JOURNAL OF COMPUTATIONAL PHYSICS 109, 115–132 (1993)

Accuracy Optimized Methods for Constrained Numerical Solutions of Hyperbolic Conservation Laws

C. CORAY AND J. KOEBBE

Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322

Received January 2, 1992; revised March 29, 1993

A general method for construction of numerical schemes for scalar conservation laws which optimizes accuracy is applied to linear advection problems and Burgers' equation. The schemes, termed accuracy optimized methods (AOMs), define and solve a quadratic programming problem at each discrete time level to minimize perturbations from higher order accurate methods subject to imposed constraints. The constraints are used to impose desired behavior on the numerical approximation of the solution of the conservation law. The resulting schemes compare favorably with other high resolution schemes for scalar conservation laws. Numerical examples are presented for linear advection of discontinuities and development and transport of shocks in Burgers' equation.

1. INTRODUCTION

The problems of numerical diffusion, resolution of shock fronts, and spurious oscillations which arise in approximating the solution of scalar hyperbolic conservation laws of the form

$$\begin{aligned} &u_t + f(u)_x = 0, \qquad t > 0, \qquad x \in \Re, \\ &u(x, 0) = u_0(x), \end{aligned} \tag{1}$$

have long been a concern of physicists, engineers, and mathematicians. It is not possible to maintain, for example, monotonicity of a numerical solution near a shock front with an unmodified higher order accurate method. One needs only to examine the second-order method of Lax and Wendroff [8] to find a scheme which may introduce spurious oscillations for a variety of initial conditions even in the linear advection problem. Many authors have contributed significantly to this general topic, including work by Boris and Book [1], Harten [5], Osher [10], Osher and Chakravarthy [11], Van Leer [14], Zalesak [16], Sweby [13], and others. For general reference, a recent comprehensive overview of the subject of numerical methods for conservation laws has been published by LeVeque [9].

The central idea in this paper is to construct a general method which "optimizes" the accuracy of the approximation subject to constraints, such as monotonicity preserving, total variation diminishing (TVD), or entropy constraints. The basic approach is to consider a higher order accurate scheme and perturb it slightly when necessary to meet imposed constraints. The perturbation is based on minimal modification of a higher order method which then defines an optimization problem. The solution of the optimization adds to the computational effort necessary to numerically approximate the solution of Eq. (1). The extra computational effort can be reduced significantly since the optimization problem can be localized to regions where the solution is not smooth. This allows the use of second (or higher) order methods over regions where numerical solution profiles permit unconstrained application of such methods. The "accuracy optimized method," hereafter referred to as AOM, requires that a constrained optimization problem be solved at each time step in those narrow regions where an unmodified method would violate imposed constraints (e.g., monotonicity preserving or TVD). In this paper we address in detail the general linear advection problem and the specific nonlinear Burgers' equation, whereas in a second paper to follow, the general nonlinear case will be examined.

A somewhat surprising and positive result is that, although the approach is entirely motivated by a minimal perturbation of a high order method required to satisfy imposed constraints, the resolution of shock fronts with this method compares favorably with the high resolution methods presented in Sweby [13] which result from addition of "antidiffusive" flux.

Imposed constraints on approximate solutions of scalar conservation laws (e.g., monotonicity preserving) are often required for mathematical or physical reasons. In the AOM approach such constraints lead to a well-posed quadratic programming problem. The resulting sum of squares objective function in this problem guarantees the existence and uniqueness of the solution to the optimization problem. To illustrate the general methodology we initially focus in this

0021-9991/93 $5.00

paper on monotonicity preserving schemes, but we emphasize that the same methodology can be applied to a variety of other imposed constraining conditions, e.g., total variation diminishing (TVD) or an entropy condition.

The purpose of this paper is to illustrate the AOM framework; thus we develop the methodology by considering the scalar case in both the linear and nonlinear problems, but we note that extension is easily made to strictly hyperbolic linear systems (see Lax [7]). Also, in this paper the AOM is illustrated by developing schemes that are perturbations of second-order methods. The AOM can be extended to schemes which are perturbations of methods which are higher than second-order accurate. These extensions will be considered in another paper.

The methods presented here are not constructed to be computationally efficient. The goal of this paper is only to present initial results and introduce the alternate AOM viewpoint for construction of numerical methods for scalar conservation laws. The issues associated with efficiency of the schemes will be addressed in future work.

In Section 2 we present the motivation for the AOM approach, identifying in a linear advection example case the nature of the perturbation we propose, and we introduce two asymmetric AOMs. In Section 3 we introduce a symmetric version of the AOM which addresses in the linear advection case some of the difficulties arising in the asymmetric methods discussed in Section 2. An AOM modified "combination" of the second-order Lax–Wendroff [8] scheme and the double upwinding scheme of Warming and Beam [15] is presented. Section 4 addresses the localization of the optimization problem. In Section 5 we discuss the general methodology as it applies to nonlinear problems and illustrate the AOM on Burgers' equation. In Section 6 we show that a numerical scheme which satisfies the monotonicity constraint introduced in Section 2 is a TVD scheme and hence AOM schemes developed using this constraint are TVD. Section 7 contains numerical results for both the linear and nonlinear advection cases and computational comparisons with other methods.

2. ACCURACY OPTIMIZED METHODS

Assuming that f is linear in Eq. (1), i.e., $f(u)=au$, $a>0$, we consider the conservation form discretization

$$u_j^{n+1}=u_j^n-(a\lambda)(u_{j+1/2}^n-u_{j-1/2}^n), \tag{2}$$

on a mesh defined by nodes x_j and grid lines $x_{j+1/2}$. Define t_n and t_{n+1} as discrete time levels, Δx and Δt as the grid size in space and time, $\lambda=\Delta t/\Delta x$, and $u_j^n=u(x_j, t_n)$. The values of $u_{j\pm 1/2}^n$ are defined by

$$u_{j\pm 1/2}^n=m_{j\pm 1/2}^n(\alpha_{j\pm 1/2}^n), \tag{3}$$

where $m_{j\pm 1/2}^n$ is a function chosen to represent the unknown u between neighboring nodes. The point of evaluation, $\alpha_{j\pm 1/2}^n$, is chosen as the point which is a characteristic distance upwind from the grid line $x_{j\pm 1/2}$. Thus the grid line values $u_{j\pm 1/2}^n$ are in fact approximations of the unknown at the grid lines $x_{j\pm 1/2}$ for the next time level, t_{n+1}. In Koebbe [6] the general background for this method of describing discretizations is discussed for the case where $m_{j\pm 1/2}^n$ are chosen to be polynomials of arbitrary degree with coefficients depending on neighboring nodal values.

The accuracy of the scheme can be determined by fixing the evaluation points $\alpha_{j\pm 1/2}^n$ and coefficients of the polynomials $m_{j\pm 1/2}^n$ to satisfy accuracy conditions based on standard truncation error analysis (see [6]). The simplest example in this framework would be to choose the constant functions $m_{j+1/2}^n=u_j^n$ and $m_{j-1/2}^n=u_{j-1}^n$, which would yield the conventional first-order upwinding scheme with $a>0$.

Another natural choice for $m_{j\pm 1/2}^n$ which will be used to illustrate the AOM throughout this paper are the linear interpolating polynomials

$$m_{j+1/2}^n(\alpha_{j+1/2}^n)=u_{j+1}^n+(u_j^n-u_{j+1}^n)\,\alpha_{j+1/2}^n \tag{4}$$

and

$$m_{j-1/2}^n(\alpha_{j-1/2}^n)=u_j^n+(u_{j-1}^n-u_j^n)\,\alpha_{j-1/2}^n. \tag{5}$$

If we set $\alpha_{j\pm 1/2}^n=\frac{1}{2}(1+a\lambda)$, the resulting discretization is exactly the second-order Lax–Wendroff [8] scheme for the linear advection case. Choosing, instead,

$$m_{j+1/2}^n(\alpha_{j+1/2}^n)=u_j^n-(u_{j-1}^n-u_j^n)\,\alpha_{j+1/2}^n \tag{6}$$

and

$$m_{j-1/2}^n(\alpha_{j-1/2}^n)=u_{j-1}^n-(u_{j-2}^n-u_{j-1}^n)\,\alpha_{j-1/2}^n. \tag{7}$$

with $\alpha_{j\pm 1/2}^n=\frac{1}{2}(1-a\lambda)$ produces the second-order, double upwinding scheme for Warming and Beam [15]. Since the CFL limits for the Lax–Wendroff and Warming and Beam schemes are 1 and 2, respectively, the range of the evaluation point, $\alpha_{j\pm 1/2}^n$, must be restricted by

$$0\leqslant\alpha_{j\pm 1/2}^n\leqslant 1$$

for the Lax–Wendroff method and

$$-\tfrac{1}{2}\leqslant\alpha_{j\pm 1/2}^n\leqslant\tfrac{1}{2}$$

for the Warming and Beam method.

