A SELF-ADAPTIVE THETA SCHEME USING DISCONTINUITY AWARE QUADRATURE FOR SOLVING CONSERVATION LAWS*

TODD ARBOGAST† AND CHIEH-SEN HUANG‡

Abstract. We present a *discontinuity aware quadrature* (DAQ) rule, and use it to develop implicit *self-adaptive theta* (SATh) schemes for the approximation of scalar hyperbolic conservation laws. Our SATh schemes require the solution of a system of two equations, one controlling the cell averages of the solution at the time levels, and the other controlling the space-time averages of the solution. These quantities are used within the DAQ rule to approximate the time integral of the hyperbolic flux function accurately, even when the solution may be discontinuous somewhere over the time interval. The result is a finite volume scheme using the theta time stepping method, with theta defined implicitly (or self-adaptively). Two schemes are developed, SATh-up for a monotone flux function using simple upstream stabilization, and SATh-LF using the Lax-Friedrichs numerical flux. We prove that DAQ is accurate to second order when there is a discontinuity in the solution and third order when it is smooth. We prove that SATh-up is unconditionally stable, provided that theta is set to be at least 1/2 (which means that SATh can be only first order accurate in general). We also prove that SATh-up satisfies the maximum principle and is total variation diminishing under appropriate monotonicity and boundary conditions. General flux functions require the SATh-LF scheme, so we assess its accuracy through numerical examples in one and two space dimensions. These results suggest that SATh-LF is also stable and satisfies the maximum principle (at least at reasonable CFL numbers). Compared to solutions of finite volume schemes using Crank-Nicolson and backward Euler time stepping, SATh-LF solutions often approach the accuracy of the former but without oscillation, and they are numerically less diffuse than the later.

Key words. hyperbolic transport, theta time stepping method, space-time average, numerical integration, maximum principle, TVD, numerical diffusion

AMS subject classifications. 41A55, 65D30, 65D32, 65M08, 65M12, 76M12

1. Introduction.

A hyperbolic conservation law posed on $\mathbb{R}^d$, $d \geq 1$, for the scalar function $u(\mathbf{x}, t)$ can be written in terms of the flux function $\mathbf{f}(u) \in \mathbb{R}^d$ as

$$u_t + \nabla \cdot \mathbf{f}(u) = 0, \quad u(\mathbf{x}, 0) = u^0(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \ t > 0. \tag{1.1}$$

The differential operator has hyperbolic scaling, which implies that a numerical scheme should use $\Delta t \sim \Delta x$. Often, a second (or higher) order diffusive operator is added to the left hand side, giving an advection-diffusion equation with parabolic scaling requiring $\Delta t \sim \Delta x^2$ (or worse). Numerical solution by explicit time stepping therefore requires extremely small time steps. The differential operator can be split into advection and diffusion subproblems and solved using methods tailored to each. A popular choice are the IMEX methods, which use implicit solution of the diffusive operator and explicit solution of the advective operator. However, in this paper, we take the point of view that we will use fully implicit methods, so that the problem can be solved without resorting to operator splitting. A practical advantage of the implicit

*The first author was supported in part by the U.S. National Science Foundation under grant DMS-1912735. The second author was supported in part by the Taiwan Ministry of Science and Technology under grant MOST 109-2115-M-110 -003 -MY3, the National Center for Theoretical Sciences, Taiwan, and the Multidisciplinary and Data Science Research Center of the National Sun Yat-sen University, Taiwan.

†Department of Mathematics, University of Texas, 2515 Speedway, C1200, Austin, TX 78712-1202 and Oden Institute for Computational Engineering and Sciences, University of Texas, 201 East 24th St., C0200, Austin, TX 78712-1229 (arbogast@oden.utexas.edu)

‡Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 804, Taiwan, R.O.C. (huangcs@math.nsysu.edu.tw)

approach, especially for nonlinear problems, is that the discretization parameters can be chosen based on one's desire to maintain numerical accuracy of the underlying physics, rather than the artificial concern of maintaining numerical stability.

A basic finite volume approximation of (1.1) uses backward Euler time stepping combined with upstream weighting for spatial stability (the BE scheme). Unfortunately, it is a low order accurate scheme that bestows on the solution excessive numerical diffusion, so that shocks, contact discontinuities, and steep fronts are smeared greatly over time. Nevertheless, the BE scheme is useful in many contexts. For some applications, it is the method of choice, since it is unconditionally stable and satisfies the maximum principle [16, 5] (or invariant domain property [6]). For other applications, it can be used in combination with a higher order scheme to improve the quality of the solution, for example in flux-limiter and flux corrected transport schemes [18, 14, 15, 13, 11, 2, 12].

The theta time stepping method is often seen to reduce numerical diffusion compared to the BE scheme. For parameter θ, the method blends the implicit backward Euler ($\theta = 1$) and explicit forward Euler ($\theta = 0$) time stepping. The implicit Crank-Nicolson method results when $\theta = 1/2$ (see (1.2) below). The resulting finite volume scheme can be viewed as a flux limiting method with the limiting parameter θ. However, it is only conditionally stable and violates the maximum principle when $\theta < 1$.

A fundamental difficulty with the usual approaches are that an assessment of accuracy is based on the analysis of smooth solutions. We develop in this paper a nonlinear, self-adaptive theta (SATh) scheme that varies θ based on estimating the location of the discontinuities in the solution. The price we pay is that on a cell I_i, we need to approximate both the spatial averages of the solution at time level t^{n+1}, $\bar{u}_i^{n+1}$, and the space-time average of the solution $\tilde{\bar{u}}_i^{n+1}$ (see (2.4)). In one space dimension, using upstream weighting for transport in the positive direction, the SATh scheme is

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{\Delta x}\big[f(\bar{u}_i^n) + \theta_i^{n+1}\big(f(\bar{u}_i^{n+1}) - f(\bar{u}_i^n)\big) - f(\bar{u}_{i-1}^n) - \theta_{i-1}^{n+1}\big(f(\bar{u}_{i-1}^{n+1}) - f(\bar{u}_{i-1}^n)\big)\big], \tag{1.2}$$

$$\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{2\Delta x}\big[f(\bar{u}_i^n) + (\theta_i^{n+1})^2\big(f(\bar{u}_i^{n+1}) - f(\bar{u}_i^n)\big) - f(\bar{u}_{i-1}^n) - (\theta_{i-1}^{n+1})^2\big(f(\bar{u}_{i-1}^{n+1}) - f(\bar{u}_{i-1}^n)\big)\big], \tag{1.3}$$

where, as we will see,

$$\theta_i^{n+1} = \frac{\tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n}{\bar{u}_i^{n+1} - \bar{u}_i^n}, \tag{1.4}$$

at least when the denominator does not vanish.

This θ_i^{n+1} will arise from an accurate approximation of a time integral using what we call *discontinuity aware quadrature* (DAQ), which is an approximate integration rule that respects a discontinuity in the solution, should one appear. The stability constraint $\theta_i^{n+1} \geq 1/2$ will also be necessary. This is not the first adaptive theta scheme to appear [3], but ours is unconditionally stable and satisfies the maximum principle in appropriate situations (see §5). Unlike explicit methods [4], the maximum principle is not so well characterized for implicit methods.

In the next section, we discuss the framework for finite volume schemes to set our notation. DAQ is developed in §3, and two SATh schemes are defined in §4. In §5, the upstream weighted scheme is proved to be unconditionally stable and satisfy

the maximum principle in the case of a monotone flow. After discussing two space dimensions in §6, we present in §7 numerical results designed to test the other, Lax-Friedrichs stabilized SATh scheme. We end with a summary of results, conclusions, and some open issues.

2. The finite volume framework. In a finite volume scheme, we fix a computational mesh of elements and time levels $0 = t^0 < t^1 < t^2 < \cdots$. We approximate the average of u over each mesh element E, which we write as

$$\bar{u}_E(t) = \frac{1}{|E|} \int_E u(\mathbf{x}, t)\, d\mathbf{x}, \tag{2.1}$$

where $|E|$ is the volume of E. (Later we abuse notation by using the symbol $\bar{u}_E$ for the *approximation* of this average.) One reason finite volume methods are popular is that the governing equation (1.1) directly controls $\bar{u}_E^{n+1} = \bar{u}_E(t^{n+1})$. This is usually derived by integrating the equation over $E \times [t^n, t^{n+1}]$ to see that

$$\begin{aligned} \bar{u}_E^{n+1} &= \bar{u}_E^n - \frac{1}{|E|} \int_{t^n}^{t^{n+1}} \int_E \nabla \cdot f(u(\mathbf{x}, t))\, d\mathbf{x}\, dt \\ &= \bar{u}_E^n - \frac{1}{|E|} \int_{t^n}^{t^{n+1}} \int_{\partial E} f(u(\mathbf{x}, t)) \cdot \nu\, d\sigma(\mathbf{x})\, dt. \end{aligned} \tag{2.2}$$

However, if the solution can be discontinuous, it is not so clear that this result is valid. Rather, one should return to the physics of the problem, which dictates mass conservation in the form

$$\bar{u}_E(t) = \bar{u}_E^n - \frac{1}{|E|} \int_{t^n}^{t} \int_{\partial E} f(u(\mathbf{x}, s)) \cdot \nu\, d\sigma(\mathbf{x})\, ds, \tag{2.3}$$

and restrict to $t = t^{n+1}$ to obtain (2.2).

One cannot estimate the location of a discontinuity in the solution u using only $\bar{u}_E^n$ and $\bar{u}_E^{n+1}$. One needs more information. As in multi-moment finite volume schemes [8, 7], we approximate another linear functional of the solution, namely, its space-time average defined by

$$\tilde{\bar{u}}_E^{n+1} = \frac{1}{\Delta t^{n+1} |E|} \int_{t^n}^{t^{n+1}} \int_E u(\mathbf{x}, t)\, d\mathbf{x}\, dt, \tag{2.4}$$

where $\Delta t^{n+1} = t^{n+1} - t^n$. This quantity is useful because it is controlled by the physics of mass conservation. Simply integrate (2.3) in time to see that

$$\begin{aligned} \tilde{\bar{u}}_E^{n+1} &= \frac{1}{\Delta t^{n+1}} \int_{t^n}^{t^{n+1}} \bar{u}_E(t)\, dt \\ &= \bar{u}_E^n - \frac{1}{\Delta t^{n+1} |E|} \int_{t^n}^{t^{n+1}} \int_{t^n}^{t} \int_{\partial E} f(u(\mathbf{x}, s)) \cdot \nu\, d\sigma(\mathbf{x})\, ds\, dt \\ &= \bar{u}_E^n - \frac{1}{\Delta t^{n+1} |E|} \int_{t^n}^{t^{n+1}} \int_{s}^{t^{n+1}} \int_{\partial E} f(u(\mathbf{x}, s)) \cdot \nu\, d\sigma(\mathbf{x})\, dt\, ds, \end{aligned}$$

which gives

$$\tilde{\bar{u}}_E^{n+1} = \bar{u}_E^n - \frac{1}{|E|} \int_{t^n}^{t^{n+1}} \int_{\partial E} f(u(\mathbf{x}, t)) \cdot \nu\, d\sigma(\mathbf{x})\, \frac{t^{n+1} - t}{\Delta t^{n+1}}\, dt. \tag{2.5}$$

For completeness, we remark that the governing equation formally gives (2.5) as well. To see this, multiply (1.1) by $w(t) = (t^{n+1} - t)/\Delta t^{n+1}$, integrate in space and time, and use integration by parts in time for the first term.

Hyperbolic stabilization will need to be incorporated into (2.2) and (2.5). Moreover, the time integrals will be evaluated accurately by using discontinuity aware quadrature (DAQ), which we develop next.

3. Discontinuity aware quadrature (DAQ). We begin by defining what we mean by an isolated discontinuity.

DEFINITION 3.1. *A function $v : [0, \Delta t] \to \mathbb{R}$ has a (potential)* isolated discontinuity *at $\tau \in (0, \Delta t)$ if there exist continuous functions $v_L(t)$ and $v_R(t)$ with $v_L(0) = v_R(\Delta t) = 0$ and constants v^0 and v^1 such that*

$$v(t) = \begin{cases} v^0 + v_L(t), & 0 \le t < \tau, \\ v^1 + v_R(t), & \tau < t \le \Delta t. \end{cases} \tag{3.1}$$

We consider approximate integration (quadrature) of a smooth function $g(t, v)$ over the interval $[0, \Delta t]$, where $v = v(t)$ has an isolated discontinuity at $t = \tau$ but is otherwise smooth. We use only the data

$$v^0 = v(0), \quad v^1 = v(\Delta t), \quad \tilde{v} = \frac{1}{\Delta t} \int_0^{\Delta t} v(t)\, dt. \tag{3.2}$$

For some $\tau^* \approx \tau$, we approximate

$$\int_0^{\Delta t} g(t, v(t))\, dt \approx \int_0^{\tau^*} g(t, v^0)\, dt + \int_{\tau^*}^{\Delta t} g(t, v^1)\, dt.$$

To determine τ^* we apply the same rule to the function $g(t, v) = v$ and assume equality, i.e.,

$$\Delta t\, \tilde{v} = \int_0^{\Delta t} v(t)\, dt = \tau^*\, v^0 + (\Delta t - \tau^*)\, v^1, \tag{3.3}$$

which implies that the location of the discontinuity is approximated by

$$\tau^* = \frac{v^1 - \tilde{v}}{v^1 - v^0} \Delta t, \quad \text{provided } v^1 \ne v^0. \tag{3.4}$$

DEFINITION 3.2. *Let $g(t, v)$ be a continuous function defined from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$, and let $v(t)$ satisfy the conditions for an isolated discontinuity at $\tau \in (0, \Delta t)$. With*

$$\tau^* = \begin{cases} \dfrac{v^1 - \tilde{v}}{v^1 - v^0} \Delta t & \text{if } v^1 \ne v^0, \\ \frac{1}{2} \Delta t & \text{if } v^1 = v^0, \end{cases} \tag{3.5}$$

the discontinuous aware quadrature (DAQ) *rule $Q_0^{\Delta t}(g)$ is*

$$\int_0^{\Delta t} g(t, v(t))\, dt \approx Q_0^{\Delta t}(g) = \int_0^{\tau^*} g(t, v^0)\, dt + \int_{\tau^*}^{\Delta t} g(t, v^1)\, dt. \tag{3.6}$$

We remark that although $\tau \in (0, \Delta t)$, we cannot conclude the same for τ^*. Moreover, when $v^1 = v^0$, one could take other values for τ^*. The value $\tau^* = \Delta t/2$ seems most reasonable at this stage, but later we will see that $\tau^* = 0$ may be preferred to emphasize v^1 (i.e., implicitness in the SATh scheme).

