L^1_{-1/2} - \dot{A}^-_{-1/2}\tilde{L}^1_{-1/2}\Big) \end{aligned} \qquad (48)$$

Using (29,45), we can rewrite C_0 as

$$\begin{aligned} C_0 &= I - \tfrac{\Delta t}{2\Delta x}\Big(\dot{A}^+_{1/2}(I + \tilde{L}^1_{1/2}) - \mathring{A}^-_{1/2} - \dot{A}^-_{1/2} + \mathring{A}^+_{-1/2} + \dot{A}^+_{-1/2} - \dot{A}^-_{-1/2}(I + \tilde{L}^1_{-1/2}) \\ &\quad -\mathring{A}^+_{1/2} + \dot{A}^-_{1/2}L^1_{1/2} + \mathring{A}^-_{-1/2} - \dot{A}^+_{-1/2}L^1_{-1/2}\Big) \\ &\ge I - \tfrac{\Delta t}{2\Delta x}[\dot{A}^+_{1/2}(I + \tilde{L}^1_{1/2}) - \mathring{A}^-_{1/2} - \dot{A}^-_{1/2}] - \tfrac{\Delta t}{2\Delta x}[\mathring{A}^+_{-1/2} + \dot{A}^+_{-1/2} - \dot{A}^-_{-1/2}(I + \tilde{L}^1_{-1/2})]. \end{aligned} \qquad (49)$$

As before the matrices in bracket are positive. Under the following CFL condition

$$\frac{\Delta t}{\Delta x}\max_{1\le i\le n, U}\mu^i \le \frac{1}{2}, \qquad (50)$$

each is $\le \frac{1}{2}I$, therefore the coefficient matrix $C_0 \ge 0$. As before the sum of three coefficient matrices is the identity. Therefore the more dissipative schemes (21,43) are positive.

We can construct now a family of schemes by combining the least dissipative positive scheme and a more dissipative one:

$$F^{\alpha,\beta}_{j+1/2} = \frac{F(U_j) + F(U_{j+1})}{2} - \frac{1}{2}\left(\alpha \mid \mathring{A}_{1/2} \mid (I - L^0_{1/2}) + \beta \mid \dot{A}_{1/2} \mid (I - L^1_{1/2})\right)(U_{j+1} - U_j) \qquad (51)$$

where α and β are constant satisfying $0 \le \alpha \le 1$ and $\alpha + \beta \ge 1$. The combined scheme (21,51) is a positive scheme under the following CFL condition

$$\frac{\Delta t}{\Delta x}\left(\alpha \max_{1\le i\le n, U} \mid \lambda^i \mid + \beta \max_{1\le i\le n, U}\mu^i\right) \le \frac{1}{2}. \qquad (52)$$

We show that as follows; from (21,51) we obtain

$$
\begin{aligned}
U_j^* &= U_j - \tfrac{\Delta t}{2\Delta x}[A_{1/2}(U_{j+1}-U_j) + A_{-1/2}(U_j - U_{j-1}) \\
&-\alpha\left(\mid \mathring{A}_{1/2} \mid (U_{j+1}-U_j) - \mathring{A}^{+}_{1/2}\tilde{L}^{0}_{1/2}(U_j - U_{j-1}) + \mathring{A}^{-}_{1/2}L^{0}_{1/2}(U_{j+1}-U_j)\right) \\
&-\beta\left(\mid \dot{A}_{1/2} \mid (U_{j+1}-U_j) - \dot{A}^{+}_{1/2}\tilde{L}^{1}_{1/2}(U_j - U_{j-1}) + \dot{A}^{-}_{1/2}L^{1}_{1/2}(U_{j+1}-U_j)\right) \\
&+\alpha\left(\mid \mathring{A}_{-1/2} \mid (U_j - U_{j-1}) - \mathring{A}^{+}_{-1/2}L^{0}_{-1/2}(U_j - U_{j-1}) + \mathring{A}^{-}_{-1/2}\tilde{L}^{0}_{-1/2}(U_{j+1}-U_j)\right) \\
&+\beta\left(\mid \dot{A}_{-1/2} \mid (U_j - U_{j-1}) - \dot{A}^{+}_{-1/2}L^{1}_{-1/2}(U_j - U_{j-1}) + \dot{A}^{-}_{-1/2}\tilde{L}^{1}_{-1/2}(U_{j+1}-U_j)\right)].
\end{aligned}
\tag{53}
$$

Hence we obtain with the help of $\mathring{A}^{+} \leq \dot{A}^{+}$ and $\alpha + \beta \geq 1$

$$
\begin{aligned}
C_1 = &\; -\tfrac{\Delta t}{2\Delta x}A_{1/2} - \tfrac{\Delta t}{2\Delta x}\alpha\left(- \mid \mathring{A}_{1/2} \mid - \mathring{A}^{-}_{1/2}L^{0}_{1/2} + \mathring{A}^{-}_{-1/2}\tilde{L}^{0}_{-1/2}\right) \\
&-\tfrac{\Delta t}{2\Delta x}\beta\left(- \mid \dot{A}_{1/2} \mid - \dot{A}^{-}_{1/2}L^{1}_{1/2} + \dot{A}^{-}_{-1/2}\tilde{L}^{1}_{-1/2}\right) \\
= &\; -\tfrac{\Delta t}{2\Delta x}\alpha\left(\mathring{A}^{-}_{1/2}(2I - L^{0}_{1/2}) + \mathring{A}^{-}_{-1/2}\tilde{L}^{0}_{-1/2}\right) \\
&-\tfrac{\Delta t}{2\Delta x}\left((1-\alpha)\mathring{A}^{-}_{1/2} + \left((1-\alpha)\mathring{A}^{+}_{1/2} - \beta\dot{A}^{+}_{1/2}\right) + \beta\dot{A}^{-}_{1/2}(I - L^{1}_{1/2}) + \beta\dot{A}^{-}_{-1/2}\tilde{L}^{1}_{-1/2}\right) \\
\geq &\; 0.
\end{aligned}
$$

Similarly the coefficient matrix $C_{-1} \geq 0$.

