Journal of Scientific Computing (2025) 104:18
https://doi.org/10.1007/s10915-025-02938-6

Further Studies on the Self-Adaptive Theta Scheme for Conservation Laws

Todd Arbogast[1,2] · **Chieh-Sen Huang**[3] · **Danielle N. King**[1]

Received: 4 November 2024 / Revised: 1 April 2025 / Accepted: 4 May 2025

Abstract

The finite volume, self-adaptive theta (SATh) scheme was defined in Arbogast and Huang, *A self-adaptive theta scheme using discontinuity aware quadrature for solving conservation laws*, IMA J. Numer. Anal. (2022). The basic scheme evolves both the local space and space-time averages of the solution in time with an implicitly defined theta parameter. Here, the scheme is extended to unstructured meshes in multiple space dimensions, general numerical flux functions, and higher (formally second) order using WENO reconstructions. Theoretical results apply to the one space dimension, upstream weighted case, in the setting of a monotone solution. In this case, if the theta parameter is bounded below by $\theta_{\min} = 0$, it is shown that SATh is stable, L-stable for the linear problem, total variation diminishing (TVD), and maximum principle preserving (MPP). These results generalize those known previously with the assumption that $\theta_{\min} = 1/2$. Numerical tests for problems with contact discontinuities, shocks, and rarefactions show that SATh performs better than finite volume schemes using backward Euler time stepping. Moreover, SATh gives solutions about as sharp as when using Crank-Nicolson time stepping, but SATh is non-oscillatory. In cases covered by the theoretical results, SATh combined with a Lax-Friedrichs numerical flux (rather than upstream weighting) appears to be TVD and MPP. SATh is non-oscillatory if $\theta_{\min} = 1/2$, but if $\theta_{\min} = 0$ and the solution is not monotone, it can develop oscillations. The higher order SATh scheme converges to order two and compares favorably with CN, but is less oscillatory.

The first and third authors were funded in part by the U.S. National Science Foundation under grant DMS-1912735. The second author was funded by the Taiwan Ministry of Science and Technology grant MOST 109-2115-M-110-003-MY3, the Taiwan National Science and Technology Council grant NSTC 112-2115-M-110-005-MY2, and the National Center for Theoretical Sciences, Taiwan.

✉ Chieh-Sen Huang
huangcs@math.nsysu.edu.tw

Todd Arbogast
arbogast@oden.utexas.edu

Danielle N. King
dking3@math.utexas.edu

[1] Department of Mathematics, University of Texas, C1200, Austin, TX 78712–1202, USA

[2] Oden Institute for Computational Engineering and Sciences C0200, University of Texas, Austin, TX 78712–1229, USA

[3] Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 804, Taiwan, R.O.C.

Keywords Theta time stepping · Space-time average · Stability · Maximum principle · TVD · WENO · Hyperbolic transport

Mathematics Subject Classification 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

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

The law describes the fundamental principles of conservation and transport of physical quantities in a wide range of dynamic systems. There are difficulties associated with solving these problems that must be dealt with carefully in the development of numerical methods, especially that the solution may develop shocks.

Finite volume schemes inherently preserve certain physical properties like conservation due to their integral form, while also doing well to capture shocks in the solution. Perhaps the simplest finite volume approximation of (1.1) uses backward Euler (BE) time-stepping combined with upstream weighting for spatial stability. This is a low order accurate scheme with excessive numerical diffusion. Physically relevant shocks, contact discontinuities, and steep fronts are diffused over time. Despite this shortcoming, BE is used because it is unconditionally stable and is maximum principle preserving (MPP) [4, 5, 7, 9]. These desirable numerical properties make BE a viable low order scheme. Moreover, it can be used in combination with a high order scheme to improve the quality of the solution, e.g., to reduce spurious oscillations by way of flux-limiting or flux corrected transport (FCT) [3, 6, 11].

One can reduce the numerical diffusion of the BE scheme by instead resorting to the theta time-stepping method. For parameter θ, the method blends the implicit backward Euler ($\theta = 1$) and explicit forward Euler ($\theta = 0$) time-stepping. The implicit Crank-Nicolson (CN) method results when $\theta = 1/2$. The 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 $\theta < 1$.

In order to improve the low order accurate BE scheme, two of the current authors presented a discontinuity aware quadrature (DAQ) rule and used it to develop an implicit self-adaptive theta (SATh) scheme in [1]. This finite volume scheme uses the theta time-stepping method, but with an implicitly defined theta parameter given in terms of the local space and space-time averages of the solution. One has the freedom to choose the type of numerical flux used and the values of various parameters. We view SATh as being a class of methods with various minor variations allowed.

It is proved in [1] that as long as one restricts $\theta \geq \theta_{\min} = 1/2$, the finite volume approximation of (1.1) in one space dimension using SATh time-stepping combined with upstream weighting (which we denote as SATh-up) is unconditionally stable in terms of the discretization parameters for monotone flux functions. Moreover, when the true solution is monotone (increasing or decreasing), the upstream weighted scheme is MPP and the numerical solution is total variation bounded (TVB) as well as total variation diminishing (TVD). It was also observed computationally, but not proven, that SATh using a Lax-Friedrichs numerical flux satisfies the same properties.

In this paper, further theoretical and numerical studies on the SATh scheme are presented. The basic scheme is extended to multiple space dimensions on unstructured meshes. General numerical flux functions are incorporated, as long as they can be split into right and left going waves.

In one space dimension using monotone flux functions, SATh-up is shown here to be theoretically stable with only the restriction that $\theta \geq \theta_{\min} = 0$ when the true solution is monotone. This result is quite surprising, because the theta time stepping method itself may be unstable when $\theta < 1/2$. Moreover, SATh-up applied to the linear equation is L-stable, i.e., it is *linearly L-stable*. Our proof relies on an algebraically equivalent formulation of the scheme that is more symmetric. Under these restrictions, the SATh-up scheme continues to be MPP, TVB, and TVD.

Numerical tests are presented for problems with contact discontinuities, shocks, and rarefactions. When the true solution is monotone, comparisons will show that the SATh scheme using $\theta_{\min} = 0$, which is denoted SATh_0, gives better solutions than when using the original $\theta_{\min} = 1/2$, denoted $\text{SATh}_{1/2}$. The SATh solutions are less diffusive than those of BE and generally comparable to CN, but without oscillation. $\text{SATh}_{1/2}$ behaves similarly for problems with nonmonotone solutions; unfortunately, SATh_0 does not. Oscillations can develop in the SATh_0 solutions, presumably because $\theta < 1/2$ does not provide enough damping to suppress them.

The SATh scheme is extended to using spatially higher order upwind values of the solution, which results in a formally second order accurate scheme in space and time away from discontinuities. Numerically speaking, the higher order SATh scheme compares favorably with CN, but is less oscillatory.

In the next section, the self-adaptive theta scheme is recalled and then extended to general meshes in multiple dimensions and to general numerical flux functions. The stability and linear L-stability properties of SATh_0-up are proved in §3 provided that the space-time averages of the solution lie between the space averages of the solution at the two time levels. Next, §4 discusses the MPP, TVB, and TVD properties of SATh_0-up for monotone solutions. Numerical results are presented in §5 and §6 for problems posed in 1D and 2D spatial domains, respectively. These sections compare the numerical solutions of SATh_0, $\text{SATh}_{1/2}$, BE, and CN. In §7, the implications of using high order spatial reconstructions are shown. Finally, §8 ends the paper with a summary of results, conclusions, and some open questions.

2 Formulation of the Self-Adaptive Theta Scheme

Partition time $0 = t^0 < t^1 < t^2 < \cdots$ and define $\Delta t^{n+1} = t^{n+1} - t^n$ and $t^{n+1/2} = (t^n + t^{n+1})/2$.

2.1 Discontinuity Aware Quadrature (DAQ)

The SATh scheme is based on the discontinuity aware quadrature (DAQ) rule derived in [1], which is reviewed briefly here. To approximate the integral of a function $g(t, v(t))$ in time over $[t^n, t^{n+1}]$, the rule uses only the information

$$v^n = v(t^n), \quad v^{n+1} = v(t^{n+1}), \quad \text{and} \quad \tilde{v}^{n+1} = \frac{1}{\Delta t^{n+1}} \int_{t^n}^{t^{n+1}} v(t)\, dt.$$

The basic assumptions is that there is a single shock passing through the interval $[t^n, t^{n+1}]$ and that the solution is otherwise smooth in the interval. This situation is approximated by viewing $v(t)$ as if it were a piece-wise constant function on $[t^n, t^{n+1}]$ with only two values, i.e., v^n and v^{n+1}, changing at τ, an approximation to the time of the shock. Therefore,

$$\tilde{v}^{n+1} = \frac{1}{\Delta t^{n+1}}\big[(\tau - t^n)v^n + (t^{n+1} - \tau)v^{n+1}\big], \tag{2.1}$$

which gives τ. We should write $\tau = \tau^{n+1}$, but the superscript is suppressed for easier reading. The DAQ rule approximation is

$$Q_{\mathrm{DAQ}}\big(g(v)\big) = \int_{t^n}^{\tau} g(t, v^n)\,dt + \int_{\tau}^{t^{n+1}} g(t, v^{n+1})\,dt \approx \int_{t^n}^{t^{n+1}} g(t, v(t))\,dt. \tag{2.2}$$

The formulation should mirror the one parameter family of theta methods, so define

$$\theta = 1 - \frac{\tau - t^n}{\Delta t^{n+1}} = \frac{\tilde{v}^{n+1} - v^n}{v^{n+1} - v^n}, \tag{2.3}$$

i.e., $\tau = (1-\theta)\Delta t^{n+1} + t^n$. Then the DAQ rule (2.2), when applied to $G(v(t))\,w(t)$ for weight function $w(t)$, becomes

$$\begin{aligned} Q_{\mathrm{DAQ}}\big(G(v)\,w\big) = \Delta t^{n+1}\Big\{ & G(v^n)\int_0^{1-\theta} w(s\Delta t^{n+1} + t^n)\,ds \\ & + G(v^{n+1})\int_{1-\theta}^{1} w(s\Delta t^{n+1} + t^n)\,ds\Big\}. \end{aligned} \tag{2.4}$$

The rule is accurate to order $\mathcal{O}(\Delta t^2)$ when there is a discontinuity [1, Theorem 3.3]), and $\mathcal{O}(\Delta t^3)$ when the solution is smooth [1, Theorem 3.4].

2.2 The SATh Scheme

Let the region of interest be partitioned into a computational mesh of cells or elements E. The skeleton set $\Gamma = \cup_E \partial E$ consists of facets e (a point in 1D, an edge in 2D, etc.). Each facet is the intersection of two adjacent elements, $e = E^- \cap E^+$ and has a unit normal $\nu_e = \nu_{E^-}$ pointing from E^- to E^+.

The cell average of u over mesh element E at time t is given by

$$\bar{u}_E(t) = \frac{1}{|E|}\int_E u(\mathbf{x}, t)\,dx. \tag{2.5}$$

Similarly, the space-time cell average of u over $E \times [t^n, t^{n+1}]$ is given by

$$\tilde{\bar{u}}_E^{n+1} = \frac{1}{\Delta t^{n+1}|E|}\int_{t^n}^{t^{n+1}}\int_E u(\mathbf{x}, t)\,dx\,dt. \tag{2.6}$$

We will later abuse notation by using these same symbols for the *approximations* of these averages.

Begin by multiplying (1.1) by a test function $w(t)$, averaging in space over E, and applying the divergence theorem. Then integrate in time over $[t^n, t^{n+1}]$ and apply integration by parts to obtain the weak form

$$\int_{t^n}^{t^{n+1}} \frac{d\bar{u}_E(t)}{dt}\, w(t)\, dt = \bar{u}_E^{n+1} w(t^{n+1}) - \bar{u}_E^n w(t^n) - \int_{t^n}^{t^{n+1}} \bar{u}_E(t)\, w'(t)\, dt$$
$$= -\frac{1}{|E|} \int_{t^n}^{t^{n+1}} \int_{\partial E} \mathbf{f}(u(\mathbf{x}, t)) \cdot \nu_E\, dS(x)\, w(t)\, dt$$
$$= -\frac{1}{|E|} \int_{t^n}^{t^{n+1}} \sum_{e \subset \partial E} \int_e \hat{F}_e\big(u_e^-(\mathbf{x}, t), u_e^+(\mathbf{x}, t)\big) \nu_e \cdot \nu_E\, dS(x)\, w(t)\, dt, \tag{2.7}$$

wherein has been introduced the numerical flux function $\hat{F}_e(u_e^-, u_e^+) \approx \mathbf{f}(u) \cdot \nu_e$ for facet $e = E^- \cap E^+$. Here u_e^- and u_e^+ are the interface values of u on e taken as the trace from E^- and E^+, respectively. These interface values will later be approximations. To apply the DAQ rule later, the numerical flux will need to be split into unidirectional waves, which is to say

$$\hat{F}_e(u_e^-, u_e^+) = \hat{F}_e^+(u_e^-) + \hat{F}_e^-(u_e^+) \tag{2.8}$$

where $(\hat{F}_e^+)'(u) \geq 0$ and $(\hat{F}_e^-)'(u) \leq 0$.