With the choice of $m_{j\pm 1/2}^n$ given in Eqs. (4) and (5), expansion of (2) yields the discretization

$$u_j^{n+1}=u_j^n+(a\lambda)\left\{\frac{1}{r_j^n}(1-\alpha_{j+1/2}^n)+\alpha_{j-1/2}\right\}(u_{j-1}^n-u_j^n), \tag{8}$$

where

$$r_j^n = \frac{u_{j-1}^n - u_j^n}{u_j^n - u_{j+1}^n}, \tag{9}$$

is the ratio of consecutive gradients of the unknown as used by Sweby [13] and others.

With the discretization in Eq. (8) we are now able to impose constraints on the numerical solution. Consider, for example, a monotonicity preserving requirement, i.e., the condition that if an initial solution profile at time level t_n is monotone, then the profile at time level t_{n+1} is also required to be monotone. Without loss of generality, we assume that

$$u_j^n \leqslant u_{j-1}^n. \tag{10}$$

The monotonicity constraint that will be imposed on the AOM scheme in this case is

$$u_j^n \leqslant u_j^{n+1} \leqslant u_{j-1}^n. \tag{11}$$

After substitution for u_j^{n+1} using Eq. (8), subtraction of u_j^n, and division by $u_{j-1}^n - u_j^n$, the discretization yields the basic inequality

$$0 \leqslant (a\lambda)\left\{\frac{1}{r_j^n}(1-\alpha_{j+1/2}^n) + \alpha_{j-1/2}^n\right\} \leqslant 1. \tag{12}$$

It is easy to show that the same inequality applies for both positive and negative ratios r_j^n which implies that all possible profiles of the unknown may be treated in the same manner.

If our objective is to preserve monotonicity, then the unmodified methods of Lax–Wendroff and Warming–Beam cannot be used. While these methods are second-order accurate, when $-1 < r_j^n < \frac{1}{3}$ the Lax–Wendroff method *will* violate the monotonicity condition for some values of $a\lambda$. Similarly, when $r_{j-1}^n < -1$ or $r_{j-1}^n > 3$ then the Warming–Beam method will also fail to be monotone. See Van Leer [14] for a thorough presentation of the precise regions where failure occurs. This potential failure of monotonicity occurs because these two methods "see" and accurately transport a polynomial of degree two which interpolates three consecutive solution values. Under the conditions described above on r_{j-1}^n or r_j^n, this quadratic polynomial transport can produce nonmonotone numerical results as illustrated in Fig. 1. The exact solution is a discontinuity, but the Lax–Wendroff method "sees" and transports the quadratic interpolatory polynomial, producing a nonmonotone result.

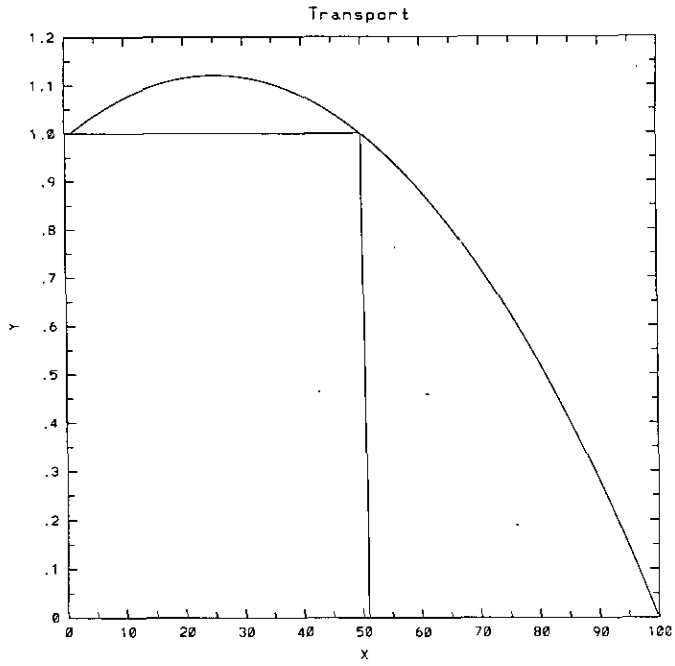

FIG. 1. The second-order Lax–Wendroff scheme transports the quadratic polynomial through the three points instead of the discontinuity. Thus nonmonotone results occur.

To preserve monotonicity with our discretization, we will modify the evaluation point of the linear polynomial approximation which defines $u_{j\pm 1/2}^n$. Instead of using $\alpha_{j\pm 1/2}^n$ in discretization (4), define the perturbed evaluation point

$$\eta_{j\pm 1/2}^n = \alpha_{j\pm 1/2}^n + \varepsilon_{j\pm 1/2}^n = \tfrac{1}{2}(1+a\lambda) + \varepsilon_{j\pm 1/2}^n \tag{13}$$

and then use the discretization which minimizes the perturbations, $\varepsilon_{j\pm 1/2}^n$, of the higher order method in a way that satisfies imposed constraints. For the work in this paper we choose to minimize $\|\varepsilon\|_2^2 = \sum (\varepsilon_{j+1/2}^n)^2$, subject to the imposed constraints. These constraints include satisfying the basic inequality for $\eta_{j-1/2}^n$,

$$0 \leqslant (a\lambda)\left\{\frac{1}{r_j^n}(1-\eta_{j+1/2}^n) + \eta_{j-1/2}^n\right\} \leqslant 1, \tag{14}$$

as well as the CFL limits which are

$$0 \leqslant \eta_{j\pm 1/2}^n \leqslant 1. \tag{15}$$

The choice of the l^2 norm has been made so that robust available algorithms for solving quadratic programming problems may be brought to bear on the minimization process. We are relying heavily on the wealth of theory underlying the solution of the quadratic programming problem. Other norms have been considered, but not implemented in the AOM framework.

Now the problem is formally one of minimizing a simple sum of squares quadratic function in $\varepsilon^n_{j\pm 1/2}$ subject to linear variable constraints. Because the nature of the objective function guarantees that a solution is unique if one exists and because choosing $\frac{1}{2}(1+a\lambda)+\varepsilon^n_{j+1/2}=1$ satisfies both the basic constraint and the CFL limit, a solution does exist (see Fletcher [3]). We are therefore left with a well-posed quadratic programming problem at each time step.

It is important to remark that this methodology is different from that of the flux-limiters described in Sweby [13]. Here we seek to "optimize" accuracy by varying minimally from a second-order method subject to constraints imposed to preserve desirable properties on the approximation of the solution rather than adding a maximal amount of "antidiffusive" flux. Additionally, instead of a local condition, this approach couples *all* grid line values if necessary. In theory, a minor variation from Lax–Wendroff at one point might have global effects. For example, a single modification imposed at a sharp front might require global modification, including some regions where the solution is smooth. As will be seen, however, the optimization problem may be localized in the sense that perturbations from the second-order method occur only over those regions where the basic inequality has forced modification. Section 4 contains a proof of this assertion.

To illustrate the AOM modification of the Lax–Wendroff method we consider the initial condition of the square wave described in Sweby [13], with $a\lambda=\frac{1}{2}$, and move the approximate solution 25 time steps. In the discussion below we will refer to resolution of the shock in terms of the number of grid blocks necessary to represent the jump between the left and right states of the discontinuity. In the numerical experiment throughout this paper, the jump will be from 0 to 1 or 1 to 0. For comparison, the same simulations are done with the compressive "Super-Bee" flux limiter of Roe [12] which is TVD and is a "high resolution" method. Figure 2 presents the results of the approximation for the AOM and Roe flux limiter method. The numerical solutions are superimposed on a graph with the exact solution. The figure illustrates two interesting and important properties. First, note that upwind of each discontinuity the shock is more highly resolved by the AOM than by the compressive flux limiter. That is, the number of grid blocks necessary to resolve the jump upwind is fewer for the AOM (two grid blocks) than the Roe flux limiter method (four grid blocks). This higher resolution upwind of the shock is an unexpected but favorable outcome of the AOM approach. Such resolution was not, however, a primary objective of this modified method. Note also that on the downwind side of each discontinuity the compressive method of Roe is superior to *this* version of the AOM in terms of shock resolution. The reason for the difference in resolution on upwind and downwind sides of the discontinuity is that, on the downwind side of a discontinuity, the method of Lax–Wendroff need not be modified at all to preserve the monotonicity property. Hence, the $\varepsilon^n_{j\pm 1/2}$ values are all zero and the method is monotonicity preserving, locally second order, and numerically diffusive. On the upwind side, the $\varepsilon^n_{j\pm 1/2}$ values must be adjusted away from zero in order to satisfy the monotonicity condition (14).

