TIME STABILITY OF STRONG BOUNDARY CONDITIONS IN FINITE-DIFFERENCE SCHEMES FOR HYPERBOLIC SYSTEMS*

NEK SHARAN†, PETER T. BRADY‡, AND DANIEL LIVESCU‡

Abstract. A framework to construct time-stable finite-difference schemes that apply boundary conditions strongly (or exactly) is presented for hyperbolic systems. A strong time-stability definition that applies to problems with homogeneous as well as non-homogeneous boundary data is introduced. Sufficient conditions for strong time stability and conservation are derived for the linear advection equation and coupled system of hyperbolic equations using the energy method. Explicit boundary stencils and norms that satisfy those sufficient conditions are derived for various order of accuracies. The discretization uses non-square derivative operators to allow stability and conservation conditions in terms of boundary data at grid points where physical boundary condition is directly injected and solution values at rest of the grid points. Various linear and non-linear numerical tests that verify the accuracy and stability of the derived stencils are presented.

Key words. time stability, conservation, boundary conditions

AMS subject classifications. 65M06, 65M12, 76M20

1. Introduction.

High-fidelity fluid dynamics simulations require stable boundary closures for long-time calculations typical of practical applications. High-order centered finite-difference schemes are commonly used for accurate turbulent flow [25, 20, 26, 30] and aeroacoustics [33, 39, 8, 11] simulations because of their non-dissipative properties, ease of implementation, and computational efficiency. However, the non-dissipative character of centered schemes also renders them susceptible to numerical instabilities when the boundary closure for a given interior scheme is not derived to satisfy stability conditions [5].

Numerical stability proofs require bounding the computational solution in terms of constants independent of grid spacing [15]. Various stability definitions exist that impose different solution bounds. The classical (Lax and G-K-S) stability definition allows non-physical solution growth in time even though the solution may converge on successive grid refinements [35, 6], which can be detrimental to long-time integrations in fluid-flow calculations. In this study, boundary stencils are, therefore, derived to satisfy the time stability (also called strict or energy stability) definition, which provides a uniform bound for the solution in time, preventing non-physical temporal growth.

Commonly used time-stable boundary treatments include the weak imposition of boundary conditions (BCs) with simultaneous-approximation-term (SAT) [6] as well as the projection method [21, 22]. The SAT approach imposes BCs using a penalty term, whereas the projection method uses a projection matrix to incorporate

*Submitted to the editors January 29, 2022.

Funding: This work was supported by the US Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20190227ER. Computational resources were provided by the LANL Institutional Computing (IC) Program and the National Science Foundation XSEDE resources under grant TG-PHY210037.

†Department of Aerospace Engineering, Auburn University, Auburn, AL 36849, USA (nsharan@auburn.edu),

‡CCS-2, Los Alamos National Laboratory, Los Alamos, NM 87545, USA (ptb@lanl.gov, livescu@lanl.gov).

BCs into the system of ordinary differential equations (ODEs) solved for the discrete solution. The extent to which the boundary point may satisfy the BC with SAT approach depends on the magnitude of the penalty parameter. A higher value may better satisfy the BC, however, it may make the ODE system stiffer. In cases of non-homogeneous boundary data, the projection method may also not satisfy the BC exactly because the projected ODE system imposes the time-derivative of boundary data, and the time-integration of the ODE system may not be exact. This work focuses on derivation of a time-stable method that enforces BCs strongly (or exactly).

Kreiss & Scherer [18] proposed a method to derive first-derivative finite-difference approximations with centered interior schemes and boundary stencils that satisfy a summation-by-parts (SBP) property of the differential equation. In general, the SBP property is not a sufficient condition for time stability with strong BCs [15, 6], but several SBP operators are time stable for scalar hyperbolic problems with homogeneous boundary data. However, as observed by Carpenter *et al.* [6], high-order schemes can lead to unphysical solution growth in time for coupled hyperbolic systems, when solved using strong BCs. In particular, for the 2×2 system discussed in [6, Section 3], and solved here in Section 4.2, Carpenter *et al.* noted at the time that no central difference scheme of order greater than two was time stable for this system. To the best of our knowledge, there are still no central finite-difference scheme of order greater than two that are time stable for this system with strong BCs. Carpenter *et al.* [6] proved time stability of SBP schemes for this system using SAT (weak) BC implementations. In this work, we derive boundary stencils for centered interior schemes up to sixth-order accurate that are time stable for this system with strong BCs.

Theoretical time-stability analyses of finite-difference schemes using weak BC implementations are widely available [12, 31]. However, similar analyses for strong BCs are hindered by the challenge of incorporating exact boundary conditions in the system of ODEs (following a method-of-lines approach) such that it also ensures a uniform solution bound (for systems with bounded energy), independent of grid spacing. An alternative approach that uses non-linear optimization to numerically examine the stability of boundary closures with strong BCs is proposed in [4]. Theoretical stability proofs provide sufficient conditions of stability, so in principle, it is possible that a numerical optimization may provide time-stable schemes that satisfy yet unknown necessary conditions of stability but not the sufficient conditions from theoretical proofs. However, at present this procedure has also not yielded time-stable schemes for the 2×2 system mentioned above.

In theoretical stability analyses, application of strong BCs is typically represented by a projection operator that omits rows in the derivative operator corresponding to grid points where the physical BCs are applied, e.g. [6, 18]. The row omissions prevent calculations at the boundary points where exact boundary data is injected. Row omissions in a derivative operator that was originally designed for calculations on the whole domain compromises the numerical properties of the full operator [15]. For example, a derivative operator that discretely satisfies the conservation condition for the scalar convection equation

$$\frac{\partial U}{\partial t} + \frac{\partial U}{\partial x} = 0, \qquad 0 \le x \le 1,\ t \ge 0, \tag{1.1}$$

given by

$$\frac{d}{dt}\int_0^1 U\,dx = -\int_0^1 \frac{\partial U}{\partial x}\,dx = U(0,t) - U(1,t), \tag{1.2}$$

is not conservative after row omission, as shown in Lemma A.1. To alleviate these issues, we consider non-square derivative operators that incorporate exact BCs to begin with and derive time-stability and conservation conditions for such operators. This is in contrast to the traditional approach where stability and conservation conditions are satisfied for square operators, which may not preserve those properties on row omission(s) for strong BC implementation.

