JOURNAL OF COMPUTATIONAL PHYSICS 112, 308–315 (1994)

On an Essentially Conservative Scheme for Hyperbolic Conservation Laws*

BAOXIA JIN

Computing Center, Academia Sinica, Beijing 100080, Peoples' Republic of China

Received April 5, 1993

In this paper, a class of essentially conservative scheme are constructed and analyzed. The numerical tests and theoretical analysis show that although these schemes can not be written in the usual conservation form, but the numerical solutions obtained with these schemes can converge, as the mesh size tends to zero, to the physical solution of conservation laws.

1. INTRODUCTION

The estimation of the total variation of numerical solutions is an important part in the proof of the convergence property of many finite difference schemes for hyperbolic conservation laws. To ensure that the total variation of numerical solutions is uniformly bounded, a variety of TVD (total variation diminishing) schemes have been constructed and widely used in the numerical computation of conservation laws [1–3]. The advantage of TVD schemes is its high resolution for shock waves. Unfortunately, the TVD property and the second-order accuracy of a difference scheme is inconsistent. In the one-dimensional case, the inconsistency means that you cannot construct a TVD scheme, meanwhile which has uniformly high-order accuracy. In the two-dimensional case, the inconsistency means that a TVD scheme is at most first-order accurate [4]. To obtain uniformly high-order schemes, a series of work has been done by Harten, Osher, Enquist, Chakravarthy, who construct an ENO (essentially non-oscillatory) type scheme [5, 6], and Shu who constructs a TVB (total variation bounded) scheme [7]. These schemes can obtain uniformly high-order accuracy and do not generate too much oscillation near strong shock waves; however, there are still some problems unsolved, among which are the difficulty to estimate the bound of the total variation of numerical solutions of ENO type schemes, the difficulty to extend the same idea to the high-dimensional case, and the difficulty to prove the entropy conditions of the schemes.

The purpose of this paper is to develop a new kind of scheme which should not only give good results in real computations but also have good theoretical properties. These schemes are called essentially conservative ones because they cannot be written in the usual conservation form; but it can be proven that the numerical solutions of these schemes converge to the weak solution of conservation laws as the mesh size tends to zero. In Section 2, we give some definitions and statements. In Section 3, we construct a uniformly second-order essentially conservative scheme and give some theoretical analysis for scalar case. Section 4 is for the system of hyperbolic conservation laws. Section 5 presents some numerical results of these essentially conservative schemes and some conclusions of this paper.

2. DEFINITIONS AND STATEMENTS

To simplify the analysis, consider the following one-dimensional conservation laws,

$$\frac{\partial u}{\partial t}+\frac{\partial f(u)}{\partial x}=0, \qquad (x, t)\in[0, T]\times R^1 \tag{2.1}$$

$$u(x, 0)=u_0(x), \qquad x\in R^1, \tag{2.2}$$

where $u=(u_1, ..., u_m)^{\mathrm{T}}, f(u)=(f_1(u), ..., f_m(u))^{\mathrm{T}}$.

Suppose the finite difference approximation of (2.1) ~ (2.2) can be written in the operator form as

$$\begin{aligned} u^{n+1}&=C(\Delta t, \Delta x, u^n)\\ u^0&=u_0, \end{aligned} \tag{2.3}$$

where $u^n=\{u_i^n\}$, Δt and Δx are the time step and the spatial mesh size, respectively.

* The work for this paper was supported by the National Natural Science Foundation of China.

0021-9991/94 $6.00

Let

$$u_\Delta(x, t) = u_i^n, \qquad \text{if} \quad (x, t) \in \Omega_\Delta,$$

$$\Omega_\Delta = \left[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}\right] \times [t^n, t^n + \Delta t], \qquad x_i = i\,\Delta x.$$

DEFINITION 2.1. Suppose $u_\Delta \xrightarrow{L_1} u$. If u is the weak solution of Eq. (2.1); i.e., for any test function $\varphi \in C_0^\infty(R^2)$, equality

$$\iint_{[0,t]\times R}\left[u\frac{\partial \varphi}{\partial t} + f(u)\frac{\partial \varphi}{\partial x}\right]dx\,dt = 0$$

is valid, then scheme (2.3) is called an essentially conservative scheme.

Obviously, the usual conservative scheme, such as a Lax scheme or a Godunov scheme, must be the essentially conservative scheme, but an essentially conservative scheme is not necessarily a conservative scheme.

If the limit solution u satisfies the entropy condition, i.e., for any entropy function $U(u)$ and entropy flux $F(u)$ of Eq. (2.1), inequality

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} \leqslant 0$$

is valid in a distribution sense; we call scheme (2.3) an entropy scheme.

In a scalar case, the x total variation of numerical solutions, for any fixed $t > 0$, is defined as

$$TV(u^n) = \sum_j |\Delta_{j+1/2}^n u|$$

$$\Delta_{j+1/2}^n u = u_{j+1}^n - u_j^n.$$

To prove the convergence of difference schemes, a bound on the total variation of u^n is often needed. Since 1983, many high resolution TVD (total variation diminishing) schemes have been constructed which satisfy

$$TV(u^{n+1}) \leqslant TV(u^n).$$

If the initial data $u_0(x) \in BV(R)$, it can be deduced that

$$TV(u^n) \leqslant TV(u_0) < \infty.$$

Thus, a convergent subsequence of numerical solutions can be obtained. Unfortunately, a depressing has been proved by Goodman and Leveque, that a two-dimensional TVD scheme is at most first-order accurate [4]. Even in a one-dimensional case, it is also difficult to construct a uniformly high-order accurate TVD scheme.