The asymmetry of the approximate solution in the method above is due to properties of the Lax–Wendroff

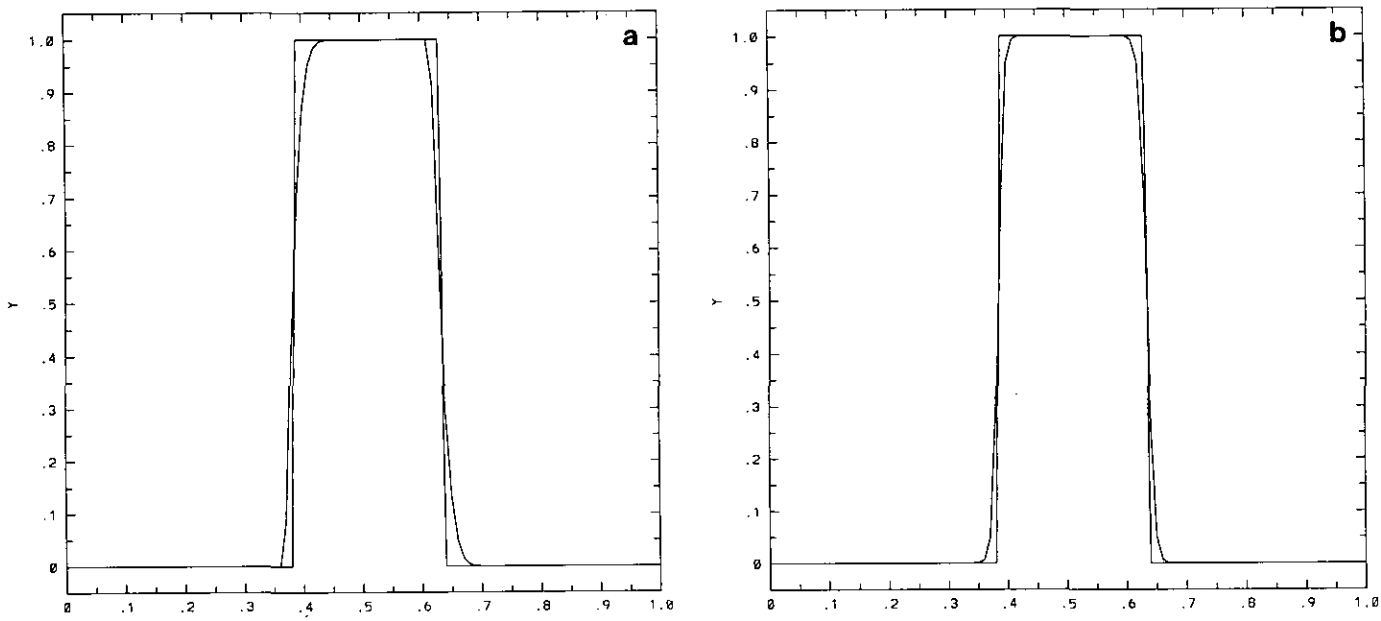

FIG. 2. Approximations of shock resolution in the transport of a square wave are shown for (a) the AOM using Lax–Wendroff as a base scheme and (b) the compressive "Super-Bee" scheme of Roe. The exact solution is plotted in each graph for comparison.

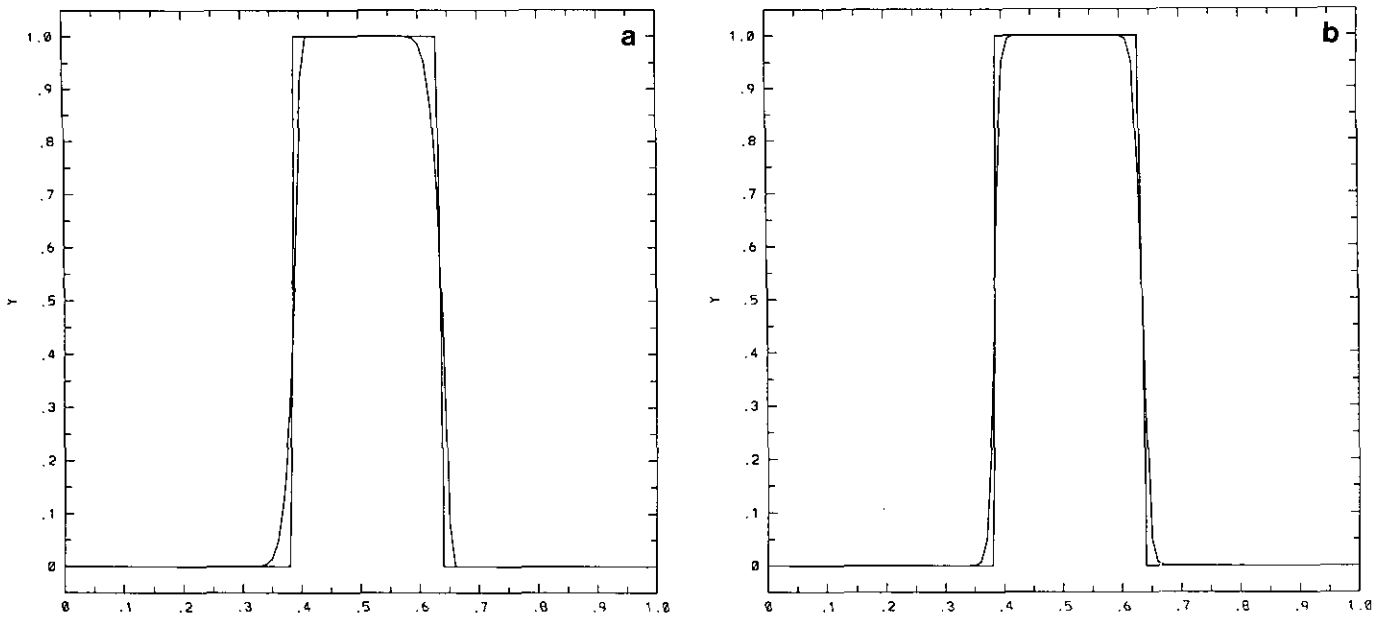

FIG. 3. Approximations of shock resolution in the transport of a square wave are shown for (a) the AOM using Warming and Beam as a base scheme and (b) the compressive "Super-Bee" scheme of Roe. The exact solution is plotted in each graph for comparison.

method and the monotonicity constraint. If, instead of using Lax–Wendroff as the basic method, we begin with the double upwinding scheme of Warming–Beam and minimally perturb it to satisfy the monotonicity constraint (11); then the results are predictably different. As before, the minimal perturbation is defined by minimizing the sum of squares of the $\varepsilon^n_{j+1/2}$, subject to our constraining conditions. The monotonicity constraint becomes

$$0 \leqslant (a\lambda)\{1 + \eta^n_{j+1/2} - r^n_{j-1}\eta^n_{j-1/2}\} \leqslant 1, \tag{16}$$

where

$$\eta^n_{j\pm 1/2} = \tfrac{1}{2}(1 - a\lambda) + \varepsilon^n_{j\pm 1/2}, \tag{17}$$

and the CFL condition on $\eta^n_{j\pm 1/2}$ remains unchanged.

Figure 3 illustrates the results when the modified Warming and Beam method is applied to the same square wave used earlier. Note that with this approach the downwind side of shock discontinuities are resolved more highly when compared with the compressive flux limiter of Roe but that the upwind side shows more diffusion. The reasons for this asymmetry are identical with the remarks made above about the modified Lax–Wendroff method. On the downwind side of steep gradients the Warming–Beam method must be modified to satisfy the constraint, while on the upwind side of steep gradients the interpolating polynomial transported by the method never violates the constraint; hence the method need not be modified and increased numerical diffusion appears.

One possible approach at this point would be to average the two outcomes developed in this section to produce a "symmetric" method. This would be analogous to the scheme of Fromm [4]. Instead, we propose in the next section a combination of the two methods *before* the optimization is carried out, which will produce distinctly different and better results.

3. A CONVEX COMBINATION AOM

Motivated by the results of the previous section a "symmetric" version of the AOM is described. In the basic conservation discretization, define

$$u^n_{j+1/2} = \tfrac{1}{2}\{u^n_{j+1} + (u^n_j - u^n_{j+1})\,\eta^n_{j+1/2} + u^n_j - (u^n_j - u^n_{j-1})\,\rho^n_{j+1/2}\} \tag{18}$$

and

$$u^n_{j-1/2} = \tfrac{1}{2}\{u^n_j + (u^n_{j-1} - u^n_j)\,\eta^n_{j-1/2} + u^n_{j-1} - (u^n_{j-1} - u^n_{j-2})\,\rho^n_{j-1/2}\}. \tag{19}$$

Here we have taken an equally weighted combination of approximate values of the $u^n_{j\pm 1/2}$ from the Lax–Wendroff and Warming–Beam methods. Note that we have added a second set of variables, $\rho^n_{j\pm 1/2}$, to the discretization. If we choose $\eta^n_{j\pm 1/2} = \frac{1}{2}(1 + a\lambda)$ as in Section 2 and $\rho^n_{j\pm 1/2} = \frac{1}{2}(1 - a\lambda)$, then the method simplifies to the average of Lax–Wendroff and Warming–Beam and is, therefore, second-order accurate. Requiring the method to satisfy the

monotonicity constraints leads, after some calculation, to the inequality

$$0 \leqslant \frac{(a\lambda)}{2}\left\{\frac{1}{r_j^n}(1-\eta_{j+1/2}^n)+\eta_{j-1/2}^n+1 + \rho_{j+1/2}^n - r_{j-1}^2\rho_{j-1/2}^n\right\} \leqslant 1. \tag{20}$$

