MATHEMATICS OF COMPUTATION
VOLUME 52, NUMBER 186
APRIL 1989, PAGES 411-435

TVB Runge-Kutta Local Projection Discontinuous Galerkin Finite Element Method for Conservation Laws II: General Framework

By Bernardo Cockburn and Chi-Wang Shu

Dedicated to Professor Eugene Isaacson on the occasion of his 70th birthday

Abstract. This is the second paper in a series in which we construct and analyze a class of TVB (total variation bounded) discontinuous Galerkin finite element methods for solving conservation laws $u_t + \sum_{i=1}^{d}(f_i(u))_{x_i} = 0$. In this paper we present a general framework of the methods, up to any order of formal accuracy, using scalar one-dimensional initial value and initial-boundary problems as models. In these cases we prove TVBM (total variation bounded in the means), TVB, and convergence of the schemes. Numerical results using these methods are also given. Extensions to systems and/or higher dimensions will appear in future papers.

1. Introduction. We consider numerical solutions of the hyperbolic conservation law

$$u_t + \sum_{i=1}^{d}(f_i(u))_{x_i} = 0 \tag{1.1}$$

equipped with suitable initial or initial-boundary conditions. Here,

$$u = (u_1, \ldots, u_m)^t, \qquad x = (x_1, \ldots, x_d),$$

and any real combination of the Jacobian matrices

$$\sum_{i=1}^{d} \xi_i \frac{\partial f_i}{\partial u}$$

has m real eigenvalues and a complete set of eigenvectors.

In this paper we treat the special case of (1.1) with $d = m = 1$. We use this as a model to present the framework of the schemes, to prove theoretical results, and to highlight the essential ingredients of the methods, while keeping in mind that these methods are naturally extendable to $d > 1$ and/or $m > 1$.

The main difficulty in solving (1.1) is that solutions may contain discontinuities even if the initial conditions are smooth. Among the successful numerical methods for solving (1.1) we mention the modern nonoscillatory conservative finite difference methods such as TVD (total variation diminishing), TVB (total variation bounded)

Received February 23, 1988.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65M60, 65N30, 35L65.
Key words and phrases. TVD, TVB, Runge-Kutta, discontinuous finite elements, conservation law.

and ENO (essentially nonoscillatory) schemes (see, e.g., [7]–[10], [20]–[28] and the references listed there). These methods are usually total variation stable (in contrast with only L^2-stability of many traditional numerical methods, which may behave erratically if discontinuities are present) for one-dimensional scalar nonlinear problems ($d = m = 1$ in (1.1)), hence can capture shocks sharply without introducing oscillations. Extensions of these methods to higher dimensions and/or systems usually work well in rectangular coordinates or in general geometries by using conformal mapping (two-dimensional case) or some other mapping. Finite difference schemes have the big advantage of being simple to code, but they usually achieve high-order accuracy by using a wide stencil, hence they may "pollute", i.e., lose accuracy in a fairly large region near shocks, and they usually are difficult to apply in complicated geometries and/or boundary conditions. The first difficulty (polluting) may be considered overcome by ENO schemes via an adaptive stencil idea, but the second difficulty remains, rendering finite difference methods not as flexible as finite element methods in treating complicated geometries and/or boundary conditions. There is also considerable ongoing research on finite element methods for solving (1.1). We mention in particular the successful streamline diffusion method introduced by Hughes et al. [11], [12], [14]–[16]. Oscillations are usually greatly reduced, and, by *assuming* an L_∞ bound of the numerical solution (in the second-order case the L_∞ bound was recently proved for Burgers' equation [16]), convergence to the entropy solution can be proven [15], [16]. However, these methods are implicit in time, even in semidiscrete formulation (i.e., discretize spatial variable only). To advance in time, some iterative method is usually needed, and the question of convergence can be very subtle (see, e.g., [1]). This might be a serious drawback in practical computations, especially for the hyperbolic problem (1.1) which, unlike parabolic problems, seems more natural for explicit methods. Another successful finite element method for solving (1.1) is the characteristic Galerkin method, e.g. [19] and the references listed there. By following the characteristic directions in the evolution step, the severe CFL restriction can be removed. High-order accuracy can be obtained by a reconstruction procedure similar in flavor to ENO schemes using cell-average approaches [10]. This reconstruction procedure again uses a wide stencil, hence requires remedies, e.g. shock recovery, to overcome oscillations and polluting [19]. We try to obtain methods which are more local, in the sense that higher orders are achieved by more moments in a cell rather than using neighboring cells. We also used the method of lines plus a TVD ODE solver. This may render the method simpler, especially when forcing terms are present so that the characteristics are no longer straight lines.

In [18], LeSaint and Raviart first introduced the discontinuous Galerkin finite element method for solving the neutron transport equation, which is a linear version of (1.1). Because of the linearity, the solution can be obtained essentially explicitly by following the characteristic directions, hence avoiding the above-mentioned difficulty. (For an error analysis when the solution is smooth, see [13], [18].) Unfortunately, this advantage no longer exists for a nonlinear problem (1.1). In [3], Chavent and Salzano modified this method and rendered it explicit in time, by using elements which are piecewise linear in space but piecewise constant in time. However, their scheme is linearly unconditionally unstable under any fixed CFL

number [2] (it is linearly stable under the very restrictive CFL condition $\Delta t/\Delta x = O(\sqrt{\Delta x})$, which is not realistic for a hyperbolic problem (1.1)). By using a local projection limiter, based on the monotonicity-preserving projections introduced by van Leer [17], Chavent and Cockburn [2] obtained a scheme (they call it the $\Lambda\Pi P^0P^1$-scheme) which can be proven TVDM (total variation diminishing in the means) and TVB, under a fixed CFL number which can be chosen close to 1/2. Convergence of a subsequence is thus guaranteed, and numerical results given in [2] indicate convergence to the correct entropy solutions. There are still some drawbacks of the scheme. One is that it is only first-order accurate in time, although second-order in space. The second is that like any other TVD scheme, the accuracy degenerates to first order at smooth extrema of the solution [20]. The last is that since it is linearly unstable under any fixed CFL number, the projection limiting has to balance the spurious oscillations in smooth regions caused by linear instability, hence may adversely affect accuracy in these regions. These drawbacks limit the practical use of the scheme and impose serious obstacles in generalizing the method to higher orders. In [5], our first paper in this series, we retained the finite element formulation in [2] for the spatial discretizations, except that we modified the local projection limiter in order to recover accuracy at extrema, but used a special TVD second-order Runge-Kutta type discretization introduced by Shu and Osher in a finite difference framework [27]. The resulting *explicit* scheme was then proven linearly stable for CFL number $\leq \frac{1}{3}$, formally uniformly second-order accurate in space and time, including at extrema, and TVBM (total variation bounded in the means). Numerical results in [5] indicate good convergence behavior—second order in smooth regions, including at extrema, sharp shock transitions (usually in one or two elements) without oscillations, and convergence to entropy solutions even for nonconvex fluxes.

In this paper we give a unified framework to generalize the scheme in [5] to higher orders, and consider initial-boundary value problems. The advantage of such methods will be fully demonstrated only in higher dimensions. Generalizations to systems and/or higher dimensions are the subject of current research. Preliminary results are very promising.

The format of the paper is as follows. In Section 2 we present the general framework of our scheme using scalar, one-dimensional initial value problems as a model. We prove TVBM, TVB, convergence of a subsequence, and discuss entropy conditions. In Section 3 we discuss initial-boundary value problems, present boundary treatments both for inflow and for outflow boundary conditions, and prove TVBM and TVB of the resulting schemes. We present some numerical results in Section 4, and give a summary in Section 5.

2. General Formulation. In this section we take $d = m = 1$ in (1.1) and consider the problem of finding an approximate solution of (1.1) with a periodic or compact supported initial condition of bounded variation,

$$u(x,0) = u^0(x), \qquad x \in \mathbf{R}. \tag{2.1}$$

We shall first discretize in the spatial variable x. In finite element methods the approximate solution u_h is chosen to belong to a finite-dimensional space U_h. Its degrees of freedom are then obtained by solving a weak formulation of (1.1) that is

usually obtained by multiplying (1.1) by a test function $v_h \in V_h$, integrating over suitable domains, and finally integrating by parts. If the space of trial functions U_h and the one of test functions V_h coincide, the method is called a Galerkin method, otherwise it is called a Petrov-Galerkin method.

Let $I_j = (x_{j-1/2}, x_{j+1/2})$, $I = \bigcup_j I_j$ be a partition of the real line. Denote $\Delta x_j = x_{j+1/2} - x_{j-1/2}$ and $h = \sup_j \Delta x_j$. The finite element method we are going to use is a Galerkin method for which

$$U_h = V_h = V_h^k = \{p \in BV \cap L^1 : p|_{I_j} \in P^k(I_j)\}, \tag{2.2}$$

where $P^k(I_j)$ is the space of polynomials of degree $\leq k$ on I_j. Note that in V_h^k, the functions are allowed to have jumps at the interfaces $x_{j+1/2}$ (this is why this method is called a discontinuous Galerkin method [18], [3], [13], [2]), hence $V_h^k \not\subset H^1$. This is one of the main differences between the discontinuous Galerkin method and the streamline diffusion method or most other finite element methods. It is exactly this difference which makes the discontinuous Galerkin method explicit in semidiscrete ODE form, hence amenable to explicit time discretizations.

One way to implement (2.2) is to use a local orthogonal basis over I_j, $\{v_l^{(j)}(x), l = 0, 1, 2, \ldots, k\}$, such that $v_l^{(j)}(x)$ is supported in I_j and $\int_{I_j} v_l^{(j)}(x) v_s^{(j)}(x)\, dx = c_l \delta_{ls}$ with $c_l \neq 0$. Notice that, up to constant coefficients, $v_l^{(j)}(x)$ are simply the Legendre polynomials over I_j. We choose, for example,

$$v_0^{(j)}(x) = 1, \quad v_1^{(j)}(x) = x - x_j, \quad v_2^{(j)}(x) = (x - x_j)^2 - \tfrac{1}{12}\Delta x_j^2, \ldots . \tag{2.3}$$

For these $v_l^{(j)}(x)$, and $x \in I_j$, we have

$$\begin{aligned} &\frac{d^r}{dx^r} v_l^{(j)}(x) = O(\Delta x_j^{l-r}) \quad \text{for } r = 0, 1, \ldots, l, \\ &\int_{I_j} (v_l^{(j)}(x))^2\, dx = O(\Delta x_j^{2l+1}), \qquad l = 0, 1, 2, \ldots, k. \end{aligned} \tag{2.4}$$

If we define the degrees of freedom as

$$u_j^{(l)} = u_j^{(l)}(t) = \frac{1}{\Delta x_j^{l+1}} \int_{I_j} u(x,t) v_l^{(j)}(x)\, dx, \qquad l = 0, 1, \ldots, k \tag{2.5}$$

(note that $u_j^{(0)} = \bar{u}_j$ is the cell average of u in I_j), then $u_j^{(l)} = O(\Delta x_j^l)$, and

$$u^h(x,t) = \sum_{l=0}^{k} a_l u_j^{(l)}(t) v_l^{(j)}(x) \quad \text{for } x \in I_j, \tag{2.6a}$$

$$a_l = \frac{\Delta x_j^{l+1}}{\int_{I_j} (v_l^{(j)}(x))^2\, dx}, \quad \text{i.e., } a_0 = 1,\ a_1 = \frac{12}{\Delta x_j},\ a_2 = \frac{180}{\Delta x_j^2}, \ldots, \tag{2.6b}$$

where $u^h(x,t)$ is the approximation of the solution $u(x,t)$ in U_h.