In our setting, we define

$$\theta = 1 - \frac{\tau^*}{\Delta t} = \frac{\tilde{v} - v^0}{v^1 - v^0} \tag{3.7}$$

and apply DAQ using $g(t, v) = f(v)$ to see that

$$\int_0^{\Delta t} f(v(t))\, dt \approx Q_0^{\Delta t}(f) = \big(f^0 + \theta\,(f^1 - f^0)\big)\Delta t, \tag{3.8}$$

where $f^0 = f(v^0)$ and $f^1 = f(v^1)$. Moreover, with $g(t, v) = f(v)\,(\Delta t - t)/\Delta t$, we see that

$$\int_0^{\Delta t} f(v(t))\,\frac{\Delta t - t}{\Delta t}\, dt \approx Q_0^{\Delta t}\left(f\,\frac{\Delta t - t}{\Delta t}\right) = \frac{1}{2}\big(f^0 + \theta^2\,(f^1 - f^0)\big)\Delta t. \tag{3.9}$$

3.1. Accuracy of DAQ in the case of an isolated discontinuity.

THEOREM 3.3. *Suppose that $v : \mathbb{R} \to \mathbb{R}$ satisfies the conditions for a (potential) isolated discontinuity on $[0, \Delta t]$ at $\tau \in (0, \Delta t)$. If v_L and v_R have bounded derivatives and $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is continuous, differentiable in the second argument, and $D_2 g$ is bounded, then*

$$\left| \int_0^{\Delta t} g(t, v(t))\, dt - Q_0^{\Delta t}(g) \right| \le C\Delta t^2, \tag{3.10}$$

where C depends only on the L^∞ norms of v_L', v_R', and $D_2 g$. Moreover, the result holds no matter how τ^* is defined when $v^0 = v^1$.

Proof. We first note that since $v_L(0) = 0$ and $v_R(\Delta t) = 0$, for $t \in [0, \Delta t]$,

$$|v_L(t)| = \left| \int_0^t v_L'(s)\, ds \right| \le \|v_L'\|_{L^\infty} \Delta t, \tag{3.11}$$

and similarly $|v_R(t)| \le \|v_R'\|_{L^\infty}\Delta t$. We compute the integral of g to the left side of the true discontinuity τ as

$$\int_0^\tau g(t, v(t))\, dt = \int_0^\tau g(t, v^0 + v_L(t))\, dt = \int_0^\tau g(t, v^0)\, dt + R_L,$$

where the absolute value of the remainder

$$\begin{aligned} |R_L| = \left| \int_0^\tau \big(g(t, v^0 + v_L(t)) - g(t, v^0)\big)\, dt \right| &\le \|D_2 g\|_{L^\infty} \|v_L\|_{L^\infty} \Delta t \\ &\le \|D_2 g\|_{L^\infty} \|v_L'\|_{L^\infty} \Delta t^2. \end{aligned}$$

We get a similar estimate of the integral of g to the right side of τ, namely

$$\int_\tau^{\Delta t} g(t, v(t))\, dt = \int_\tau^{\Delta t} g(t, v^1)\, dt + R_R, \quad |R_R| \le \|D_2 g\|_{L^\infty} \|v_R'\|_{L^\infty} \Delta t^2.$$

Therefore, the quadrature error is

$$
\begin{aligned}
(3.12)\qquad &\left| \int_0^{\Delta t} g(t,v(t))\,dt - \int_0^{\tau^*} g(t,v^0)\,dt - \int_{\tau^*}^{\Delta t} g(t,v^1)\,dt \right| \\
&\quad = \left| \int_0^{\tau} g(t,v^0)\,dt + R_L + \int_{\tau}^{\Delta t} g(t,v^1)\,dt + R_R \right. \\
&\qquad \left. - \int_0^{\tau^*} g(t,v^0)\,dt - \int_{\tau^*}^{\Delta t} g(t,v^1)\,dt \right| \\
&\quad = \left| \int_{\tau^*}^{\tau} \big(g(t,v^0) - g(t,v^1)\big)\,dt + R_L + R_R \right| \\
&\quad \le \|D_2 g\|_{L^\infty} |(v^0 - v^1)(\tau - \tau^*)| + \|D_2 g\|_{L^\infty} \big(\|v_L'\|_{L^\infty} + \|v_R'\|_{L^\infty}\big)\Delta t^2.
\end{aligned}
$$

It remains only to estimate $\tau - \tau^*$. We note that

$$
\Delta t\,\tilde{v} = \int_0^{\Delta t} v(t)\,dt = \tau\, v^0 + (\Delta t - \tau)\, v^1 + \int_0^{\tau} v_L(t)\,dt + \int_{\tau}^{\Delta t} v_R(t)\,dt,
$$

so using (3.11),

$$
\big|\Delta t\,\tilde{v} - \tau\, v^0 - (\Delta t - \tau)\, v^1\big| \le \big(\|v_L'\|_{L^\infty} + \|v_R'\|_{L^\infty}\big)\Delta t^2.
$$

Recalling (3.3), we conclude that

$$
\big|(\tau - \tau^*)(v^1 - v^0)\big| \le \big(\|v_L'\|_{L^\infty} + \|v_R'\|_{L^\infty}\big)\Delta t^2.
$$

Combining this with (3.12) completes the proof. ∎

We remark that the Theorem holds even in the case that v is actually continuous, i.e., $v^0 + v_L(\tau) = v^1 + v_R(\tau)$. When v and g are two times differentiable, we can improve the result.

3.2. Accuracy of DAQ in the case of smooth functions.

THEOREM 3.4. *If $v : \mathbb{R} \to \mathbb{R}$ has two bounded derivatives and $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is twice differentiable with bounded derivatives, then*

$$
(3.13)\qquad \left| \int_0^{\Delta t} g(t,v(t))\,dt - Q_0^{\Delta t}(g) \right| \le C\Delta t^3,
$$

where C depends only on the L^∞ norms of v', v'', Dg, and $D^2 g$. Moreover, the result holds no matter how τ^* is defined when $v^0 = v^1$.

Proof. We first recall that the trapezoidal rule applied to a function $\varphi(t)$ satisfies

$$
\int_0^{\Delta t} \varphi(t)\,dt = \tfrac{1}{2}\big(\varphi(0) + \varphi(\Delta t)\big)\Delta t + R_T(\varphi), \quad |R_T(\varphi)| \le \tfrac{1}{12}\|\varphi''\|_{L^\infty}\Delta t^3.
$$

Therefore the DAQ error is

$$
\begin{aligned}
E = \int_0^{\Delta t} g(t,v(t))\,dt - Q_0^{\Delta t}(g) &= \tfrac{1}{2}\big(g(0,v^0) + g(\Delta t, v^1)\big)\Delta t - Q_0^{\Delta t}(g) + R_T(g(\cdot,v)), \\
|R_T(g(\cdot,v))| &\le \tfrac{1}{12}\big\|D^2\big(g(\cdot,v(\cdot))\big)\big\|_{L^\infty}\Delta t^3.
\end{aligned}
$$

If $v^0 = v^1$, then E simplifies independently of the value of τ^* to

$$\begin{aligned} E &= \tfrac{1}{2}\big(g(0,v^0) + g(\Delta t, v^0)\big)\Delta t - Q_0^{\Delta t}(g(\cdot,v^0)) + R_T(g(\cdot,v)) \\ &= \tfrac{1}{2}\big(g(0,v^0) + g(\Delta t, v^0)\big)\Delta t - \int_0^{\Delta t} g(t,v^0)\,dt + R_T(g(\cdot,v)) \\ &= R_T(g(\cdot,v^0)) + R_T(g(\cdot,v)), \end{aligned}$$

and we conclude the bound stated in the theorem.

If $v^0 \neq v^1$, we compute

$$\begin{aligned} E &= \tfrac{1}{2}\big(g(0,v^0) + g(\Delta t, v^1)\big)\Delta t - \int_0^{\tau^*} g(t,v^0)\,dt - \int_{\tau^*}^{\Delta t} g(t,v^1)\,dt + R_T(g(\cdot,v)) \\ &= \int_0^{\Delta t/2} \big(g(0,v^0) - g(t,v^0)\big)\,dt + \int_{\Delta t/2}^{\Delta t} \big(g(\Delta t,v^1) - g(t,v^1)\big)\,dt \\ &\quad - \int_{\Delta t/2}^{\tau^*} \big(g(t,v^0) - g(t,v^1)\big)\,dt + R_T(g(\cdot,v)) \\ &= E_1 + E_2 + E_3 + R_T(g(\cdot,v)), \end{aligned}$$

respectively.

We first estimate $E_1 + E_2$ by computing

$$\begin{aligned} E_1 + E_2 &= \int_0^{\Delta t/2} \big(g(0,v^0) - g(t,v^0)\big)\,dt + \int_{\Delta t/2}^{\Delta t} \big(g(\Delta t,v^1) - g(t,v^1)\big)\,dt \\ &= -\int_0^{\Delta t/2}\int_0^t D_1 g(s,v^0)\,ds\,dt + \int_{\Delta t/2}^{\Delta t}\int_t^{\Delta t} D_1 g(s,v^1)\,ds\,dt \\ &= \int_0^{\Delta t/2}\int_0^t \big(D_1 g(\Delta t - s, v^1) - D_1 g(s,v^0)\big)\,ds\,dt, \end{aligned}$$

using the change of variables $\hat{t} = \Delta t - t$ and $\hat{s} = \Delta t - s$ on the second integral of the middle line (and replacing $\hat{t}$ by t and $\hat{s}$ by s). The mean value theorem in two dimensions then gives

$$\begin{aligned} |E_1 + E_2| &\le \Big| \int_0^{\Delta t/2}\int_0^t \|DD_1 g\|_{L^\infty}\big(|\Delta t - 2s| + |v^1 - v^0|\big)\,ds\,dt \Big| \\ &\le \tfrac{1}{2}\|D^2 g\|_{L^\infty}\big(1 + \|v'\|_{L^\infty}\big)\Delta t^3. \end{aligned}$$

Now for E_3, we see that

$$|E_3| = \Big| \int_{\Delta t/2}^{\tau^*} \big(g(t,v^0) - g(t,v^1)\big)\,dt \Big| \le \|D_2 g\|_{L^\infty} |(v^0 - v^1)(\tau^* - \Delta t/2)|.$$

Moreover,

$$\begin{aligned} \big(v^0 - v^1)(\tau^* - \tfrac{1}{2}\Delta t\big) &= (v^0 - v^1)\Big(\frac{v^1 - \tilde{v}}{v^1 - v^0} - \frac{1}{2}\Big)\Delta t \\ &= \big(\tilde{v} - \tfrac{1}{2}(v^0 + v^1)\big)\Delta t = R_T(v), \end{aligned} \tag{3.14}$$

since this is a trapezoidal approximation of $\tilde{v}$. Thus

$$|E_3| \le \tfrac{1}{12}\|D_2 g\|_{L^\infty}\|v''\|_{L^\infty}\Delta t^3,$$

and the proof is complete. ∎

We remark that when the solution is smooth, (3.14) implies that

$$\tau^* = \tfrac{1}{2}\Delta t + \frac{\tilde{v} - \frac{1}{2}(v^0+v^1)}{v^0 - v^1}\Delta t. \tag{3.15}$$

The absolute value of the deviation of τ^* from $\Delta t/2$ could be quite large, for example at a minimum or maximum in v, where $\tilde{v} \ne \frac{1}{2}(v^0+v^1)$ but v^0 could be arbitrarily close to v^1.

4. Derivation of self-adaptive theta (SATh) schemes in one space dimension. We restrict to one space dimension; that is, to the governing equation

$$u_t + \big(f(u)\big)_x = 0, \quad x \in \mathbb{R},\ t > 0. \tag{4.1}$$

Our computational mesh is defined by grid points $\cdots < x_{i-1/2} < x_{i+1/2} < x_{i+3/2} < \cdots$ and elements (or grid cells) $I_i = [x_{i-1/2}, x_{i+1/2}]$. For simplicity, we replace subscript E by i, rather than I_i.

We introduce the numerical flux function $\hat{f}$ and restrict (2.2) and (2.5) to one space dimension. Denoting $\hat{f}_{i+1/2} = \hat{f}|_{x_{i+1/2}}$, the result is

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{1}{\Delta x_i}\int_{t^n}^{t^{n+1}} \big(\hat{f}_{i+1/2} - \hat{f}_{i-1/2}\big)\,dt, \tag{4.2}$$

$$\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n - \frac{1}{\Delta x_i}\int_{t^n}^{t^{n+1}} \big(\hat{f}_{i+1/2} - \hat{f}_{i-1/2}\big)\frac{t^{n+1}-t}{\Delta t^{n+1}}\,dt. \tag{4.3}$$

4.1. The SATh scheme using upstream weighting. When the flux function f is monotone in u, say $f'(u) > 0$, then we can use simple one point upstream weighting stabilization, i.e., $\hat{f}_{i+1/2} = \bar{f}_i = f(\bar{u}_i)$. We apply DAQ to the integrals in (4.2)–(4.3) to obtain the self-adaptive theta upstream weighted (SATh-up) scheme

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{\Delta x_i}\big[\bar{f}_i^n + \theta_i^{n+1}(\bar{f}_i^{n+1} - \bar{f}_i^n) - \bar{f}_{i-1}^n - \theta_{i-1}^{n+1}(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\big], \tag{4.4}$$

$$\begin{aligned}\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{2\Delta x_i}\big[&\bar{f}_i^n + (\theta_i^{n+1})^2(\bar{f}_i^{n+1} - \bar{f}_i^n)\\ &- \bar{f}_{i-1}^n - (\theta_{i-1}^{n+1})^2(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\big],\end{aligned} \tag{4.5}$$

where, for some $\epsilon \ge 0$,

$$\theta_i^{n+1} = \begin{cases} \max\left(\dfrac{1}{2}, \dfrac{\tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n}{\bar{u}_i^{n+1} - \bar{u}_i^n}\right) & \text{if } |\bar{u}_i^{n+1} - \bar{u}_i^n| > \epsilon, \\ \theta^* & \text{if } |\bar{u}_i^{n+1} - \bar{u}_i^n| \le \epsilon. \end{cases} \tag{4.6}$$

The restriction $\theta_i^{n+1} \ge 1/2$ will be explained in §5.1. We take ϵ very small (even zero) and $\theta^* = 1$ (backward Euler) or possibly $\theta^* = 1/2$ (Crank-Nicolson). We will discuss these issues in §7 on numerical results.