The SATh scheme in multiple space dimensions is given below in (2.9)–(2.12). Consider the test function $w(t) = 1$ in (2.7) and apply the DAQ rule (2.4) to each piece of the numerical flux (2.8) see that

$$\bar{u}_E^{n+1} = \bar{u}_E^n - \frac{\Delta t^{n+1}}{|E|} \sum_{e \subset \partial E} \nu_e \cdot \nu_E \int_e \Big[(1 - \theta_e^-)\hat{F}_e^+(u_e^{n,-}) + \theta_e^- \hat{F}_e^+(u_e^{n+1,-})$$
$$+ (1 - \theta_e^+)\hat{F}_e^-(u_e^{n,+}) + \theta_e^+ \hat{F}_e^-(u_e^{n+1,+})\Big] dS(x). \tag{2.9}$$

The test function $w(t) = (t^{n+1} - t)/\Delta t^{n+1}$ leads similarly to

$$\tilde{\bar{u}}_E^{n+1} = \bar{u}_E^n - \frac{\Delta t^{n+1}}{2|E|} \sum_{e \subset \partial E} \nu_e \cdot \nu_E \int_e \Big[\big(1 - (\theta_e^-)^2\big)\hat{F}_e^+(u_e^{n,-}) + (\theta_e^-)^2 \hat{F}_e^+(u_e^{n+1,-})$$
$$+ \big(1 - (\theta_e^+)^2\big)\hat{F}_e^-(u_e^{n,+}) + (\theta_e^+)^2 \hat{F}_e^-(u_e^{n+1,+})\Big] dS(x). \tag{2.10}$$

The definition of θ in (2.3) results in

$$\theta_e^\pm = \theta_e^{n+1,\pm} = \begin{cases} \max\left(\theta_{\min}, \dfrac{\tilde{u}_e^{n+1,\pm} - u_e^{n,\pm}}{u_e^{n+1,\pm} - u_e^{n,\pm}}\right) & \text{if } |u_e^{n+1,\pm} - u_e^{n,\pm}| \\ & \quad > \epsilon\big(|\tilde{u}_e^{n+1,\pm} - u_e^{n,\pm}| + 1\big), \\ \theta^* & \text{otherwise,} \end{cases} \tag{2.11}$$

where $\tilde{u}_e^{n+1,\pm}$ is the trace of the time average of u on e. The restriction $\theta_e^\pm \geq \theta_{\min}$ is the stability constraint. The theory and numerical examples will illuminate when $\theta_{\min} = 0$ or $\theta_{\min} = 1/2$ is the appropriate choice. The parameter θ^* is a default value given to θ when it cannot be defined with relative accuracy due to division by a small number. It should be set to a value which is known to maintain stability, i.e., a value in the interval $[1/2, 1]$, such as $1/2$ or 1. The parameter $\epsilon \geq 0$ is a small parameter (say 1e-6 in practice) needed to avoid division by zero and numerical instability. We remark that previously in [1], ϵ was used in place of $\epsilon\big(|\tilde{\bar{u}}_j^{n+1} - u_j^n| + 1\big)$. This scaled version can reduce numerical instability, especially in the Newton solution procedure.

We generally use simple constant values to approximate the cell boundary interface values $u_e^\pm$ and $\tilde{u}_e^\pm$, i.e.,

$$u_e^\pm = \bar{u}_{E^\pm} \quad \text{and} \quad \tilde{u}_e^\pm = \tilde{\bar{u}}_{E^\pm}, \tag{2.12}$$

where $e = E^- \cap E^+$. However, the implications of using high order spatial reconstructions is discussed in §7.

2.3 The SATh-up and SATh-LF Schemes in 1D

To aid the reader, here is expressed the SATh scheme for the one space dimension equation

$$u_t + f(u)_x = 0, \quad x \in \mathbb{R}, \quad t > 0. \tag{2.13}$$

Define a computational mesh by grid points $\cdots < x_{i-1/2} < x_{i+1/2} < x_{i+3/2} < \cdots$ with grid cells $E_i = [x_{i-1/2}, x_{i+1/2}]$ of length $\Delta x_i = x_{i+1/2} - x_{i-1/2}$. The cell subscript E_i will be denoted simply as i. Given a quantity ψ that depends on space and/or time, denote $\psi^n_{i+1/2} = \psi(x_{i+1/2}, t^n)$ (and similarly when ψ is composed with u, $\psi^n_{i+1/2} = \psi(u(x_{i+1/2}, t^n))$).

The scheme (2.9)–(2.10) simplifies to

$$\begin{aligned}\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{\Delta x_i}\Big\{ & \hat{F}^{+,n}_{i+1/2} + \theta^-_{i+1/2}\big(\hat{F}^{+,n+1}_{i+1/2} - \hat{F}^{+,n}_{i+1/2}\big) \\ & + \hat{F}^{-,n}_{i+1/2} + \theta^+_{i+1/2}\big(\hat{F}^{-,n+1}_{i+1/2} - \hat{F}^{-,n}_{i+1/2}\big) \\ & - \hat{F}^{+,n}_{i-1/2} + \theta^-_{i-1/2}\big(\hat{F}^{+,n+1}_{i-1/2} - \hat{F}^{+,n}_{i-1/2}\big) \\ & - \hat{F}^{-,n}_{i-1/2} + \theta^+_{i-1/2}\big(\hat{F}^{-,n+1}_{i-1/2} - \hat{F}^{-,n}_{i-1/2}\big)\Big\},\end{aligned} \tag{2.14}$$

$$\begin{aligned}\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{2\Delta x_i}\Big\{ & \hat{F}^{+,n}_{i+1/2} + (\theta^-_{i+1/2})^2\big(\hat{F}^{+,n+1}_{i+1/2} - \hat{F}^{+,n}_{i+1/2}\big) \\ & + \hat{F}^{-,n}_{i+1/2} + (\theta^+_{i+1/2})^2\big(\hat{F}^{-,n+1}_{i+1/2} - \hat{F}^{-,n}_{i+1/2}\big) \\ & - \hat{F}^{+,n}_{i-1/2} + (\theta^-_{i-1/2})^2\big(\hat{F}^{+,n+1}_{i-1/2} - \hat{F}^{+,n}_{i-1/2}\big) \\ & - \hat{F}^{-,n}_{i-1/2} + (\theta^+_{i-1/2})^2\big(\hat{F}^{-,n+1}_{i-1/2} - \hat{F}^{-,n}_{i-1/2}\big)\Big\}.\end{aligned} \tag{2.15}$$

Using simple constant values on the cell boundary interfaces (2.12) results in

$$u^-_{j+1/2} = \bar{u}_j, \quad u^+_{j+1/2} = \bar{u}_{j+1}, \quad \tilde{u}^-_{j+1/2} = \tilde{\bar{u}}_j, \quad \text{and} \quad \tilde{u}^+_{j+1/2} = \tilde{\bar{u}}_{j+1}, \tag{2.16}$$

and in this case $\theta^-_{j+1/2} = \theta^+_{j-1/2} = \theta_j$. Moreover, θ as given by (2.11) becomes

$$\theta_j = \theta_j^{n+1} = \begin{cases} \max\left(\theta_{\min}, \dfrac{\tilde{\bar{u}}_j^{n+1} - \bar{u}_j^n}{\bar{u}_j^{n+1} - \bar{u}_j^n}\right) & \text{if } |\bar{u}_j^{n+1} - \bar{u}_j^n| \\ & \quad > \epsilon\big(|\tilde{\bar{u}}_j^{n+1} - u_j^n| + 1\big), \\ \theta^* & \text{otherwise.} \end{cases} \tag{2.17}$$

The Lax-Friedrichs stabilized numerical flux will be used. It can be split as

$$\hat{F}^{\pm}_{j+1/2} = \frac{1}{2}\big[f(u^{\mp}_{j+1/2}) \pm \alpha_{\mathrm{LF}} u^{\mp}_{j+1/2}\big], \tag{2.18}$$

where $\alpha_{\mathrm{LF}} = \max_u |f'(u)|$ is the maximum wave speed. We denote the scheme as SATh-LF when using this numerical flux.

The upstream weighted numerical flux will also be used, assuming the waves are traveling to the right so $f'(u) \geq 0$. It is split as

$$\hat{F}^+_{j+1/2} = f(u^-_{j+1/2}) \quad \text{and} \quad \hat{F}^-_{j+1/2} = 0. \tag{2.19}$$

Then (2.14)–(2.15) gives the SATh-up scheme

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{\Delta x_i}\big[(1-\theta_i)\bar{f}_i^n + \theta_i \bar{f}_i^{n+1} - (1-\theta_{i-1})\bar{f}_{i-1}^n - \theta_{i-1}\bar{f}_{i-1}^{n+1}\big], \tag{2.20}$$

$$\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t^{n+1}}{2\Delta x_i}\big[(1-\theta_i^2)\bar{f}_i^n + \theta_i^2 \bar{f}_i^{n+1} - (1-\theta_{i-1}^2)\bar{f}_{i-1}^n - \theta_{i-1}^2\bar{f}_{i-1}^{n+1}\big], \tag{2.21}$$

where $\bar{f}_i^n = f(\bar{u}_i^n)$.

3 Stability of SATh$_0$

It is shown in [1] that, under a mild monotonicity condition on the flux function, the SATh-up scheme in 1D, (2.20)–(2.21), is unconditionally stable in terms of the discretization parameters, provided one takes the lower bound $\theta_{\min} = 1/2$ in (2.17). We present a different proof that shows that in fact one may require only $\theta_{\min} \geq 0$ to achieve stability, but at the expense of the additional hypothesis that $\tilde{\bar{u}}_i^{n+1}$ lies between $\bar{u}_i^n$ and $\bar{u}_i^{n+1}$ for all $n \geq 0$ and $i \geq 1$.

The SATh-up scheme (2.20)–(2.21) is algebraically equivalent to (2.20) and (2.21) minus one half of (2.20), which is

$$\tilde{\bar{u}}_i^{n+1} = \frac{1}{2}(\bar{u}_i^n + \bar{u}_i^{n+1}) + \frac{\Delta t^{n+1}}{2\Delta x_i}\big[\theta_i(1-\theta_i)(\bar{f}_i^{n+1} - \bar{f}_i^n) - \theta_{i-1}(1-\theta_{i-1})(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\big]. \tag{3.1}$$

For the theoretical results, assume that $\epsilon = 0$ in the definition of θ, (2.17).

Theorem 1 *Assume that $f'(u) \geq 0$ and $\epsilon = 0$ in (2.17). If $\theta_{\min} = \theta^* = 1/2$, then the upstream weighted scheme (SATh-up) in 1D is unconditionally stable for the nonlinear problem. If $\theta_{\min} = \theta^* = 0$ and $\tilde{\bar{u}}_i^{n+1}$ lies between $\bar{u}_i^n$ and $\bar{u}_i^{n+1}$ for all $n \geq 0$ and $i \geq 1$, then SATh-up in 1D is unconditionally stable for the nonlinear problem and unconditionally L-stable for the linear problem.*

The first part of the theorem was shown in [1]. For the second part of the theorem, Corollary 2 gives conditions under which $\tilde{\bar{u}}_i^{n+1}$ is guaranteed to lie between $\bar{u}_i^n$ and $\bar{u}_i^{n+1}$, i.e., when the flow in monotone in the sense of Theorem 2.

3.1 Proof of Linear L-Stability of SATh$_0$-up (and SATh$_0$-LF)

We will begin by showing that the SATh$_0$-up scheme is L-stable when applied to the linear problem, for which $f(u) = \alpha_{\text{LF}} u$. Observe that the upstream weighted and Lax-Friedrichs numerical fluxes coincide for the linear problem.

Under the hypotheses, we conclude from (2.17) that $\theta_i(\bar{u}_i^{n+1} - \bar{u}_i^n) = \tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n$. The SATh-up scheme (2.20) and (3.1) applied to the linear problem is

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \hat{\lambda}\big[\tilde{\bar{u}}_i^{n+1} - \tilde{\bar{u}}_{i-1}^{n+1}\big], \tag{3.2}$$

$$\tilde{\bar{u}}_i^{n+1} = \frac{1}{2}\left(\bar{u}_i^{n+1} + \bar{u}_i^n\right) + \frac{\hat{\lambda}}{2}\big[\theta_i(\bar{u}_i^{n+1} - \tilde{\bar{u}}_i^{n+1}) - \theta_{i-1}(\bar{u}_{i-1}^{n+1} - \tilde{\bar{u}}_{i-1}^{n+1})\big], \tag{3.3}$$

where $\hat{\lambda} = \alpha_{\text{LF}} \Delta t^{n+1} / \Delta x_i > 0$. Following [1], view the scheme in matrix form, $A\xi^{n+1} = B\xi^n$, using the variables

$$\xi^n = \left(\ldots, \bar{u}_{i-1}^n, \tilde{\bar{u}}_{i-1}^n, \bar{u}_i^n, \tilde{\bar{u}}_i^n, \ldots \right)^T. \tag{3.4}$$

The matrices A and B are block 2×2 lower triangular. The eigenvalues of the matrix $A^{-1}B$ are the eigenvalues of $A_d^{-1} B_d$, where the subscript d indicates the diagonal blocks. It is straightforward to see that the i-th blocks are

$$A_d = \begin{bmatrix} 1 & \hat{\lambda} \\ -\frac{1}{2}(1 + \hat{\lambda}\theta) & 1 + \frac{\hat{\lambda}}{2}\theta \end{bmatrix} \quad \text{and} \quad B_d = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 0 \end{bmatrix},$$

where the subscript i has been repressed for readability. Assuming $\theta \geq \theta_{\min} = 0$, the determinant of A_d is nonzero, since

$$|A_d| = 1 + \frac{\hat{\lambda}}{2}(1 + \theta) + \frac{\hat{\lambda}^2}{2}\theta > 0. \tag{3.5}$$

Thus, A_d is invertible and

$$A_d^{-1} = \frac{1}{|A_d|} \begin{bmatrix} 1 + \frac{\hat{\lambda}}{2}\theta & -\hat{\lambda} \\ \frac{1}{2}(1 + \hat{\lambda}\theta) & 1 \end{bmatrix} \quad \text{so that} \quad A_d^{-1} B_d = \frac{1}{|A_d|} \begin{bmatrix} 1 - \frac{\hat{\lambda}}{2}(1 - \theta) & 0 \\ 1 + \frac{\hat{\lambda}}{2}\theta & 0 \end{bmatrix}.$$

The eigenvalues of $A_d^{-1} B_d$ are 0 and

$$\mu = \frac{1 - \frac{\hat{\lambda}}{2}(1 - \theta)}{1 + \frac{\hat{\lambda}}{2}(1 + \theta) + \frac{\hat{\lambda}^2}{2}\theta}. \tag{3.6}$$

In order to conclude that our new formulation is L-stable for the linear problem, it must be shown that the absolute value of the eigenvalues are bounded by unity and tend to zero as the step size goes to infinity. This need only be shown for the non-zero eigenvalue. To see that $|\mu| \leq 1$ when $\theta \geq \theta_{\min} = 0$, note that the lower bound $-1 \leq \mu$ is equivalent to

$$-1 - \frac{\hat{\lambda}}{2}(1 + \theta) - \frac{\hat{\lambda}^2}{2}\theta \leq 1 - \frac{\hat{\lambda}}{2}(1 - \theta) \iff 0 \leq 2 + \hat{\lambda}\theta + \frac{\hat{\lambda}^2}{2}\theta,$$

which holds true, and the upper bound $\mu \leq 1$ is equivalent to

$$1 - \frac{\hat{\lambda}}{2}(1 - \theta) \leq 1 + \frac{\hat{\lambda}}{2}(1 + \theta) + \frac{\hat{\lambda}^2}{2}\theta \iff 0 \leq \hat{\lambda} + \frac{\hat{\lambda}^2}{2}\theta,$$

which also holds. We conclude that $|\mu| \leq 1$ and the scheme is unconditionally stable to rounding error. Moreover, the scheme is L-stable because $\mu \to 0$ as $\hat{\lambda} \to \infty$, since μ is quadratic in $\hat{\lambda}$ in the denominator and linear in $\hat{\lambda}$ in the numerator. In fact, one can see that these two conditions and (3.5) hold if merely

$$\theta \geq \theta_{\min} \geq \frac{-4}{\hat{\lambda}(2 + \hat{\lambda})}. \tag{3.7}$$

This lower bound tends to zero as $\lambda \overset{\rightarrow}{\infty}$. Since we have assumed that $\epsilon = 0$ in (2.17), and preferring a uniform lower bound, we simply take $\theta_{\min} = 0$ in practice.