The paper is organized as follows. Time-stability and conservation constraints for finite-difference schemes imposing strong BCs are derived in Section 2 for a hyperbolic scalar equation as well as coupled system of equations. For non-homogeneous boundary data, a definition of strong time-stability is introduced, in addition to the time-stability definition for homogeneous boundary data. Steps in the construction of boundary stencils to satisfy the time-stability and conservation constraints are discussed in Section 3. The stability and the accuracy of the derived schemes is evaluated for various linear and non-linear problems in Section 4. Application of the derived schemes to the Euler equations with characteristic boundary conditions is discussed in Section 5 and the conclusions are provided in Section 6.

2. Numerical approach and proof of stability.

This section derives the constraints on boundary stencils for time-stable enforcement of strong BCs to solve a hyperbolic scalar equation (Section 2.1) and hyperbolic system of equations (Section 2.2). The derived constraints are then used to obtain schemes of various order of accuracies in Section 3.

2.1. The hyperbolic scalar problem.

Consider the scalar hyperbolic equation (1.1) with the initial and the boundary condition given by

$$U(x,0) = f(x), \qquad U(0,t) = g(t). \tag{2.1}$$

On a domain with $n+1$ equidistant grid points $(x_0 = 0, x_1, \cdots, x_{n-1}, x_n = 1)$, a semi-discretization of (1.1)–(2.1) using strong boundary conditions can be written as

$$\begin{aligned} \frac{d\tilde{\mathbf{u}}}{dt} &= -D\mathbf{u}, \\ \mathbf{u}(0) &= \mathbf{f}, \end{aligned} \tag{2.2}$$

where $\mathbf{u}(t) = \begin{bmatrix} u_0(t) & \cdots & u_n(t) \end{bmatrix}^T$, with $u_0(t) \equiv g(t)$, is the discrete solution vector. $\tilde{\mathbf{u}}(t) = \begin{bmatrix} u_1(t) & \cdots & u_n(t) \end{bmatrix}^T$ is the solution vector without the first element, which corresponds to the grid point where the boundary data is injected. D, a matrix of size $n\times(n+1)$, denotes the derivative operator. The entries of D are denoted by d_{ij}, where $1 \le i \le n$ and $0 \le j \le n$. Its non-square structure prevents computation at the first grid point, where physical boundary condition is applied. $\mathbf{f} = \begin{bmatrix} f(x_0) & \cdots & f(x_n) \end{bmatrix}^T$ denotes the discrete initial data.

Define a scalar product and norm for discrete real-valued vector functions $\mathbf{v} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^T$ and $\mathbf{w} = \begin{bmatrix} w_1 & \cdots & w_n \end{bmatrix}^T$ by (*e.g.* [18])

$$(\mathbf{v},\mathbf{w})_H = \mathbf{v}^T H \mathbf{w} = \sum_{i,j=1}^{\kappa} h_{ij} v_i w_j \Delta x + \sum_{i=\kappa+1}^{n-\kappa} v_i w_i \Delta x + \sum_{i,j=n-\kappa+1}^{n} h_{ij} v_i w_j \Delta x, \tag{2.3}$$

$$\|\mathbf{v}\|_H = \sqrt{(\mathbf{v}, \mathbf{v})_H}, \tag{2.4}$$

where Δx denotes the grid spacing, κ represents the depth of boundary stencil, and $h_{i,j}$ are the coefficients of a symmetric positive-definite (norm) matrix H.

Multiplying (2.2) by $\tilde{\mathbf{u}}^T H$, where H is a norm matrix of size $n \times n$, and adding its transpose yields

$$\frac{d}{dt} \|\tilde{\mathbf{u}}\|_H^2 = -\tilde{\mathbf{u}}^T H D \mathbf{u} - (D\mathbf{u})^T H \tilde{\mathbf{u}}. \tag{2.5}$$

Using Definition 2.13 of [15], time stability is defined as:

Definition 1. *The approximation (2.2) is time stable if for $g = 0$, there is a unique solution $\tilde{\mathbf{u}}(t)$ satisfying*

$$\|\tilde{\mathbf{u}}\|_H \leq K \left\|\tilde{\mathbf{f}}\right\|_H, \qquad or \qquad \frac{d}{dt} \|\tilde{\mathbf{u}}\|_H^2 \leq 0, \tag{2.6}$$

where K is independent of Δx, $\mathbf{f}$ and t. $\tilde{\mathbf{f}}$ denotes the vector $\mathbf{f}$ without its first element, following the definition of $\tilde{\mathbf{u}}$.

For $g = 0$, the first element of vector $\mathbf{u}$ is zero, *i.e.* $u_0 = 0$. Substituting $u_0 = 0$ in (2.5) yields

$$\frac{d}{dt} \|\tilde{\mathbf{u}}\|_H^2 = -\tilde{\mathbf{u}}^T H \tilde{D} \tilde{\mathbf{u}} - \left(\tilde{D}\tilde{\mathbf{u}}\right)^T H \tilde{\mathbf{u}} = \tilde{\mathbf{u}}^T \left[HM + (HM)^T\right] \tilde{\mathbf{u}}, \tag{2.7}$$

where $M = -\tilde{D}$ and $\tilde{D}$ is a square $(n \times n)$ matrix containing all columns of D except the first. If the approximation (2.2) is time stable, *i.e.* (2.6) is true, then the following result about the eigenvalues of M can be stated.

Theorem 1. *If there exists a positive definite matrix H such that $HM + (HM)^T$ is negative definite (semi-definite), then the real part of all eigenvalues of M are negative (non-positive).*

Proof. See [10, Lemma 3.1.1]. $\square$

(2.6) defines time stability for homogeneous boundary data, *i.e.* $g = 0$. For $g \neq 0$, following the Definition 2.12 of [15] for strong stability, we define strong time stability as:

Definition 2. *The approximation (2.2) is strongly time stable if there is a unique solution $\tilde{\mathbf{u}}(t)$ satisfying*

$$\|\tilde{\mathbf{u}}\|_H^2 \leq K \left(\left\|\tilde{\mathbf{f}}\right\|_H^2 + \int_0^t |g(\tau)|^2 \, d\tau \right), \qquad or \qquad \frac{d}{dt} \|\tilde{\mathbf{u}}\|_H^2 \leq K |g|^2, \tag{2.8}$$

where K is independent of Δx, $\mathbf{f}$, g and t.

Remark. The time-stability definition (2.6) differs from the classical stability definition [15, Definition 2.11] in requiring a uniform solution bound, independent of time [35, 6]. The energy estimates derived for the SBP operators in [18] ensure classical stability (see [18, Theorem 1.1]), but may not ensure time stability [6, 15]. The diagonal- and restricted full-norm SBP first-derivative operators of [32] on omitting their first row for strong BC implementation with semi-discretization (2.2) satisfy (2.6), for homogeneous boundary data, but do not guarantee (2.8) for non-zero boundary data. Moreover, row omission introduces an $\mathcal{O}(1)$ conservation error, as shown in Lemma A.1.