4.2. The SATh scheme using Lax-Friedrichs stabilization. For a general flux function f, we can use Lax-Friedrichs stabilization, posed in terms of the maximum wave speed

$$\alpha_{\mathrm{LF}} = \max_u |f'(u)|. \tag{4.7}$$

The numerical flux is

$$\hat{f}(u^-, u^+) = \frac{1}{2}\big[f(u^-) + f(u^+) - \alpha_{\mathrm{LF}}(u^+ - u^-)\big], \tag{4.8}$$

where at a point in space, u^- and u^+ are left and right limits of the solution (allowing for discontinuities). We use simple one point upstream weighting to define these quantities, so

$$u^-_{i+1/2} = \bar{u}_i \quad \text{and} \quad u^+_{i+1/2} = \bar{u}_{i+1}. \tag{4.9}$$

In this case, we approximate (4.2)–(4.3) by applying the DAQ integration formulas (3.8)–(3.9) to obtain the self-adaptive theta Lax-Friedrichs (SATh-LF) scheme

$$\begin{aligned}\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{2\Delta x_i}\Big\{ & \\ & \bar{f}^n_{i+1} + \theta^{n+1}_{i+1}(\bar{f}^{n+1}_{i+1} - \bar{f}^n_{i+1}) - \bar{f}^n_{i-1} - \theta^{n+1}_{i-1}(\bar{f}^{n+1}_{i-1} - \bar{f}^n_{i-1}) \\ & - \alpha_{\mathrm{LF}}\big[\bar{u}^n_{i+1} + \theta^{n+1}_{i+1}(\bar{u}^{n+1}_{i+1} - \bar{u}^n_{i+1}) - 2\bar{u}^n_i - 2\theta^{n+1}_i(\bar{u}^{n+1}_i - \bar{u}^n_i) \\ & \qquad + \bar{u}^n_{i-1} + \theta^{n+1}_{i-1}(\bar{u}^{n+1}_{i-1} - \bar{u}^n_{i-1})\big]\Big\},\end{aligned} \tag{4.10}$$

$$\begin{aligned}\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{4\Delta x_i}\Big\{ & \\ & \bar{f}^n_{i+1} + (\theta^{n+1}_{i+1})^2(\bar{f}^{n+1}_{i+1} - \bar{f}^n_{i+1}) - \bar{f}^n_{i-1} - (\theta^{n+1}_{i-1})^2(\bar{f}^{n+1}_{i-1} - \bar{f}^n_{i-1}) \\ & - \alpha_{\mathrm{LF}}\big[\bar{u}^n_{i+1} + (\theta^{n+1}_{i+1})^2(\bar{u}^{n+1}_{i+1} - \bar{u}^n_{i+1}) - 2\bar{u}^n_i - 2(\theta^{n+1}_i)^2(\bar{u}^{n+1}_i - \bar{u}^n_i) \\ & \qquad + \bar{u}^n_{i-1} + (\theta^{n+1}_{i-1})^2(\bar{u}^{n+1}_{i-1} - \bar{u}^n_{i-1})\big]\Big\},\end{aligned} \tag{4.11}$$

where (4.6) defines θ_i^{n+1}.

5. Properties of the upstream weighted scheme (SATh-up). Analysis of the scheme using Lax-Friedrichs stabilization (4.10)–(4.11) is complicated by the fact that waves can move in both directions. However, the upstream weighted scheme (4.4)–(4.5) is amenable to a more straightforward analysis. We analyze the upstream weighted scheme under a mild monotonicity condition on the flux function f, namely, that $f(0) = 0$ and $f'(u) > 0$ for $u \neq 0$.

5.1. Nonlinear stability for a monotone flux. We show that the upstream weighted scheme (4.4)–(4.5) is stable, provided only that one sets a lower bound on $\theta_i^{n+1} \geq 1/2$. That is, it is unconditionally stable in terms of the discretization parameters. Both the backward Euler method and the Crank-Nicolson method are stable, so the proof does not require that we analyze carefully the definition of θ_i^{n+1}.

THEOREM 5.1. *Assume that $f(0) = 0$ and $f'(u) \geq 0$ for all $u \in \mathbb{R}$ (but not identically zero). If the SATh-up scheme (4.4)–(4.5) is posed on a finite interval with a boundary condition imposed on the left, then the scheme is stable provided that*

$$\theta_i^{n+1} \geq \frac{1}{2} - \frac{\Delta x_i}{\alpha_{\mathrm{LF}}\Delta t^{n+1}},$$

where the maximum wave speed α_{LF} *is defined in* (4.7)*. Moreover, the scheme is unconditionally stable and independent of all problem parameters provided that* $\theta_i^{n+1} \geq 1/2$.

Proof. Let

$$\delta_i^n = \begin{cases} \dfrac{\bar{f}_i^n}{\alpha_{\mathrm{LF}} \bar{u}_i^n} & \text{if } \bar{u}_i^n \neq 0, \\ 0 & \text{if } \bar{u}_i^n = 0, \end{cases}$$

and note that $\delta_i^n \in [0,1]$, since $\delta_i^n = \dfrac{f(\bar{u}_i^n) - f(0)}{\alpha_{\mathrm{LF}}(\bar{u}_i^n - 0)}$ when $u_i^n \neq 0$. The scheme (4.4)–(4.5) can be expanded to three equations in three unknowns $\bar{u}_i^{n+1}$, $\tilde{\bar{u}}_i^{n+1}$, and $\bar{f}_i^{n+1}$ by including the equation

$$\bar{f}_i^{n+1} = \alpha_{\mathrm{LF}}\, \delta_i^{n+1}\, \bar{u}_i^{n+1}. \tag{5.1}$$

In this expanded form, we may view the scheme in terms of the variables

$$\xi^n = (\ldots, \bar{u}_{i-1}^n, \bar{f}_{i-1}^n, \tilde{\bar{u}}_{i-1}^n, \bar{u}_i^n, \bar{f}_i^n, \tilde{\bar{u}}_i^n, \ldots)^T \tag{5.2}$$

as $A\xi^{n+1} = B\xi^n + b^{n+1}$, where b^{n+1} represents the boundary condition. The matrices A and B are block 3×3 lower triangular. The eigenvalues of the matrix $A^{-1}B$ are the eigenvalues of $A_d^{-1}B_d$, where A_d and B_d are the diagonal blocks. In terms of $\hat{\lambda}_i = \Delta t^{n+1}/\Delta x_i$, the ith blocks are

$$A_d = \begin{bmatrix} 1 & \hat{\lambda}_i \theta_i^{n+1} & 0 \\ -\alpha_{\mathrm{LF}} \delta_i^{n+1} & 1 & 0 \\ 0 & \hat{\lambda}_i (\theta_i^{n+1})^2/2 & 1 \end{bmatrix}, \quad B_d = \begin{bmatrix} 1 & -\hat{\lambda}_i(1 - \theta_i^{n+1}) & 0 \\ 0 & 0 & 0 \\ 1 & -\hat{\lambda}_i\big(1 - (\theta_i^{n+1})^2\big)/2 & 0 \end{bmatrix}. \tag{5.3}$$

It is not difficult to compute the eigenvalues of $A_d^{-1}B_d$, and they are 0, 0, and

$$\frac{1 - \alpha_{\mathrm{LF}} \hat{\lambda}_i (1 - \theta_i^{n+1}) \delta_i^{n+1}}{1 + \alpha_{\mathrm{LF}} \hat{\lambda}_i \theta_i^{n+1} \delta_i^{n+1}}, \tag{5.4}$$

provided that the denominator does not vanish (the denominator is $\det A_d$, so the condition is simply that A_d is invertible). The scheme is stable to rounding error provided that the absolute values of the eigenvalues are bounded by 1; that is, if $1 + \alpha_{\mathrm{LF}} \hat{\lambda}_i \theta_i^{n+1} \delta_i^{n+1} > 0$ and

$$-1 - \alpha_{\mathrm{LF}} \hat{\lambda}_i \theta_i^{n+1} \delta_i^{n+1} \leq 1 - \alpha_{\mathrm{LF}} \hat{\lambda}_i (1 - \theta_i^{n+1}) \delta_i^{n+1} \leq 1 + \alpha_{\mathrm{LF}} \hat{\lambda}_i \theta_i^{n+1} \delta_i^{n+1}.$$

The upper bound holds trivially, and the lower bound holds if and only if $\delta_i^{n+1} = 0$ (and θ_i^{n+1} is unconstrained) or

$$\theta_i^{n+1} \geq \frac{\alpha_{\mathrm{LF}} \hat{\lambda}_i \delta_i^{n+1} - 2}{2 \alpha_{\mathrm{LF}} \hat{\lambda}_i \delta_i^{n+1}} = \frac{1}{2} - \frac{1}{\alpha_{\mathrm{LF}} \hat{\lambda}_i \delta_i^{n+1}}.$$

In either case, $1 + \alpha_{\mathrm{LF}} \hat{\lambda}_i \theta_i^{n+1} \delta_i^{n+1} > 0$. Since $\delta_i^{n+1} \in [0,1]$, the proof is complete. ∎

5.2. Satisfaction of the maximum principle and TVB/TVD property in a monotone setting. The maximum principle does not hold for the Crank-Nicolson method, so we must analyze carefully the way in which θ_i^{n+1} is set within the overall scheme. For the analysis, we need to assume a monotone flow, as would occur for a Riemann shock or rarefaction problem.

THEOREM 5.2. *Assume that f is strictly monotone increasing and $\epsilon = 0$ in (4.6), and that the SATh-up scheme (4.4)–(4.5), (4.6) is posed on a finite interval with a boundary condition imposed on the left (so $\bar{u}_0^n$ is given for all n). If the boundary and initial conditions of the flow satisfy the monotone decreasing property*

$$\bar{u}_0^n \le \bar{u}_0^{n+1} \quad \forall n \ge 0 \quad \textit{and} \quad \bar{u}_i^0 \le \bar{u}_{i-1}^0 \quad \forall i \ge 1, \tag{5.5}$$

then the scheme satisfies the maximum principle in the sense that

$$\bar{u}_i^n \le \bar{u}_i^{n+1} \le \bar{u}_{i-1}^{n+1} \quad \forall n \ge 0,\ i \ge 1. \tag{5.6}$$

Moreover, if the inequalities involving $\bar{u}$ are reversed, so that

$$\bar{u}_0^n \ge \bar{u}_0^{n+1} \quad \forall n \ge 0 \quad \textit{and} \quad \bar{u}_i^0 \ge \bar{u}_{i-1}^0 \quad \forall i \ge 1, \tag{5.7}$$

then

$$\bar{u}_i^n \ge \bar{u}_i^{n+1} \ge \bar{u}_{i-1}^{n+1} \quad \forall n \ge 0,\ i \ge 1. \tag{5.8}$$

Proof. We prove the theorem by an inductive argument. The result (5.6) holds initially where $(i,n) = (i,-1)\ \forall i \ge 1$ and on the boundary where $(i,n) = (0,n)\ \forall n \ge 0$, provided we define $\bar{u}_i^{-1} = \bar{u}_i^0$ and $\bar{u}_{-1}^{n+1} = \bar{u}_0^{n+1}$. We need to show that if it holds for $(i,n-1)$ and $(i-1,n)$, then it also holds for (i,n), which will give the result for all i and n. To be specific, for fixed $i \ge 1$ and $n \ge 0$, we make the induction hypothesis

$$\bar{u}_i^n \le \bar{u}_{i-1}^n \quad \text{and} \quad \bar{u}_{i-1}^n \le \bar{u}_{i-1}^{n+1} \quad (\text{i.e., } \bar{u}_i^n \le \bar{u}_{i-1}^n \le \bar{u}_{i-1}^{n+1}), \tag{5.9}$$

and we show (5.6) for the same i and n. By (strict) monotonicity of f, we also have

$$\bar{f}_i^n \le \bar{f}_{i-1}^n \le \bar{f}_{i-1}^{n+1}. \tag{5.10}$$

In the case that $\bar{u}_i^{n+1} = \bar{u}_i^n$, it is trivial to check that the induction continues. So we consider the case when $\bar{u}_i^{n+1} \ne \bar{u}_i^n$. To handle the nonlinearity in f, we define

$$\delta_i = \frac{\bar{f}_i^{n+1} - \bar{f}_i^n}{\bar{u}_i^{n+1} - \bar{u}_i^n} > 0,$$

suppressing the index n. To handle the lower bound on θ_i^{n+1} in (4.6), we define

$$\eta_i = \theta_i^{n+1} - \frac{\tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n}{\bar{u}_i^{n+1} - \bar{u}_i^n} \ge 0, \tag{5.11}$$

so that $\tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n = w_i(\theta_i^{n+1} - \eta_i)$, where we find it convenient to define $w_i = \bar{u}_i^{n+1} - \bar{u}_i^n$, and also $\hat{\lambda} = \Delta t^{n+1}/\Delta x_i$. Then (4.4)–(4.5) can be written as

$$(1 + \hat{\lambda}\theta_i^{n+1}\delta_i)w_i = -\hat{\lambda}\big[(\bar{f}_i^n - \bar{f}_{i-1}^n) - \theta_{i-1}^{n+1}(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\big] = -\hat{\lambda}A, \tag{5.12}$$

$$\begin{aligned} 2w_i\big[\theta_i^{n+1} - \eta_i + (\hat{\lambda}/2)(\theta_i^{n+1})^2\delta_i\big] & \\ = -\hat{\lambda}\big[(\bar{f}_i^n - \bar{f}_{i-1}^n) - (\theta_{i-1}^{n+1})^2(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\big] & \quad = -\hat{\lambda}B. \end{aligned} \tag{5.13}$$

By (5.10), we conclude that $A \le 0$ and $B \le 0$, and in fact $A = 0$ if and only if $B = 0$. Now (5.12) implies that $w_i \ge 0$, hence $w_i > 0$. That is, $\bar{u}_i^n < \bar{u}_i^{n+1}$, which gives half of what must be shown in (5.6).

Substitute w_i from (5.12) into (5.13) to obtain that

$$2\big[\theta_i^{n+1} - \eta_i + (\hat{\lambda}/2)(\theta_i^{n+1})^2\delta_i\big]A = (1 + \hat{\lambda}\theta_i^{n+1}\delta_i)B.$$

This is a quadratic equation in θ_i^{n+1},

$$(\hat{\lambda}A\delta_i)(\theta_i^{n+1})^2 + (2A - \hat{\lambda}B\delta_i)\theta_i^{n+1} - (2A\eta_i + B) = 0.$$

If $A = 0$, then (5.12) implies that $w_i = 0$. But we are working in the case where $w_i \ne 0$, so we conclude that $A < 0$, and so also $B < 0$. Since $\delta_i > 0$, the equation for θ_i^{n+1} is strictly quadratic. The two solutions are

$$\begin{aligned}\theta_i^{n+1} &= \frac{1}{2\hat{\lambda}A\delta_i}\Big[-2A + \hat{\lambda}B\delta_i \pm \sqrt{(2A - \hat{\lambda}B\delta_i)^2 + 4\hat{\lambda}A\delta_i(2A\eta_i + B)}\Big] \\ &= \frac{1}{2\hat{\lambda}A\delta_i}\Big[-2A + \hat{\lambda}B\delta_i \pm \sqrt{(2A)^2 + (\hat{\lambda}B\delta_i)^2 + 8\hat{\lambda}A^2\delta_i\eta_i}\Big].\end{aligned}$$

The solution which adds the square root would yield a negative θ_i^{n+1}, but $\theta_i^{n+1} \ge 1/2$, so we must take the solution which subtracts the square root. Then

$$1 + \hat{\lambda}\theta_i^{n+1}\delta_i = \frac{\hat{\lambda}B\delta_i}{2A} + \sqrt{1 + \Big(\frac{\hat{\lambda}B\delta_i}{2A}\Big)^2 + 2\hat{\lambda}\delta_i\eta_i} > \frac{\hat{\lambda}B\delta_i}{A} > 0. \tag{5.14}$$

Returning to (5.12), we have that

$$\begin{aligned}w_i &= -\frac{\hat{\lambda}A}{1 + \hat{\lambda}\theta_i^{n+1}\delta_i} < -\frac{A^2}{B\delta_i} \\ &= \frac{\big[\theta_{i-1}^{n+1}(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) - (\bar{f}_i^n - \bar{f}_{i-1}^n)\big]^2}{\big[(\theta_{i-1}^{n+1})^2(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) - (\bar{f}_i^n - \bar{f}_{i-1}^n)\big]\delta_i} = \frac{1}{\delta_i}g(\theta_{i-1}^{n+1}).\end{aligned}$$

The function $g(\theta) = \dfrac{(a\theta + b)^2}{a\theta^2 + b}$ in our case has $a = (\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) \ge 0$ and $b = -(\bar{f}_i^n - \bar{f}_{i-1}^n) \ge 0$, and at least one of a and b is strictly positive (since $A < 0$ and $B < 0$). If $a = 0$, $g(\theta) = b = a + b$, and otherwise the maximum of g on $[1/2, \infty)$ occurs at $\theta = 1$. The maximum is $a + b$ in either case, so

$$w_i \le \frac{1}{\delta_i}\big[(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) - (\bar{f}_i^n - \bar{f}_{i-1}^n)\big] = \frac{\bar{f}_{i-1}^{n+1} - \bar{f}_i^n}{\bar{f}_i^{n+1} - \bar{f}_i^n}w_i,$$

and we conclude that $1 \le \dfrac{\bar{f}_{i-1}^{n+1} - \bar{f}_i^n}{\bar{f}_i^{n+1} - \bar{f}_i^n}$. Since the numerator is positive by (5.10), so also is the denominator, and we conclude that

$$\bar{f}_i^{n+1} - \bar{f}_i^n \le \bar{f}_{i-1}^{n+1} - \bar{f}_i^n \quad \text{and then} \quad \bar{f}_i^{n+1} \le \bar{f}_{i-1}^{n+1}.$$

This then implies that $\bar{u}_i^{n+1} \le \bar{u}_{i-1}^{n+1}$, and the other half of (5.6) has been shown. This completes the induction.