The AOM uses perturbation terms $\varepsilon_{j\pm 1/2}^n$ and $\delta_{j\pm 1/2}^n$ of the second-order accurate values of $\eta_{j\pm 1/2}^n$ and $\rho_{j\pm 1/2}^n$; i.e., define

$$\eta_{j\pm 1/2}^n = \tfrac{1}{2}(1+a\lambda)+\varepsilon_{j\pm 1/2}^n \tag{21}$$

and

$$\rho_{j\pm 1/2}^n = \tfrac{1}{2}(1-a\lambda)+\delta_{j\pm 1/2}^n. \tag{22}$$

The addition of such perturbations will allow the method to preserve monotonicity and, as before, permit us to "optimize" accuracy by minimally perturbing the second-order method in such a way as to meet imposed constraints. Carrying out this additional calculation yields the fundamental monotonicity inequality

$$-2 \leqslant \frac{1}{r_j^n}\left\{1-\frac{1}{2}(1+a\lambda)-\varepsilon_{j+1/2}^n\right\}+\varepsilon_{j-1/2}^n+\delta_{j+1/2}^n - r_{j-1}^n\left\{\frac{1}{2}(1-a\lambda)+\delta_{j-1/2}^n\right\} \leqslant \frac{2}{a\lambda}-2. \tag{23}$$

Formally, we state this particular symmetric AOM as

$$u_j^{n+1} = u_j^n - a\lambda(u_{j+1/2}^n - u_{j-1/2}^n), \tag{24}$$

where $u_{j\pm 1/2}^n$ are defined as in Eqs. (18) and (19). The values of $\eta_{j\pm 1/2}^n$ and $\rho_{j\pm 1/2}^n$ are defined in Eqs. (21) and (22) and are chosen to minimize

$$\sum[(\varepsilon_{j+1/2}^n)^2+(\delta_{j+1/2}^n)^2], \tag{25}$$

subject to the monotonicity constraint of Eq. (23) and the two CFL constraints implied by the CFL limits on the two unmodified methods. In the linear advection case this is, as before, a well-posed quadratic programming problem with linear variable constraints. The form of the quadratic objective function, a simple sum of squares, coupled with the known existence of a feasible point, guarantees the uniqueness of the solution of the optimization problem.

As a secondary feature, we observe that in every case tested this method has very high resolution shock-capturing properties which compare favorably with the "Super Bee" limiter of Roe. This feature will be illustrated in Section 7. Further, we note that the procedure of combining Lax–Wendroff and Warming–Beam in a convex combination and then optimizing is mathematically different from optimizing each separately and then combining afterwards. As will be discussed in Section 7, the extra play which comes from having four variables per combined constraint equation, as opposed to the two variables in each of the separate methods, produces a different set of perturbation values and better resolution of every shock front tested.

The advantage in using the symmetric method on problems can be seen in the numerical experiments presented below. The first two methods presented both have some dissipative behavior on one side of the shock. The symmetric method produces results that retain the best resolution properties of each of the one-sided methods while eliminating the dissipative effects of both the one-sided methods. Note that if the one-sided methods produce undesirable dissipative behavior in linear advection problems then these methods cannot be expected to produce better results in nonlinear problems. Thus it seems appropriate to use the symmetric method over either of the one-sided methods for both linear and nonlinear problems.

Finally, an error analysis of this combined method leads to the result that the general perturbation from a second-order method is given by

$$\frac{-a\lambda}{2}(\Delta x(u_x)_j^n)[-\varepsilon_{j+1/2}^n+\varepsilon_{j-1/2}^n-\delta_{j+1/2}^n-\delta_{j-1/2}^n] + O(\Delta x)^2. \tag{26}$$

It should be expected that constraining the numerical method as we have precludes a method which is second order everywhere. Note, however, that if the $\varepsilon_{j\pm 1/2}^n$ and $\delta_{j\pm 1/2}^n$ are zero in the above equation then entire first-order perturbation term disappears and the method is, as expected, second order.

4. LOCALIZATION OF THE OPTIMIZATION PROBLEM

One potential problem with the approach of the AOM is that the global optimization problem applied to the entire domain may be computationally expensive. However, the optimization process can be localized to those regions with steep gradients, thereby avoiding excessive computation which might arise from large scale discretizations and a global optimization process. We formally state the result as a theorem.

THEOREM 1. *For the convex combination AOM addressed in Section* 3, *and on those regions where both ratios satisfy*

$$\tfrac{1}{2} \leqslant |r_j^n, r_{j-1}^n| \leqslant 2, \tag{27}$$

the global *optimization process need not be utilized, and in fact the solution to the global optimization over the entire discretization problem is identical to that obtained by using local optimization processes only where the ratios violate the constraint above.*

Proof. Consider the global optimization problem, without regard to local ratios. It has a unique solution from the form of the quadratic objective functional and the linearity of variable constraints. If there is a region where all ratios satisfy the inequality (27) then arbitrary replacement of the values of $\varepsilon^n_{j\pm 1/2}$ and $\delta^n_{j\pm 1/2}$ by zeros in that region will satisfy both the inequality constraint in Eq. (23), the CFL limits, and it will not increase the objective function value. In fact, if these variables were not zero in the global optimization solution, the objective function would, upon replacement by such zero values, decrease, which is a violation of the existence and uniqueness of the global optimization solution. It is a simple process to show that for the ratio limits described in (27) zero values assigned to $\varepsilon^n_{j\pm 1/2}$ and $\delta^n_{j\pm 1/2}$ will not violate any inequality constraints defined by (23) or CFL constraints and hence they are feasible. ∎

Therefore, in such regions where the solution is smooth we may ignore the optimization process, preserve the second-order scheme directly, and use a local optimization process for regions where sharp profiles exist.

We conclude this section with three remarks about the theorem. First, note that with some extra work one can alter the end points of the region in Eq. (27) as a function of $a\lambda$, as in Van Leer [14]. However, the values given are the sharpest possible for all values of $a\lambda$, where all we assume is that $a\lambda \leqslant 1$. Second, the endpoints in (27) are precisely the left and right endpoints of Sweby's flux limiter region [13], where the limiter definition changes. Finally, note that each disjoint region where perturbations are necessary can be treated independently. Thus disjoint optimization problems can be solved in parallel.

5. NONLINEAR PROBLEMS—BURGERS' EQUATION

In this section we briefly discuss the fundamental problem (1), in the case where f is nonlinear and assume $f' > 0$. As in Section 2 the AOM begins with the numerical scheme written in the fundamental conservation form

$$u_j^{n+1} = u_j^n - \lambda\{f(u^n_{j+1/2}) - f(u^n_{j-1/2})\}, \tag{28}$$

where $u^n_{j\pm 1/2}$ are defined as in Eq. (4), i.e.,

$$u^n_{j+1/2} = u^n_{j+1} + (u^n_j - u^n_{j+1})\,\eta^n_{j+1/2} \tag{29}$$

and

$$u^n_{j-1/2} = u^n_j + (u^n_{j-1} - u^n_j)\,\eta^n_{j-1/2}. \tag{30}$$

One approach in this nonlinear case is to define the $\eta^n_{j\pm 1/2}$ by

$$\eta^n_{j\pm 1/2} = \tfrac{1}{2}\{1 + \lambda f'(u^n_{j\pm 1/2})\} + \varepsilon^n_{j\pm 1/2}. \tag{31}$$

Combining the above leads to

$$u^n_{j\pm 1/2} = u^n_{j+1} + (u^n_j - u^n_{j+1}) \times \{\tfrac{1}{2}\{1 + \lambda f'(u^n_{j+1/2})\} + \varepsilon^n_{j+1/2}\}, \tag{32}$$

and a similar equation in $u^n_{j-1/2}$. The particular discretization above is the nonlinear analog of the Lax–Wendroff scheme within the AOM framework.

There are two major points to be made with this discretization. First, $u^n_{j\pm 1/2}$ are defined implicitly, involving f' evaluated at $u^n_{j\pm 1/2}$. In most cases this implicit definition requires an iterative loop at each time step in order to solve for the $u^n_{j\pm 1/2}$. If one makes reasonable assumptions about the smoothness of f, then a fixed point iteration will converge for small enough λ (Koebbe [6]). If we define

$$F(u) = u - m^n_{j\pm 1/2}(\alpha^n_{j\pm 1/2}(u)),$$

then one can compute Newton iterates using

$$(u^n_{j+1/2})^{k+1} = (u^n_{j+1/2})^k - \frac{(u^n_{j+1/2})^k - m^n_{j+1/2}(\alpha^n_{j+1/2}((u^n_{j+1/2})^k))}{1 - m^n_{j+1/2}{}'(\alpha^n_{j+1/2}((u^n_{j+1/2})^k))},$$

where the derivative of $m^n_{j+1/2}(\alpha^n_{j\pm 1/2}(u))$ is straightforward to compute.

