TVB Boundary Treatment for Numerical Solutions of Conservation Laws

By Chi-Wang Shu

Abstract. In the computation of hyperbolic conservation laws \( u_t + f(u)_x = 0 \), TVD (total-variation-diminishing) and TVB (total-variation-bounded) schemes have been very successful for initial value problems. But most of the existing boundary treatments are only proved to be linearly stable, hence the combined initial-boundary scheme may not be TVB. In this paper we describe a procedure of boundary treatment which uses the original high-order scheme up to the boundary, plus extrapolation and upwind treatment at the boundary. The resulting scheme is proved to be TVB for the scalar nonlinear case and for linear systems.

1. Introduction. In this paper we consider the numerical solutions to the hyperbolic conservation law

\[
\begin{align*}
\frac{du}{dt} + A(u)\frac{du}{dx} &= 0, \\
\frac{d}{dx}u(x, 0) &= u_0(x).
\end{align*}
\]

Here \( u = (u_1, \ldots, u_s)^T \), and the Jacobian matrix \( A(u) = \frac{\partial f}{\partial u} \) has \( s \) real eigenvalues

\[
\lambda_1(u) \leq \lambda_2(u) \leq \cdots \leq \lambda_s(u)
\]

and a complete set of eigenvectors.

On a computational grid \( x_j = j\Delta x, \ t_n = n\Delta t \), we use \( u^n_j \) to denote the computed approximation to the exact solution \( u(x_j, t_n) \) of (1.1).

For pure initial value problems, i.e., problems with \( u_0(x) \) in (1.1b) to be either periodic or to have a compact support, the recently introduced TVD (total-variation-diminishing) and TVB (total-variation-bounded) schemes have been very successful. See, e.g., [1], [2], [3], [4], [7], and the references listed there. The total variation of a discrete scalar solution is defined by

\[
TV(u) = \sum_j |u_{j+1} - u_j|,
\]

and if

\[
TV(u^{n+1}) \leq TV(u^n)
\]

we say the scheme is TVD; while if

\[
TV(u^n) \leq B
\]
for some fixed \( B > 0 \) depending only on \( u^0 \) and \( \text{TV}(u^0) \) and all possible \( n \) and \( \Delta t \) such that \( n\Delta t \leq T \), we say that the scheme is TVB in \( 0 \leq t \leq T \). Clearly TVD implies TVB.

One major advantage of TVB schemes is that there is a convergent (in \( L^1_{\text{local}} \)) subsequence as \( \Delta x \rightarrow 0 \) to a weak solution of (1.1). If an additional entropy condition, which implies uniqueness of weak solution to (1.1), is satisfied, then the scheme is convergent. See, e.g., [2].

For initial-boundary value problems we hope the boundary treatment can still retain the TVB property of the scheme. The usual method is to use a lower-order scheme near the boundary. This not only reduces the order of the scheme near the boundary, but also makes any theoretical results about TVB of the initial-boundary scheme very hard to prove. In this paper we present an approach to the treatment of boundaries which uses the same high-order scheme up to the boundary, plus extrapolation and an upwind treatment at the boundary. The resulting scheme is proved to be TVB for the scalar nonlinear case (Section 2) and for linear systems (Section 3).

Our boundary treatment is based on the globally high-order TVB schemes discussed in [4], [6], [7], and [8]. These schemes have natural upwind-downwind decompositions which help us to implement and prove the TVB boundary treatments.

We include several numerical results in the appendix to demonstrate the usefulness of the TVB boundary treatment in Sections 2 and 3.

2. Scalar Case. In this paper we use the following three equivalent forms of an \( r \) th order (in space and time) TVB scheme (see [4], [6], [7], and [8] for details):

\[
(2.1a1) \quad u_j^{n+1} = \sum_{k=0}^{m} \left[ \alpha_k u_j^{n-k} + \beta_k \lambda \left( C_{j+1/2}^{(n-k)} \Delta u_j^{n-k} - D_{j-1/2}^{(n-k)} \Delta u_j^{n-k} \right) \right] + \text{cor}_j^n,
\]

where

\[
(2.1a2) \quad C_{j+1/2} \geq 0, \quad D_{j+1/2} \geq 0, \quad 1 - \lambda \left( C_{j+1/2} + D_{j+1/2} \right) \geq 0,
\]

\[
(2.1a3) \quad |\text{cor}_j^n| \leq B\Delta x^2.
\]