3.2 Proof of Nonlinear Stability of SATh$_0$-up

We now analyze the stability properties of the SATh$_0$-up scheme applied to the nonlinear problem. For a flux f, define the parameter ψ by

$$\psi = \begin{cases} \dfrac{f(\bar{u}^{n+1}) - f(\bar{u}^n)}{\alpha_{\mathrm{LF}}(\bar{u}^{n+1} - \bar{u}^n)}, & \text{if } |\bar{u}^{n+1} - \bar{u}^n| > 0, \\ \psi^*, & \text{otherwise,} \end{cases} \tag{3.8}$$

for $\psi^* \in [0, 1]$. Note that when $f'(u) \geq 0$ and $\alpha_{\mathrm{LF}} = \max_u |f'(u)|$, the bound $0 \leq \psi \leq 1$ holds. Observe that if the linear flux function $f(u) = \alpha_{\mathrm{LF}} u$ is substituted, then $\psi = 1$, regardless of the denominator when $\psi^* = 1$ is taken.

The hypotheses imply that $(1 - \theta_i)(\bar{u}_i^{n+1} - \bar{u}_i^n) = (\bar{u}_i^{n+1} - \tilde{\bar{u}}_i^{n+1})$, and so

$$(1 - \theta_i)(\bar{f}_i^{n+1} - \bar{f}_i^n) = (1 - \theta_i)\alpha_{\mathrm{LF}}\psi_i(\bar{u}_i^{n+1} - \bar{u}_i^n) = \alpha_{\mathrm{LF}}\psi_i(\bar{u}_i^{n+1} - \tilde{\bar{u}}_i^{n+1}),$$

even in the case $\psi_i = \psi_i^*$, i.e., $\bar{u}_i^{n+1} - \bar{u}_i^n = 0$, since then $\bar{u}_i^{n+1} = \bar{u}_i^n = \tilde{\bar{u}}_i^{n+1}$. Reformulating (2.20) and (3.1) in terms of ψ gives

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \lambda\big[\bar{f}_i^{n+1} - \alpha_{\mathrm{LF}}\left(\bar{u}_i^{n+1} - \tilde{\bar{u}}_i^{n+1}\right)\psi_i - \bar{f}_{i-1}^{n+1} + \alpha_{\mathrm{LF}}\left(\bar{u}_{i-1}^{n+1} - \tilde{\bar{u}}_{i-1}^{n+1}\right)\psi_{i-1}\big], \tag{3.9}$$

$$\tilde{\bar{u}}_i^{n+1} = \frac{1}{2}\left(\bar{u}_i^{n+1} + \bar{u}_i^n\right) + \frac{\lambda\alpha_{\mathrm{LF}}}{2}\big[\theta_i^{n+1}\left(\bar{u}_i^{n+1} - \tilde{\bar{u}}_i^{n+1}\right)\psi_i - \theta_{i-1}^{n+1}\left(\bar{u}_{i-1}^{n+1} - \tilde{\bar{u}}_{i-1}^{n+1}\right)\psi_{i-1}\big], \tag{3.10}$$

where $\lambda = \Delta t^{n+1}/\Delta x_i$.

In order to express the scheme in matrix form, rearrange the definition of ψ_i to get a third equation, namely,

$$\bar{f}_i^{n+1} - \alpha_{\mathrm{LF}}\bar{u}_i^{n+1}\psi_i = \bar{f}_i^n - \alpha_{\mathrm{LF}}\bar{u}_i^n\psi_i. \tag{3.11}$$

Our system $A\xi^{n+1} = B\xi^n$ now uses the variables

$$\xi^n = \left(\ldots,\ \bar{u}_{i-1}^n,\ \bar{f}_{i-1}^n,\ \tilde{\bar{u}}_{i-1}^n,\ \bar{u}_i^n,\ \bar{f}_i^n,\ \tilde{\bar{u}}_i^n, \ldots\right)^T. \tag{3.12}$$

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 the subscript d indicates the diagonal blocks.

It is straightforward to see that the i-th blocks are

$$A_d = \begin{bmatrix} 1 - \lambda\alpha_{\mathrm{LF}}\psi & \lambda & \lambda\alpha_{\mathrm{LF}}\psi \\ -\alpha_{\mathrm{LF}}\psi & 1 & 0 \\ -\frac{1}{2}(1 + \lambda\alpha_{\mathrm{LF}}\psi\theta) & 0 & 1 + \frac{\lambda}{2}\alpha_{\mathrm{LF}}\psi\theta \end{bmatrix} \quad \text{and} \quad B_d = \begin{bmatrix} 1 & 0 & 0 \\ -\alpha_{\mathrm{LF}}\psi & 1 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix},$$

where the subscript i has been repressed for readability. Note that the determinant of A_d is positive; that is, with $\hat{\lambda} = \lambda\alpha_{\mathrm{LF}} > 0$,

$$|A_d| = 1 + \frac{\hat{\lambda}}{2}\psi\theta + \frac{\hat{\lambda}}{2}\psi(1 + \lambda\hat{\psi}\theta) \geq 1, \tag{3.13}$$

since $\theta \geq \theta_{\min} \geq 0$ and $\psi \in [0, 1]$. Hence, A_d is invertible and

$$A_d^{-1} = \frac{1}{|A_d|}\begin{bmatrix} 1 + \frac{1}{2}\hat{\lambda}\psi\theta & -\lambda\big(1 + \frac{1}{2}\hat{\lambda}\psi\theta\big) & -\hat{\lambda}\psi \\ \alpha_{\mathrm{LF}}\psi\big(1 + \frac{1}{2}\hat{\lambda}\psi\theta\big) & |A_d|\mu_* & -\hat{\lambda}\alpha_{\mathrm{LF}}\psi^2 \\ \frac{1}{2}(1 + \hat{\lambda}\psi\theta) & -\frac{1}{2}\lambda(1 + \hat{\lambda}\psi\theta) & 1 \end{bmatrix},$$

where

$$\mu_* = \frac{(1-\hat{\lambda}\psi)\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big)+\frac{1}{2}\hat{\lambda}\psi(1+\hat{\lambda}\psi\theta)}{|A_d|}.$$

Now

$$A_d^{-1}B_d = \frac{1}{|A_d|}\begin{bmatrix} |A_d| & -\lambda\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big) & 0 \\ 0 & |A_d|\mu_* & 0 \\ |A_d| & -\frac{1}{2}\lambda(1+\hat{\lambda}\psi\theta) & 0 \end{bmatrix}.$$

The eigenvalues of $A_d^{-1}B_d$ are $\mu = 0$, 1, and μ_*. The nontrivial eigenvalue is well-defined, because $|A_d| > 0$, and can be expressed as

$$\mu_* = \frac{(1-\hat{\lambda}\psi)\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big)+\frac{1}{2}\hat{\lambda}\psi(1+\hat{\lambda}\psi\theta)}{|A_d|} = \frac{|A_d|-\hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big)}{|A_d|}. \tag{3.14}$$

First, we show that the absolute value of (3.14) is bounded by unity so that the scheme is stable up to rounding error. Begin with the evident upper bound,

$$\mu_* \le 1 \iff |A_d| - \hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big) \le |A_d| \iff -\hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big) \le 0,$$

which holds since $\hat{\lambda}$, ψ, and θ are all non-negative. Now, the lower bound,

$$-1 \le \mu_* \iff 0 \le 2|A_d| - \hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big),$$

is less obvious. To show that the inequality holds, expand the right-hand side to see

$$\begin{aligned} 2|A_d| - \hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big) &= 2\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big) + \hat{\lambda}\psi\big(1+\hat{\lambda}\psi\theta\big) - \hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big) \\ &= 2 + \hat{\lambda}\psi\theta + \frac{1}{2}(\hat{\lambda}\psi)^2\theta, \end{aligned}$$

which is clearly non-negative. It is concluded that the scheme is unconditionally stable for $\theta_{\min} = 0$.

Furthermore, substituting (3.13) into (3.14) and taking the limit gives

$$\begin{aligned} \lim_{\lambda\to\infty}\mu_* &= \lim_{\lambda\to\infty}\frac{|A_d|-\hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big)}{|A_d|} \\ &= 1 - \lim_{\lambda\to\infty}\frac{\hat{\lambda}\psi\big(1+\frac{1}{2}\hat{\lambda}\psi\theta\big)}{1+\frac{1}{2}\hat{\lambda}\psi\theta+\frac{1}{2}\hat{\lambda}\psi(1+\hat{\lambda}\psi\theta)} = 0, \end{aligned}$$

However, the unit eigenvalue prevents us from concluding that SATh$_0$-up is L-stable for the nonlinear problem.

4 Satisfaction of the Maximum Principle in a Monotone Setting

The SATh-up scheme was proven in [1, Theorem 5.2] to satisfy the maximum principle in a monotone flow setting when $\theta_{\min} = 1/2$. This restriction was taken in that paper because it was assumed to be the requirement for stability. A careful reading of the proof of [1, Theorem 5.2] shows that with only minor changes, the result continues to hold when $\theta_{\min} = 0$. We now state the result.

Theorem 2 *Suppose that f is strictly monotone increasing, and suppose that in the definition (2.17) of θ, $\epsilon = 0$ and $\theta_{\min} \geq 0$. Pose the SATh-up scheme on a finite interval with a boundary condition on the left, i.e., let $\bar{u}_0^n$ be given for all n. If the boundary and initial conditions of the flow satisfy the monotone decreasing property*

$$\bar{u}_0^n \leq \bar{u}_0^{n+1} \quad \forall n \geq 0 \quad \text{and} \quad \bar{u}_i^0 \leq \bar{u}_{i-1}^0 \quad \forall i \geq 1, \tag{4.1}$$

then the scheme satisfies the maximum principle,

$$\bar{u}_i^n \leq \bar{u}_i^{n+1} \leq \bar{u}_{i-1}^{n+1} \quad \forall n \geq 0, \ i \geq 1. \tag{4.2}$$

Moreover, in the monotone increasing case,

$$\bar{u}_0^n \geq \bar{u}_0^{n+1} \quad \forall n \geq 0 \quad \text{and} \quad \bar{u}_i^0 \geq \bar{u}_{i-1}^0 \quad \forall i \geq 1, \tag{4.3}$$

then

$$\bar{u}_i^n \geq \bar{u}_i^{n+1} \geq \bar{u}_{i-1}^{n+1} \quad \forall n \geq 0, \ i \geq 1. \tag{4.4}$$

Also obtained are the two corollaries derived from this result in [1]. Recall that the total variation is

$$\mathrm{TV}(\bar{u}^n) = \sum_{i=1}^{\infty} |\bar{u}_{i-1}^n - \bar{u}_i^n|. \tag{4.5}$$

We state the result [1, Corollary 5.4] as follows, and the proof holds verbatim.

Corollary 1 *Assume the hypotheses of Theorem 2 and that for some constant $M \geq 0$, $|u_0^n| \leq M$ and $|u_i^0| \leq M$ for all $n \geq 0$ and $i \geq 0$. Then $TV(\bar{u}^n) \leq 2M$, so the SATh-up is total variation bounded (TVB) for $\theta_{\min} \geq 0$. Moreover, if also $\bar{u}_0^{n+1} = \bar{u}_0^n$ for all $n \geq 0$ (i.e., oscillation is not introduced at the boundary), then the scheme is total variation decreasing (TVD), meaning*

$$TV(\bar{u}^{n+1}) \leq TV(\bar{u}^n). \tag{4.6}$$

The other corollary [1, Corollary 5.3] is seen to hold for $\theta_{\min} \geq 0$ after a careful reading of the proof. In fact, the result can be generalized a bit, as we now state.

Corollary 2 *Assume the hypotheses of Theorem 2 and that $\theta^* \in [\theta_{\min}, 1]$ in (2.17). 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 $\theta_i = \theta_i^{n+1} \in [\theta_{\min}, 1]$ $\forall n \geq 0, \ i \geq 1$. Moreover, $\tilde{\bar{u}}_i^{n+1}$ lies between $\bar{u}_i^n$ and $\bar{u}_i^{n+1}$ $\forall n \geq 0, \ \forall i \geq 1$.*

Proof A hypothesis has been removed for the final result regarding $\tilde{\bar{u}}_i^{n+1}$, so only this part of the corollary requires attention. The proof for a monotone increasing flow is similar to that for a monotone decreasing flow, so consider only for the latter case. Assume (4.1), so that also (4.2) and $\bar{f}_i^n \leq \bar{f}_i^{n+1} \leq \bar{f}_{i-1}^{n+1}$ hold (since f is monotone increasing). First consider the case that $\bar{u}_i^{n+1} \neq \bar{u}_i^n$. Since

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

infer the bound $\tilde{\bar{u}}_i^{n+1} \leq \bar{u}_i^{n+1}$. The other bound is proved by contradiction. Suppose that $\tilde{\bar{u}}_i^{n+1} < \bar{u}_i^n$. Then $\theta_i = \theta_{\min}$. From (2.21), conclude that

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

wherein one should recall that $\lambda = \Delta t^{n+1}/\Delta x_i$. Since $(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^{n}) \geq 0$, estimate

$$\begin{aligned}\tilde{\bar{u}}_i^{n+1} &\geq \bar{u}_i^n - \frac{\lambda}{2}\Big[(\bar{f}_i^n - \bar{f}_{i-1}^n) + \theta_{\min}^2(\bar{f}_i^{n+1} - \bar{f}_i^n) - \theta_{\min}^2(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\Big] \\ &= \bar{u}_i^n - \frac{\lambda}{2}\Big[(1-\theta_{\min}^2)(\bar{f}_i^n - \bar{f}_{i-1}^n) + \theta_{\min}^2(\bar{f}_i^{n+1} - \bar{f}_{i-1}^{n+1})\Big] \\ &\geq \tilde{\bar{u}}_i^n,\end{aligned}$$

since $(\bar{f}_i^n - \bar{f}_{i-1}^n) \leq 0$ and $(\bar{f}_i^{n+1} - \bar{f}_{i-1}^{n+1}) \leq 0$, and a contradiction has been drawn. This establishes that $\bar{u}_i^n \leq \tilde{\bar{u}}_i^{n+1} \leq \bar{u}_i^{n+1}$ when $\bar{u}_i^{n+1} \neq \bar{u}_i^n$.