To obtain uniformly high-order accurate scheme in one- or two-dimensional case, the restriction on the total variation of numerical solutions must be relaxed. In this paper we only ask that the numerical schemes satisfy

$$TV(u^n) \leqslant K,$$

where K is a constant which does not depend on Δt and Δx, this kind of scheme is called a TVB (total variation bounded) scheme [4].

3. THE SCALAR CASE

We start our construction from the following scalar hyperbolic equation:

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0. \tag{3.1}$$

The first-order three points TVD scheme of Eq. (3.1) can be written as

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2} - h_{j-1/2}) \tag{3.2}$$

where $\lambda = \Delta t/\Delta x$; $h_{j+1/2} = h(u_j, u_{j+1})$ is the numerical flux, which is consistent with Eq. (3.1) in the sense that $h(u, u) = f(u)$.

Let $\phi_{j+1/2}$ be the difference between $h_{j+1/2}$ and the Lax–Wendroff numerical flux denoted as $h_{j+1/2}^{\mathrm{L-W}}$; i.e.,

$$\phi_{j+1/2} = h_{j+1/2}^{\mathrm{L-W}} - h_{j+1/2}, \tag{3.3}$$

where $h_{j+1/2}^{\mathrm{L-W}} = \frac{1}{2}[f_j + f_{j+1} - \lambda(a_{j+1/2})^2\,\Delta_{j+1/2}u]$ and $a_{j+1/2}$ is the local characteristic speed

$$a_{j+1/2} = \begin{cases} \dfrac{f_{j+1} - f_j}{\Delta_{j+1/2}u}, & \text{if} \quad \Delta_{j+1/2}u \neq 0, \\ \left.\dfrac{\partial f}{\partial u}\right|_{u=u_j}, & \text{if} \quad \Delta_{j+1/2}u = 0. \end{cases}$$

In smooth regions of u, $\phi_{j+1/2}$ satisfies

$$\phi_{j+1/2} - \phi_{j-1/2} = O(\Delta x^2). \tag{3.4}$$

Consider the following finite difference approximation of Eq. (2.1):

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2} - h_{j-1/2}) + \lambda Q_j, \tag{3.5}$$

$$Q_j = -\min\left\{M \max\left(1, \frac{|\Delta_{j+1/2}^n u|}{\Delta x}, \frac{|\Delta_{j-1/2}^n u|}{\Delta x}\right) \times \Delta x^2, |\phi_{j+1/2} - \phi_{j-1/2}|\right\} S_j \tag{3.6}$$

$S_j = \operatorname{sign}(\phi_{j+1/2} - \phi_{j-1/2})$, M is a positive number.

Q_j can be looked upon as an improving term to the first-order scheme (3.2). Scheme (3.5)–(3.6) cannot be written in conservation form, but it is essentially conservative.

THEOREM 3.1. *The finite difference scheme* (3.5)–(3.6) *is an essentially conservative scheme.*

Proof. Scheme (3.5)–(3.6) can be written as

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{h_{j+1/2}-h_{j-1/2}}{\Delta x}=\frac{Q_j}{\Delta x}. \tag{3.7}$$

Choose a test function $\varphi(x,t)\in C_0^\infty([0,t]\times R^1)$, multiply (3.7) with $\Delta t\,\Delta x\,\varphi_j^n$, and sum to j and n; we have

$$-\sum_{j,n}\left(u_j^n\frac{\varphi_j^n-\varphi_j^{n-1}}{\Delta t}+h_{j+1/2}^n\frac{\varphi_{j+1}^n-\varphi_j^n}{\Delta x}\right)\Delta x\,\Delta t$$
$$=\sum_{j,n}\varphi_j^n\frac{Q_j}{\Delta x}\Delta x\,\Delta t.$$

From the definition of Q_j, it can be deduced that

$$\left|\sum_{j,n}\varphi_j^n\frac{Q_j}{\Delta x}\Delta x\,\Delta t\right|\leqslant M\sum_{j,n}|\varphi_j^n|(\Delta x+|\Delta_{j+1/2}u|+|\Delta_{j-1/2}u|)\,\Delta x\,\Delta t.$$

So, if $\{u_j^n\}\xrightarrow{L_1}u$, we have

$$\left|\sum_{j,n}\varphi_j^n\frac{Q_j}{\Delta x}\Delta x\,\Delta t\right|\to 0\qquad(\text{as }\Delta t\to 0,\ \Delta x\to 0),$$

which implies that

$$-\sum_{j,n}\left(u_j^n\frac{\varphi_j^n-\varphi_j^{n-1}}{\Delta t}+h_{j+1/2}^n\frac{\varphi_{j+1}^n-\varphi_j^n}{\Delta x}\right)\Delta x\,\Delta t\to 0$$
$$(\text{as}\quad \Delta t\to 0,\quad \Delta x\to 0).$$

The consistency of $h_{j+1/2}$ gives

$$\iint_{[0,T]\times R^1}\left(u\frac{\partial\varphi}{\partial t}+f(u)\frac{\partial\varphi}{\partial x}\right)dx\,dt=0.\quad\blacksquare$$