In the following, we derive the constraints on the entries of the derivative operator, D, for the solution of (2.2) to satisfy the strong time-stability definition (2.8) and a discrete conservation condition. To simplify algebra, the non-square operator $Q = HD$ can be decomposed such that

$$\tilde{\mathbf{u}}^T HD\mathbf{u} = \tilde{\mathbf{u}}^T Q\mathbf{u} = \tilde{\mathbf{u}}^T \tilde{Q}\tilde{\mathbf{u}} + \tilde{\mathbf{u}}^T \mathbf{q}_0 g, \tag{2.9}$$

where $\tilde{Q}$ is a square $(n \times n)$ matrix containing all the columns of Q except the first and vector $\mathbf{q}_0$ is the first column of Q. $u_0(t) \equiv g(t)$ is substituted in the second term of the r.h.s. of (2.9). The entries of Q, like D, are denoted by q_{ij}, where $1 \leq i \leq n$ and $0 \leq j \leq n$. Substituting (2.9) in the r.h.s. of (2.5) provides the strong time-stability condition that respects (2.8):

$$-\tilde{\mathbf{u}}^T HD\mathbf{u} - (D\mathbf{u})^T H\tilde{\mathbf{u}} = -\tilde{\mathbf{u}}^T \left(\tilde{Q} + \tilde{Q}^T\right)\tilde{\mathbf{u}} - 2\tilde{\mathbf{u}}^T \mathbf{q}_0 g \leq K\left|g\right|^2. \tag{2.10}$$

In addition to the above time-stability condition, we seek a discrete conservation condition. A discrete version of (1.2) is given by

$$\frac{d}{dt}\int_0^1 U\,dx \approx \frac{d}{dt}\sum_{i=1}^{n}\left(H\tilde{\mathbf{u}}\right)_i = -\sum_{i=1}^{n}\left(HD\mathbf{u}\right)_i = g(t) - u_n(t), \tag{2.11}$$

where the notation $(\mathbf{v})_i$ denotes the i-th component of a vector $\mathbf{v} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^T$ and the entries of H constitute a quadrature for the domain $0 \leq x \leq 1$.

In terms of the operators defined in (2.9), condition (2.11) translates to

$$\sum_{i=1}^{n}\left(\mathbf{q}_0\right)_i = -1, \qquad \sum_{i=1}^{n} q_{ij} = \begin{cases} 1 & j = n \\ 0 & \text{otherwise} \end{cases}, \tag{2.12}$$

where $(\mathbf{q}_0)_i \equiv q_{i0}$.

We seek derivative approximations, D, and norm matrices, H, that satisfy the strong time-stability condition (2.10) and the discrete conservation condition (2.11) for various order of accuracies. The derivation proceeds by assuming an extent of non-zero elements in vector $\mathbf{q}_0$, denoted by β, *i.e.*, let

$$\mathbf{q}_0 = \begin{bmatrix} q_{10} & \cdots & q_{\beta 0} & 0 & \cdots & 0 \end{bmatrix}^T. \tag{2.13}$$

In other words, $\beta > 0$ represents the depth of boundary stencils that use the physical boundary point, where the boundary data is injected, for derivative approximation. A non-zero (row) entry in $\mathbf{q}_0$ requires a corresponding non-zero diagonal entry in $\tilde{Q}$ to satisfy (2.10), as shown in the following theorem.

THEOREM 2. (a) *The strong time-stability condition (2.10) is satisfied if, for $1 \leq i, j \leq n$ and $\beta > 0$,*

$$q_{ij} \begin{cases} = -q_{ji} & \textit{if} \quad i \neq j, \\ > 0 & \textit{if} \quad i = j \leq \beta, \\ \geq 0 & \textit{if} \quad i = j > \beta. \end{cases} \tag{2.14}$$

(b) *The conservation condition (2.12) is concurrently satisfied if the latter two conditions in (2.14), for the diagonal entries of $\tilde{Q}$, are replaced by the stricter conditions, given by*

$$q_{ij} = \begin{cases} -q_{ji} & \text{if} \quad i \neq j, \\ -\frac{1}{2}q_{i0} > 0 & \text{if} \quad i = j \leq \beta, \\ \;\; 0 & \text{if} \quad \beta < i = j < n, \\ \;\; \frac{1}{2} & \text{if} \quad i = j = n, \end{cases} \tag{2.15}$$

and $\sum\limits_{i=1}^{n} q_{i0} = \sum\limits_{i=1}^{\beta} q_{i0} = -1.$

Proof. Matrix $\tilde{Q}$ with entries satisfying $q_{ij} = -q_{ji}$ for $i \neq j$ yields

$$\frac{\tilde{Q} + \tilde{Q}^T}{2} = \operatorname{diag}\left(q_{11}, \cdots, q_{\beta\beta}, \cdots, q_{nn}\right), \tag{2.16}$$

whose substitution in (2.10), with $\mathbf{q}_0 = \begin{bmatrix} q_{10} & \cdots & q_{\beta 0} & 0 & \cdots & 0 \end{bmatrix}^T$, provides

$$\begin{aligned} -\tilde{\mathbf{u}}^T\left(\tilde{Q} + \tilde{Q}^T\right)\tilde{\mathbf{u}} - 2\tilde{\mathbf{u}}^T\mathbf{q}_0 g &= -\sum_{i=1}^{n} 2q_{ii}u_i^2 - \sum_{i=1}^{\beta} 2q_{i0}u_i g \\ &= \sum_{i=1}^{\beta}\left[-2q_{ii}\left(u_i + \frac{q_{i0}}{2q_{ii}}g\right)^2 + \frac{q_{i0}^2}{2q_{ii}}g^2\right] - \sum_{i=\beta+1}^{n} 2q_{ii}u_i^2 \leq K_1 g^2, \end{aligned} \tag{2.17}$$

where the last inequality holds if $q_{ii} > 0$ for $1 \leq i \leq \beta$ and $q_{ii} \geq 0$ for $\beta < i \leq n$ (the conditions in (2.14)), and $K_1 = \sum\limits_{i=1}^{\beta} \frac{q_{i0}^2}{2q_{ii}}$. This proves the (a) part of the theorem.

For the (b) part of the theorem, note first that conditions in (2.15) satisfy (2.14), which ensures strong time stability. This can also be seen by substituting (2.15) in (2.17), and using $\sum\limits_{i=1}^{\beta} q_{i0} = -1$. It remains to be shown that (2.15) also satisfies the conservation condition (2.12).