$$
\begin{aligned}
C_0 = &\; I - \tfrac{\Delta t}{2\Delta x}\left(-A_{1/2} + A_{-1/2}\right) \\
&-\tfrac{\Delta t}{2\Delta x}\alpha\left(\mid \mathring{A}_{1/2} \mid + \mid \mathring{A}_{-1/2} \mid + \mathring{A}^{+}_{1/2}\tilde{L}^{0}_{1/2} - \mathring{A}^{-}_{-1/2}\tilde{L}^{0}_{-1/2} + \mathring{A}^{-}_{1/2}L^{0}_{1/2} - \mathring{A}^{+}_{-1/2}L^{0}_{-1/2}\right) \\
&-\tfrac{\Delta t}{2\Delta x}\beta\left(\mid \dot{A}_{1/2} \mid + \mid \dot{A}_{-1/2} \mid + \dot{A}^{+}_{1/2}\tilde{L}^{1}_{1/2} - \dot{A}^{-}_{-1/2}\tilde{L}^{1}_{-1/2} + \dot{A}^{-}_{1/2}L^{1}_{1/2} - \dot{A}^{+}_{-1/2}L^{1}_{-1/2}\right) \\
\geq &\; I - \tfrac{\Delta t}{2\Delta x}\alpha\left(-A_{1/2} + A_{-1/2} + \mid \mathring{A}_{1/2} \mid + \mid \mathring{A}_{-1/2} \mid + \mathring{A}^{+}_{1/2}\tilde{L}^{0}_{1/2} - \mathring{A}^{-}_{-1/2}\tilde{L}^{0}_{-1/2}\right) \\
&-\tfrac{\Delta t}{2\Delta x}\left(-(1-\alpha)A_{1/2} + (1-\alpha)A_{-1/2} + \beta \mid \dot{A}_{1/2} \mid + \beta \mid \dot{A}_{-1/2} \mid + \beta\dot{A}^{+}_{1/2}\tilde{L}^{1}_{1/2} - \beta\dot{A}^{-}_{-1/2}\tilde{L}^{1}_{-1/2}\right) \\
\geq &\; I - \tfrac{\Delta t}{2\Delta x}\alpha[-2\mathring{A}^{-}_{1/2} + \mathring{A}^{+}_{1/2}\tilde{L}^{0}_{1/2}] - \tfrac{\Delta t}{2\Delta x}\alpha[2\mathring{A}^{+}_{-1/2} - \mathring{A}^{-}_{-1/2}\tilde{L}^{0}_{-1/2}] \\
&-\tfrac{\Delta t}{2\Delta x}[-(1-\alpha)\mathring{A}^{-}_{1/2} - \beta\dot{A}^{-}_{1/2} + \beta\dot{A}^{+}_{1/2} + \beta\dot{A}^{+}_{1/2}\tilde{L}^{1}_{1/2}] \\
&-\tfrac{\Delta t}{2\Delta x}[(1-\alpha)\mathring{A}^{+}_{-1/2} + \beta\dot{A}^{+}_{-1/2} - \beta\dot{A}^{-}_{-1/2} - \beta\dot{A}^{-}_{-1/2}\tilde{L}^{1}_{-1/2}].
\end{aligned}
$$

The matrices in the brackets are positive, therefore $C_0 \geq 0$ under the CFL condition (52). As before we can read off from (53) that $C_1 + C_0 + C_{-1} = I$. The proof is completed.

To achieve second order accuracy in time we use the following second order accurate energy preserving

Runge-Kutta method, due to Shu:

$$
\begin{aligned}
U_j^* &= U_j^m - \tfrac{\Delta t}{\Delta x}[F_{j+1/2}^{\alpha,\beta} - F_{j-1/2}^{\alpha,\beta}],\\
U_j^{**} &= U_j^* - \tfrac{\Delta t}{\Delta x}[F_{j+1/2}^{\alpha,\beta,*} - F_{j-1/2}^{\alpha,\beta,*}],\\
U_j^{m+1} &= \tfrac{1}{2}U_j^m + \tfrac{1}{2}U_j^{**}.
\end{aligned}
$$

Here F^* abbreviates the positive numerical flux evaluated at U^*. If we assume that the positive schemes (51) does not increase the l^2 norm, it follows that neither does Shu's Runge-Kutta method:

$$\|U^{**}\| \leq \|U^*\| \leq \|U^m\|,$$

and so

$$\|U^{m+1}\| \leq \frac{1}{2}\|U^m\| + \frac{1}{2}\|U^{**}\| \leq \|U^m\|.$$

Remark 1: An essential difference of our schemes from others is that we combine schemes using two different limiters. This makes a difference for systems. One of our scheme is the least dissipative positive scheme; the other contains more dissipation. We remark that we can construct positive schemes with as much dissipation as needed.

Remark 2: The 2nd order accurate numerical flux $F_{j+1/2}^{\alpha,\beta}$ (51) can be expressed in a simple compact form as

$$F_{j+1/2}^{\alpha,\beta} = \frac{F(U_j) + F(U_{j+1})}{2} - \frac{1}{2}R\left(\alpha \mid \Lambda \mid (I - \Phi^0) + \beta diag(\mu^i)(I - \Phi^1)\right) R^{-1}(U_{j+1} - U_j). \tag{54}$$

The leading cost of computing one flux in the above compact form is $3n^2$ scalar multiplications, where n is the order of the system.

Remark 3: For one-dimensional systems, we can mix Lax-Wendroff with upwind and obtain the second family of positive schemes as follows

$$F_{j+1/2}^{\alpha,\beta} = \frac{F(U_j) + F(U_{j+1})}{2} - \frac{1}{2}\left(\alpha \mid \mathring{A} \mid (I - L^0) + \beta \mid \dot{A} \mid (I - L^1) + \frac{\Delta t}{\Delta x}A^2\right)(U_{j+1} - U_j). \tag{55}$$

The numerical flux (55) is more dissipative than (54) by the amount of $\frac{\Delta t}{\Delta x} A^2 (U_{j+1} - U_j)$. *The family of schemes (21,55) are positive under the following CFL condition*

$$\frac{\Delta t}{\Delta x}\left(\alpha \max_{1\le i\le n, U} \mid \lambda^i \mid + \beta \max_{1\le i\le n, U} \mu^i + \frac{1}{2}\frac{\Delta t}{\Delta x} \max_{1\le i\le n, U} \mid \lambda^i \mid^2 \right) \le \frac{1}{2}. \tag{56}$$