For example, in the case of the Lax–Wendroff analogy for the nonlinear problem

$$1 - m^n_{j+1/2}{}'(\alpha^n_{j+1/2}((u^n_{j+1/2})^k)) = 1 - \tfrac{1}{2}(u^n_j - u^n_{j+1})\,\lambda f''((u^n_{j+1/2})^k).$$

Thus if bounds are known on the function f'' the range of λ can be chosen so that the denominator is never zero. Note that the computation of these bounds is dependent on the steepest gradient in the profile of the unknown u at the old time level t_n.

Second, it can be shown after considerable calculation that if $\varepsilon^n_{j\pm 1/2}$ are all identically zero then the method defined in Eq. (28) is formally second-order accurate. The computations required rely heavily on the implicit definition of the evaluation point in Eq. (31). Expansion of the derivative of f in the evaluation point allows cancellation of terms in

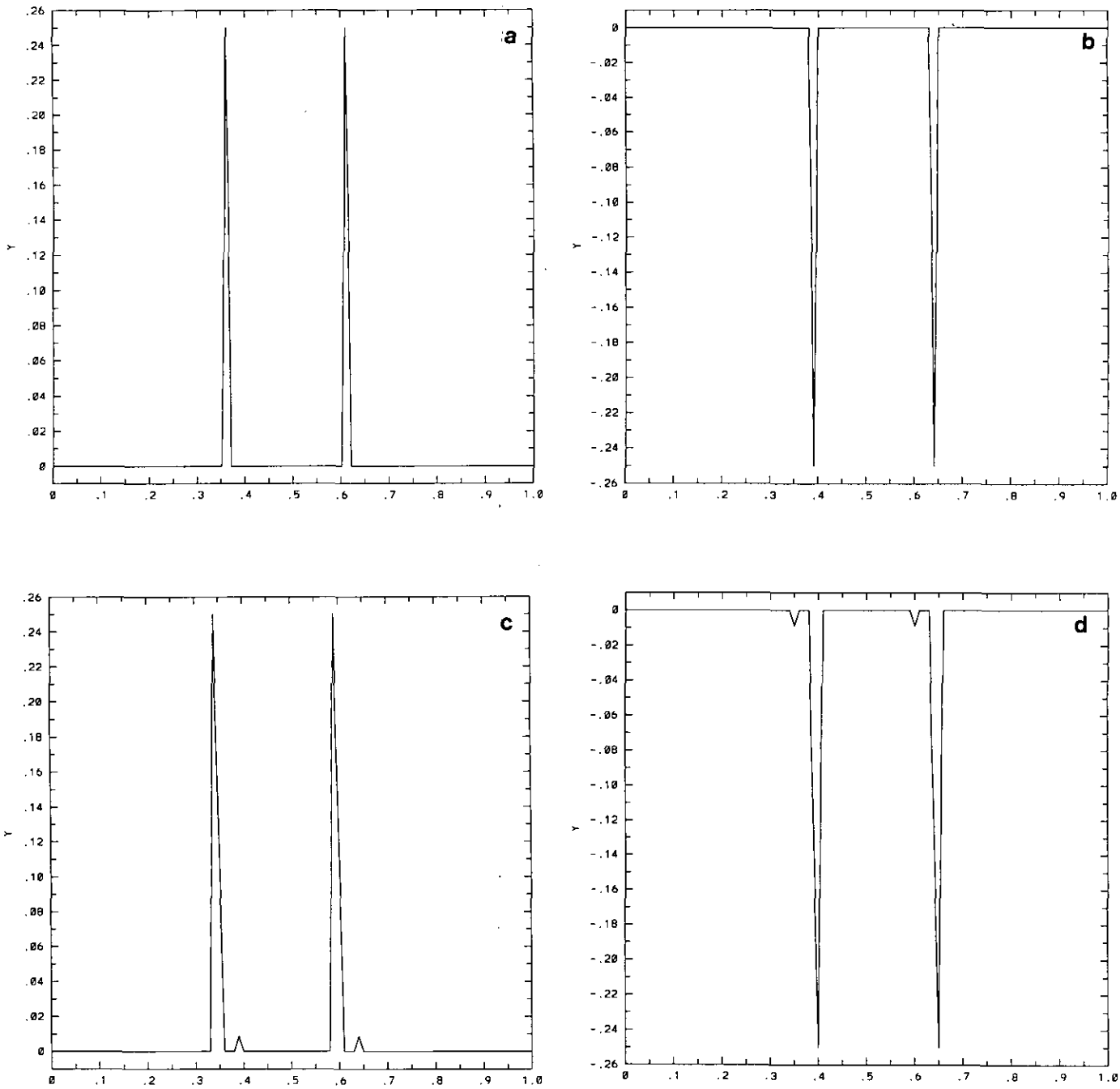

FIG. 6. The values of the perturbations in the domain of the simulation for (a) $\varepsilon^n_{j+1/2}$ in the Lax–Wendroff based AOM, (b) $\varepsilon^n_{j+1/2}$ in the Warming and Beam based AOM, and (c) $\varepsilon^n_{j+1/2}$ and (d) $\delta^n_{j+1/2}$ in the combined AOM. Note the small "bumps" near the large spikes in (c) and (d) which are consequences of combining the methods before optimization.

$\lambda = 0.25$ was chosen and the experiment was run on 20, 40, and 60 nodes. Tables I and II each show the computational convergence rates for the l_∞, l_1, and l_2 norms for both the symmetric AOM and the symmetric base method for the AOM without optimization at 10 time levels in the simulation. The last time level in both tables corresponds to taking 300 steps with 40 nodes. The results show degradation in convergence for some steps, but this also happens at the same time levels for the unmodified scheme. Also, note that the convergence rates were computed for fairly coarse grids.

Finally, note that the errors were computed for time levels beyond the time necessary to transport the wave

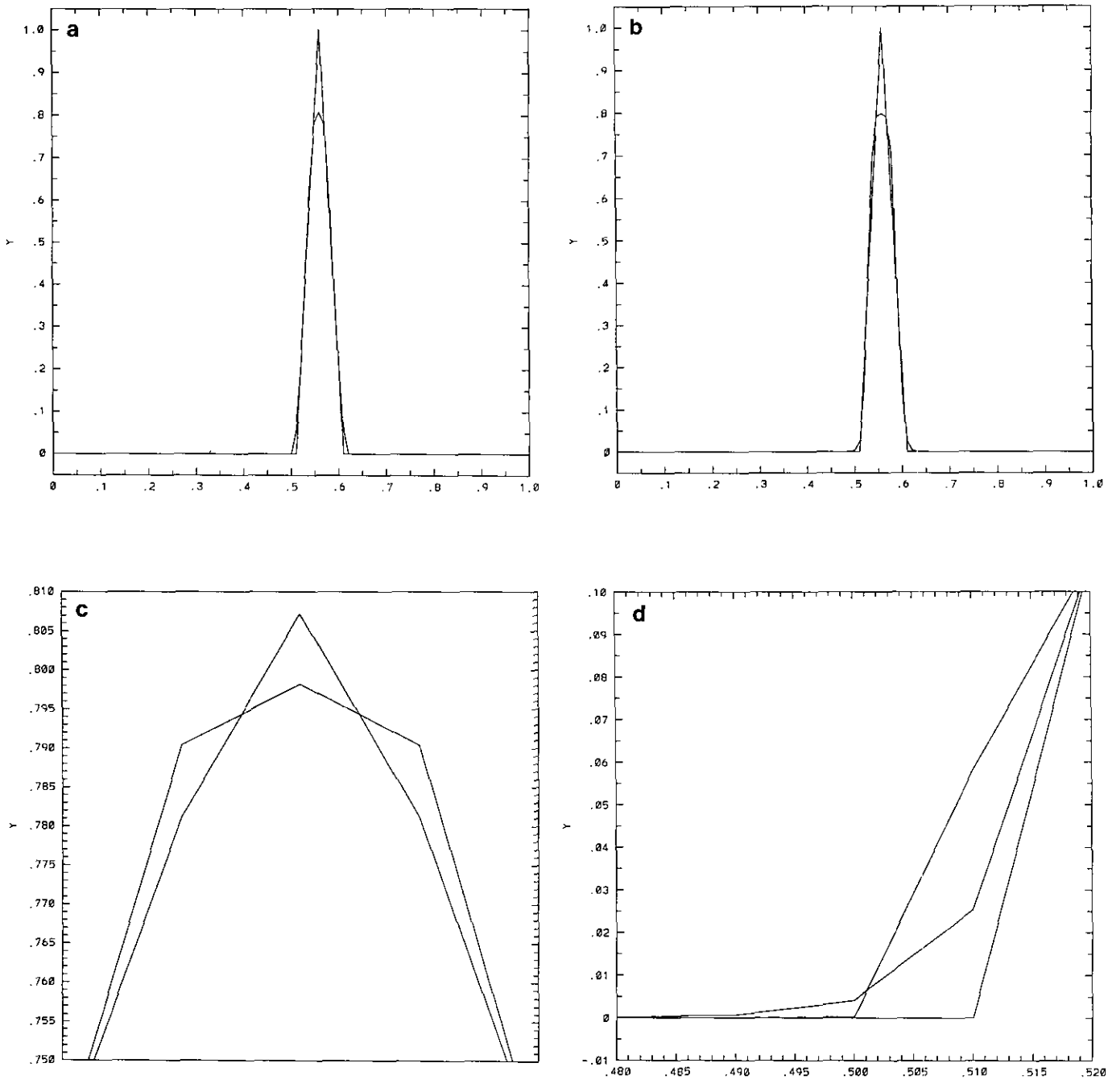

FIG. 7. Approximations of the solution of the linear advection problem with spike initial data using the (a) AOM of Section 3, (b) the "Super-Bee" flux limited scheme. The methods are compared at the peak of the spike in (c), and at the tails in (d). The exact solution is shown in both (a) and (b) for comparison.

one period. Any problems that might occur due to the right boundary should have appeared by the end of the simulation.