The rows of a derivative approximation, D, sum to zero and, hence, the rows of $Q = HD$ also sum to zero (for proof, see Lemma A.2 in Appendix A), *i.e.*,

$$\sum_{j=0}^{n} q_{ij} = q_{i0} + q_{ii} + \sum_{\substack{j=1 \\ j\neq i}}^{n} q_{ij} = 0 \qquad \forall\ 1 \leq i \leq n, \tag{2.18}$$

where, from (2.13), $q_{i0} = 0$ for $i > \beta$. Using $q_{ij} = -q_{ji}$ for $i \neq j$ (this structure is typical of centered finite-difference scheme in the interior) yields

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} q_{ij} = -\sum_{\substack{j=1 \\ j\neq i}}^{n} q_{ji} \qquad \forall\ 1 \leq i \leq n. \tag{2.19}$$

Adding $-q_{ii}$ to both sides of (2.19) and using (2.18) provides

$$-\sum_{j=1}^{n} q_{ji} = \sum_{j=1}^{n} q_{ij} - 2q_{ii} = -q_{i0} - 2q_{ii} \qquad \forall\ 1 \leq i \leq n. \tag{2.20}$$

(2.12) is then satisfied if $q_{ii} = -\frac{1}{2}q_{i0}$ for $1 \leq i < n$ and $q_{ii} = \frac{1}{2} - \frac{1}{2}q_{i0}$ for $i = n$. From (2.13), $q_{i0} = 0$ for $i > \beta$, which yields $q_{ii} = \frac{1}{2}$ for $i = n$ and $q_{ii} = -\frac{1}{2}q_{i0} = 0$ for $\beta < i < n$ (the conditions in (2.15)). This completes the proof. □

To summarize the above theorem, a skew-symmetric $\tilde{Q}$ except at the top-left and the bottom-right corner satisfies the conservation condition (2.12) at the interior points and it leads to cancellations of interior point terms in (2.10) for time stability. The skew-symmetric structure prescribes centered derivative approximations in the interior. The top-left and the bottom-right corner of Q (that comprises $\mathbf{q}_0$ and $\tilde{Q}$, see (2.9)) determine behavior at the inflow and the outflow boundary, respectively. The conditions in (2.15) for the outflow boundary, where no physical boundary condition is required, satisfy the summation-by-parts (SBP) formula [18] and, hence, SBP stencils are used at the outflow boundary in the proposed scheme. At the inflow boundary, new stencils that satisfy (2.14) are derived in Section 3 for various centered interior schemes.

2.2. The coupled hyperbolic system. This section discusses the time-stability conditions for the semi-discretization of a one-dimensional hyperbolic system using strong boundary conditions. A hyperbolic system coupled at the boundaries, considered by Carpenter *et al.* [6] and by Abarbanel & Chertock [1] to prove time stability of finite-difference schemes with SAT (weak) BCs, is considered here with strong BCs.

The system, with domain $0 \leq x \leq 1$ and $t \geq 0$, is given by

$$\frac{\partial \mathbf{U}^I}{\partial t} + \Lambda^I \frac{\partial \mathbf{U}^I}{\partial x} = 0, \tag{2.21}$$

$$\frac{\partial \mathbf{U}^{II}}{\partial t} + \Lambda^{II} \frac{\partial \mathbf{U}^{II}}{\partial x} = 0, \tag{2.22}$$

where

$$\mathbf{U}^I = \begin{bmatrix} U^1(x,t) & \cdots & U^k(x,t) \end{bmatrix}^T \qquad \text{and} \qquad \Lambda^I = \text{diag}\,(\lambda_1, \cdots, \lambda_k)$$

for $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$ describe a system of right-moving waves and

$$\mathbf{U}^{II} = \begin{bmatrix} U^{k+1}(x,t) & \cdots & U^r(x,t) \end{bmatrix}^T \qquad \text{and} \qquad \Lambda^{II} = \text{diag}\,(\lambda_{k+1}, \cdots, \lambda_r)$$

for $0 > \lambda_{k+1} > \lambda_{k+2} > \cdots > \lambda_r$ describe a system of left-moving waves. The system (2.21)-(2.22) is well-posed for boundary conditions given by

$$\mathbf{U}^I(0,t) = L\mathbf{U}^{II}(0,t) + \mathbf{g}^I(t), \tag{2.23}$$

$$\mathbf{U}^{II}(1,t) = R\mathbf{U}^I(1,t) + \mathbf{g}^{II}(t), \tag{2.24}$$

where L and R are constant matrices of size $k \times (r-k)$ and $(r-k) \times k$, respectively, and $\mathbf{g}^I$ and $\mathbf{g}^{II}$ are vectors of size k and $r-k$, respectively. The system (2.21)-(2.24) has a non-growing solution in time if $\mathbf{g}^I$ and $\mathbf{g}^{II}$ are zero and (see [6, Theorem 2.1])

$$\|L\|\,\|R\| \leq 1. \tag{2.25}$$

The matrix norm for real matrices is defined by $\|L\|^2 = \rho\left(L^T L\right)$, where $\rho(\cdot)$ denotes the spectral radius. For the system (2.21)-(2.22) to be coupled, the norms $\|L\|$ and $\|R\|$ should be non-zero.

A semi-discretization of (2.21)-(2.24) using strong boundary conditions can be written as

$$\frac{d\mathbf{w}}{dt} = -\mathcal{D}\mathbf{w} + \mathbf{b}, \tag{2.26}$$

where $\mathbf{w}(t) = \begin{bmatrix} \tilde{\mathbf{u}}^I(t) & \tilde{\mathbf{u}}^{II}(t) \end{bmatrix}^T$ with $\tilde{\mathbf{u}}^I(t) = \begin{bmatrix}\tilde{\mathbf{u}}^1(t) & \cdots & \tilde{\mathbf{u}}^k(t)\end{bmatrix}$ and $\tilde{\mathbf{u}}^{II}(t) = \begin{bmatrix}\tilde{\mathbf{u}}^{k+1}(t) & \cdots & \tilde{\mathbf{u}}^r(t)\end{bmatrix}$. The unknowns for each equation in the system are given by (assuming a discretization with $n+1$ grid points, as described in Section 2.1) $\tilde{\mathbf{u}}^\phi(t) = \begin{bmatrix}u_1^\phi(t) & \cdots & u_n^\phi(t)\end{bmatrix}^T$ for $1 \le \phi \le k$ and $\tilde{\mathbf{u}}^\phi(t) = \begin{bmatrix}u_0^\phi(t) & \cdots & u_{n-1}^\phi(t)\end{bmatrix}^T$ for $k+1 \le \phi \le r$, where $\tilde{\mathbf{u}}^\phi(t)$ is the solution vector without the element corresponding to the grid point where the boundary data is injected. Therefore, the solution vectors for the first k equations do not contain the element corresponding to the first grid point and the rest do not contain the element corresponding to the last grid point. The derivative operator, $\mathcal{D}$, is then given by

$$\mathcal{D} = \Lambda \mathcal{H}^{-1}\mathcal{Q}, \tag{2.27}$$

where $\Lambda = \text{diag}\,(\lambda_1, \cdots, \lambda_r)$,

$$\mathcal{H} = \begin{bmatrix}\mathcal{H}_{11} & 0 \\ 0 & \mathcal{H}_{22}\end{bmatrix}, \qquad \text{and} \qquad \mathcal{Q} = \begin{bmatrix}\mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{21} & \mathcal{Q}_{22}\end{bmatrix}.$$