In case $\bar{u}_i^{n+1} = \bar{u}_i^n$, also $\bar{f}_i^{n+1} = \bar{f}_i^n$, and (2.20)–(2.21) reduce to

$$\begin{aligned}\bar{u}_i^{n+1} &= \bar{u}_i^n - \lambda\big[(\bar{f}_i^n - \bar{f}_{i-1}^n) - \theta_{i-1}(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\big], \\ \tilde{\bar{u}}_i^{n+1} &= \bar{u}_i^n - \frac{\lambda}{2}\big[(\bar{f}_i^n - \bar{f}_{i-1}^n) - \theta_{i-1}^2(\bar{f}_{i-1}^{n+1} - \bar{f}_{i-1}^n)\big].\end{aligned}$$

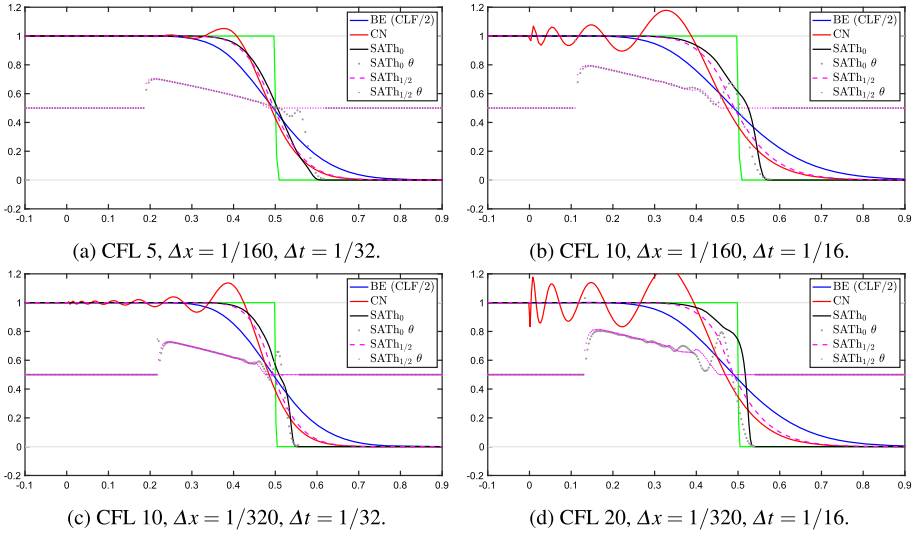
The expression in square brackets in the first of these two equations must vanish, and in fact each of its two terms must vanish, since it is a sum of two non-positive terms. Therefore the second equation reduces to $\tilde{\bar{u}}_i^{n+1} = \bar{u}_i^n = \bar{u}_i^{n+1}$. □

5 Numerical Investigation in 1D

This section is devoted to an investigation of the numerical performance of the general SATh scheme in one space dimension. Comparisons are made between four specific schemes, all using the same spatial discretization but differing in their time stepping. They are SATh_0 (with the stability bound $\theta_{\min} = 0$), $\text{SATh}_{1/2}$ (with the stability bound $\theta_{\min} = 1/2$), backward Euler (BE), and Crank-Nicolson (CN). Because SATh requires the solution of both $\bar{u}$ and $\tilde{\bar{u}}$, while BE only solves for $\bar{u}$, we take twice as many BE time steps as SATh. In other words, in this section BE always uses half the CFL number stated for SATh and CN. The computational mesh will be uniform of spacing $h = \Delta x = 1/m > 0$.

In this section, the schemes use constant values on the cell boundary interfaces (2.16). Either the upstream (2.19) or Lax-Friedrichs (2.18) numerical fluxes will be used. Although there is no theory for the Lax-Friedrichs stabilized scheme, we will see that it satisfies the results obtained by the upstream weighted scheme. It is mentioned in [1] that the values of ϵ and θ^* in (2.17) have little effect on the solution of the $\text{SATh}_{1/2}$ schemes. We take $\epsilon = 10^{-6}$ and $\theta^* = 1/2$, unless explicitly stated otherwise.

The nonlinear problem is solved for $w_i = \bar{u}_i^{n+1} - \bar{u}_i^n$ and $v_i = \tilde{\bar{u}}_i^{n+1} - \bar{u}_i^n$, since θ_i is defined by their ratio v_i/w_i. It can sometimes be difficult or impossible to solve the nonlinear equations defining the SATh scheme. In the case of a high CFL number, the Jacobian matrix may have an extreme condition number or become singular. In other cases, poor Newton convergence may be due partly to the difficulty of handling the case where $\bar{u}$ does not change, so that θ is poorly and nonsmoothly defined in (2.17). We do not use any special version of Newton's method to handle the nonsmooth nonlinearity. However, it can be useful to incorporate a damped Newton update to facilitate convergence. A damping factor around 0.75 or 0.85 seems to work well. In other words, each Newton iteration uses only 0.75 or 0.85 times the predicted Newton step update. In addition, as we remarked earlier, the factor multiplying ϵ in (2.17) improves the Newton convergence.



(a) CFL 5, $\Delta x = 1/160$, $\Delta t = 1/32$.

(b) CFL 10, $\Delta x = 1/160$, $\Delta t = 1/16$.

(c) CFL 10, $\Delta x = 1/320$, $\Delta t = 1/32$.

(d) CFL 20, $\Delta x = 1/320$, $\Delta t = 1/16$.

Fig. 1 Linear problem (5.1) with discontinuous initial condition (5.2) at $t = 0.5$. Shown are the solutions and θ for SATh_0-up and $\text{SATh}_{1/2}$-up, using $\epsilon = 10^{-6}$ and $\theta^* = 1/2$. Also shown are the CN and BE solutions (BE uses half the CFL) and the true solution (in green)

5.1 Linear Transport

First consider the linear problem; that is,

$$u_t + u_x = 0, \quad L_0 < x < L_1, \quad t > 0, \tag{5.1}$$

with unit speed $\alpha_{\text{LF}} = 1$. The Lax-Friedrichs and upstream weighted schemes are the same for the linear problem (up to rounding error).

5.1.1 A Contact Discontinuity in a Monotone Solution

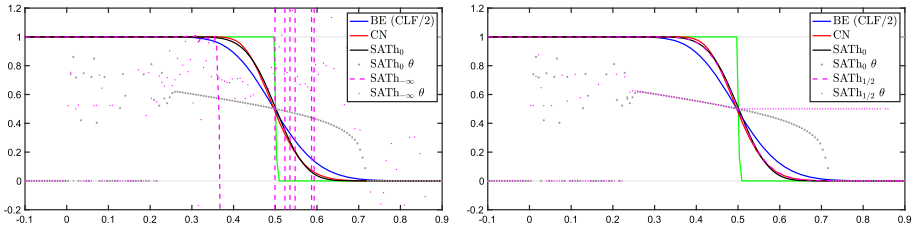
The first test is for a contact discontinuity. Let $L_0 = -0.1$, $L_1 = 0.9$, $u(L_0, t) = 1$, $u(L_1, t) = 0$, and let the initial condition be a step function, i.e.,

$$u_0(x) = \begin{cases} 1, & x < 0, \\ 0, & x > 0. \end{cases} \tag{5.2}$$

The true solution is $u(x, t) = u_0(x - t)$, and it remains monotone decreasing.

Figure 1a shows the true solution and computed solutions of BE, CN, SATh_0, and $\text{SATh}_{1/2}$ at time $t = 0.5$ using $\Delta x = 1/160$ and $\Delta t = 1/32$; that is, with a moderate CFL number of 5 (2.5 for BE using $\Delta t = 1/64$). All schemes conserve mass up to rounding error, so all these schemes have the discontinuity in the correct location. The two SATh solutions compare favorably to CN, although CN oscillates unacceptably. The BE solution shows excessive numerical diffusion.

Compared to $\text{SATh}_{1/2}$-up, the solution for SATh_0-up displays less numerical diffusion due to its ability to use θ values less than one half. These values are also shown in the figure, and the two schemes have similar θ values except near the front at $x = 0.5$. SATh_0-up generates values of θ below $1/2$ near the front, resulting in a sharpening of the solution compared to



(a) $\theta_{\min} = -\infty$, $\theta^* = 0$, CFL 2.5, $\Delta x = 1/160$. (b) $\theta^* = 0$, CFL 2.5, $\Delta x = 1/160$.

Fig. 2 Linear problem (5.1) with discontinuous initial condition (5.2) at $t = 0.5$, using $\Delta x = 1/160$ and CFL 2.5. Shown are the solutions and θ for SATh-up, using $\epsilon = 10^{-6}$ and $\theta^* = 0$, taking $\theta_{\min} = -\infty$, 0, or $1/2$. Also shown are the CN and BE solutions (BE uses half the CFL) and the true solution (in green)

$\text{SATh}_{1/2}$-up. To the right of the front the solution is constant, resulting in $\theta = \theta_*$ for both schemes. To the left of the front, the solution is essentially constant, and ideally $\theta = \theta_*$ would result. However, θ is computed using very small differences in the solution. In this example, values above $1/2$ result in a region from about $x = 0.2$ or $x = 0.1$ to the front. If a somewhat larger value is taken for ϵ (rather than 1e-6), this region decreases.

Before continuing, note that the profile of the BE solution in Figure 1 is symmetric on either side of the true contact discontinuity at $x = 0.5$, since the scheme is stabilized with linear numerical diffusion. The $\text{SATh}_{1/2}$-up scheme has a similar symmetry, even though it is an inherently nonlinear scheme. On the other hand, the SATh_0-up solution is asymmetric about $x = 0.5$. With $\Delta x = 1/160$ but an increased CFL from 5 to 10, the effect is noticeably exacerbated, as shown in Figures 1a –1b . However, increasing resolution to $\Delta x = 1/320$ while keeping CFL 10 reduces the effect nicely, as shown in Figure 1c . Figure 1d shows the effect for the finer resolution but CFL 20.

The observant reader may notice that in Figure 1d , the SATh_0-up scheme generates a value of $\theta > 1$. This is not a violation of Corollary 2 for two reasons. First, the corollary does not account for rounding error. Second, $\epsilon \neq 0$ in (2.17), since taking $\epsilon = 0$ generally causes numerical problems associated to defining θ when its denominator is nearly zero. Nonetheless, the SATh-up schemes provide $\bar{u}$ and $\tilde{\bar{u}}$ values that satisfy the maximum principle. Moreover, the total variation remains 1 for all the schemes except CN.

The new stability bound is essentially sharp. This can be seen in Figure 2a , where we attempt to solve the unconstrained problem using $\theta_{\min} = -\infty$. The $\text{SATh}_{-\infty}$-up solution using $\theta^* = 1/2$ shows no difficulties, since the solution is computed without trying to make $\theta < 0$. However, if we change $\theta^* = 0$, we see the results in the figure. The $\text{SATh}_{-\infty}$-up solution appears to go unstable ahead of the discontinuity, where some of the values of θ are negative.

In Figure 2b , we see that both SATh_0-up and $\text{SATh}_{1/2}$-up have values of θ that oscillate behind the discontinuity. This is due to the facts that $\theta^* = 0$ and the solution is constant in that region, so θ is poorly defined because its denominator is nearly zero. Taking $\theta^* = 1/2$ avoids this numerical difficulty.