It follows that the condition below is sufficient for positivity:

$$\frac{\Delta t}{\Delta x}\left(\alpha \max_{1\le i\le n, U} \mid \lambda^i \mid + \beta \max_{1\le i\le n, U} \mu^i \right) \le \sqrt{2} - 1. \tag{57}$$

Note that the family of positive schemes (21,55) are second order accurate both in space and time. No Runge-Kutta is needed, which saves $1 - \frac{1}{4(\sqrt{2}-1)} \approx 40\%$ *of computations.*

Remark 4: For one-dimensional systems, we can also mix Lax-Wendroff with upwind in another way and obtain a positive scheme as follows

$$\begin{aligned} F^{\alpha,\beta}_{j+1/2} = \quad & \tfrac{F(U_j)+F(U_{j+1})}{2} \\ & -\tfrac{1}{2}\left(\alpha \mid \mathring{A} \mid (I - L^0) + \beta \mid \dot{A} \mid (I - L^1) + \tfrac{\Delta t}{\Delta x} A^2 (\alpha L^0 + (1-\alpha) L^1)\right)(U_{j+1} - U_j). \end{aligned} \tag{58}$$

This scheme is positive under the following CFL condition

$$\frac{\Delta t}{\Delta x}\left(\alpha \max_{1\le i\le n, U} \mid \lambda^i \mid + \beta \max_{1\le i\le n, U} \mu^i + \frac{\alpha+1}{2}\frac{\Delta t}{\Delta x} \max_{1\le i\le n, U} \mid \lambda^i \mid^2 \right) \le \frac{1}{2}. \tag{59}$$

The proof is similar and is omitted here.

2.3 Positive Schemes for Multi-dimensional Systems of Conservation Laws

We consider multi-dimensional hyperbolic systems of conservation laws (17). In each dimension x_s, $s = 1, \cdots, d$, we construct a family of fluxes exactly as in one-dimension (51), for $0 \le \alpha_s \le 1$ and $\alpha_s + \beta_s \ge 1$,

$$F^{\alpha_s,\beta_s}_{J+1/2e_s} = \frac{F_s(U_{J+e_s}) + F_s(U_J)}{2} - \frac{1}{2}\alpha_s \mid \mathring{A} \mid (I - L^0)(U_{J+e_s} - U_J) - \frac{1}{2}\beta_s \mid \dot{A} \mid (I - L^1)(U_{J+e_s} - U_J). \tag{60}$$

Here α_s, β_s could be different for each dimension.

The family of schemes (18, 60) are positive under the following CFL condition

$$\sum_{s=1}^{d} \frac{\Delta t}{\Delta x_s} \left(\alpha_s \max_{1 \le i \le n, U} \mid \lambda^{s,i} \mid + \beta_s \max_{1 \le i \le n, U} \mu^{s,i} \right) \le \frac{1}{2}. \tag{61}$$

We use the same 2nd order accurate energy preserving Runge-Kutta of Shu to achieve 2nd order accuracy in time: for $m = 0, 1, \cdots$,

$$\begin{aligned} U_J^* &= U_J^m - \sum_{s=1}^{d} \tfrac{\Delta t}{\Delta x_s} [F_{J+1/2e_s}^{\alpha_s,\beta_s} - F_{J-1/2e_s}^{\alpha_s,\beta_s}], \\ U_J^{**} &= U_J^* - \sum_{s=1}^{d} \tfrac{\Delta t}{\Delta x_s} [F_{J+1/2e_s}^{\alpha_s,\beta_s,*} - F_{J-1/2e_s}^{\alpha_s,\beta_s,*}], \\ U_J^{m+1} &= \tfrac{1}{2} U_J^m + \tfrac{1}{2} U_J^{**}. \end{aligned}$$

2.4 Symmetrizable Systems

A system of conservation laws (17) is symmetrizable if the Jacobian of the flux vectors

$$A_s = \bigtriangledown F_s,$$

can be symmetrized by the same similarity transformation:

$$S A_s S^{-1}.$$

The matrix S depends on U. It is well known that the equations of compressible flow, and of MHD, are symmetrizable. As explained as in the introduction, our schemes (51) (55) and their multi-dimensional analogue (60) can also be made symmetric and positive by the same similarity transformation, modulo terms that can be expressed as spatial differences of U. Therefore the positivity philosophy applies to symmetrizable systems as well.

3 Numerical Experiments

In the applications described in this section no special problem-dependent techniques were used and no post-processing were used.

In all of our numerical experiments we choose $| \dot{A} |= \max_{1 \leq i \leq n, U} | \lambda^i | I$ for the more dissipative scheme. The least dissipative positive scheme $\alpha = 1, \beta = 0$ gives us sharp shocks but admits entropy violating solutions, which are not physically relevant. The more dissipative scheme $\alpha = 0, \beta = 1$ gives us smeary shocks but seems to converge to physically relevant solutions. The compromise $\alpha = 0.9, \beta = 0.1$ seems to combine the best features of both; these are the schemes used in experiments described here unless otherwise stated.

3.1 One-dimensional Euler Equations of Gas Dynamics

In this subsection we apply our schemes (21,51) and (21,55) to the Euler equation of gas dynamics for a polytropic gas.

$$U_t + F(U)_x = 0$$
$$U = (\rho, m, E)^T$$
$$F(U) = (m, \rho q^2 + P, q(E + P))^T$$
$$P = (\gamma - 1)(E - \tfrac{1}{2}\rho q^2)$$
$$m = \rho q$$

In the following computation $\gamma = 1.4$. For details of the Jacobian $F^{'}(U)$, its eigenvalues, eigenvectors, etc., see [7].

We used the scheme (21,51) in example 1 and the scheme (21,55) in examples 2 and 3.

Example 1. We consider the following Riemann problems:

$$U_0(x) = \begin{cases} U_l & x < 0 \\ U_r & x > 0. \end{cases}$$

Two sets of initial date are used. One is proposed by Sod in [19]:

$$(\rho_l, q_l, P_l) = (1, 0, 1); \quad (\rho_r, q_r, P_r) = (0.125, 0, 0.1).$$

The other is used by Lax [12]:

$$(\rho_l, q_l, P_l) = (0.445, 0.698, 3.528); \quad (\rho_r, q_r, P_r) = (0.5, 0, 0.571).$$

The numerical results, showed in Figure 1a, are obtained by using the positive scheme (21,51) based on centered differencing with 100 points as in [7] and [18]. We used the superbee limiter function

$$\phi^0(\theta) = \begin{cases} 0, & \theta \le 0 \\ 2\theta & 0 < \theta \le 1/2 \\ 1 & 1/2 < \theta \le 1 \\ \theta & 1 < \theta \le 2 \\ 2 & \theta > 2 \end{cases}$$

for ϕ^0. We observe that the positive scheme does much better than the 4-th order ENO scheme at corners of rarefactions (discontinuities in derivatives) and better at contact discontinuity; but is smearer at shocks. Hence our numerical results are comparable to the 4-th order ENO scheme [7].

