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

TODD ARBOGAST† AND CHIEH-SEN HUANG‡

Version: June 26, 2020

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), 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 \( u(x,t) \) can be written in terms of the flux function \( f(u) \in \mathbb{R}^d \) as

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

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.

*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 107-2115-M-110-004-MY2, 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)
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 [13, 4] (or invariant domain property [5]). 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 [11, 12, 10, 9, 1].

The theta time stepping method is often seen to reduce numerical diffusion compared to the BE scheme. For parameter $\theta$, 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 $\theta$. 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 $\theta$ 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 $t^{n+1}$, $\tilde{u}^{n+1}_i$, and the space time average of the solution $\tilde{\bar{u}}^{n+1}_i$. In one space dimension, using upstream weighting for transport in the positive direction, the SATH scheme is

\begin{equation}
\tilde{u}^{n+1}_i = \bar{u}_i^n - \frac{\Delta t}{\Delta x} \left[ f(\bar{u}_i^n) + \theta^{n+1}_i (f(\tilde{u}^{n+1}_i) - f(\bar{u}_i^n)) - f(\bar{u}^{n+1}_{i-1}) - \theta^{n+1}_i (f(\tilde{u}^{n+1}_{i-1}) - f(\bar{u}^{n+1}_{i-1})) \right],
\end{equation}

\begin{equation}
\tilde{\bar{u}}^{n+1}_i = \bar{u}_i^n - \frac{2\Delta t}{\Delta x} \left[ f(\bar{u}_i^n) + \theta^{n+1}_i (f(\tilde{u}^{n+1}_i) - f(\bar{u}_i^n)) - f(\bar{u}^{n+1}_{i-1}) - \theta^{n+1}_i (f(\tilde{u}^{n+1}_{i-1}) - f(\bar{u}^{n+1}_{i-1})) \right],
\end{equation}

where, as we will see,

\begin{equation}
\theta^{n+1}_i = \frac{\tilde{u}^{n+1}_i - \bar{u}_i^n}{\tilde{\bar{u}}^{n+1}_i - \bar{u}_i^n},
\end{equation}

at least when the denominator does not vanish.

This $\theta^{n+1}_i$ 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 $\theta^{n+1}_i \geq 1/2$ will also be necessary. This is not the first adaptive theta scheme to appear [2], but ours is unconditionally stable and satisfies the maximum principle in appropriate situations (see §5). Unlike explicit methods [3], 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 and conclusions.
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

\[
\tilde{u}_E(t) = \frac{1}{|E|} \int_E u(x,t) \, dx,
\]

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

\[
\tilde{u}^{n+1}_E = \tilde{u}^n_E - \frac{1}{|E|} \int_{t^n}^{t^{n+1}} \int_E \nabla \cdot f(u(x,t)) \, dx \, dt
\]

\[
= \tilde{u}^n_E - \frac{1}{|E|} \int_{t^n}^{t^{n+1}} \int_{\partial E} f(u(x,t)) \cdot \nu \, d\sigma(x) \, dt.
\]

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

\[
\tilde{u}_E(t) = \tilde{u}^n_E - \frac{1}{|E|} \int_{t^n}^{t} \int_{\partial E} f(u(x,s)) \cdot \nu \, d\sigma(x) \, ds,
\]

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 \(\tilde{u}^n_E\) and \(\tilde{u}^{n+1}_E\). One needs more information. As in multi-moment finite volume schemes [7, 6], we approximate another linear functional of the solution, namely, its space-time average defined by

\[
\bar{\mathcal{E}}^{n+1}_E = \frac{1}{\Delta t^{n+1}|E|} \int_{t^n}^{t^{n+1}} \int_E u(x,t) \, dx \, dt,
\]

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

\[
\bar{\mathcal{E}}^{n+1}_E = \frac{1}{\Delta t^{n+1}|E|} \int_{t^n}^{t^{n+1}} \tilde{u}_E(t) \, dt
\]

\[
= \tilde{u}^n_E - \frac{1}{\Delta t^{n+1}|E|} \int_{t^n}^{t^{n+1}} \int_{\partial E} f(u(x,s)) \cdot \nu \, d\sigma(x) \, ds \, dt
\]

\[
= \tilde{u}^n_E - \frac{1}{\Delta t^{n+1}|E|} \int_{t^n}^{t^{n+1}} \int_s^{t^{n+1}} f(u(x,s)) \cdot \nu \, d\sigma(x) \, ds \, dt,
\]

which gives

\[
\bar{\mathcal{E}}^{n+1}_E = \tilde{u}^n_E - \frac{1}{|E|} \int_{t^n}^{t^{n+1}} \int_{\partial E} f(u(x,t)) \cdot \nu \frac{t^{n+1} - t}{\Delta t^{n+1}} \, dt.
\]

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

\[
v(t) = \begin{cases} v^0 + v_L(t), & 0 \leq t < \tau, \\ v^1 + v_R(t), & \tau < t \leq \Delta t. \end{cases}
\]

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

\[
v^0 = v(0), \quad v^1 = v(\Delta t), \quad \tilde{v} = \frac{1}{\Delta t} \int_0^{\Delta t} v(t) \, dt.
\]

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 \( \tau^* \) we apply the same rule to the function \( g(t,v) = v \) and assume equality, i.e.,

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

which implies that the location of the discontinuity is approximated by

\[
\tau^* = \frac{v^1 - \tilde{v}}{v^1 - v^0} \Delta t, \quad \text{provided } v^1 \neq v^0.
\]

**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 \( t = \tau \in (0, \Delta t) \). With

\[
\tau^* = \begin{cases} \frac{v^1 - \tilde{v}}{v^1 - v^0} \Delta t & \text{if } v^1 \neq v^0, \\ \frac{1}{2} \Delta t & \text{if } v^1 = v^0,
\end{cases}
\]

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

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

We remark that although \( \tau \in (0, \Delta t) \), we cannot conclude the same for \( \tau^* \). Moreover, when \( v^1 = v^0 \), one could take other values for \( \tau^* \). 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

\[
\theta = 1 - \frac{\tau^*}{\Delta t} = \frac{\tilde{v} - v^0}{v^1 - v^0}.
\]
and apply DAQ using \( g(t, v) = f(v) \) to see that

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

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

\[
\int_0^{\Delta t} f(v(t)) \, dt \approx Q_0^{\Delta t}(f) = (f^0 + \theta^2 (f^1 - f^0)) \Delta t. \tag{3.9}
\]

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

**Theorem 3.3.** Suppose that \( v : \mathbb{R} \rightarrow \mathbb{R} \) 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 \( g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \) is continuous, differentiable in the second argument, and \( D_2g \) is bounded, then

\[
|v_L(t)| = \left| \int_0^t v'_L(s) \, ds \right| \leq \|v'_L\|_{L^\infty} \Delta t, \tag{3.10}
\]

where \( C \) depends only on the \( L^\infty \) norms of \( v'_L \), \( v'_R \), and \( D_2g \). Moreover, the result holds no matter how \( \tau^* \) is defined when \( v^0 = v^1 \).