In order to determine the degrees of freedom of u^h, we multiply (1.1) by $v^h \in V_h^k$, integrate over I_j and replace the exact solution u by its approximation u^h:

$$\int_{I_j} \left(\frac{d}{dt} u^h(x,t)\right) v^h(x)\, dx + \int_{I_j} \left(\frac{d}{dx} f(u^h(x,t))\right) v^h(x)\, dx = 0 \quad \forall v^h \in V_h^k. \tag{2.7}$$

Denoting $u_{j+1/2} = u^h(x_{j+1/2}, t)$, and formally integrating (2.7) by parts, we arrive at an ODE which the degrees of freedom $u_j^{(l)} = u_j^{(l)}(t)$ must satisfy:

$$\begin{aligned}\frac{d}{dt}u_j^{(l)} &+ \frac{1}{\Delta x_j^{l+1}}[\Delta_+(v_l^{(j)}(x_{j-1/2})f(u_{j-1/2}))] \\ &- \frac{1}{\Delta x_j^{l+1}}\int_{I_j} f(u^h(x,t))\frac{d}{dx}v_l^{(j)}(x)\,dx = 0, \qquad l = 0, 1, \ldots, k,\end{aligned} \tag{2.8}$$

where here and in the following, $\Delta_\pm$ are the usual difference operators,

$$\Delta_+ w_j = w_{j+1} - w_j, \qquad \Delta_- w_j = w_j - w_{j-1}.$$

Note the ambiguity in (2.8): $f(u_{j+1/2}) = f(u^h(x_{j+1/2}, t))$ is not defined at the cell interfaces $x_{j+1/2}$, hence we still have quite some freedom in (2.8). This freedom comes from the fact that we have more parameters in the piecewise polynomial solution than equations to define them. It is this freedom which gives us a chance to adopt the successful finite difference nonoscillatory methodology. If we write out the first equation in (2.8),

$$\frac{d}{dt}u_j^{(0)} + \frac{1}{\Delta x_j}[f(u_{j+1/2}) - f(u_{j-1/2})] = 0, \tag{2.9}$$

we can see that it is very similar to a conservative finite difference scheme for the means $u_j^{(0)}$. This suggests that we use $h_{j+1/2} = h(u_{j+1/2}^-, u_{j+1/2}^+)$ in place of $f(u_{j+1/2})$, where $u_{j+1/2}^\pm = u^h(x_{j+1/2}^\pm, t)$, and $h(\cdot,\cdot)$ is some "average" flux. The only requirement for accuracy is consistency, $h(u,u) = f(u)$, plus Lipschitz continuity of $h(\cdot,\cdot)$. For total variation stability we need more restrictions on $h(\cdot,\cdot)$, and also some modifications on $u_{j+1/2}^\pm$. We will use a Lipschitz continuous "monotone" flux as defined in [6], i.e., $h(\cdot,\cdot)$ is nondecreasing in its first argument and nonincreasing in its second argument. See Remark 2.4 below. Scheme (2.8) now reads

$$\begin{aligned}\frac{d}{dt}u_j^{(l)} &+ \frac{1}{\Delta x_j^{l+1}}[\Delta_+(v_l^{(j)}(x_{j-1/2})h_{j-1/2})] \\ &- \frac{1}{\Delta x_j^{l+1}}\int_{I_j} f(u^h(x,t))\frac{d}{dx}v_l^{(j)}(x)\,dx = 0, \qquad l = 0, 1, \ldots, k,\end{aligned} \tag{2.10}$$

i.e.,

$$\frac{d}{dt}u_j^{(0)} + \frac{1}{\Delta x_j}(h_{j+1/2} - h_{j-1/2}) = 0, \tag{2.10a}$$

$$\frac{d}{dt}u_j^{(1)} + \frac{1}{2\Delta x_j}(h_{j+1/2} + h_{j-1/2}) - \frac{1}{\Delta x_j^2}\int_{I_j} f(u^h(x,t))\,dx = 0, \tag{2.10b}$$

$$\begin{aligned}\frac{d}{dt}u_j^{(2)} &+ \frac{1}{6\Delta x_j}(h_{j+1/2} - h_{j-1/2}) \\ &- \frac{2}{\Delta x_j^3}\int_{I_j} f(u^h(x,t))v_1^{(j)}(x)\,dx = 0,\end{aligned} \tag{2.10c}$$

etc.

We recall that $h_{j+1/2} = h(u^-_{j+1/2}, u^+_{j+1/2})$, where $u^\pm_{j+1/2} = u^h(x^\pm_{j+1/2}, t)$ are defined by (2.6), i.e.,

(2.11a) First order ($k = 0$): $u^-_{j+1/2} = u^+_{j-1/2} = u^{(0)}_j$;

(2.11b) Second order ($k = 1$): $u^-_{j+1/2} = u^{(0)}_j + 6u^{(1)}_j, u^+_{j-1/2} = u^{(0)}_j - 6u^{(1)}_j$;

(2.11c) Third order ($k = 2$): $u^-_{j+1/2} = u^{(0)}_j + 6u^{(1)}_j + 30u^{(2)}_j$,
$u^+_{j-1/2} = u^{(0)}_j - 6u^{(1)}_j + 30u^{(2)}_j$;

etc.

The integrations in (2.10) can be approximated by a suitable quadrature rule, see Remark 2.6 below. For linear problems $f(u) = au$ this is unnecessary, because the integrations can be performed exactly.

We still need to modify $u^\pm_{j+1/2}$ in (2.11) by some local projection limiter to achieve total variation stability. To this end, we write

$$u^-_{j+1/2} = u^{(0)}_j + \tilde{u}_j, \qquad u^+_{j-1/2} = u^{(0)}_j - \tilde{\tilde{u}}_j. \tag{2.12}$$

Comparing with (2.6), we see

$$\tilde{u}_j = \sum_{l=1}^{k} a_l u^{(l)}_j v^{(j)}_l(x_{j+1/2}), \qquad \tilde{\tilde{u}}_j = -\sum_{l=1}^{k} a_l u^{(l)}_j v^{(j)}_l(x_{j-1/2}), \tag{2.13}$$

i.e.,

(2.13a) First order ($k = 0$): $\tilde{u}_j = \tilde{\tilde{u}}_j = 0$;

(2.13b) Second order ($k = 1$): $\tilde{u}_j = \tilde{\tilde{u}}_j = 6u^{(1)}_j$;

(2.13c) Third order ($k = 2$): $\tilde{u}_j = 6u^{(1)}_j + 30u^{(2)}_j$, $\tilde{\tilde{u}}_j = 6u^{(1)}_j - 30u^{(2)}_j$;

etc.

Then (2.10a), the scheme for the means $u^{(0)}_j$, becomes

$$\frac{d}{dt}u^{(0)}_j = -\frac{1}{\Delta x_j}\Delta_- h(u^{(0)}_j + \tilde{u}_j, u^{(0)}_{j+1} - \tilde{\tilde{u}}_{j+1}). \tag{2.14}$$

We also consider the Euler forward version of (2.14),

$$(u^{(0)}_j)^{n+1} = (u^{(0)}_j)^n - \frac{\Delta t}{\Delta x_j}\Delta_- h((u^{(0)}_j + \tilde{u}_j)^n, (u^{(0)}_{j+1} - \tilde{\tilde{u}}_{j+1})^n). \tag{2.15}$$

Note that (2.14) resembles a MUSCL type finite difference scheme considered by Osher in [20].

First we have the following lemma.

LEMMA 2.1. *(TVDM) Assume*

$$-\theta \le \frac{\Delta_+ \tilde{\tilde{u}}_j}{\Delta_+ u^{(0)}_j} \le 1, \qquad -\theta \le -\frac{\Delta_+ \tilde{u}_j}{\Delta_+ u^{(0)}_j} \le 1. \tag{2.16}$$

Then,

(i) *Scheme* (2.14)) *is TVD for any* $\theta \ge 0$;

(ii) *Scheme* (2.15) *is TVD under the CFL restriction*

$$\lambda(h_1 - h_2) \le \frac{1}{1+\theta}, \tag{2.17}$$

where h_1 and $-h_2$ are the Lipschitz coefficients of $h(\cdot,\cdot)$ with respect to the first and second arguments, respectively. Note that in general $h_1 - h_2 \approx \max|f'(u)|$. Also $\lambda = \Delta t/h$.

Proof. (i) We follow [20, Lemma 2.4] and write

$$-\Delta_- h(u_j^{(0)} + \tilde{u}_j, u_{j+1}^{(0)} - \tilde{\tilde{u}}_{j+1}) = C_{j+1/2}\Delta_+ u_j^{(0)} - D_{j-1/2}\Delta_- u_j^{(0)}, \tag{2.18}$$

where

$$C_{j+1/2} = -h_2 \cdot \left(1 - \frac{\Delta_+ \tilde{\tilde{u}}_j}{\Delta_+ u_j^{(0)}}\right), \qquad D_{j-1/2} = h_1 \left(1 + \frac{\Delta_- \tilde{u}_j}{\Delta_- u_j^{(0)}}\right) \tag{2.18a}$$

with

$$h_1 = \frac{h(u_j^{(0)} + \tilde{u}_j, u_j^{(0)} - \tilde{\tilde{u}}_j) - h(u_{j-1}^{(0)} + \tilde{u}_{j-1}, u_j^{(0)} - \tilde{\tilde{u}}_j)}{(u_j^{(0)} + \tilde{u}_j) - (u_{j-1}^{(0)} + \tilde{u}_{j-1})}, \tag{2.18b}$$

$$-h_2 = -\frac{h(u_j^{(0)} + \tilde{u}_j, u_{j+1}^{(0)} - \tilde{\tilde{u}}_{j+1}) - h(u_j^{(0)} + \tilde{u}_j, u_j^{(0)} - \tilde{\tilde{u}}_j)}{(u_{j+1}^{(0)} - \tilde{\tilde{u}}_{j+1}) - (u_j^{(0)} - \tilde{\tilde{u}}_j)} \tag{2.18c}$$

being the (local) Lipschitz coefficients of $h(\cdot,\cdot)$ with respect to the two arguments. Now monotonicity of $h(\cdot,\cdot)$ and (2.16) guarantee

$$C_{j+1/2}, D_{j-1/2} \geq 0. \tag{2.19}$$

The rest of the proof can be found in [20].