Or,

\[
(2.1b1) \quad u_j^{n+1} = \sum_{k=0}^{m} \left[ \alpha_k u_j^{n-k} + \beta_k \lambda \left( C_{j+1/2}^{(n-k)} \Delta u_j^{n-k} - D_{j-1/2}^{(n-k)} \Delta u_j^{n-k} \right) \right],
\]

where

\[
(2.1b2) \quad C_{j+1/2}^{(n-k)} \Delta u_j^{n-k} = C_{j+1/2}^{(n-k)} \Delta u_j^{n-k} + \text{cor}_j^{1^{n-k}},
\]

\[
(2.1b3) \quad D_{j-1/2}^{(n-k)} \Delta u_j^{n-k} = D_{j-1/2}^{(n-k)} \Delta u_j^{n-k} + \text{cor}_j^{2^{n-k}},
\]

with

\[
(2.1b4) \quad |\text{cor}_j^{1^{n-k}}| \leq B\Delta x^2, \quad |\text{cor}_j^{2^{n-k}}| \leq B\Delta x^2, \quad \text{cor}_j^{1^{n-k}} + \text{cor}_j^{2^{n-k}} = \text{cor}_j^{n-k}.
\]

Or

\[
(2.1c1) \quad u_j^{n+1} = \sum_{k=0}^{m} \left[ \alpha_k u_j^{n-k} + \beta_k \lambda \left( (Df^-)_j^{n-k} - (Df^+_j)^{n-k} \right) \right]
\]
with
\[
Df_j^- = \lambda \tilde{C}_{j+1/2} \Delta_+ u_j, \quad Df_j^+ = \lambda \tilde{D}_{j-1/2} \Delta_- u_j.
\]

In (2.1), \( \alpha_k \) and \( \beta_k \) are such that
\[
\alpha_k \geq 0, \quad k = 0, 1, \ldots, m,
\]
\[
\alpha_k \geq 0, \quad k = 0, 1, \ldots, m,
\]
\[
\beta_0 - \sum_{k=1}^{m} (k \alpha_k - \beta_k) = 1,
\]
\[
(-1)^l \sum_{k=1}^{m} k^{l-1}(k \alpha_k - l \beta_k) = 1, \quad l = 2, 3, \ldots, r.
\]

The CFL condition is
\[
\lambda \leq \lambda_0 \cdot \min_k \left( \frac{\alpha_k}{|\beta_k|} \right),
\]
where \( \lambda_0 \) satisfies (2.1a2).

We now consider the equation (1.1) defined on
\[
0 < x < +\infty, \quad t > 0,
\]
and still assume that \( u_0(x) \) in (1.1b) is zero in \( x > L \).

Notice that (2.1) has a more than three-point stencil in \( x \) only because of the \( C \)'s and the \( D \)'s. Near the boundary \( x = 0 \) we may not have the necessary \( u_{-1}, u_{-2}, \ldots \) etc., to define \( C \) and \( D \). In this paper we always use extrapolation with order \( r \) to get the necessary \( u_{-j} \). For example, if \( r = 2 \), we may use \( u_{-1} = 3u_0 - 3u_1 + u_2 \), etc. It is well known that extrapolation may cause instability, but due to the upwind-biased property of our scheme (2.1), we shall prove that the resulting initial-boundary scheme is still TVB. With the help of extrapolation, the boundary problem simplifies to that of determining \( u_0 \) only. From the well-known properties of the hyperbolic equation (1.1), we know that we should prescribe \( u(0, t) = g(t) \) on the boundary \( x = 0 \) if \( f'(u(0, t)) > 0 \) (corresponding to case (a) of Theorem 2.1 below), and should prescribe nothing at \( x = 0 \) if \( f'(u(0, t)) < 0 \) (corresponding to case (b) of the theorem).

**Theorem 2.1.** In the region defined by (2.7):

(a) The scheme (2.1) for \( u^n_j, j \geq 1 \), with the above-mentioned extrapolations for computing the \( C \)'s and \( D \)'s, and initial condition \( u^n_j = u(x_j, 0) \) and boundary condition
\[
u^n_0 = g(t^n),
\]
where \( g(t) \) is a function of bounded variation in \( 0 \leq t \leq T \), is TVB in \( 0 \leq t \leq T \).

(b) Define
\[
\bar{D}f^n_0 = \min(|Df^n_0|, |Df^n_0| + K\Delta t) \cdot \text{sign}(Df^n_0)
\]
in (2.1c), where \( K > 0 \) is a constant. Then the scheme (2.1), as in part (a) where \( u^{n+1}_0 \) is computed by (2.1c) with \( Df^n_0^+ \) replaced by \( \bar{D}f^n_0^+ \), is TVB in \( 0 \leq t \leq T \).
Proof. (a) We use form (2.1a):

\[
\Delta^+ u^{n+1}_j = \sum_{k=0}^{m} \alpha_k \left( 1 - \frac{\beta_k}{\alpha_k} \lambda \left( C_{j+1/2}^{(n-k)} + D_{j+1/2}^{(n-k)} \right) \right) \left( \Delta^+ u^{n-k}_j \right) + \beta_k |\lambda| C_{j+3/2}^{(n-k)} \left( \Delta^+ u^{n-k}_j \right) + \beta_k |\lambda| D_{j-1/2}^{(n-k)} \left( \Delta^+ u^{n-k}_{j+1} \right) + \text{cor}^{n+1}_j - \text{cor}^n_j
\]