For the reverse inequalities (5.7), $A \ge 0$ and $B \ge 0$, and an entirely similar argument gives the result (5.8). ∎

As a corollary of the proof, we have the following result.

COROLLARY 5.3. *Assume the hypotheses of Theorem 5.2 and that* $\theta^* \in [1/2, 1]$ *in (4.6). If* $\tilde{\bar{u}}_0^{n+1}$ *satisfies the monotonicity property that it lies between* $\bar{u}_0^n$ *and* $\bar{u}_0^{n+1}$ $\forall n \geq 0$, *then* $\theta_i^{n+1} \in [1/2, 1]$ $\forall n \geq 0,\ i \geq 1$. *Moreover, if* $\theta^* = 1$, *then* $\tilde{\bar{u}}_i^{n+1}$ *lies between* $\bar{u}_i^n$ *and* $\bar{u}_i^{n+1}$ $\forall n \geq 0,\ \forall i \geq 1$.

The corollary does not hold in general, but it holds in the case of a monotone flow, i.e., when either (5.5) or (5.7) holds.

Proof. We prove $\theta_i^{n+1} \leq 1$ for all $n \geq 0$ and $i \geq 1$ by induction on i. By the monotonicity assumption on $\tilde{\bar{u}}_0^{n+1}$, $\theta_0^{n+1} \leq 1$ $\forall n \geq 0$. So assume by induction that $\theta_{i-1}^{n+1} \leq 1$. The case of $\bar{u}_i^n = \bar{u}_i^{n+1}$ leads to $\theta_i^{n+1} = \theta^* \leq 1$, so consider the case $\bar{u}_i^n \neq \bar{u}_i^{n+1}$. We conclude that the A and B defined in (5.12)–(5.13) satisfy

$$\frac{B}{2A} = \frac{(\theta_{i-1}^{n+1})^2(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) - (\bar{f}_i^n - \bar{f}_{i-1}^n)}{2\big[\theta_{i-1}^{n+1}(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) - (\bar{f}_i^n - \bar{f}_{i-1}^n)\big]} \leq \frac{1}{2}$$

(recalling that $a = (\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)$ and $b = -(\bar{f}_i^n - \bar{f}_{i-1}^n)$ have the same sign and both do not vanish). We remark that if $\theta_i^{n+1} > 1/2$, then $\eta_i = 0$ in (5.11). Since we are trying to eliminate the case that $\theta_i^{n+1} > 1$, we can assume that $\eta_i = 0$. Returning to (5.14), we have that

$$\theta_i^{n+1} = \frac{B}{2A} + \sqrt{\Big(\frac{1}{\hat{\lambda}\delta_i}\Big)^2 + \Big(\frac{B}{2A}\Big)^2} - \frac{1}{\hat{\lambda}\delta_i} \leq \frac{1}{2} + \sqrt{\Big(\frac{1}{\hat{\lambda}\delta_i}\Big)^2 + \frac{1}{4}} - \frac{1}{\hat{\lambda}\delta_i} \leq 1,$$

and the induction is complete.

We now prove the result for $\tilde{\bar{u}}_i^{n+1}$. Provided that $\bar{u}_i^{n+1} \neq \bar{u}_i^n$,

$$1 \geq \theta_i^{n+1} \geq \frac{\tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n}{\bar{u}_i^{n+1} - \bar{u}_i^n}.$$

To continue, assume the monotonicity condition (5.5), so that also (5.6) holds. We conclude that $\tilde{\bar{u}}_i^{n+1} \leq \bar{u}_i^{n+1}$, one side of the bound on $\tilde{\bar{u}}_i^{n+1}$. For the other bound, suppose to the contrary that $\tilde{\bar{u}}_i^{n+1} < \bar{u}_i^n$ so that $\theta_i^{n+1} = 1/2$. Then (4.5) implies that

$$\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n - \frac{\hat{\lambda}}{2}\Big[(\bar{f}_i^n - \bar{f}_{i-1}^n) + \frac{1}{4}(\bar{f}_i^{n+1} - \bar{f}_i^n) - (\theta_{i-1}^{n+1})^2(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\Big].$$

The right side is minimized by $\theta_{i-1}^{n+1} = 1/2$, so

$$\tilde{\bar{u}}_i^{n+1} \geq \bar{u}_i^n - \frac{\hat{\lambda}}{2}\Big[\frac{1}{4}(\bar{f}_i^{n+1} - \bar{f}_{i-1}^{n+1}) + \frac{3}{4}(\bar{f}_i^n - \bar{f}_{i-1}^n)\Big] \geq \bar{u}_i^n.$$

Thus when $\bar{u}_i^{n+1} \neq \bar{u}_i^n$, we conclude that $\bar{u}_i^n \leq \tilde{\bar{u}}_i^{n+1} \leq \bar{u}_i^{n+1}$. Assuming the monotonicity condition (5.7) leads similarly to the opposite inequalities.

In case $\bar{u}_i^{n+1} = \bar{u}_i^n$, the right hand side of (4.4) reduces to $\bar{u}_i^n$. But, by assumption, $\theta_i^{n+1} = \theta^* = 1$ in this case, so the right hand side of (4.5) also reduces to $\bar{u}_i^n$, which shows that $\bar{u}_i^n = \tilde{\bar{u}}_i^{n+1} = \bar{u}_i^{n+1}$. ∎

In the monotone decreasing or increasing cases of Theorem 5.2, it is straightforward to compute the total variation (TV) of $\bar{u}^n$. In the monotone decreasing case

(5.5), it is

$$\mathrm{TV}(\bar{u}^n) = \sum_{i=1}^{\infty} |\bar{u}_{i-1}^n - \bar{u}_i^n| = \sum_{i=1}^{\infty} (\bar{u}_{i-1}^n - \bar{u}_i^n). \tag{5.15}$$

As a corollary of Theorem 5.2, we can then show that the scheme is total variation bounded (TVB) and total variation diminishing (TVD) under appropriate hypotheses.

COROLLARY 5.4. *Assume the hypotheses of Theorem 5.2. If there is a constant $M \geq 0$ such that $|u_0^n| \leq M$ and $|u_i^0| \leq M$ for all $n \geq 0$ and $i \geq 0$, then the SATh-up scheme is TVB, i.e.,*

$$\mathrm{TV}(\bar{u}^n) \leq 2M. \tag{5.16}$$

Moreover, if also $\bar{u}_0^{n+1} = \bar{u}_0^n$, then the scheme is TVD, i.e.,

$$\mathrm{TV}(\bar{u}^{n+1}) \leq \mathrm{TV}(\bar{u}^n). \tag{5.17}$$

Proof. In the monotone decreasing case (5.5)–(5.6), $\bar{u}_i^n \leq \bar{u}_i^{n+1}$, which implies that for all i, $-M \leq \bar{u}_i^0 \leq \bar{u}_i^1 \leq \cdots \leq \bar{u}_i^n$. The sum in (5.15) collapses, so

$$\mathrm{TV}(\bar{u}^n) = \lim_{i_{\max}\to\infty} \sum_{i=1}^{i_{\max}} (\bar{u}_{i-1}^n - \bar{u}_i^n) \leq \bar{u}_0^n - \liminf_{i_{\max}\to\infty} \bar{u}_i^n \leq 2M.$$

Moreover, when $\bar{u}_0^{n+1} = \bar{u}_0^n$,

$$\begin{aligned}
\mathrm{TV}(\bar{u}^n) - \mathrm{TV}(\bar{u}^{n+1}) &= \lim_{i_{\max}\to\infty} \sum_{i=1}^{i_{\max}} \left[(\bar{u}_{i-1}^n - \bar{u}_i^n) - (\bar{u}_{i-1}^{n+1} - \bar{u}_i^{n+1}) \right] \\
&= \lim_{i_{\max}\to\infty} \sum_{i=1}^{i_{\max}} \left[(\bar{u}_{i-1}^n - \bar{u}_{i-1}^{n+1}) - (\bar{u}_i^n - \bar{u}_i^{n+1}) \right] \\
&\geq (\bar{u}_0^n - \bar{u}_0^{n+1}) - \limsup_{i_{\max}\to\infty} (\bar{u}_{i_{\max}}^n - \bar{u}_{i_{\max}}^{n+1}) \\
&= \liminf_{i_{\max}\to\infty} (\bar{u}_{i_{\max}}^{n+1} - \bar{u}_{i_{\max}}^n) \geq 0.
\end{aligned}$$

The monotone increasing case (5.7)–(5.8) is shown in a similar way. ∎

6. Extension to higher space dimensions. Extension of low order finite volume methods to general meshes in higher dimensions is nontrivial, even using backward Euler time stepping, since the classic two point flux may not be orthogonal to the mesh element edge. However, it is easy to extend to rectangular meshes. While this could be done using Strang splitting [17] into one dimensional problems, we discuss here a genuine multidimensional extension of the SATh scheme. We illustrate the ideas for the scalar equation in two space dimensions, namely,

$$u_t + \big(f(u)\big)_x + \big(g(u)\big)_y = 0, \quad (x,y) \in \mathbb{R}^2,\ t > 0. \tag{6.1}$$

We fix a rectangular mesh of grid points by choosing $\cdots < x_{i-1/2} < x_{i+1/2} < x_{i+3/2} < \cdots$ and $\cdots < y_{j-1/2} < y_{j+1/2} < y_{j+3/2} < \cdots$ for each coordinate direction, and we let $I_{ij} = [x_{i-1/2}, x_{i+1/2}] \times [y_{j-1/2}, y_{j+1/2}]$, $\Delta x_i = x_{i+1/2} - x_{i-1/2}$, and $\Delta y_j = y_{j+1/2} - y_{j-1/2}$. For simplicity, we replace subscript E by ij, rather than I_{ij}.

Suppose that there is a shock or contact discontinuity within the space-time cell $I_{ij} \times [t^n, t^{n+1}]$ at time $\tau_{ij}(x,y)$. Since we expect a low order of approximation, we simply approximate $\tau_{ij}(x,y)$ by a constant τ_{ij}^*. Moreover, we assume that the solution is constant in space and in time on either side of τ_{ij}^*. To determine τ_{ij}^*, consider that

$$\Delta t^{n+1} \tilde{\bar{u}}_{ij}^{n+1} = \int_{t^n}^{t^{n+1}} \bar{u}_{ij}(t)\, dt \approx (\tau_{ij}^* - t^n)\, \bar{u}_{ij}^n + (t^{n+1} - \tau_{ij}^*)\, \bar{u}_{ij}^{n+1},$$

which implies that

$$\tau_{ij}^* - t^n = \frac{\bar{u}_{ij}^{n+1} - \tilde{\bar{u}}_{ij}^{n+1}}{\bar{u}_{ij}^{n+1} - \bar{u}_{ij}^n} \Delta t^{n+1} \quad \text{and} \quad \theta_{ij} = \frac{t^{n+1} - \tau_{ij}^*}{\Delta t^{n+1}} = \frac{\tilde{\bar{u}}_{ij}^{n+1} - \bar{u}_{ij}^n}{\bar{u}_{ij}^{n+1} - \bar{u}_{ij}^n}. \tag{6.2}$$

Equation (2.2) posed over I_{ij} reduces to

$$\bar{u}_{ij}^{n+1} = \bar{u}_{ij}^n - \frac{1}{\Delta x_i \Delta y_j} \int_{t^n}^{t^{n+1}} \left\{ \int_{y_{j-1/2}}^{y_{j+1/2}} \left[f(u(x_{i+1/2}, y, t) - f(u(x_{i-1/2}, y, t)) \right] dy \right.$$
$$\left. + \int_{x_{i-1/2}}^{x_{i+1/2}} \left[g(u(x, y_{j+1/2}, t)) - g(u(x, y_{j-1/2}, t)) \right] dx \right\} dt.$$

We use midpoint quadrature on each spatial interval to conclude (approximately) that

$$\bar{u}_{ij}^{n+1} = \bar{u}_{ij}^n - \frac{1}{\Delta x_i} \int_{t^n}^{t^{n+1}} \left[f(u(x_{i+1/2}, y_j, t) - f(u(x_{i-1/2}, y_j, t)) \right] dt$$
$$- \frac{1}{\Delta y_j} \int_{t^n}^{t^{n+1}} \left[g(u(x_i, y_{j+1/2}, t)) - g(u(x_i, y_{j-1/2}, t)) \right] dt. \tag{6.3}$$

We have reduced the flux integrals to two terms, each one varying only in one space dimension. It is now straightforward to incorporate a numerical flux and apply the DAQ rule as in the case for one dimension. Moreover, a similar procedure can be used for (2.5).

7. Numerical results. We have theoretical proof that the SATh-up scheme works well, but it is restricted to monotone flux functions. We show some numerical results for SATh-up, but concentrate on the more general SATh-LF scheme (4.10)–(4.11). Although we have no theory for this scheme, we will see that it satisfies the results obtained for SATh-up. As should be expected, when the SATh-up scheme can be used, it produces less numerical diffusion than SATh-LF.

