STABILITY OF STRUCTURE-AWARE TAYLOR METHODS FOR TENTS

JAY GOPALAKRISHNAN AND ZHENG SUN

Abstract. Structure-aware Taylor (SAT) methods are a class of timestepping schemes designed for propagating linear hyperbolic solutions within a tent-shaped spacetime region. Tents are useful to design explicit time marching schemes on unstructured advancing fronts with built-in locally variable timestepping for arbitrary spatial and temporal discretization orders. The main result of this paper is that an s-stage SAT timestepping within a tent is weakly stable under the time step constraint $\Delta t \leq Ch^{1+1/s}$, where Δt is the time step size and h is the spatial mesh size. Improved stability properties are also presented for high-order SAT time discretizations coupled with low-order spatial polynomials. A numerical verification of the sharpness of proven estimates is also included.

1. Introduction

Spacetime methods for solving evolution equations can easily incorporate widely varying spatio-temporal grid sizes and discretization orders. However, to be competitive with standard timestepping methods, spacetime methods must have memory requirements and coupling of degrees of freedom that are comparable to standard timestepping methods. Such competitive spacetime methods can indeed be constructed for hyperbolic systems by partitioning the spacetime region into tent-shaped subregions satisfying causality: one then propagates numerical solutions asynchronously across an unstructured advancing front. The process of creating a mesh of tents by advancing spacetime fronts is referred to as "tent pitching" and recent methods like the Mapped Tent Pitching (MTP) schemes [6, 7] have proven themselves to be competitive tent-based alternatives to standard timestepping schemes, especially on complex geometries. Many previous works, both in the engineering and the mathematics literature, have constructed tent-based numerical methods [1, 5, 13, 14].

The purpose of this paper is to provide a complete stability analysis of the structure-aware Taylor (SAT) methods, a class of timestepping methods suitable for the above-mentioned MTP schemes applied to linear hyperbolic systems. To understand the origins of the SAT timestepping, consider the computational drawbacks that can arise from the lack of tensor-product structure in tent-shaped domains, including the inability to use standard discretizations combined with timestepping within a tent. A proposal to overcome such difficulties was made in [7]. The idea is to map the non-tensor-product tent region to a tensor-product cylindrical region. This made fully explicit timestepping within spacetime tents, combined with a standard discontinuous Galerkin (DG) spatial discretization, possible. Indeed, after the semidiscretization, the unknown function of a pseudotime variable $\hat{t}$, introduced later as

2020 *Mathematics Subject Classification.* 65M12.

Key words and phrases. Structure aware Taylor method, tent pitching, linear hyperbolic equations, stability analysis, discontinuous Galerkin methods.

$\hat{u}_h(\hat{t})$, satisfies the ordinary differential equation (ODE)

$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}}\left(M(\hat{t})\hat{u}_h\right) = A\,\hat{u}_h \tag{1.1}$$

with a time-dependent mass operator $M(\hat{t})$ and a differential operator A, defined later in (2.6). Introducing $y(\hat{t}) = M(\hat{t})\hat{u}_h(\hat{t})$, this ODE can be restated as

$$\frac{\mathrm{d}y}{\mathrm{d}\hat{t}} = AM(\hat{t})^{-1}y. \tag{1.2}$$

Although high-order standard Runge–Kutta timestepping can be applied to solve (1.2), the resulting solutions were not observed to achieve the expected high orders of accuracy, as reported in [6, 8]. New timestepping methods, incorporating the structure of the time-dependent mass matrix arising from tents, were then developed. Specifically, the SAT timestepping was proposed in [6] to address this issue for linear hyperbolic systems. Its extension to nonlinear hyperbolic systems, named SARK timestepping, was proposed in [8].

Energy-type stability estimates for these new timestepping schemes remained unknown until [4], where a framework for the stability and error analysis of the MTP methods for linear hyperbolic systems was constructed. For SAT schemes, the stability of the first- and the second-order methods were proved in [4]. In particular, the analysis requires a 3/2-Courant–Friedrichs–Levy (CFL) condition for the stability of the second-order methods. Here, as in [2], η-CFL condition refers to the time step constraint $\Delta t \leq Ch^{\eta}$, for some fixed constant C. Δt is the time step size and h is the spatial mesh size. Furthermore, based on the numerical tests in [8, Section 6.1], extrapolating from the provable cases, it was conjectured in [4] that the SAT method is stable under a $(1 + 1/s)$-CFL condition for MTP schemes, where s is the order of the SAT time discretization. This more restrictive CFL condition is also required by and known to be necessary for standard explicit Runge–Kutta methods when applied to hyperbolic equations in certain cases [2, 27, 26, 25]. For the SAT methods, the rigorous stability proof is only available for $s \leq 2$. The analysis for the higher-order methods remained open.

In this paper, we prove the above-mentioned conjecture for any s through an energy-type stability analysis. Since naive eigenvalue analyses with the stability regions may lead to insufficient and even misleading results [10, 12, 24, 20], energy arguments are widely used for stability analysis, especially for systems resulting from discretizations of partial differential equations. For implicit time marching methods or dissipative equations, universal stability analysis is well documented [3, 9]. However for hyperbolic type problems with high-order explicit methods, a systematic analysis was not available until recently. Based on the techniques developed in the analysis of the fourth order Runge–Kutta methods [20, 18], in [21], Sun and Shu proposed a general framework on analyzing the strong stability of explicit Runge–Kutta (Taylor) methods of arbitrary order. This work also relates to the fully discrete analysis of Runge–Kutta DG methods in [26, 25]. We also refer to [22, 23] on further extensions of the work in [21]. Results on nonlinear or non-autonomous problems can be found in [16, 17].

The main challenge in the analysis of the SAT method is to appropriately handle the mass matrix that is affine-linear in a pseudotime variable arising from the mapping. It leads to the following complications that have not been encountered in the analysis of standard Runge–Kutta (Taylor) methods. First, the numerical dissipation will depend on the time

derivative of the mass matrix. Second, the high-order spatial derivatives are defined via a recursive formula, rather than a simple matrix power. Third, there are extra tail terms arising in the simplification of energy equality, and finding an appropriate way of grouping the terms becomes an issue. The key ingredient for solving these issues is to introduce a novel discrete integration by parts formula for the MTP schemes (developed in Lemma 3.1 below). The analysis of the SAT method can be viewed as a generalization of the framework developed in [21]. In the special case of constant mass matrix, many results in Section 3 reduce to the known estimates for the standard Runge–Kutta or Taylor methods.

Furthermore, we show that when a low-order spatial discretization is coupled with a high-order SAT timestepping method, the fully discrete scheme may exhibit improved stability properties with a relaxed CFL condition. Consider symmetric linear hyperbolic systems with constant coefficients, we show that with spatially piecewise constant elements ($p = 0$), the SAT scheme is strongly stable under the usual CFL condition for any order. When the spatial polynomial degree satisfies $0 < p \leq (s-1)/2$, then we show that the numerical method is weakly stable under the $(1 + 1/(2s - 2p))$-CFL condition. The key step of the proof is to give an explicit characterization of the derivative operator in the SAT scheme (found in Lemma 4.1 below). The estimates are verified to be sharp within a subtent using the linear advection equation in one dimension (in Section 5). This investigation of improved stability when employing low-order polynomials is inspired by a similar study of the standard Runge–Kutta DG methods for linear advection by Xu et al. in [26]. It turns out that the SAT-DG methods in this paper exhibit stability properties that are different from those of the standard Runge–Kutta DG methods for linear autonomous equations—for the latter, strong stability can be achieved for $p > 0$ when sufficiently high-order timestepping methods are used. This is not the case for SAT-DG methods for (1.1), whose weak stability properties seem more in line with those of nonautonomous equations, which is perhaps not surprising since (1.2) is not autonomous.

The rest of the paper is organized as follows. In Section 2, we briefly outline the MTP scheme and state the corresponding ordinary differential equation (ODE) system arising from the semidiscretization after the tent mapping. The weak stability of the SAT method under the $(1 + 1/s)$-CFL condition is proved in Section 3. In Section 4, we prove the improved stability properties of SAT-DG schemes with low-order spatial polynomials. Then we show the sharpness of our estimates in Section 5 using the one-dimensional example. Proofs of all the lemmas in these sections are presented in Section 6 in the same order they appeared previously. Finally, conclusions and future work are discussed in Section 7.

2. Tents, Maps, and the SAT timestepping

In this section, we describe a model symmetric linear hyperbolic problem and how one constructs an advancing front solution on unstructured meshes using spacetime tents. Here we collect preliminary results from elsewhere that we need for the subsequent stability analysis.

Let $\Omega \subseteq \mathbb{R}^d$ represent the spatial domain of the simulation and let $u = u(x,t) : \Omega \times [0, t_{\max}] \to \mathbb{R}^b$ be a vector-valued function. Our goal is to solve the symmetric linear hyperbolic system

$$\partial_t g(u) + \operatorname{div}_x f(u) = 0, \tag{2.1}$$

with

$$[g(u)]_l = \sum_{k=1}^{b} \mathcal{G}_{lk} u_k, \quad [f(u)]_{lj} = \sum_{k=1}^{b} \mathcal{L}^{(j)}_{lk} u_k, \quad l = 1, \cdots, b, \quad j = 1, \cdots d. \tag{2.2}$$

Here $\mathcal{G} = [\mathcal{G}_{lk}] : \Omega \to \mathbb{R}^{b\times b}$ and $\mathcal{L}^{(j)} = [\mathcal{L}^{(j)}_{lk}] : \Omega \to \mathbb{R}^{b\times b}$ are symmetric bounded matrix-valued functions, and $\mathcal{G}$ is uniformly positive definite on $\overline{\Omega}$. Furthermore, let us assume that $\sum_{j=1}^{d} \partial_j \mathcal{L}^{(j)} = 0$ in the sense of distributions [4, Subsection 2.1], so that the (weighted) L^2 energy of (2.1) is nonincreasing in time. To avoid unnecessary technicalities, we consider periodic boundary conditions or compactly supported solutions in this paper, although more general boundary conditions can be handled using the techniques outlined in [4, Subsection 2.1]. We proceed to build a spacetime mesh of tents atop Ω.

First, we mesh the spatial domain. Let $\mathcal{T}$ denote a shape regular and conforming simplicial mesh of the spatial domain Ω. Let h be the mesh size parameter equaling the maximal diameter of elements in $\mathcal{T}$. We march forward in time by considering a sequence of advancing fronts $\varphi_i : \Omega \to \mathbb{R}$, $i = 0, \ldots, m$. Here $\{\varphi_i\}_{i=0}^{m}$ are continuous piecewise linear functions, specifically the lowest-order Lagrange finite element functions, on the mesh $\mathcal{T}$. In particular, we have $\varphi_0(x) \equiv 0$ and $\varphi_m(x) \equiv t_{\max}$. Given a vertex v, we define Ω^{v} to be the vertex patch which includes spatial simplices connecting to v. We advance from φ_i to φ_{i+1} over Ω^{v} by erecting a spacetime tent pole at the vertex v and forming the tent

$$T_i^{\mathrm{v}} = \{(x,t) : \ x \in \Omega^{\mathrm{v}}, \ \varphi_i(x) \le t \le \varphi_{i+1}(x)\}.$$

To ensure that each spacetime tent encloses the domain of dependence of all its points, we employ the "causality condition"

$$\|(\mathrm{grad}_x \varphi_i)(x)\|_2 < \frac{1}{c_{\max}}, \quad x \in \Omega, \quad i = 0, \ldots, m, \tag{2.3}$$

where $c_{\max}$ is a strict upper bound of the maximal hyperbolic wave speed. For a graphical illustration of the tent-pitching meshing process, we refer to [7, Figure 1].