for \( j = 1, 2, \ldots \)

and

\[
\Delta^+ u^{n+1}_0 = u^{n+1}_0 - g(t^{n+1})
\]

\[
= \sum_{k=0}^{m} \left[ \alpha_k u^{n-k}_1 + \beta_k |\lambda| \left(C_{j+1/2}^{(n-k)} \Delta^+ u^{n-k}_1 - D_{j+1/2}^{(n-k)} \Delta^+ u^{n-k}_1 \right) \right] + \text{cor}^{n}_1 - g(t^{n+1})
\]

\[
= \sum_{k=0}^{m} \left[ \alpha_k (u^{n-k}_1 - u^{n-k}_0) + \beta_k |\lambda| \left(C_{j+1/2}^{(n-k)} \Delta^+ u^{n-k}_1 - D_{j+1/2}^{(n-k)} \Delta^+ u^{n-k}_1 \right) \right] + \text{cor}^{n}_1 - \sum_{k=0}^{m} \alpha_k (g(t^{n+1}) - g(t^{n-k})).
\]

Hence, by (2.1a2) and the CFL condition (2.6), we have

\[
\text{TV}(u^{n+1}) = \sum_{j=0}^{\infty} \left| \Delta^+ u^{n+1}_j \right|
\]

\[
\leq \sum_{j=1}^{\infty} \sum_{k=0}^{m} \left\{ \alpha_k \left[ 1 - \frac{\beta_k}{\alpha_k} \lambda \left( C_{j+1/2}^{(n-k)} + D_{j+1/2}^{(n-k)} \right) \right] \left| \Delta^+ u^{n-k}_j \right| + \beta_k |\lambda| C_{j+3/2}^{(n-k)} \left| \Delta^+ u^{n-k}_j \right| + \beta_k |\lambda| D_{j-1/2}^{(n-k)} \left| \Delta^+ u^{n-k}_{j+1} \right| + \text{cor}^{n+1}_j + \text{cor}^n_j \right\}
\]

\[
+ \sum_{k=0}^{m} \alpha_k \left( 1 - \frac{\beta_k}{\alpha_k} \lambda D_{j/2}^{(n-k)} \right) \left| \Delta^+ u^{n-k}_0 \right| + \beta_k |\lambda| C_{j/2}^{(n-k)} \left| \Delta^+ u^{n-k}_1 \right|
\]

\[
+ \sum_{k=0}^{m} \alpha_k |g(t^{n+1}) - g(t^{n-k})|
\]

\[
\leq \sum_{j=0}^{\infty} \sum_{k=0}^{m} \alpha_k \left| \Delta^+ u^{n-k}_j \right| + 2BNAX^2 + \sum_{k=0}^{m} \alpha_k |g(t^{n+1}) - g(t^{n-k})|
\]

\[
\leq \sum_{k=0}^{m} \alpha_k \text{TV}(u^{n-k}) + \overline{B} \Delta t + (m + 1) \sum_{k=0}^{m} |g(t^{n-k+1}) - g(t^{n-k})|,
\]

where \( N = L/\Delta x, \overline{B} = 2BL/\lambda. \)

Hence clearly

\[
\text{TV}(u^n) \leq \max_{0 \leq k \leq m} \text{TV}(u^k) + \overline{B} T + (m + 1)^2 \text{TV}(g)
\]

for all \( n\Delta t \leq T. \)
(b) Similarly to part (a), we have
\[
TV(u^{n+1}) \leq \sum_{j=0}^{m} \sum_{k=0}^{m} \alpha_k |D_{u_j^{n-k}}| + \sum_{k=0}^{m} \alpha_k \left( Df_0^+ - Df_0^- \right) + 2BN\Delta x^2
\]
\[
\leq \sum_{k=0}^{m} \alpha_k TV(u^{n-k}) + \left( \bar{B} + K \right) \Delta t
\]
(see (2.9)).

Hence clearly
\[
TV(u^n) \leq \max_{0 \leq k \leq m} TV(u^k) + \left( \bar{B} + K \right) T
\]
for all \( n\Delta t \leq T \). □

**Remark 2.1.** From Definition (2.9) we see that
\[
(2.10) \quad Df_0^+ = Df_0^+
\]
if and only if
\[
(2.11) \quad |Df_0^+| \leq |Df_0^-| + K\Delta t.
\]

Since scheme (2.1) is upwind-biased, we should expect the “downwind” part \( |Df_0^-| \) to be less than the “upwind” part \( |Df_0^+| \) if \( f'(u(0, t)) < 0 \). With the help of \( K\Delta t \), we may be very safe to expect (2.11), hence (2.10), to be almost always valid. Hence, since we used extrapolation of sufficient accuracy to approximate \( u_{-j} \), the accuracy remains \( r \)th order up to the boundary. This is superior to the usual way of using a lower-order scheme with narrower stencil near the boundary.