7.2. *Nonlinear Advection–Burgers' Equation Examples*

For the first example using the nonlinear Burgers' equation discussed in Section 5 we used the initial data

$$u_0(x) = u(x, 0) = \begin{cases} 1.0, & 0.0 \leqslant x < 0.3 \\ -5x + 2.5, & 0.3 \leqslant x < 0.5 \\ 0.0, & 0.5 \leqslant x \leqslant 1.0. \end{cases}$$

In this example the solution was advanced 52 time steps, with $\Delta x = 0.01$ and $\lambda = 0.5$. This number of time steps permits comparison of the methods through shock formation

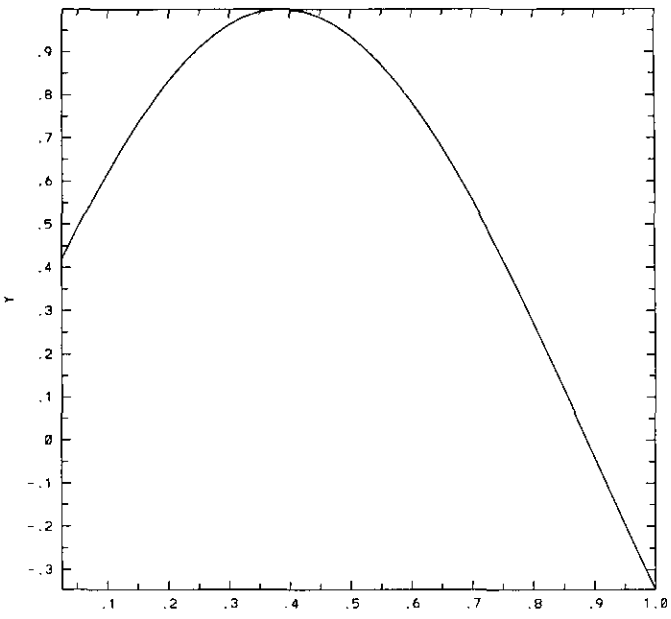

FIG. 8. Results from transport of a sine wave, $u(x, 0) = \sin(\pi x)$, by the symmetric AOM plotted against the analytic solution on the interval $[0, 1]$ with 40 nodes and Courant number 0.25 after 300 time steps. There is virtually no difference between the curves.

TABLE I

Computational Convergence Rates for the Second-Order Symmetric Base Method with and without Optimization

	Symmetric AOM			Symmetric base scheme		
Time level	l_∞ error	l_1 error	l_2 error	l_∞ error	l_1 error	l_2 error
1	1.59819	2.44867	2.20196	2.05573	2.10069	2.09816
2	2.21231	2.20251	2.24067	2.05873	2.13543	2.11736
3	1.43680	2.12129	2.00977	2.06152	2.07100	2.08848
4	1.65184	2.17060	2.10055	2.16902	2.01565	2.03341
5	1.90220	2.26906	2.21909	2.00418	1.99502	1.97135
6	2.03653	2.09261	2.08751	1.81232	2.05628	1.99737
7	1.47902	2.00424	1.78617	2.21534	2.18637	2.19825
8	1.35335	1.81416	1.62176	2.17955	2.16051	2.18591
9	1.61260	2.17024	2.09271	2.27502	2.05943	2.09889
10	2.21526	2.15741	2.20735	2.07279	2.00853	2.01447

Note. The convergence rates are computed between cases with 20 and 40 nodes on a unit interval at corresponding time steps. In this case $\lambda = 0.25$ was chosen and the last time level represents 300 time steps for the 40-node case and 150 time steps with 20 nodes.

(40 time steps) and transport of the resulting shock for 12 additional steps. Additional time steps produced no increase in the shock resolution width. With this scale the exact shock speed is 0.25 units per time step, and this is accurately represented by the numerical result. Figure 10 presents the approximation of the solution for the unmodified Lax–Wendroff scheme and the AOM using the Lax–Wendroff scheme as the base scheme for three time levels. The first two time levels occur during shock formation and the third is after the shock is formed and has been transported. Note the significant oscillations which have appeared in the unmodified scheme on the upwind side of the shock. The small diffusive behavior downwind of the shock in

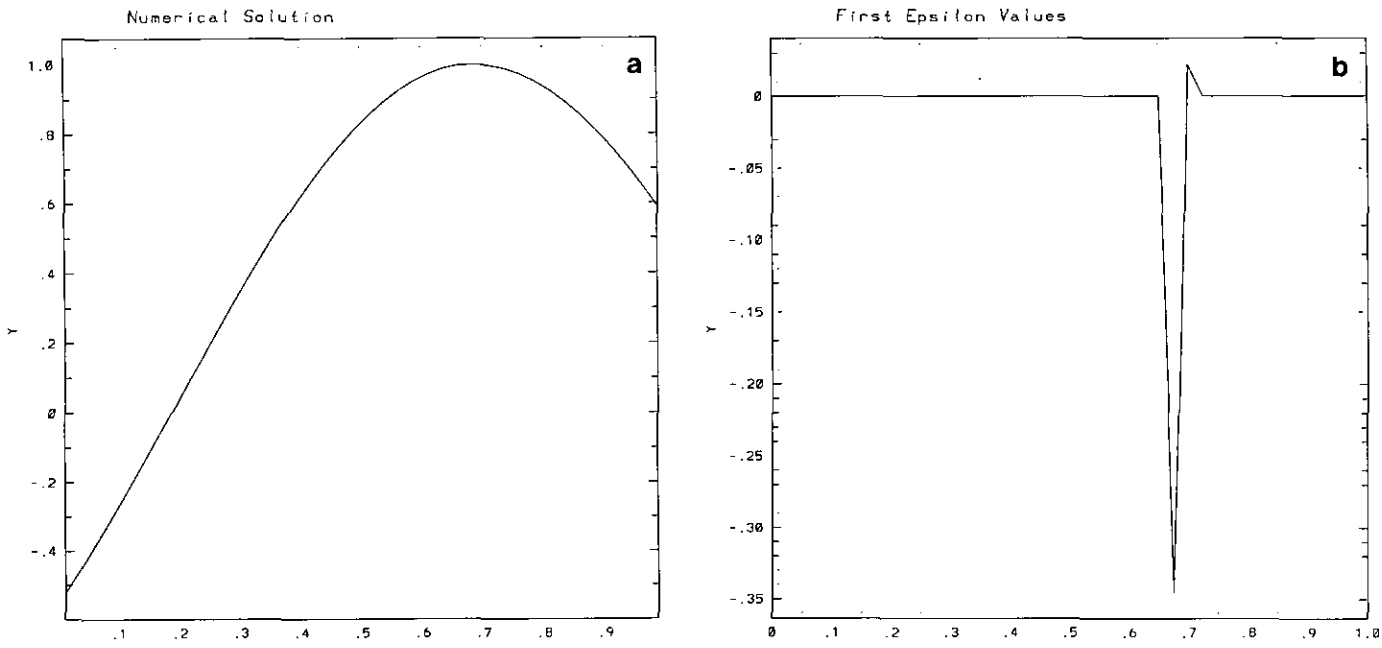

FIG. 9. Transport of $u(x, 0) = \sin(\pi x)$ after 30 time steps is shown in (a). The perturbations necessary at the time step for the Lax–Wendroff portion of the symmetric method are shown in (b). The perturbation is very local and occurs due to a flat spot in the representation of u at the time level.