In MTP schemes, one maps the tents to domains which are a tensor product of a spatial vertex patch and a "pseudotime" interval in order to gain efficiency and to allow reutilization of common spatial discretization tools and tensor-product techniques like timestepping. Consider a single tent over any given vertex patch Ω^{v},

$$T = \{(x,t) : x \in \Omega^{\mathrm{v}}, \varphi_{\mathrm{bot}} \le t \le \varphi_{\mathrm{top}}\}.$$

Here φ_{bot} and φ_{top} are restrictions of φ_i and φ_{i+1} over Ω^{v}. They are also continuous and piecewise linear. The goal is to solve (2.1) locally within the tent T from $t = \varphi_{\mathrm{bot}}$ to $t = \varphi_{\mathrm{top}}$ using a timestepping technique. To this end, we transform T into a tensor product domain $\hat{T} = \Omega^{\mathrm{v}} \times [0,1]$. See [7, Figure 2]. The required change of variables is given by $(x,t) = (x, \varphi(x,\hat{t}))$, where

$$\varphi(x,\hat{t}) = (1-\hat{t})\varphi_{\mathrm{bot}}(x) + \hat{t}\varphi_{\mathrm{top}}(x) = \varphi_{\mathrm{bot}}(x) + \hat{t}\delta(x).$$

Here, $\delta(x) = \varphi_{\mathrm{top}}(x) - \varphi_{\mathrm{bot}}(x)$ and $\hat{t}$ is what we referred to above as the pseudotime variable. From the causality condition (2.3), we know that $\delta \le Ch$ for some constant C depending on the wavespeed. In [7, Theorem 3.1] it is shown that the transformed unknown $\hat{u}(x,\hat{t}) = u(x,t)$ solves the equation

$$\partial_{\hat{t}} \left(g(\hat{u}) - f(\hat{u})\mathrm{grad}_x \varphi\right) + \mathrm{div}_x \left(\delta f(\hat{u})\right) = 0, \quad (x,\hat{t}) \in \hat{T}. \tag{2.4}$$

Hence MTP schemes proceed by first semidiscretizing (2.4) in space and then timestepping in pseudotime.

Let us now freeze the pseudotime variable and introduce notations associated with the spatial discretization. For the spatial discretization we use the standard DG space

$$V_h = \{v : v|_K \in [P_p(K)]^b, \text{ for all } K \in \mathcal{T} \text{ and } K \subseteq \Omega^{\mathrm{v}}\}.$$

Here $P_p(K)$ is the space of polynomials on K of degree less than or equal to p. Let $\mathcal{F}^{\mathrm{v}}$ be the set of facets on the spatial vertex patch Ω^{v}. On each facet F, let $\nu = [\nu_1, \cdots, \nu_d]$ denote a spatial unit normal vector, whose direction is currently irrelevant. Across each facet F, we define the jump $[\![v]\!] = \lim_{\varepsilon\to 0^+} v(x+\varepsilon\nu) - v(x-\varepsilon\nu)$ and the average $\{v\} = \lim_{\varepsilon\to 0^+}(v(x+\varepsilon\nu) + v(x-\varepsilon\nu))/2$. Furthermore, we introduce the following notations

$$(v,w) = \sum_{K\subseteq\Omega^{\mathrm{v}}} \int_K v\cdot w \,\mathrm{d}x, \qquad \text{for vector-valued functions } v, w : \Omega^{\mathrm{v}} \to \mathbb{R}^b,$$

$$\langle v,w\rangle = \sum_{K\subseteq\Omega^{\mathrm{v}}} \int_K \left(\sum_{i=1}^{b}\sum_{j=1}^{d} v_{ij}w_{ij}\right) \mathrm{d}x, \qquad \text{for matrix-valued functions } v=[v_{ij}], w=[w_{ij}] : \Omega^{\mathrm{v}} \to \mathbb{R}^{b\times d},$$

$$(v,w)_{\mathcal{F}^{\mathrm{v}}} = \int_{\mathcal{F}^{\mathrm{v}}} v\cdot w \,\mathrm{d}x, \qquad \text{for vector-valued functions } v, w : \mathcal{F}^{\mathrm{v}} \to \mathbb{R}^b,$$

and $\|\cdot\| = (\cdot,\cdot)^{1/2}$. Note that the vertex patch Ω^{v} is omitted in the notation for the $L^2(\Omega^{\mathrm{v}})$-norm and inner product to lighten the notation since a substantial part of our analysis will be carried out on a single given Ω^{v}. Given a selfadjoint operator $B : V_h \to V_h$, let $(v,w)_B = (Bv,w)$, $\|v\|_B = \sqrt{(v,v)_B}$ if B is positive definite, and $|v|_B = \sqrt{(v,v)_B}$ if B is positive semidefinite.

Applying standard DG discretization techniques to (2.4), we obtain the following semidiscrete scheme: find $\hat{u}_h(\cdot,\hat{t}) \in V_h$ such that

$$(\partial_{\hat{t}}[g(\hat{u}_h) - f(\hat{u}_h)\mathrm{grad}_x\varphi], v) = \langle \delta f(\hat{u}_h), \mathrm{grad}_x v\rangle + \left(\delta\hat{F}^{\nu}, [\![v]\!]\right)_{\mathcal{F}^{\mathrm{v}}}, \quad \text{for all } v \in V_h, \tag{2.5}$$

where the numerical flux $\hat{F}^{\nu}$ is given by

$$\hat{F}^{\nu} = \mathcal{D}\{\hat{u}_h\} - S[\![\hat{u}_h]\!],$$

using the matrix functions $\mathcal{D} = \sum_{j=1}^{d} \nu_j \mathcal{L}^{(j)}$ and S, a $b\times b$ constant symmetric positive semidefinite stabilization matrix. Let the operators $M_0, M_1, A : V_h \to V_h$ be such that their action on any given $\hat{u}_h \in V_h$ is defined by

$$(M_0\hat{u}_h, v) = (g(\hat{u}_h) - f(\hat{u}_h)\mathrm{grad}_x\varphi_{\mathrm{bot}}, v), \tag{2.6a}$$

$$(M_1\hat{u}_h, v) = (f(\hat{u}_h)\mathrm{grad}_x\delta, v), \tag{2.6b}$$

$$(A\hat{u}_h, v) = \langle \delta f(\hat{u}_h), \mathrm{grad}_x v\rangle + \left(\delta\hat{F}^{\nu}, [\![v]\!]\right)_{\mathcal{F}^{\mathrm{v}}}, \tag{2.6c}$$

for all $v \in V_h$. Furthermore, let $M(\hat{t}) = M_0 - \hat{t}M_1$. Then the DG scheme in (2.5) can be written as

$$(M\hat{u}_h)_{\hat{t}} = A\hat{u}_h, \quad \hat{t} \in [0,1]. \tag{2.7}$$

where we have denoted the derivative $\mathrm{d}(M\hat{u}_h)/\mathrm{d}\hat{t}$ by $(M\hat{u}_h)_{\hat{t}}$. Note that M_0, M_1 and A are independent of $\hat{t}$, while $M = M(\hat{t})$ is an affine linear function of $\hat{t}$. Since δ vanishes along

the boundary of Ω^{v}, there is no coupling through the numerical fluxes between Ω^{v} and its neighboring vertex patches. Hence the system (2.7) is defined locally within Ω^{v}, allowing us to localize all stability considerations. We will need the following properties of the above-defined operators established in [4, Lemmas 3.1 and 3.2]. We remark that although [4] additionally assumed what we state later in (4.1), the proofs of the propositions we list in this section, found there, do not use that assumption.

Proposition 2.1.

(1) *The operators M_0, M_1, and M are selfadjoint. In addition, the causality condition implies that M_0 and M are positive definite.*
(2) *The operator*
$$D := -(A^\top + A + M_1) \geq 0 \tag{2.8}$$
is positive semidefinite. Here $A^\top$ is the adjoint operator of A under $(\cdot,\cdot)$.

The semidefiniteness of D was crucially exploited in the stability analyses of [4] and will also be crucial in this paper. To understand why this is important, we reproduce a simple argument essentially contained in [4, Lemma 3.3].

Proposition 2.2 (Stability of the semidiscrete scheme)**.** *Solutions of* (2.7) *are stable in the weighted L^2-like norm $\|\cdot\|_M$, specifically,*
$$\frac{\mathrm{d}}{\mathrm{d}\hat{t}} \|\hat{u}_h\|_M^2 \leq 0.$$

Proof.
$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\hat{t}} (M\hat{u}_h, \hat{u}_h) &= ((M\hat{u}_h)_{\hat{t}}, \hat{u}_h) + (M\hat{u}_h, (\hat{u}_h)_{\hat{t}}) \\
&= ((M\hat{u}_h)_{\hat{t}}, \hat{u}_h) + (\hat{u}_h, M(\hat{u}_h)_{\hat{t}}) && \text{since } M \text{ is selfadjoint,} \\
&= ((M\hat{u}_h)_{\hat{t}}, \hat{u}_h) + (\hat{u}_h, (M\hat{u}_h)_{\hat{t}}) - (\hat{u}_h, M_{\hat{t}}\hat{u}_h) \\
&= (A\hat{u}_h, \hat{u}_h) + (\hat{u}_h, A\hat{u}_h) + (\hat{u}_h, M_1\hat{u}_h) && \text{by (2.7) and } M_{\hat{t}} = -M_1, \\
&= \big((A + A^\top + M_1)\hat{u}_h, \hat{u}_h\big) && \text{by definition of } A^\top, \\
&= -|\hat{u}_h|_D^2 && \text{by definition of } D \text{ and } |\cdot|_D,
\end{aligned}$$
so the result follows from (2.8). □

Finally, we turn to the full discretization by SAT timestepping. The SAT approximation of (2.7) at $\hat{t} = \tau$ is given by $\hat{u}_h(\tau) \approx R_s \hat{u}_h(0)$, where
$$R_s v = S_1 v + M^{-1} M_0 S_2 v, \text{ with} \tag{2.9}$$
$$S_1 v = \sum_{i=0}^{s-1} (i!)^{-1} X_i v, \qquad S_2 v = (s!)^{-1} X_s v, \tag{2.10}$$
and X_i is defined recursively by
$$X_0 = I \quad \text{and} \quad X_i = \tau M_0^{-1}(A + iM_1)X_{i-1} \quad \text{for } i \geq 1. \tag{2.11}$$
To ensure stability, we usually need τ to be sufficiently small. Therefore, for time marching on $\hat{T}$, we will need to divide $\hat{T}$ into several "subtents" and use the propagation operator (2.9) on each subtent, as we shall see later in Section 3.2.

Remark 2.3. In the special case $M_1 = 0$, we have

$$R_s v = \sum_{i=0}^{s} (i!)^{-1} \tau^i \tilde{A}^i v, \tag{2.12}$$

where $\tilde{A} = M_0^{-1} A$ is a negative semidefinite operator under the inner product $(\cdot, \cdot)_{M_0}$. Thus R_s reduces to a high-order Runge–Kutta (Taylor) operator for linear problems that has been analyzed in [21]. It is no surprise therefore that our analysis in Section 3 is substantially guided by the techniques in [21]. In addition, when $\tilde{A}$ in (2.12) represents the DG operator for linear advection, the energy estimate of (2.12) is essentially the stability estimate of the standard Runge–Kutta DG methods, which has been systematically studied by Xu et al. in [26, 25]. Some parts of our analyses are also inspired by their work, especially the improved estimate with low-order polynomials in Section 4.

For analyzing R_s, we need bounds on the norms of the various operators that go into building R_s. The following bounds are gathered from [4, Lemmas 3.1 and 4.4].

Proposition 2.4. *There is a $C_1 > 0$ independent of h such that*

$$\max\left(\|M_0\|, \left\|M_0^{-1}\right\|, \|M_1\|, \|M\|, \left\|M^{-1}\right\|, \|A\|, \|D\|\right) \le C_1.$$

Proposition 2.5. *There is an h-independent constant $C_2 > 0$ such that for any $j \ge i$,*

$$\|X_j v\|_{M_0} \le C_2 \tau^{j-i} \|X_i v\|_{M_0}, \tag{2.13}$$

$$|X_j v|_{\tau D} \le C_2 \tau^{j-i+\frac{1}{2}} \|X_i v\|_{M_0}, \tag{2.14}$$

$$\tau\left(M_0 S_2 v, M^{-1} M_1 S_2 v\right) \le C_2 \tau^{2s+1} \|v\|_{M_0}^2. \tag{2.15}$$

Proof. This follows immediately from the recursive definition of X_i in (2.11). □

This completes our review of the tent-based discretization whose stability we now proceed to analyze.