### 3. Linear Systems

The techniques discussed in Section 2 can be generalized, via Godunov’s, Osher’s or Roe’s field-by-field decompositions, to nonlinear systems (1.1). See, e.g., [4], [5]. Even for pure initial value problems, at present there is no theory about total variation boundedness for general nonlinear systems. For a linear system
\[
(3.1) \quad u_t + Au_x = 0,
\]
where, for simplicity, \( A \) is assumed a constant matrix (we may generalize the theory to the case \( A = A(x) \)), we have a similar theory as in Theorem 2.1.

Assume \( A \) has \( s \) nonzero real eigenvalues
\[
(3.2) \quad \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{s_1} < 0 < \lambda_{s_1+1} \leq \cdots \leq \lambda_s
\]
with a complete set of right eigenvectors \( r_1, r_2, \ldots, r_s \). Then we can write
\[
(3.3a) \quad u(x, t) = \sum_{p=1}^{s} \delta^{(p)}(x, t)r_p,
\]
\[
(3.3b) \quad Au(x, t) = \sum_{p=1}^{s} \lambda_p \delta^{(p)}(x, t)r_p,
\]
where the scalar functions \( \delta^{(p)} \) satisfy the scalar conservation laws
\[
(3.4) \quad \delta_t^{(p)} + \lambda_p \delta_x^{(p)} = 0.
\]
Hence a generalization of the scalar TVB scheme (2.1) is just using (2.1) on (3.4) for each \( p \):

\[
    u_j^{n+1} = \sum_{p=1}^{s} \sum_{k=0}^{m} \left[ \alpha_k (\delta^{(p)})_{j-k}^{n-k} 
    + |\beta_k| \left( ((D\delta^{(p)})_{j-k}^{n-k} - (D\delta^{(p)})_{j-k}^{n-k}) \right) \right] r_p.
\]

(3.5)

For pure initial value problems, the scheme (3.5) decouples to \( s \) scalar schemes, hence TVB of (3.5), with TV defined by

\[
    TV(u) = \sum_{j} \sum_{p=1}^{s} \left| \delta_{j+1}^{(p)} - \delta_{j}^{(p)} \right|,
\]

(3.6)

is an immediate consequence of TVB of each of the decoupled scalar schemes.

Now we consider the initial-boundary value problem defined on (2.7).

By (3.2), on the differential equation level, it is well known that a boundary condition

\[
    \left( \delta_{j+1}(0,t) \right) = S(t) \left( \delta_{j+1}(0,t) \right) + g(t)
\]

(3.7)

is well posed.

We obtain the discrete form of (3.7) via the following procedure:

(i) Use extrapolation of order \( r \) to find the necessary \( \delta_{j+1}^{(p)} \) in order to compute \( C \)'s and \( D \)'s near \( j = 0 \);

(ii) Let

\[
    D\delta_{j+1}^{(p)} = \min \left( |D\delta_{j}^{(p)}|, (1 - \varepsilon)|D\delta_{j}^{(p)}| + K_1 \Delta t \right) \cdot \text{sign}(D\delta_{j}^{(p)}),
\]

(3.8)

where \( K_1 > 0, 0 < \varepsilon \ll 1 \) are constants and \( p = 1, 2, \ldots, s_1 \).

Notice that (i) and (ii) are similar to the scalar case treatment. By Remark 2.1 we should expect that most of the time \( D\delta_{j+1}^{(p)} = D\delta_{j}^{(p)} \).

(iii) Define

\[
    \left( \delta_{j+1}^{(s_1+1)}(0,t) \right) = S(t) \left( \delta_{j}^{(s_1+1)}(0,t) \right) + g(t),
\]

(3.9)

where \( S(t) = (S_{ij}(t)) \) is a \((s - s_1) \times s_1 \) matrix function with each of its elements being Lipschitz continuous in \( 0 \leq t \leq T \):

\[
    |S_{ij}(t_1) - S_{ij}(t_2)| \leq L|t_1 - t_2| \quad \text{for } 0 \leq t_1, t_2 \leq T
\]

(3.10)

and \( g(t) \) is a \((s - s_1) \times 1 \) vector function of bounded variation in \( 0 \leq t \leq T \).

(3.9) is just the discretization of (3.7).

Notice that the scheme (3.5) is coupled by the boundary condition (3.9).