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

\[
|v_L(t)| = \left| \int_0^t g(t, v^0 + v_L(t)) \, dt \right| \leq \|D_2g\|_{L^\infty} \|v'_L\|_{L^\infty} \Delta t.
\]

and similarly \( |v_R(t)| \leq \|v'_R\|_{L^\infty} \Delta t \). We compute the integral of \( g \) to the left side of the true discontinuity \( \tau \) 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

\[
|R_L| = \left| \int_0^\tau (g(t, v^0 + v_L(t)) - g(t, v^0)) \, dt \right| \leq \|D_2g\|_{L^\infty} \|v'_L\|_{L^\infty} \Delta t
\]

\[
\leq \|D_2g\|_{L^\infty} \|v'_R\|_{L^\infty} \Delta t^2.
\]

We get a similar estimate of the integral of \( g \) to the right side of \( \tau \), namely

\[
\int_\tau^{\Delta t} g(t, v(t)) \, dt = \int_\tau^{\Delta t} g(t, v^1) \, dt + R_R, \quad |R_R| \leq \|D_2g\|_{L^\infty} \|v'_R\|_{L^\infty} \Delta t^2.
\]

Therefore, the quadrature error is

\[
|\int_0^{\Delta t} g(t, v(t)) \, dt - \int_0^{\Delta t} g(t, v^0) \, dt - \int_\tau^{\Delta t} g(t, v^1) \, dt| = |\int_0^\tau g(t, v^0) \, dt + \int_\tau^{\Delta t} g(t, v^1) \, dt + R_L - \int_0^{\Delta t} g(t, v^0) \, dt - \int_\tau^{\Delta t} g(t, v^1) \, dt|
\]

\[
\leq \|D_2g\|_{L^\infty} (|v^0 - v^1| (\tau - \tau^*) + \|D_2g\|_{L^\infty} (\|v'_L\|_{L^\infty} + \|v'_R\|_{L^\infty})) \Delta t^2.
\]
It remains only to estimate $\tau - \tau^*$. We note that
\[
\Delta t \hat{v} = \int_0^{\Delta t} v(t) \, dt = \tau v^0 + (\Delta t - \tau) v^1 + \int_0^\tau v_L(t) \, dt + \int_\tau^{\Delta t} v_R(t) \, dt,
\]
so using (3.11),
\[
|\Delta t \hat{v} - \tau v^0 - (\Delta t - \tau) v^1| \leq (\|v_L'\|_{L^\infty} + \|v_R'\|_{L^\infty}) \Delta t^2.
\]
Recalling (3.3), we conclude that
\[
\|\hat{R} - \tau v^0\|_{L^\infty} \leq (\|v_L'\|_{L^\infty} + \|v_R'\|_{L^\infty}) \Delta t^2.
\]
Combining this with (3.12) completes the proof. \(\Box\)

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.

\[\text{3.2. Accuracy of DAQ in the case of smooth functions.}\]

**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

\[
(3.13) \quad \left| \int_0^{\Delta t} g(t, v(t)) \, dt - Q_0^{\Delta t}(g) \right| \leq C \Delta t^3,
\]

where \(C\) depends only on the \(L^\infty\) norms of \(v', v'', Dg, \text{ and } D^2g\). Moreover, the result holds no matter how \(\tau^*\) is defined when \(v^0 = v_1\).

**Proof.** We first recall that the trapezoidal rule applied to a function \(\varphi(t)\) satisfies
\[
\int_0^{\Delta t} \varphi(t) \, dt = \frac{1}{2} (\varphi(0) + \varphi(\Delta t)) \Delta t + R_T(\varphi), \quad |R_T(\varphi)| \leq \frac{1}{12} \|\varphi''\|_{L^\infty} \Delta t^3.
\]

Therefore the DAQ error is
\[
E = \int_0^{\Delta t} g(t, v(t)) \, dt - Q_0^{\Delta t}(g) = \frac{1}{2} \left( g(0, v^0) + g(\Delta t, v^1) \right) \Delta t - Q_0^{\Delta t}(g) + R_T(g(\cdot, v)),
\]
\[
|R_T(g(\cdot, v))| \leq \frac{1}{12} \|D^2(g(\cdot, v))\|_{L^\infty} \Delta t^3.
\]

If \(v^0 = v^1\), then \(E\) simplifies independently of the value of \(\tau^*\) to
\[
E = \frac{1}{2} \left( g(0, v^0) + g(\Delta t, v^0) \right) \Delta t - Q_0^{\Delta t}(g(\cdot, v^0)) + R_T(g(\cdot, v))
\]
\[
= \frac{1}{2} \left( g(0, v^0) + g(\Delta t, v^0) \right) \Delta t - \int_0^{\Delta t} g(t, v^0) \, dt + R_T(g(\cdot, v))
\]
\[
= R_T(g(\cdot, v)) + R_T(g(\cdot, v)),
\]

and we conclude the bound stated in the theorem.

If \(v^0 \neq v^1\), we compute
\[
E = \frac{1}{2} \left( g(0, v^0) + g(\Delta t, v^1) \right) \Delta t - \int_0^{\Delta t} g(t, v^0) \, dt - \int_0^{\Delta t} g(t, v^1) \, dt + R_T(g(\cdot, v))
\]
\[
= \int_0^{\Delta t/2} \left( g(0, v^0) - g(t, v^0) \right) \, dt + \int_{\Delta t/2}^{\Delta t} \left( g(\Delta t, v^1) - g(t, v^1) \right) \, dt
\]
\[
- \int_{\Delta t/2}^{\Delta t} \left( g(t, v^0) - g(t, v^1) \right) \, dt + R_T(g(\cdot, v))
\]
\[
= E_1 + E_2 + E_3 + R_T(g(\cdot, v)),
\]
respectively.

We first estimate $E_1 + E_2$ by computing

\[
E_1 + E_2 = \int_0^{\Delta t/2} (g(0, v^0) - g(t, v^0)) \, dt + \int_{\Delta t/2}^{\Delta t} (g(\Delta t, v^1) - g(t, v^1)) \, dt \\
= -\int_0^{\Delta t/2} \int_0^t D_1g(s, v^0) \, ds \, dt + \int_{\Delta t/2}^{\Delta t} \int_t^1 D_1g(s, v^1) \, ds \, dt \\
= \int_0^{\Delta t/2} (D_1g(\Delta t - s, v^1) - D_1g(s, v^0)) \, ds \, dt,
\]

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

\[
|E_1 + E_2| \leq \int_0^{\Delta t/2} \int_0^t |DD_1g|_{L^\infty} (|\Delta t - 2s| + |v^1 - v^0|) \, ds \, dt \\
\leq \frac{1}{4} |D^2g|_{L^\infty} (1 + ||v'||_{L^\infty}) \Delta t^3.
\]

Now for $E_3$, we see that

\[
|E_3| = \left| \int_{\Delta t/2}^{\tau^*} (g(t, v^0) - g(t, v^1)) \, dt \right| \leq ||D_2g||_{L^\infty} |(v^0 - v^1)(\tau^* - \Delta t/2)|.
\]

Moreover,

\[
(v^0 - v^1)(\tau^* - \frac{1}{2} \Delta t) = (v^0 - v^1) \left( \frac{v^1 - \hat{v}}{v^1 - v^0} - \frac{1}{2} \right) \Delta t \\
= \left( \hat{v} - \frac{1}{2}(v^0 + v^1) \right) \Delta t = R_T(v),
\]

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

\[
|E_3| \leq \frac{1}{12} ||D_2g||_{L^\infty} ||v'||_{L^\infty} \Delta t^3,
\]

and the proof is complete. $\square$

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

\[
\tau^* = \frac{1}{2} \Delta t + \frac{\hat{v} - \frac{1}{2}(v^0 + v^1)}{v^0 - v^1} \Delta t.
\]