To simplify the analysis, we choose $h_{j+1/2}$ to be the first-order Roe's scheme, i.e.,

$$h_{j+1/2}=\tfrac{1}{2}(f_j+f_{j+1}-|a_{j+1/2}|)\,\Delta_{j+1/2}u;$$

the analysis for other cases is similar. We have

$$\phi_{j+1/2}=\tfrac{1}{2}|a_{j+1/2}|\,(1-\lambda|a_{j+1/2}|)\,\Delta_{j+1/2}u.$$

In the smooth regions of function u, if the inequality

$$\frac{\partial}{\partial x}\left\{|a|\,(1-\lambda\,|a|)\frac{\partial u}{\partial x}\right\}<M \tag{3.8}$$

is valid, then Q_j satisfies

$$Q_j=-(\phi_{j+1/2}-\phi_{j-1/2})$$

as Δt and Δx are sufficiently small. Thus

$$u_j^{n+1}=u_j^n-\lambda(h_{j+1/2}^{\mathrm{L-W}}-h_{j-1/2}^{\mathrm{L-W}});$$

i.e., the scheme is second-order accurate. In the computation we can choose M large enough so that inequality (3.8) is valid everywhere in smooth regions of u; then the scheme is second-order accurate everywhere in smooth regions of u.

To keep it nonoscillatory in the computation of discontinuous solutions, we modify $\phi_{j+1/2}$ as

$$\phi_{j+1/2}=\text{minmod}(\sigma_{j-1/2}+D_{j+1/2}^L,\ \sigma_{j+1/2},\ \sigma_{j+3/2}-D_{j+1/2}^R) \tag{3.9}$$
$$\sigma_{j+1/2}=\tfrac{1}{2}|a_{j+1/2}^n|(1-\lambda\,|a_{j+1/2}^n|)\,\Delta_{j+1/2}^n u$$
$$D_{j+1/2}^L=\text{minmod}\times(\sigma_{j-1/2}-\sigma_{j-3/2},\ \sigma_{j+1/2}-\sigma_{j-1/2}) \tag{3.10}$$
$$D_{j+1/2}^R=\text{minmod}\times(\sigma_{j+3/2}-\sigma_{j+1/2},\ \sigma_{j+5/2}-\sigma_{j+3/2}), \tag{3.11}$$

where

$$\text{minmod}(x_1,\ldots,x_k)=\begin{cases}\text{sign}(x_1)\min(|x_1|,\ldots,|x_k|),\\ \qquad\text{if}\quad x_1,\ldots,x_n\text{ have same sign}\\ 0,\qquad\text{else}\end{cases}$$

(3.9)–(3.11) imply that, in smooth regions,

$$\phi_{j+1/2}=\tfrac{1}{2}|a_{j+1/2}^n|\,(1-\lambda\,|a_{j+1/2}^n|)\,\Delta_{j+1/2}^n u+O(\Delta x^3)$$
(see [8] for details).

Thus scheme (3.5)–(3.6) with $\phi_{j+1/2}$ being modified as (3.9)–(3.11) is still second-order accurate even at the local extreme points of u.

THEOREM 3.2. *Scheme* (3.5)–(3.6) *with* $\phi_{j+1/2}$ *being modified as* (3.9)–(3.11) *is monotonicity preserving under the CFL restriction*

$$\mu=\max_j\lambda\,|a_{j+1/2}|\leqslant 0.5. \tag{3.12}$$

Proof. Q_j can be written as

$$Q_j=-\varepsilon(\phi_{j+1/2}-\phi_{j-1/2}),\qquad 0\leqslant\varepsilon\leqslant 1,$$

so that scheme (3.5)–(3.6) can be written as

$$u_j^{n+1} = u_j^n + C^-_{j+1/2}\,\Delta^n_{j+1/2}u - C^+_{j-1/2}\,\Delta^n_{j-1/2}u - \varepsilon\phi_{j+1/2} + \varepsilon\phi_{j-1/2}, \quad (3.13)$$

$$C^{\pm}_{j+1/2} = \frac{\lambda}{2}(|a_{j+1/2}| \pm a_{j+1/2}). \quad (3.14)$$

Therefore,

$$\Delta^{n+1}_{j+1/2}u = (1 - C^+_{j+1/2} - C^-_{j+1/2})\,\Delta^n_{j+1/2}u + C^-_{j+3/2}\,\Delta^n_{j+1/2}u + C^+_{j-1/2}\,\Delta^n_{j-1/2}u - A_1\phi_{j+3/2} + 2A_2\phi_{j+1/2} - A_3\phi_{j-1/2},$$

where $0 \leqslant A_1, A_2, A_3 \leqslant 1$, (3.10), (3.11), and (3.13) imply that

$$A_1\phi_{j+1/2} = \varepsilon_1\sigma_{j+1/2}$$
$$A_2\phi_{j+3/2} = \varepsilon_2\sigma_{j+1/2} - \varepsilon_3\sigma_{j-1/2}$$
$$A_3\phi_{j-1/2} = \varepsilon_4\sigma_{j+1/2} - \varepsilon_5\sigma_{j+3/2},$$

here $0 \leqslant \varepsilon_1, \varepsilon_3, \varepsilon_5 \leqslant 1$, $0 \leqslant \varepsilon_2, \varepsilon_4 \leqslant 2$. So