(ii) For (2.15) we need (2.19) plus

$$\lambda(C_{j+1/2} + D_{j+1/2}) \leq 1 \tag{2.20}$$

for TVD (see [7]). Clearly, (2.16), (2.17) and (2.18) imply (2.20). □

Since (2.14) and (2.15) are schemes for the means $u_j^{(0)}$, we say the scheme is TVDM (total variation diminishing in the means).

We now try to modify $\tilde{u}_j$ and $\tilde{\tilde{u}}_j$ to achieve (2.16). Observe that the middle quantities in (2.16), in smooth regions, are $O(h)$, so the limiting required is very mild. One simple way to accomplish this is to define, for $\theta = 1$ in (2.16),

$$\tilde{u}_j^{(\mathrm{mod})} = m(\tilde{u}_j, \Delta_+ u_j^{(0)}, \Delta_- u_j^{(0)}), \qquad \tilde{\tilde{u}}_j^{(\mathrm{mod})} = m(\tilde{\tilde{u}}_j, \Delta_+ u_j^{(0)}, \Delta_- u_j^{(0)}), \tag{2.21}$$

where m is the usual minmod function [7],

$$m(a_1, a_2, \ldots, a_n) = \begin{cases} s \cdot \min_{1\leq i\leq n} |a_i| & \text{if } \operatorname{sign}(a_1) = \cdots = \operatorname{sign}(a_n) = s, \\ 0 & \text{otherwise.} \end{cases} \tag{2.22}$$

We then define $u_{j+1/2}^{\pm(\mathrm{mod})}$ as in (2.12),

$$u_{j+1/2}^{-(\mathrm{mod})} = u_j^{(0)} + \tilde{u}_j^{(\mathrm{mod})}, \qquad u_{j-1/2}^{+(\mathrm{mod})} = u_j^{(0)} - \tilde{\tilde{u}}_j^{(\mathrm{mod})}, \tag{2.23}$$

and define scheme (2.10) by using $h_{j+1/2} = h(u_{j+1/2}^{-(\mathrm{mod})}, u_{j+1/2}^{+(\mathrm{mod})})$.

Note that Taylor expansions reveal

$$\Delta_\pm u_j^{(0)} = (u_x)_j \frac{\Delta x_j + \Delta x_{j\pm 1}}{2} + O(h^2),$$

while $\tilde{u}_j$, $\tilde{\tilde{u}}_j = \frac{1}{2}(u_x)_j \Delta x_j + O(h^2)$. So in smooth regions away from critical points, $\tilde{u}_j^{(\mathrm{mod})} = \tilde{u}_j, \tilde{\tilde{u}}_j^{(\mathrm{mod})} = \tilde{\tilde{u}}_j$, by the definitions (2.21) and (2.22), hence the order of accuracy of the scheme is not affected by the limiting in these regions. Unfortunately, like in any other TVD scheme, accuracy degenerates to first order at critical points [21]. To overcome this difficulty, we apply the idea of Shu [24] and change the minmod function (2.22) to

$$\tilde{m}(a_1, a_2, \ldots, a_n) = \begin{cases} a_1 & \text{if } |a_1| \le Mh^2, \\ m(a_1, a_2, \ldots, a_n) & \text{otherwise,} \end{cases} \tag{2.24}$$

where $M > 0$ is a constant. For the choice of M we have the following lemma.

LEMMA 2.2. *If*

$$M = \frac{2}{3} M_2, \tag{2.25a}$$

or

$$M = M_j = \frac{2}{9}(3 + 10M_2) \cdot M_2 \cdot \frac{h^2}{h^2 + |\Delta_+ u_j^{(0)}| + |\Delta_- u_j^{(0)}|}, \tag{2.25b}$$

then the limiting (2.21)–(2.24) *does not affect accuracy in any region where* $u \in C^2$ *and* $|u_{xx}| \le M_2$.

Proof. First we observe that accuracy is guaranteed if the limiter (2.21) returns the first argument. For simplicity we take $\Delta x_j \equiv h$. By Taylor expansion,

$$\Delta_+ u_j^{(0)} = u_x(x_j)h + \tfrac{1}{2}u_{xx}(\xi_1)h^2, \qquad \Delta_- u_j^{(0)} = u_x(x_j)h - \tfrac{1}{2}u_{xx}(\xi_2)h^2,$$
$$\tilde{u}_j = \tfrac{1}{2}u_x(x_j)h + \tfrac{1}{12}u_{xx}(\xi_3)h^2, \qquad \tilde{\tilde{u}}_j = \tfrac{1}{2}u_x(x_j)h - \tfrac{1}{12}u_{xx}(\xi_4)h^2.$$

Hence, in any region where $u \in C^2$, $|u_{xx}| \le M_2$, we have either

(i) $|u_x(x_j)| > \frac{7}{6}M_2 h$ in which case

$$|\Delta_\pm u_j^{(0)}| \ge |u_x(x_j)h| - \tfrac{1}{2}M_2h^2 \ge \tfrac{1}{2}|u_x(x_j)h| + \tfrac{7}{12}M_2h^2 - \tfrac{1}{2}M_2h^2$$
$$= \tfrac{1}{2}|u_x(x_j)h| + \tfrac{1}{12}M_2h^2 \ge |\tilde{u}_j|, |\tilde{\tilde{u}}_j|,$$

and (2.21) returns the first argument; or

(ii) $|u_x(x_j)| \le \frac{7}{6}M_2 h$, in which case

$$|\tilde{u}_j|, |\tilde{\tilde{u}}_j| \le \tfrac{1}{2} \cdot \tfrac{7}{6}M_2h^2 + \tfrac{1}{12}M_2h^2 = \tfrac{2}{3}M_2h^2,$$

and

$$\frac{h^2}{h^2 + |\Delta_+ u_j^{(0)}| + |\Delta_- u_j^{(0)}|} \ge \frac{h^2}{h^2 + 2 \cdot (\frac{7}{6} + \frac{1}{2})M_2h^2} = \frac{1}{1 + \frac{10}{3}M_2},$$

hence by (2.24)–(2.25), (2.21) again returns the first argument. □

The expression (2.25b) is better than (2.25a) in that it is very small except near critical points, hence the correction from (2.22) is minimal but enough to recover accuracy. In practice, we should choose $M_2 = \max_J |u_{xx}^0|$, where J is some neighborhood of smooth critical points of $u^0(x)$ in (2.1).

For the correction (2.24) we have the following lemma.

LEMMA 2.3. (*TVBM*) *The conclusions of Lemma* 2.1 *hold, with* m *in* (2.22) *replaced by* $\tilde{m}$ *in* (2.24), *and TVD replaced by TVB.*

Proof. Follow the lines of proof in [24, Theorem 2.2]. □

Since (2.21) is only a modification on some combinations of $u_j^{(l)}$, $l \geq 1$, see (2.13), we still need to define $u_j^{(l)(\mathrm{mod})}$, $l \geq 1$, for example, by

(2.26a) Second order ($k = 1$): $u_j^{(1)(\mathrm{mod})} = \frac{1}{6}\tilde{u}_j^{(\mathrm{mod})} = \frac{1}{6}\tilde{\tilde{u}}_j^{(\mathrm{mod})}$;

(2.26b) Third order ($k = 2$): $u_j^{(1)(\mathrm{mod})} = \frac{1}{12}(\tilde{u}_j^{(\mathrm{mod})} + \tilde{\tilde{u}}_j^{(\mathrm{mod})})$,
$u_j^{(2)(\mathrm{mod})} = \frac{1}{60}(\tilde{u}_j^{(\mathrm{mod})} - \tilde{\tilde{u}}_j^{(\mathrm{mod})})$.

Beginning at fourth order ($k \geq 3$), the $u_j^{(l)(\mathrm{mod})}$ are no longer uniquely determined by $\tilde{u}_j^{(\mathrm{mod})}$ and $\tilde{\tilde{u}}_j^{(\mathrm{mod})}$. We may then use, e.g., first $u_j^{(l)(\mathrm{mod})} = u_j^{(l)}$ for all $l \geq 3$, then determine $u_j^{(1)(\mathrm{mod})}$ and $u_j^{(2)(\mathrm{mod})}$ by $\tilde{u}_j^{(\mathrm{mod})}$, $\tilde{\tilde{u}}_j^{(\mathrm{mod})}$ and (2.13), and then, if either $u_j^{(1)(\mathrm{mod})} \neq u_j^{(1)}$ or $u_j^{(2)(\mathrm{mod})} \neq u_j^{(2)}$, reset $u_j^{(l)(\mathrm{mod})} = 0$ for $l \geq 3$, and redetermine $u_j^{(1)(\mathrm{mod})}$ and $u_j^{(2)(\mathrm{mod})}$ again by (2.26b). There are, of course, other ways to accomplish this. We must have

$$|u_j^{(l)}| \leq c_1 \min\{|\Delta_+ u_j^{(0)}|, |\Delta_- u_j^{(0)}|\} + c_2 h^2, \tag{2.27}$$

as is true for all the cases discussed above, because then the good properties of the means $u_j^{(0)}$ (e.g. TVBM) can be passed to the whole solution $u^h(x,t)$ (see Proposition 2.11 below).