**Theorem 3.1.** The scheme (3.5) with boundary treatment (i)--(iii) is TVB in \( 0 \leq t \leq T \).
**Proof.** Let

\[ Q = \max \left( 1, \frac{2}{\varepsilon} \max_{0 \leq k \leq m} \frac{|\beta_k|}{\alpha_k} \left\| S \right\|_1 \right) = \max \left( 1, \frac{2}{\varepsilon} \max_{0 \leq k \leq m} \frac{|\beta_k|}{\alpha_k} \max_{0 \leq i \leq T} \sum_{i=1}^{s-s_i} |S_{i, t}(t)| \right) \]

and define

\begin{equation}
\overline{TV}(u) = \sum_j \left( Q \sum_{p=1}^{s_i} + \sum_{p=s_i+1}^s \right) |\delta_j^{(p)} - \delta_j^{(p)}|.
\end{equation}

Clearly,

\begin{equation}
TV(u) \leq \overline{TV}(u) \leq Q \cdot TV(u).
\end{equation}

Since

\[(\delta^{s_1+p})^n_0 = \sum_{i=1}^{s_1} S^n_{pi}(\delta^{(i)})^n_0 + g_n(t^n),\]

we have

\[\Delta_+ (\delta^{(s_1+p)})^{n+1}_0 = (\delta^{(s_1+p)})^{n+1}_1 - (\delta^{(s_1+p)})^{n+1}_0\]

\[= \sum_{k=0}^{s_i} \alpha_k \left( \delta^{(s_1+p)} \right)^{n-k}_1 + \left| \beta_k \right| \lambda \left( (C^{(s_1+p)})^{n-k}_{1/2} \right) \Delta_+ (\delta^{(s_1+p)})^{n-k}_1 - (D^{(s_1+p)})^{n-k}_{1/2} \Delta_-(\delta^{(s_1+p)})^{n-k}_1 + \left( \sum_{i=1}^{s_i} S^n_{pi} \delta^{(i)}^{n+1}_0 + g_n(t^{n+1}) \right) + \text{cor}_n^n \]

\[= \sum_{k=0}^{m} \left( \alpha_k \left( 1 - \frac{\left| \beta_k \right|}{\alpha_k} \right) \lambda (D^{(s_1+p)})^{n-k}_{1/2} \right) \Delta_+ (\delta^{(s_1+p)})^{n-k}_0 + \left( \sum_{i=1}^{s_i} \sum_{k=0}^{m} \alpha_k \left( S^n_{pi} \delta^{(i)}^{n+1}_0 - S^n_{pi} \delta^{(i)}^{n-k}_0 \right) \right) + \left( \sum_{k=0}^{m} \alpha_k \left( g_n(t^{n+1}) - g_n(t^{n-k}) \right) \right) + \text{cor}_n^n \]

Notice that

\[S^n_{pi} \delta^{(i)}^{n+1}_0 - S^n_{pi} \delta^{(i)}^{n-k}_0 = S^n_{pi} \delta^{(i)}^{n+1}_0 - (\delta^{(i)})^{n-k}_0 + (\delta^{(i)})^{n-k}_0 \delta^{(i)}^{n+1}_0 - (\delta^{(i)})^{n-k}_0 \delta^{(i)}^{n-k}_0,\]

and, since \(u^0(x)\) is zero for large \(x\) and the numerical solution has a finite speed of propagation, we have \((\delta^{(i)})^n_j = 0\) for large enough \(J\), so

\[\sum_{i=1}^{s_i} \left| \delta^{(i)}^n_0 \right| = \sum_{i=1}^{s_i} \left| \delta^{(i)}_j^+ - \delta^{(i)}_0 \right| \leq \sum_{i=1}^{s_i} \sum_{j=0}^{s_i} \left| \delta^{(i)}_j^+ - \delta^{(i)}_j \right| \leq TV(u^n) \leq \overline{TV}(u^n).
\]

(This is the maximum principle for TVB schemes. See Remark 3.1 below.)
So, following the lines of the proof of Theorem 2.1, we have

\[
\widetilde{TV}(u^{n+1}) = \sum_{j \geq 0} \left( Q \sum_{p=1}^{s_1} + \sum_{p=s_1+1}^{s} \right) |\Delta_+ (\delta(p))_{j}^{n+1}|
\]

\[
\leq Q \sum_{p=1}^{s_1} \left[ \sum_{j \geq 0} \left( \sum_{k=0}^{m} \alpha_k |\Delta_+ (\delta(p))_j^{n-k}| \right) + K_1 \Delta t - \lambda \epsilon (C(p))_{1/2} |\Delta_+ (\delta(p))_0^{n-k}| \right] + \sum_{p-1}^{s-s_1} \left[ \sum_{j \geq 0} \left( \sum_{k=0}^{m} \alpha_k |\Delta_+ (\delta(s_1+p))_j^{n-k}| \right) + \sum_{i=1}^{s_1} \left[ \sum_{k=0}^{m} \beta_k \left| \lambda (C(i))_{1/2} |\Delta_+ (\delta(i))_0^{n-k}| \right| + (D(i))_{1/2} |\Delta_+ (\delta(i))_0^{n-k}| \right] + \sum_{k=0}^{m} \alpha_k |(\delta(i))_0^{n-k} ||S_{p_i}^{n+1} - S_{p_i}^{n-k}| \right] + \left( \sum_{k=0}^{m} \alpha_k |g_p(t^{n+1}) - g_p(t^{n-k})| \right) + 2N\Delta x^2
\]