The absolute value of the deviation of $\tau^*$ from $\Delta t/2$ could be quite large, for example at a minimum or maximum in $v$, where $\hat{v} \neq \frac{1}{2}(v^0 + v^1)$ 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

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

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 $I_i$. 

\[
E
\]
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

\begin{align}
\bar{u}_{i+1}^n &= \bar{u}_i^n - \frac{1}{\Delta x_i} \int_{t^n}^{t^{n+1}} (\hat{f}_{i+1/2} - \hat{f}_{i-1/2}) \, dt, \\
\tilde{\bar{u}}_{i+1}^n &= \bar{u}_i^n - \frac{1}{\Delta x_i} \int_{t^n}^{t^{n+1}} (\hat{f}_{i+1/2} - \hat{f}_{i-1/2}) \frac{\Delta t^{n+1} - t}{\Delta t^{n+1}} \, dt.
\end{align}

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

\begin{align}
\bar{u}_{i+1}^n &= \bar{u}_i^n - \Delta t^{n+1} \Delta x_i \left[ \bar{f}_i^n + \theta^{n+1}_i (\bar{f}_{i+1}^n - \bar{f}_{i-1}^n) - \bar{f}_{i-1}^{n+1} - \theta^{n+1}_i (\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) \right], \\
\tilde{\bar{u}}_{i+1}^n &= \bar{u}_i^n - \frac{\Delta t^{n+1}}{2 \Delta x_i} \left[ \bar{f}_i^n + (\theta^{n+1}_i)^2 (\bar{f}_{i+1}^n - \bar{f}_{i}^n) - \bar{f}_{i-1}^{n+1} - (\theta^{n+1}_i)^2 (\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n) \right],
\end{align}

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

\begin{align}
\theta^n_i &= \begin{cases} 
\max \left( \frac{1}{2}, \frac{\bar{u}_{i+1}^n - \bar{u}_i^n}{\bar{u}_{i+1}^n - \bar{u}_i^n} \right) & \text{if } |\bar{u}_{i+1}^n - \bar{u}_i^n| > \epsilon, \\
\theta^* & \text{if } |\bar{u}_{i+1}^n - \bar{u}_i^n| \leq \epsilon.
\end{cases}
\end{align}

The restriction $\theta_i^{n+1} \geq 1/2$ will be explained in §5.1. We take $\epsilon$ 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. The numerical flux is

\begin{equation}
\hat{f}(u^-, u^+) = \frac{1}{2} \left[ f(u^-) + f(u^+) - \alpha_{LF}(u^+ - u^-) \right], \quad \alpha_{LF} = \max_u |f'(u)|,
\end{equation}

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

\begin{align}
\tilde{u}_{i+1/2}^- &= \bar{u}_i \quad \text{and} \quad \tilde{u}_{i+1/2}^+ = \bar{u}_{i+1}.
\end{align}
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

\[ \begin{align*}
\bar{u}^{n+1}_i &= \bar{u}_i^n - \frac{\Delta t^{n+1}}{2\Delta x_i} \left\{ \tilde{f}^{n+1}_i + \delta_{i+1} \bar{f}^{n+1}_{i+1} - \tilde{f}^n_i - \theta_{i+1} f^{n+1}_{i+1} (\tilde{f}^{n+1}_i - \tilde{f}^n_i) \\
&\quad - \alpha_{LF} \left[ \bar{u}^{n+1}_{i+1} + \theta_{i+1} (\bar{u}^{n+1}_{i+1} - \bar{u}^n_{i+1}) - 2\bar{u}^n_i - 2\theta_i^{n+1} (\bar{u}^{n+1}_i - \bar{u}_i^n) \\
&\quad + \bar{u}_i^{n+1} + \theta_i^{n+1} (\bar{u}_i^{n+1} - \bar{u}_i^n) \right] \right\}, \\
\hat{u}^{n+1}_i &= \hat{u}_i^n - \frac{\Delta t^{n+1}}{4\Delta x_i} \left\{ \tilde{f}^{n+1}_i + (\theta_{i+1}^{n+1})^2 (\bar{f}^{n+1}_{i+1} - \tilde{f}^{n+1}_i) - \tilde{f}^n_i - (\theta_{i+1}^{n+1})^2 (\bar{f}^{n+1}_{i+1} - \tilde{f}^n_i) \\
&\quad - \alpha_{LF} \left[ \bar{u}^{n+1}_{i+1} + (\theta_{i+1}^{n+1})^2 (\bar{u}^{n+1}_{i+1} - \bar{u}^n_{i+1}) - 2\bar{u}^n_i - 2(\theta_i^{n+1})^2 (\bar{u}_i^{n+1} - \bar{u}_i^n) \\
&\quad + \bar{u}_i^{n+1} + (\theta_i^{n+1})^2 (\bar{u}_i^{n+1} - \bar{u}_i^n) \right] \right\},
\end{align*} \]

where (4.6) defines \( \theta_i^{n+1} \).

5. Properties of the upstream weighted scheme (SATh-up). Analysis of the scheme using Lax-Friedrichs stabilization (4.9)–(4.10) 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 \( \theta_i^{n+1} \).

**Theorem 5.1.** Assume that \( f(0) = 0 \) and \( f'(u) > 0 \) for \( u \neq 0 \). 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

\[ \theta_i^{n+1} \geq \frac{1}{2} - \frac{\Delta x_i}{\alpha_{LF} \Delta t^{n+1}}. \]

Moreover, the scheme is unconditionally stable and independent of all problem parameters provided that \( \theta_i^{n+1} \geq 1/2 \).

**Proof.** Let

\[ \delta_i^n = \begin{cases} \frac{\tilde{f}^n_i}{\alpha_{LF} \bar{u}_i^n} & \text{if } \bar{u}_i^n \neq 0, \\ 0 & \text{otherwise}, \end{cases} \]

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

\[ \tilde{f}_i^{n+1} = \alpha_{LF} \delta_i^n \bar{u}_i^{n+1}. \]
In this expanded form, we may view the scheme in terms of the variables

\[ \xi^n = (\ldots, \tilde{u}^n_{i-1}, \tilde{f}^n_{i-1}, \tilde{u}^n_i, \tilde{f}^n_i, \tilde{u}^n_i, \ldots) \]

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 \( 3 \times 3 \) 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 \( i \)th blocks are

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

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

\[
\frac{1 - \alpha_{LF} \hat{\lambda}_i (1 - \theta_i^{n+1}) \delta_i^{n+1}}{1 + \alpha_{LF} \hat{\lambda}_i \theta_i^{n+1} \delta_i^{n+1}}.
\]

The scheme is stable to rounding error provided that the absolute values of the eigenvalues are bounded by 1; that is, if

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

The upper bound holds trivially, and the lower bound holds if and only if

\[ \theta_i^{n+1} \geq \frac{\alpha_{LF} \hat{\lambda}_i \delta_i^{n+1} - \frac{1}{2} - \frac{1}{\alpha_{LF} \hat{\lambda}_i \delta_i^{n+1}}}{2 \alpha_{LF} \hat{\lambda}_i \delta_i^{n+1}}. \]

Since \( \delta_i^n \in [0, 1] \), the proof is complete. \( \square \)