In most of our figures, we plot the solution to the SATh scheme $\bar{u}$ as a black line and $\tilde{\bar{u}}$ as a black dotted line. We compare SATh-LF with two other schemes, both stabilized with the same Lax-Friedrichs numerical flux and using equivalent space discretization, but the time stepping is either Crank-Nicolson (CN), shown in red, or backward Euler (BE), shown in blue. Since SATh solves for both $\bar{u}$ and $\tilde{\bar{u}}$, we take twice as many steps using the BE scheme (so it uses half the CFL number reported for SATh). We also show the value of θ in magenta. For reference, we sometimes give light green horizontal lines at the minimum and maximum values that u may take (usually, but not always, at $u = 0$ and $u = 1$). In the case of linear transport, we also compare SATh-LF with an implicit TVD scheme due to Yee [18], and plot its results in purple. Like BE, we use twice as many time steps for Yee to present a fairer comparison to SATh.

Yee presents several TVD flux corrected transport schemes. They are finite difference schemes based on central differencing in space and use the theta method for time stepping. We use the linearized scheme which is fully implicit, since the partially implicit schemes (i.e., those with $\theta < 1$) require a CFL-like time step constraint to achieve the TVD property. The limiter function Q blends the higher order scheme that uses central differencing in space with BE, which is simply upstream differencing for the linear transport problem. We use the limiter function

$$Q_{i+1/2} = \operatorname{minmod}\Big(1, \frac{u_i - u_{i-1}}{u_{i+1} - u_i}\Big) + \operatorname{minmod}\Big(1, \frac{u_{i+2} - u_{i+1}}{u_{i+1} - u_i}\Big) - 1.$$

The unlimited scheme is unstable, but it is formally second order accurate in space and first order in time.

Before proceeding, we remark that we will see test cases in which $\theta > 1$. This seems counterintuitive to our development of DAQ, but such values of θ arise in at least two situations. First, when the denominator in (3.7) is close to zero, rounding error can produce $\theta > 1$. This possibility has no effect on the solution, however, since the integrals (3.8)–(3.9) are well approximated. Second, $\theta > 1$ may occur at a peak or valley in the solution. In the notation of (3.7), a peak or valley will lead to a mesh element where $\tilde{v}$ is larger or smaller than the endpoint values v^0 and v^1, and so $\theta > 1$ results. This possibility does not cause problems for the SATh scheme. The explanation and development of DAQ given in Section 3 tacitly assumed that $0 \le \theta \le 1$, but the theoretical results do not in fact require this assumption. The integrals in (3.8)–(3.9) are accurately approximated by an extrapolation procedure rather than through interpolation when $\theta > 1$.

Let $H(x)$ denote the Heaviside function, which is zero for $x < 0$ and one for $x > 0$.

7.1. Implementation. Before presenting the results, we make a few comments on the implementation of the scheme. We restrict the problem to $[L_0, L_1] \times [0, t_{\max}]$, and we impose either a periodic boundary condition or a Dirichlet boundary condition on both sides of the spatial interval (for simplicity of implementation). The scheme almost always develops instabilities if $\theta \ge 1/2$ is not enforced. We define $w_i = \bar{u}_i^{n+1} - \bar{u}_i^n$ and $v_i = \tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n$ and solve for these variables using a straightforward implementation of Newton's method with the initial guess $w_i = 0$ and $v_i = \tilde{\bar{u}}_i^n - \bar{u}_i^n$.

Recall the two parameters ϵ and θ^* in the definition of θ, (4.6). We found that the value of ϵ has little effect on the solution, as long as it is small (say $\epsilon \le 10^{-6}$). The value $\epsilon = 0$ seems to work well, but in principle it could lead to floating point overflow. We therefore took $\epsilon = 10^{-100}$. We also found that the value of θ^* has little effect on the converged solution. However, the first Newton iteration will use $\theta = \theta^*$ (since then $w_i = 0$). We found that for this first iteration, taking $\theta^* = 1$ enabled Newton's method to converge faster. After the first Newton iteration, one can revert to $\theta^* = 1/2$, say, if one wishes.

We define θ_i as in (4.6) using a cut-off function κ. That is, we let

$$\theta_i = \begin{cases} \kappa(\tilde{\theta}_i),\ \tilde{\theta}_i = v_i/w_i & \text{if } |w_i| > \epsilon, \\ \theta^* & \text{if } |w_i| \le \epsilon. \end{cases}$$

With the choice $\kappa(\tilde{\theta}) = \max(1/2, \tilde{\theta})$, we recover the stated definition (4.6). Since then $\kappa'(\tilde{\theta}) = 0.5(1 + \operatorname{sign}(\tilde{\theta} - 0.5))$ is not continuous, we took smoothed versions of the function κ, with the intent to improve the Newton convergence. We found that smoothing κ had little effect on the number of iterations. However, whatever

instantiation of κ was chosen, it was seen to be important to use its correct derivative, even when this derivative is discontinuous. The results we present below use no smoothing.

In Newton's method, it is important that the implementation of the derivatives (i.e., the Jacobian matrix) of the function (4.10)–(4.11) can handle division by w in $\tilde{\theta} = v/w$, since w can be quite small and even vanish. When $|w| \le \epsilon$, we simply fix $\theta = \theta^*$ and consider its derivatives with respect to w and v to be zero. So consider the case when $|w| > \epsilon$. The terms involving $\tilde{\theta} = v/w$ in the function (4.10)–(4.11) have a form like $T = \theta^p\big(f(w+\bar{u}^n) - f(\bar{u}^n)\big)$, $p = 1, 2$, so the derivatives can be computed as

$$\frac{\partial T}{\partial w} = \theta^p\, f'(w+\bar{u}^n) - p\theta^{p-1}\kappa'(\tilde{\theta})\,\tilde{\theta}\,\frac{f(w+\bar{u}^n) - f(\bar{u}^n)}{w},$$
$$\frac{\partial T}{\partial v} = p\theta^{p-1}\kappa'(\tilde{\theta})\,\frac{f(w+\bar{u}^n) - f(\bar{u}^n)}{w}.$$

One should implement the derivatives this way, since the quantity $\dfrac{f(w+\bar{u}^n) - f(\bar{u}^n)}{w}$ is the derivative of f at some point between $w+\bar{u}^n$ and $\bar{u}^n$, and so it is reasonable in size.

We determined Newton's method had converged when the size of the Newton update met a tolerance of 10^{-6} times the quantity one plus the initial size of the residual. Overall, our SATh-LF scheme converged in about 1-4 more iterations per time step than the backward Euler scheme.

7.2. Linear transport in one space dimension. We consider first the linear equation

$$u_t + u_x = 0, \quad L_0 < x < L_1,\ t > 0, \tag{7.1}$$

with unit speed, so $\alpha_{\mathrm{LF}} = 1$. In this case, SATh-LF and SATh-up are the same scheme, up to rounding error.

7.2.1. A contact discontinuity. We consider the linear advection equation (7.1) with $L_0 = -0.1$ and $L_1 = 0.9$, the boundary condition $u(L_0, t) = 1$ (and $u(L_1, t) = 0$, but the mass does not propagate that far in the simulation), and the initial jump condition $u(x, 0) = 1 - H(x)$ ($H(x)$ is the Heaviside function). The solution should be a contact discontinuity at $x = t$. We show the solution at time $t = 0.5$ for four tests in Figure 7.1, using $\Delta x = 1/80, 1/160, 1/320$ and $\Delta t = 5\Delta x$ (so the CFL number is 5 and we use 8, 16, and 32 time steps) as well as $\Delta x = 1/160$ and $\Delta t = 10\Delta x$ (CFL 10). As mentioned above, we plot the solution to the SATh-LF scheme $\bar{u}$ as a black line and $\tilde{\bar{u}}$ as a dotted line. We compare the solution to standard Lax-Friedrichs stabilized Crank-Nicolson (CN) time stepping in red. We also compare to backward Euler (BE) in blue and Yee's TVD scheme in purple, both of which use twice as many time steps, i.e., CFL 2.5 and CFL 5.

We can see that Crank-Nicolson, although stable, displays excessive oscillation, and both backward Euler and Yee's scheme display excessive numerical diffusion. The SATh-LF scheme, however, shows no oscillation, nearly the accuracy of CN, and much less numerical diffusion compared to backward Euler and TVD flux-limited scheme of Yee. The solution remains stable and monotone (cf. Theorems 5.1 and 5.2). For reference, the value of θ is shown in magenta. For this problem, SATh-LF uses Crank-Nicolson ($\theta = 1/2$) over most of the domain, but improves on backward Euler by maintaining $1/2 \le \theta < 1$. (That $\theta \le 1$ is consistent with Corollary 5.3.)

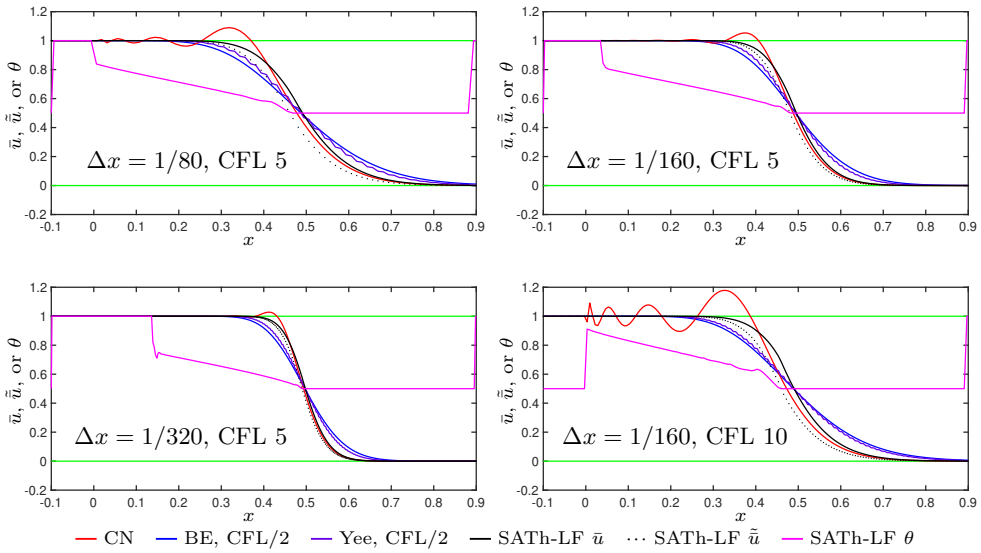

FIG. 7.1. *Contact discontinuity at $t = 0.5$, using $\Delta x = 1/80$ (top left), $\Delta x = 1/160$ (top right), and $\Delta x = 1/320$ (bottom left) with $\Delta t = 5\Delta x$ (CFL 5), as well as $\Delta x = 1/160$ (bottom right) with $\Delta t = 10\Delta x$ (CFL 10) (although BE and Yee use CFL/2).*

TABLE 7.1
Contact discontinuity at $t = 0.5$, error and convergence order for SATh, BE, CN, and Yee using $m = 1/\Delta x$ cells and $\Delta t = 5\Delta x$ (CFL 5, but BE and Yee use CFL 2.5).

	SATh		BE		CN		Yee	
m	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order
40	1.045e-1	0.414	1.528e-1	0.328	1.195e-1	0.523	1.356e-1	0.353
80	7.348e-2	0.508	1.140e-1	0.423	7.740e-2	0.626	9.971e-2	0.444
160	5.066e-2	0.536	8.198e-2	0.475	4.984e-2	0.635	7.120e-2	0.486
320	3.490e-2	0.538	5.835e-2	0.491	3.313e-2	0.589	5.103e-2	0.481
640	2.411e-2	0.533	4.141e-2	0.495	2.264e-2	0.549	3.657e-2	0.481

We computed the total variation of SATh-LF, backward Euler, and Yee's scheme. As we would hope from Corollary 5.4, the total variation remained one for all three schemes.

The computed discrete L^1 error and convergence order are given in Table 7.1. It shows an order of convergence $\mathcal{O}(\Delta x^{1/2})$ for all the schemes, as one should expect for a pure contact discontinuity. In spite of the convergence theory for DAQ, we do not see $\mathcal{O}(\Delta x)$ for SATh since $\theta \geq 1/2$ is enforced to maintain stability.

7.2.2. Convergence for a smooth problem. We test our scheme, CN, BE, and Yee in the simple case of constant linear transport (7.1) with $L_0 = 0$ and $L_1 = 2$, the initial condition $u_0(x) = 0.5 + \sin(\pi x)$, and periodic boundary conditions. We observe in Table 7.2 a first order rate of convergence for the schemes in the discrete L^1 norm at moderate (5) CFL. At high (20) CFL, SATh and CN converge a bit better, while BE and Yee converge a bit worse. Results in the L^∞ norm are also given in the table.

In light of Theorem 3.4, one might have expected to see second order convergence for SATh-LF, and moreover, third order for CN and second order for Yee. However,

the accuracy of SATh-LF and CN are limited to first order accuracy since we use simple one point upstream weighting. The unlimited Yee scheme (central differencing in space) is an unstable scheme, and so flux limiting is required to maintain stability and achieve the TVD property even for a smooth problem, making it behave more like BE. Moreover, SATh is also limited by the stability requirement that $\theta \geq 1/2$, so the full accuracy of DAQ is *not* realized. We remark that we have observed second order convergence in some tests when a higher order flux stabilization is used (such as one based on WENO reconstruction); however, the maximum principle is lost in those cases.

TABLE 7.2
Smooth linear transport error and convergence order for SATh-LF, BE, CN, and Yee at $t = 2$, using $m = 2/\Delta x$ mesh cells.

	SATh-LF CFL 5		BE CFL 2.5		CN CFL 5		Yee CFL 2.5	
m	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order
80	3.114e-1	1.041	7.319e-1	0.510	2.854e-1	0.937	6.244e-1	0.598
160	1.531e-1	1.025	4.453e-1	0.717	1.486e-1	0.941	3.664e-1	0.769
320	7.711e-2	0.989	2.470e-1	0.850	7.626e-2	0.963	1.992e-1	0.879
640	3.883e-2	0.990	1.303e-1	0.923	3.868e-2	0.979	1.042e-1	0.935
m	$L^\infty_{\Delta x}$ err.	order	$L^\infty_{\Delta x}$ err.	order	$L^\infty_{\Delta x}$ err.	order	$L^\infty_{\Delta x}$ err.	order
80	2.953e-1	0.877	5.747e-1	0.509	2.241e-1	0.934	4.944e-1	0.593
160	1.500e-1	0.977	3.497e-1	0.717	1.167e-1	0.941	2.891e-1	0.774
320	7.371e-2	1.025	1.940e-1	0.850	5.990e-2	0.963	1.570e-1	0.881
640	3.571e-2	1.046	1.023e-1	0.923	3.038e-2	0.979	8.203e-2	0.936
	CFL 20		CFL 10		CFL 20		CFL 10	
m	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order	$L^1_{\Delta x}$ err.	order
80	9.159e-1	0.503	1.205e-0	0.161	1.103e-0	0.964	1.184e-0	0.187
160	4.149e-1	1.142	9.381e-1	0.361	3.799e-1	1.538	9.042e-1	0.389
320	1.448e-1	1.519	6.237e-1	0.589	1.224e-1	1.634	5.947e-1	0.605
640	4.781e-2	1.598	3.655e-1	0.771	4.590e-2	1.415	3.437e-1	0.791
m	$L^\infty_{\Delta x}$ err.	order	$L^\infty_{\Delta x}$ err.	order	$L^\infty_{\Delta x}$ err.	order	$L^\infty_{\Delta x}$ err.	order
80	6.921e-1	0.571	9.467e-1	0.159	8.656e-1	0.967	9.318e-1	0.186
160	3.819e-1	0.858	7.368e-1	0.362	2.984e-1	1.537	7.126e-1	0.387
320	1.805e-1	1.081	4.898e-1	0.589	9.615e-2	1.634	4.675e-1	0.608
640	8.228e-2	1.133	2.871e-1	0.771	3.605e-2	1.415	2.705e-1	0.789

7.2.3. Shu's linear test. We next consider a standard test problem [9], often called Shu's linear test. The initial profile is defined over $x \in [0, 2]$, contains discontinuous jumps and smooth regions, and imposes periodic boundary conditions. The test is designed for high order methods, so we should not expect to see particularly good results, but only some improvement for SATh over the backward Euler results.