5.1.2 A Nonmonotone Sine Wave

Now consider (5.1) with $L_0 = 0$, $L_1 = 1$, using the smooth initial condition

$$u_0(x) = \frac{1 + \sin(2\pi x)}{2}, \quad 0 < x < 1, \tag{5.3}$$

Table 1 Nonmonotone sine wave linear transport error and convergence order at $t = 0.5$ using CFL 4, $m = 1/\Delta x$ cells, and $\Delta t = 4h$ (except BE uses $\Delta t = 2h$)

	BE		CN		$SATh_0$-up		$SATh_{1/2}$-up	
m	L_h^1-error	order	L_h^1-error	order	L_h^1-error	order	L_h^1-error	order
80	9.77e-02	0.74	3.71e-02	0.93	3.75e-02	0.97	3.87e-02	1.05
160	5.36e-02	0.87	1.91e-02	0.96	1.91e-02	0.97	1.93e-02	1.00
320	2.81e-02	0.93	9.67e-03	0.98	9.68e-03	0.98	9.72e-03	0.99
640	1.44e-02	0.97	4.87e-03	0.99	4.87e-03	0.99	4.88e-03	0.99
m	L_h^∞-error	order	L_h^∞-error	order	L_h^∞-error	order	L_h^∞-error	order
80	1.54e-01	0.74	5.82e-02	0.93	6.10e-02	1.10	7.89e-02	0.97
160	8.43e-02	0.87	2.99e-02	0.96	3.06e-02	0.99	3.86e-02	1.03
320	4.42e-02	0.93	1.52e-02	0.98	1.54e-02	0.99	1.86e-02	1.06
640	2.26e-02	0.97	7.65e-03	0.99	7.73e-03	1.00	8.92e-03	1.06

Table 2 Nonmonotone sine wave linear transport error and convergence order at $t = 0.5$ using CFL 10, $m = 1/\Delta x$ cells, and $\Delta t = 10h$ (except BE uses $\Delta t = 5h$)

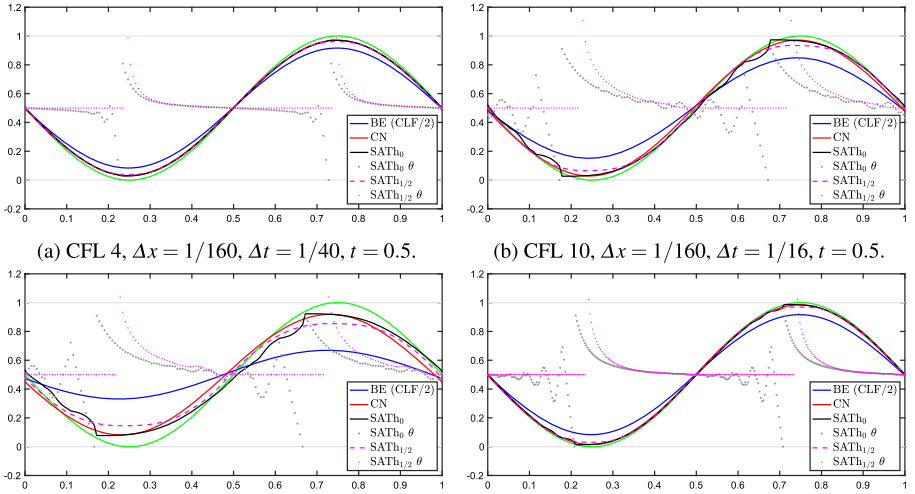
	BE		CN		$SATh_0$-up		$SATh_{1/2}$-up	
m	L_h^1-error	order	L_h^1-error	order	L_h^1-error	order	L_h^1-error	order
80	1.63e-01	0.55	5.56e-02	1.41	5.51e-02	1.25	6.53e-02	1.27
160	9.76e-02	0.74	2.22e-02	1.33	2.18e-02	1.34	2.36e-02	1.47
320	5.36e-02	0.86	1.01e-02	1.13	9.85e-03	1.15	1.03e-02	1.20
640	2.81e-02	0.93	4.93e-03	1.04	4.89e-03	1.01	4.96e-03	1.05
m	L_h^∞-error	order	L_h^∞-error	order	L_h^∞-error	order	L_h^∞-error	order
80	2.56e-01	0.55	8.73e-02	1.41	1.49e-01	1.67	1.37e-01	0.90
160	1.53e-01	0.74	3.48e-02	1.33	5.51e-02	1.44	6.63e-02	1.05
320	8.42e-02	0.86	1.59e-02	1.13	2.25e-02	1.29	3.05e-02	1.12
640	4.42e-02	0.93	7.74e-03	1.04	9.49e-03	1.25	1.37e-02	1.15

and periodic boundary conditions. The true solution $u(x,t) = u_0(x-t)$ is obviously not monotone.

All schemes (BE, CN, $SATh_0$-up, and $SATh_{1/2}$-up) are stabilized with upstream weighting and use equivalent space discretization, which limits the convergence order to one. Convergence results are shown in Tables 1–2, and all four schemes give first order accuracy for a moderate CFL (4, except 2 for BE) and a high CFL (10, except 5 for BE). BE exhibits the most error, while the two SATh-up schemes are fairly comparable to CN. Moreover, the errors for $SATh_0$-up are a bit smaller than the errors for $SATh_{1/2}$-up.

Some of the solutions are shown in Figure 3. The results for moderate CFL 4 in Figure 3a are perhaps what would be expected from the convergence data. This is also true of the results for BE, CN, and $SATh_{1/2}$-up at the high CFL 10 in Figure 3b .

The solution of $SATh_0$-up in Figure 3b exhibits an unfortunate feature in this nonmonotone setting. A clear oscillation appears in θ behind the extrema in the solution at around $x = 0.2$ and 0.7, resulting in a jagged variation in the solution. The Newton procedure of the numerical



(a) CFL 4, $\Delta x = 1/160$, $\Delta t = 1/40$, $t = 0.5$.

(b) CFL 10, $\Delta x = 1/160$, $\Delta t = 1/16$, $t = 0.5$.

(c) CFL 10, $\Delta x = 1/160$, $\Delta t = 1/16$, $t = 1.5$.

(d) CFL 10, $\Delta x = 1/320$, $\Delta t = 1/32$, $t = 0.5$.

Fig. 3 Linear problem (5.1) with initial condition (5.2) and periodic boundary conditions. Shown are the solutions and θ for SATh_0-up and $\text{SATh}_{1/2}$-up, using $\epsilon = 10^{-6}$ and $\theta* = 1/2$. Also shown are the CN and BE solutions (BE uses half the CFL) and the true solution (in green)

solution appears to be converged, and a smaller ϵ does not help. Figure 3c shows the solutions at time $t = 1.5$, and the profiles of the schemes' solutions are nearly identical to that at $t = 0.5$, except for some reduction in the total variation. The SATh_0-up solution does not appear to be unstable and its oscillation does not grow with time. Refinement of the mesh while maintaining CFL 10 yields a better profile for all schemes, as shown in Figure 3d , and especially for SATh_0-up. This is consistent with the convergence rate already observed in Table 2.

5.2 Nonlinear Transport: Burgers Equation

Next consider Burgers' equation with flux function $f(u) = u^2/2$, that is, the equation

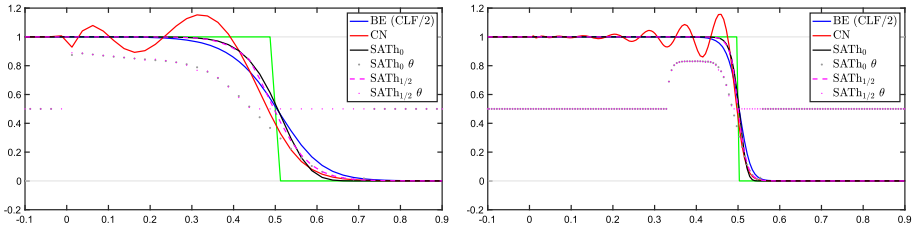
$$u_t + uu_x = 0, \quad L_0 < x < L_1, \quad t > 0. \tag{5.4}$$

Going forward in this section, the more versatile Lax-Friedrichs numerical flux (2.18) is used. All test problems in this subsection have true solutions between -1 and 1, so $\alpha_{\text{LF}} = 1$.

5.2.1 A Riemann Shock

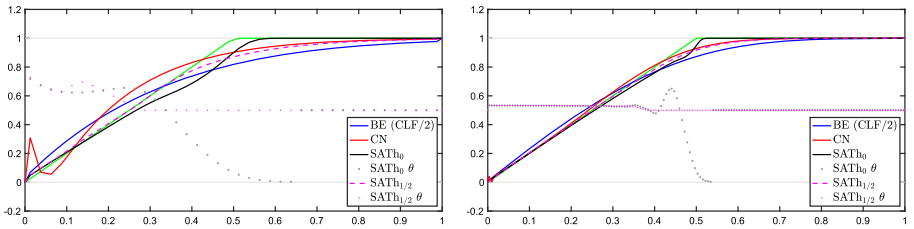
The first test is a Riemann shock with $L_0 = -0.1$, $L_1 = 0.9$, $u(L_0, t) = 1$, $u(L_1, t) = 0$, and the initial condition (5.2). The solution is monotone decreasing.

A comparison of the schemes using Lax-Friedrichs numerical fluxes in given in Figure 4 using CFL 10 and both $\Delta x = 1/40$ and $\Delta x = 1/160$. One immediately notices that the spreading of the profiles of the solutions are all relatively symmetric about the shock. This was not the case in Section 5.1.1, where especially the solution of the SATh_0-up scheme applied to the linear problem exhibited an significantly asymmetric profile. The difference is likely due to the fact that shocks are self-sharpening.



(a) CFL 10, $\Delta x = 1/40$, $\Delta t = 1/4$. (b) CFL 10, $\Delta x = 1/160$, $\Delta t = 1/16$.

Fig. 4 Burger's equation with Riemann initial condition (5.2) at time $t = 1$ using CFL 10 (BE uses half the time-step). The true solution is shown in green



(a) CFL 10, $\Delta x = 1/40$, $\Delta t = 1/4$. (b) CFL 10, $\Delta x = 1/160$, $\Delta t = 1/16$.

Fig. 5 Burger's Riemann rarefaction wave at time $t = 0.5$ using CFL 10 (BE uses half the time-step). The true solution is shown in green

The SATh_0-LF and $\text{SATh}_{1/2}$-LF schemes have significantly less numerical diffusion compared to BE. They predict the shock about as well as or better than CN, which oscillates unacceptably. Behind the shock, SATh_0-LF and $\text{SATh}_{1/2}$-LF have similar profiles. Ahead of the shock, SATh_0-LF is sharper than $\text{SATh}_{1/2}$-LF and also CN. This is due to the fact that SATh_0-LF is able to and does use $\theta < 1/2$ in this problem. At both resolutions, the SATh_0-LF, $\text{SATh}_{1/2}$-LF, and BE schemes remain stable and monotone. No oscillation is observed in the solution and the total variation remains 1.

5.2.2 A Riemann Rarefaction

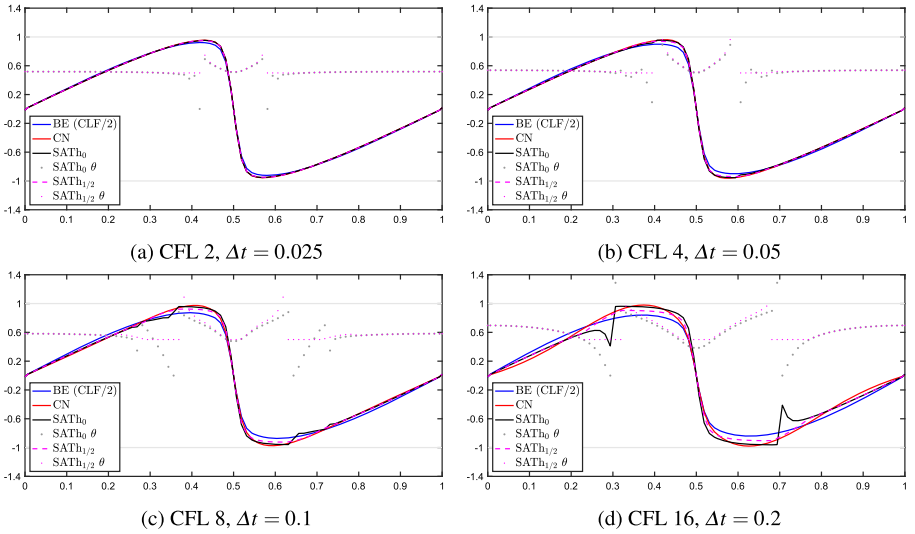
Consider Burger's equation (5.4) with a Riemann rarefaction, for which we take $L_0 = 0$, $L_1 = 1$, $u(L_0, t) = 0$, $u(L_1, t) = 1$, and $u(x, 0) = 1$. For this test, the solution is monotone increasing.

The results at time $t = 0.5$ and CFL 10 are shown in Figure 5 using $\Delta x = 1/40$ and $\Delta x = 1/160$. Note that only 2 time steps are taken in the former case, and 8 in the latter, to get to time 0.5. As in previous examples, CN may oscillate unacceptably and BE is the most diffused. The SATh_0-LF and $\text{SATh}_{1/2}$-LF solutions appear fairly sharp while remaining stable and monotone. In this example, SATh_0-LF outperforms all the schemes, including CN, in maintaining the rarefaction profile.

5.2.3 Shock Formation in a Nonmonotone Sine Wave

Finally, we simulate shock formation by taking the initial condition

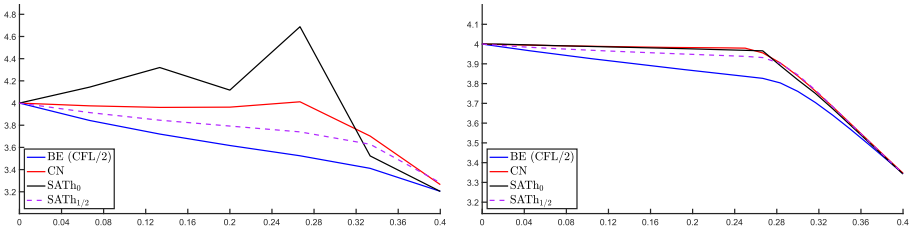
$$u_0(x) = \sin(2\pi x), \quad 0 < x < 1, \tag{5.5}$$



(a) CFL 2, $\Delta t = 0.025$ (b) CFL 4, $\Delta t = 0.05$

(c) CFL 8, $\Delta t = 0.1$ (d) CFL 16, $\Delta t = 0.2$

Fig. 6 Burger's equation with smooth initial condition $\sin(2\pi x)$ at time $t = 0.2$ using $\Delta x = 1/80$ and various Δt (BE uses half the time-step)



(a) CFL 10, $\Delta x = 1/150$, $\Delta t = 0.4/6$ (b) CFL 10, $\Delta x = 1/600$, $\Delta t = 0.4/24$

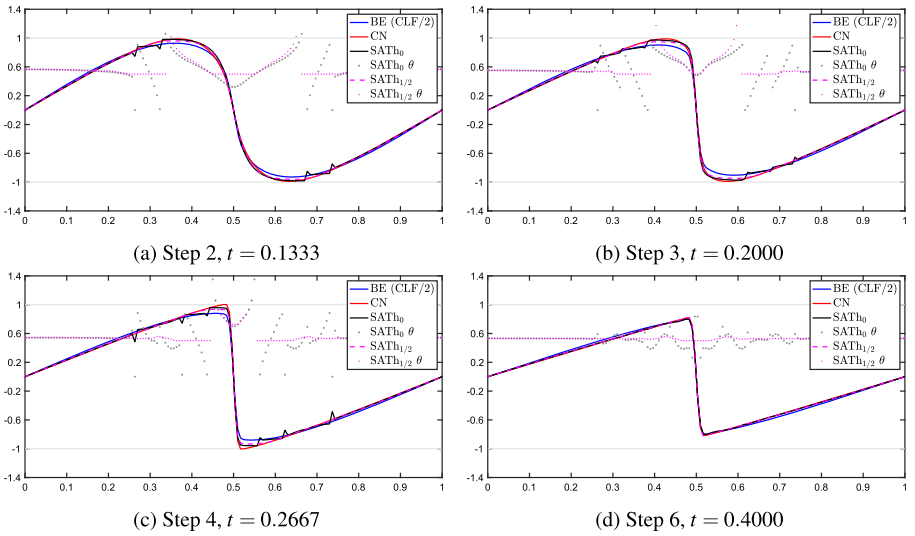
Fig. 7 Burger's equation with smooth initial condition $\sin(2\pi x)$. Total variation to time $t = 0.4$

and imposing periodic boundary conditions. This problem *requires* the Lax-Friedrichs numerical flux, since mass moves to the right for $x < 0.5$ and to the left for $x > 0.5$. The shock forms at time $t = 1/(2\pi) \approx 0.1592$.