3. Stability analysis

The goal of our energy-type stability analysis is to show that $\|R_s v\|_M$ is appropriately bounded by $\|v\|_{M_0}$. We first obtain a bound on $\|R_s v\|_M$ in terms of $\|v\|_{M_0}$ and the pseudotime τ in Theorem 3.7. This then leads to the identification of a CFL condition and the main stability result of this section, Theorem 3.10.

Before proceeding, let us remark an important consequence of the stability estimates. As shown in [4]—see also Remark 3.12 below—if we define the "energy" at the tent's top and bottom as $\|R_s v\|_M$ and $\|v\|_{M_0}$, respectively, then one can combine such stability bounds with local truncation error estimates to obtain bounds for the global error at the final time, even on unstructured meshes.

3.1. Key ideas of the analysis.

Our proof of the above-mentioned Theorem 3.7 requires a number of quite technical steps. To ease entry into these technicalities, we identify and motivate the key ideas as lemmas here, whose proofs are postponed to Section 6. Using the lemmas, we can prove Theorem 3.7 at the end of this subsection.

To consider how we may bound $\|R_s v\|_M$ by $\|v\|_{M_0}$ for any $v \in V_h$, we begin by squaring both sides of (2.9). Since M and M_0 are selfadjoint, obvious manipulations yield

$$\begin{aligned}\|R_s v\|_M^2 &= (MR_s v, R_s v)\\ &= \left(MS_1 v + M_0 S_2 v, S_1 v + M^{-1} M_0 S_2 v\right)\\ &= \|S_1 v\|_M^2 + 2\left(S_1 v, S_2 v\right)_{M_0} + \left(M_0 S_2 v, M^{-1} M_0 S_2 v\right).\end{aligned}$$

Since $\|S_1 v\|_M^2 = \|S_1 v\|_{M_0}^2 - (S_1 v, S_1 v)_{\tau M_1}$,

$$\begin{aligned}\|R_s v\|_M^2 &= \|S_1 v\|_{M_0}^2 + 2\left(S_1 v, S_2 v\right)_{M_0} + \|S_2 v\|_{M_0}^2\\ &\quad + \left(M_0 S_2 v, (M^{-1} M_0 - I) S_2 v\right) - (S_1 v, S_1 v)_{\tau M_1}\\ &= \|S_1 v + S_2 v\|_{M_0}^2 + \left(M_0 S_2 v, (M^{-1}(M_0 - M)) S_2 v\right) - (S_1 v, S_1 v)_{\tau M_1}\\ &= \Big\| \sum_{i=0}^{s} (i!)^{-1} X_i v \Big\|_{M_0}^2 - (S_1 v, S_1 v)_{\tau M_1} + \tau \left(M_0 S_2 v, M^{-1} M_1 S_2 v\right). \qquad (3.1)\end{aligned}$$

For ease of notation, we introduce

$$F_{ij} = (X_i v, X_j v)_{\tau M_1}, \quad G_{ij} = (X_i v, X_j v)_{M_0}, \quad \text{and} \quad H_{ij} = (X_i v, X_j v)_{\tau D}.$$

These terms can be thought of as entries of symmetric matrices F, G, and H. Moreover, the diagonal entries of G and H are non-negative. Using F and G, we rewrite (3.1) as

$$\|R_s v\|_M^2 = \sum_{i,j=0}^{s} \frac{G_{ij}}{i!j!} - \sum_{i,j=0}^{s-1} \frac{F_{ij}}{i!j!} + \tau \left(M_0 S_2 v, M^{-1} M_1 S_2 v\right). \qquad (3.2)$$

From (2.15), it is clear that the last term is a high-order term in τ. The remaining lower-order terms above must be carefully sorted out to obtain a stability estimate.

To this end, a critical observation is that the off-diagonal entries G_{ij} for $j > i$ can be expressed in terms of closer-to-diagonal entries of G, F, and H, as stated next.

Lemma 3.1. *We have*

$$G_{ij} = -\frac{1}{2} H_{ii} + \left(i + \frac{1}{2}\right) F_{ii}, \qquad \text{if } j = i + 1, \qquad (3.3a)$$

$$G_{ij} = -G_{i+1,j-1} - H_{i,j-1} + (i + j) F_{i,j-1}, \qquad \text{if } j > i + 1. \qquad (3.3b)$$

We give a short proof in Section 6, which is simple, but obscures the origins of such identities. It is illustrative to draw an analogy with the (non-tent) case of $M_1 = 0$ and $X_i = (\tau M^{-1} A)^i$ is an approximation of $(\tau \partial_x)^i$, which corresponds to the special case when (2.1) represents one-dimensional transport. Then $G_{ij} = (X_i v, X_j v)_{M_0} \approx \tau^{i+j} (\partial_x^i v, \partial_x^j v)$ can be manipulated by integration by parts to obtain identities like that of the lemma. We may therefore view the identities of Lemma 3.1 as having originated in some discrete analog of integration by parts.

One can apply Lemma 3.1 recursively to simplify the first sum of (3.2). Indeed, an even more general sum can be rearranged as stated in the next lemma.

Lemma 3.2. *For any numbers α_{ij} with $\alpha_{ij} = \alpha_{ji}$, the identity*

$$\sum_{i,j=0}^{s} \alpha_{ij} G_{ij} = \sum_{i=0}^{s} \beta_i G_{ii} + \sum_{i,j=0}^{s-1} \gamma_{ij} H_{ij} + \sum_{i,j=0}^{s-1} \delta_{ij} F_{ij} \qquad (3.4)$$

holds with

$$\beta_i = \sum_{q=\max\{0,2i-s\}}^{\min\{2i,s\}} \alpha_{q,2i-q}(-1)^{i-q}, \tag{3.5a}$$

$$\gamma_{ij} = \sum_{q=\max\{0,i+j+1-s\}}^{\min\{i,j\}} (-1)^{\min\{i,j\}+1-q}\alpha_{q,i+j+1-q}, \tag{3.5b}$$

$$\delta_{ij} = \sum_{q=\max\{0,i+j+1-s\}}^{\min\{i,j\}} (-1)^{\min\{i,j\}-q}\alpha_{q,i+j+1-q}(i+j+1). \tag{3.5c}$$

When applying Lemma 3.2 to treat the first sum of (3.2), the case of interest is $\alpha_{ij} = (i!j!)^{-1}$. In this case, by a few combinatorial identities, we obtain the following explicit expressions for some of the coefficients introduced in Lemma 3.2.

Lemma 3.3. *When* $\alpha_{ij} = (i!j!)^{-1}$,

$$\beta_0 = 1 \quad \text{and} \quad \beta_i = 0, \qquad \text{for } 1 \le i \le s/2, \tag{3.6}$$

$$\gamma_{ij} = -(i!j!(i+j+1))^{-1}, \qquad \text{for } i+j \le s-1, \tag{3.7}$$

$$\delta_{ij} = (i!j!)^{-1}, \qquad \text{for } i+j \le s-1. \tag{3.8}$$

To motivate the next result, first substitute (3.4) into (3.2) to get

$$\|R_s v\|_M^2 = \sum_{i=0}^{s} \beta_i G_{ii} + \sum_{i,j=0}^{s-1} \gamma_{ij} H_{ij} + \sum_{i,j=0}^{s-1} \tilde{\delta}_{ij} F_{ij} + \tau \left(M_0 S_2 v, M^{-1} M_1 S_2 v\right), \tag{3.9}$$

where $\tilde{\delta}_{ij} = \delta_{ij} - (i!j!)^{-1}$. A number of terms in the first and last sums are zero by virtue of (3.6) and (3.8), respectively. One might also anticipate from (3.7) that a partial sum of the second sum in (3.9) is negative. Keeping these considerations in view, we introduce the following definition of critical indices to ease the bookkeeping.

Definition 3.4. Let

(1) $\zeta \le s$ be the positive integer such that $\beta_\zeta \ne 0$ and $\beta_i = 0$ for all $1 \le i < \zeta$,
(2) $\rho \le s$ be the largest integer such that the $\rho \times \rho$ principal submatrix $\Gamma_\rho = (\gamma_{ij})_{0 \le i,j \le \rho-1}$ is negative definite,
(3) $\sigma \le 2s$ be the largest integer such that $\tilde{\delta}_{ij} = 0$ for all $i + j \le \sigma$, and
(4) $\kappa = \min(2\zeta, 2\rho+1, \sigma+2)$.

Explicit expressions or estimates are obtained for ζ, ρ, σ, and κ in the case $\alpha_{ij} = (i!j!)^{-1}$ in the next result. We also list the numerical values of ζ, ρ, σ and κ for $1 \le s \le 20$ in Table 3.1.

Lemma 3.5. *When* $\alpha_{ij} = (i!j!)^{-1}$,

$$\zeta = \lfloor s/2 \rfloor + 1, \quad \rho \ge \lfloor (s+1)/2 \rfloor, \quad \sigma = s-1, \quad \text{and} \quad \kappa = s+1. \tag{3.10}$$

In the rest of the paper, we use τ_0 and C (with or without subscripts) to denote a constant that is independent of h and τ, but generally dependent on the order of SAT method s, the polynomial degree p, the mesh regularity constant, the norm of $g(u)$ and $f(u)$ in (2.2), the

s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ζ	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11
ρ	1	2	2	2	3	4	4	4	5	6	6	6	7	8	8	8	9	10	10	10
σ	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
κ	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

TABLE 3.1. Values of ζ, ρ, σ and κ in Definition 3.4 for some s.

constant in the causality condition $c_{\max}$, etc. The same symbol C may represent different values at different places. We will also extensively use the fact $\tau \leq 1$ to simplify our estimates.

Now, consider how we might attempt to bound the right hand side of (3.9). The next result, Lemma 3.6, which is proved using Lemma 3.5, tells us which low-order terms can be ignored while doing so. Proposition 2.5 tells us how the remaining high-order terms in (3.9) can be bounded by low-order ones. These ideas complete the analysis as shown next.

Lemma 3.6. *There exists positive constants τ_0, $C_{\beta,+}$, $C_{\gamma,+}$ and $C_{\delta,+}$, and a negative constant $C_{\gamma,-}$, such that for all $\tau \leq \tau_0$,*

$$\sum_{i=0}^{s} \beta_i G_{ii} \leq \beta_0 G_{00} + (\beta_\zeta + C_{\beta,+}\tau)\, G_{\zeta\zeta}, \tag{3.11}$$

$$\sum_{i,j=0}^{s-1} \gamma_{ij} H_{ij} \leq C_{\gamma,+}\tau G_{\rho\rho} + C_{\gamma,-} \sum_{l=0}^{\rho-1} H_{ll}, \tag{3.12}$$

$$\sum_{i,j=0}^{s-1} \tilde{\delta}_{ij} F_{ij} \leq C_{\delta,+}\tau^{\sigma+2} G_{00}. \tag{3.13}$$

Theorem 3.7. *There exists a constant τ_0 such that for all $\tau \leq \tau_0$, we have*

$$\|R_s v\|_M \leq \left(1 + C\tau^{s+1}\right) \|v\|_{M_0}, \quad \text{for all } v \in V_h. \tag{3.14}$$

Proof. Using the estimates of Lemma 3.6 in (3.9),

$$\begin{aligned} \|R_s v\|_M^2 \leq (1 + C_{\delta,+}\tau^{\sigma+2})\|v\|_{M_0}^2 + (\beta_\zeta + C_{\beta,+}\tau)\, G_{\zeta\zeta} + C_{\gamma,+}\tau G_{\rho\rho} \\ + \tau \left(M_0 S_2 v, M^{-1} M_1 S_2 v\right). \end{aligned} \tag{3.15}$$

Here we have used that $C_{\gamma,-} < 0$, $\beta_0 = 1$ (by Lemma 3.3), and $G_{00} = \|v\|_{M_0}^2$. Next, we use a consequence of Proposition 2.5, $G_{ii} = \|X_i v\|_{M_0}^2 \leq C\tau^{2i} \|v\|_{M_0}^2$, for indices $i = \zeta$ and ρ in (3.15). The result, when combined with an application of (2.15), yields

$$\|R_s v\|_M^2 \leq \left(1 + C\tau^{\min(2\zeta, 2\rho+1, \sigma+2, 2s+1)}\right) \|v\|_{M_0}^2. \tag{3.16}$$

By Lemma 3.5, $\kappa = \min(2\zeta, 2\rho+1, \sigma+2) = s+1$. Hence the theorem follows after taking the square root on both sides in (3.16). □

Remark 3.8. An equivalent way of stating (3.14) is via the following operator norm of R_s

$$\|R_s\|_{L(M_0,M)} := \sup_{0 \neq v \in V_h} \frac{\|R_s v\|_M}{\|v\|_{M_0}}. \tag{3.17}$$

Clearly, (3.14) is equivalent to $\|R_s\|_{L(M_0,M)} \le 1 + C\tau^{s+1}$.