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 \( \theta_i^{n+1} \) 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 \( f'(u) > 0 \) and \( \epsilon = 0 \) in (4.6), and that the SATh-up scheme (4.4)–(4.5) is posed on a finite interval with a boundary condition imposed on the left (so \( \tilde{u}^0_n \) is given for all \( n \)). If the boundary and initial conditions of the flow satisfy the monotone decreasing property

\[
\tilde{u}^0_n \leq \tilde{u}^{n+1}_n \quad \forall n \geq 0 \quad \text{and} \quad \tilde{u}^0_i \leq \tilde{u}^0_{i-1} \quad \forall i \geq 1,
\]

then the scheme satisfies the maximum principle in the sense that

\[
\tilde{u}^n_i \leq \tilde{u}^{n+1}_i \leq \tilde{u}^{n+1}_{i-1} \quad \forall n \geq 0, \quad i \geq 1.
\]

Moreover, if the inequalities involving \( \tilde{u} \) are reversed, so that

\[
\tilde{u}^0_n \geq \tilde{u}^{n+1}_n \quad \forall n \geq 0 \quad \text{and} \quad \tilde{u}^0_i \geq \tilde{u}^0_{i-1} \quad \forall i \geq 1,
\]

then

\[
\tilde{u}^n_i \geq \tilde{u}^{n+1}_i \geq \tilde{u}^{n+1}_{i-1} \quad \forall n \geq 0, \quad i \geq 1.
\]
Proof. We prove the theorem by an inductive argument. The result (5.6) holds initially where \((i, n) = (i, -1)\) \(\forall i \geq 1\) and on the boundary where \((i, n) = (0, n)\) \(\forall n \geq 0\), provided we define \(\bar{\tilde{u}}_{i}^{-1} = \bar{u}_{0}^{i}\) and \(\bar{\tilde{u}}_{i}^{n+1} = \bar{u}_{0}^{n+1}\). 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 \(i \geq 1\) and \(n \geq 0\), we make the induction hypothesis

\[
\bar{\tilde{u}}_{i}^{-1} \leq \bar{u}_{i}^{-1} \quad \text{and} \quad \bar{\tilde{u}}_{i,n}^{n} \leq \bar{u}_{i,n}^{n} \quad \text{(i.e.,} \quad \bar{\tilde{u}}_{i}^{-1} \leq \bar{u}_{i,n}^{n} \leq \bar{u}_{i,n}^{n+1})
\]

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

\[
\bar{\tilde{f}}_{i,n}^{-1} \leq \bar{\tilde{f}}_{i,n}^{n} \leq \bar{\tilde{f}}_{i,n}^{n+1}.
\]

In the case that \(\bar{\tilde{u}}_{i,n}^{n+1} = \bar{u}_{i,n}^{n}\), it is trivial to check that the induction continues. So we consider the case when \(\bar{\tilde{u}}_{i,n}^{n+1} \neq \bar{u}_{i,n}^{n}\). To handle the nonlinearity in \(f\), we define

\[
\delta_{i} = \frac{\bar{\tilde{f}}_{i,n}^{-1} - \bar{\tilde{f}}_{i,n}^{n}}{\bar{\tilde{u}}_{i,n}^{n+1} - \bar{\tilde{u}}_{i,n}^{n}} > 0,
\]

suppressing the index \(n\). To handle the lower bound on \(\theta_{i,n}^{n+1}\), we define

\[
\eta_{i} = \theta_{i,n}^{n+1} - \frac{\bar{\tilde{u}}_{i,n}^{n+1} - \bar{\tilde{u}}_{i,n}^{n}}{\bar{\tilde{u}}_{i,n}^{n+1} - \bar{\tilde{u}}_{i,n}^{n}} \geq 0,
\]

so that \(\bar{\tilde{u}}_{i,n}^{n+1} - \bar{\tilde{u}}_{i,n}^{n} = w_{i}(\theta_{i,n}^{n+1} - \eta_{i})\), where we find it convenient to define \(w_{i} = \bar{\tilde{u}}_{i,n}^{n+1} - \bar{\tilde{u}}_{i,n}^{n}\), and also \(\lambda = \Delta t^{n+1}/\Delta x_{i}\). Then (4.4)–(4.5) can be written as

\[
\begin{align*}
(1 + \lambda \theta_{i,n}^{n+1} \delta_{i}) w_{i} = & -\lambda [((\bar{\tilde{f}}_{i,n}^{n}) - \bar{\tilde{f}}_{i,n}^{n+1}) - \theta_{i,n}^{n+1} (\bar{\tilde{f}}_{i,n}^{n+1} - \bar{\tilde{f}}_{i,n}^{n})] = -\lambda A, \\
2 w_{i} [\theta_{i,n}^{n+1} - \eta_{i} + (\lambda/2)(\theta_{i,n}^{n+1})^{2} \delta_{i}] = & -\lambda [(\bar{\tilde{f}}_{i,n}^{n}) - (\theta_{i,n}^{n+1})^{2} (\bar{\tilde{f}}_{i,n}^{n+1} - \bar{\tilde{f}}_{i,n}^{n})] = -\lambda B.
\end{align*}
\]

By (5.10), we conclude that \(A \leq 0\) and \(B \leq 0\). Moreover, (5.11) implies that \(w_{i} \geq 0\), i.e., \(\bar{\tilde{u}}_{i,n}^{n} \leq \bar{\tilde{u}}_{i,n}^{n+1}\), which is half of what must be shown in (5.6).

Substitute \(w_{i}\) from (5.11) into (5.12) to obtain that

\[
2 [\theta_{i,n}^{n+1} - \eta_{i} + (\lambda/2)(\theta_{i,n}^{n+1})^{2} \delta_{i}] A = (1 + \lambda \theta_{i,n}^{n+1} \delta_{i}) B.
\]

This is a quadratic equation in \(\theta_{i,n}^{n+1}\),

\[
(\lambda A \delta_{i})(\theta_{i,n}^{n+1})^{2} + (2A - \lambda B \delta_{i}) \theta_{i,n}^{n+1} - (2A \eta_{i} + B) = 0.
\]

If \(A = 0\), then (5.11) implies that \(w_{i} = 0\). If \(A \neq 0\) but \(\delta_{i} = 0\), then \(\bar{\tilde{f}}_{i,n}^{n+1} = \bar{\tilde{f}}_{i,n}^{n}\), and \(\bar{\tilde{u}}_{i,n}^{n+1} = \bar{\tilde{u}}_{i,n}^{n}\) (i.e., \(w_{i} = 0\)). But we are working in the case where \(w_{i} \neq 0\), so we conclude that \(A < 0\) and \(\delta_{i} > 0\), and that the equation for \(\theta_{i,n}^{n+1}\) is strictly quadratic. The two solutions are

\[
\theta_{i,n}^{n+1} = \frac{1}{2\lambda A \delta_{i}} \left[ -2A + \lambda B \delta_{i} \pm \sqrt{(2A - \lambda B \delta_{i})^{2} + 4\lambda A \delta_{i}(2A \eta_{i} + B)} \right] = \frac{1}{2\lambda A \delta_{i}} \left[ -2A + \lambda B \delta_{i} \pm \sqrt{(2A)^{2} + (\lambda B \delta_{i})^{2} + 8\lambda A \delta_{i}(2A \eta_{i} + B)} \right].
\]
Since \( \theta_i^{n+1} \geq 1/2 \), we must take the solution which subtracts the square root. Then

\[
1 + \lambda \theta_i^{n+1} \delta_i = \frac{\lambda B \delta_i}{2A} + \sqrt{1 + \left( \frac{\lambda B \delta_i}{2A} \right)^2 + 2\lambda \delta_i \eta_i} \geq \frac{\lambda B \delta_i}{A} > 0.
\]