$$\Delta^{n+1}_{j+1/2}u = (1 - \lambda|a_{j+1/2}| - \varepsilon_2 C^0_{j+1/2} - \varepsilon_4 C^0_{j+1/2})\,\Delta^n_{j+1/2}u + (C^-_{j+3/2} + \varepsilon_5 C^0_{j+3/2})\,\Delta^n_{j+3/2}u + (C^+_{j-1/2} + \varepsilon_3 C^0_{j-1/2})\,\Delta^n_{j-1/2}u + 2\varepsilon_1 C^0_{j+1/2}\,\Delta^n_{j+1/2}u,$$

where $C^0_{j+1/2} = \frac{1}{2}\lambda|a_{j+1/2}|(1 - \lambda|a_{j+1/2}|)$. When CFL condition (3.13) being satisfied, we have

$$C^0_{j+1/2} \geqslant 0, \qquad C^{\pm}_{j+1/2} \geqslant 0, \qquad j = 0, \pm 1, \pm 2, \ldots,$$

and

$$1 - \lambda|a_{j+1/2}| - \varepsilon_2 C^0_{j+1/2} - \varepsilon_4 C^0_{j+1/2} \geqslant 1 - \lambda|a_{j+1/2}| - \tfrac{1}{2} \geqslant 0,$$

which imply that if $\Delta^n_{j+1/2}u$ ($j = 0, \pm 1, \pm 2, \ldots$), have same sign, then $\Delta^{n+1}_{j+1/2}u$ ($j = 0, \pm 1, \pm 2, \ldots$), also have same sign as $\Delta^n_{j+1/2}u$, that means the numerical solutions can preserve the monotonicity of the initial function $u_0(x)$. ∎

Remark. In the proof of Theorem 3.2, if $C^-_{j+3/2}$ and $C^-_{j-1/2}$ do not equal zero at same time, for example, $C^-_{j+3/2} \neq 0$, i.e., $a_{j+3/2} < 0$, then

$$\phi_{j+3/2} = \varepsilon_6 C^0_{j+3/2}\,\Delta u^n_{j+3/2},$$

where $0 \leqslant \varepsilon_6 \leqslant 1$; thus

$$\Delta^{n+1}_{j+1/2}u = (1 - \lambda|a_{j+1/2}| - \varepsilon_4\sigma_{j+1/2})\,\Delta^n_{j+1/2}u + 2\varepsilon_1 C^0_{j+1/2}\,\Delta^n_{j+1/2}u + (C^-_{j+3/2} - \varepsilon_6 C^0_{j+3/2} + \varepsilon_5 C^0_{j+3/2})\,\Delta^n_{j+3/2}u + C^+_{j-1/2}\,\Delta^n_{j-1/2}u$$

so, only if

$$\mu = \max_j \lambda|a_{j+1/2}| \leqslant 1,$$

we have

$$1 - \lambda|a_{j+1/2}| - \varepsilon_4 C^0_{j+1/2} \geqslant (1 - \lambda|a_{j+1/2}|)^2 \geqslant 0$$

and

$$C^-_{j+3/2} - \varepsilon_6 C^0_{j+3/2} + \varepsilon_5 C^0_{j+3/2} \geqslant \lambda|a_{j+3/2}| - \tfrac{1}{2}\lambda|a_{j+3/2}|(1 - \lambda|a_{j+3/2}|) \geqslant 0$$

so the CFL restriction (3.12) should be satisfied only when $C^-_{j+3/2}$ and $C^-_{j-1/2}$ equal zero at the same time, which occurs at sonic points in the rarefactive regions, where the characteristic speed $|a_{j+1/2}|$ is near zero. So in real computation, we can choose μ larger than 0.5, even near one.

THEOREM 3.3. *Assume that* $u_0(x) \in L^1(R^1) \cap L^\infty(R^1) \cap BV(R^1)$, *and*

$$\frac{\partial u_0}{\partial x} = 0, \qquad \text{if} \quad |x| > B;$$

then under the CFL restriction (3.12), *we have*

1. *The numerical solution of scheme* (3.5)–(3.6) *satisfies*

$$\max_{-\infty < j < \infty} |u_j^n| \leqslant K_0(T).$$

2. *Scheme* (3.5)–(3.6) *is TVB.*

Proof. Scheme (3.5)–(3.6) can be written as

$$u_j^{n+1} = u_j^n + C^+_{j+1/2}\,\Delta^n_{j+1/2}u - C^-_{j-1/2}\,\Delta^n_{j-1/2}u + \lambda Q_j = (1 - C^+_{j+1/2} - C^-_{j-1/2})\,u_j^n + C^+_{j+1/2}u^n_{j+1} + C^-_{j-1/2}u^n_{j-1} + \lambda Q_j.$$

$C^{\pm}_{j\pm 1/2}$ are defined as (3.14). When (3.12) is satisfied we have

$$1 - C^+_{j+1/2} - C^-_{j-1/2} \geqslant 0, \qquad C^+_{j+1/2} \geqslant 0, \qquad C^-_{j-1/2} \geqslant 0.$$

Thus

$$\begin{aligned}|u_j^{n+1}| &\leqslant (1 - C^+_{j+1/2} - C^-_{j-1/2})\,|u_j^n| \\ &\quad + C^+_{j+1/2}\,|u^n_{j+1}| + C^-_{j-1/2}\,|u^n_{j-1}| + \lambda\,|Q_j| \\ &\leqslant \max_{-\infty<j<\infty} |u_j^n| + \lambda\,|Q_j|,\end{aligned}$$

so

$$\max_{-\infty<j<\infty} |u_j^{n+1}| \leqslant \max_{-\infty<j<\infty} |u_j^n| + \lambda\,|Q_j|.$$