3.2. **Subtents obeying a CFL condition.** Theorem 3.7 allows us to identify a CFL condition and a practical subdivision of the range of pseudotime that theoretically guarantees weak stability, as we shall now see. We divide the reference tent into r subtents, with

$$T_{[l]} = \{(x,t) : x \in \Omega^{\mathrm{v}}, \varphi(x,\hat{t}^{[l]}) \le t \le \varphi(x,\hat{t}^{[l+1]})\}, \quad l = 1, \cdots, r.$$

Here $\hat{t}^{[l]} = (l-1)/r$. The time step size for each subtent is $\tau = r^{-1}$. In the lth subtent, the propagator is defined as

$$R_{[l],s}v = \sum_{k=0}^{s-1} (k!)^{-1} X_k^{[l]} v + (M^{[l]})^{-1} M_0^{[l]} (s!)^{-1} X_s^{[l]} v.$$

Here $M_0^{[l]} = M_0 - (l-1)\tau M_1$, $M^{[l]} = M_0^{[l]} - \tau M_1$ and

$$X_0^{[l]} = I, \quad \text{and} \quad X_i^{[l]} = \tau (M_0^{[l]})^{-1}(A + kM_1) X_{i-1}^{[l]}, \quad \text{for all } i \ge 1.$$

The final solution operator at $\hat{t} = 1$ is given by

$$R_{r,s} = R_{[r],s} \circ R_{[r-1],s} \circ \cdots \circ R_{[1],s}.$$

Theorem 3.9. *If $\tau = r^{-1} \le Ch^{1/s}$ for C sufficiently small, then*

$$\|R_{r,s}v\|_{M(1)} \le (1 + Ch) \|v\|_{M_0}. \tag{3.18}$$

Proof. Note that $M_0^{[l]}$, $M^{[l]}$ and $X_j^{[l]}$ still satisfy Propositions 2.4 and 2.5, hence the estimate (3.14) of Theorem 3.7 holds after replacing R_s with $R_{[l],s}$, namely $\left\|R_{[l],s}\right\|_{L(M^{[l-1]},M^{[l]})} \le 1 + C\tau^{s+1}$, for all $1 \le l \le r$. As a result, for all $v \in V_h$, employing an analog of the operator norm in (3.17), we have

$$\begin{aligned}
\|R_{r,s}v\|_{M(1)} &\le \left\|R_{[r],s}\right\|_{L(M^{[r-1]},M^{[r]})} \cdots \left\|R_{[1],s}\right\|_{L(M^{[0]},M^{[1]})} \|v\|_{M_0} \\
&\le \left(1 + C\tau^{s+1}\right)^r \|v\|_{M_0} \\
&\le \left(1 + C\tau^{s+1} r\right) \|v\|_{M_0} \\
&\le (1 + C\tau^s) \|v\|_{M_0} \\
&\le (1 + Ch) \|v\|_{M_0}.
\end{aligned}$$

Here we have used the fact $\tau = r^{-1}$ in the second last inequality and $\tau \le Ch^{1/s}$ in the last inequality. □

Recall that $\delta \le Ch$. In the physical domain, the constraint $\tau \le Ch^{1/s}$ for the pseudotime coordinate should be interpreted as $\Delta t = \tau\delta \le Ch^{1+1/s}$, which leads to the following summary of the main result we have proven.

Theorem 3.10. *The SAT timestepping for the hyperbolic equation* (2.1) *is weakly stable—in the sense of* (3.18)*—under the* $(1 + 1/s)$*-CFL condition* $\Delta t \le Ch^{1+1/s}$ *whenever a spatial discretization satisfying the conclusions of Propositions 2.1 and 2.4 is used.*

Remark 3.11. From (3.18), it can be deduced that

$$\|\hat{u}_h^m\|_{L^2(\Omega)} \le \exp(Ct_{\max}) \left\|\hat{u}_h^0\right\|_{L^2(\Omega)}, \quad t_{\max} = m\Delta t.$$

In other words, the L^2 norm of the solution at the final time is bounded by a scalar multiple of that of the initial data—see [4, Remark 3.14] for further details.

Remark 3.12. Based on the stability result, one can prove a high-order error estimate for the fully discrete SAT-DG scheme for linear hyperbolic systems—see [4, Theorem 4.19].

4. Improved stability with low-order elements

In this section, we study the improved stability properties when a high-order SAT method is coupled with a low-order DG spatial discretization. We now proceed under the additional assumption that

$$\mathcal{G} \text{ and } \mathcal{L}^{(j)} \text{ are constant on each mesh element } K \in \mathcal{T}. \tag{4.1}$$

4.1. Key ideas for the low-order case.

Again, the analysis is based on the identity (3.9). This time however, we will observe that many terms there can simply be bounded, in the low-order case, by the dissipation terms $H_{ii} = |X_i v|^2_{\tau D}$, resulting in improved stability.

The first idea towards making this precise is the identification of a high-order spatial derivative term in X_i. To define this derivative term, first note that under the assumption (4.1), M_0 and M_1 reduce to point-wise linear operators

$$M_0 w = g(w) - f(w)\,\mathrm{grad}_x \varphi_{\mathrm{bot}}, \qquad M_1 w = f(w)\,\mathrm{grad}_x \delta, \tag{4.2}$$

for any $w \in V_h$. Furthermore, using integration by parts in (2.6c), as in [4, Lemma 3.2], it can be verified that for all $v, w \in V_h$,

$$(Aw, v) = -\left(\mathrm{div}_x(\delta f(w)), v\right) - \left(\delta \mathcal{D}[\![w]\!], \{v\}\right)_{\mathcal{F}^{\mathrm{v}}} - \left(\delta S[\![w]\!], [\![v]\!]\right)_{\mathcal{F}^{\mathrm{v}}}.$$

Here and throughout, differential operators like div_x and grad_x above, are applied element by element. By the product rule on each element, we see that $\mathrm{div}_x(\delta f(w)) = \delta \mathrm{div}_x f(w) + f(w)\mathrm{grad}_x \delta$ is in $P_p(K)^b$ for any $w \in V_h$ due to (4.1). Hence, letting $A_1 w = -\delta \mathrm{div}_x f(w)$, and defining a lifting operator L by $(Lw, v) := \left(\delta \mathcal{D}[\![w]\!], \{v\}\right)_{\mathcal{F}^{\mathrm{v}}} + \left(\delta S[\![w]\!], [\![v]\!]\right)_{\mathcal{F}^{\mathrm{v}}}$ for all $v, w \in V_h$, we can rewrite A as a sum of three linear operators,

$$A = A_1 - M_1 - L. \tag{4.3}$$

Let us define the operator K such that

$$Kw = -\tau M_0^{-1} \mathrm{div}_x f(w)$$

for any $w \in V_h$. On each element, Kw has one degree less than w. The next lemma rewrites the X_k defined in (2.11) using powers of K, which represent higher-order spatial derivative operators. As before, all lemmas are proved in Section 6.

Lemma 4.1. *For all $i \geq 0$, we have*

$$X_i = \delta^i K^i + Z_i, \tag{4.4}$$

where Z_i is defined recursively by

$$Z_0 = 0 \quad \text{and} \quad Z_i = \tau M_0^{-1}(A + iM_1 + L)Z_{i-1} - \tau M_0^{-1} L X_{i-1}. \tag{4.5}$$

The message of Lemma 4.1 is that X_i can be decomposed as a scalar multiple of a high-order spatial derivative (namely K^i) plus Z_i. When $i \geq p+1$, we have $K^i v = 0$ and $X_i = Z_i$.

The next key idea is that the norm of Z_i can be bounded by the sum of H_{ll}, which arises from the dissipation due to DG jumps, as shown in the next lemma. When combined with Lemma 4.1 and the observation that $X_i = Z_i$ for $i \geq p+1$, this then yields bounds for some of the numbers G_{ii} and F_{ij} in the subsequent result, Lemma 4.3.

Lemma 4.2. *For all $i \geq 0$, we have*

$$\|Z_i v\|_{M_0}^2 \leq C\tau \sum_{l=0}^{i-1} H_{ll}. \tag{4.6}$$

Lemma 4.3. *For all $i \geq p+1$, we have*

$$G_{ii} \leq C\tau \sum_{l=0}^{p} H_{ll}, \tag{4.7}$$

$$|F_{ij}| \leq C\varepsilon^{-1}\tau^{2(i+j-p)+1} G_{00} + \varepsilon \sum_{l=0}^{p} H_{ll}, \quad \text{for all } \varepsilon > 0. \tag{4.8}$$

It now only remains to apply the above inequalities in Lemma 3.6 and use the resulting bounds in (3.9) to obtain improved stability estimates. We proceed to discuss this separately for $p = 0$ in Subsection 4.2 and for $0 < p \leq (s-1)/2$ in Subsection 4.3.

4.2. **Strong stability for the lowest-order case.** In this section, we show that when the DG spatial discretization is used with $p = 0$, the SAT scheme is strongly stable under the usual CFL condition for any temporal order $s \geq 1$.

Theorem 4.4. *If $p = 0$, then there exists a constant τ_0 such that for all $\tau \leq \tau_0$, we have*

$$\|R_s v\|_M \leq \|v\|_{M_0}, \quad \text{for all } v \in V_h. \tag{4.9}$$

Proof. The proof proceeds by bounding the terms in the identity (3.9). By Proposition 2.5,

$$G_{jj} = \|X_j v\|_{M_0}^2 \leq C \|X_1 v\|_{M_0}^2 = C G_{11} \quad \text{for all } j \geq 1. \tag{4.10}$$

Hence by Lemma 3.6,

$$\sum_{i=0}^{s} \beta_i G_{ii} + \sum_{i,j=0}^{s-1} \gamma_{ij} H_{ij} \leq \beta_0 G_{00} + C G_{11} + C_{\gamma,-} H_{00}, \tag{4.11}$$

where we have used the fact that $C_{\gamma,-} < 0$ and $H_{ll} \geq 0$ to drop the high-order H_{ll} terms. Again, by Proposition 2.5, $|F_{ij}| \leq C G_{ii}^{1/2} G_{jj}^{1/2}$, which when combined with (4.10), yields

$$\sum_{i,j=0}^{s-1} \tilde{\delta}_{ij} F_{ij} + \tau \left(M_0 S_2 v, M^{-1} M_1 S_2 v \right) \leq C G_{11}. \tag{4.12}$$

Here we have used the fact that $\tilde{\delta}_{i0} = \tilde{\delta}_{0j} = 0$ for all $0 \leq i, j \leq s-1$ (recall that $\sigma = s-1$) and $S_2 v = X_s v / s!$ (see (2.10)) with $s \geq 1$. Using (4.11) and (4.12) in (3.9) and recalling

that $\beta_0 G_{00} = \|v\|_{M_0}^2$, we get

$$\|R_s v\|_M^2 \le \|v\|_{M_0}^2 + CG_{11} + C_{\gamma,-} H_{00}.$$

Applying (4.7) of Lemma 4.3 with $p = 0$, we have $G_{11} \le C\tau H_{00}$, and hence

$$\|R_s v\|_M^2 \le \|v\|_{M_0}^2 + (C_{\gamma,-} + C\tau)\, H_{00}.$$

Since $C_{\gamma,-} < 0$, by taking τ sufficiently small, we obtain (4.9). □

Repeatedly applying Theorem 4.4 on successive subtents, we obtain the following analogue of Theorem 3.9.