Returning to (5.11), we have that

\[
w_i = -\frac{\lambda A}{1 + \lambda \theta_i^{n+1} \delta_i} \leq -\frac{A^2}{B \delta_i}
\]

\[
= \left[ \frac{\theta_i^{n+1} (\bar{f}_i^{n+1} - \bar{f}_i^n) - (f_i^n - f_i^{n-1})}{(\theta_i^{n+1})^2 (f_i^{n+1} - f_i^n) - (f_i^n - f_i^{n-1})} \right] \delta_i = \frac{1}{\delta_i} g(\theta_i^{n+1}).
\]

The function \( g(\theta) = \frac{(a \theta + b)^2}{a \theta^2 + b} \) (where in our case \( a = (\bar{f}_i^{n+1} - \bar{f}_i^n) > 0 \) and \( b = -(f_i^n - f_i^{n-1}) \geq 0 \)) achieves its maximum on \([1/2, \infty)\) at \( \theta = 1 \). Therefore

\[
w_i \leq \frac{1}{\delta_i} \left( |\bar{f}_i^{n+1} - \bar{f}_i^n| - |f_i^n - f_i^{n-1}| \right) = \frac{\bar{f}_i^{n+1} - \bar{f}_i^n}{\bar{f}_i^{n+1} - f_i^n} w_i,
\]

and we conclude that \( 1 \leq \frac{\bar{f}_i^{n+1} - \bar{f}_i^n}{\bar{f}_i^{n+1} - f_i^n} \). Since the numerator is positive by (5.10), so also is the denominator, and we conclude that

\[
\bar{f}_i^{n+1} - f_i^n \leq \bar{f}_i^{n+1} - f_i^n \text{ and then } \bar{f}_i^{n+1} \leq \bar{f}_i^{n+1}.
\]

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

For the reverse inequalities (5.7), \( A > 0 \) and \( B \geq 0 \), and an entirely similar argument gives the result (5.8). \( \Box \)

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

**Corollary 5.3.** Assume the hypotheses of Theorem 5.2. If \( \bar{u}_i^{n+1} \) satisfies the monotonicity property that it lies between \( \bar{u}_i^n \) and \( \bar{u}_i^{n+1} \) \( \forall n \geq 0 \), then \( \theta_i^{n+1} \in [1/2, 1] \) \( \forall n \geq 0 \), \( i \geq 1 \). Moreover, if \( \theta_i = 1 \) in (4.6), then \( \theta_i^{n+1} \) 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.** If \( \theta_i^{n+1} > 1/2 \), then \( \eta_i = 0 \). By the monotonicity assumption on \( \bar{u}_i^{n+1} \), \( \theta_0^{n+1} \leq 1 \forall n \geq 0 \), so we conclude by induction on \( i \) that

\[
B = \left( \frac{\theta_i^{n+1}}{2A} \right)^2 (\bar{f}_i^{n+1} - \bar{f}_i^n) - (f_i^n - f_i^{n-1}) \leq \frac{1}{2},
\]

where \( A \) and \( B \) are defined in (5.11)–(5.12). Returning to (5.13), we have that

\[
\theta_i^{n+1} = \frac{B}{2A} + \sqrt{\left( \frac{1}{\lambda \delta_i} \right)^2 + \left( \frac{B}{2A} \right)^2} - \frac{1}{\lambda \delta_i} \leq \frac{1}{2} + \sqrt{\left( \frac{1}{\lambda \delta_i} \right)^2 + \frac{1}{4}} - \frac{1}{\lambda \delta_i} \leq 1.
\]

Now, provided that \( \bar{u}_i^{n+1} \neq \bar{u}_i^n \),

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

\[
\tilde{u}_i^{n+1} = \tilde{u}^{n}_i - \frac{\lambda}{2} \left( f_i^n - \hat{f}_{i-1}^{n} + \frac{1}{4} (f_i^{n+1} - \hat{f}_i^n) - (\theta_i^{n+1})^2 (\hat{f}_{i+1}^{n+1} - \hat{f}_{i-1}^{n}) \right).
\]

The right side is minimized by \( \theta_i^{n+1} = 1/2 \), so

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

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

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

In the monotone decreasing or increasing cases of Theorem 5.2, it is straightforward to compute the total variation (TV) of \( \tilde{u}^n \). In the monotone decreasing case (5.5), it is

\[
TV(\tilde{u}^n) = \sum_{i=1}^{\infty} |\tilde{u}^{n}_{i-1} - \tilde{u}^n_i| = \sum_{i=1}^{\infty} (\tilde{u}^{n}_{i-1} - \tilde{u}^{n}_i).
\]

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 \( |\tilde{u}_i^n| \leq M \) and \( |\tilde{u}_i^0| \leq M \) for all \( n \geq 0 \) and \( i \geq 0 \), then the SATh-up scheme is TVB, i.e.,

\[
TV(\tilde{u}^n) \leq 2M.
\]

Moreover, if also \( \tilde{u}_i^{n+1} = \tilde{u}^n_i \), then the scheme is TVD, i.e.,

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

**Proof.** In the monotone decreasing case (5.5), \( \tilde{u}_i^n \leq \tilde{u}_i^{n+1} \), which implies that for all \( i, -M \leq \tilde{u}_i^0 \leq \tilde{u}_i^1 \leq \cdots \leq \tilde{u}_i^n \). The sum in (5.14) collapses, so

\[
TV(\tilde{u}^n) = \lim_{i_{\max} \to \infty} \sum_{i=1}^{i_{\max}} (\tilde{u}^{n}_{i-1} - \tilde{u}^n_i) \leq \tilde{u}^0_{i_{\max}} - \liminf_{i_{\max} \to \infty} \tilde{u}_i^n \leq 2M.
\]

Moreover, when \( \tilde{u}_i^n = \tilde{u}^n_i \),

\[
TV(\tilde{u}^n) - TV(\tilde{u}_i^{n+1}) = \lim_{i_{\max} \to \infty} \sum_{i=1}^{i_{\max}} [(\tilde{u}^{n}_{i-1} - \tilde{u}^n_i) - (\tilde{u}^{n+1}_{i-1} - \tilde{u}^{n+1}_i)] \\
= \lim_{i_{\max} \to \infty} \sum_{i=1}^{i_{\max}} [(\tilde{u}^{n}_{i-1} - \tilde{u}^{n+1}_{i-1}) - (\tilde{u}^{n}_i - \tilde{u}^{n+1}_i)] \\
\geq (\tilde{u}^n_0 - \tilde{u}_0^{n+1}) - \limsup_{i_{\max} \to \infty} (\tilde{u}^{n}_{i_{\max}} - \tilde{u}^{n+1}_{i_{\max}}) \\
= \liminf_{i_{\max} \to \infty} (\tilde{u}^{n+1}_{i_{\max}} - \tilde{u}^n_{i_{\max}}) \geq 0.
\]

The monotone increasing case (5.7) is shown in a similar way. \( \square \)
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. We extend our SATh scheme by illustrating the ideas for the scalar equation in two space dimensions, namely,

\begin{equation}
    u_t + (f(u))_x + (g(u))_y = 0, \quad (x, y) \in \mathbb{R}^2, \quad t > 0.
\end{equation}

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 \( \tau_{ij}^* \). Moreover, we assume that the solution is constant in space and in time on either side of \( \tau_{ij}^* \). To determine \( \tau_{ij}^* \), consider that

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

which implies that

\begin{equation}
\tau_{ij}^* - t^n = \frac{\tilde{u}_{ij}^{n+1} - \tilde{u}_{ij}^n}{\tilde{u}_{ij}^{n+1} - \tilde{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{u}_{ij}^{n+1} - \tilde{u}_{ij}^n}{\tilde{u}_{ij}^{n+1} - \tilde{u}_{ij}^n}.
\end{equation}

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

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

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

\begin{equation}
\tilde{u}_{ij}^{n+1} = \tilde{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.
\end{equation}

We have reduced the flux integrals to two terms, each one varying only in one space dimension. We can now 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).