Consider first a fixed mesh with $\Delta x = 1/80$ and various Δt. Results at time $t = 0.2$ after the shock has formed appear in Figure 6. Overall BE is the most diffusive. The two SATh-LF schemes give a sharp shock profile comparable to CN. SATh_0-LF gives the sharpest shock.

In this nonmonotone setting, except for CFL 2, SATh_0 exhibits a jagged profile similar to what was observed for the linear problem in Section 5.1.2. The jaggedness worsens as the CFL number increases. The solution for CFL 2 is clean, CFL 4 has slight wiggles, CFL 8 has noticeable wiggles, and CFL 16 has a pronounced kink. Clearly SATh_0-LF is neither TVD nor MPP in this case when the CFL number is too large.

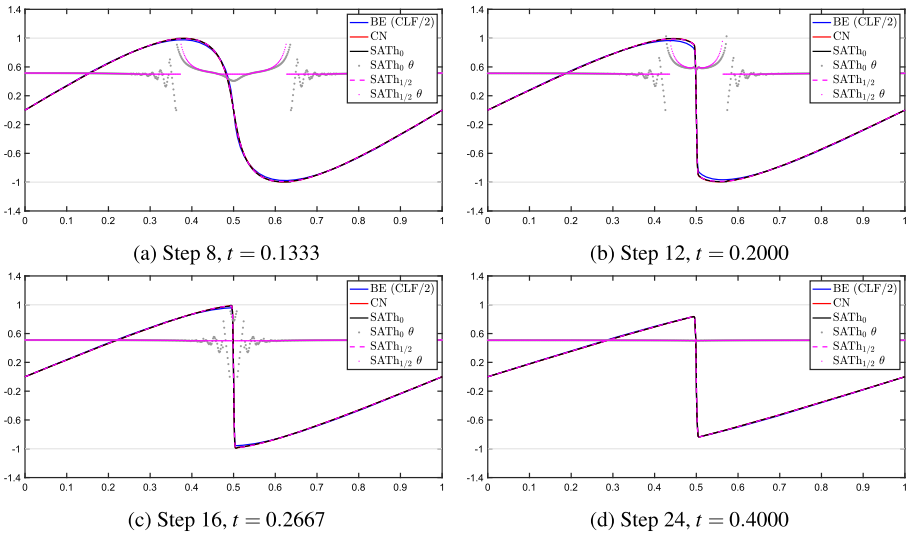
Next consider time evolution of the solution to the final time $t = 0.4$ using CFL 10 and two resolutions, $\Delta x = 1/150$ (so $\Delta t = 1/15$) and $\Delta x = 1/600$ (so $\Delta t = 1/60$). The total variations of the computed solutions are shown in Figure 7. We see that BE and $\text{SATh}_{1/2}$-LF are TVD at both resolutions, but CN and SATh_0-LF are only TVD for the finer resolution.



(a) Step 2, $t = 0.1333$ (b) Step 3, $t = 0.2000$

(c) Step 4, $t = 0.2667$ (d) Step 6, $t = 0.4000$

Fig. 8 Burger's equation with smooth initial condition $\sin(2\pi x)$ using CFL 10, $\Delta x = 1/150$, and $\Delta t = 0.0667 = 0.4/6$ (BE uses half the time-step)

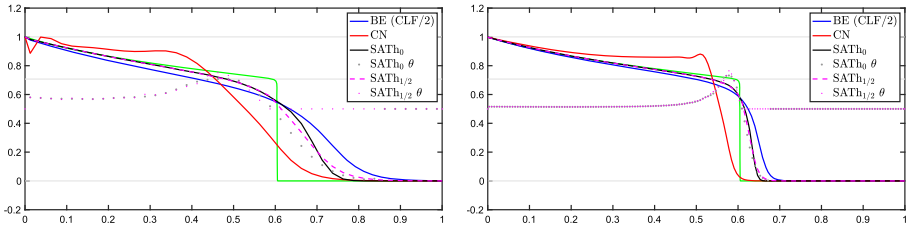


(a) Step 8, $t = 0.1333$ (b) Step 12, $t = 0.2000$

(c) Step 16, $t = 0.2667$ (d) Step 24, $t = 0.4000$

Fig. 9 Burger's equation with smooth initial condition $\sin(2\pi x)$ using CFL 10, $\Delta x = 1/600$, and $\Delta t = 0.0167 = 0.4/24$ (BE uses half the time-step)

In Figure 8, one can see the shock develop at the lower resolution. The jagged profile of the SATh_0-LF solution is apparent, but it dies out by time $t = 0.4$. As in the case of the linear problem, a refined mesh removes the jagged profile. The results of the finer resolution are given in Figure 9.



(a) CFL 10, $\Delta x = 1/40$, $\Delta t = 1/4$. (b) CFL 10, $\Delta x = 1/160$, $\Delta t = 1/16$.

Fig. 10 Buckley-Leverett equation with Riemann initial condition (5.2) at time $t = 0.5$ using CFL 10 (CFL 5 for BE). The true solution is shown in green

5.3 Nonlinear Transport: Buckley-Leverett Equation

Finally, consider the Buckley-Leverett equation with flux function

$$f(u) = \frac{u^2}{u^2 + (1-u)^2} \tag{5.6}$$

on the interval $0 < x < 1$, for which $\alpha_{\mathrm{LF}} = 2$ when $u \in [0, 1]$. Impose the initial condition (5.2). The solution maintains a monotone decreasing profile with a shock and a trailing rarefaction wave. The transition between them appears at the value of u where $f'(u) = f(u)/u$, which is about 0.707.

The results at time $t = 0.5$ appear in Figure 10 for CFL 10 using both $\Delta x = 1/40$ and $\Delta x = 1/160$. The true solution (plotted in green) is computed using simple forward Euler time-stepping and $\Delta x = \Delta t = 1/4000$. Interestingly, CN provides the worst and most diffuse profile, and it oscillates behind the shock. BE is more diffusive than the two SATh-LF schemes, which give remarkably good solutions. Of these two, SATh_0-LF provides the sharper solution.

6 Numerical Investigation in 2D

In any number of space dimensions, the algorithms (SATh_0, $\mathrm{SATh}_{1/2}$, backward Euler (BE), and Crank-Nicolson (CN)) can use arbitrary meshes. Two types of meshes are used here. First are logically rectangular meshes of quadrilaterals generated from uniform meshes by randomly perturbing the vertices by a factor of 0.25 times the unperturbed mesh spacing. Second are general meshes of polygons in two space dimensions generated by (a slightly modified version of) the software package *PolyMesher* [10] using only three smoothing iterations.

In these tests, all the schemes use constant values on the cell boundary interfaces (2.12) and the Lax-Friedrichs numerical flux. This flux is defined on facet $e = E^+ \cap E^-$ by (2.8) and

$$\hat{F}_e^{\pm}(u_e^{\mp}) = \mathbf{f}(\bar{u}_{E^{\mp}}) \cdot \nu_e \pm \alpha_{\mathrm{LF}}\, \bar{u}_{E^{\mp}}, \tag{6.1}$$

where the maximum wave speed $\alpha_{\mathrm{LF}} = 1$ in our tests, except the last one which uses $\alpha_{\mathrm{LF}} = 2.5$. The parameters are set as $\epsilon = 10^{-6}$ and $\theta^* = 1/2$.

The time step Δt is taken as a fixed constant for these tests. Rarely, the Newton procedure does not converge to a relative tolerance of 1e-5 (1e-4 for the tests of Section 6.2.1). In those cases, the time step is completed by using two sub-steps with half the original value of Δt.

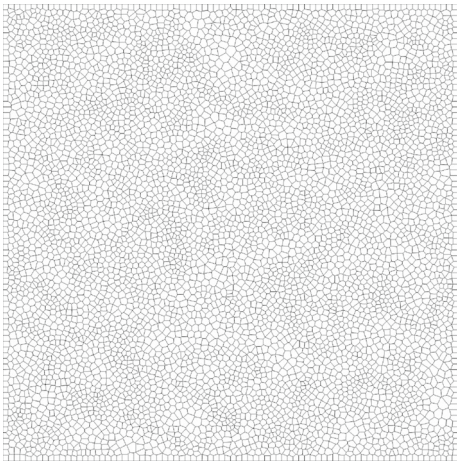


Fig. 11 Periodic polygonal mesh of 6400 elements on $[0, 1]^2$

Table 3 Linear transport in 2D with a squared sine wave initial condition. Errors are computed from the exact solution (6.3) at time $t = 1.0$ (one period) using $\Delta t = 0.025$ (CFL 3.982). Errors are measured in the discrete L_h^1 and L_h^∞ norms

	BE	CN	$\text{SATh}_{1/2}$	SATh_0
L_h^1	1.32e-01	6.21e-02	6.38e-02	6.23e-02
L_h^∞	4.31e-01	2.21e-01	2.37e-01	2.21e-01

Again, the equations are solved for the differences $w_E = \bar{u}_E^{n+1} - \bar{u}_E^n$ and $v_E = \tilde{\bar{u}}_E^{n+1} - \bar{u}_E^n$. In some cases, using a damped Newton update with damping factor around 0.75 after the first several unmodified iterations can help the SATh schemes to converge. We also impose a maximum allowable derivative of θ_E with respect to v_E and w_E set at 1e+6.

6.1 Linear Transport in 2D: Transport of a Squared Sine Wave

Consider first the linear transport equation

$$u_t + u_x + u_y = 0, \quad 0 < x < 1,\ 0 < y < 1,\ t > 0, \tag{6.2}$$

on a periodic domain with an initial condition given by a squared sine wave. The exact solution to this problem is

$$u(x, y, t) = \sin^2(\pi(x - t))\ \sin^2(\pi(y - t)). \tag{6.3}$$

We use a periodic mesh of 12,170 vertices and 6400 polygonal elements having up to 9 sides. (It has as many elements as a mesh of 80×80 squares.) The mesh is depicted in Figure 11.

Solutions are computed up to time $t = 1.0$, which gives a transport of exactly one period. The time step $\Delta t = 0.025$ gives the CFL number 3.982 on the mesh. The initial condition and solutions to BE, CN, $\text{SATh}_{1/2}$, and SATh_0 are displayed in Figure 13. We see excessive and asymmetric smearing for BE. The other three low order methods give generally good and very similar results. The computed errors measured in the discrete L_h^1 and L_h^∞ norms are given in Table 3, and these confirm our observations.

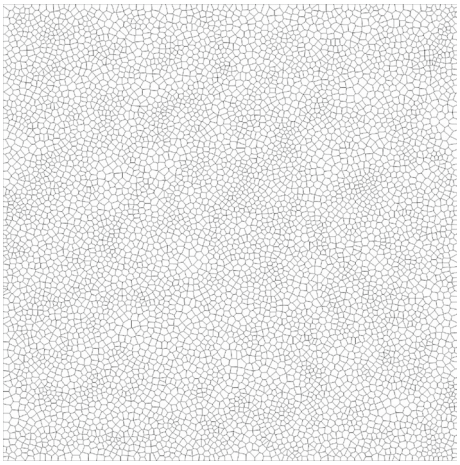
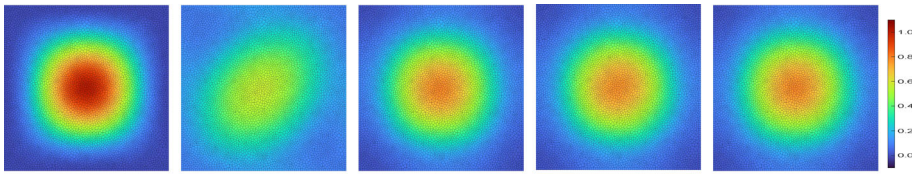


Fig. 12 Nonperiodic polygonal mesh of 6400 elements on $[0, 1]^2$



(a) $t = 0$ (b) BE (c) CN (d) $\text{SATh}_{1/2}$ (e) SATh_0

Fig. 13 Linear transport in 2D with a squared sine wave initial condition (see (6.3)) at time $t = 1.0$ (one period) using $\Delta t = 0.025$, which is CFL 3.982

6.2 Nonlinear Transport in 2D: Burgers Equation

Consider next the nonlinear Burgers equation

$$u_t + uu_x + uu_y = 0, \quad L_0 < x < L_1, \ L_2 < y < L_3, \ t > 0, \tag{6.4}$$

with some initial and boundary conditions yet to be imposed.

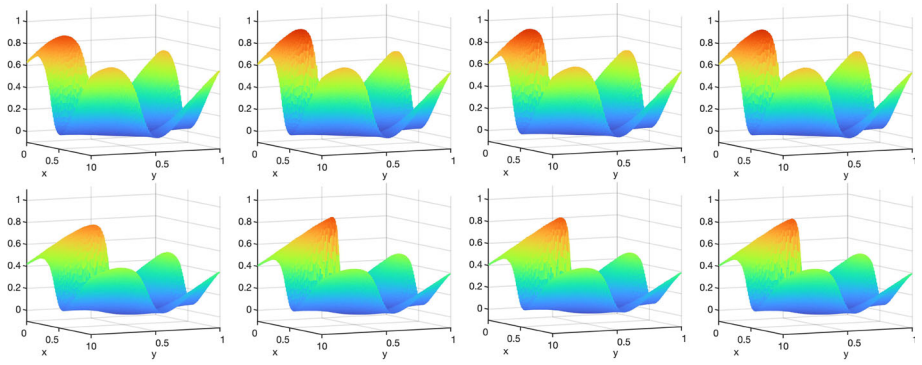
6.2.1 Burgers Transport of a Squared Cosine Wave in 2D

For the first test of Burgers equation, use the domain $(0, 1)^2$ and impose periodic boundary conditions. Take a squared cosine initial condition

$$u(x, y, 0) = \cos^2(\pi x) \cos^2(\pi y). \tag{6.5}$$

We use the periodic 6400 polygonal element mesh depicted in Figure 11 and used in the previous section for the linear equation. Recall that this mesh is akin to an 80×80 mesh. The time step is $\Delta t = 0.025$ (CFL 3.982). The simulation is run to a time after which a shock has developed in the solution.