Theorem 4.5. *If $p = 0$, then for any s, there exists a constant τ_0 such that when $\tau = r^{-1} \le \tau_0$, we have*

$$\|R_{r,s} v\|_{M(1)} \le \|v\|_{M_0}, \qquad \text{for all } v \in V_h.$$

As a result, the SAT-DG0 scheme is strongly stable under the usual CFL condition $\Delta t \le Ch$.

4.3. **Improved weak stability for other low-order cases.** We now prove a better weak stability result for lower-order DG discretizations beyond the lowest-order case. The improvement is visible when comparing the powers of τ in (3.14) and (4.13), and the consequent less restrictive CFL condition in Theorem 4.7.

Theorem 4.6. *If $0 < p \le (s-1)/2$, then there exists a constant τ_0, such that for all $\tau \le \tau_0$,*

$$\|R_s v\|_M \le (1 + C\tau^{2s-2p+1}) \|v\|_{M_0}, \qquad \text{for all } v \in V_h. \tag{4.13}$$

Proof. By (3.10) of Lemma 3.5, the given condition on p implies that $\zeta = \lfloor s/2 \rfloor + 1 \ge p+1$ and $\rho \ge \lfloor (s+1)/2 \rfloor \ge p+1$. Therefore, we can apply Lemma 4.3 to estimate $G_{\zeta\zeta}$ and $G_{\rho\rho}$ in (3.11) and (3.12) to get

$$\sum_{i=0}^{s} \beta_i G_{ii} \le \beta_0 G_{00} + C\tau \sum_{l=0}^{p} H_{ll}, \tag{4.14}$$

$$\sum_{i,j=0}^{s-1} \gamma_{ij} H_{ij} \le C\tau \sum_{l=0}^{p} H_{ll} + C_{\gamma,-} \sum_{l=0}^{\rho-1} H_{ll}. \tag{4.15}$$

To estimate (3.13), note $\tilde{\delta}_{ij} = 0$ for $i + j \le \sigma = s - 1$. While for $i + j \ge s$, we have

$$\min_{i,j,i+j\ge s} \max\{i,j\} \ge \left\lceil \frac{i+j}{2} \right\rceil \ge \left\lceil \frac{s}{2} \right\rceil \ge p+1.$$

Therefore, one can invoke (4.8) for each term in the summation (3.13), which gives

$$\sum_{i,j=0}^{s-1} \tilde{\delta}_{ij} F_{ij} = \sum_{i+j\ge s, i,j\le s-1} \tilde{\delta}_{ij} F_{ij} \le C\varepsilon^{-1}\tau^{2s-2p+1} G_{00} + \varepsilon \sum_{l=0}^{p} H_{ll}. \tag{4.16}$$

With this we have bounded all terms on the right hand side of (3.9) except the last. For the last term in (3.9), we can use Proposition 2.4 and Lemma 4.1 to obtain

$$\left(M_0 S_2 v, M^{-1} M_1 S_2 v\right) \le C \|X_s v\|_{M_0}^2 = CG_{ss} \le C\tau \sum_{l=0}^{p} H_{ll}. \tag{4.17}$$

Using the bounds of (4.14), (4.15), (4.16) and (4.17) in (3.9), we obtain

$$\|R_s v\|_M^2 \le \left(\beta_0 + C\varepsilon^{-1}\tau^{2s-2p+1}\right) G_{00} + \left(C_{\gamma,-} + C\tau + \varepsilon\right) \sum_{i=0}^{\rho-1} H_{ii}.$$

Here we have used that $p \le \rho - 1$. Note that $\beta_0 = 1$ and $G_{00} = \|v\|_{M_0}^2$. Since $C_{\gamma,-} < 0$, we prove (4.13) by taking τ and ε to be so small such that $C_{\gamma,-} + C\tau + \varepsilon < 0$. □

Theorem 4.7. *Suppose $0 < p \le (s-1)/2$. Then there exists a constant C such that when $\tau = r^{-1} \le Ch^{1/(2s-2p)}$, we have*

$$\|R_{r,s} v\|_{M(1)} \le (1 + Ch) \|v\|_{M_0}, \quad \text{for all } v \in V_h.$$

As a result, the SAT-DG method is weakly stable under the $(1 + 1/(2s-2p))$-CFL condition $\Delta t \le Ch^{1+1/(2s-2p)}$.

Proof. Following along the lines of the proof of Theorem 3.9, we have

$$\begin{aligned}\|R_{r,s} v\|_{M(1)} \le& \left(1 + C\tau^{2s-2p+1}\right)^r \|v\|_{M_0} \le (1 + C\tau^{2s-2p+1} r) \|v\|_{M_0} \\ \le& (1 + C\tau^{2s-2p}) \|v\|_{M_0} \le (1 + Ch) \|v\|_{M_0}.\end{aligned}$$

Here we have used the fact $\tau = r^{-1} \le Ch^{1/(2s-2p)}$ in the last two inequalities. □

5. Illustration using linear advection

In this section, we will use the one-dimensional linear advection

$$\partial_t u + \partial_x u = 0 \tag{5.1}$$

with the upwind DG discretization to illustrate that the estimates of Theorems 3.7, 4.4, and 4.6 cannot generally be improved. For the simple equation (5.1), the associated matrices are of moderate sizes and many quantities can be evaluated analytically.

5.1. **Matrix form and the norm of the SAT operator.** We consider a uniform mesh partition of mesh size h of the one-dimensional domain with the origin $x = 0$ as a mesh point. Setting the pitch vertex v to the spatial point $x = 0$, we pitch a tent from $t = 0$, corresponding to $\varphi_{\text{bot}} = 0$, over the vertex patch $\Omega^{\text{v}} = I_0 \cup I_1$. Here $I_0 = (-h, 0]$ and $I_1 = (0, h]$. We march forward in time to the point $(x, t) = (0, h)$, thus making a spacetime tent, and consider a subtent of it where pseudotime $\tau < 1$. Since the causality condition is $\Delta t < h$ for this problem, $M(\tau)$ is invertible for $\tau < 1$.

Focusing on this single tent, we introduce a spatial basis of Legendre polynomials $L_i(x)$, which is defined recursively as

$$L_0(x) = 1, \quad L_1(x) = x, \quad \text{and} \quad L_{i+1}(x) = \frac{2i+1}{i+1} x L_i(x) + \frac{i}{i+1} L_{i-1}(x), \quad \text{for all } i \ge 2.$$

The piecewise polynomial basis over Ω^{v} for the DG space is defined using normalized Legendre polynomials on I_0 and I_1, namely

$$\begin{aligned} b_{0i}(x) &= \sqrt{\frac{2i+1}{h}} \bar{b}_{0i}(x), && \text{with } \bar{b}_{0i}(x) = L_i\left(\frac{x + h/2}{h/2}\right) 1_{[-h,0]}(x), \\ b_{1i}(x) &= \sqrt{\frac{2i+1}{h}} \bar{b}_{1i}(x), && \text{with } \bar{b}_{1i}(x) = L_i\left(\frac{x - h/2}{h/2}\right) 1_{[0,h]}(x). \end{aligned}$$

In this basis, functions w in V_h are represented by their vector $\mathtt{w}$ of coefficients in the basis expansion, i.e.,

$$w(x) = \sum_{i=0}^{p} w_{0i} b_{0i}(x) + \sum_{i=0}^{p} w_{1i} b_{1i}(x)$$

is represented as the vector $\mathtt{w} = [w_{00}, \cdots w_{0p}, w_{10}, \cdots, w_{1p}]^\top$. In the same basis, keeping the same block partitioning corresponding to the two spatial intervals, we have the following matrix representations of M, M_0, M_1 and A:

$$\mathtt{M} = \mathtt{M}_0 - \tau \mathtt{M}_1, \quad \mathtt{M}_0 = \mathtt{I}_{2p+2}, \quad \mathtt{M}_1 = \begin{bmatrix} \mathtt{I}_{p+1} & \mathtt{O}_{p+1} \\ \mathtt{O}_{p+1} & -\mathtt{I}_{p+1} \end{bmatrix}, \quad \mathtt{A} = \begin{bmatrix} \mathtt{A}_{00} & \mathtt{A}_{01} \\ \mathtt{A}_{10} & \mathtt{A}_{11} \end{bmatrix}, \tag{5.2}$$

where $\mathtt{I}_l$ and $\mathtt{O}_l$ are the lth order identity and zero matrices, respectively, and

$$(\mathtt{A}_{00})_{ij} = \int_{-h}^{0} (x+h)\, b_{0j} \partial_x b_{0i} \mathrm{d}x - h b_{0j}(0) b_{0i}(0), \qquad (\mathtt{A}_{01})_{ij} = 0,$$
$$(\mathtt{A}_{10})_{ij} = h b_{0j}(0) b_{1i}(0), \qquad (\mathtt{A}_{11})_{ij} = \int_{0}^{h} (h-x)\, b_{1j} \partial_x b_{1i} \mathrm{d}x.$$

One can show that $\mathtt{A}$ is independent of h (in accordance with Proposition 2.4).

The matrix representation of the SAT propagation operator can now be written down using $\mathtt{X}_k$, defined recursively by $\mathtt{X}_0 = \mathtt{I}_{2p+2}$ and $\mathtt{X}_i = \tau \mathtt{M}_0^{-1}(\mathtt{A} + i\mathtt{M}_1)\mathtt{X}_{i-1}$. The vector representation of $R_s w$ for any $w \in V_h$ equals $\mathtt{R}_s \mathtt{w}$ where

$$\mathtt{R}_s = \sum_{k=0}^{s-1} (i!)^{-1} \mathtt{X}_i + \mathtt{M}^{-1} \mathtt{M}_0 (s!)^{-1} \mathtt{X}_s.$$

Following [8, Section 6.1], the norm $\|R_s\|_{L(M_0,M)}$ defined in (3.17), can be computed by

$$\|R_s\|_{L(M_0,M)} = \sup\{|\lambda|^{\frac{1}{2}} : 0 \neq \mathtt{w} \in \mathbb{R}^{2p+2}, \mathtt{R}_s^\top \mathtt{M} \mathtt{R}_s \mathtt{w} = \lambda \mathtt{M}_0 \mathtt{w}\}. \tag{5.3}$$

5.2. **Numerical illustration of stability inequalities.** Since the matrix $\mathtt{R}_s$ depends only on τ, the operator norm in (5.3) is a one-variable function of τ. We use Mathematica© for evaluation of (5.3) and conduct a Taylor expansion of $\|R_s\|_{L(M_0,M)}$ with respect to τ. The leading terms in these Taylor series are documented in Table 5.2 with different values of s and p. Here is a summary of observations in the table:

(1) For $p = 0$ and for any s, we observe that $\|R_s\|_{L(M_0,M)} = 1$.
(2) For $0 < p \leq (s-1)/2$, we observe that $\|R_s\|_{L(M_0,M)} \leq 1 + C\tau^{2s-2p+1}$.
(3) In general, we observe that $\|R_s\|_{L(M_0,M)} \leq 1 + C\tau^{s+1}$.

These observations indicate that our analysis in the previous sections is sharp.

	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$	$s=7$	$s=8$
$p=0$	0	0	0	0	0	0	0	0
$p=1$	$2\tau^2$	$0.08\tau^3$	$0.25\tau^5$	$0.25\tau^7$	$0.25\tau^9$	$0.25\tau^{11}$	$0.25\tau^{13}$	$0.25\tau^{15}$
$p=2$	$13.32\tau^2$	$0.25\tau^3$	$0.08\tau^4$	$7.31\tau^5$	$3.19\tau^7$	$6\tau^9$	$9.91\tau^{11}$	$15\tau^{13}$
$p=3$	$46.20\tau^2$	$1.16\tau^3$	$0.09\tau^4$	$156.15\tau^5$	$7.72\tau^6$	$0.27\tau^7$	$27.19\tau^9$	$45.31\tau^{11}$
$p=4$	$117.66\tau^2$	$3.53\tau^3$	$0.10\tau^4$	$1511.49\tau^5$	$204.33\tau^6$	$2.81\tau^7$	$7.81\tau^8$	$21.47\tau^9$

TABLE 5.2. Leading terms of Taylor expansions in $\|R_s\|_{L(M_0,M)} - 1$ with respect to τ for (5.1) as $\tau \ll 1$. For example, the entry $2\tau^2$ with $p=1$ and $s=1$ corresponds to $\|R_s\|_{L(M_0,M)} = 1 + 2\tau^2 + \mathcal{O}(\tau^3)$.