6.1. The SATh scheme using upstream stabilization in two space dimensions. Suppose that we have a monotone flux in the sense that \( f'(u) > 0 \) and \( g'(u) > 0 \). In that case, we can use simple one point upstream weighting stabilization. With the notation \( f_{ij}^* = f(\bar{u}(x_i, y_j, t^n)) \), \( g_{ij}^* = g(\bar{u}(x_i, y_j, t^n)) \), \( \hat{\lambda}_x = \Delta t^{n+1}/\Delta x_i \), and
\[ \dot{\lambda}_y = \frac{\Delta t^{n+1}}{\Delta y_i}, \]  
the SATh-up scheme is

\begin{equation}
\tilde{u}_{ij}^{n+1} = \bar{u}_{ij}^n - \dot{\lambda}_x \left[ f_{ij}^n + \theta_{ij}^{n+1}(\bar{f}_{ij}^{n+1} - \bar{f}_{ij}^n) - f_{i-1,j}^n - \theta_{i-1,j}^{n+1}(\bar{f}_{i-1,j}^{n+1} - \bar{f}_{i-1,j}^n) \right] \\
- \dot{\lambda}_y \left[ \bar{g}_{ij}^n + \theta_{ij}^{n+1}(\bar{g}_{ij}^{n+1} - \bar{g}_{ij}^n) - \bar{g}_{i,j-1}^n - \theta_{i,j-1}^{n+1}(\bar{g}_{i,j-1}^{n+1} - \bar{g}_{i,j-1}^n) \right],
\end{equation}

\begin{equation}
\tilde{\tilde{u}}_{ij}^{n+1} = \bar{u}_{ij}^n \\
- \frac{1}{2} \dot{\lambda}_x \left[ f_{ij}^n + (\theta_{ij}^{n+1})^2(\bar{f}_{ij}^{n+1} - \bar{f}_{ij}^n) - f_{i-1,j}^n - (\theta_{i-1,j}^{n+1})^2(\bar{f}_{i-1,j}^{n+1} - \bar{f}_{i-1,j}^n) \right] \\
- \frac{1}{2} \dot{\lambda}_y \left[ \bar{g}_{ij}^n + (\theta_{ij}^{n+1})^2(\bar{g}_{ij}^{n+1} - \bar{g}_{ij}^n) - \bar{g}_{i,j-1}^n - (\theta_{i,j-1}^{n+1})^2(\bar{g}_{i,j-1}^{n+1} - \bar{g}_{i,j-1}^n) \right].
\end{equation}

Now

\begin{equation}
\theta_{ij}^{n+1} = \left\{ \begin{array}{ll}
\max \left( \frac{1}{2}, \frac{\tilde{u}_{ij}^{n+1} - \bar{u}_{ij}^n}{\bar{u}_{ij}^n - \bar{u}_{ij}^n} \right) & \text{if } |\bar{u}_{ij}^{n+1} - \bar{u}_{ij}^n| > \epsilon, \\
\theta^* & \text{if } |\bar{u}_{ij}^{n+1} - \bar{u}_{ij}^n| \leq \epsilon.
\end{array} \right.
\end{equation}

6.2. The SATh scheme using Lax-Friedrichs stabilization in two space dimensions. Lax-Friedrichs stabilization can be applied component-wise. Let

\begin{equation}
\alpha_{LF} = \max_u |f(u)| \quad \text{and} \quad \beta_{LF} = \max_u |g'(u)|.
\end{equation}

We use one point upstream weighting to define the one-sided limits of the solution on the right, left, bottom, and top of \( I_{ij} \). Applying DAQ, (2.2) (i.e., (6.3)) and (2.5) become

\begin{equation}
\tilde{u}_{ij}^{n+1} = \bar{u}_{ij}^n - \frac{\Delta t^{n+1}}{2\Delta x_i} \mathcal{F}_{ij}^n(\theta^{n+1}) - \frac{\Delta t^{n+1}}{2\Delta y_j} \mathcal{G}_{ij}^n(\theta^{n+1}),
\end{equation}

\begin{equation}
\tilde{\tilde{u}}_{ij}^{n+1} = \bar{u}_{ij}^n - \frac{\Delta t^{n+1}}{4\Delta x_i} \mathcal{F}_{ij}^n((\theta^{n+1})^2) - \frac{\Delta t^{n+1}}{4\Delta y_j} \mathcal{G}_{ij}^n((\theta^{n+1})^2),
\end{equation}

where

\[ \mathcal{F}_{ij}^n(\theta) = f_{i+1,j}^n + \theta_{i+1,j}(f_{i+1,j}^n - f_{i,j+1}^n) - f_{i-1,j}^n - \theta_{i-1,j}(f_{i-1,j}^n - f_{i,j}^n) \\
- \alpha_{LF} \left[ \bar{u}_{i+1,j}^n + \theta_{i+1,j} (\bar{u}_{i+1,j}^n - \bar{u}_{i,j+1}^n) - 2\bar{u}_{i,j}^n - 2\theta_{ij} (\bar{u}_{i,j}^n - \bar{u}_{ij}^n) \right] \\
+ \bar{u}_{i-1,j}^n + \theta_{i-1,j} (\bar{u}_{i-1,j}^n - \bar{u}_{i,j-1}^n), \]

\[ \mathcal{G}_{ij}^n(\theta) = \bar{g}_{i,j+1}^n + \theta_{i,j+1} (\bar{g}_{i,j+1}^n - \bar{g}_{i,j}^n) - \bar{g}_{i,j-1}^n - \theta_{i,j-1} (\bar{g}_{i,j-1}^n - \bar{g}_{i,j}^n) \\
- \beta_{LF} \left[ \bar{u}_{i,j+1}^n + \theta_{i,j+1} (\bar{u}_{i,j+1}^n - \bar{u}_{i,j}^n) - 2\bar{u}_{i,j}^n - 2\theta_{ij} (\bar{u}_{i,j}^n - \bar{u}_{ij}^n) \right] \\
+ \bar{u}_{i,j-1}^n + \theta_{i,j-1} (\bar{u}_{i,j-1}^n - \bar{u}_{i,j-1}^n). \]

7. Numerical results. We have theoretical proof that the SATh-up scheme works well, but it is restricted to monotone flux functions. We therefore show numerical results for the more general SATh-LF scheme (4.9)–(4.10), or (6.8)–(6.9) in two space dimensions. Although we have no theory for this scheme, we will see that it satisfies the results obtained for SATh-up.