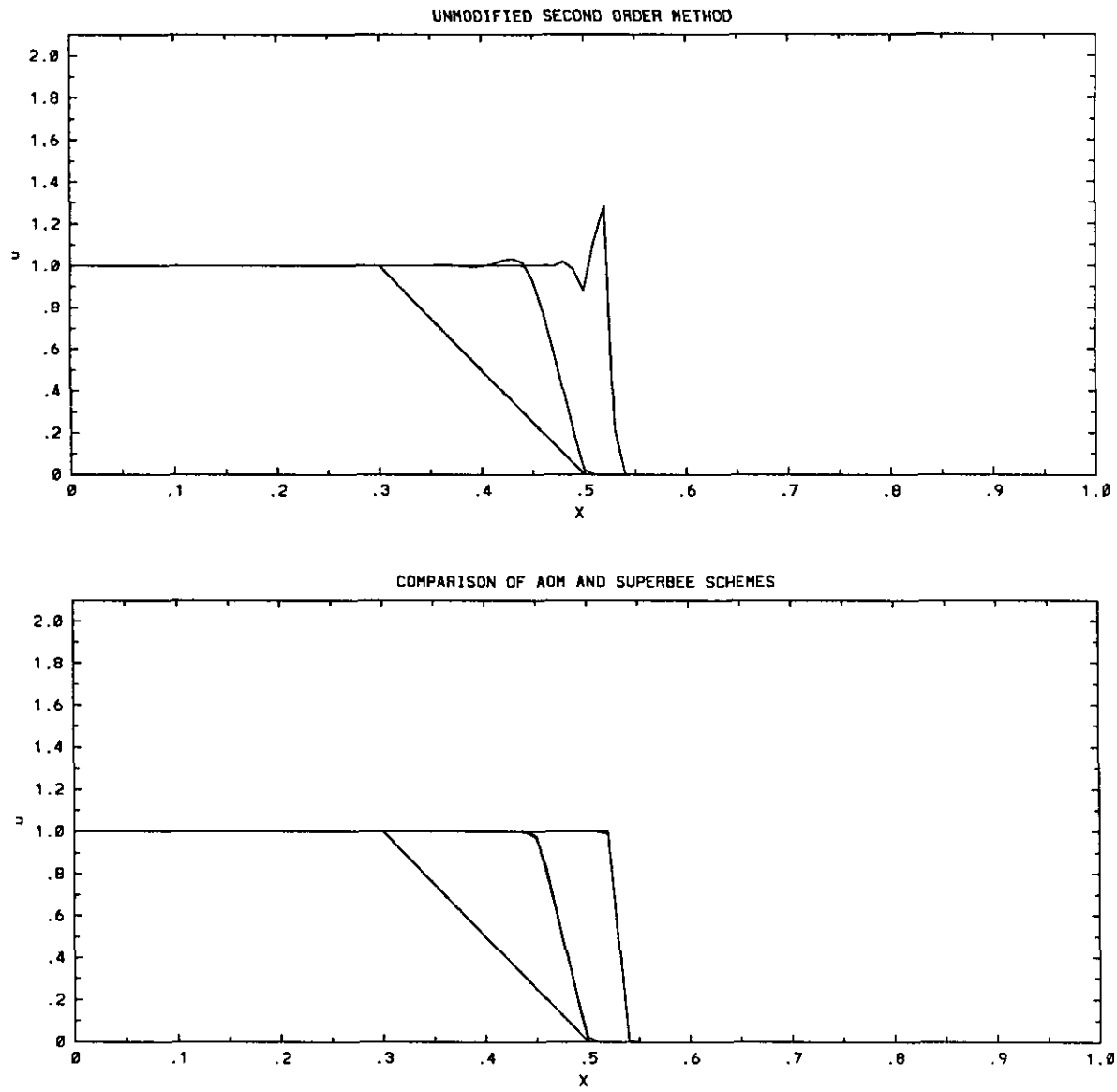

FIG. 10. The development and transport of a shock for Burgers' equation using the unmodified higher order method (top) and using the AOM based on the Lax–Wendroff method and the Super-Bee flux-limited methods (bottom). The graphs show several time steps on the same plot to illustrate the evolution of the approximations. The difference in the two methods in (b) is the same as shown in the square wave example in the linear advection problems.

the results produced by the AOM method is due to the Lax–Wendroff base scheme. The same behavior was observed in the linear advection problem discussed in Sections 2 and 3.

In the second example we applied the symmetric AOM using the initial data

$$u(x, 0) = |\sin(2\pi x)|, \qquad x \in \Re, \qquad t > 0,$$

with $\lambda = 0.25$, 99 nodes, and a total of 100 time steps. Figure 11 shows the results of the simulation at various time steps. Figure 11a shows the initial condition, Fig. 11b shows the approximation just before the expected shock develops, Fig. 11c shows the approximation as the shock is forming, and Fig. 11d shows the approximation after the shock has fully formed and has begun to move. The "lip" in Fig. 11c shows that the shock is forming in a nonlinear fashion, which is expected, due to the form of the initial condition.

To try to understand the cause of the "lip" in this example several other initial conditions were tested. These included the top half of an ellipse repeated periodically to match this example and cases where only a single hump of the sine wave and ellipse were used. In each case a "lip" appeared. However, the worst case is that shown in Fig. 12. In the

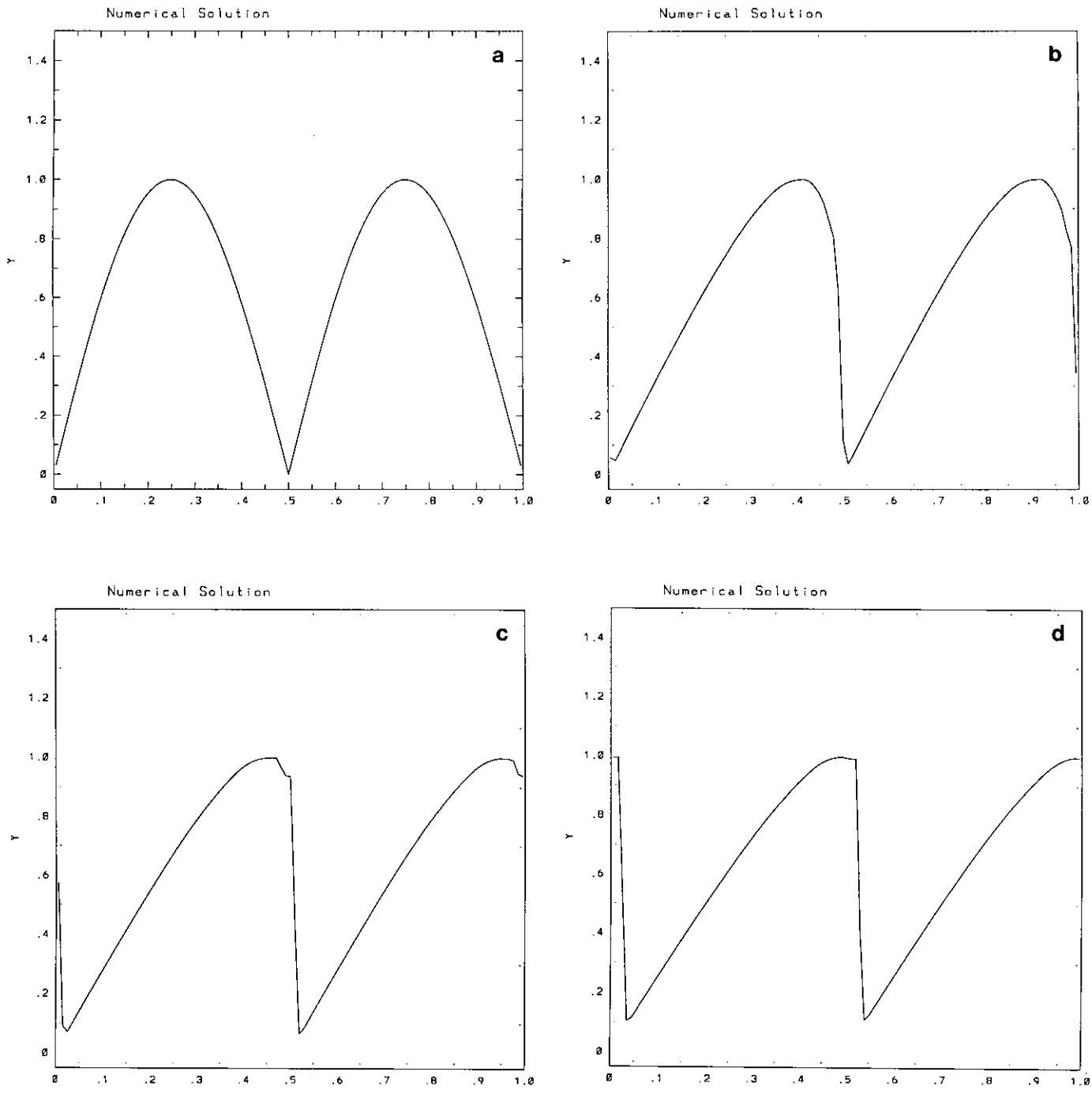

FIG. 11. Transport of $|\sin(2\pi x)|$ in Burgers' equation using an AOM with a second-order base scheme. The four figures include (a) the initial condition, (b) the approximation just before shock formation, (c) the approximation as the shock is forming, and (d) the approximation after the shock is fully formed and being transported.

other three cases the "lip" was restricted to one node and could be attributed to a representation error on the grid. Unfortunately, the best way for seeing this effect cannot be presented here. In animations of the numerical results it is easy to see the effect. In the two cases with sine wave initial conditions the wave breaks after a finite time and, due to the steepness of the initial condition, the shock grows very rapidly. In the ellipse cases the shock starts building immediately and does not build as rapidly. Thus the effect is reduced in the ellipse cases. Finally, the initial condition