6. Proofs of the lemmas

6.1. **Proof of Lemma 3.1.** When $j = i+1$, by the definition of X_{i+1}, we have

$$\begin{aligned} G_{ij} &= (X_i v, X_{i+1} v)_{M_0} = \tau \left(X_i v, (A + (i+1)M_1) X_i v\right) \\ &= \tau \left(X_i v, \left(\frac{1}{2}\left(A + A^\top + M_1\right) + \left(i + \frac{1}{2}\right) M_1 \right) X_i v \right) \\ &= -\frac{1}{2} |X_i v|^2_{\tau D} + \left(i + \frac{1}{2}\right) (X_i v, X_i v)_{\tau M_1} , \end{aligned}$$

which proves the first identity of the lemma. In the case $j > i+1$, using the fact that M_1 is selfadjoint, we have

$$\begin{aligned} G_{ij} &= (X_i v, X_j v)_{M_0} = \tau \left(X_i v, (A + jM_1) X_{j-1} v\right) \\ &= \tau \left(X_i v, \left(-(A + (i+1)M_1)^\top + (A + A^\top + M_1) + (i+j)M_1 \right) X_{j-1} v \right) \\ &= -\left(\tau (A + (i+1)M_1) X_i v, X_{j-1} v\right) - (X_i v, X_{j-1} v)_{\tau D} + (i+j) (X_i v, X_{j-1} v)_{\tau M_1} \\ &= -(X_{i+1} v, X_{j-1} v)_{M_0} - (X_i v, X_{j-1} v)_{\tau D} + (i+j) (X_i v, X_{j-1} v)_{\tau M_1} , \end{aligned}$$

which proves the second identity of the lemma. □

6.2. **Proof of Lemma 3.2.** We proceed to prove Lemma 3.2 using inductive applications of Lemma 3.1. It will be convenient to denote

$$\nu_{lm} = \begin{cases} 1/2, & l = m \\ 1, & l \neq m \end{cases}$$

and adopt the convention that $G_{\phi\phi} = 0$ if ϕ is not an integer (so when $i+j$ is not even, the quantity $G_{\frac{i+j}{2},\frac{i+j}{2}}$ below vanishes).

Lemma 6.1. *For any $j \geq i$, we have*

$$G_{ij} = (-1)^{\frac{j-i}{2}} G_{\frac{i+j}{2},\frac{i+j}{2}} + \sum_{k=0}^{\lfloor \frac{j-i-1}{2} \rfloor} (-1)^{k+1} \nu_{i+k,j-k-1} H_{i+k,j-k-1} + (i+j) \sum_{k=0}^{\lfloor \frac{j-i-1}{2} \rfloor} (-1)^{k} \nu_{i+k,j-k-1} F_{i+k,j-k-1}.$$

Proof. The identity is trivial for the diagonal entries with $j = i$. For the superdiagonal entries, where $j = i+1$, the stated identity is the same as (3.3a) of Lemma 3.1. When $j - i \geq 2$, the identity can be proved by applying Lemma 3.1 recursively and formalizing by mathematical induction. If $j - i$ is even, then the recursion terminates in the obvious diagonal case. If $j - i$ is odd, then the recursion instead terminates in the superdiagonal case of (3.3a). In both cases we obtain the stated identity. □

We will use the identity of Lemma 6.1 to expand $\sum_{ij} \alpha_{ij} G_{ij}$ to prove Lemma 3.2. A few more preparations on rearrangements of sums will be helpful for the proof.

Lemma 6.2. *For any numbers μ_{ij}, the variable change $m = i + j$ and $q = i$ yields*

$$\sum_{i,j=0}^{s} \mu_{ij} = \sum_{m=0}^{2s} \sum_{q=\max\{0,m-s\}}^{\min\{m,s\}} \mu_{q,m-q}. \tag{6.1}$$

In particular, if $\mu_{ij} = 0$ when $i + j$ is odd, then the variable change $m = 2l$ in (6.1) *gives*

$$\sum_{i,j=0}^{s} \mu_{ij} = \sum_{l=0}^{s} \sum_{q=\max\{0,2l-s\}}^{\min\{2l,s\}} \mu_{q,2l-q}. \tag{6.2}$$

Proof. The sum over the discrete square region $0 \leq i \leq s, 0 \leq j \leq s$, under the given variable change $m = i + j$ and $q = i$, becomes a sum over the discrete parallelogram region $P = \{(q,m) \in \mathbb{Z}^2 : \ 0 \leq q \leq s, \ 0 \leq m - q \leq s\}$, i.e.,

$$\sum_{i,j=0}^{s} \mu_{ij} = \sum_{(q,m)\in P} \mu_{q,m-q}.$$

It is easy to see (considering the boundaries of the parallelogram) that $P = \{(q,m) \in \mathbb{Z}^2 : 0 \leq m \leq 2s, \ \max(0, m-s) \leq q \leq \min(m,s)\}$, so (6.1) follows. Finally, (6.2) can be obtained by dropping terms in (6.1) with an odd m and substituting in $m = 2l$. □

Lemma 6.3. *For any numbers μ_{ijk}, the variable change $l = i + k$, $m = j - k - 1$ and $q = i$ yields*

$$\sum_{i,j=0,\, j>i}^{s} \sum_{k=0}^{\lfloor \frac{j-i-1}{2} \rfloor} \mu_{ijk} = \sum_{l,m=0}^{s-1} \sum_{m\geq l}^{l} \sum_{q=\max\{0,l+m+1-s\}} \mu_{q,l+m+1-q,l-q}. \tag{6.3}$$

Proof. The left sum is over the region $\{(i,j,k) \in \mathbb{Z}^3 : 0 \leq i < j \leq s, \ 0 \leq k \leq (j-i-1)/2\}$. The change of variable $i = q$, $j = l + m + 1 - q$ and $k = l - q$, obviously transforms the

region to

$$T_1 = \left\{ (q,l,m) \in \mathbb{Z}^3 :\ 0 \le q < l+m+1-q \le s,\ 0 \le l-q \le \frac{1}{2}(l+m) - q \right\}.$$

The sum on the right hand side of (6.3) is over

$$T_2 = \{(q,l,m) \in \mathbb{Z}^3 :\ 0 \le l \le m \le s-1,\ 0 \le q,\ \ l+m+1-s \le q \le l\}, \tag{6.4}$$

so it is enough to show that $T_1 = T_2$.

Let $(q,l,m) \in T_1$. Then since $0 \le q < l+m+1-q \le s$, we have

$$0 \le q \quad \text{and} \quad l+m+1-s \le q, \tag{6.5}$$

two inequalities needed for membership in T_2. Moreover, since $0 \le l-q \le \frac{1}{2}(l+m)-q$, we have $q \le l$ and $l \le m$, which together with (6.5) implies that $0 \le l$ and $l+m+1-s \le q \le l$. The latter, in particular, implies $m \le s-1$. Thus, having obtained all the inequalities in (6.4), we conclude that $T_1 \subseteq T_2$. It is also easy to show that $T_2 \subseteq T_1$, so $T_1 = T_2$. □

Proof of Lemma 3.2. By Lemma 6.1 and the symmetry of α_{ij},

$$\sum_{i,j=0}^{s} \alpha_{ij} G_{ij} = \sum_{i=0}^{s} \alpha_{ii} G_{ii} + 2 \sum_{i,j=0,\ j>i}^{s} \alpha_{ij} G_{ij} = S_\beta + S_\gamma + S_\delta,$$

where

$$S_\beta = \sum_{i,j=0}^{s} \alpha_{ij} (-1)^{\frac{j-i}{2}} G_{\frac{i+j}{2},\frac{i+j}{2}}, \quad S_\gamma = 2 \sum_{i,j=0,j>i}^{s} \alpha_{ij} \sum_{k=0}^{\lfloor \frac{j-i-1}{2} \rfloor} (-1)^{k+1} \nu_{i+k,j-k-1} H_{i+k,j-k-1},$$

$$S_\delta = 2\,(i+j) \sum_{i,j=0,\ j>i}^{s} \alpha_{ij} \sum_{k=0}^{\lfloor \frac{j-i-1}{2} \rfloor} (-1)^{k} \nu_{i+k,j-k-1} F_{i+k,j-k-1}.$$

By the variable change (6.2) in Lemma 6.2, we have

$$S_\beta = \sum_{l=0}^{s} \left(\sum_{q=\max\{0,2l-s\}}^{\min\{2l,s\}} \alpha_{q,2l-q}\,(-1)^{l-q} \right) G_{ll} = \sum_{l=0}^{s} \beta_l G_{ll}$$

where β_l is as defined in (3.5a). Next, apply the variable change of Lemma 6.3 to S_γ and S_δ. Then

$$\begin{aligned} S_\gamma &= 2 \sum_{l,m=0,m\ge l}^{s-1} \left(\sum_{q=\max\{0,l+m+1-s\}}^{\min\{l,m\}} (-1)^{\min\{l,m\}+1-q} \alpha_{q,l+m+1-q} \right) \nu_{lm} H_{lm} \\ &= \sum_{l,m=0}^{s-1} \left(\sum_{q=\max\{0,l+m+1-s\}}^{\min\{l,m\}} (-1)^{\min\{l,m\}+1-q} \alpha_{q,l+m+1-q} \right) H_{lm} = \sum_{l,m=0}^{s-1} \gamma_{lm} H_{lm}. \end{aligned}$$

Here we have used the identity $2\sum_{l,m=0,\, m\geq l}^{s-1}\nu_{lm}\mu_{lm} = \sum_{l,m=0}^{s-1}\mu_{lm}$, which holds for all μ_{lm} satisfying $\mu_{lm} = \mu_{ml}$. Similarly, for S_δ, we have

$$S_\delta = \sum_{l,m=0}^{s-1}\left(\sum_{q=\max\{0,l+m+1-s\}}^{\min\{l,m\}}(-1)^{\min\{l,m\}-q}\alpha_{q,l+m+1-q}\,(l+m+1)\right)F_{lm} = \sum_{l,m=0}^{s-1}\delta_{lm}F_{lm}.$$

The sum of these expressions for S_β, S_γ, and S_δ proves (3.4). □

6.3. **Proof of Lemma 3.3.** In this proof, we shall use the following combinatorial identities.

Lemma 6.4.

$$\sum_{q=0}^{2i}\left(q!(2i-q)!\right)^{-1}(-1)^{i-q} = 0, \quad \textit{for any integer } i \geq 1, \tag{6.6}$$

$$\sum_{q=0}^{i}\binom{i+j+1}{q}(-1)^{i-q} = \binom{i+j}{i} \quad \textit{for any integers } i,j \geq 0. \tag{6.7}$$

Proof. To prove (6.6), we use the following binomial expansion for real x,

$$x^i(1+x)^{2i} = x^i\sum_{q=0}^{2i}\binom{2i}{q}x^q = (2i)!\sum_{q=0}^{2i}\left(q!(2i-q)!\right)^{-1}x^{i+q}.$$

The result follows by choosing $x=-1$ and replacing $(-1)^{i+q}$ with $(-1)^{i-q}$.