The results are shown in Figure 7.2, where we have used $\Delta x = 1/320$ and advanced to time $t = 2$, which is one period. The initial profile is shown in green (and is the exact solution at $t = 2$). The left plot uses CFL 0.5, and SATh-LF and CN give essentially the same solution (i.e., the red line is covered by the black line). We also show forward Euler (FE) in cyan, which of course is more accurate than most of the implicit schemes, although Yee's scheme is slightly better than FE in this test (recall that Yee uses twice the number of time steps).

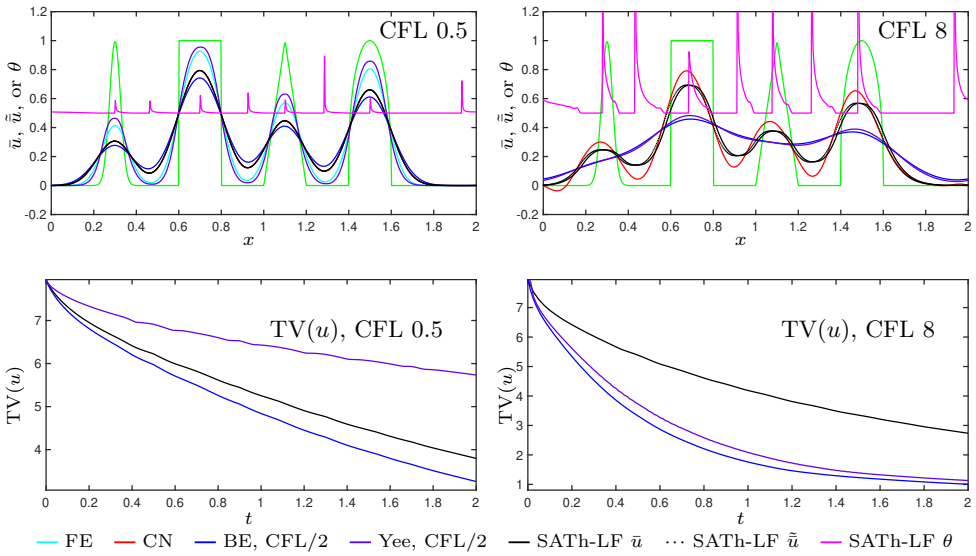

FIG. 7.2. *Shu's linear test at $t = 2$, using $\Delta x = 1/320$ and $\Delta t = \Delta x/2$ (CFL 0.5) on the left and $\Delta t = 8\Delta x$ (CFL 8) on the right (although BE and Yee use half the CFL step). The true solution is shown in green, and forward Euler (FE) results on the left are shown in cyan. The total variation for SATh-LF, BE, and Yee are shown on the bottom.*

The right plot uses CFL 8. As expected, there is significant numerical diffusion; however, we see considerable improvement for SATh-LF (black) over BE (blue). Moreover, SATh-LF shows little degradation from CN (red) for the larger CFL, but this is not so for BE. Yee's scheme (purple) reverts closer and closer to BE as the CFL increases (and more limiting is needed). One should note that $\theta > 1$ often occurs at a local extrema in the solution.

The total variation for SATh-LF and BE are shown on the bottom line of Figure 7.2. The SATh-LF, BE, and Yee schemes display the TVD property for this example, with SATh-LF dissipating the total variation at a better (i.e., slower) rate than BE, and much better than Yee's scheme at the higher CFL.

7.3. Burgers equation in one space dimension. Next we consider Burgers equation with the flux function $f(u) = u^2/2$, i.e.,

$$u_t + u\,u_x = 0, \quad L_0 < x < L_1,\ t > 0. \tag{7.2}$$

7.3.1. A Riemann shock. The first test is for a Riemann shock, implemented as in the case of a contact discontinuity above ($L_0 = -0.1$, $L_1 = 0.9$, $u(x,0) = 1 - H(x)$, $u(L_0,t) = 1$, and $u(L_1,t) = 0$). For this problem, $\alpha_{\mathrm{LF}} = 1$. We show the results in Figure 7.3, for a test at CFL 4 and both low ($\Delta x = 1/20$), medium ($\Delta x = 1/40$), and high $\Delta x = 1/80$) resolution. We also show $\Delta x = 1/40$ at CFL 8.

We see results similar to the contact discontinuity. The three schemes correctly predict the speed of the shock. SATh-LF has less numerical diffusion compared to backward Euler, and predicts the shock about as well as CN, which oscillates unacceptably. The SATh-LF solution remains stable and monotone (as suggested by Theorems 5.1 and 5.2). The total variation also remains 1 for both SATh-LF and BE at CFL 4 and 8. At the higher CFL, we see a degradation in the overall approximation for all three schemes, but the comparisons remain the same (i.e., SATh-LF is the

most accurate without introducing oscillatory behavior).

At very high CFL (greater than about 10), we have difficulty solving the equations, but it appears that the SATh-LF solution may not be TVD. A slight oscillation arises at the location of the shock at the first time step, with the total variation being 1.005. The second time step appears to be fine, and the solution is TVD from then on. If SATh-LF is TVD, it may be so only with some conditions.

We remark that we also ran this example with the SATh-up scheme. We found that Theorem 5.2 and Corollary 5.4 hold as expected. The solution remains monotone and the total variation is one, even with tests using CFL 200.

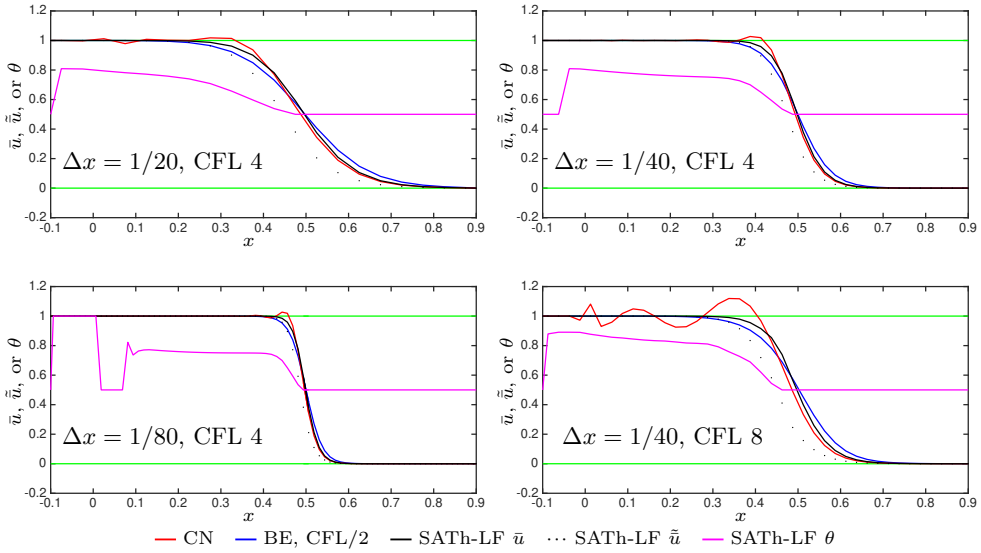

FIG. 7.3. *Burgers Riemann shock discontinuity at $t = 1$, using $\Delta x = 1/20$ (top left), $\Delta x = 1/40$ (top right), and $\Delta x = 1/80$ (bottom left) with $\Delta t = 4\Delta x$ (CFL 4), as well as $\Delta x = 1/40$ (bottom right) with $\Delta t = 5\Delta x$ (CFL 8) (although BE uses CFL/2).*

The contact discontinuity and the Riemann shock differ in the observed convergence rate. As shown in Table 7.3 for CFL 4, the SATh-LF scheme convergences with first order accuracy. As is well known, the shock is in some sense self-sharpening (since characteristics converge at the shock), and so this problem is actually better behaved than the contact discontinuity of Table 7.1.

TABLE 7.3

Burgers Riemann shock at $t = 1.0$, error and convergence order for SATh-LF, BE, and CN using $m = 1/\Delta x$ and $\Delta t = 4\Delta x$ (CFL 4, BE uses CFL 2).

	SATh-LF		BE		CN	
m	$L^1_{\Delta x}$ error	order	$L^1_{\Delta x}$ error	order	$L^1_{\Delta x}$ error	order
20	8.724e-02	—	1.047e-01	—	8.150e-02	—
40	4.443e-02	0.973	5.529e-02	0.921	4.105e-02	0.989
80	2.225e-02	0.998	2.792e-02	0.986	2.055e-02	0.998
160	1.112e-02	1.001	1.396e-02	1.000	1.028e-02	0.999
320	5.562e-03	0.999	6.982e-03	1.000	5.138e-03	1.001
640	2.781e-03	1.000	3.491e-03	1.000	2.569e-03	1.000

TABLE 7.4
Burgers Riemann shock at $t = 1.0$, computational efficiency study.

Scheme	Δt	CFL	CPU time (μs)	Average number of Newton iterations
SATh-LF	1/20	4	1364.65	5.05
BE	1/40	2	770.90	4.00
BE	1/100	0.8	1471.15	3.03

We use this problem to conduct a study of the computational efficiency of SATh-LF versus BE. For this test, we implemented the schemes using the C++ programming language and used identical programming style for each scheme. The linear systems are solved using the LAPACK [1] double precision banded solver routine. We ran the simulations using a fixed $\Delta x = 1/80$ and final simulation time $t = 1.0$. (The solutions were depicted earlier in Fig. 7.3, lower left.)

We report measurements in Table 7.4, which gives the overall CPU time used by the computer program for each scheme (in microseconds) and the average number of Newton iterations used per time step. The CPU time varies from run to run, but we reported typical average values. We compare SATh-LF using $\Delta t = 1/20$ (CFL 4) and BE using both $\Delta t = 1/40$ (CFL 2) and $\Delta t = 1/100$ (CFL 0.08). The latter test was chosen because the SATh-LF and BE schemes produce nearly the same solution for $\Delta t = 1/100$ (although there is still slightly more numerical dissipation for BE), while $\Delta t = 1/80$ (CFL 1) gave noticeably different solutions. As one can see, the SATh-LF scheme is relatively efficient compared to BE. For the same quality of solution, it is a little faster than BE, which requires a CFL *below* 1 in this test.

TABLE 7.5
Burgers Riemann rarefaction at $t = 0.25$, L^1 and L^∞ error and convergence order for SATh-LF, BE, and CN using $m = 1/\Delta x$ cells and $\Delta t = 5\Delta x$ (CFL 5, BE uses CFL 2.5).

	SATh-LF		BE		CN	
m	$L^1_{\Delta x}$ error	order	$L^1_{\Delta x}$ error	order	$L^1_{\Delta x}$ error	order
40	4.658e-02	0.759	5.068e-02	0.736	3.958e-02	0.862
80	2.716e-02	0.778	2.978e-02	0.767	2.247e-02	0.817
160	1.562e-02	0.798	1.799e-02	0.727	1.325e-02	0.762
320	8.858e-03	0.818	1.104e-02	0.704	7.661e-03	0.790
640	4.962e-03	0.836	6.761e-03	0.707	4.359e-03	0.814
m	$L^\infty_{\Delta x}$ error	order	$L^\infty_{\Delta x}$ error	order	$L^\infty_{\Delta x}$ error	order
40	1.892e-01	0.315	2.221e-01	0.218	1.671e-01	1.388
80	1.495e-01	0.340	1.839e-01	0.272	1.265e-01	0.402
160	1.146e-01	0.384	1.478e-01	0.315	1.016e-01	0.316
320	8.568e-02	0.420	1.159e-01	0.351	7.890e-02	0.365
640	6.285e-02	0.447	8.919e-02	0.378	5.936e-02	0.411

7.3.2. A Riemann rarefaction. We also consider Burgers equation with a Riemann rarefaction, implemented as $u(0,t) = 0$, $u(1,t) = 1$, and $u(x,0) = 1$. Again $\alpha_{\mathrm{LF}} = 1$. We show the results for CFL 5 in Figure 7.4 using $\Delta x = 1/40$ and $\Delta x = 1/80$ resolution. All three schemes work reasonably well, although CN oscillates unacceptably and SATh-LF has less numerical diffusion than backward Euler. The SATh-LF solution remains stable and monotone (as suggested by Theorems 5.1 and 5.2 and

Corollary 5.3). The total variation also remains 1 for both SATh-LF and BE. The rate of convergence of SATh-LF in both the discrete L^1 and L^∞ norms is given in Table 7.5 for CFL 5. It appears to be approaching a convergence rate of 1 in L^1 and $1/2$ in L^∞ as Δx is refined.

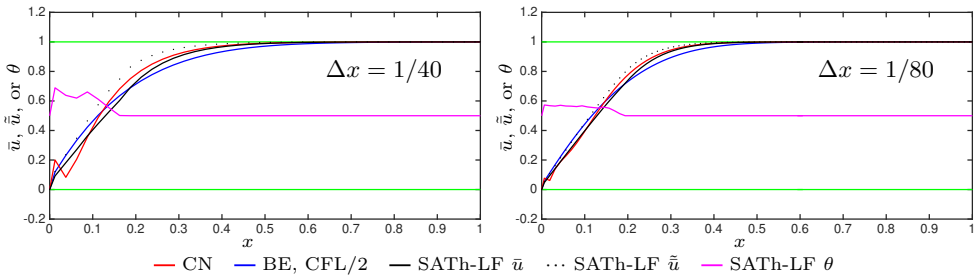

FIG. 7.4. *Burgers Riemann rarefaction at $t = 0.25$, using $\Delta x = 1/40$ (left) and $\Delta x = 1/80$ (right) and $\Delta t = 5\Delta x$ (CFL 5) (although BE uses half the CFL step).*

7.3.3. Shock formation. Finally, we can simulate shock formation by, e.g., imposing periodic boundary conditions and the initial sine wave condition $u_0(x) = 0.5 + \sin(\pi x)$ over $x \in [0, 2]$. For this problem, the solution lies in the interval $[-0.5, 1.5]$, so $\alpha_{\text{LF}} = 1.5$ and the characteristics move in both the positive and negative directions. The shock forms at time $1/\pi = 0.318$. Results are shown in Figure 7.5 for $\Delta x = 1/100$ and CFL 5 at times 0.2, 0.4, and 0.6. The shock forms cleanly, with SATh-LF giving a solution about as accurate as CN, although the CN solution oscillates a bit, and the BE solution is more diffuse. The total variation should remain constant until the shock forms (it reduces a little), and it should reduce after the shock forms (as it does). In both regimes, however, SATh-LF improves on the BE results.