From the definition of Q_j, we have

$$|Q_j| \leqslant \lambda M(\Delta x + |\Delta^n_{j+1/2}u| + |\Delta^n_{j-1/2}|)\,\Delta x,$$

so

$$\begin{aligned}\max_{-\infty<j<\infty} |u_j^{n+1}| &\leqslant \max_{-\infty<j<\infty} |u_j^n| + \lambda M\,\Delta x^2 \\ &\quad + 4\lambda M\,\Delta x \max_{-\infty<j<\infty} |u_j^n| \\ &= (1 + 4M\,\Delta t) \max_{-\infty<j<\infty} |u_j^n| + M\,\Delta x\,\Delta t.\end{aligned}$$

Let $K_1 = 4M$, then we have

$$\begin{aligned}\max_{-\infty<j<\infty} |u_j^n| &\leqslant (1 + K_1\,\Delta t)^n \max_{-\infty<j<\infty} |u_j^0| \\ &\quad + \frac{\Delta x}{4}\,[(1 + K_1\,\Delta t)^n - 1]; \\ &\leqslant K_0(T)\end{aligned}$$

here $K_0(T) = e^{K_1 T}\,\|u_0\|_{L^\infty} + (\Delta x/4)(e^{K_1 T} - 1)$.

Now, we prove 2. Let $\tilde{u}_j = u_j^n - \lambda(h^n_{j+1/2} - h^n_{j-1/2})$; then

$$TV(\tilde{u}) \leqslant TV(u^n).$$

Thus

$$\begin{aligned}TV(u^{n+1}) &\leqslant TV(u^n) + 2\lambda \sum_j |Q_j| \\ &\leqslant TV(u^n) + 2\lambda M \\ &\quad \times \sum_{|j\,\Delta x| \leqslant B + 6n\,\Delta x} (\Delta x + 2\,|\Delta^n_{j+1/2}u|)\,\Delta x \\ &\leqslant TV(u^n) + 4M(B + 6n\,\Delta x)\,\Delta t \\ &\quad + 4M\,\Delta t \sum_j |\Delta^n_{j+1/2}u| \\ &= (1 + 4M\,\Delta t)\,TV(u^n) \\ &\quad + 4M(B + 6n\,\Delta x)\,\Delta t \\ &\leqslant (1 + 4M\,\Delta t)\,TV(u^n) \\ &\quad + 4M\left(B + \frac{T}{6\mu} \sup_{|u|<K_0} \left|\frac{\partial f}{\partial u}\right|\right)\Delta t \\ &\leqslant (1 + K_1\,\Delta t)\,TV(u^n) + K_2\,\Delta t,\end{aligned}$$

where $K_2 = 4M(B + (T/6\mu)\sup_{|u|<K_0}|\partial f/\partial u|)$. So

$$\begin{aligned}TV(u^n) &\leqslant (1 + K_1\,\Delta t)^n\,TV(u_0) + \frac{K_2}{K_1}\,[(1 + K_1\,\Delta t)^n - 1] \\ &\leqslant e^{K_1 T} TV(u_0) + \frac{K_2}{K_1}\,(e^{K_1 T} - 1);\end{aligned}$$

i.e., the total variation of numerical solutions is uniformly bounded. ∎

About the entropy condition of scheme (3.5)–(3.6), we have

THEOREM 3.4. *Suppose scheme* (3.2) *is a montone scheme; then the a.e. bounded limit, as* $\Delta t \to 0$ *and* $\Delta x \to 0$, *of the numerical solutions of scheme* (3.5)–(3.6) *is the unique solution of Eq.* (3.1). *We still assume that* $u_0(x) \in L^1(R^1) \cap L^\infty(R^1) \cap BV(R^1)$, *and*

$$\frac{\partial u_0}{\partial x} = 0, \qquad \text{if} \quad |x| > B,$$

Proof. We only need to prove that the limit solution u satisfies inequality

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} \leqslant 0,$$

in the weak sense, where

$$\begin{aligned}U(u) &= \operatorname{sign}(u - c)(u - c), \\ F(u) &= \operatorname{sign}(u - c)(f(u) - f(c)).\end{aligned}$$

Define

$$c \vee u = \max\{c, u\}, \qquad c \wedge u = \min\{c, u\}$$

and

$$\begin{aligned}H_{j+1/2} = H(u_j, u_{j+1}) &= h(c \vee u_j, c \vee u_{j+1}) \\ &\quad - h(c \wedge u_j, c \wedge u_{j+1});\end{aligned}$$

then we have

$$H(u, u) = F(u) = \operatorname{sign}(u - c)(f(u) - f(c)).$$

Let

$$\tilde{u}_j = u_j^n - \lambda(h^n_{j+1/2} - h^n_{j-1/2}).$$

Since scheme (3.2) is a monotone scheme, thus

$$\frac{1}{\Delta t}\,(U(\tilde{u}_j) - U(u_j^n)) + \frac{1}{\Delta x}\,(H_{j+1/2} - H_{j-1/2}) \leqslant 0.$$

See [11] for further details.