The solutions $z = u(x, y, t)$ for the four schemes are plotted in Figure 14. We see that each has a considerably sharpened shape by time $t = 0.25$, and a sharp shock has developed by time $t = 0.5$. BE shows more numerical diffusion than the other three schemes, which show results similar to each other.



(a) BE (b) CN (c) $\text{SATh}_{1/2}$ (d) SATh_0

Fig. 14 Burgers transport in 2D with a squared cosine initial condition (6.5) at times $t = 0.25$ (top) and $t = 0.5$ (bottom) using $\Delta t = 0.025$, which is CFL 3.982

6.2.2 A 1D Riemann Shock and Rarefaction in 2D

Let the domain be $(0, 3) \times (0, 0.1)$ and the initial contition be

$$u(x, y, 0) = \begin{cases} 1, & 0.5 < x < 1.5, \\ 0, & \text{elsewhere.} \end{cases} \tag{6.6}$$

A trailing rarefaction wave and a leading shock in the x-direction results, with the exact solution

$$u(x, y, t) = \begin{cases} 0, & x < 0.5, x \geq 0.5t + 1.5, \\ (x - 0.5)/t, & 0.5 \leq x < t + 0.5, \\ 1, & t + 0.5 \leq x < 0.5t + 1.5, \end{cases} \tag{6.7}$$

up to time $t = 2$ when the rarefaction reaches the shock.

We solve this problem on meshes of quadrilaterals given by distorting the mesh points of a uniform mesh randomly up to 0.25 times the undistorted mesh spacing. Numerical results for this problem on the 240×8 mesh appear in Figure 15 at time $t = 1.0$ using $\Delta t = 0.05$ (the CFL number is 6.913). The results for BE are diffused more than the others, and CN results exhibit overshoots near the leading shock. The two SATh results are clean.

Errors as measured in the discrete L^1 norm and the orders of convergence appear in Table 4 for the four schemes in terms of the unperturbed mesh spacing. The distortion of the mesh causes the CFL number to increase with h. Nevertheless, the convergence order is seen to be one for this problem. BE has the most error and CN the least, but the two SATh schemes have errors close to CN.

6.2.3 A Flooding Scenario in 2D

Burgers equation is posed on the domain $(0, 1)^2$ and the initial condition is taken to be $u(x, y, 0) = 0$. The boundary condition is given on the inflow sides and taken to be $u(0, y, t) = u(x, 0, t) = 1$. There is no boundary condition along the outflow sides $(1, y, t)$ and $(x, 1, t)$. In this scenario, fluid has an initial discontinuity at the inflow sides of the boundary which evolves into a flood advancing diagonally into the domain.

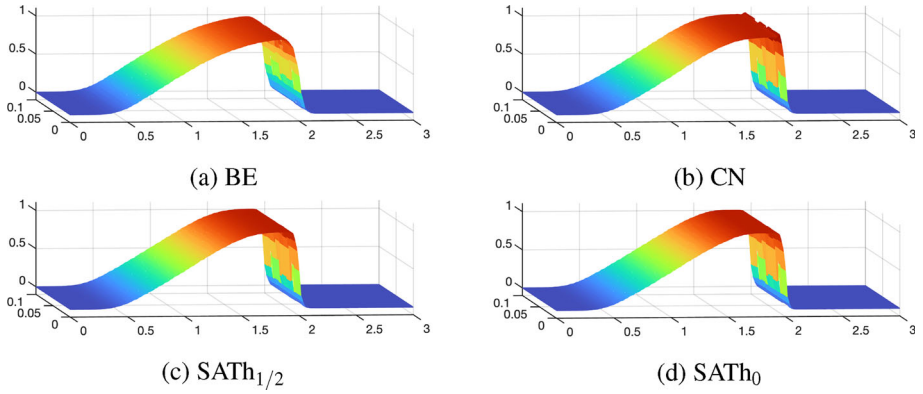


Fig. 15 Burgers Riemann shock and rarefaction in 1D at time $t = 1.0$ using $\Delta t = 0.05$, which is CFL 6.913

Table 4 Burgers Riemann shock and rarefaction in 1D at time $t = 1.0$. Errors in the L_h^1 norm for various distorted quadrilateral meshes with unperturbed spacing h and $\Delta t = 4h$

		BE		CN		SATh$_{1/2}$		SATh$_0$	
$1/h$	CFL	error	order	error	order	error	order	error	order
40	6.577	5.97e-02	—	3.33e-02	—	3.78e-02	—	3.82e-02	—
80	6.913	3.84e-02	0.64	2.04e-02	0.71	2.27e-02	0.74	2.24e-02	0.77
160	7.241	2.20e-02	0.80	1.09e-02	0.90	1.22e-02	0.90	1.20e-02	0.90
320	7.530	1.31e-02	0.75	5.78e-03	0.92	6.15e-03	0.99	6.12e-03	0.97

The problem is solved on the nonperiodic polygonal mesh of 6400 elements with up to 9 sides depicted in Figure 12. We use a time step of $\Delta t = 0.025$ (CFL 4.281). The solution is shown in Figure 16, at times $t = 0.375$ and $t = 1.0$. We see that the BE result is more diffused than that from the other three schemes, that the CN result exhibits overshoot along the line $x = y$. The two SATh schemes produce results comparable to CN, but with no overshoot. To see this more clearly, the solutions are shown in profile in Figure 17.

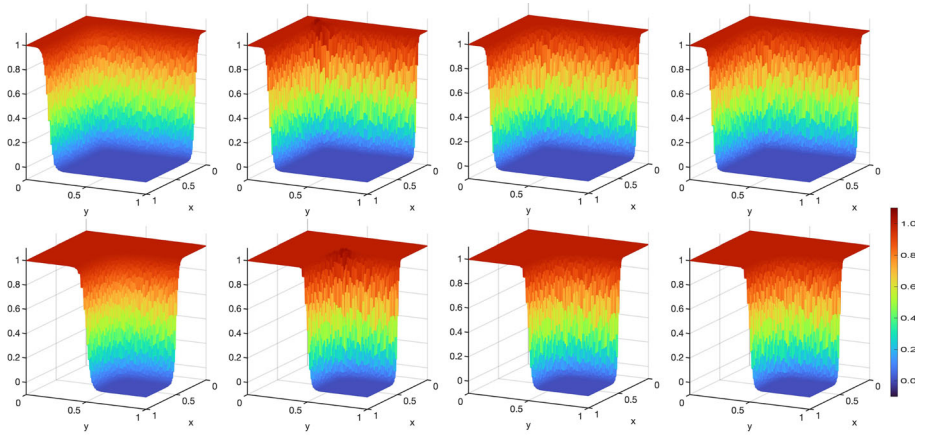
6.3 Nonlinear Transport in 2D: Buckley-Leverett Equation with Gravity

Finally, consider the Buckley-Leverett equation with vector flux function

$$f(u) = \frac{u^2}{u^2 + (1-u)^2} \begin{pmatrix} 1 \\ 1 - 5(1-u)^2 \end{pmatrix} \tag{6.8}$$

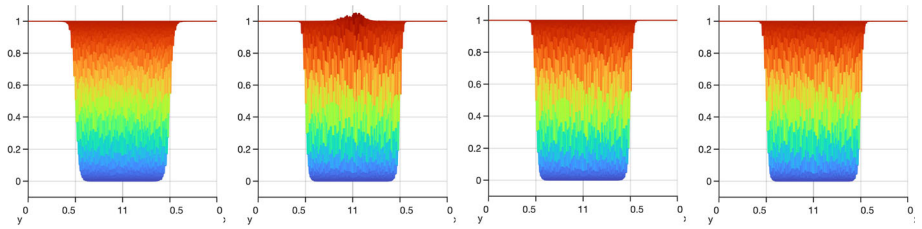
on the domain $(-1.5, 1.5)^2$. The usual Buckley-Leverett flux has been modified in the y-direction to model the effects of gravity. Because this flux function is nonconvex, the problem is quite challenging and there is less theory regarding the behavior of the numerical schemes.

For the numerical results, we take $\alpha_{\mathrm{LF}} = 2.5$. The solution should remain between 0 and 1, which makes the maximum wave speed a bit over 3.3. However, we do not use a local Lax-Friedrichs flux, so the full maximum wave speed results in excessively smoothed solutions. The value $\alpha_{\mathrm{LF}} = 2$ leads to some oscillation in the solution, but the value we take $\alpha_{\mathrm{LF}} = 2.5$ gives nonoscillatory solutions.



(a) BE (b) CN (c) $\mathrm{SATh}_{1/2}$ (d) SATh_0

Fig. 16 Burgers flooding scenario in 2D at times $t = 0.375$ (top) and $t = 1.0$ (bottom) using $\Delta t = 0.025$, which is CFL 4.281



(a) BE (b) CN (c) $\mathrm{SATh}_{1/2}$ (d) SATh_0

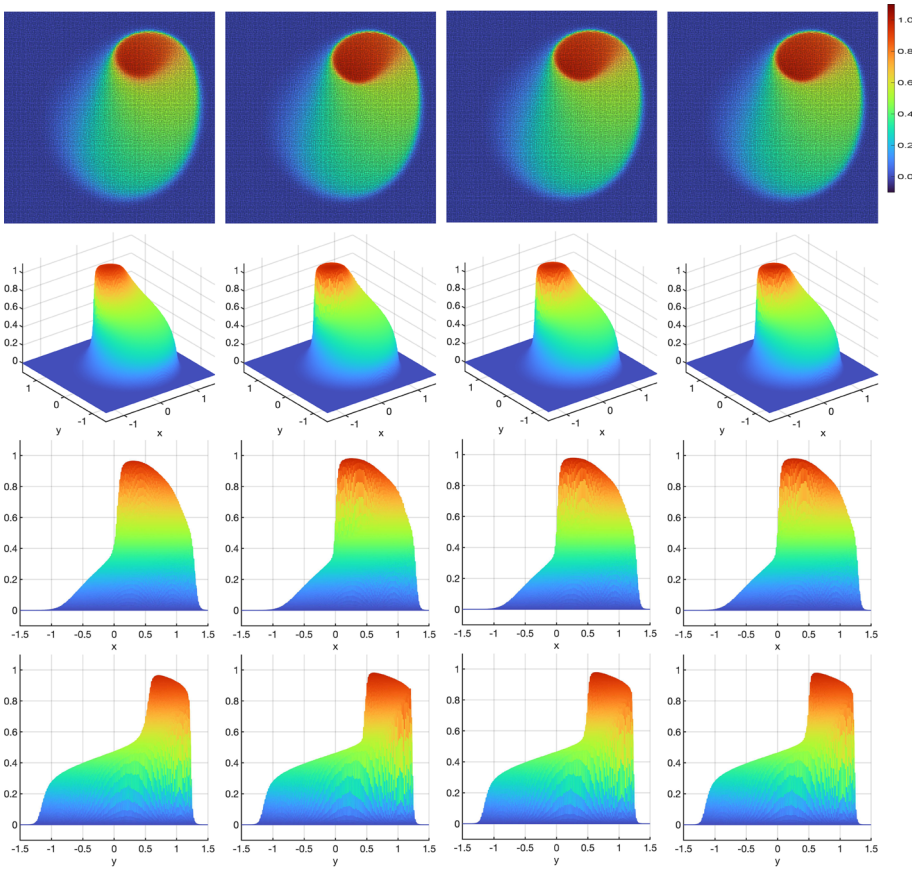
Fig. 17 Burgers flooding scenario in 2D at time $t = 1.0$ using $\Delta t = 0.025$, which is CFL 4.281. Profile of the solution showing overshoot for CN but not for the others

Tests are conducted using a quadrilateral mesh of 160×160 elements and time step $\Delta t = 0.0125$, giving CFL 3.144. Results at $t = 0.5$ are given in Figure 18. Again, the BE results show more numerical diffusion, and the CN and SATh results are fairly comparable, and similar to results seen in the literature.

7 A Higher Order Scheme

We return to the 1D equation. As noted in §2.1, the discontinuity aware quadrature rule is third order in the case of smooth functions [1]. We should postulate second order accuracy of the SATh scheme over time. However, Tables 1–2 clearly show first order convergence. This is due at least partly to the fact that the DAQ theory assumes θ is unrestricted, but $\theta \geq \theta_{\min}$ is required in the numerical scheme. However, θ is restricted only occasionally, and so a convergence above one is still to be expected.

A more prevalent issue is the use of constant cell boundary interface values (2.12), which limits the spatial accuracy to first order everywhere. To justify our postulate, we show the convergence rate in terms of Δt for the problem of Section 5.1.2 calculated using $\Delta t = \Delta x^{1/2}$



(a) BE (b) CN (c) $\text{SATh}_{1/2}$ (d) SATh_0

Fig. 18 Buckley-Leverett Equation with gravity in 2D at time $t = 0.5$ using $\Delta t = 0.0125$, which is CFL 3.144. Shown are the contour plot of the solutions, and the 3D plot and x and y cross-sections of the solution

in Table 5. Second order convergence in time is seen for CN, SATh_0-up, and $\text{SATh}_{1/2}$-up. Moreover, the SATh_0-up solutions are non-oscillatory in this test.