Remark 2.4. In the above discussion we may use any two-point Lipschitz continuous monotone flux $h(\cdot,\cdot)$. Some possible choices are

(i) Engquist-Osher:

$$h^{\mathrm{EO}}(a,b) = \int_0^b \min(f'(s),0)\,ds + \int_0^a \max(f'(s),0)\,ds + f(0); \tag{2.28a}$$

(ii) Godunov:

$$h^{\mathrm{G}}(a,b) = \begin{cases} \min_{a\leq u\leq b} f(u) & \text{if } a \leq b, \\ \max_{a\geq u\geq b} f(u) & \text{if } a > b; \end{cases} \tag{2.28b}$$

(iii) Lax-Friedrichs:

$$h^{\mathrm{LF}}(a,b) = \tfrac{1}{2}[f(a) + f(b) - \alpha(b-a)], \qquad \alpha = \max|f'(u)|, \tag{2.28c}$$

where the maximum is taken over the whole region in which a, b varies, e.g., in $[\inf u^0(x), \sup u^0(x)]$, where $u^0(x)$ is the initial function;

(iv) Local Lax-Friedrichs:

$$h^{\mathrm{LLF}}(a,b) = \tfrac{1}{2}[f(a)+f(b)-\beta(b-a)], \qquad \beta = \max_{\min(a,b)\leq u\leq\max(a,b)} |f'(u)|. \tag{2.28d}$$

For convex $f, f'' \geq 0$, one has $\beta = \max(|f'(a)|, |f'(b)|)$;

(v) Roe with entropy fix (this is an E-flux [21], which is a generalization of monotone fluxes):

$$h^{\mathrm{RF}}(a,b) = \begin{cases} f(a) & \text{if } f'(u) \geq 0 \text{ for } u \in [\min(a,b), \max(a,b)], \\ f(b) & \text{if } f'(u) \leq 0 \text{ for } u \in [\min(a,b), \max(a,b)], \\ h^{\mathrm{LLF}}(a,b) & \text{otherwise.} \end{cases} \tag{2.28e}$$

We can use h^{G} if f is not too complicated. However, for convex f we suggest using h^{LLF} or h^{RF}, because of their simplicity and good numerical results (based on our experience in finite difference computations. See also Section 4). □

Remark 2.5. Note that even if there are $k+1$ degrees of freedom $u_j^{(l)}$ in (2.10) to compute, for $l \geq 1$ we only need to evaluate one integration (by quadrature) to obtain each $u_j^{(l)}$ for nonlinear f, and none for linear f, because the $h_{j\pm 1/2}$ are available from the computation of $u_j^{(0)}$ in (2.10a). In terms of cost, this compares favorably, for multi-dimensional problems, with finite difference methods using cell average formulations (e.g., ENO schemes in [10],[8], in which point values are reconstructed from cell averages). □

Remark 2.6. To compute the integrations in (2.10) for nonlinear f, quadratures of correct order must be used so that the error of approximating

$$\frac{1}{\Delta x_j^{l+1}} \int_{I_j} f(u^h(x,t)) \frac{d}{dx} v_l^{(j)}(x)\, dx$$

is $O(h^{k+1})$. Considering (2.4), we need a quadrature whose error is $O(h^{k+l+2})$, or $O(h^{2k+2})$, to work for every l. For example, Simpson's rule (the error is $O(h^5)$) is enough for $k = 1$ (second-order scheme), and three-point Gaussian (the error is $O(h^6)$) is needed for $k = 2$ (third-order scheme). Since we can use $h_{j\pm 1/2}$ as $f(u^h(x_{j\pm 1/2}, t))$ in the quadratures to save function evaluations, Gauss-Lobatto quadratures, which use end points of the interval as nodes, are preferred. □

Remark 2.7. We choose degrees of freedom as in (2.5) only for easy presentation. The essential ingredients are the weak formulation (2.7), the choice of spaces (2.2), and the local projection limiting (2.21). Other degrees of freedom can of course be chosen for various purposes, e.g., for some physical considerations, or to save cost. One possibility is to choose point values of $u^h(x,t)$ at some Gaussian points as degrees of freedom, to save the cost of evaluating (2.6a) when quadrature is used to approximate the integration. □

We now turn our attention to entropy conditions. First we have the following lemma.

LEMMA 2.8. *If $\hat{f}_{j+1/2} = h_{j+1/2} + c_{j+1/2}$, where $h_{j+1/2}$ is a monotone flux and $|c_{j+1/2}| \leq ch^\alpha$ for some fixed $c, \alpha > 0$, then a TVB scheme*

$$u_j^{n+1} = u_j^n - \lambda(\hat{f}_{j+1/2} - \hat{f}_{j-1/2})$$

is an entropy scheme, i.e., it always converges to entropy solutions. Moreover, for every bounded domain Ω,

$$\|u - u_h\|_{L^\infty(0,T;(L^1(\Omega))} \leq C(\Omega) h^{\min\{1,\alpha\}/2}.$$

Proof. Similar to the proof of [4, I, Theorem 3.2], hence omitted. □

By Lemma 2.8, if we can tolerate an *explicit* Ch^α appearance in the projection limiter (2.21),

$$\begin{aligned} \tilde{u}_j^{(e,\mathrm{mod})} &= m(\tilde{u}_j, \Delta_+ u_j^{(0)}, \Delta_- u_j^{(0)}, \mathrm{sign}(\tilde{u}_j)(Ch^\alpha)), \\ \tilde{\tilde{u}}_j^{(e,\mathrm{mod})} &= m(\tilde{\tilde{u}}_j, \Delta_+ u_j^{(0)}, \Delta_- u_j^{(0)}, \mathrm{sign}(\tilde{\tilde{u}}_j)(Ch^\alpha)), \end{aligned} \tag{2.29}$$

then schemes (2.14) and (2.15) are both entropy schemes.

Explicitly inserting Ch^α into the limiter to enforce entropy conditions is not an intrinsic procedure: the parameters C and α cannot be automatically adjusted by the scheme using the numerical solution, but must be tuned for each individual problem, and forcing convergence to entropy solutions is heavily dependent on the choice of C and α. If h is not in the asymptotic regime, inappropriate choices of C and α will either smear the profile too much, or fail to correct a weak entropy violating shock. We note, however, that for $\alpha \le 1$, the accuracy in smooth regions is not affected by the Ch^α term in any region where $u \in C^1$ and $|u_x| \le 2C$.

If f in (1.1) is convex, then we can avoid the explicit Ch^α term by using the following lemma in [20, Theorem 3.1].

LEMMA 2.9 (OSHER). *If condition* (2.16) *is strengthened to*

$$\frac{\tilde{u}_j}{\Delta_+ u_j^{(0)}}, \qquad \frac{\tilde{\tilde{u}}_{j+1}}{\Delta_+ u_j^{(0)}} \le \frac{1}{2}, \tag{2.30a}$$

and, when $u_j^{(0)} > u_{j+1}^{(0)}$, *denoting*

$$\hat{u}_{j+1/2} = \frac{\int_{u_j^{(0)}}^{u_{j+1}^{(0)}} w f'(w)\, dw}{f(u_{j+1}^{(0)}) - f(u_j^{(0)})},$$

there holds

$$\begin{cases} \tilde{u}_j \ge \hat{u}_{j+1/2} - u_j^{(0)}, \tilde{\tilde{u}}_j \ge u_{j+1}^{(0)} - \hat{u}_{j+1/2} & \text{if } u_j^{(0)} > \hat{u}_{j+1/2} > u_{j+1}^{(0)}, \\ \tilde{u}_j = \tilde{\tilde{u}}_{j+1} = 0 & \text{otherwise,} \end{cases} \tag{2.30b}$$

then scheme (2.14) *satisfies one entropy condition with the particular entropy* $V(u) = u^2/2$ *if* f *is convex. Hence, for convex* f, (2.14) *is an entropy scheme.* □

Unlike (2.29), the limiting in (2.30) does affect accuracy in smooth regions. Moreover, Lemma 2.9 for the Euler forward version (2.15) is not easy to prove. See [22] for a proof of the second-order case.

In our preliminary computations, which include some nonconvex f, we simply use the limiter (2.21)–(2.24) without any entropy corrections and always observed convergence to the correct entropy solution. See [5] and Section 4 for details.

Next we turn to time discretizations of (2.10). This is the easy part because scheme (2.10), unlike most other finite element schemes, is explicit, hence readily amenable to the TVD time discretization techniques for the ODE method of lines introduced in [26], [27].

We first rewrite (2.10) in a concise ODE form,

$$\frac{d}{dt} u^h = L_h(u^h, t), \tag{2.31}$$

and apply, e.g., the high-order TVD Runge-Kutta methods in [27]:

$$(u^h)^{(i)} = \sum_{l=0}^{i-1} [\alpha_{il}(u^h)^{(l)} + \beta_{il}\Delta t L_h((u^h)^{(l)}, t^n + d_l \Delta t)], \qquad i = 1, \ldots, r, \tag{2.32a}$$

$$(u^h)^{(0)} = (u^h)^n, \qquad (u^h)^{(r)} = (u^h)^{n+1}. \tag{2.32b}$$

Here, r is related to the order of the scheme. For $r \leq 4$, (2.32) can be made rth-order accurate. We have included the time variable in L_h in case we have a time-dependent forcing term or time-dependent boundary conditions, as is the case in Section 3.

Scheme (2.32) is TVD (or TVB) under the CFL restriction

$$\lambda \leq \lambda_r = \lambda_0 \min_{i,l} \frac{\alpha_{il}}{|\beta_{il}|} \tag{2.33}$$

if the Euler forward version of (2.31) is TVD (or TVB) for $\lambda \leq \lambda_0$, provided $\alpha_{il} \geq 0$, and when $\beta_{il} < 0$, L_h is replaced by $\tilde{L}_h$ where $-\tilde{L}_h$ is a TVD (or TVB) operator approximating $(-f(u))_x$ in $u_t - (f(u))_x = 0$. For details, see [26], [27].

Up to fourth order, the coefficients $\alpha_{il}, \beta_{il}, d_l$ are rather simple. For example,

Second order ($r = 2$):

$$\begin{aligned} &\alpha_{10} = \beta_{10} = 1, \quad \alpha_{20} = \alpha_{21} = \beta_{21} = \frac{1}{2}, \quad \beta_{20} = 0; \\ &d_0 = 0, \quad d_1 = 1; \quad \text{CFL: } \lambda_2 = 1; \end{aligned} \tag{2.34a}$$

Third order ($r = 3$):

$$\begin{aligned} &\alpha_{10} = \beta_{10} = 1, \quad \alpha_{20} = \frac{3}{4}, \quad \beta_{20} = 0, \quad \alpha_{21} = \beta_{21} = \frac{1}{4}, \\ &\alpha_{30} = \frac{1}{3}, \quad \beta_{30} = \alpha_{31} = \beta_{31} = 0, \quad \alpha_{32} = \beta_{32} = \frac{2}{3}; \\ &d_0 = 0, \quad d_1 = 1, \quad d_2 = \frac{1}{2}; \quad \text{CFL: } \lambda_3 = 1; \end{aligned} \tag{2.34b}$$

etc.

Finally, we turn our attention to the linear stability of the scheme (2.32), (2.10) without using the projection limiter (2.21). In [2] it was proven that the Euler forward version of (2.10) is linearly unconditionally unstable for any fixed CFL number. (The proof is for the second-order case, but it goes through for higher orders.) However, linear stability becomes much better with (2.32), (2.10):

LEMMA 2.10. *Scheme* (2.32), (2.10) *is linearly stable under the CFL condition*

$$\lambda \max |f'(u)| \leq \frac{1}{2k+1} \tag{2.35}$$

for $k = 1, 2$ (*second-order and third-order schemes*).

Proof. See [5], [2] for the second-order case. The proof for the third-order case is similar but is more technical, and is thus omitted. □

We do not know whether (2.35) is the correct CFL condition for linear stability when $k \geq 3$. However, we believe there exist fixed CFL numbers for $k \geq 3$ for which (2.32), (2.10) is linearly stable.

We summarize the results of this section in the following proposition, in which "order of accuracy" is in the sense of local truncation errors.