Choose a test function $\varphi(x, t) \in C_0^\infty$ and $\varphi \geqslant 0$; we have

$$\begin{aligned}
&\sum_{j,n} \varphi_j^n \left[\frac{U(u_j^{n+1}) - U(u_j^n)}{\Delta t} + \frac{H_{j+1/2} - H_{j-1/2}}{\Delta x} \right] \Delta t\, \Delta x \\
&\quad = \sum_{j,n} \varphi_j^n \left[\frac{U(u_j^{n+1}) - U(\tilde{u}_j^n)}{\Delta t} + \frac{U(\tilde{u}_j^n) - U(u_j^n)}{\Delta t} \right. \\
&\qquad \left. + \frac{H_{j+1/2} - H_{j-1/2}}{\Delta x} \right] \Delta t\, \Delta x \\
&\quad \leqslant \sum_{j,n} \varphi_j^n \frac{U(u_j^{n+1}) - U(\tilde{u}_j^n)}{\Delta t} \Delta t\, \Delta x \\
&\quad \leqslant \sum_{j,n} \varphi_j^n \frac{|Q_j|}{\Delta t} \Delta t\, \Delta x,
\end{aligned}$$

so

$$\begin{aligned}
&-\sum_{j,n} \left[U(u_j^{n+1}) \frac{\varphi_j^{n+1} - \varphi_j^n}{\Delta t} + H_{j+1/2} \frac{\varphi_{j+1}^n - \varphi_j^n}{\Delta x} \right] \Delta x\, \Delta t \\
&\quad \leqslant \sum_{j,n} \varphi_j^n \frac{|Q_j|}{\Delta t} \Delta t\, \Delta x.
\end{aligned}$$

Similar to the proof of Theorem 3.1, we have

$$\sum_{j,n} \varphi_j^n \frac{|Q_j|}{\Delta t} \Delta x\, \Delta x \to 0 \qquad (\text{as } \Delta t \to 0,\ \Delta x \to 0).$$

Thus

$$-\iint_{[0,T] \times R} \left[U(u) \frac{\partial \varphi}{\partial t} + F(u) \frac{\partial \varphi}{\partial x} \right] dx\, dt \leqslant 0;$$

i.e.

$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} \leqslant 0.$$

in the weak sense. ∎

4. THE SYSTEM OF HYPERBOLIC CONSERVATION LAWS

Consider the system of conservation laws,

$$\begin{aligned}
&\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0, \\
&\qquad u = (u_1, \ldots, u_m)^{\mathrm{T}}, \qquad f(u) = (f_1, \ldots, f_m)^{\mathrm{T}}.
\end{aligned} \tag{4.1}$$

We now extend the essentially conservative scheme to approximate Eq. (4.1). Suppose that $v(u_j, u_{j+1})$ is some kind of averaging, such as Roe's averaging [9], of u_j and u_{j+1}, $A(v) = \partial f/\partial u|_{u=v}$ is the Jacobian matrix, $\alpha_{j+1/2}^k$ and $R_{j+1/2}^k$ are the kth eigenvalue and the eigenvector of $A(v)$, respectively. Let $d_{j+1/2}^k$ be the kth component of $\Delta_{j+1/2}^n u = u_{j+1}^n - u_j^n$ along $R_{j+1/2}^k$; i.e.,

$$\Delta_{j+1/2}^n u = \sum_{k=1}^m d_{j+1/2}^k R_{j+1/2}^k,$$

then the first-order scheme can be written as

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2}^n - h_{j-1/2}^n), \tag{4.2}$$

$$h_{j+1/2}^n = \tfrac{1}{2}\left(f_j^n + f_{j+1}^n - \sum_{k=1}^m |a_{j+1/2}^k|\, d_{j+1/2}^k R_{j+1/2}^k \right), \tag{4.3}$$

where $\lambda = \Delta t/\Delta x$. The difference between (4.3) and the Lax–Wendroff numerical flux is

$$\phi_{j+1/2} = \tfrac{1}{2} \sum_{k=1}^m |a_{j+1/2}^k| \,(1 - \lambda\, |a_{j+1/2}^k|)\, d_{j+1/2}^k R_{j+1/2}^k. \tag{4.4}$$

Then we can construct a second-order essentially conservative approximation of Eq. (4.1) as

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2}^n - h_{j-1/2}^n) + \lambda Q_j, \tag{4.5}$$

$$Q_j = (Q_j^1, \ldots, Q_j^m)^{\mathrm{T}}, \tag{4.6}$$

$$Q_j^k = -\min\left\{ M \max\left(1, \frac{|d_{j+1/2}^k|}{\Delta x}, \frac{|d_{j-1/2}^k|}{\Delta x}\right) \times \Delta x^2, |\phi_{j+1/2}^k - \phi_{j-1/2}^k| \right\} S_j^k, \tag{4.7}$$

$S_j^k = \operatorname{sign}(\phi_{j+1/2}^k - \phi_{j-1/2}^k)$, M is a positive number, $k = 1$ to m;

$\phi_{j+1/2}^k$ is kth component of $\tilde{\phi}_{j+1/2}$, where

$$\tilde{\phi}_{j+1/2} = \tfrac{1}{2} \sum_{k=1}^m |a_{j+1/2}^k| \,(1 - \lambda\, |a_{j+1/2}^k|)\, g_{j+1/2}^k R_{j+1/2}^k$$

$$g_{j+1/2}^k = \operatorname{minmod}(d_{j-1/2}^k + D_{j+1/2}^L, d_{j+1/2}^k, d_{j+3/2}^k - D_{j+1/2}^R)$$

$$D_{j+1/2}^L = \operatorname{minmod}(d_{j-1/2}^k - d_{j-3/2}^k, d_{j+1/2}^k - d_{j-1/2}^k)$$

$$D_{j+1/2}^R = \operatorname{minmod}(d_{j+3/2}^k - d_{j+1/2}^k, d_{j+5/2}^k - d_{j+3/2}^k)$$

Similarly as in scalar case, we can prove

THEOREM 4.1. *Scheme* (4.5)–(4.7) *is a second-order essentially conservative approximation of Eq.* (4.1).