In the numerical results shown in Figure 1b we used Van Leer's limiter function

$$\phi^0(\theta) = \frac{\theta + \mid \theta \mid}{1 + \mid \theta \mid}.$$

The positive scheme does better than the flux based 3-rd order ENO scheme [18] at corners of rarefactions (discontinuities in derivatives) and comparable at contact discontinuity and shocks in Figure 1b. Hence our numerical results are comparable the flux based 3-rd order ENO scheme without sharpening [18]. With sharpening, ENO resolves the contact discontinuity much better. At extrema in smooth regions of the solutions, ENO schemes have high order accuracy; our positive schemes have only first order accuracy.

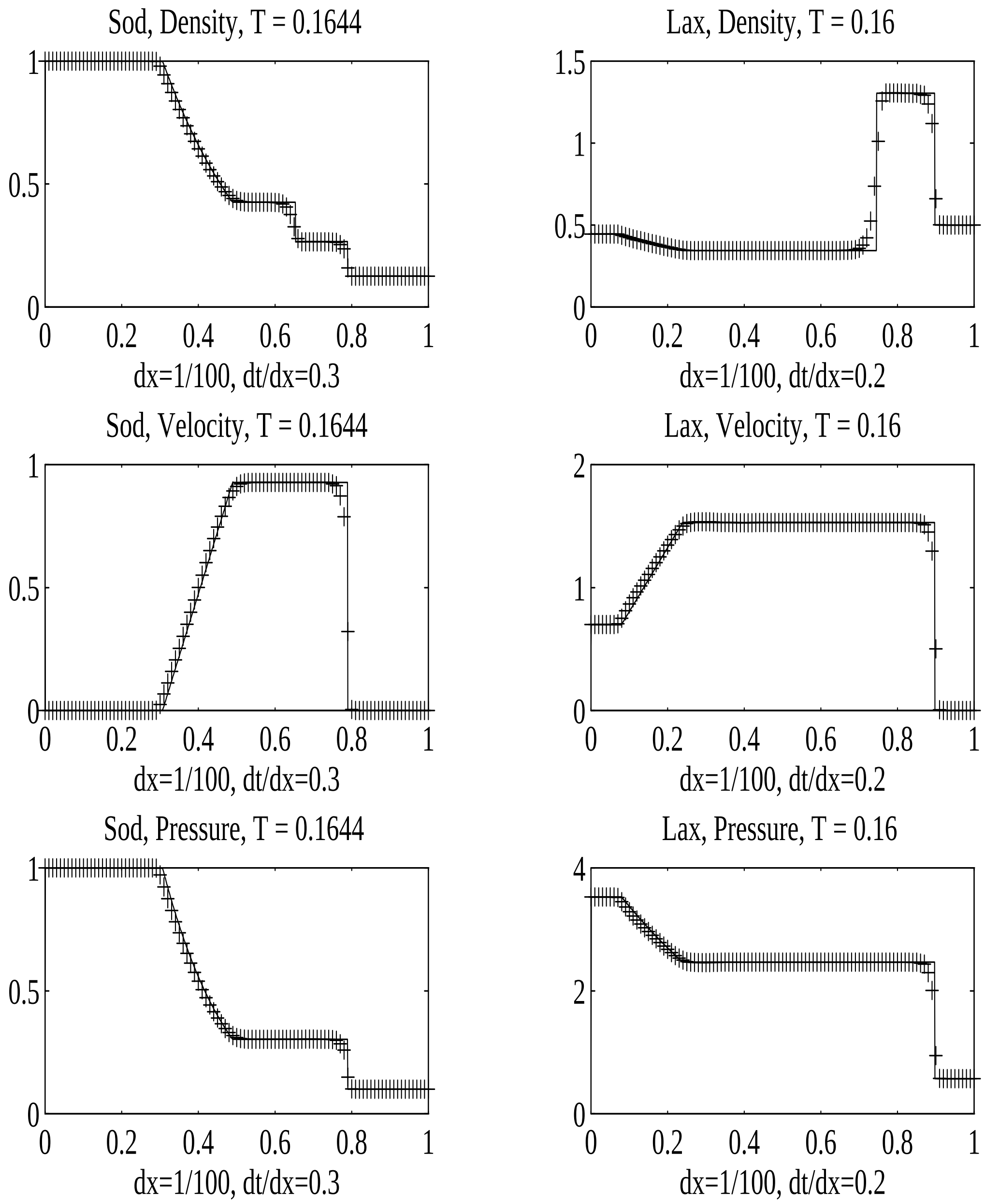
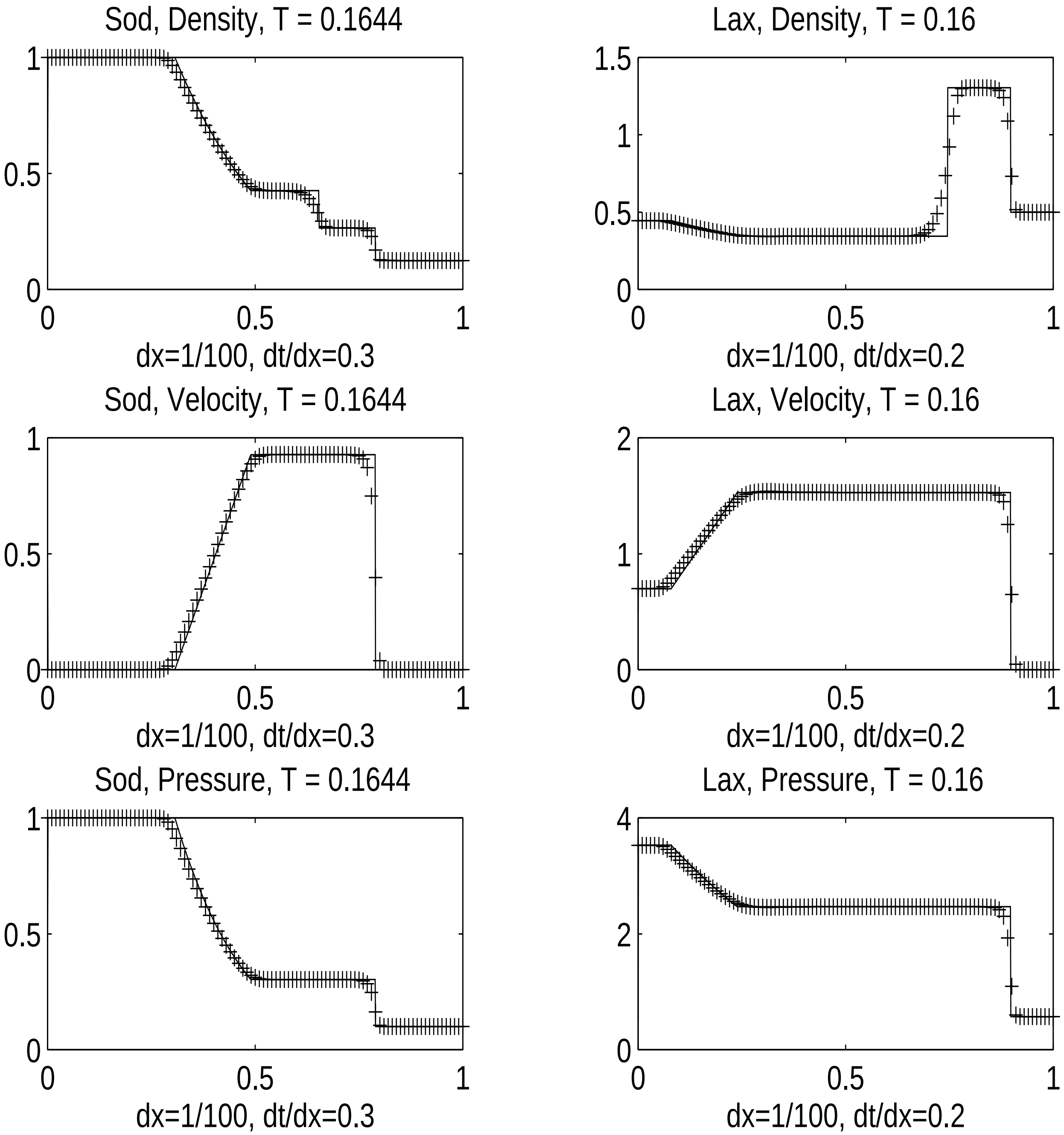