PROPOSITION 2.11. (i) *Scheme* (2.32), (2.10) *is linearly stable under the CFL condition* (2.35) *for* $k = 1, 2$.

(ii) *Scheme* (2.32), (2.10), *with local projection limiting* (2.21)–(2.22), *is* $(k+1)$*th-order accurate, except near critical points, is TVDM and TVB under the CFL condition* (2.17) *with* $\theta = 1$, *and thus has a convergent subsequence.*

(iii) *Scheme* (2.32), (2.10), *with local projection limiting* (2.21)–(2.24)–(2.25), *is uniformly* $(k+1)$*th-order in any region where* $u \in \mathscr{C}^2$, $|u_{xx}| \le M_2$, *is TVBM and TVB under* (2.17) *with* $\theta = 1$, *and thus has a convergent subsequence.*

(iv) *Scheme* (2.31), (2.10), *with local projection limiting* (2.21)–(2.24)–(2.30), *is TVBM and TVB, satisfies one entropy condition for the particular entropy* $V(u) = u^2/2$ *for convex* f, *hence is convergent to the entropy solution when* f *is convex.*

(v) *Scheme* (2.32), (2.10), *with local projection limiting* (2.29)–(2.24)–(2.25), *is uniformly* $(k+1)$*th-order in any region where* $u \in \mathscr{C}^2, |u_x| \le 2C, |u_{xx}| \le M_2$, *is TVBM and TVB under* (2.17) *with* $\theta = 1$, *satisfies all entropy conditions, hence is convergent to the entropy solution.*

Proof. Using the lemmas in this section. To prove TVB from TVBM or TVDM, or to pass entropy conditions from the means $u^{(0)}$ to u^h, note (2.27) and (2.6). The accuracy in terms of local truncation errors is an exercise in Taylor expansion. □

In practice, we suggest using the schemes in (iii) above.

3. Initial-Boundary Value Problems. In this section we again take $d = m = 1$ in (1.1) and consider initial-boundary value problems. We solve (1.1) in $a < x < b$ subject to the initial condition (2.1) and suitable boundary conditions at $x = a$ and $x = b$.

For simplicity, we take $b = \infty$ and consider one boundary at $x = a$ only. The case of two boundaries can be treated similarly. It is well known that on the differential equation level, we should prescribe

$$u(a,t) = g(t) \tag{3.1}$$

if $f'(u(a,t)) > 0$, and prescribe nothing if $f'(u(a,t)) < 0$. We assume $g(t)$ in (3.1) has bounded variation.

We put the boundary at $x_{-1/2} = a$. Because of the "local" property of our schemes (2.32), (2.10) (i.e., higher orders are achieved through more moments within the cell rather than through using many cells), the boundary treatment here is rather simple: we only need to consider prescribing $u^-_{-1/2}$, if necessary, and modifying the local projection limiter (2.21) on the boundary cell $j = 0$. Comparing with [25], where a similar boundary treatment was done for high-order wide stencil finite difference TVB schemes, using extrapolation and upwinding, we can see the advantage of finite element methods in dealing with boundaries.

Notice that if a pure upwind monotone flux $h(\cdot,\cdot)$, for example (2.28a, b, e), is used in (2.10), then when $f'(u(a,t)) < 0$ there is no need to assign $u^-_{-1/2}$, since $h_{-1/2} = h(u^-_{-1/2}, u^+_{-1/2}) = f(u^+_{-1/2})$. For general monotone flux, e.g. (2.28c, d), this is not the case, but $u^-_{-1/2}$ will have little effect on $h_{-1/2}$ because $-h_2 \gg h_1$, where $h_1, -h_2$, as in (2.18), are the local Lipschitz coefficients of $h(\cdot,\cdot)$ with respect to the two arguments. On the other hand, if $f'(u(a,t)) > 0$, it is natural to prescribe $u^-_{-1/2}(t) = g(t)$. We thus have the following two boundary treatments:

$$u^-_{-1/2} = u^+_{-1/2}, \quad \tilde{u}_0^{(\mathrm{mod})} = m(\tilde{u}_0, \Delta_+ u_0^{(0)}), \quad \tilde{\tilde{u}}_0^{(\mathrm{mod})} = m(\tilde{\tilde{u}}_0, \Delta_+ u_o^{(0)}) \tag{3.2a}$$

and

$$\begin{aligned} u^-_{-1/2} = u^-_{-1/2}(t) = g(t), \quad & \tilde{u}_0^{(\mathrm{mod})} = m(\tilde{u}_0, \Delta_+ u_0^{(0)}, 2(u_0^{(0)} - g(t))), \\ & \tilde{\tilde{u}}_0^{(\mathrm{mod})} = m(\tilde{\tilde{u}}_0, \Delta_+ u_0^{(0)}). \end{aligned} \tag{3.2b}$$

We note that in both cases the limiting does not affect accuracy, as is easily revealed by Taylor expansion, i.e., $\tilde{u}_0^{(\mathrm{mod})} = \tilde{u}_0, \tilde{\tilde{u}}_0^{(\mathrm{mod})} = \tilde{\tilde{u}}_0$ if the solution is smooth near the boundary. The coefficient 2 before the $u_0^{(0)} - g(t)$ term in the limiting is chosen for accuracy considerations.

Similar to Proposition 2.11, we have

PROPOSITION 3.1. *Scheme* (2.32), (2.10), *with local projection limiting* (2.21)–(2.24)–(2.25), *and boundary treatment* (3.2a) *or* (3.2b), *is TVBM and TVB, under the CFL condition* (2.17) *with* $\theta = 1$ *for* (3.2a) *and* $\theta = 2$ *for* (3.2b).

Proof. We only need to consider (2.15). The only difference between here and Lemma 2.1 is the boundary term $-\Delta_- h_{1/2}$. Instead of (2.18), we write it, for (3.2a), as

$$\begin{aligned} -\Delta_- h_{1/2} &= -\left[h(u_0^{(0)} + \tilde{u}_0, u_1^{(0)} - \tilde{\tilde{u}}_1) - h(u_0^{(0)} - \tilde{\tilde{u}}_0, u_0^{(0)} - \tilde{\tilde{u}}_0)\right] \\ &= C_{1/2}\Delta_+ u_0^{(0)}, \end{aligned} \tag{3.3a}$$

with

$$C_{1/2} = -h_2\left(1 - \frac{\tilde{\tilde{u}}_1 - (1 - \frac{h_1}{-h_2})\tilde{\tilde{u}}_0 + \frac{h_1}{-h_2}\tilde{u}_0}{\Delta_+ u_0^{(0)}}\right), \tag{3.3b}$$

where h_1 and $-h_2$ are the local Lipschitz coefficients of $h(\cdot,\cdot)$ defined in (2.18b,c).

We pause here to remark that if an upwind flux (2.28a,b,e) is used, then $h_1 = 0$, hence (3.2a) clearly implies $C_{1/2} \geq 0$ in (3.3b) and (2.20) under (2.17) with $\theta = 1$; if a general monotone flux like (2.28c,d) is used, we need to assume $h_1/(-h_2) \leq \frac{1}{3}$, which is reasonable for any upwind biased monotone flux, and then limit $\tilde{\tilde{u}}_1$ in a slightly more restrictive way:

$$\tilde{\tilde{u}}_1^{(\mathrm{mod})} = m(\tilde{\tilde{u}}_1, \Delta_+ u_1^{(0)}, \tfrac{2}{3}\Delta_- u_1^{(0)}). \tag{3.4}$$

Note that (3.4) does not affect accuracy. We then have $C_{1/2} \geq 0$, and standard arguments [7], [25] lead to TVDM or TVBM, depending on whether (2.22) or (2.25) is used, under the TV definition

$$\mathrm{TV}(u^{(0)}) = \sum_{j \geq 0} |u_{j+1}^{(0)} - u_j^{(0)}|. \tag{3.5}$$

For (3.2b), we again have

$$\begin{aligned} -\Delta_- h_{1/2} &= -\left[h(u_0^{(0)} + \tilde{u}_0, u_1^{(0)} - \tilde{\tilde{u}}_1) - h(g(t), u_0^{(0)} - \tilde{\tilde{u}}_0)\right] \\ &= C_{1/2}\Delta_+ u_0^{(0)} - D_{-1/2}\Delta_- u_0^{(0)}, \end{aligned} \tag{3.6a}$$

where

$$\begin{aligned} C_{1/2} &= -h_2\left(1 - \frac{\Delta_+ \tilde{\tilde{u}}_j}{\Delta_+ u_j^{(0)}}\right), \\ D_{-1/2} &= h_1\left(1 + \frac{\tilde{u}_0}{\Delta_- u_j^{(0)}}\right), \quad \text{and} \quad u_{-1}^{(0)} \equiv g(t). \end{aligned} \tag{3.6b}$$

Now (3.2b) clearly implies $C_{1/2}$, $D_{-1/2} \geq 0$ and $\lambda D_{-1/2} \leq 1$ under (2.17) with $\theta = 2$. We then follow the arguments in [25, Theorem 2.1] to obtain

$$\text{TV}((u^{(0)})^{n+1}) \leq \text{TV}((u^{(0)})^n) + |g(t^{n+1}) - g(t^n)| + \tilde{M}h, \tag{3.7}$$

with TV defined by

$$\text{TV}(u^{(0)}) = \sum_{j \geq -1} |u_{j+1}^{(0)} - u_j^{(0)}|, \qquad u_{-1}^{(0)} \equiv g(t). \tag{3.8}$$

□

Remark 3.2. In the above proof we see that the boundary treatment itself does not increase the total variation; i.e., if $g \equiv$ constant and limiter (2.22) is used, then the resulting schemes with boundary treatments (3.2a) and (3.2b) are both TVDM. □

4. Numerical Results. In this section we use some model problems to numerically test our schemes. We use the third-order scheme (2.32), (2.10) (i.e., $r = 3$ in (2.34b), $k = 2$ in (2.10)). The local projection limiter is (2.21)–(2.24)–(2.25b). We do not use any of the entropy forcing limiters (2.29) or (2.30). The E-flux we use is the Roe flux with entropy correction (2.28e). Comparing with [5], we can see the improvements (mainly in smooth region accuracy) by increasing the order of the scheme from 2 to 3; and comparing with [24], [25], [27] we can see that the results here are comparable to nonoscillatory finite difference schemes.

Example 1. We solve the Burgers' equation with a periodic boundary condition:

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \tag{4.1a}$$

$$u(x,0) = u_0(x) = \frac{1}{4} + \frac{1}{2}\sin \pi x. \tag{4.1b}$$

The exact solution is smooth up to $t = 2/\pi$, then it develops a moving shock which interacts with rarefaction waves. Observe that there is a sonic point. For details, see [10].

At $t = 0.3$ the solution is still smooth. We list the errors in Table 1. Note that we have the full order of accuracy, 3, in both L_1 and L_∞ norms.