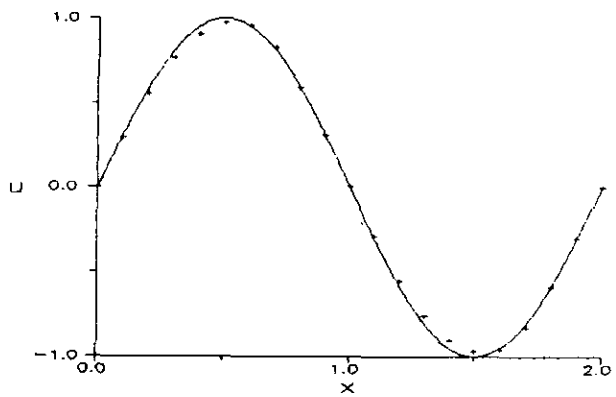

FIGURE 5.1

5. NUMERICAL RESULTS

In this section we give some numerical results to show the performance of the essentially conservative scheme constructed in this paper.

1. *Initial problem of a scalar linear equation.* We choose the following problem to show the accuracy of the scheme in smooth region of solutions:

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}=0, \qquad (x, t)\in[0, \infty)\times(-\infty, \infty), \tag{5.1}$$

$$u(x, 0)=\sin \pi(x+1), \qquad x\in(-\infty, \infty). \tag{5.2}$$

The computation is carried out in $[0, 2]$. Periodic boundary conditions are given at $x=0$ and $x=2$. Twenty-one mesh points are equally spaced in $[0, 2]$. The CFL number $=0.5$ and the constant M is chosen to be five. Figure 5.1 is the result at $t=4$, which shows that the essentially conservative scheme gives good result even at the local extreme points. The solid line is the exact solution and the symbol "+" represents the numerical solution (other figures are same).

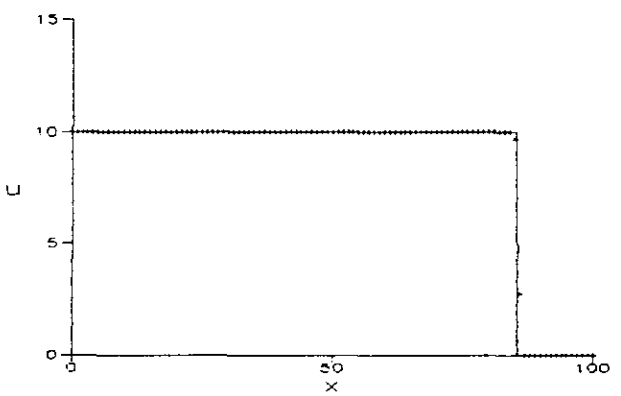

FIGURE 5.2

2. *Scalar moving shock wave problem.* The following problem shows that the essentially conservative scheme can obtain the correct shock wave position. Consider the scalar inviscid burger's equation

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=0, \qquad (x, t)\in[0, \infty)\times(-\infty, \infty), \tag{5.3}$$

$$u(x, 0)=\begin{cases}10, & \text{if } x<0\\ 0, & \text{if } x>0.\end{cases} \tag{5.4}$$

The CFL number is 0.5, the constant M is 5, and the space mesh size $\Delta x=1$. Figure 5.2 is the result at the 351th time step. From the result it can be shown that the scheme can give a very correct shock wave position even after a long computation time. When M is chosen larger than 5 the scheme becomes conservative; the numerical results not presented here are similar to Fig. 5.2.

3. *Shock wave tube problem* [10]. Consider the Euler equations of gas dynamics with a discontinuous initial value,

$$\frac{\partial u}{\partial t}+\frac{\partial f(w)}{\partial x}=0, \qquad (x, t)\in[0, \infty)\times(-\infty, \infty), \tag{5.5}$$

$$w(x, 0)=\begin{cases}w_L, & \text{if } x<0\\ w_R, & \text{if } x>0.\end{cases} \tag{5.6}$$

Where $w=(u, \rho u, \rho E)^{\mathrm{T}}, f(w)=(\rho u, p+\rho u^2, u(p+\rho E)^{\mathrm{T}}$; u, p, ρ, E are velocity, pressure, density, and total energy of gas in unit mass. We use G. A. Sod's initial value to show the performance of the scheme in shock wave capturing,

$$(p_L, \rho_L, u_L)=(1, 1, 0),$$
$$(p_R, \rho_R, u_R)=(0.1, 0.125, 0)\ [10].$$

To improve the resolution of shock waves and contact discontinuities, we introduce here an artificial compression technique. In the scalar case, the technique can be implemented as follows: apply the scheme to a modified equation of Eq. (3.1). The modified one can be written in form

$$\frac{\partial u}{\partial t}+\frac{\partial(f(u)+g(u))}{\partial x}=0, \tag{5.7}$$

where

$$g_j=\max\{0, |\operatorname{minmod}(2L_{j-1/2}, L_{j+1/2})|, |\operatorname{minmod}(L_{j-1/2}, 2L_{j+1/2})|\}\operatorname{sign}(L_{j+1/2})$$

and

$$L_{j+1/2}=\tfrac{1}{2}|a^n_{j+1/2}|\,(1-\lambda\,|a^n_{j+1/2}|)\times(\Delta_{j+1/2}u-\operatorname{minmod}(\Delta_{j-1/2}u, \Delta_{j+1/2}u, \Delta_{j+3/2}u)).$$