To prove (6.7), we will first show that given $l \geq 1$

$$\sum_{q=0}^{i}\binom{l}{q}(-1)^{i-q} = \binom{l-1}{i}, \quad \text{for all } i \leq l. \tag{6.8}$$

Here a binomial coefficient $\binom{k}{i}$ is to be considered as zero when $i > k$, as happens for the $i = l$ case above. Obviously, (6.8) holds for $l = 1$. To use induction on l, suppose (6.8) holds for $l = k$ and for any $i \leq k$. Then using the identity $\binom{k+1}{q} = \binom{k}{q} + \binom{k}{q-1}$ and the induction hypothesis, we have for $l = k+1$ and $i \leq k$,

$$\begin{aligned}\sum_{q=0}^{i}\binom{k+1}{q}(-1)^{i-q} &= \sum_{q=0}^{i}\binom{k}{q}(-1)^{i-q} + \sum_{q=1}^{i}\binom{k}{q-1}(-1)^{i-q}\\ &= \sum_{q=0}^{i}\binom{k}{q}(-1)^{i-q} + \sum_{q=0}^{i-1}\binom{k}{q}(-1)^{i-1-q} = \binom{k-1}{i} + \binom{k-1}{i-1} = \binom{k}{i},\end{aligned}$$

i.e., (6.8) holds for $l = k+1$ and $i \leq k$. The identity also holds for $l = k+1$ and $i = k+1$, as can be seen by choosing $x = -1$ in the binomial expansion of $(1+x)^{k+1}$. So we have shown that (6.8) holds for $l = k+1$ and $i \leq k+1$. Hence, by induction, (6.8) holds for any $l \geq 1$. The identity (6.7) follows by setting $l = i+j+1$ in (6.8). □

Proof of Lemma 3.3. From (3.5a) of Lemma 3.2, it is obvious that $\beta_0 = \alpha_{00} = 1$. When $1 \leq i \leq s/2$, substituting $\alpha_{ij} = (i!j!)^{-1}$ into (3.5a) and using the (6.6) of Lemma 6.4, we

obtain

$$\beta_i = \sum_{q=0}^{2i} \alpha_{q,2i-q}(-1)^{i-q} = \sum_{q=0}^{2i} (q!(2i-q)!)^{-1}(-1)^{i-q} = 0,$$

thus proving (3.6). To prove (3.7), due to the symmetry of γ_{ij}, we proceed assuming without loss of generality that $j \geq i$. Then, for $i + j \leq s - 1$, (3.5b) yields

$$\begin{aligned}\gamma_{ij} &= \sum_{q=0}^{i} (-1)^{i+1-q}\alpha_{q,i+j+1-q} = \sum_{q=0}^{i} \frac{(-1)^{i+1-q}}{q!(i+j+1-q)!} \\ &= -\left((i+j+1)!\right)^{-1} \sum_{q=0}^{i} \binom{i+j+1}{q} (-1)^{i-q} \\ &= -\left((i+j+1)!\right)^{-1} \binom{i+j}{i} = -(i!j!(i+j+1))^{-1},\end{aligned}$$

where we have used (6.7) of Lemma 6.4. This proves (3.7). Finally, to prove (3.8), note that (3.5c) implies that $\delta_{ij} = -(i+j+1)\gamma_{ij}$. Hence the result $\delta_{ij} = (i!j!)^{-1}$ for $i + j \leq s - 1$ follows immediately from the just established expression for γ_{ij}. □

6.4. **Proof of Lemma 3.5.**

Step 1. It is immediate from (3.6) and the definition of ζ that $\zeta > s/2$. In fact $\zeta = \lfloor s/2 \rfloor + 1$. To see this, apply (3.5a) with $i = \lfloor s/2 \rfloor + 1$, noting that $2i$ is either $s + 1$ or $s + 2$. Then applying (6.6) to (3.5a) gives

$$\beta_{\lfloor s/2 \rfloor + 1} = \left(\sum_{q=0}^{2i} - \sum_{q=0}^{2i-(s+1)} - \sum_{q=s+1}^{2i} \right) \alpha_{q,2i-q}(-1)^{i-q} = -\left(\sum_{q=0}^{2i-(s+1)} + \sum_{q=s+1}^{2i} \right) \alpha_{q,2i-q}(-1)^{i-q}.$$

With the variable change $l = 2i - q$ and the symmetry $\alpha_{ij} = (i!j!)^{-1} = \alpha_{ji}$, one can get

$$\beta_{\lfloor s/2 \rfloor + 1} = -\sum_{l=s+1}^{2i} \alpha_{2i-l,l}(-1)^{l-i} - \sum_{q=s+1}^{2i} \alpha_{q,2i-q}(-1)^{i-q} = -2 \sum_{q=s+1}^{2i} \alpha_{q,2i-q}(-1)^{i-q}.$$

This is a sum of one or two terms which can be easily verified to be nonzero. Hence $\beta_{\lfloor s/2 \rfloor + 1} \neq 0$, so $\zeta = \lfloor s/2 \rfloor + 1$.

Step 2. From (3.7), we know that in particular, for all $0 \leq i, j \leq (s-1)/2$, we have $\gamma_{ij} = -(i!j!(i+j+1))^{-1}$. Let $\Lambda = \operatorname{diag}\left(0!, 1!, 2!, \cdots, \left\lfloor \frac{s-1}{2} \right\rfloor !\right)$. Then

$$\Gamma_{\left\lfloor \frac{s+1}{2} \right\rfloor} = -\Lambda^{-1}\mathcal{H}_{\left\lfloor \frac{s+1}{2} \right\rfloor}\Lambda^{-1} \tag{6.9}$$

where $\mathcal{H}_m$ denotes the $m \times m$ Hilbert matrix. Since Hilbert matrices are positive definite, the matrix in (6.9) is negative definite. Hence $\rho \geq \lfloor (s+1)/2 \rfloor$.

Step 3. Since $\tilde{\delta}_{ij} = \delta_{ij} - (i!j!)^{-1}$, using (3.8), we have $\tilde{\delta}_{ij} = 0$ when $i + j \leq s - 1$. Hence we have $\sigma \geq s - 1$. Using (3.5c) and (6.7), it is easy to check that $\tilde{\delta}_{ij}$ is nonzero when $i + j = s$. Hence $\sigma = s - 1$.

Step 4. Note that $2\zeta \geq s + 1$, $2\rho + 1 \geq s + 1$, and $\sigma + 2 = s + 1$. As a result, we have $\kappa = \min(2\zeta, 2\rho + 1, \sigma + 2) = \sigma + 2 = s + 1$. □

6.5. **Proof of Lemma 3.6.** By Lemma 3.5, $\sum_{i=0}^{s}\beta_i G_{ii} = \beta_0 G_{00} + \sum_{i=\zeta}^{s}\beta_i G_{ii}$, so the first inequality of the lemma (3.11) is immediately obtained by applying (2.13) of Proposition 2.5.

To prove (3.12), since $\Gamma_\rho < 0$ is negative definite, there exists a constant $C_- < 0$ such that $\Gamma_\rho - C_- I_\rho < 0$ remains negative definite. A simple argument (see [20, Lemma 2.3]) then proves that $\sum_{i,j=0}^{\rho-1}[\Gamma_\rho - C_- I_\rho]_{ij}\,(X_i v, X_j v)_{\tau D} \le 0$. Hence

$$\sum_{i,j=0}^{\rho-1}\gamma_{ij}H_{ij} \le C_- \sum_{l=0}^{\rho-1} H_{ll}. \tag{6.10}$$

The remaining summands on the left hand side of (3.12) involve indices with $\max(i,j) \ge \rho$. By the Cauchy–Schwarz inequality $\left|(X_i v, X_j v)_{\tau D}\right| \le |X_i v|_{\tau D}|X_j v|_{\tau D}$, and if $i \ge \rho$, then (2.14) of Proposition 2.5 yields $|X_i v|_{\tau D} \le C\tau^{1/2}\left\|X_\rho v\right\|_{M_0}$. Thus

$$|\gamma_{ij}H_{ij}| \le C\tau G_{\rho\rho}, \quad \text{for } i \ge \rho \text{ and } j \ge \rho. \tag{6.11}$$

In case only one of i or j is greater than or equal to ρ, say $i \le \rho-1$ and $j \ge \rho$ without loss of generality, then in addition to the Cauchy–Schwarz inequality, we also apply the inequality $ab \le \varepsilon a^2 + (4\varepsilon)^{-1}b^2$, for any $0 < \varepsilon < \varepsilon_0$ with ε_0 to be specified, to get $\left|(X_i v, X_j v)_{\tau D}\right| \le |X_i v|_{\tau D}|X_j v|_{\tau D} \le \varepsilon|X_i v|^2_{\tau D} + (4\varepsilon)^{-1}|X_j v|^2_{\tau D}$. Bounding the term with the larger index using (2.14), we have

$$|H_{ji}| = |H_{ij}| \le \varepsilon H_{ii} + C\varepsilon^{-1}\tau G_{\rho\rho}, \quad \text{for } i \le \rho-1 \text{ and } j \ge \rho. \tag{6.12}$$

Combining (6.10), (6.11), and (6.12), it gives

$$\sum_{i,j=0}^{s-1}\gamma_{ij}H_{ij} \le C\left(1+\varepsilon^{-1}\right)\tau G_{\rho\rho} + (C_- + C\varepsilon)\sum_{l=0}^{\rho-1}H_{ll}.$$

Choosing ε_0 small enough so that $C_- + C\varepsilon_0 \le C_-/2$, we have proven (3.12) with $C_{\gamma,+} = C\left(1+\varepsilon^{-1}\right)$ and $C_{\gamma,-} = C_-/2$.

It only remains to prove (3.13). In its left hand sum, by definition of σ in Definition 3.4, only summands with indices in $T_\sigma = \{(i,j) \in \mathbb{Z}^2 : i+j > \sigma \text{ and } 0 \le i,j \le s-1\}$ are nontrivial. The summands can be bounded by the estimates of Proposition 2.4 and 2.5 $F_{ij} \le C\tau\left\|X_i v\right\|_{M_0}\left\|X_j v\right\|_{M_0} \le C\tau^{i+j+1}\|v\|^2_{M_0}$. Hence

$$\sum_{i,j=0}^{s-1}\tilde{\delta}_{ij}F_{ij} = \sum_{(i,j)\in T_\sigma}\tilde{\delta}_{ij}F_{ij} \le C\sum_{(i,j)\in T_\sigma}\tau^{i+j+1}G_{00}.$$

Since $i+j+1 \ge \sigma+2$ for $(i,j) \in T_\sigma$, the inequality (3.13) follows. □

6.6. **Proof of Lemma 4.1.** We use induction. Note that $X_0 = I$ admits the described form (4.4). Assuming that (4.4) holds for $i = k$, we need to prove that it holds for $i = k+1$. Subtracting the recursive defining equation (4.5) of Z_{k+1} from that of X_{k+1} (namely (2.11)), and using (4.3),

$$X_{k+1} - Z_{k+1} = \tau M_0^{-1}\left(A_1 + kM_1\right)(X_k - Z_k).$$

Using the induction hypothesis $X_k - Z_k = \delta^k K^k$ and the definitions of A_1 and M_1, we obtain, for any $v \in V_h$,

$$\begin{aligned}(X_{k+1} - Z_{k+1})\, v &= -\tau M_0^{-1} \left(\delta \operatorname{div}_x \left(f(\delta^k K^k v)\right) - k\, f(\delta^k K^k v) \operatorname{grad}_x \delta\right) \\ &= -\tau M_0^{-1} \left(\delta \operatorname{div}_x \left(\delta^k f(K^k v)\right) - k\, \delta^k f(K^k v) \operatorname{grad}_x \delta\right) \\ &= -\tau M_0^{-1} \left(\delta^{k+1} \operatorname{div}_x f(K^k v)\right).\end{aligned}$$

Here we have used the fact that f is homogeneous of degree 1 (recall (2.2)) in the second equality and the product rule for differentiation in the third equality. Since M_0 acts pointwise (see (4.2)) the last expression is the same as $\delta^{k+1}(-\tau M_0^{-1}\operatorname{div}_x f) \circ K^k v = \delta^{k+1} K^{k+1} v$, thus establishing the formula (4.4) for $i = k+1$. □

6.7. **Proof of Lemma 4.2.** We start by proving a preparatory bound on the norm of $\|Lv\|$.