This observation motivates the use of higher order interface values. Within the schemes (SATh and CN), in this section we use a weighted essentially non-oscillatory (WENO) reconstruction in space to define the interface values. We use a standard WENO(3,2) reconstruction [8], but a WENO-AO(3,2) reconstruction [2] behaves similarly. In either case, the WENO reconstruction of $\bar{u}_{i-1}$, $\bar{u}_i$, and $\bar{u}_{i+1}$ on the cell E_i will be denoted by $R_i(x)$. The same reconstruction but using the data $\tilde{\bar{u}}_{i-1}$, $\tilde{\bar{u}}_i$, and $\tilde{\bar{u}}_{i+1}$ will be denoted by $\tilde{R}_i(x)$. Then instead of (2.12), we define for $e = x_{i+1/2}$,

$$u^-_{i+1/2} = R_i(x_{i+1/2}) \quad \text{and} \quad \tilde{u}^-_{i+1/2} = \tilde{R}_i(x_{i+1/2}), \tag{7.1}$$

and for $e = x_{i-1/2}$,

$$u^+_{i-1/2} = R_i(x_{i-1/2}) \quad \text{and} \quad \tilde{u}^+_{i-1/2} = \tilde{R}_i(x_{i-1/2}). \tag{7.2}$$

These expressions replace (2.16), and they are used within the schemes presented in Section 2.3 to define the interface fluxes, and also in the definition of θ, (2.11).

Table 5 Nonmonotone sine wave linear transport error and convergence order at $t = 0.5$ using $m = 1/\Delta x$ cells and $\Delta t = \Delta x^{1/2}$. The convergence rate reported is in terms of Δt

		CN		$SATh_0$-up		$SATh_{1/2}$-up	
m	CFL	L_h^1-error	order	L_h^1-error	order	L_h^1-error	order
32	2	8.31e-02	1.37	8.57e-02	1.42	8.74e-02	1.39
128	4	2.37e-02	1.81	2.38e-02	1.85	2.42e-02	1.86
512	8	6.12e-03	1.95	6.10e-03	1.96	6.16e-03	1.97
2048	16	1.54e-03	1.99	1.53e-03	1.99	1.55e-03	2.00
m	CFL	L_h^∞-error	order	L_h^∞-error	order	L_h^∞-error	order
32	2	1.30e-01	1.33	1.33e-01	1.34	1.50e-01	1.23
128	4	3.71e-02	1.81	3.80e-02	1.81	4.88e-02	1.62
512	8	9.61e-03	1.95	1.10e-02	1.80	1.53e-02	1.67
2048	16	2.42e-03	1.99	3.17e-03	1.79	5.01e-03	1.61

We remark that SATh-LF schemes have issues with Newton convergence at higher CFLs when high order WENO reconstruction is used. We found that using damped Newton and limiting the derivative of θ can be used to improve the convergence behavior.

7.1 Higher Order Linear Transport

As in Section 5.1.2, consider first the linear equation (5.1) with $L_0 = 0$, $L_1 = 1$, using the nonmonotone sine wave smooth initial condition (5.3) and periodic boundary conditions.

Convergence results for CFL 4 using the upstream numerical flux are shown in Table 6. The three schemes (CN, $SATh_0$-up, and $SATh_{1/2}$-up WENO) give overall second order accuracy in the L_h^1 norm. $SATh_{1/2}$-up has a convergence rate a bit below the other schemes due to its restriction $\theta \geq \theta_{\min} = 1/2$. The two schemes CN and $SATh_0$-up appear to be second order in the L_h^∞ norm, but $SATh_{1/2}$-up WENO converges at about only order 1.3 due to a slight flattening of the extrema. CN has the least error, and the $SATh_0$-up WENO solutions exhibit wiggles for the lower resolutions. Similar results for CFL 10 appear in Table 7; although, the $SATh_0$-up WENO solution is oscillatory at these resolutions.

7.2 Burgers Equation: Higher Order Shock Formation

Now consider Burgers equation (5.4) for $x \in (0, 1)$ with the smooth initial condition (5.3) and periodic boundary conditions. For this problem, the shocks form at time $t = 1/\pi \approx 0.318$, before which the solution is smooth, albeit nonmonotone.

The Lax-Friedrichs numerical flux is used. Tests using CFL 4 and a uniform mesh up to time $t = 0.25$ give the convergence results in Table 8. The CN and $SATh_0$-LF WENO schemes appear to have second order convergence in the L_h^1 norm, and $SATh_{1/2}$-LF WENO has nearly the same convergence. In the L_h^∞ norm, CN appears to maintain second order convergence, but the SATh schemes perhaps drop to order 1.5. The afore-mentioned difficulties with $SATh_0$-LF appear when using WENO reconstructions: the solution shows oscillation behind the steep front, which improves as the resolution increases. The problem also worsens as the CFL is increased.

Table 6 Nonmonotone sine wave linear transport error and convergence order at $t = 0.5$ using CFL 4, $m = 1/\Delta x$ cells, and $\Delta t = 4\Delta x$. The schemes use WENO(3,2) reconstruction of the interface values

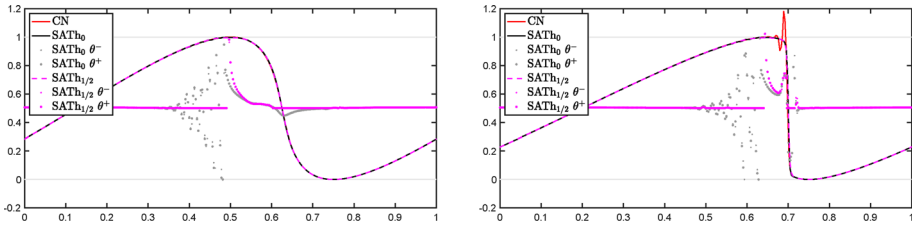
	CN WENO		SATh$_0$-up WENO		SATh$_{1/2}$-up WENO	
m	L_h^1-error	order	L_h^1-error	order	L_h^1-error	order
160	2.04e-03	1.97	4.52e-03	1.82	3.82e-03	1.83
320	5.16e-04	1.98	9.45e-04	2.26	1.04e-03	1.87
640	1.29e-04	1.99	1.90e-04	2.31	2.79e-04	1.90
1280	3.24e-05	2.00	4.14e-05	2.20	7.35e-05	1.93
m	L_h^∞-error	order	L_h^∞-error	order	L_h^∞-error	order
160	3.23e-03	1.95	1.60e-02	1.52	1.50e-02	1.27
320	8.11e-04	1.99	6.73e-03	1.25	6.07e-03	1.30
640	2.02e-04	2.01	1.73e-03	1.96	2.41e-03	1.33
1280	5.05e-05	2.00	5.05e-04	1.78	9.64e-04	1.32

Table 7 Nonmonotone sine wave linear transport error and convergence order at $t = 0.5$ using CFL 10, $m = 1/\Delta x$ cells, and $\Delta t = 10\Delta x$. The schemes use WENO(3,2) reconstruction of the interface values

	CN WENO				SATh$_{1/2}$-up WENO			
m	L_h^1-error	order	L_h^∞-error	order	L_h^1-error	order	L_h^∞-error	order
160	1.25e-02	1.91	1.97e-02	1.91	1.91e-02	1.36	4.62e-02	1.18
320	3.19e-03	1.98	5.03e-03	1.97	5.58e-03	1.77	1.96e-02	1.24
640	8.01e-04	1.99	1.26e-03	2.00	1.52e-03	1.87	8.00e-03	1.29
1280	2.01e-04	2.00	3.15e-04	2.00	4.20e-04	1.86	3.22e-03	1.31

Table 8 Burgers' equation (5.4) with periodic initial condition (5.3) at time $t = 0.25$, before shocks develop, using CFL 4 and $m = 1/\Delta x$ cells. The schemes use WENO(3,2) reconstructions

	CN WENO		SATh$_0$-LF WENO		SATh$_{1/2}$-LF WENO	
m	L_h^1-error	order	L_h^1-error	order	L_h^1-error	order
160	3.69e-03	1.56	2.33e-03	1.58	3.25e-03	1.53
320	1.07e-03	1.79	6.30e-04	1.89	1.01e-03	1.69
640	2.81e-04	1.93	1.64e-04	1.95	2.94e-04	1.78
1280	7.11e-05	1.98	4.16e-05	1.98	8.45e-05	1.80
m	L_h^∞-error	order	L_h^∞-error	order	L_h^∞-error	order
160	1.71e-02	1.36	9.08e-03	1.46	1.69e-02	1.23
320	5.45e-03	1.65	3.24e-03	1.49	5.45e-03	1.63
640	1.51e-03	1.85	1.05e-03	1.63	1.51e-03	1.86
1280	3.88e-04	1.96	3.38e-04	1.63	5.98e-04	1.33



(a) CFL 4, $\Delta x = 1/320$, $t = 0.25$ (b) CFL 4, $\Delta x = 1/320$, $t = 0.4$

Fig. 19 Burger's equation (5.4) with periodic initial condition (5.3) at time $t = 0.25$ (before the shock forms) and 0.4 (after the shock forms) using CFL 4 and $m = 1/\Delta x = 320$ mesh cells. The schemes use WENO(3,2) reconstructions. The true solution is shown in green on the left, but it is covered by the approximate solutions

Solutions are shown in Figure 19 at times $t = 0.25$, before the shocks form, and $t = 0.4$, after the shocks have formed. The resolution is sufficient that the SATh$_0$-LF WENO solution does not wiggle noticeably. All three schemes perform well at time $t = 0.25$, but CN WENO oscillates unacceptably immediately behind the shock at time $t = 0.4$.

8 Summary, Conclusions, and Open Problems

We presented further theoretical and numerical studies on the self-adaptive theta (SATh) scheme [1] for solving scalar conservation laws. We extended the scheme to unstructured meshes in multiple space dimensions, to numerical flux functions that can be split into right and left going waves, and to a higher order scheme by using WENO reconstructions for obtaining the values of the solution on the cell interfaces. An open problem is to properly define the scheme for systems of equations.

Theoretical results were given for problems in one space dimension with monotone increasing fluxes. The results in [1] made the assumption that $\theta_{\min} = 1/2$. These results were generalized here to allow $\theta_{\min} = 0$, at the expense of requiring that $\tilde{\bar{u}}_i^{n+1}$ lies between $\bar{u}_i^n$ and $\bar{u}_i^{n+1}$ for all $n \geq 0$ and $i \geq 1$. If the SATh scheme uses the upstream numerical flux and $\epsilon = 0$ in (2.17), it was shown that the scheme is stable for any $\theta_{\min} \geq 0$ and L-stable for the linear problem (Theorem 1). Furthermore, if the flux is strictly monotone increasing and the true solution is monotone (and $\theta_{\min} \geq 0$), then the SATh solution $\bar{u}$ remains monotone and the scheme is TVB and TVD (Theorem 2 and Corollary 1); moreover, θ and $\tilde{\bar{u}}$ are well behaved (Corollary 2).

Numerical tests in one space dimension were presented. It was noted that it can be difficult to solve the system of equations defining the SATh scheme using Newton's method. An open problem is to find a way to improve the convergence behavior.

Test problems with contact discontinuities, shocks, and rarefactions showed the behavior of SATh compared to finite volume schemes using backward Euler (BE) and Crank-Nicolson (CN) time stepping. As in [1], it was seen that SATh$_{1/2}$ pe'rforms better than the theory predicts. The SATh$_{1/2}$ solutions were less diffusive than those of BE and less oscillatory than those of CN while being about as sharp. Moreover, SATh$_{1/2}$-LF (i.e., SATh using a Lax-Friedrichs numerical flux) appears to be TVD and MPP. An major open problem is to prove this conjecture, as well as that SATh$_{1/2}$-LF is stable. When the true solution is monotone, it was seen that SATh$_0$ is an improvement over SATh$_{1/2}$.

However, it was discovered that using $\theta_{\min} = 0$ can be problematic when applied to problems outside the bounds of the theory, and especially to nonmonotone flows. In particular, SATh$_0$-LF solutions became oscillatory when the spatial resolution was too low, and thus

the MPP and TVD properties failed to hold. This observation suggests that one should use $\theta_{\min} = 1/2$ in multiple space dimensions and in one space dimension unless the true solution is known to be monotone (or the resolution can be made sufficiently fine).

Numerical tests in two space dimensions showed that BE is more diffusive than the other three schemes. The results for CN, SATh$_{1/2}$-LF, and SATh$_0$-LF were of similar accuracy. However, CN may show oscillation and violation of the maximum principle in some problems. SATh$_{1/2}$-LF and SATh$_0$-LF showed no oscillation in the tests conducted here (but the 1D tests demonstrate that at least SATh$_0$-LF does not satisfy the maximum principle in general).

The higher order SATh scheme was seen to converge to order two and compare favorably with CN, and SATh$_{1/2}$-LF proved to be less oscillatory than CN.

Data Availability Data will be made available on request.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Arbogast, T., Huang, C.S.: A self-adaptive theta scheme using discontinuity aware quadrature for solving conservation laws. IMA J. Numer. Anal. **42**(4), 3430–3463 (2022). https://doi.org/10.1093/imanum/drab071
2. Arbogast, T., Huang, C.S., Zhao, X.: Accuracy of WENO and adaptive order WENO reconstructions for solving conservation laws. SIAM J. Numer. Anal. **56**(3), 1818–1847 (2018). https://doi.org/10.1137/17M1154758
3. Boris, J.P., Book, D.L.: Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. Journal of computational physics **11**(1), 38–69 (1973). https://doi.org/10.1016/0021-9991(73)90147-2
4. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics. Springer-Verlag, Berlin Heidelberg (2005)
5. Hesthaven, J.S.: Numerical Methods for Conservation Laws: From Analysis to Algorithms. Computational Science and Engineering. Society for Industrial and Applied Mathematics, Philadelphia (2018)
6. Kuzmin, D., Löhner, R., Turek, S.: Flux-Corrected Transport: Principles, Algorithms, and Applications. Springer (2012)
7. LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge Univ. Press, Cambridge, England (2002)
8. Shu, C.W.: Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In: B. Cockburn, A. Quarteroni (eds.) Advanced numerical approximation of nonlinear hyperbolic equations, *Lecture Notes in Mathematics*, vol. 1697, chap. 4, pp. 325–432. Springer-Verlag, Berlin (1998)
9. Smoller, J.: Shock Waves and Reaction-Diffusion Equations, vol. 258. Springer Science & Business Media (2012)
10. Talischi, C., Paulino, G.H., Pereira, A., Menezes, I.F.M.: Polymesher: a general-purpose mesh generator for polygonal elements written in Matlab. Structural and Multidisciplinary Optimization **45**, 309–328 (2012)

11. Zalesak, S.T.: Fully multidimensional flux-corrected transport algorithms for fluids. Journal of Computational Physics **31**(3), 335–362 (1979). https://doi.org/10.1016/0021-9991(79)90051-2