At $t = 2/\pi$ the shock just begins to form; at $t = 1.1$ the reaction between the shock and the rarefaction waves is over, and the solution becomes monotone between the shocks. In Figures 1–4 we can see the excellent behavior of our scheme in both cases. The errors 0.1 away from the shock (i.e., $|x$-shock location$| \geq 0.1$) are listed in Table 2. These errors are of the same magnitude as in the smooth case of Table 1.

This example illustrates the results in Section 2: uniformly high order in smooth regions, including at critical and sonic points, and good shock transitions. □

Example 2. We solve the same problem as in Example 1, except that we drop "periodic" and impose boundary conditions:

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \qquad -1 < x < 1, \tag{4.2a}$$

$$u(x,0) = u_0(x) = \frac{1}{4} + \frac{1}{2}\sin \pi x, \qquad -1 \leq x \leq 1, \tag{4.2b}$$

$$u(-1,t) = g(t) = v(-1,t), \tag{4.2c}$$

where $v(x,t)$ is the exact solution of (4.1).

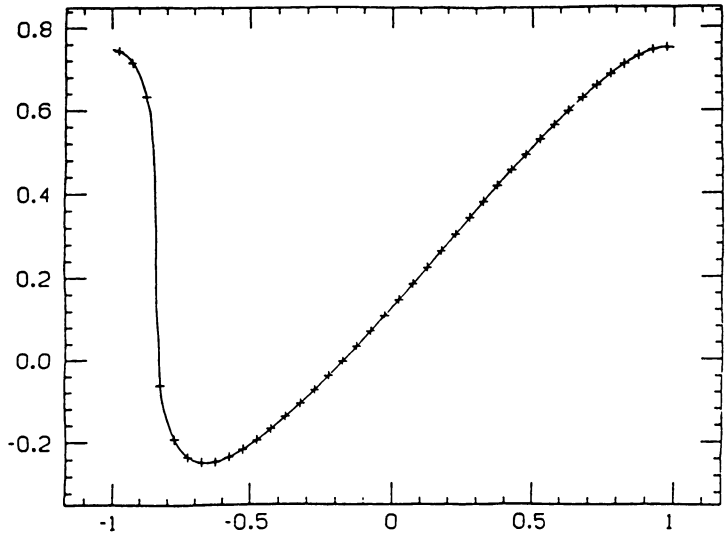

FIGURE 1

Burgers' equation, initial value problem (4.1), $t = \frac{2}{\pi}, \Delta x = \frac{1}{20}$.
(In all the figures, the solid lines are exact solutions, the '+' signs are numerical solutions—one point per cell only.)

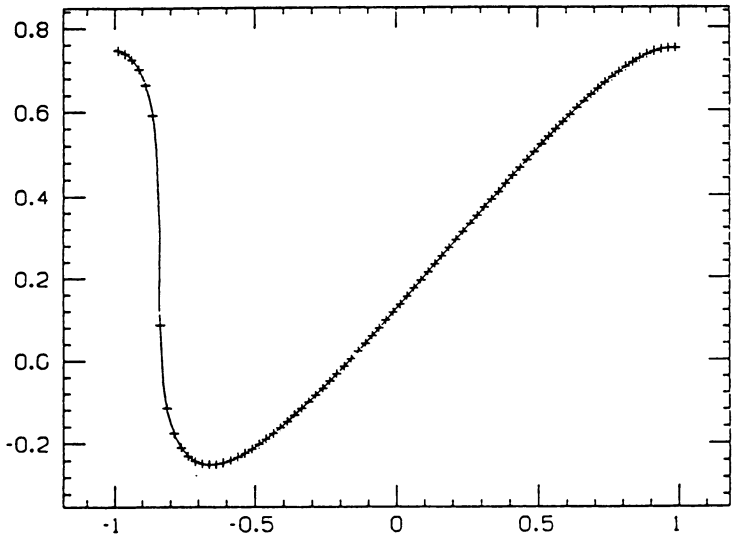

FIGURE 2

Burgers' equation, initial value problem (4.1), $t = \frac{2}{\pi}, \Delta x = \frac{1}{40}$.

Note that (4.2) is well posed, because $x = -1$ is an inflow and $x = 1$ an outflow boundary.

We use the boundary treatment (3.2b) for $x = -1$, and (3.2a) for $x = 1$.

Table 3 contains the errors at $t = 0.3$. Table 4 contains the errors 0.1 away from the shock, at $t = 2/\pi$ and $t = 1.1$. Comparing with Tables 1 and 2, we see almost identical results. Figures 5–6 are the shock transitions for $t = 2/\pi$ and $t = 1.1$. They are also almost identical to Figures 2 and 4 for initial value problems.

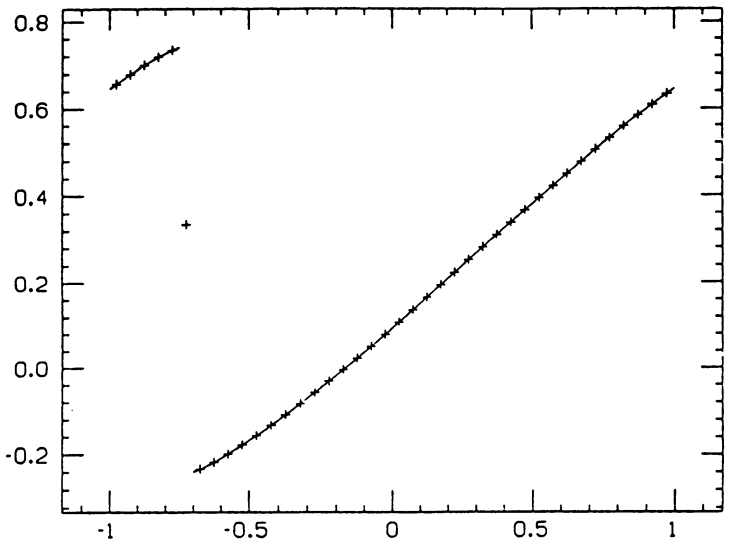

FIGURE 3
Burgers' equation, initial value problem (4.1), $t = 1.1, \Delta x = \frac{1}{20}$.

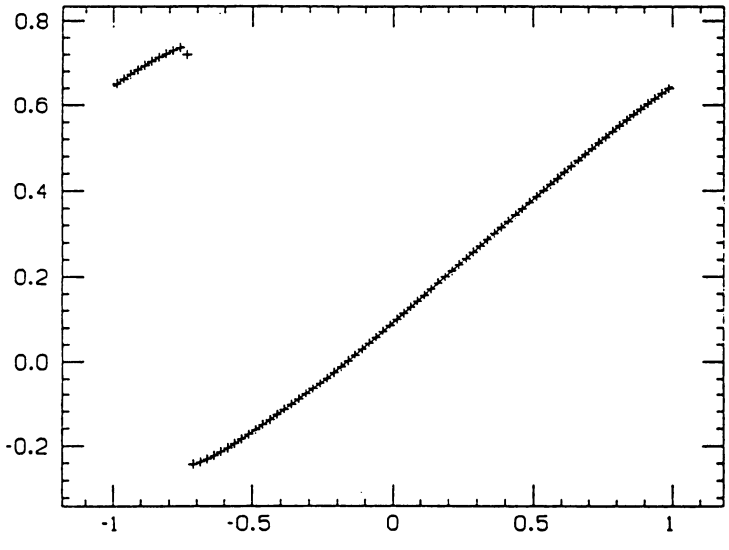

FIGURE 4
Burgers' equation, initial value problem (4.1), $t = 1.1, \Delta x = \frac{1}{40}$.

Since we only compute up to $t = 1.1$, the $g(t)$ in (4.2c) is still smooth. In order to see the behavior of the boundary treatment (3.2b) when $g(t)$ is discontinuous, we shift (4.2) by 0.2:

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \qquad -1 \le x \le 1, \tag{4.3a}$$

$$u(x,0) = \tilde{u}_0(x) = \frac{1}{4} + \frac{1}{2}\sin\pi(x+0.2), \qquad -1 \le x \le 1, \tag{4.3b}$$

$$u(-1,t) = \tilde{g}(t) = v(-0.8,t), \tag{4.3c}$$

where $v(x,t)$ is again the exact solution of (4.1).

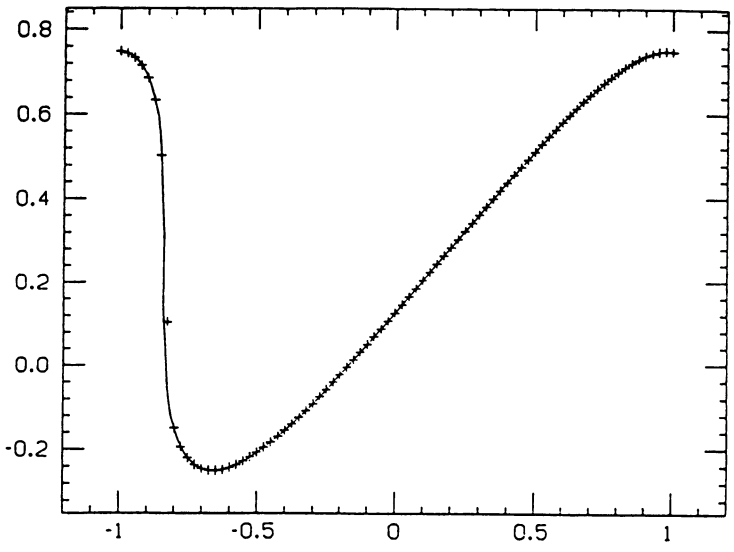

FIGURE 5
Burgers' equation, initial-boundary value problem (4.2), $t = \frac{2}{\pi}, \Delta x = \frac{1}{40}$.

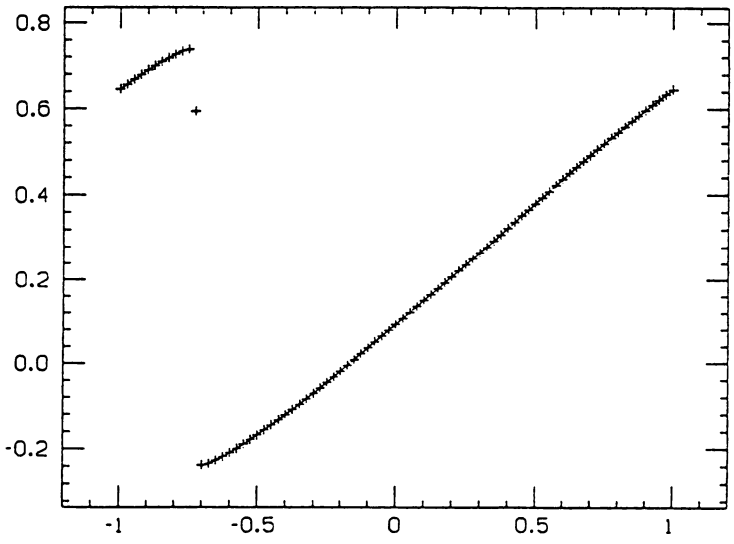

FIGURE 6
Burgers' equation, initial-boundary value problem (4.2), $t = 1.1, \Delta x = \frac{1}{40}$.