In most of our figures, we plot the solution to the SATh-LF scheme \( \bar{u} \) as a black line and \( \tilde{u} \) as a black dotted line. We also compare this solution with the standard Lax-Friedrichs stabilized scheme using Crank-Nicolson (CN) time stepping in red and using backward Euler (BE) in blue. Since we solve for both \( \bar{u} \) and \( \tilde{u} \), we take twice
as many steps using the backward Euler scheme (so it uses half the CFL number reported for SATh-LF). We also show the value of $\theta$ 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 $u = 1$).

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 $[L_0, L_1] \times [0, t_{\text{max}}]$, 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 $\theta \geq 1/2$ is not enforced. We define $w_i = \tilde{u}_i^{n+1} - \tilde{u}_i^n$ and $v_i = \tilde{u}_i^{n+1} - \tilde{u}_i^n$ and solve for these variables using a straightforward implementation of Newton’s method with the initial guess $w_i = 0$ and $v_i = \tilde{u}_i^n - \tilde{u}_i^n$.

Recall the two parameters $\epsilon$ and $\theta^*$ in the definition of $\theta$, (4.6) or (6.6). We found that the value of $\epsilon$ has little effect on the solution, as long as it is small (say $\epsilon \leq 10^{-6}$). 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 $\theta^*$ 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 $\theta_i$ as in (4.6) using a cut-off function $\kappa$. That is, we let

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

With the choice $\kappa(\tilde{\theta}) = \max(1/2, \tilde{\theta})$, we recover the stated definition (4.6). Since then $\kappa'(\tilde{\theta}) = 0.5(1 + \text{sign}(\tilde{\theta} - 0.5))$ is not continuous, we took smoothed versions of the function $\kappa$, with the intent to improve the Newton convergence. We found that smoothing $\kappa$ had little effect on the number of iterations. However, whatever instantiation of $\kappa$ 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 (4.9)–(4.10) can handle division by $w$ in $\tilde{\theta} = v/w$, since $w$ can be quite small and even vanish. When $|w| \leq \epsilon$, 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 (4.9)–(4.10) have a form like $T = \theta^p(f(w + \tilde{u}^n) - f(\tilde{u}^n))$, $p = 1, 2$, so the derivatives can be computed as

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

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

One should implement the derivatives this way, since the quantity $f(w + \tilde{u}^n) - f(\tilde{u}^n)$ is the derivative of $f$ at some point between $w + \tilde{u}^n$ and $\tilde{u}^n$, 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 2-4 more iterations per
time step than the backward Euler scheme.

7.2. Linear transport in one space dimension. We consider first the linear
equation
\[ u_t + u_x = 0, \quad L_0 < x < L_1, \quad t > 0, \]
with unit speed, so \( \alpha_{LF} = 1 \).

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
\( t = 0.5 \) in Figure 7.1, using \( \Delta x = 1/100 \) and \( \Delta t = 1/20 \) (so the CFL number is 5
and we use 10 time steps). 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 and backward
Euler (BE) in blue (which uses twice as many time steps, so it uses \( \Delta t = 1/40 \) and
CFL 2.5).

We computed the total variation of both SATh-LF (black) and backward Euler
(blue). As we would hope from Corollary 5.4, the total variation remained one for
both schemes.

We can see that Crank-Nicolson, although stable, displays excessive oscillation,
and backward Euler displays 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. The solution remains stable and monotone
(cf. Theorems 5.1 and 5.2). For reference, the value of \( \theta \) 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 \( 1/2 \leq \theta < 1 \). (That \( \theta \leq 1 \) is consistent
with Corollary 5.3.)

We computed the total variation of both SATh-LF (black) and backward Euler
(blue). As we would hope from Corollary 5.4, the total variation remained one for
both schemes.

The computed discrete \( L^1 \) error and convergence order are given in Table 7.1.
It shows an order of convergence \( \mathcal{O}(\Delta x^{1/2}) \), as one should expect for a pure contact
discontinuity. We do not see \( \mathcal{O}(\Delta x) \), since \( \theta \geq 1/2 \) needs to be enforced for stability.

### Table 7.1: Contact discontinuity at \( t = 0.5 \), error and convergence order for SATh-LF using
\( m = 1/\Delta x \) cells and \( \Delta t = 5\Delta x \) (CFL 5).

<table>
<thead>
<tr>
<th>( m )</th>
<th>( \mathcal{L} \Delta x ) error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.505E-01</td>
<td>0.374</td>
</tr>
<tr>
<td>40</td>
<td>1.091E-01</td>
<td>0.465</td>
</tr>
<tr>
<td>80</td>
<td>7.558E-02</td>
<td>0.529</td>
</tr>
<tr>
<td>160</td>
<td>5.171E-02</td>
<td>0.548</td>
</tr>
<tr>
<td>320</td>
<td>3.542E-02</td>
<td>0.546</td>
</tr>
</tbody>
</table>
7.2.2. Convergence for a smooth problem. We test our scheme 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 scheme in the discrete \( L^1 \) and \( L^\infty \) norms at moderate (5) and high (20) CFL. In light of Theorem 3.4, one might have expected to see second order convergence. However, the accuracy of the scheme is limited to first order accuracy, due both to the requirement that \( \theta \geq 1/2 \) and to the fact that we use simple one point upstream weighting.

Table 7.2: Ex 1. Smooth linear transport error and convergence order for SATh-LF at \( t = 2 \), using \( m = 1/\Delta x \) mesh cells.

<table>
<thead>
<tr>
<th>( m )</th>
<th>( \Delta t = 5\Delta x )</th>
<th>( \Delta t = 20\Delta x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L^1_{\Delta x} )</td>
<td>( L^\infty_{\Delta x} )</td>
<td>( L^1_{\Delta x} )</td>
</tr>
<tr>
<td>40</td>
<td>3.965E-1 0.761</td>
<td>5.063E-1 0.647</td>
</tr>
<tr>
<td>160</td>
<td>1.582E-1 0.942</td>
<td>1.363E-1 0.973</td>
</tr>
<tr>
<td>640</td>
<td>4.122E-2 0.979</td>
<td>3.376E-2 1.012</td>
</tr>
<tr>
<td>1280</td>
<td>2.076E-2 0.989</td>
<td>1.674E-2 1.012</td>
</tr>
</tbody>
</table>

7.2.3. Shu’s linear test. We next consider a standard test problem [8], 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 on the backward Euler results.

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 uses 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 and BE are shown on the bottom.
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. We also show forward Euler (FE) in cyan, which of course is more accurate than the implicit schemes. The right plot uses CFL 8. As expected, there is significant numerical diffusion; however, we see significant improvement for SATH-LF (black) over BE (blue). Moreover, SATH-LF and CN (red) agree for the lower CFL example, and for the larger CFL example SATH-LF shows little degradation, but not so for BE. 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 the figure. Both the BE and SATH-LF schemes display the TVD property for this example, with SATH-LF dissipating the total variation at a better (i.e., slower) rate.

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

$$u_t + uu_x = 0, \quad L_0 < x < L_1, \; t > 0.$$  

(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, $\alpha_{LF} = 1$. 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 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.

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.

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 $\alpha_{LF} = 1$. 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
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).

Table 7.3: Burgers Riemann shock at \( t = 1 \), error and convergence order for SATh-LF, \( m = 1/\Delta x \) and \( \Delta t = 4\Delta x \) (CFL 4).