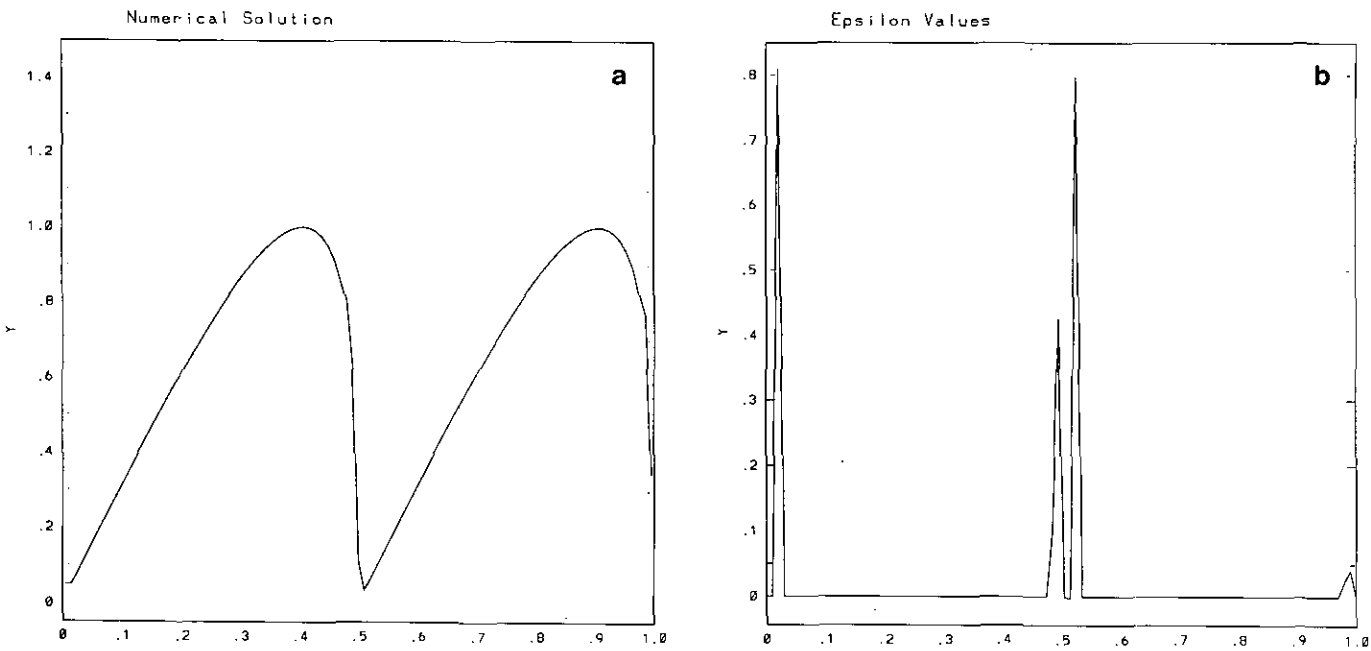

FIG. 12. Transport of $|\sin(2\pi x)|$ in Burgers' equation using AOM at the same time step as in Fig. 11b with perturbation values over the entire domain.

with the single sine wave hump was not as bad as the repeated case. This indicates that the interaction of successive humps in the periodic case is contributing to this behavior.

In this test problem the same localization occurs which was seen in the linear advection examples. Figure 12b shows the graph of the perturbations over the entire domain.

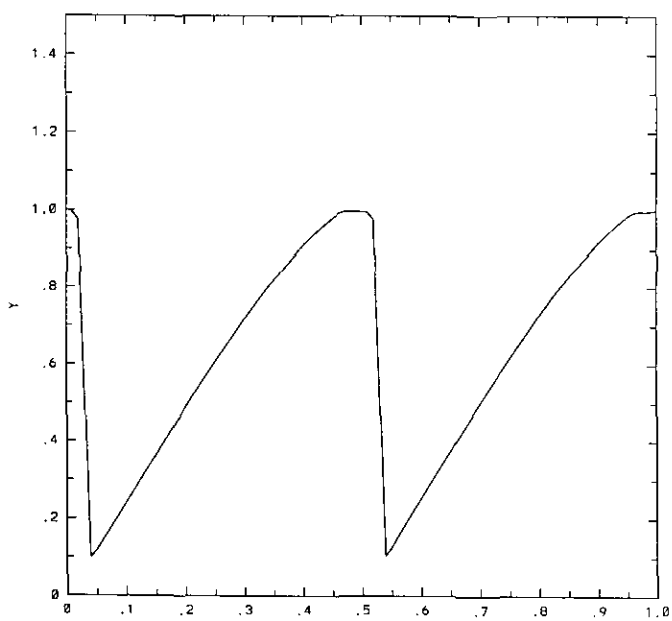

FIG. 13. Results from applying the Roe flux-limited method to the initial condition $|\sin(2\pi x)|$ in Burgers' equation. Note that the method flattens out the tops of the sine wave unlike the AOM method results.

Figure 12a is the same as Fig. 11a and is included to ease comparison of the perturbation locations and the numerical solution. Again the nonzero values of the perturbation are localized to small bands. The localization of the optimization will be addressed in a paper on the AOM applied to nonlinear problems. Finally, the results of the same simulation using the Roe flux-limited scheme are shown in Fig. 13. Note that this method tends to flatten off the wave at the top, unlike the AOM scheme.

TABLE II

Computational Convergence Rates for the Second-Order Symmetric Base Method with and without Optimization

	Symmetric AOM			Symmetric base scheme		
Time level	l_∞ error	l_1 error	l_2 error	l_∞ error	l_1 error	l_2 error
1	2.57528	2.52997	2.50473	2.02300	2.04479	2.04561
2	1.49990	2.20596	1.96964	2.02398	2.06801	2.06221
3	1.96935	2.12120	2.17618	2.01949	2.04137	2.04837
4	1.97333	2.21246	2.16833	2.02195	2.00166	2.01596
5	1.86730	2.08442	1.95579	2.01873	1.98250	1.97761
6	1.36742	2.00064	1.62783	1.91111	2.00202	1.96876
7	1.90622	1.95567	1.83797	1.81501	2.07989	2.08847
8	1.40433	2.08267	1.90834	2.09656	2.09840	2.10524
9	2.50421	2.23529	2.30559	2.16458	2.02907	2.05377
10	1.61054	2.17912	2.03851	2.05311	1.99083	2.00162

Note. The convergence rates are computed between cases with 40 and 60 nodes on a unit interval at corresponding time steps. In this case $\lambda = 0.25$ was chosen and the last time level represents 300 time steps for the 40-node case and 450 time steps with 60 nodes.

8. SUMMARY

In this paper a method for the construction of constrained numerical approximations of scalar conservation laws has been introduced. The AOM methods optimize accuracy of the approximation, subject to imposed constraints. The methods require the solution of a well-posed optimization problem at each time step. The optimization problems can be localized to regions where the solution is not smooth. All examples of the AOMs presented here are TVD and produce high resolution approximations. The methods were tested on linear advection examples and Burgers' equation.

ACKNOWLEDGMENTS

The authors acknowledge the thorough reading given this paper by the referees. The proof of Theorem 2 was suggested by one of the reviewers. The proof related the material in this paper to existing TVD literature better than the original proof.

REFERENCES

1. J. P. Boris and D. L. Book, *J. Comput. Phys.* **11**, 38 (1973).
2. C. S. Coray and J. V. Koebbe, *SIAM J. Sci. Stat. Comput.*, submitted.
3. R. Fletcher, *Practical Methods of Optimization*, 2nd ed. (Wiley, New York, 1987).
4. J. E. Fromm, *J. Comput. Phys.* **3**, 176 (1968).
5. A. Harten, *J. Comput. Phys.* **49**, 357 (1983).
6. J. V. Koebbe, Ph.D. thesis, University of Wyoming, August 1988.
7. P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves* (SIAM, Philadelphia, 1973).
8. P. D. Lax and B. Wendroff, *Commun. Appl. Math.* **13**, 217 (1960).
9. R. J. LeVeque, *Numerical Methods for Conservation Laws* (Birkhauser, Basel, 1990).
10. S. Osher, *SIAM J. Numer. Anal.* **21** (2), 217 (1984).
11. S. Osher and S. Chakravarthy, High resolution schemes and the entropy condition. *SIAM J. Numer. Anal.* **21** (5), 955 (1984).
12. P. L. Roe, "Some Contributions to the Modelling of Discontinuous Flows," Proceedings, AMS/SIAM Seminar, San Diego, 1983.
13. P. K. Sweby, *SIAM J. Numer. Anal.* **21** (5), 995 (1984).
14. B. Van Leer, *J. Comput. Phys.* **14**, 361 (1974).
15. R. F. Warming and R. M. Beam, *AIAA J.* **14**, 1241 (1976).
16. S. T. Zalesak, *J. Comput. Phys.* **31**, 335 (1979).