This time, $g(t)$ is discontinuous before $t = 1.1$. Nevertheless, we see good behavior of our schemes in Figures 7–8. (Notice that in Figure 7 the shock is only one cell from the boundary.)

This example illustrates the results in Section 3: the boundary treatments (3.2) are accurate and stable even in the presence of a shock emitting from the boundary. □

Example 3. Since we do not use any of the entropy enforcing limiters, we use two nonconvex fluxes to test the convergence to entropy solutions of our scheme.

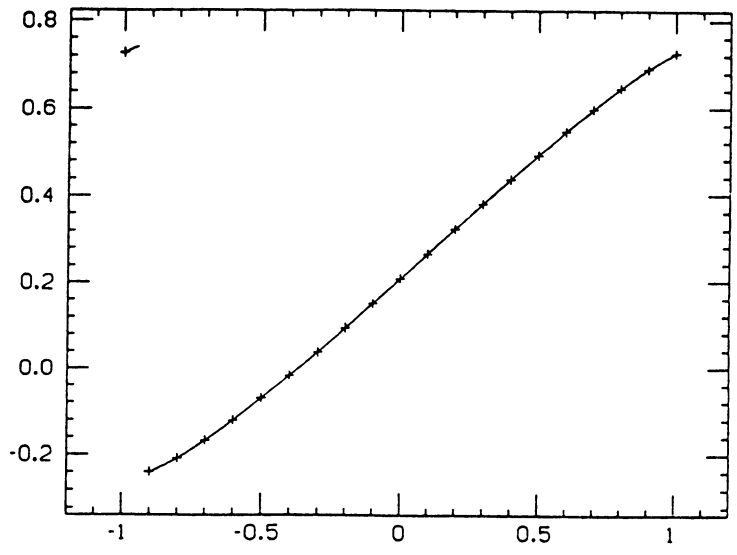

FIGURE 7
Burgers' equation, initial-boundary value problem (4.3), $t = 1.1, \Delta x = \frac{1}{10}$.

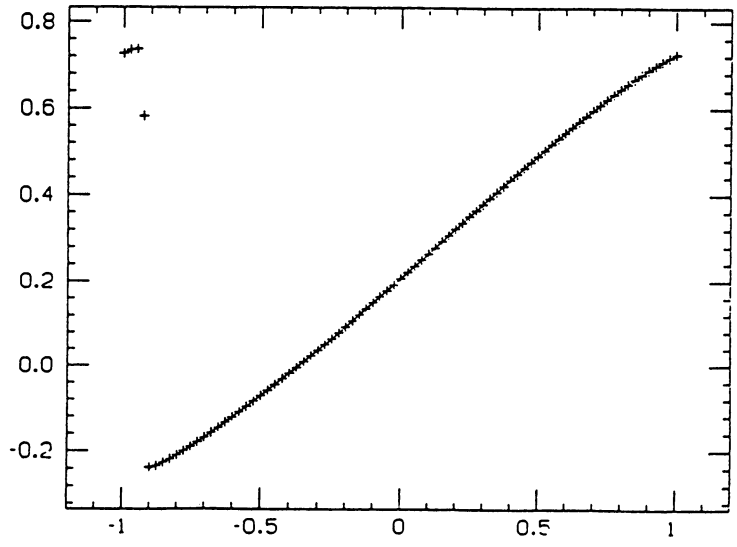

FIGURE 8
Burgers' equation, initial-boundary value problem (4.3), $t = 1.1, \Delta x = \frac{1}{40}$.

The first is a Riemann problem with

$$f(u) = \tfrac{1}{4}(u^2 - 1)(u^2 - 4) \tag{4.4}$$

and initial condition

$$u(x,0) = \begin{cases} u_L, & x < 0, \\ u_R, & x > 0. \end{cases}$$

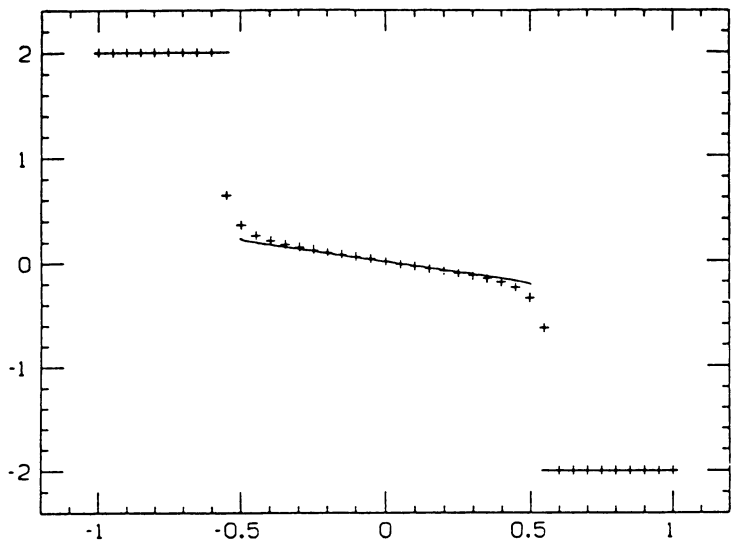

FIGURE 9
Nonconvex flux (4.4), *Riemann problem* $u_L = 2$, $u_R = -2, t = 1, \Delta x = \frac{1}{20}$.

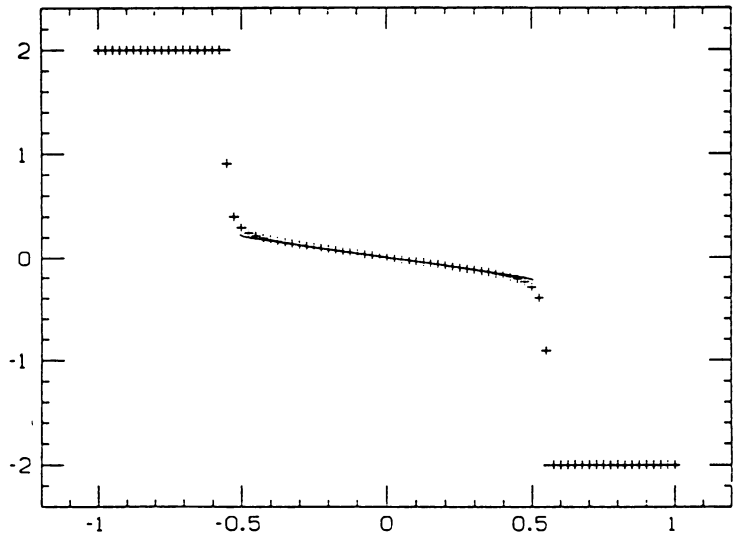

FIGURE 10
Nonconvex flux (4.4), *Riemann problem* $u_L = 2$, $u_R = -2$, $t = 1$, $\Delta x = \frac{1}{40}$.

The two cases we test are (i) $u_L = 2$, $u_R = -2$, Figures 9–10; (ii) $u_L = -3, u_R = 3$, Figures 11–12. For details of this problem, see [10].

The second flux is the Buckley-Leverett flux

$$f(u) = \frac{4u^2}{4u^2 + (1-u)^2}, \tag{4.5}$$

with initial data $u = 1$ in $[-\frac{1}{2}, 0]$ and $u = 0$ elsewhere. The results are in Figures 13–14.

In all cases, we see convergence with good resolution to the entropy solutions.

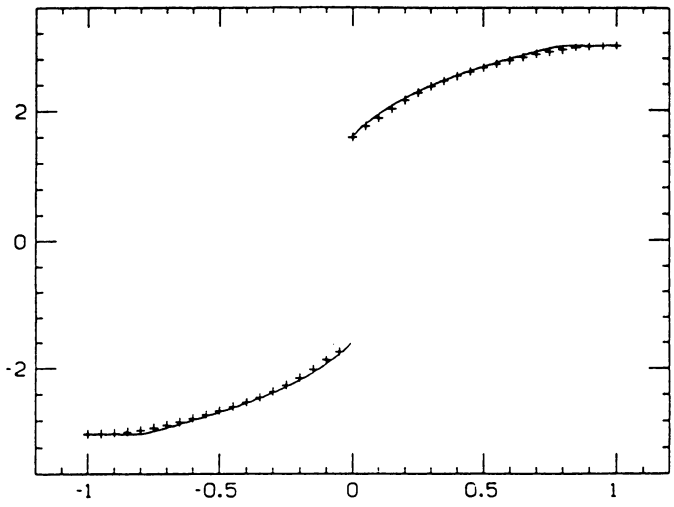

FIGURE 11

Nonconvex flux (4.4), *Riemann problem* $u_L = -3$, $u_R = 3$, $t = 0.04$, $\Delta x = \frac{1}{20}$.

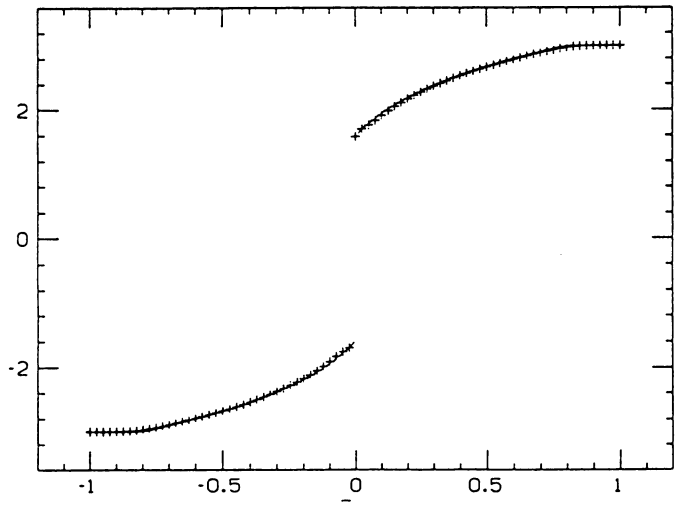

FIGURE 12

Nonconvex flux (4.4), *Riemann problem* $u_L = -3$, $u_R = 3, t = 0.04$, $\Delta x = \frac{1}{40}$.

This example illustrates that our scheme converges to entropy solutions even for nonconvex flux f. □

TABLE 1 (*Example* 1)

$t = 0.3$, *initial value problem* (4.1).