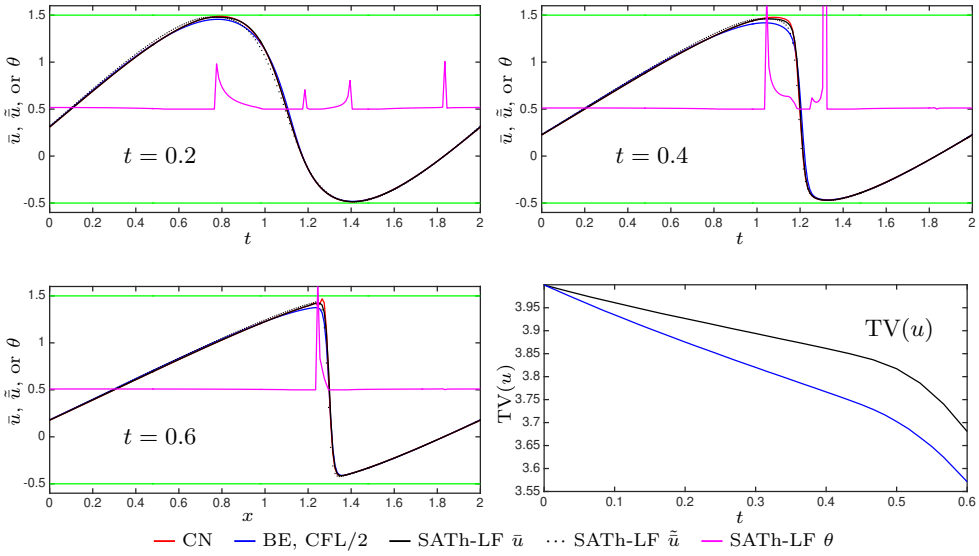

FIG. 7.5. *Burgers shock formation, using $\Delta x = 1/100$ and $\Delta t = 1/30$ (CFL 5) at times $t = 0.2$, $t = 0.4$, and $t = 0.6$, and the total variation. In this figure, the horizontal green reference lines are at $u = -0.5, 1.5$.*

TABLE 7.6
Before Burgers shock formation, L^1 error and convergence order for SATh-LF at $t = 0.3$ (just before the true shock forms at $t = 1/\pi = 0.318$) using $m = 2/\Delta x$ cells and CFL 5 and CFL 25.

	CFL 5				CFL 25			
m	$L^1_{\Delta x}$ error	order	$L^\infty_{\Delta x}$ error	order	$L^1_{\Delta x}$ error	order	$L^\infty_{\Delta x}$ error	order
100	5.388e-2	——	2.136e-1	——	1.053e-1	——	3.235e-1	——
200	3.018e-2	0.836	1.561e-1	0.453	5.472e-2	0.944	2.779e-1	0.219
400	1.601e-2	0.915	1.078e-1	0.534	2.555e-2	1.099	1.995e-1	0.478
800	8.380e-3	0.934	7.076e-2	0.608	1.140e-2	1.164	1.228e-1	0.700

In Table 7.6 we give the discrete L^1 and L^∞ errors and convergence order for the SATh-LF scheme. The results are at $t = 0.3$, which is just before the true shock forms at $t = 1/\pi = 0.318$. Results for CFL 5 and CFL 25 are presented, and both sets of results show first order convergence in L^1. The L^∞ convergence order seems to approach one as the mesh is refined (i.e., as the smooth but steep front is resolved).

7.4. Buckley-Leverett equation in one space dimension. Next we consider the Buckley-Leverett equation

$$u_t + f(u)_x = 0, \quad 0 < x < 1, \ t > 0, \quad \text{where } f(u) = \frac{u^2}{u^2 + (1-u)^2}. \tag{7.3}$$

We consider two problems with $u \in [0, 1]$, so $\alpha_{\mathrm{LF}} = 2$.

7.4.1. A Riemann problem. We apply an initial jump at $x = 0$ by setting $u(0, t) = 1$ and $u(x, 0) = 0$, and a shock followed by a rarefaction is produced. The results at $t = 0.5$ are shown in Figure 7.6 using $\Delta x = 1/40$ and $\Delta t = \Delta x$ (CFL 2) and $\Delta x = 1/80$ and $\Delta t = 2.5\Delta x$ (CFL 5). The three schemes perform similarly for the low CFL test, although BE is more diffusive. For the higher CFL test, the higher order CN scheme is able to capture the transition from the rarefaction to the shock (occurring at about $x = 0.6$) better than the low order methods. However, the CN solution has an unphysical oscillation there. The SATh-LF scheme clearly outperforms BE (and we remind the reader, BE is using half the CFL number).

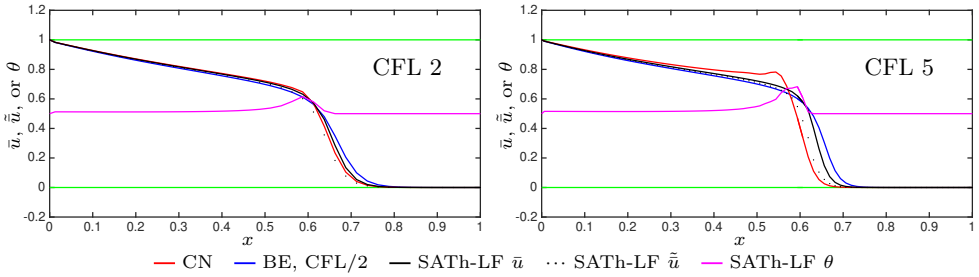

FIG. 7.6. *A Buckley-Leverett rarefaction and shock at $t = 0.5$, using $\Delta x = 1/40$ and $\Delta t = \Delta x$ (CFL 2) on the left and $\Delta x = 1/80$ and $\Delta t = 2.5\Delta x$ (CFL 5) on the right (although BE uses half the CFL step).*

7.4.2. A problem of merging pulses. The next example for the Buckley-Leverett flux function uses the initial condition

$$u_0(x) = \begin{cases} 1 - 20x & \text{for } 0 \le x \le 0.05, \\ 0.5 & \text{for } 0.25 \le x \le 0.4, \\ 0 & \text{otherwise.} \end{cases} \tag{7.4}$$

Two pulses merge over time, which gives rise to an interaction of shocks and rarefactions. We use $\Delta x = 1/120$ grid elements and $\Delta t = 1.5\Delta x$ (CFL 3). The results at times $t = 0.15, 0.3, 0.45$ are shown in Fig. 7.7 for the Lax-Friedrichs schemes. We also show in Fig. 7.8 results for the upstream schemes, which are less diffuse and so give better results. The fine scale CN-up scheme ($\Delta x = 1/1200$, $\Delta t = \Delta x$) is shown in light green, and it is considered the reference solution. All six schemes handle the merging of the two pulses reasonably well. For each stabilization (LF or upstream), the CN results are sharpest, but the solution oscillates, and much worse so as the CFL number increases.

Fine scale simulations show that the right-most pulse is overtaken by the left one at about $t = 0.5$ (i.e., two shocks merge into one at this time). The BE results are so diffuse that the second pulse is lost at $t = 0.4$, while SATh and CN lose it at about $t = 0.45$. Moreover, BE dissipates the total variation faster than SATh, although both are TVD. Overall, in this test SATh-LF and SATh-up reproduce the solution to adequate accuracy without oscillation.

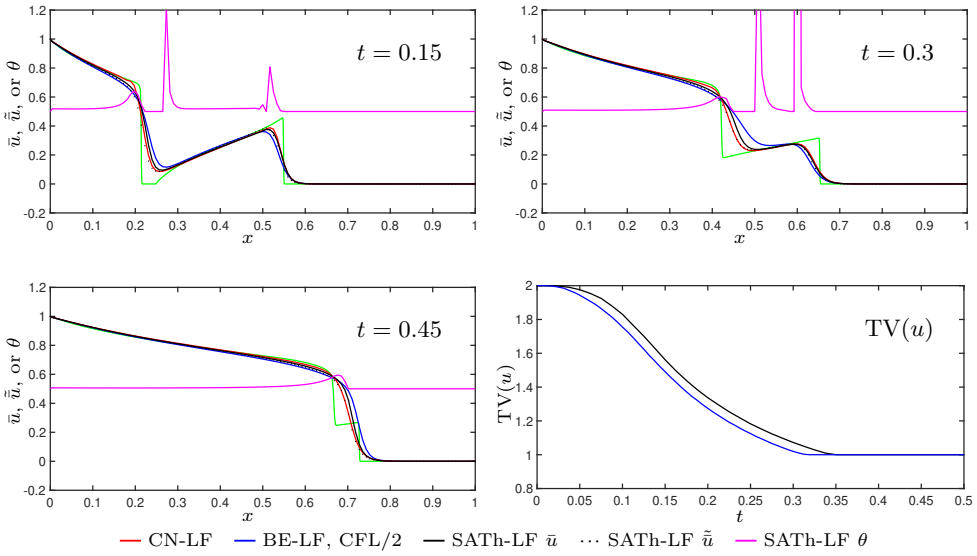

FIG. 7.7. *A Buckley-Leverett example of merging pulses using Lax-Freidrichs stabilization, $\Delta x = 1/120$ and $\Delta t = 1.5\Delta x$ (CFL 3, BE uses half the CFL step). The fine scale CN-up scheme ($\Delta x = 1/1200$, $\Delta t = \Delta x$) is used to produce the reference solution, shown in light green. Also shown is the total variation for BE-LF and SATh-LF.*

7.5. A non-monotone flux function in one space dimension. The theory we developed for SATh-up depended on the monotonicity of the flux function. We consider next a flux function that is not monotone, namely,

$$u_t + f(u)_x = 0, \quad -0.1 < x < 0.9, \ t > 0, \quad \text{where } f(u) = \frac{64}{39}\Big(u^3 - \frac{3}{2}u^2 + \frac{39}{64}u\Big). \tag{7.5}$$

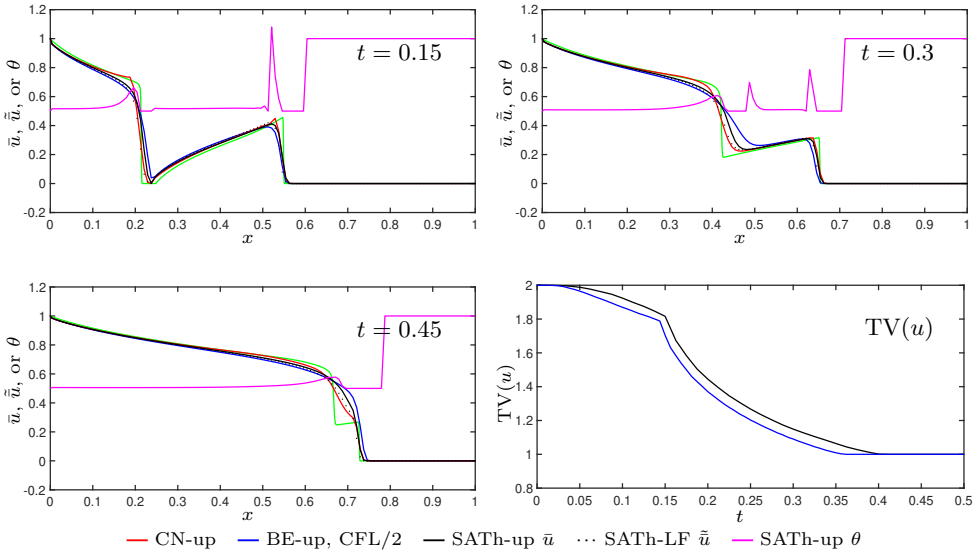

FIG. 7.8. *A Buckley-Leverett example of merging pulses using upstream stabilization,* $\Delta x = 1/120$ *and* $\Delta t = 1.5\Delta x$ *(CFL 3, BE uses half the CFL step). The fine scale CN-up scheme (*$\Delta x = 1/1200$*,* $\Delta t = \Delta x$*) is used to produce the reference solution, shown in light green. Also shown is the total variation for BE-up and SATh-up.*

When $u \in [0,1]$, one can verify that $\alpha_{\mathrm{LF}} = 1$. The graph of $f(u)$ appears in Figure 7.9 on the top right. The flux is far from being monotone; moreover, it is not convex.

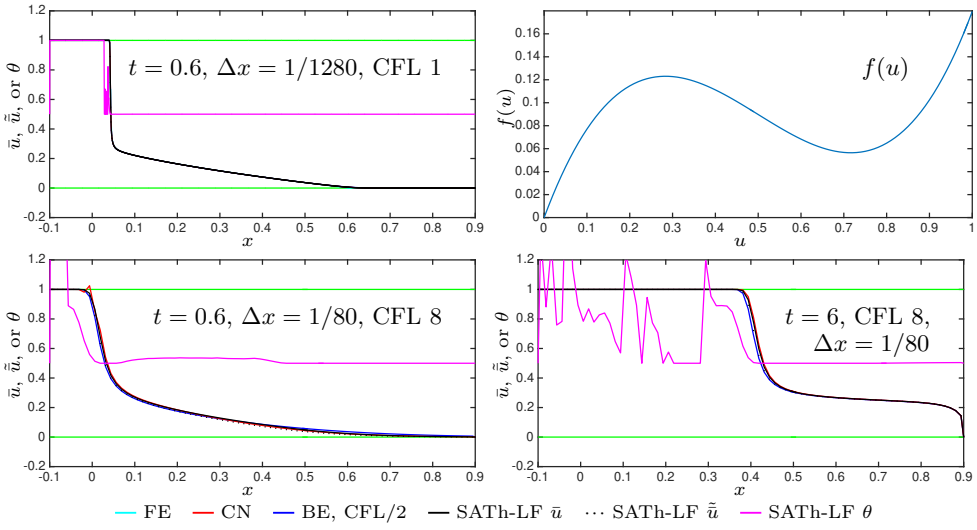

FIG. 7.9. *A non-monotone flux. On the top left is the reference solution at* $t = 0.6$ *using* $\Delta x = 1/1280$ *and* $\Delta t = \Delta x$ *(CFL 1). Forward Euler (FE) time stepping gives the same result, plotted in cyan (although it is not really visible). On the top right is the flux function. On the bottom is the solution at* $t = 0.6$ *(left) and* $t = 6$ *(right) using* $\Delta x = 1/80$ *and* $\Delta t = 8\Delta x$ *(CFL 8), although BE uses half the CFL step.*

The results also appear in Figure 7.9 for the standard jump problem $u(-0.1, t) = 1$, $u(0.9, t) = 0$, and $u(x, 0) = 1 - H(x)$. The top left can be considered as the reference

solution at $t = 0.6$. It uses $\Delta x = \Delta t = 1/1280$ (CFL 1), and all schemes produce the same solution, including forward Euler. Results for CFL 8 are given in the bottom left, where one sees that BE is the most diffuse and CN overshoots. However, the solution is well behaved for all schemes to $t = 6$, shown in the bottom right.