The submatrices are given by

$$\begin{aligned} &\mathcal{H}_{11} = I_k \otimes H, \qquad \mathcal{H}_{22} = I_{r-k} \otimes H^\#, \\ &\mathcal{Q}_{11} = I_k \otimes \tilde{Q}, \qquad \mathcal{Q}_{12} = L \otimes Q_0, \qquad \mathcal{Q}_{21} = -R \otimes Q_0^\#, \qquad \mathcal{Q}_{22} = -I_{r-k} \otimes \tilde{Q}^\#, \end{aligned} \tag{2.28}$$

where I_m denotes an identity matrix of size $m \times m$ and $\otimes$ denotes the Kronecker product. The superscript $^\#$ denotes the matrix and vector transformations $M^\# = \mathcal{J}^{-1}M\mathcal{J}$ and $\mathbf{m}^\# = \mathcal{J}^{-1}\mathbf{m}$, respectively, where

$$\mathcal{J} = \mathcal{J}^{-1} = \begin{bmatrix}0 & & 1 \\ & \iddots & \\ 1 & & 0\end{bmatrix}. \tag{2.29}$$

The transformation yields matrix/vector "rotated" by 180^o, for example,

$$\begin{bmatrix}a & b \\ c & d\end{bmatrix}^\# = \begin{bmatrix}d & c \\ b & a\end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix}a \\ b\end{bmatrix}^\# = \begin{bmatrix}b \\ a\end{bmatrix}. \tag{2.30}$$

Q_0 is a $n \times n$ matrix with $\mathbf{q}_0$ as the first column and remaining columns zero. The vector $\mathbf{q}_0$ and matrices H and $\tilde{Q}$ are as described in Section 2.1. Vector $\mathbf{b}$ incorporates the boundary data $\mathbf{g}^I$ and $\mathbf{g}^{II}$, and is given by

$$\mathbf{b} = \Lambda \mathcal{H}^{-1} \begin{bmatrix}\mathbf{g}^I \otimes \mathbf{q}_0 \\ -\mathbf{g}^{II} \otimes \mathbf{q}_0^\#\end{bmatrix}. \tag{2.31}$$

Let the discrete energy be defined as (*e.g.*, [6, 1])

$$E(t) = \sum_{\phi=1}^{k} \frac{\|R\|}{\lambda_\phi} \left(\tilde{\mathbf{u}}^\phi\right)^T H \tilde{\mathbf{u}}^\phi + \sum_{\phi=k+1}^{r} \frac{\|L\|}{|\lambda_\phi|} \left(\tilde{\mathbf{u}}^\phi\right)^T H^\# \tilde{\mathbf{u}}^\phi, \tag{2.32}$$

which provides

$$\frac{dE}{dt} = \sum_{\phi=1}^{k} \frac{\|R\|}{\lambda_\phi} \frac{d}{dt} \left(\tilde{\mathbf{u}}^\phi\right)^T H \tilde{\mathbf{u}}^\phi + \sum_{\phi=k+1}^{r} \frac{\|L\|}{|\lambda_\phi|} \frac{d}{dt} \left(\tilde{\mathbf{u}}^\phi\right)^T H^\# \tilde{\mathbf{u}}^\phi. \tag{2.33}$$

The time-stability condition, assuming $\mathbf{g}^I = 0$ and $\mathbf{g}^{II} = 0$ in (2.23)-(2.24), is defined as

$$\frac{dE}{dt} \leq 0, \tag{2.34}$$

and the strong time-stability condition for non-zero $\mathbf{g}^I$ and $\mathbf{g}^{II}$ is defined as

$$\frac{dE}{dt} \leq K_I \left\|\mathbf{g}^I\right\|^2 + K_{II} \left\|\mathbf{g}^{II}\right\|^2, \tag{2.35}$$

where $\|\mathbf{v}\| = \sqrt{\mathbf{v}^T\mathbf{v}}$ for a vector $\mathbf{v}$. The conservation condition for the system (2.21)-(2.22) is the same as that for the scalar equation (1.1), since the system comprises of scalar advection equations. The conservation condition for the operators used in the semi-discretization (2.26) is, therefore, given by (2.12). The numerical flux should "telescope" across the domain to the boundaries without loss, consistent with the continuous flux behavior.

The following theorem provides sufficient conditions for the semi-discretization (2.26) to satisfy the strong time-stability and conservation conditions.

THEOREM 3. (a) *The strong time-stability condition (2.35) is satisfied if, for* $1 \leq i, j \leq n$ *and* $\beta > 0$*,*

$$q_{ij} \begin{cases} = -q_{ji} & \text{if} \quad i \neq j, \\ > \frac{q_{i0}^2}{4 q_{nn} a_i} \|L\| \, \|R\| & \text{if} \quad i = j \leq \beta, \\ \geq 0 & \text{if} \quad \beta < i = j < n, \\ > 0 & \text{if} \quad i = j = n, \end{cases} \tag{2.36}$$

where $a_i > 0$ *and* $\sum_{i=1}^{\beta} a_i = 1$.

(b) *The conservation condition (2.12) is concurrently satisfied if (2.15) is true with* $\sum_{i=1}^{n} q_{i0} = \sum_{i=1}^{\beta} q_{i0} = -1$.