Sod, Density, T = 0.1644
1
0.5
0
0
0.2
0.4
0.6
0.8
1
dx=1/100, dt/dx=0.3
Lax, Density, T = 0.16
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1
dx=1/100, dt/dx=0.2
Sod, Velocity, T = 0.1644
1
0.5
0
0
0.2
0.4
0.6
0.8
1
dx=1/100, dt/dx=0.3
Lax, Velocity, T = 0.16
2
1
0
0
0.2
0.4
0.6
0.8
1
dx=1/100, dt/dx=0.2
Sod, Pressure, T = 0.1644
1
0.5
0
0
0.2
0.4
0.6
0.8
1
dx=1/100, dt/dx=0.3
Lax, Pressure, T = 0.16
4
2
0
0
0.2
0.4
0.6
0.8
1
dx=1/100, dt/dx=0.2

Figure 1a



Sod, Density, T = 0.1644

Lax, Density, T = 0.16

Sod, Velocity, T = 0.1644

Lax, Velocity, T = 0.16

Sod, Pressure, T = 0.1644

Lax, Pressure, T = 0.16

Figure 1b

Example 2. Interacting Blast Wave [22] with initial data,

$$U_0(x) = \begin{cases} U_l & 0 \le x < 0.1, \\ U_m & 0.1 \le x < 0.9, \\ U_r & 0.9 \le x < 1. \end{cases}$$

$$(\rho_l, m_l, E_l) = (1, 0, 1000); \quad (\rho_m, m_m, E_m) = (1, 0, 0.01); \quad (\rho_r, m_r, E_r) = (1, 0, 100).$$

Reflecting boundary conditions are applied at both ends. The numerical results were obtained by scheme (21,55) based on Lax-Wendroff. In the calculations displayed in Figure 2a, we used Van Leer's limiter for ϕ^0. The circled lines are numerical solution with 200 points, and the solid lines are numerical solution with 1200 points, same as the resolutions in [22]. The positive scheme captures important structures on the fine grid, and the results on both grids are comparable to MUSCL scheme in [22]. The cost of the positive scheme is about 60% cost of MUSCL scheme. In Figure 2b, we show the numerical result of the positive scheme with the superbee limiter for ϕ^0; the solid lines are the same as in Figure 2a. The resolution is improved dramatically; this shows that the numerical resolution of this problem depends very much on the limiter.

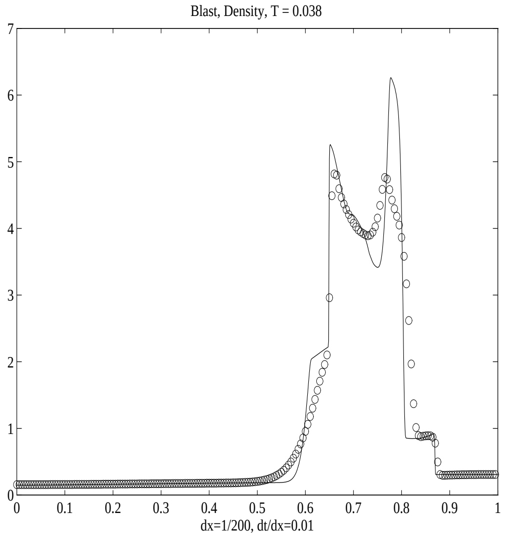


Figure 2a

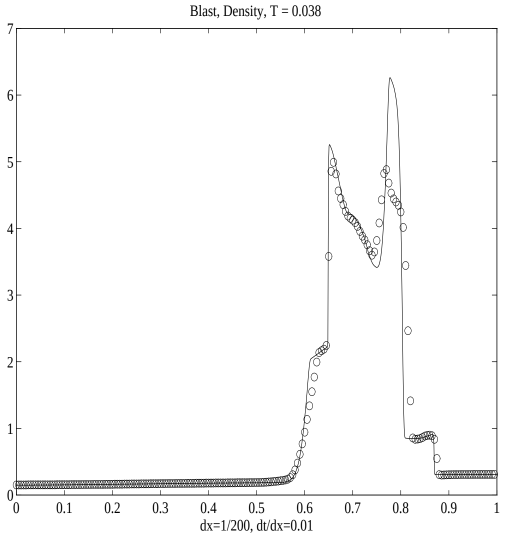


Figure 2b

Example 3. Low density and internal energy Riemann problem [2] with initial data,