<table>
<thead>
<tr>
<th>( m )</th>
<th>( L^1_{\Delta x} ) error</th>
<th>order</th>
<th>( L^\infty_{\Delta x} ) error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>8.724E-02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>4.443E-02</td>
<td>0.973</td>
<td>1.892E-01</td>
<td>0.315</td>
</tr>
<tr>
<td>80</td>
<td>2.225E-02</td>
<td>0.998</td>
<td>1.496E-01</td>
<td>0.339</td>
</tr>
<tr>
<td>160</td>
<td>1.112E-02</td>
<td>1.000</td>
<td>1.146E-01</td>
<td>0.384</td>
</tr>
<tr>
<td>320</td>
<td>5.562E-03</td>
<td>1.000</td>
<td>8.568E-02</td>
<td>0.420</td>
</tr>
<tr>
<td>640</td>
<td>2.781E-03</td>
<td>1.000</td>
<td>6.285E-02</td>
<td>0.447</td>
</tr>
</tbody>
</table>

Table 7.4: Burgers Riemann rarefaction at \( t = 0.25 \), \( L^1 \) and \( L^\infty \) error and convergence order for SATh-LF using \( m = 1/\Delta x \) cells and \( \Delta t = 5\Delta x \) (CFL 5).

<table>
<thead>
<tr>
<th>( m )</th>
<th>( L^1_{\Delta x} ) error</th>
<th>order</th>
<th>( L^\infty_{\Delta x} ) error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7.885E-02</td>
<td></td>
<td>2.353E-01</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>4.658E-02</td>
<td>0.760</td>
<td>1.892E-01</td>
<td>0.315</td>
</tr>
<tr>
<td>80</td>
<td>2.716E-02</td>
<td>0.778</td>
<td>1.496E-01</td>
<td>0.339</td>
</tr>
<tr>
<td>160</td>
<td>1.562E-02</td>
<td>0.798</td>
<td>1.146E-01</td>
<td>0.384</td>
</tr>
<tr>
<td>320</td>
<td>8.858E-03</td>
<td>0.818</td>
<td>8.568E-02</td>
<td>0.420</td>
</tr>
<tr>
<td>640</td>
<td>4.962E-03</td>
<td>0.836</td>
<td>6.285E-02</td>
<td>0.447</td>
</tr>
</tbody>
</table>

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 \( L^\infty \) norms is given in Table 7.4 for CFL 5. It appears to be approaching a convergence rate of 1 in \( L^1 \) and \( 1/2 \) in \( L^\infty \) as \( \Delta x \) is refined.

**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 \( \alpha_{LF} = 1.5 \) and the characteristics move in both the positive and negative directions. The shock forms at time \( t/\pi = 0.318 \). 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.

![Burgers Riemann rarefaction](image1)

**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).

In Table 7.5 we give the discrete $L^1$ and $L^\infty$ 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 $t = 1/\pi = 3.18$. Results for CFL 5 and CFL 25 are presented, and both sets of results show first order convergence in $L^1$. The $L^\infty$ convergence order seems to approach one as the mesh is resolved (i.e., as the smooth but steep front is resolved).

![Burgers shock formation](image2)

**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.

In Table 7.5 we give the discrete $L^1$ and $L^\infty$ 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 $t = 1/\pi = 3.18$. Results for CFL 5 and CFL 25 are presented, and both sets of results show first order convergence in $L^1$. The $L^\infty$ convergence order seems to approach one as the mesh is resolved (i.e., as the smooth but steep front is resolved).
Table 7.5: 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.

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

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

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

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

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

$$u_0(x) = \begin{cases} 
1 - 20x & \text{for } 0 \leq x \leq 0.05, \\
0.5 & \text{for } 0.25 \leq x \leq 0.4, \\
0 & \text{otherwise.}
\end{cases}$$
Two pulses merge over time, which gives rise to an interaction of shocks and rarefactions. We use $\Delta x = 1/80$ grid elements and $\Delta t = 2\Delta x$ (CFL 4). The results at times $t = 0.2, 0.3, 0.5$ are shown in Fig. 7.7. The fine scale CN scheme ($\Delta x = 1/1280$, $\Delta t = \Delta x$) is shown in light green and is considered the reference solution. All three schemes handle the merging of the two pulses reasonably well. The CN results are sharpest, but the solution oscillates (and much worse so as the CFL number increases). All three schemes lose the second pulse by $t = 0.4$ rather than by $t = 0.5$, but the BE results are so diffuse that the second pulse is lost at $t = 0.3$. Moreover, BE dissipates the total variation faster than SATh, although both are TVD. Overall, SATh-LF reproduces the solution to adequate accuracy without oscillation even on the relatively low resolution mesh used in this test.

![Fig. 7.7: A Buckley-Leverett example of merging pulses at $t = 0.2, 0.3, 0.5$, using $\Delta x = 1/80$ and $\Delta t = 2\Delta x$ (CFL 4, although BE uses half the CFL step). The fine scale CN scheme ($\Delta x = 1/1280$, $\Delta t = \Delta x$) is used to produce the reference solution, shown in light green. Also shown is the total variation for BE and SATh.](image)

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,

$$u_t + f(u)_x = 0, \quad -0.1 < x < 0.9, \quad t > 0,$$

where

$$f(u) = \frac{64}{39} \left( u^3 - \frac{3}{2} u^2 + \frac{39}{64} u \right).$$

When $u \in [0, 1]$, one can verify that $\alpha_{LF} = 1$. The graph of $f(u)$ appears in Figure 7.8 on the top right. The flux is far from being monotone; moreover, it is not convex.

The results also appear in Figure 7.8 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, and one can see that CN has produced the wrong front speed. This is due to
Fig. 7.8: 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.

Wild oscillations that appear in the first few time steps. This trend continues to $t = 6$, shown in the bottom right.

Interestingly, the value of $\theta$ 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 $\bar{u}^{n+1}_i - \bar{u}^n_i$ is zero to rounding error and $\theta^{n+1}_i$ 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 $\theta$ rather than reverting to $\theta^*$. If one uses $\epsilon = 10^{-6}$, there are no oscillations in $\theta$ (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 $\dot{\bar{u}}$. 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. Burgers equation in two space dimensions. Finally, we consider a problem in two space dimensions, using a uniformly spaced rectangular mesh. We report results of the Lax-Friedrichs stabilized scheme SATh-LF (6.8)–(6.9) for the two dimensional Burgers equation

\begin{equation}
    u_t + \left(\frac{u^2}{2}\right)_x + \left(\frac{u^2}{2}\right)_y = 0 \quad \text{for } 0 < x < 1, \ 0 < y < 1.
\end{equation}

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, $\alpha_{LF} = \beta_{LF} = 1$.

Results for SATh-LF are shown in Figure 7.9, 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
Fig. 7.9: Burgers equation in two space dimensions. Shown is $\tilde{u}(x,y,t)$ 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.

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 linear transport and the Buckley-Leverett flux.

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 $O(\Delta t^2)$ when there is a discontinuity (Theorem 3.3), and $O(\Delta t^3)$ when the solution is smooth (Theorem 3.4), even when $v(\Delta t) = v(0)$ and $\tau$ 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, $\tilde{u}^{n+1}_{i}$, and the local space-time average, $\tilde{u}^{n+1}$. 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 $\theta$, i.e.,

$$\theta^{n+1}_i = \max \left( 1, \frac{1}{2} \frac{\tilde{u}^{n+1}_{i} - \tilde{u}^n}{\overline{u}^{n+1}_{i} - \overline{u}^n} \right).$$

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 (4.9)–(4.10). 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, suitable for direct use or in higher order flux-limited or flux corrected transport schemes.

REFERENCES