Proof. The individual terms in summations of (2.33), that denote the contribution from each equation of the system, are given by

$$\begin{aligned} \frac{d}{dt} \left(\tilde{\mathbf{u}}^\phi\right)^T H \tilde{\mathbf{u}}^\phi = \frac{d}{dt} \left\|\tilde{\mathbf{u}}^\phi\right\|_H^2 = {} & -\lambda_\phi \left(\tilde{\mathbf{u}}^\phi\right)^T \left(\tilde{Q} + \tilde{Q}^T\right) \tilde{\mathbf{u}}^\phi \\ & -2\lambda_\phi \left(\tilde{\mathbf{u}}^\phi\right)^T \mathbf{q}_0 \left(L \tilde{\mathbf{u}}_0^{II} + \mathbf{g}^I\right)_\phi, \end{aligned} \tag{2.37}$$

for $1 \leq \phi \leq k$, and by

$$\begin{aligned} \frac{d}{dt} \left(\tilde{\mathbf{u}}^\phi\right)^T H^\# \tilde{\mathbf{u}}^\phi = \frac{d}{dt} \left\|\tilde{\mathbf{u}}^\phi\right\|_{H^\#}^2 = {} & -\lambda_\phi \left(\tilde{\mathbf{u}}^\phi\right)^T \left(\tilde{Q}^\# + \left(\tilde{Q}^\#\right)^T\right) \tilde{\mathbf{u}}^\phi \\ & -2\lambda_\phi \left(\tilde{\mathbf{u}}^\phi\right)^T \mathbf{q}_0^\# \left(R \tilde{\mathbf{u}}_n^{I} + \mathbf{g}^{II}\right)_\phi, \end{aligned} \tag{2.38}$$

for $k+1 \leq \phi \leq r$, where $\tilde{\mathbf{u}}_0^{II} = \left[u_0^{k+1}(t) \ \cdots \ u_0^r(t)\right]^T$ and $\tilde{\mathbf{u}}_n^I = \left[u_n^1(t) \ \cdots \ u_n^k(t)\right]^T$. Assuming $q_{ij} = -q_{ji}$, for $i \neq j$ in matrix $\tilde{Q}$, the contribution to (2.33) from the first term in the r.h.s. of (2.37) and (2.38) can be calculated from, respectively,

$$\sum_{\phi=1}^{k} \left(\tilde{\mathbf{u}}^\phi\right)^T \left(\tilde{Q} + \tilde{Q}^T\right) \tilde{\mathbf{u}}^\phi = 2\sum_{i=1}^{n} q_{ii} \sum_{\phi=1}^{k} \left(u_i^\phi\right)^2 = 2\sum_{i=1}^{n} q_{ii} \left\|\tilde{\mathbf{u}}_i^I\right\|^2, \tag{2.39}$$

$$\begin{aligned} \sum_{\phi=k+1}^{r} \left(\tilde{\mathbf{u}}^\phi\right)^T \left(\tilde{Q}^\# + \left(\tilde{Q}^\#\right)^T\right) \tilde{\mathbf{u}}^\phi & \\ = -2\sum_{i=1}^{n} q_{ii} \sum_{\phi=k+1}^{r} \left(u_{n-i}^\phi\right)^2 = -2\sum_{i=1}^{n} q_{ii} \left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\|^2, \end{aligned} \tag{2.40}$$

where $\left\|\tilde{\mathbf{u}}_i^I\right\|^2 = \sum_{\phi=1}^{k} \left(u_i^\phi\right)^2$ and $\left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\|^2 = \sum_{\phi=k+1}^{r} \left(u_{n-i}^\phi\right)^2$. Further, assuming $\mathbf{q}_0 = \left[q_{10} \quad \cdots \quad q_{\beta 0} \quad 0 \quad \cdots \quad 0\right]^T$, as in (2.13), the contribution to (2.33) from the second term in the r.h.s. of (2.37) and (2.38) are, respectively,

$$\sum_{\phi=1}^{k} \left(\tilde{\mathbf{u}}^\phi\right)^T \mathbf{q}_0 \left(L\tilde{\mathbf{u}}_0^{II} + \mathbf{g}^I\right)_\phi = \sum_{i=1}^{\beta} q_{i0} \sum_{\phi=1}^{k} u_i^\phi \left(L\tilde{\mathbf{u}}_0^{II} + \mathbf{g}^I\right)_\phi, \tag{2.41}$$

$$\sum_{\phi=k+1}^{r} \left(\tilde{\mathbf{u}}^\phi\right)^T \mathbf{q}_0^\# \left(R\tilde{\mathbf{u}}_n^I + \mathbf{g}^{II}\right)_\phi = -\sum_{i=1}^{\beta} q_{i0} \sum_{\phi=k+1}^{r} u_{n-i}^\phi \left(R\tilde{\mathbf{u}}_n^I + \mathbf{g}^{II}\right)_\phi. \tag{2.42}$$

Using the Cauchy-Schwarz inequality,

$$\sum_{\phi=1}^{k} u_i^\phi \left(L\tilde{\mathbf{u}}_0^{II}\right)_\phi \leq \left\|\tilde{\mathbf{u}}_i^I\right\| \left\|L\right\| \left\|\tilde{\mathbf{u}}_0^{II}\right\|, \qquad \sum_{\phi=1}^{k} u_i^\phi \left(\mathbf{g}^I\right)_\phi \leq \left\|\tilde{\mathbf{u}}_i^I\right\| \left\|\mathbf{g}^I\right\|, \tag{2.43}$$

and

$$\sum_{\phi=k+1}^{r} u_{n-i}^\phi \left(R\tilde{\mathbf{u}}_n^I\right)_\phi \leq \left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\| \left\|R\right\| \left\|\tilde{\mathbf{u}}_n^I\right\|, \qquad \sum_{\phi=k+1}^{r} u_{n-i}^\phi \left(\mathbf{g}^{II}\right)_\phi \leq \left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\| \left\|\mathbf{g}^{II}\right\|. \tag{2.44}$$

Substituting (2.43) and (2.44) in (2.41) and (2.42), respectively, and, in turn, using (2.37)-(2.38) with (2.39)-(2.42) in (2.33), assuming $q_{ii} \geq 0$ for $\beta < i < n$, yields

$$\begin{aligned} \frac{dE}{dt} \leq & \left\{\sum_{i=1}^{\beta} \left(-2q_{ii} \left\|R\right\| \left\|\tilde{\mathbf{u}}_i^I\right\|^2 + 2\left|q_{i0}\right| \left\|L\right\| \left\|R\right\| \left\|\tilde{\mathbf{u}}_i^I\right\| \left\|\tilde{\mathbf{u}}_0^{II}\right\|\right) - 2q_{nn} \left\|L\right\| \left\|\tilde{\mathbf{u}}_0^{II}\right\|^2\right\} \\ & + \left\{\sum_{i=1}^{\beta} \left(-2q_{ii} \left\|L\right\| \left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\|^2 + 2\left|q_{i0}\right| \left\|L\right\| \left\|R\right\| \left\|\tilde{\mathbf{u}}_n^I\right\| \left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\|\right) - 2q_{nn} \left\|R\right\| \left\|\tilde{\mathbf{u}}_n^I\right\|^2\right\} \\ & + \sum_{i=1}^{\beta} \left(2\left|q_{i0}\right| \left\|R\right\| \left\|\tilde{\mathbf{u}}_i^I\right\| \left\|\mathbf{g}^I\right\| + 2\left|q_{i0}\right| \left\|L\right\| \left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\| \left\|\mathbf{g}^{II}\right\|\right). \end{aligned} \tag{2.45}$$