Interestingly, the value of θ seems to oscillate at $t = 6$ after the steep front, i.e., to the left of $x = 0.4$. However, the solution is constant in this region, so $\bar{u}_i^{n+1} - \bar{u}_i^n$ is zero to rounding error and θ_i^{n+1} is poorly defined by (4.6). In fact, the solution oscillates a bit on the order of rounding error, and since we used $\epsilon = 10^{-100}$, we compute a value for θ rather than reverting to θ^*. If one uses $\epsilon = 10^{-6}$, there are no oscillations in θ (i.e., $\theta = \theta^*$) and we observe less rounding error in the solution. But this issue has no effect on the quality of the solution $\bar{u}$ and $\tilde{\bar{u}}$. The SATh-LF scheme is stable for this problem, and the solution has no oscillation in the sense that the total variation for SATh-LF (and BE) remains 1 to rounding error for all time. This test suggests that our theory might extend to non-monotone fluxes.

7.6. Problems in two space dimensions. Finally, we consider problems in two space dimensions (6.1). We report results of the Lax-Friedrichs stabilized scheme SATh-LF and BE using a uniformly spaced rectangular mesh. The maximum wave speeds in each direction are $\alpha_{\mathrm{LF}} = \max_u |f'(u)|$ and $\beta_{\mathrm{LF}} = \max_u |g'(u)|$.

7.6.1. Linear transport in two space dimensions. We consider the problem

$$u_t + u_x + u_y = 0 \quad \text{for } 0 < x < 1,\ 0 < y < 1, \tag{7.6}$$

$$u(x, y, 0) = \begin{cases} 1 & \text{for } 0 < x < 0.25,\ 0 < y < 0.25, \\ 0 & \text{otherwise,} \end{cases} \tag{7.7}$$

and impose $u(x, y, t) = 0$ on the boundary. For this problem, $\alpha_{\mathrm{LF}} = \beta_{\mathrm{LF}} = 1$. The true solution is a square of height one that moves diagonally across the domain.

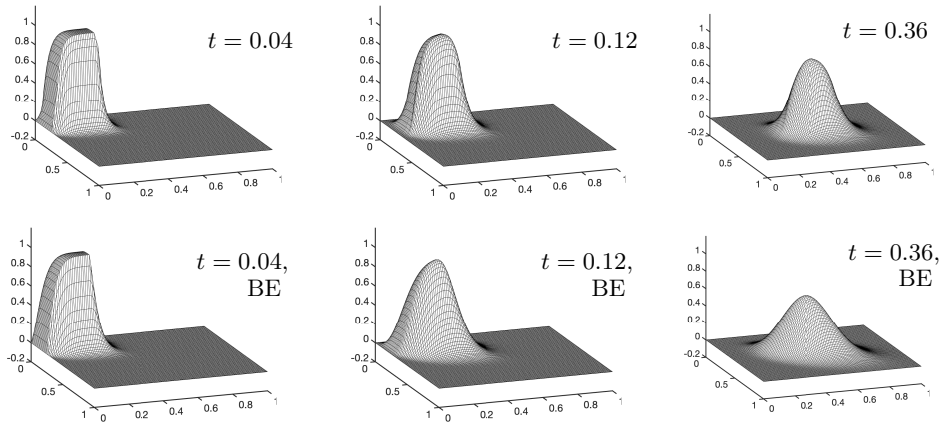

FIG. 7.10. *Linear transport in two space dimensions. Shown is* $\bar{u}(x, y, t)$ *with* $\Delta x = \Delta y = 1/100$ *at* $t = 0.04, 0.12, 0.36$ *for SATh-LF (top row,* $\Delta t = 1/25$*, CFL = 4, at step 1, 3, and 9) and BE (bottom row,* $\Delta t = 1/12.5$*, CFL = 2, at step 2, 6, and 18).*

The results are given in Fig. 7.10 using $\Delta x = \Delta y = 1/100$. SATh-LF uses $\Delta t = 1/25$ (CFL = 4) and we see the solution at steps 1, 3, and 9 (times 0.04, 0.12, and 0.36). BE uses half the time step (CFL=2) but shows the solution at the same times. Clearly the BE results display much more numerical diffusion. In fact, the height of the solution is 0.834 for SATh-LF but only 0.663 for BE.

7.6.2. Burgers equation in two space dimensions. We now consider the two dimensional Burgers equation

$$u_t + (u^2/2)_x + (u^2/2)_y = 0 \quad \text{for } 0 < x < 1,\ 0 < y < 1. \tag{7.8}$$

We impose the initial condition $u(x, y, 0) = 0$ and a boundary condition imitating a Riemann shock, namely, $u(0, y, t) = u(x, 0, t) = 1$ and $u(1, y, t) = u(x, 1, t) = 0$. For this problem, $\alpha_{\mathrm{LF}} = \beta_{\mathrm{LF}} = 1$.

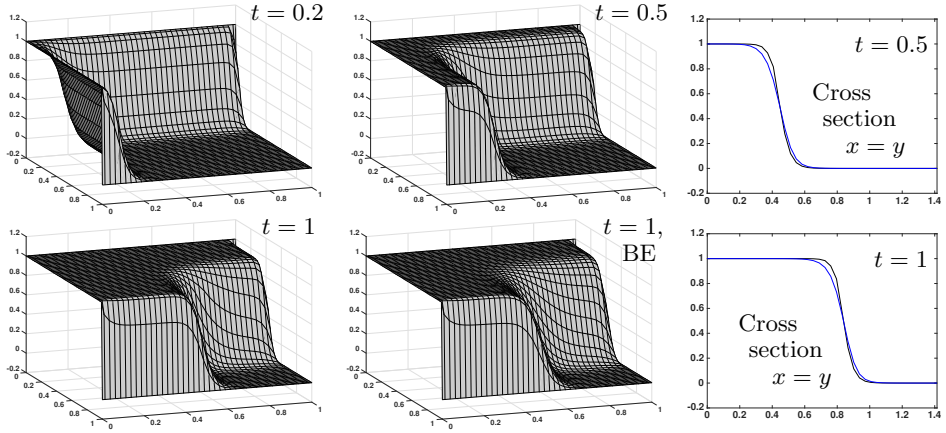

FIG. 7.11. *Burgers equation in two space dimensions. Shown is $\bar{u}(x, y, t)$ for SATh-LF at $t = 0.2, 0.5, 1.0$, and the backward Euler result at $t = 1.0$. The test uses $\Delta x = \Delta y = 1/40$, $\Delta t = 1/10$ (CFL = 4). Also shown are cross section comparisons of the front at times $t = 0.5$ and $t = 1.0$, SATh-LF in black and BE in blue.*

Results for SATh-LF are shown in Figure 7.11, using $\Delta x = \Delta y = 1/40$, $\Delta t = 1/10$ (CFL 4). The solution is shown at times $t = 0.2, 0.5, 1$. The solution of the scheme never goes above one nor below zero. Moreover, there is a bit less numerical diffusion compared to backward Euler, shown at time $t = 1$. Also shown are the $x = y$ cross sections at $t = 0.5$ and $t = 1$ for both schemes. We remark that similar results are obtained for this problem using linear transport and the Buckley-Leverett flux.

Finally, we impose a more challenging initial condition given by Jiang and Tadmor [10] involving the "oblique" data given on $[0, 1]^2$ by

$$u(x, y, 0) = \begin{cases} 0.5, & x < 0.5, y < 0.5, \\ 0.8, & x > 0.5, y < 0.5, \\ -0.2, & x < 0.5, y > 0.5, \\ -1.0, & x > 0.5, y > 0.5. \end{cases} \tag{7.9}$$

To avoid effects of the boundary conditions, we used the larger region $[0, 2]^2$ and adjusted the initial condition to center the transition lines. We report only the interior $[0.5, 1.5]^2$. The challenge of this Riemann problem with shocks and rarefactions is to avoid oscillation. Indeed, good results are obtained, as shown in Fig. 7.12 at $t = 0.5$ on a 160×160 mesh and $\Delta t = 4\Delta x$ (CFL 4) for SATh-LF and $\Delta t = 2\Delta x$ (CFL 2) for BE. No oscillation whatsoever is observed. Moreover, the contour plot in Fig. 7.12 shows that BE is more diffusive.

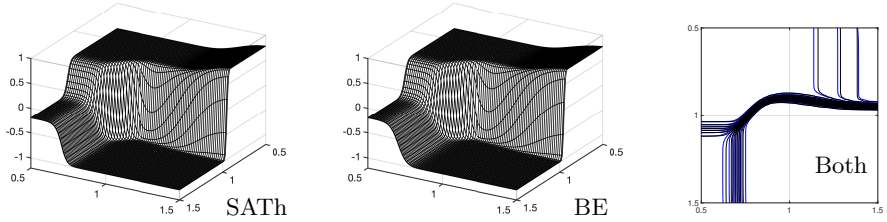

FIG. 7.12. *Burgers equation with "oblique" initial data* (7.9) *at* $t = 0.5$ *using* $m = 160$ *and CFL* 4 *for SATh-LF and CFL* 2 *for BE. Also shown is a contour plot of both superimposed with BE in blue (with* 20 *lines from* -0.9 *to* 0.9*).*

8. Summary and conclusions. We developed a *discontinuity aware quadrature* (DAQ) rule (Definition 3.2). It uses the values of the (potentially) discontinuous function $v(t)$ at the ends of the interval of integration as well as its average value. For a smooth function $g(t, v)$,

$$\int_0^{\Delta t} g(t, v(t))\, dt \approx \int_0^{\tau} g(t, v(0))\, dt + \int_{\tau}^{\Delta t} g(t, v(\Delta t))\, dt, \quad \tau = \frac{\tilde{v} - v(0)}{v(\Delta t) - v(0)}.$$

The rule is accurate to order $\mathcal{O}(\Delta t^2)$ when there is a discontinuity (Theorem 3.3), and $\mathcal{O}(\Delta t^3)$ when the solution is smooth (Theorem 3.4), even when $v(\Delta t) = v(0)$ and τ cannot be defined.

The hyperbolic conservation law (expressed either by the principle of mass conservation (2.3) or the differential equation (1.1)) controls both the local space averages at specific times, $\bar{u}_i^{n+1}$, and the local space-time average, $\tilde{\bar{u}}_i^{n+1}$. With these quantities, the DAQ rule can be applied (implicitly) to a finite volume approximation of the conservation law. The result is a theta time stepping scheme with an implicit definition of θ, i.e.,

$$\theta_i^{n+1} = \max\left(\frac{1}{2}, \frac{\tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n}{\bar{u}_i^{n+1} - \bar{u}_i^n}\right).$$

Two versions of the *self-adaptive theta* (SATh) scheme were presented, SATh-up using upstream numerical stabilization (4.4)–(4.5), and SATh-LF using the Lax-Friedrichs numerical flux function (4.10)–(4.11). These schemes were also extended to two space dimensions on rectangular meshes (§6).

For a monotone flux function, the upstream weighted scheme was amenable to analysis. We showed that SATh-up is unconditionally stable (provided only that $\theta_i^{n+1} \geq 1/2$, Theorem 5.1). If the initial and boundary conditions satisfy a monotone decreasing or increasing property, then SATh-up will satisfy the maximum principle, i.e., it gives an approximate solution that has the monotonicity property for all space and time (Theorem 5.2). Moreover, the numerical solution is TVB, and TVD if the boundary conditions do not initiate oscillation (Corollary 5.4).

For general flows one needs to use the SATh-LF scheme. We assessed its accuracy through numerical examples in one and two space dimensions. These results suggested that SATh-LF is also stable and satisfies the maximum principle, possibly even for non-monotone flux functions, at least for reasonable CFL numbers. We compared SATh-LF solutions to those of the schemes using Crank-Nicolson (CN) and backward Euler (BE) time stepping. In general, CN was oscillatory and BE was numerically diffuse, while SATh-LF gave solutions that were often near the accuracy

of CN but without oscillation, and less diffuse than BE. One might say that SATh should be viewed as a *better backward Euler* scheme, in the sense that it has reduced numerical dispersion while continuing to satisfy stability, the maximum principle, and TVD/TVB properties. It is suitable for direct use, or for use in higher order flux-limited or flux corrected transport schemes. We plan to explore this latter use in a forthcoming paper.

We end with a brief discussion of the issues regarding extension of SATh to handle systems of hyperbolic equations. The different system variables of the true solution all produce a shock at the same place in space and time, so τ in Section 3 would be well defined. However, the numerical solution may not have this property. Each system variable may predict a shock in a somewhat different location than the other variables. This leads to multiple numerical predictions of τ^*, and the resulting θ's. We expect that the issue can be overcome by using an average θ, or by further resolving the shock structure (in time). We plan to explore this issue in a future paper.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, third ed., 1999.

[2] T. Arbogast, C.-S. Huang, X. Zhao, and D. N. King, *A third order, implicit, finite volume, adaptive Runge-Kutta WENO scheme for advection-diffusion equations*, Comput. Methods Appl. Mech. Engrg., 368 (2020). DOI https://doi.org/10.1016/j.cma.2020.113155.

[3] M. Berzins and R. M. Furzeland, *An adaptive theta method for the solution of stiff and nonstiff differential equations*, Applied Numerical Mathematics, 9 (1992), pp. 1–19.

[4] S. Clain, *Finite volume maximum principle for hyperbolic scalar problems*, SIAM. J. Numer. Anal., 51 (2013), pp. 467–490.

[5] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin Heidelberg, 2005.

[6] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489.

[7] C.-S. Huang, F. Xiao, and T. Arbogast, *Fifth order multi-moment WENO schemes for hyperbolic conservation laws*, J. Sci. Comput., 64 (2015), pp. 477–507. DOI 10.1007/s10915-014-9940-z.

[8] S. Ii and F. Xiao, *Cip/multi-moment finite volume method for Euler equations, a semi-Lagrangian characteristic formulation*, J. Comput. Phys., 222 (2007), pp. 849–871.

[9] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.

[10] G. S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.

[11] D. Kuzmin, M. Q. de Luna, and C. E. Kees, *A partition of unity approach to adaptivity and limiting in continuous finite element methods*, Computers & Mathematics with Applications, 78 (2019), pp. 944–957.

[12] D. Kuzmin, M. Q. de Luna, D. I. Ketcheson, and J. Grüll, *Bound-preserving convex limiting for high-order Runge-Kutta time discretizations of hyperbolic conservation laws*, submitted, (2020).

[13] D. Kuzmin, R. Löhner, and S. Turek, *Flux-Corrected Transport: Principles, Algorithms, and Applications*, Springer, 2012.

[14] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, 2nd ed., 1992.

[15] ———, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Univ. Press, Cambridge, England, 2002.

[16] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, vol. 258, Springer Science & Business Media, 2012.

[17] G. Strang, *On the construction and comparison of difference schemes*, SIAM. J. Numer. Anal., 5 (1968), pp. 506–517.

[18] H. C. Yee, *Construction of explicit and implicit symmetric TVD schemes and their applications*, J. Comput. Phys., 68 (1987), pp. 151–179.