\[
\leq \sum_{k=0}^{m} \alpha_k \widetilde{TV}(u^{n-k})
\]

\[
- \sum_{p=1}^{s_1} \sum_{k=0}^{m} \alpha_k \left[ Q \epsilon - 2 \beta_k \sum_{i=1}^{s_1} |S_{p_i}^{n+1}| \lambda (C(p))_{1/2} |\Delta_+ (\delta(p))_0^{n-k}| \right] + (m + 1) L \Delta t \sum_{k=0}^{m} \alpha_k \widetilde{TV}(u^{n-k})
\]

\[
+ (m + 1) \sum_{p=1}^{s_1} \sum_{k=0}^{m} |g_p(t^{n-k+1}) - g_p(t^{n-k})| + H\Delta t
\]

\[
\leq \sum_{k=0}^{m} \alpha_k (1 + \tilde{L}\Delta t) \widetilde{TV}(u^{n-k})
\]

\[
- \sum_{p=1}^{s_1} \sum_{k=0}^{m} |g_p(t^{n-k+1}) - g_p(t^{n-k})| + H\Delta t.
\]

Hence, clearly,

\[
\widetilde{TV}(u^n) \leq (1 + \tilde{L}\Delta t)^n \max_{0 \leq k \leq m} \widetilde{TV}(u^k)
\]

\[
+ \sum_{i=k}^{n-1} (1 + \tilde{L}\Delta t)^{n-i-i} \left[ (m + 1) \sum_{p=1}^{s_1} \sum_{k=0}^{m} |g_p(t^{i-k+1}) - g_p(t^{i-k})| + H\Delta t \right]
\]

\[
\leq e^{L_T} \max_{0 \leq k \leq m} \widetilde{TV}(u^k) + e^{L_T} [(m + 1)^2 TV(g) + HT].
\]

This, together with (3.12), proves that

\[
TV(u^n) \leq B \cdot \max_{0 \leq k \leq m} TV(u^k) + \overline{B}
\]

for some $B, \overline{B} > 0$. \qed
Remark 3.1. With the help of the conservation form, we do not need to require that \( u_0(x) \) have compact support to prove the maximum principle for a TVB scheme in the half-plane (2.7). Taking a scalar conservation TVB scheme

\[
 u_{j+1}^n = u_j^n - (h_{j+1/2}^n - h_{j-1/2}^n)
\]
as an example (the same idea certainly works for the general TVB scheme (3.5)), we have

\[
 \max_{j \geq 0} (u_j^n) - \min_{j \geq 0} (u_j^n) \leq \text{TV}(u^n) \leq B \cdot \text{TV}(u^0).
\]

Let

\[
 s_N = \frac{1}{N} \sum_{j=0}^{N-1} u_j^n;
\]

then

\[
 s_N = \frac{1}{N} \sum_{j=0}^{N-1} \left[ u_j^{n-1} - (h_{j+1/2}^{n-1} - h_{j-1/2}^{n-1}) \right]
\]

\[
 = \frac{1}{N} \sum_{j=0}^{N-1} \left[ u_j^{n-1} + h_{j-1/2}^{n-1} - h_{N-1/2}^{n-1} \right]
\]

\[
 = \cdots = \frac{1}{N} \left[ \sum_{j=0}^{n-1} u_0^j + \sum_{k=0}^{n-1} (h_{k+1/2}^k - h_{k-1/2}^k) \right].
\]

Since \( h \) is Lipschitz continuous in its arguments and we may assume (inductively) that \( u_k^i \) is bounded for \( k < n \), the sum

\[
 \sum_{k=0}^{n-1} (h_{k+1/2}^k - h_{k-1/2}^k)
\]
is bounded independent of \( N \). Hence

\[
 \lim_{N \to \infty} |s_N| = \lim_{N \to \infty} \left| \frac{1}{N} \sum_{j=0}^{N-1} u_j^0 \right| \leq \|u^0\|_{\infty}.
\]

Clearly,

\[
 \min_j (u_j^n) \leq s_N \leq \max_j (u_j^n).
\]

So

\[
 \max_j (u_j^n) \leq B \cdot \text{TV}(u^0) + \min_j (u_j^n) \leq B \cdot \text{TV}(u^0) + s_N \leq B \cdot \text{TV}(u^0) + \|u^0\|_{\infty}.
\]

Similarly,

\[
 \min_j (u_j^n) \geq \max_j (u_j^n) - B \cdot \text{TV}(u^0) \geq s_N - B \cdot \text{TV}(u^0) \geq -\left[ B \cdot \text{TV}(u^0) + \|u^0\|_{\infty} \right].
\]

Hence

\[
 \max_j |u_j^n| \leq B \cdot \text{TV}(u^0) + \|u^0\|_{\infty}.
\]
Acknowledgment. The author would like to thank Professor Stanley Osher for his valuable help and suggestions.