The time-stability condition (2.34), where $\mathbf{g}^I = 0$ and $\mathbf{g}^{II} = 0$ is assumed, is satisfied if both curly brackets in (2.45) are non-positive. Introducing $\sum_{i=1}^{\beta} a_i = 1$, where $a_i > 0$, the last terms in the curly brackets can be written as

$$2q_{nn}\,\|L\|\,\left\|\tilde{\mathbf{u}}_0^{II}\right\|^2 = 2\sum_{i=1}^{\beta} a_i q_{nn}\,\|L\|\,\left\|\tilde{\mathbf{u}}_0^{II}\right\|^2, \tag{2.46}$$

$$2q_{nn}\,\|R\|\,\left\|\tilde{\mathbf{u}}_n^{I}\right\|^2 = 2\sum_{i=1}^{\beta} a_i q_{nn}\,\|R\|\,\left\|\tilde{\mathbf{u}}_n^{I}\right\|^2. \tag{2.47}$$

Substituting (2.46)-(2.47) in (2.45), the two curly brackets in (2.45) are non-positive if

$$q_{ii} \geq \frac{q_{i0}^2}{4q_{nn}a_i}\,\|L\|\,\|R\| \qquad \text{or} \qquad q_{ii} = s + \frac{q_{i0}^2}{4q_{nn}a_i}\,\|L\|\,\|R\|\,, \qquad 1 \leq i \leq \beta, \tag{2.48}$$

where $s \geq 0$. Substituting q_{ii} from (2.48) in (2.45) ensures that the terms in the curly brackets are non-positive and yields for $s > 0$,

$$\begin{aligned}
\frac{dE}{dt} \leq& \sum_{i=1}^{\beta} \Big(-2s\,\|R\|\,\left\|\tilde{\mathbf{u}}_i^{I}\right\|^2 + 2\,|q_{i0}|\,\|R\|\,\left\|\tilde{\mathbf{u}}_i^{I}\right\|\,\left\|\mathbf{g}^{I}\right\| - 2s\,\|L\|\,\left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\|^2 \\
&\qquad\qquad + 2\,|q_{i0}|\,\|L\|\,\left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\|\,\left\|\mathbf{g}^{II}\right\|\Big) \\
=& \sum_{i=1}^{\beta} \Bigg(-\|R\|\left[\sqrt{2s}\,\left\|\tilde{\mathbf{u}}_i^{I}\right\| - \frac{|q_{i0}|}{\sqrt{2s}}\left\|\mathbf{g}^{I}\right\|\right]^2 - \|L\|\left[\sqrt{2s}\,\left\|\tilde{\mathbf{u}}_{n-i}^{II}\right\| - \frac{|q_{i0}|}{\sqrt{2s}}\left\|\mathbf{g}^{II}\right\|\right]^2 \\
&\qquad\qquad + \frac{|q_{i0}|^2}{2s}\left\{\|R\|\left\|\mathbf{g}^{I}\right\|^2 + \|L\|\left\|\mathbf{g}^{II}\right\|^2\right\}\Bigg) \\
\leq& \frac{\sum_{i=1}^{\beta} |q_{i0}|^2}{2s}\left(\|R\|\left\|\mathbf{g}^{I}\right\|^2 + \|L\|\left\|\mathbf{g}^{II}\right\|^2\right).
\end{aligned} \tag{2.49}$$

Thus, $s > 0$ in (2.48) ensures both strong time stability, defined by (2.35), and time stability, defined by (2.34), while $s = 0$ ensures time stability but not strong time stability. This proves the (a) part of the theorem.

Theorem 2(b) shows that (2.15) with $\sum_{i=1}^{n} q_{i0} = -1$, where $q_{i0} \leq 0$, satisfies the discrete conservation condition (2.12) for the scalar advection equation. As already mentioned, the discrete conservation condition for the system (2.21)-(2.22) is the same as that for the scalar advection equation. Therefore, a stencil satisfying (2.15) provides a conservative scheme for the system (2.21)-(2.22). It remains to be shown that (2.15) also satisfies the strong time-stability condition (2.35).

Using $a_i = -q_{i0}$ and $q_{nn} = \frac{1}{2}$ in (2.36), (2.15) automatically satisfies (2.36) since

$$-\frac{1}{2}q_{i0} = \frac{q_{i0}^2}{4q_{nn}a_i} > \frac{q_{i0}^2}{4q_{nn}a_i}\,\|L\|\,\|R\| \tag{2.50}$$

for $\|L\|\,\|R\| < 1$. This completes the proof. $\square$

Remark. The energy estimate (2.49) obtained in terms of the matrix norms $\|L\|$ and $\|R\|$ is an artifact of the energy definition (2.32) used for the proof. This definition simplifies the proof of stability, and from the equivalence of norms over a finite-dimensional vector space, it can be shown that the energy defined simply by the square of the Euclidean norm, $\widetilde{E}(t) = \sum_{\phi=1}^{r} \left(\tilde{\mathbf{u}}^{\phi}\right)^T \tilde{\mathbf{u}}^{\phi}$, is bounded by

$$c_1 E(t) \leq \widetilde{E}(t) \leq c_2 E(t), \tag{2.51}$$

where c_1, $c_2 > 0$ are real constants.

Boundary stencil derivation for various order of accuracies is discussed in the next section. The goal is to satisfy the stability and conservation conditions of Theorems 2 and 3, which follows if a stencil satisfies (2.15) with $\sum_{i=1}^{n} q_{i0} = -1$. In cases where stencils that satisfy (2.15) could not be found, stencils that ensure (2.36) are derived, which also ensures (2.14) is satisfied, providing a strongly time-stable scheme for the scalar problem (1.1)-(2.1) as well as for the hyperbolic system (2.21)-(2.24). The strong time-stability condition (2.36), however, does not ensure that the conservation condition (2.11) is satisfied.