$$U_0(x) = \begin{cases} U_l & 0 \leq x < 0.5, \\ U_r & 0.5 \leq x \leq 1. \end{cases}$$

$$(\rho_l, m_l, E_l) = (1, -2, 3); \qquad (\rho_r, m_r, E_r) = (1, 2, 3).$$

It has been observed that schemes which are based on a linearized Riemann solver (e.g., Roe's scheme) may lead to an unphysical negative density or internal energy during the computational process [2]. The positive scheme (21, 55) with $\alpha = 0.9, \beta = 0.1$ does suffer from negative internal energy. However after we add more dissipation by choosing $\alpha = 1, \beta = 4$, this positive scheme does succeed. In Figure 3, the dotted lines are numerical solution with 200 points, and the solid lines are numerical solution with 1600 points.

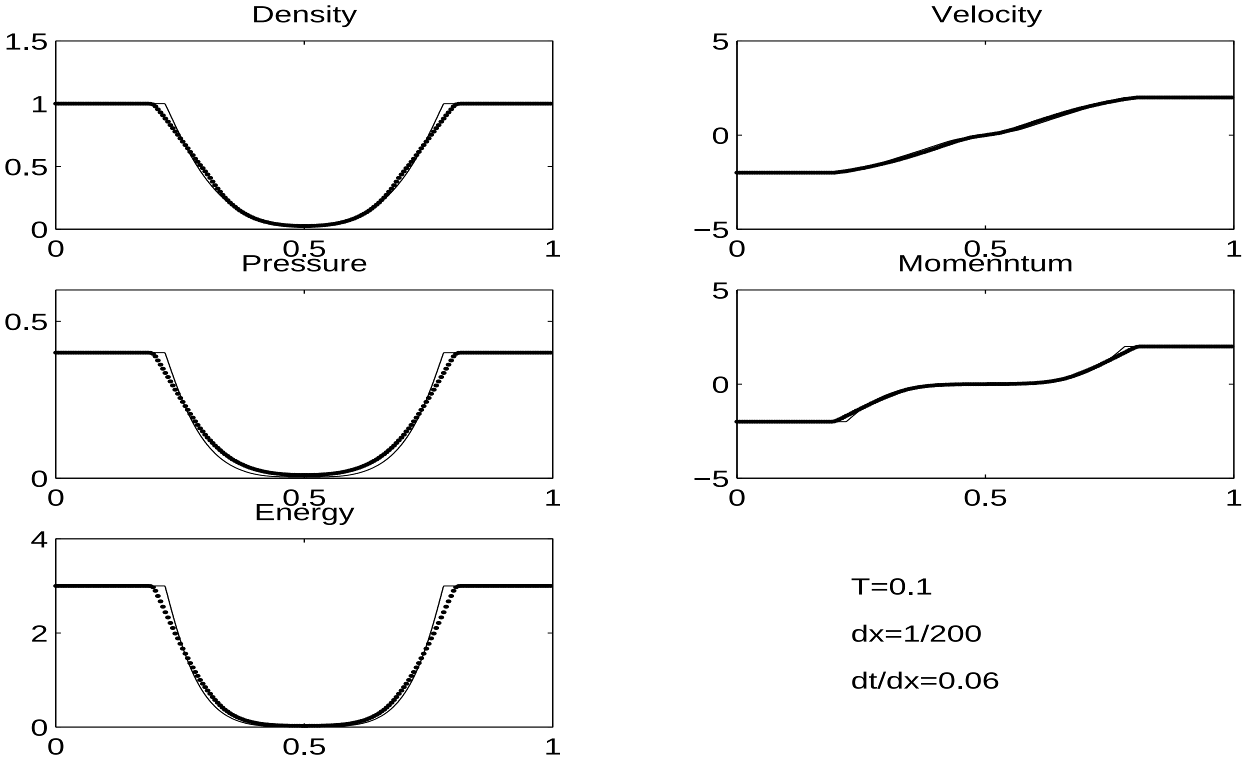


Figure 3

3.2 Multi-dimensional Euler equations of Gas Dynamics

We approximate solutions of the two-dimensional Euler equations of Gas Dynamics,

$$U_t + F_1(U)_x + F_2(U)_y = 0$$

$$U = (\rho, m, n, E)^T$$

$$F_1(U) = (m, \rho u^2 + P, \rho uv, u(E+P))^T$$

$$F_2(U) = (n, \rho uv, \rho v^2 + P, v(E+P))^T$$

$$P = (\gamma - 1)(E - \tfrac{1}{2}\rho(u^2 + v^2))$$

$$m = \rho u \qquad n = \rho v.$$

For details of the Jacobian $F_1^{'}(U)$ and $F_2^{'}(U)$, their eigenvalues, eigenvectors, etc., see [18].

Example 4: Double Mach Reflection. A planar shock is incident on an oblique wedge at a 60^o angle. The test problem involves a Mach 10 shock in air, $\gamma = 1.4$. The undisturbed air ahead of the shock has a density of 1.4 and a pressure of 1. We use the boundary conditions described in [22]. Van Leer's limiter function is used for ϕ^0. The flow at time $t = 0.2$, computed by our positive scheme, is plotted in Figure 4 with $\Delta x = \Delta y = 1/120$, $\Delta t = \frac{5}{3} \times 10^{-4}$. In each plot 30 equally spaced contours are shown. There is no visible oscillation behind the strong shock. There are three difficulties in computing this flow mentioned in [22]. The first difficulty is the rather weak second Mach shock; it dies out entirely by the time it reaches the contact discontinuity from the first Mach reflection. Figure 4 shows that the second Mach shock is perfectly captured. The second difficulty is the jet, formed when the flow of the denser fluid is deflected by a pressure gradient built up in the region where the first contact discontinuity approaches the reflecting wall. Figure 4 shows that the jet is extremely well captured. The third difficulty is caused by the region bounded by the second Mach shock, the curved reflected shock, and the reflecting wall. The double Mach reflection contains both steady and unsteady structures. The curved reflected shock is moving rapidly at its right end and is not moving at all at its left end; this causes oscillations for many difference schemes, including PPM, MUSCL and Jin and Xin's ingenious relaxing scheme [11]. For the positive scheme there is no oscillation at all; thus the positive scheme overcomes extremely well all three numerical difficulties.