Appendix: Numerical Results. The numbers in this appendix are often written in exponential forms, e.g., $4.2 \times 10^{-3}$ means $4.2 \times 10^{-3}$.

All the tables are collected at the end.

Example 1. We use the boundary treatment (2.8)–(2.9) and 3-3 (third order in space and time) TVD and TVB schemes discussed in [4], [6], [7], and [8] to solve the following scalar initial-boundary value problem:

$$
\begin{align*}
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} &= 0, & -1 \leq x \leq 1, \\
u(-1, t) &= \sin \pi t, & u(x, 0) = \sin \pi x.
\end{align*}
$$

The boundary condition given is well posed.

The 3-3 TVD and TVB schemes we use are (2.1c) with

$$
\begin{align*}
m &= 3; & a_j &= \frac{16}{27}, 0, 0, \frac{11}{27}, & \beta_j &= \frac{16}{9}, 0, 0, \frac{44}{27}; \\
Df^+_{j+1/2} &= -\lambda \left[ df^+_{j+1/2} + \frac{1}{3} \Delta_-(df^+_{j+1/2})^{(+)} + \frac{1}{6} \Delta_-(df^+_{j-1/2})^{(+)} \right], \\
Df^-_{j+1/2} &= -\lambda \left[ df^-_{j+1/2} - \frac{1}{3} \Delta_+(df^-_{j+1/2})^{(-)} + \frac{1}{6} \Delta_+(df^-_{j-1/2})^{(-)} \right],
\end{align*}
$$

where

$$
\begin{align*}
df_{j+1/2} &= f(u_{j+1}) - h_{j+1/2}; & df_{j+1/2} &= h_{j+1/2} - f(u_j),
\end{align*}
$$

with

$$
\begin{align*}
h_{j+1/2} &= \frac{1}{2} \left[ f(u_{j+1}) + f(u_j) - \alpha \Delta_+ u_j \right], \\
\alpha &= \max_{\min, u \leq u \leq \max, u} |f'(u)|,
\end{align*}
$$

being the first-order monotone Lax-Friedrichs flux (we may also use any other smooth monotone flux here), and the limited quantities defined by

$$
\begin{align*}
y_{j+1/2}^{(+)} &= m \left( y_{j+1/2}, by_{j+3/2} \right); & y_{j+1/2}^{(-)} &= m \left( y_{j+1/2}, by_{j-1/2} \right).
\end{align*}
$$

Here $b = 4$, and the function “$m$” is defined by

$$
\begin{align*}
m(\alpha, \beta) &= \min \text{mod}(\alpha, \beta) = (\text{sign} \alpha) \max(0, \min(|\alpha|, \beta \text{ sign} \alpha))
\end{align*}
$$

in the TVD case, and

$$
\begin{align*}
m(\alpha, \beta) &= \min \text{mod}(\alpha, \beta + M \Delta x^2 \text{ sign} \alpha)
\end{align*}
$$

in the TVB case. (We choose $M = 50$.)

We use (2.8) at the left boundary $x = -1$ and (2.9) at the right boundary $x = 1$. The necessary extrapolations are done to third order. We use $K = 1.0$ in (2.9). The numerical results are listed in Table 1.
By comparing Table 1 with the results in [7] (same scheme for pure initial value problem), we see that the boundary treatments (2.8) and (2.9) work quite well: We get almost the same accuracy as in the pure initial value calculations.

**Example 2.** The same 3-3-TVB scheme as in Example 1 with the TVB boundary treatment (3.8)–(3.9) is applied to the problem

\[
\begin{pmatrix}
u \\ v
\end{pmatrix}_t = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_x, \quad 0 \leq x \leq 1,
\]

(3.8)

\[u(x, 0) = v(x, 0) = \sin 2\pi x,\]

\[u(0, t) = -v(0, t), \quad v(1, t) = -u(1, t).\]

We apply (3.8)–(3.9) to both boundaries. \( \varepsilon \) and \( K_1 \) in (3.8) are taken to be 1.(-2) and 1, respectively. The exact solutions of (A.8) are

\[u(x, t) = \sin 2\pi(x - t); \quad v(x, t) = \sin 2\pi(x + t).\]

We list the numerical errors at \( t = 2 \) in Table 2.

Table 2 shows that the boundary treatment (3.8)–(3.9) works very well for smooth problems.