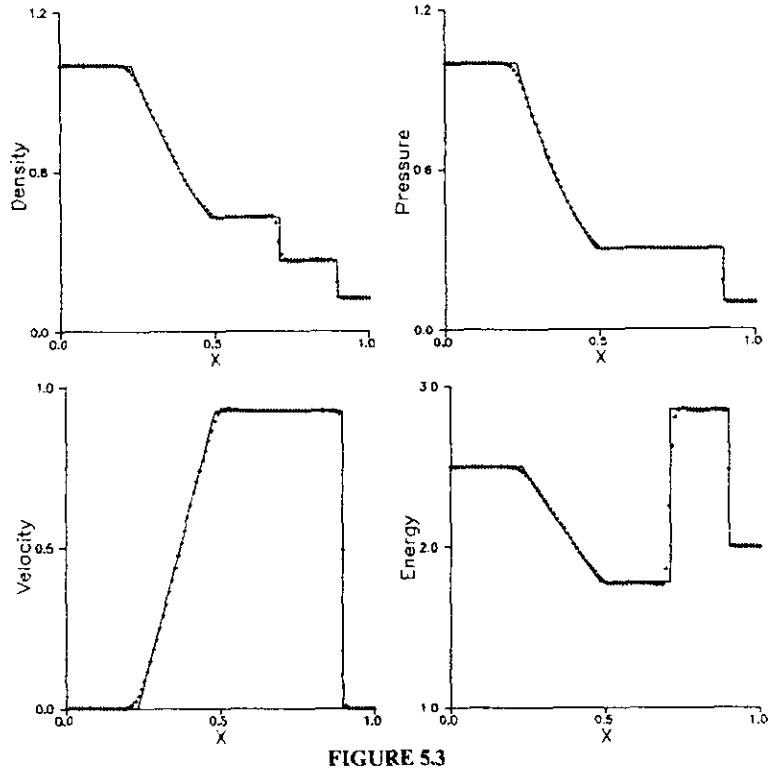

FIGURE 5.3

The artificial compression scheme can be written as

$$u_j^{n+1} = u_j^n - \lambda(h_{j+1/2} - h_{j-1/2}) + \lambda Q_j$$

$$h_{j+1/2} = \tfrac{1}{2}(f_j + f_{j+1} - |\tilde{a}_{j+1/2}|\, \Delta_{j+1/2} u)$$

$$\tilde{a}_{j+1/2} = \begin{cases} \dfrac{f_{j+1} + g_{j+1} - f_j - g_j}{u_{j+1} - u_j}, & u_j \neq u_{j+1}, \\ a(u_j), & u_j = u_{j+1}; \end{cases}$$

Q_j is same as (3.6). By applying above technique to each characteristic field, one can obtain an artificial compression scheme for system cases.

Figures 5.3a–d give the results at the 50th time step with an artificial compression scheme. 101 mesh points are equally placed in [0, 1], and the CFL number = 0.95. M is chosen to be 50. The numerical results show that the essentially conservative scheme with artificial compression gives the correct positions of shock wave and contact discontinuity and the resolution of the shock wave and contact discontinuity is also quite good.

6. CONCLUSION

The numerical results and theoretical analysis show that the essentially conservative scheme of this paper not only have good properties but also they can get good results in real computations. From the construction of the scheme, it can be seen that when M is chosen large enough for a fixed grid, the scheme tends to a second-order scheme of conservative form. So in real computations, the conservation error would not increase when M is chosen to be sufficiently large. From the numerical results, we can see that the scheme can give correct positions of discontinuities and does not generate too much oscillation even we choose a CFL number near one, which verifies the analysis in the remark of Theorem 3.2. As further research, we will consider the construction of a two-dimensional essentially conservative scheme. For two-dimensional cases, it has been proven in [4] that a second-order TVD conservative scheme does not exist. By the idea of this paper, we can construct an essentially conservative second-order TVB scheme in the two-dimensional case.

REFERENCES

1. A. Harten, *J. Comput. Phys.* **49**, 357 (1983).
2. H. Yee, *J. Comput. Phys.* **68**, 151 (1987).
3. P. K. Sweby, *SIAM J. Numer. Anal.* **21**, 995 (1984).
4. J. B. Goodman and R. J. LeVeque, *Math. Comput.* **45**, 15 (1985).
5. A. Harten and S. Osher, *SIAM J. Numer. Anal.* **24**, 279 (1987).
6. A. Harten, B. Enquist, S. Osher, and S. Chakraverthy, *J. Comput. Phys.* **71**, 231 (1987).
7. C. W. Shu, *Math. Comput.* **49**, 105 (1987).
8. B. Jin, Construction of a class uniformly second order accurate nonoscillatory schemes for hyperbolic conservation laws, *Chinese J. Numer. Math. Appl.* **13**, 79 (1991). [English]
9. P. Roe, *J. Comput. Phys.* **43**, 357 (1981).
10. G. A. Sod, *J. Comput. Phys.* **27**, 1 (1978).
11. M. Crandall and A. Majda, *Math. Comput.* **34**, 1 (1980).