Lemma 6.5. *For all $v \in V_h$,*

$$\|Lv\| \le C\tau^{-\frac{1}{2}} |v|_{\tau D} \le C \|v\|.$$

Proof. The second inequality can be obtained by applying Proposition 2.5. We now prove the first inequality. Using inverse estimates and the fact $\|\delta\|_{L^\infty} \le Ch$, it can be seen that

$$(\delta\{w\}, \{w\})_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}} \le C \|w\| \quad \text{and} \quad (\delta[\![w]\!], [\![w]\!])_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}} \le C \|w\|.$$

Hence using the Cauchy–Schwarz inequality and the fact that $\mathcal{D}$ and S are bounded, we get

$$\begin{aligned}(Lv, w) \le& C\, (\delta[\![v]\!], [\![v]\!])_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}}\, (\delta\{w\}, \{w\})_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}} + C\, (\delta[\![v]\!], [\![v]\!])_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}}\, (\delta[\![w]\!], [\![w]\!])_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}} \\ \le& C\, (\delta[\![v]\!], [\![v]\!])_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}} \|w\|.\end{aligned}$$

Taking $w = Lv$, we deduce that

$$\|Lv\| \le C\, (\delta[\![v]\!], [\![v]\!])_{\mathcal{F}^{\mathrm{v}}}^{\frac{1}{2}}. \tag{6.13}$$

By [4, Lemma 3.2], we have $|v|_{\tau D}^2 = 2\tau\, (\delta S[\![v]\!], [\![v]\!])_{\mathcal{F}^{\mathrm{v}}}$. Letting $\lambda > 0$ denote the smallest eigenvalue of the positive definite matrix S, we have $\delta[\![v]\!] \cdot [\![v]\!] \le \lambda^{-1} \delta S[\![v]\!] \cdot [\![v]\!]$. Integrating and using (6.13),

$$C \|Lv\|^2 \le (\delta[\![v]\!], [\![v]\!])_{\mathcal{F}^{\mathrm{v}}} \le \frac{1}{\lambda} (\delta S[\![v]\!], [\![v]\!]) = \frac{1}{2\lambda\tau} |v|_{\tau D}^2.$$

□

Proof of Lemma 4.2. It suffices to show that

$$\|Z_i v\| \le C\tau^{\frac{1}{2}} \sum_{l=0}^{i-1} |X_l v|_{\tau D}, \quad \text{for all } i \ge 0. \tag{6.14}$$

Indeed, (6.14) implies

$$\|Z_i v\|_{M_0}^2 \le C \|Z_i v\|^2 \le C \left(\tau^{\frac{1}{2}} \sum_{l=0}^{i-1} |X_l v|_{\tau D}\right)^2 \le C\tau \sum_{l=0}^{i-1} |X_l v|_{\tau D}^2 = C\tau \sum_{l=0}^{i-1} H_{ll},$$

thus completing the proof of (4.6).

To prove (6.14), we use induction on i. Since $Z_0 = 0$, the inequality (6.14) certainly holds for the base case $i = 0$. Assume (6.14) holds for $i = k$. By the definition of Z_{k+1} in (4.5) and the triangle inequality, we have

$$\begin{aligned}\|Z_{k+1}v\| &\le \tau \left\|M_0^{-1}\left(A + (k+1)M_1\right)\right\| \|Z_k v\| + \tau \left\|M_0^{-1}\right\| \|LZ_k v\| + \tau \left\|M_0^{-1}LX_k v\right\| \\ &\le C\|Z_k v\| + C\|LZ_k v\| + C\tau\|LX_k v\| \\ &\le C\|Z_k v\| + C\tau^{\frac{1}{2}}|X_k v|_{\tau D} \\ &\le C\tau^{\frac{1}{2}} \sum_{l=0}^{k} |X_l v|_{\tau D}.\end{aligned}$$

Here we have applied Proposition 2.4 (and $\tau \le 1$) in the second inequality, Lemma 6.5 in the third inequality, and the induction hypothesis in the last inequality. Therefore, (6.14) holds for $i = k + 1$ and hence for all $i \ge 0$. □

6.8. **Proof of Lemma 4.3.** We first prove (4.7). Since $K^{p+1}v = 0$ for $v \in P_p(K)$, we have $X_{p+1} = Z_{p+1}$. Therefore, using Proposition 2.5 and Lemma 4.2, it can be shown that for any $i \ge p + 1$,

$$\begin{aligned}G_{ii} &= \|X_i v\|_{M_0}^2 \le C\tau^{2i-2(p+1)} \|X_{p+1}v\|_{M_0}^2 \\ &= C\tau^{2i-2(p+1)} \|Z_{p+1}v\|_{M_0}^2 \le C\tau^{2i-2p-1} \sum_{l=0}^{p} H_{ll} \le C\tau \sum_{l=0}^{p} H_{ll}.\end{aligned} \tag{6.15}$$

Next, to prove (4.8), we apply the Cauchy–Schwarz inequality, Proposition 2.4, and (2.13) to get

$$|F_{ij}| = \left|(X_i v, X_j v)_{\tau M_1}\right| \le C\tau \|X_i v\|_{M_0} \|X_j v\|_{M_0} \le C\tau^{i+j-p} \|v\|_{M_0} \|X_{p+1} v\|_{M_0}.$$

Invoking (6.15) with $i = p + 1$,

$$|F_{ij}| \le C\tau^{i+j-p} G_{00}^{\frac{1}{2}} G_{p+1,p+1}^{\frac{1}{2}} \le \left(C\tau^{i+j-p+\frac{1}{2}} G_{00}^{\frac{1}{2}}\right) \left(\sum_{l=0}^{p} H_{ll}\right)^{\frac{1}{2}}, \tag{6.16}$$

which yields (4.8) after applying the inequality $ab \le (4\varepsilon)^{-1}a^2 + \varepsilon b^2$. □

7. Conclusion

We have presented a systematic stability analysis of the SAT methods for MTP schemes for solving linear hyperbolic equations. We proved the conjecture formulated in [4], that the SAT method is weakly stable under the $(1 + 1/s)$-CFL condition, is true. The analysis in this paper generalizes the results in [21] by including an affine linear time-dependent mass matrix. Furthermore, improved stability estimates are obtained for symmetric linear hyperbolic systems with piecewise constant coefficients and with DG discretizations. With P_0-DG spatial discretization, the SAT timestepping was proved to be strongly stable under the usual CFL condition for any temporal order s. With P_p-DG spatial discretization, the SAT scheme is weakly stable under the $(1 + 1/(2s - 2p))$-CFL condition when $0 < p \le (s - 1)/2$. The estimates are numerically verified to be sharp in each subtent for the one-dimensional linear advection equation. Finally, it is our hope that the new understanding presented in our analysis will inspire further ideas to improve numerical strategies for explicit

time-stepping on unstructured advancing fronts using tents. Of particular interest is the development of a tent-based scheme that is strongly stable under the usual CFL condition. Stabilization techniques with artificial viscosity [22, 15] and the relaxation time stepping methods [11, 19] may be promising avenues.

Acknowledgments

The work of the first author was partially supported by the NSF grant DMS-1912779. The work of the second author was partially supported by the NSF grant DMS-2208391. We thank Dr. Jin Jin at John Hopkins University for helpful discussions that motivated the proof of Lemma 4.1.

References

[1] R. Abedi and R. B. Haber. Spacetime simulation of dynamic fracture with crack closure and frictional sliding. *Advanced Modeling and Simulation in Engineering Sciences*, 5(1):1–22, 2018.

[2] E. Burman, A. Ern, and M. A. Fernández. Explicit Runge–Kutta schemes and finite elements with symmetric stabilization for first-order linear PDE systems. *SIAM Journal on Numerical Analysis*, 48(6):2019–2042, 2010.

[3] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2016.

[4] D. Drake, J. Gopalakrishnan, J. Schöberl, and C. Wintersteiger. Convergence analysis of some tent-based schemes for linear hyperbolic systems. *Mathematics of Computation*, 91(334):699–733, 2022.

[5] R. S. Falk and G. R. Richter. Explicit finite element methods for symmetric hyperbolic equations. *SIAM Journal on Numerical Analysis*, 36(3):935–952, 1999.

[6] J. Gopalakrishnan, M. Hochsteger, J. Schöberl, and C. Wintersteiger. An explicit mapped tent pitching scheme for Maxwell equations. In S. J. Sherwin, D. Moxey, J. Peiró, P. E. Vincent, and C. Schwab, editors, *Spectral and High Order Methods for Partial Differential Equations: ICOSAHOM 2018*, volume 134 of *Lecture Notes in Computational Science and Engineering*, pages 359–369, 2020.

[7] J. Gopalakrishnan, J. Schöberl, and C. Wintersteiger. Mapped tent pitching schemes for hyperbolic systems. *SIAM Journal on Scientific Computing*, 39(6):B1043–B1063, 2017.

[8] J. Gopalakrishnan, J. Schöberl, and C. Wintersteiger. Structure aware Runge–Kutta time stepping for spacetime tents. *SN Partial Differential Equations and Applications*, 1(4):1–24, 2020.

[9] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43(1):89–112, 2001.

[10] A. Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 2009.

[11] D. I. Ketcheson. Relaxation Runge–Kutta methods: Conservation and stability for inner-product norms. *SIAM Journal on Numerical Analysis*, 57(6):2850–2870, 2019.

[12] D. Levy and E. Tadmor. From semidiscrete to fully discrete: Stability of Runge–Kutta schemes by the energy method. *SIAM Review*, 40(1):40–73, 1998.

[13] S. T. Miller and R. B. Haber. A spacetime discontinuous Galerkin method for hyperbolic heat conduction. *Computer Methods in Applied Mechanics and Engineering*, 198(2):194–209, 2008.

[14] P. Monk and G. R. Richter. A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media. *Journal of Scientific Computing*, 22(1):443–477, 2005.

[15] P. Öffner, J. Glaubitz, and H. Ranocha. Analysis of artificial dissipation of explicit and implicit time-integration methods. *International Journal of Numerical Analysis and Modeling*, 17(3):332–349, 2020.

[16] H. Ranocha. On strong stability of explicit Runge–Kutta methods for nonlinear semibounded operators. *IMA Journal of Numerical Analysis*, 41(1):654–682, 2021.

[17] H. Ranocha and D. I. Ketcheson. Energy stability of explicit Runge–Kutta methods for nonautonomous or nonlinear problems. *SIAM Journal on Numerical Analysis*, 58(6):3382–3405, 2020.

[18] H. Ranocha and P. Öffner. L_2 stability of explicit Runge–Kutta schemes. *Journal of Scientific Computing*, 75(2):1040–1056, 2018.

[19] H. Ranocha, M. Sayyari, L. Dalcin, M. Parsani, and D. I. Ketcheson. Relaxation Runge–Kutta methods: Fully discrete explicit entropy-stable schemes for the compressible Euler and Navier–Stokes equations. *SIAM Journal on Scientific Computing*, 42(2):A612–A638, 2020.

[20] Z. Sun and C.-W. Shu. Stability of the fourth order Runge–Kutta method for time-dependent partial differential equations. *Annals of Mathematical Sciences and Applications*, 2(2):255–284, 2017.

[21] Z. Sun and C.-W. Shu. Strong stability of explicit Runge–Kutta time discretizations. *SIAM Journal on Numerical Analysis*, 57(3):1158–1182, 2019.

[22] Z. Sun and C.-W. Shu. Enforcing strong stability of explicit Runge-Kutta methods with superviscosity. *Communications on Applied Mathematics and Computation*, pages 1–30, 2021.

[23] Z. Sun, Y. Wei, and K. Wu. On energy laws and stability of Runge–Kutta methods for linear seminegative problems. *SIAM Journal on Numerical Analysis*, 60(5):2448–2481, 2022.

[24] E. Tadmor. From semidiscrete to fully discrete: Stability of Runge–Kutta schemes by the energy method. II. *Collected Lectures on the Preservation of Stability under Discretization*, Lecture Notes from Colorado State University Conference, Fort Collins, CO, 2001 (D. Estep and S. Tavener, eds.), Proceedings in Applied Mathematics, SIAM, 109:25–49, 2002.

[25] Y. Xu, X. Meng, C.-W. Shu, and Q. Zhang. Superconvergence analysis of the Runge–Kutta discontinuous Galerkin methods for a linear hyperbolic equation. *Journal of Scientific Computing*, 84(1):1–40, 2020.

[26] Y. Xu, Q. Zhang, C.-W. Shu, and H. Wang. The L^2-norm stability analysis of Runge–Kutta discontinuous Galerkin methods for linear hyperbolic equations. *SIAM Journal on Numerical Analysis*, 57(4):1574–1601, 2019.

[27] Q. Zhang and C.-W. Shu. Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM Journal on Numerical Analysis*, 42(2):641–666, 2004.

Portland State University, PO Box 751, Portland, OR 97207, USA
Email address: gjay@pdx.edu

(Corresponding author.) Department of Mathematics, The University of Alabama, Box 870350, Tuscaloosa, AL 35487, USA
Email address: zsun30@ua.edu