**Example 3.** The same scheme and the same boundary treatment are applied to the same equation as in Example 2 with a discontinuous initial-boundary condition:

\[
u(x, 0) = \begin{cases} 1, & \text{if } \frac{1}{2} \leq x \leq \frac{3}{4}, \\ 0, & \text{otherwise}, \end{cases} \quad v(x, 0) \equiv 0 \quad \text{for } x \geq 0;
\]

(3.9)

\[u(0, t) = v(0, t).\]

We use the 3-3-TVD scheme (in order to see how the boundary treatment itself affects the total variation) and the 3-3-TVB scheme with \( M = 200 \) in (A.7b). For \( \Delta x = 1/20 \) the solutions at \( t = 0.5 \) and \( t = 1 \) are printed (the exact solutions have values 0 and 1 at two sides of the star*):

(i) 3-3-TVD:

\[
t = 0.5, \quad u: \quad 0.98, \quad 0.97, \quad 0.87, \quad 0.57, \quad 0.25, \quad 4.3(-2), \quad 0, \quad 0, \ldots
\]

\[
t = 0.5, \quad v: \quad 0.98, \quad 0.98, \quad 0.85, \quad 0.59, \quad 0.29, \quad 7.1(-2), \quad 8.3(-3), \quad 8.8(-4), \quad 0, \quad 0, \ldots
\]

\[
t = 1.0, \quad u: \quad -6.4(-4), \quad 1.1(-3), \quad 1.6(-3), \quad 3.4(-3), \quad 1.4(-2), \quad 9.6(-2), \quad .32, \quad *60, \quad .84, \quad .95, \quad .96, \quad .94, \quad .81, \quad 0.7(-2), \quad 7.2(-3), \quad 0, \quad 0, \ldots
\]

(ii) 3-3-TVB with \( M = 200 \):

\[
t = 0.5, \quad u: \quad 1.06, \quad 1.08, \quad 0.85, \quad 0.53, \quad 0.23, \quad 5.2(-2), \quad -1.8(-2), \quad -2.3(-2), \quad -1.0(-2), \quad -1.1(-3), \quad 1.4(-3), \quad 9.0(-4), \ldots
\]

\[
t = 0.5, \quad v: \quad 1.09, \quad 1.01, \quad 0.84, \quad 0.60, \quad 0.33, \quad 8.3(-2), \quad -4.9(-2), \quad -5.8(-2), \quad -1.4(-2), \quad 1.2(-2), \quad 7.8(-3), \ldots
\]

\[
t = 1.0, \quad u: \quad 4.2(-3), \quad -7.9(-3), \quad -5.1(-2), \quad -7.1(-2), \quad -2.1(-2), \quad .12, \quad .33, \quad *.59, \quad .84, \quad 1.03, \quad 1.11, \quad 1.02, \quad .81, \quad 0.54, \quad .28, \quad 9.7(-2), \quad -9.0(-4), \quad -3.1(-2), \quad -2.6(-2), \quad -1.2(-2)
\]

We can see that 3-3-TVD has essentially no overshoots or undershoots. This implies that the boundary treatment itself does not increase the total variation in this case. For an \( M \) as big as 200 and a \( \Delta x \) which is not too small, we still get reasonable results for the 3-3-TVB scheme.
### Table 1 (Example 1)

<table>
<thead>
<tr>
<th>$\Delta x$</th>
<th>TVD</th>
<th>$r$</th>
<th>TVB</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/10</td>
<td>6.5(-2)</td>
<td>9.4(-3)</td>
<td>1.7(-2)</td>
<td>3.5(-3)</td>
</tr>
<tr>
<td>1/20</td>
<td>2.2(-2)</td>
<td>1.59</td>
<td>1.3(-3)</td>
<td>2.80</td>
</tr>
<tr>
<td>1/40</td>
<td>7.1(-3)</td>
<td>1.61</td>
<td>2.0(-4)</td>
<td>2.75</td>
</tr>
</tbody>
</table>

$r$: numerical order

### Table 2 (Example 2)

$L_{\infty}$: $L_{\infty}$-error; $L_1$: $L_1$-error; $r$: numerical order

<table>
<thead>
<tr>
<th>$\Delta x$</th>
<th>$u$</th>
<th>$v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/10</td>
<td>6.5(-2)</td>
<td>3.6(-2)</td>
</tr>
<tr>
<td>1/20</td>
<td>1.1(-2)</td>
<td>2.53</td>
</tr>
<tr>
<td>1/40</td>
<td>1.7(-3)</td>
<td>2.72</td>
</tr>
<tr>
<td>1/80</td>
<td>2.4(-4)</td>
<td>2.83</td>
</tr>
</tbody>
</table>

$\Delta x$: numerical order

Institute for Mathematics and Its Applications
University of Minnesota
Minneapolis, Minnesota 55455

6. C. Shu, “TVD time discretization II—time dependent problems.” (Preprint.)