	L^1		L^∞	
Δx	$10^5 \cdot$ error	order	$10^5 \cdot$ error	order
1/10	7.10		4.80	
1/20	0.94	2.92	0.66	2.86
1/40	0.12	2.97	0.09	2.87

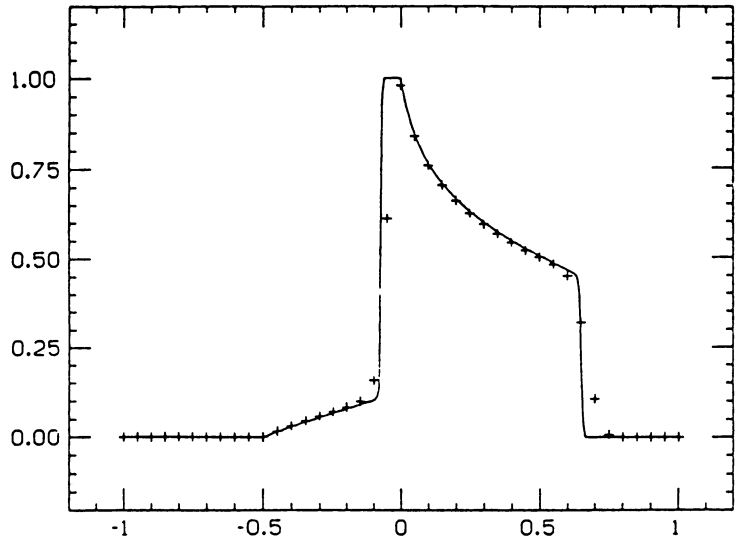

FIGURE 13
Nonconvex flux (4.5), $u = 1$ *in* $[-\frac{1}{2}, 0]$, $u = 0$ *elsewhere*, $t = 0.4$, $\Delta x = \frac{1}{20}$.

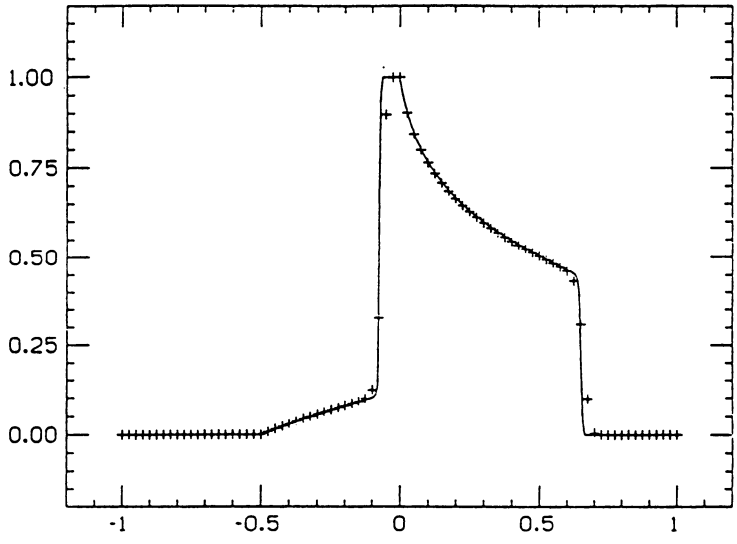

FIGURE 14
Nonconvex flux (4.5), $u = 1$ *in* $[-\frac{1}{2}, 0]$, $u = 0$ *elsewhere*, $t = 0.4$, $\Delta x = \frac{1}{40}$.

TABLE 2 (*Example* 1)
Errors in smooth region $|x - \text{shock}| \geq 0.1$; $\Delta x = \frac{1}{40}$, *initial value problem* (4.1).

L^∞		L^1	
$t = \frac{2}{\pi}$	$t = 1.1$	$t = \frac{2}{\pi}$	$t = 1.1$
0.13×10^{-4}	0.61×10^{-6}	0.11×10^{-5}	0.14×10^{-6}

TABLE 3 (*Example* 2)
$t = 0.3$, *initial value problem* (4.2).

	L^1		L^∞	
Δ_x	$10^5 \cdot$ error	order	$10^5 \cdot$ error	order
1/10	6.82	–	3.82	–
1/20	0.92	2.90	0.56	2.77
1/40	0.12	2.94	0.09	2.78

TABLE 4 (*Example* 2)
Errors in smooth region $|x - \text{shock}| \geq 0.1; \Delta x = \frac{1}{40}$, *initial-boundary value problem*(4.2).

L^∞		L^1	
$t = \frac{2}{\pi}$	$t = 1.1$	$t = \frac{2}{\pi}$	$t = 1.1$
0.15×10^{-4}	0.65×10^{-6}	0.11×10^{-5}	0.13×10^{-6}

5. Summary. We present a general framework, by using one-dimensional scalar initial value problems and initial-boundary value problems as models, of constructing and analyzing a class of TVB discontinuous Galerkin finite element methods for solving conservation laws. This new class of methods differs from most finite element methods in that they are explicit in time, hence can be implemented with high-order TVD Runge-Kutta type time discretizations. By using a local projection limiter, which does not affect accuracy in smooth regions, we can prove TVBM and TVB, hence achieve convergence without oscillations for shock calculations. Comparing with finite difference methods, these methods retain the advantage of finite element methods, i.e., achieving high accuracy by using more information within a cell rather than using a wide stencil, hence are easier to apply for boundary conditions (discussed in this paper) and complicated geometries (will be discussed in future papers). Numerical results are given to illustrate the good convergence behavior in several test problems.

Acknowledgments. We thank the referee for helpful comments on the first version of this paper. The second author also wants to thank Mr. Wei Cai for helpful discussions. This research was supported in part by the Institute for Mathematics and its Applications.

School of Mathematics
University of Minnesota
Minneapolis, Minnesota 55455
E-mail: cockburn%umn-cs-fsa@umn-cs.cs.umn.edu

Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912
E-mail: am508000@brownvm.bitnet

1. A. BOURGEAT & B. COCKBURN, *The TVD-projection method for solving implicit numerical schemes for scalar conservation laws: A numerical study of a simple case*, IMA Preprint Series #311, University of Minnesota, April 1987. To appear in *SIAM J. Sci. Statist. Comput.*, March 1989.

2. G. CHAVENT & B. COCKBURN, *The local projection P^0P^1-discontinuous-Galerkin finite element method for scalar conservation laws*, IMA Preprint Series #341, University of Minnesota, September 1987, to appear in M^2AN.

3. G. CHAVENT & G. SALZANO, "A finite element method for the 1D water flooding problem with gravity," *J. Comput. Phys.*, v. 45, 1982, pp. 307–344.

4. B. COCKBURN, *The quasi-monotone schemes for scalar conservation laws*, I, II, *and* III, IMA Preprint Series #263, 268, 277, University of Minnesota, September and October 1986.

5. B. COCKBURN & C.-W. SHU, *The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws*, IMA Preprint Series #388, University of Minnesota, 1988.

6. M. CRANDALL & A. MAJDA, "Monotone difference approximations for scalar conservation laws," *Math. Comp.*, v. 34, 1980, pp. 1–21.

7. A. HARTEN, "On a class of high resolution total-variation-stable finite-difference schemes," *SIAM J. Numer. Anal.*, v. 21, 1984, pp. 1–23.

8. A. HARTEN, *Preliminary results on the extension of ENO schemes to two-dimensional problems*, Proc. Internat. Conf. on Hyperbolic Problems, Saint-Etienne, January 1986.

9. A. HARTEN & S. OSHER, "Uniformly high-order accurate nonoscillatory schemes, I," *SIAM J. Numer. Anal.*, v. 24, 1987, pp. 279–309.

10. A. HARTEN, B. ENGQUIST, S. OSHER & S. CHAKRAVARTHY, "Uniformly high order accurate essentially non-oscillatory schemes, III," *J. Comput. Phys.*, v. 71, 1987, pp. 231–303.

11. T. HUGHES & A. BROOK, "Streamline upwind-Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations," *Comput. Methods Appl. Mech. Engrg.*, v. 32, 1982, pp. 199–259.

12. T. HUGHES, L. P. FRANCA, M. MALLET & A. MISUKAMI, "A new finite element formulation for computational fluid dynamics, I, II, III and IV," *Comput. Methods Appl. Mech. Engrg.*, v. 54, 58, 1986, pp. 223–234, 341–355; pp. 305–328, 329–336.

13. C. JOHNSON & J. PITKÄRANTA, "An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation," *Math. Comp.*, v. 46, 1986, pp. 1–26.

14. C. JOHNSON & J. SARANEN, "Streamline diffusion methods for problems in fluid mechanics," *Math. Comp.*, v. 47, 1986, pp. 1–18.

15. C. JOHNSON & A. SZEPESSY, "On the convergence of a finite element method for a nonlinear hyperbolic conservation law," *Math. Comp.*, v. 49, 1987, pp. 427–444.

16. C. JOHNSON, A. SZEPESSY & P. HANSBO, "On the convergence of shock capturing streamline diffusion finite element methods for hyperbolic conservation laws," preprint.

17. B. VAN LEER, "Towards the ultimate conservation difference scheme, II and V," *J. Comput. Phys.*, v. 14 and 32, 1974 and 1979, pp. 361–376 and 101–136.

18. P. LESAINT & P.-A. RAVIART, "On a finite element method for solving the neutron transport equation," in *Mathematical Aspects of Finite Elements in Partial Differential Equations* (C. de Boor, ed.), Academic Press, 1974, pp. 89–145.

19. K. W. MORTON & P. K. SWEBY, "A comparison of flux limited difference methods and characteristic Galerkin methods for shock modelling," *J. Comput. Phys.*, v. 73, 1987, pp. 203–230.

20. S. OSHER, "Convergence of generalized MUSCL schemes," *SIAM J. Numer. Anal.*, v. 22, 1985, pp. 947–961.

21. S. OSHER & S. CHAKRAVARTHY, "High resolution schemes and the entropy condition," *SIAM J. Numer. Anal.*, v. 21, 1984, pp. 955–984.

22. S. OSHER & E. TADMOR, "On the convergence of difference approximations to scalar conservation laws," *Math. Comp.*, v. 50, 1988, pp. 19-51.

23. R. SANDERS, "A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation laws," *Math. Comp.*, v. 51, 1988, pp. 535–558.

24. C.-W. SHU, "TVB uniformly high-order schemes for conservation laws," *Math. Comp.*, v. 49, 1987, pp. 105–121.

25. C.-W. SHU, "TVB boundary treatment for numerical solutions of conservation laws," *Math. Comp.*, v. 49, 1987, pp. 123–134.

26. C.-W. SHU, "Total-Variation-Diminishing time discretizations," *SIAM J. Sci. Statist. Comput.*, v. 9, 1988, pp. 1073–1084.

27. C.-W. SHU & S. OSHER, "Efficient implementation of essentially non-oscillatory shock-capturing schemes," *J. Comput. Phys.*, v. 77, 1988, pp. 439–471.

28. P. SWEBY, "High resolution schemes using flux limiters for hyperbolic conservation laws," *SIAM J. Numer. Anal.*, v. 21, 1984, pp. 995–1011.