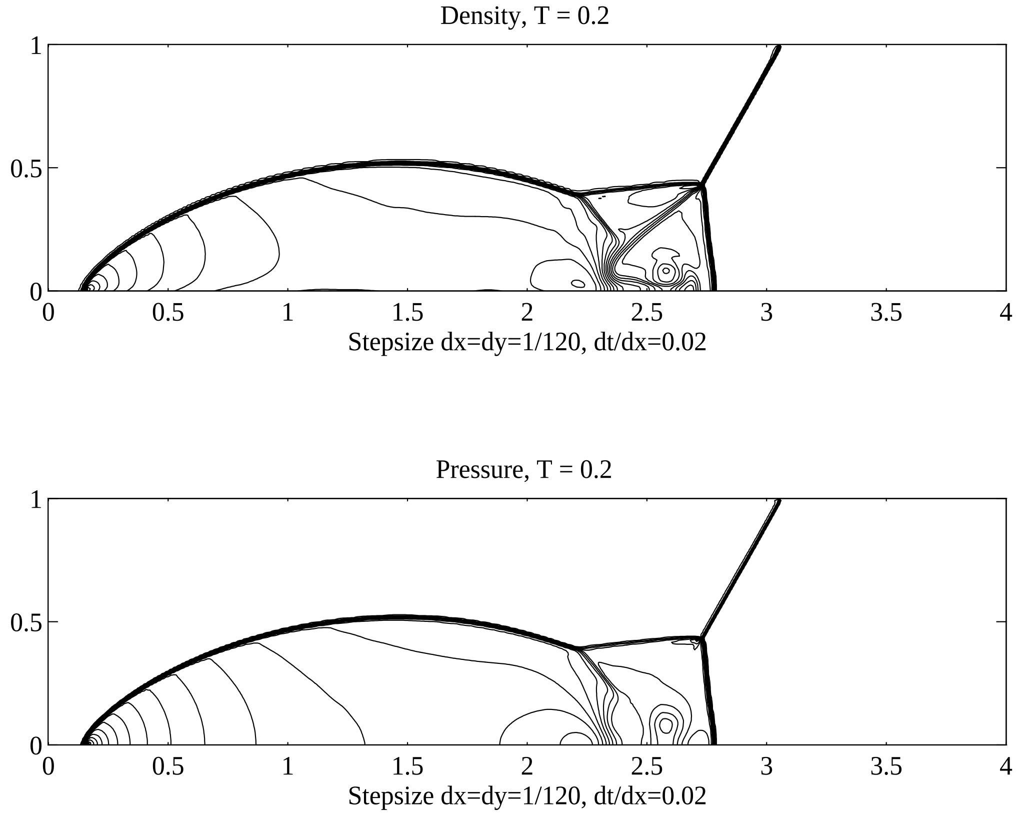


Figure 4

Example 5: A Mach 3 *Wind Tunnel with a Step.* This problem has been a useful test for schemes for many years. The tunnel is 3 length units long and 1 length unit wide, with a step which is 0.2 length units high and 0.6 length units away from the left end of the tunnel. The state behind the incoming shock is density 1.4, pressure 1.0, and velocity 3 from left to right. These are used as boundary condition at the left; at the right all horizontal gradients are assumed to vanish. Along the walls of the tunnel and the obstacle reflecting conditions are applied in the perpendicular direction. The corner of the step is the center of a rarefaction fan and hence is a singular point of the flow. At the corner we reflect the solution using a "wall" which is 45^o both to the top and the side of the step. No problem dependent boundary conditions are used as in PPM. Van Leer's limiter function is used for ϕ^0. The density and pressure contours in the tunnel at time 4 are displayed in Figure 5 with $\Delta x = \Delta y = 1/80$, $dt = \frac{3}{16} \times 10^{-2}$. The flow at time 4 is still unsteady.

The general position and shape of the shocks are accurate. The shocks are well captured. There is no numerical noise. The contact discontinuities are resolved and are only slightly broader than ones of PPM and MUSCL. The weak oblique shock is resolved.

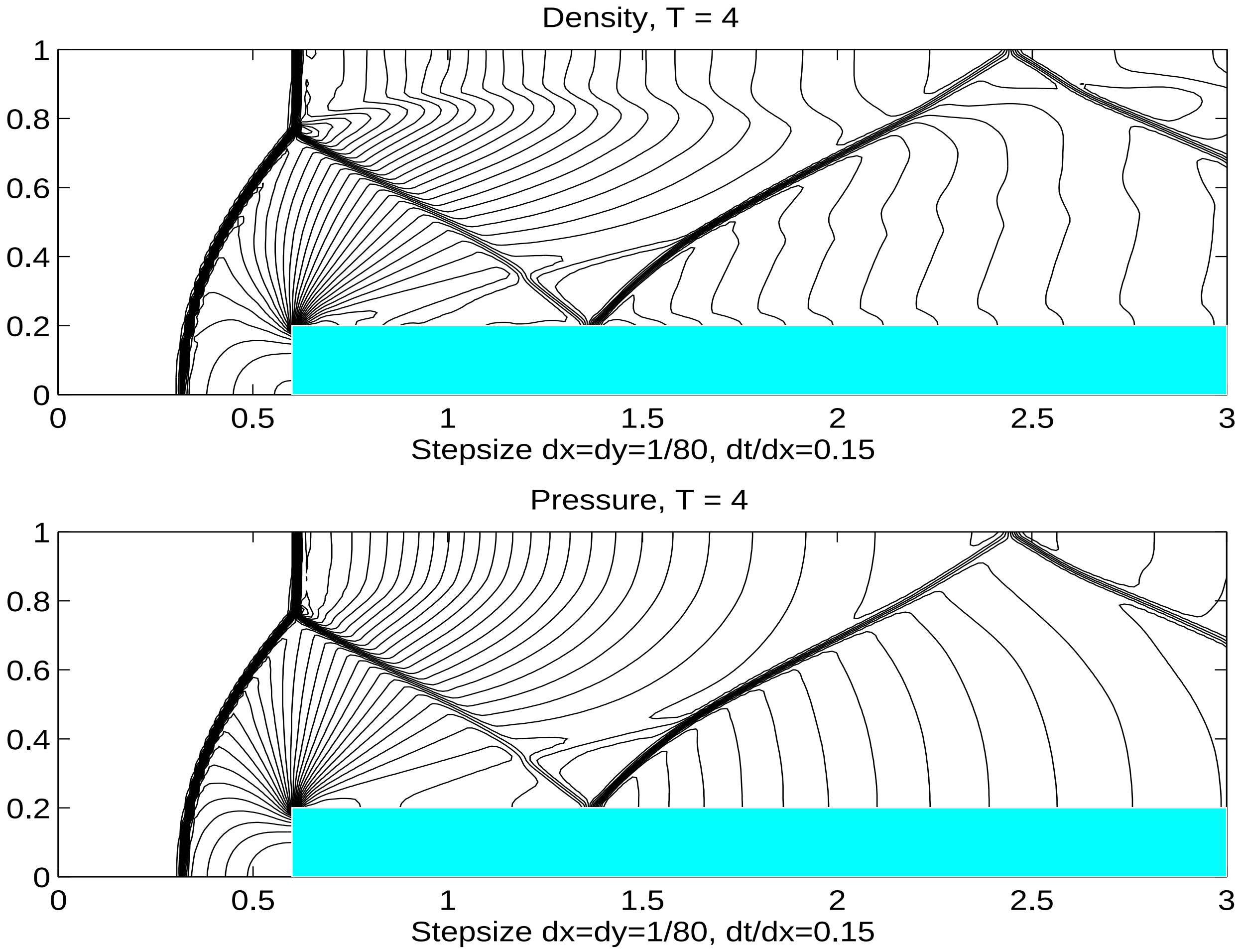


Figure 5

References

[1] J. P. Boris and D. L. Book, "Flux Corrected Transport I, SHASTA, A Fluid Transport Algorithm that Works," *J. Comput. Phys.,* 11(1973), pp. 38-69.

[2] B. Einfeldt, C.D. Munz, P.L. Roe, and B. Sjogreen, "On Godunov-Type Methods near Low Densities," *Journal of Computational Physics*, Vol. 92, 1991, pp. 273-295.

[3] K. O. Friedrichs, "Symmetric Hyperbolic Linear Differential Equations," *Commun. Pure Appl. Math.* 7, (1954), pp. 345-392.

[4] A. Harten, "On a Class of High Resolution Total-Variation-Stable Finite-Difference Schemes," *SIAM J. Numer. Anal.*, Vol. 21, No. 1, 1984, pp. 1-23.

[5] A. Harten, "High Resolution Schemes for Hyperbolic Conservation Laws," *J. Comput. Phys.*, Vol 49, 1983, pp.357-393.

[6] A. Harten, "On the Symmetric Form of Systems of Conservation Lwas with Entropy," *Journal of Computational Physics*, V49, pp. 151-164, 1983.

[7] A. Harten, B. Engquist, S. Osher and S. Chakravarthy, "Uniformly High Order Accurate Essentially Non-Oscillatory Schemes III," *Journal of Computational Physics*, V71, pp. 231-303, 1987; also ICASE Report No. 86-22, April 1986.

[8] A. Harten and G. Zwas, "Self-Adjusting Hybrid Schemes for Shock Computations," *Journal of Computational Physics*, V9, pp.568-583, 1972.

[9] A. Jameson, "Artificial Diffusion, Upwind Biasing, Limiters and Their Effect on Accuracy and Multigrid Convergence in Transonic and Hypersonic Flows," *AIAA-93-3359.*

[10] G.-S. Jiang and C.W. Shu, "Efficient Implementation of Weighted ENO Schemes," in preparing.

[11] S. Jin and Z. Xin, "The Relaxing Schemes for Systems of Conservation Laws in Arbitrary Space Dimensions," to appear in CPAM.

[12] P. Lax. “Weak Solutions of Nonlinear Hyperbolic Equations and Their Numerical Computation,” *Commun. Pure Appl. Math.* 7, (1954), pp. 159-193.

[13] R. LeVeque, “Numerical Methods for Conservation Laws,” Lectures in Mathematics, Birkhäuser Verlag.

[14] X.-D. Liu, S. Osher and T. Chan, “Weighted Essentially Non-oscillatory Schemes,” *J. Comput. Phys.* 115, 1994, pp. 200-212.

[15] C. Morawetz, “Potential Theory for Regular and Mach Reflection of a Shock at a Wedge,” *Commun. Pure Appl. Math.* Vol XLVII, (1994), pp. 593-624.

[16] P.L. Roe, “Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes,” *J. Comput. Phys.* 43, 1981, pp. 357-372.

[17] C.-W. Shu, “Total-Variation-Diminishing Time Discretizations,” *SIAM J. Stat. Comput.* Vol. 9, No. 6, November 1988, pp. 1073-1084.

[18] C.-W. Shu, Stanley Osher, “Efficient Implementation of Essentially Non-oscillatory Shock-Capturing Schemes, II,” *J. Comput. Phys.*, V83 , 1989, pp. 32-78.

[19] G. Sod, “A Survey of Several Finite Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws,” *J. Comput. Phys.* 27, 1 (1978), pp. 1-31.

[20] P.K. Sweby, “High Resolution Schemes Using Flux Limiters for Hyperbolic Conservation Laws” *SIAM J. Numer. Anal.*, Vol 21, No. 5, 1984, pp.995-1011.

[21] E. Tadmor, “Skew-Selfadjoint Form for Systems of Conservation Laws”, *Journal of Mathematical Analysis and Applications* Vol. 103, No. 2, October 30, 1984, pp.428-442.

[22] P. Woodward and P. Colella, “The Numerical Simulation of Two-Dimensional Fluid Flow with Strong Shocks,” *J. Comput. Phys.*, V54, pp.115-173.

[23] H. C. Yee, “Construction of Explicit and Implicit Symmetric TVD Schemes and Their Applications,” *J. Comput. Phys.*, V68, pp.151-179, 1987.