The condition (2.11) with an $\mathcal{O}(\Delta x)$ error, given by

$$\frac{d}{dt}\int_0^1 U\,dx \approx \sum_{i=1}^{n} \left(\frac{d}{dt} H\tilde{\mathbf{u}}\right)_i = -\sum_{i=1}^{n} \left(HD\mathbf{u}\right)_i = g(t) - u_n(t) + \mathcal{O}(\Delta x), \tag{2.52}$$

can be satisfied, concurrently with (2.36), if

$$q_{ij} \begin{cases} = -q_{ji} & \text{if} \quad i \neq j, \\ > \frac{q_{i0}^2}{4 q_{nn} a_i}\,\|L\|\,\|R\| & \text{if} \quad i = j \leq \beta, \\ = 0 & \text{if} \quad \beta < i = j < n, \\ = \frac{1}{2} & \text{if} \quad i = j = n, \end{cases} \tag{2.53}$$

and $\sum_{j=0}^{\kappa}\sum_{i=1}^{n} q_{ij} = -1$, where a_i is as in Theorem 3 and κ is the depth of the boundary block in H and $\tilde{Q}$ (as denoted in (2.3) and further described in Section 3). Obviously, condition (2.52) converges to (2.11) as $\Delta x \to 0$.

For brevity, the above-derived conditions will be referred to in the following sections as:

- Condition I: if a stencil satisfies (2.15) with $\sum_{i=1}^{n} q_{i0} = -1$,
- Condition II: if a stencil satisfies (2.53) with $\sum_{j=0}^{\kappa}\sum_{i=1}^{n} q_{ij} = -1$.

Both conditions ensure strong time stability for the scalar problem (1.1)-(2.1) as well as for the hyperbolic system (2.21)-(2.24). But while Condition I directly satisfies the conservation condition (2.11), Condition II satisfies the conservation condition to within an $\mathcal{O}(\Delta x)$ error, given by (2.52).

Remark. To put the $\mathcal{O}(\Delta x)$ error in context, the commonly-used approach of strong BC implementation [18, 15], using a projection operator that omits rows (corresponding to the grid points where the boundary data is injected) in a square derivative operator, introduces an $\mathcal{O}(1)$ conservation error, as shown in Lemma A.1 of Appendix A for the scalar problem (1.1), for example.

3. Stencil construction for various order of accuracies. The derived schemes are denoted by $p_b - p_i - p_b$, where p_b and p_i are the order-of-accuracy of boundary and interior stencils, respectively. If an energy estimate exists, the global order-of-accuracy of a $p_b - p_i - p_b$ scheme, where $p_b < p_i$, is expected to be $p_b + 1$ for first-order hyperbolic systems [13, 14]. The structure of the operators Q and H that determine the derivative approximation D are as described in the previous section. Q is of size $n \times (n+1)$, as defined in (2.9), and it can be written as

$$Q = \left[\begin{array}{c|c} \mathbf{q}_0 & \tilde{Q} \end{array}\right], \qquad \tilde{Q} = \left[\begin{array}{c|c|c} B_u^q & S & 0 \\ \hline -S^T & C & \left(S^T\right)^{\#} \\ \hline 0 & -S^{\#} & B_l^q \end{array}\right], \tag{3.1}$$

where $\mathbf{q}_0$ is the first column of Q given by (2.13) and $\tilde{Q}$ is a square $(n \times n)$ matrix with the upper-left and the lower-right boundary blocks given by

$$B_u^q = \begin{bmatrix} q_{11} & \cdots & \cdots & q_{1\kappa} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ q_{\kappa 1} & \cdots & \cdots & q_{\kappa\kappa} \end{bmatrix}, \qquad B_l^q = \begin{bmatrix} q_{n-\kappa+1,n-\kappa+1} & \cdots & \cdots & q_{n-\kappa+1,n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ q_{n,n-\kappa+1} & \cdots & \cdots & q_{nn} \end{bmatrix}, \tag{3.2}$$

and the interior blocks given by

$$C = \begin{bmatrix} 0 & c_1 & \cdots & c_w & & & \\ -c_1 & 0 & c_1 & \cdots & c_w & & \\ \cdots & -c_1 & 0 & c_1 & \cdots & c_w & \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & -c_w & \cdots & -c_1 & 0 & c_1 & \cdots \\ & & -c_w & \cdots & -c_1 & 0 & c_1 \\ & & & -c_w & \cdots & -c_1 & 0 \end{bmatrix}, \quad S = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ c_w & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ c_1 & \cdots & c_w & 0 & \cdots & 0 \end{bmatrix}. \tag{3.3}$$

The entries of B_u^q and B_l^q are the unknowns that will be determined to satisfy the stability and conservation conditions of Theorems 2 and 3. The entries of C and S are the centered scheme coefficients

$$c_k = -\frac{(-1)^k \, (w!)^2}{k \, (w+k)! \, (w-k)!} \qquad \text{for} \qquad 1 \le k \le w, \tag{3.4}$$

with half-stencil width $w = p_i/2$. Theorems 2 and 3 assume a real symmetric positive-definite matrix H. If the matrix H is diagonal, the corresponding stencil is referred to as a diagonal-norm stencil and if H has a block structure at the boundaries, the stencil is referred to as a full-norm stencil. H can be written as

$$H = \Delta x \begin{bmatrix} B_u^h & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & B_l^h \end{bmatrix}, \tag{3.5}$$

where $B_u^h = \operatorname{diag}(h_{11}, \cdots, h_{\kappa\kappa})$ and $B_l^h = \operatorname{diag}(h_{n-\kappa+1,n-\kappa+1}, \cdots, h_{nn})$ for a diagonal norm and

$$B_u^h = \begin{bmatrix} h_{11} & \cdots & \cdots & h_{1\kappa} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ h_{1\kappa} & \cdots & \cdots & h_{\kappa\kappa} \end{bmatrix}, \qquad B_l^h = \begin{bmatrix} h_{n-\kappa+1,n-\kappa+1} & \cdots & \cdots & h_{n-\kappa+1,n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ h_{n-\kappa+1,n} & \cdots & \cdots & h_{n,n} \end{bmatrix} \tag{3.6}$$

for a full norm. The unknowns $B_u^{q,h}$ and $B_l^{q,h}$ are determined using Algorithm 3.1 to satisfy the stability and conservation conditions, described in brevity by Condition I and II in the previous section.

The algorithm was executed in Mathematica [16] using $N_\kappa = 8$ for high-order cases and the non-linear optimization to maximize $\|L\|\,\|R\|$ was performed using the IPOPT library [36]. The $1-2-1$ scheme obtained from the algorithm is

$$D = \frac{1}{\Delta x} \begin{bmatrix} -\frac{2}{3} & \frac{1}{3} & \frac{1}{3} & & & \\ & -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & & -1 & 1 \end{bmatrix}, \qquad H = \Delta x \begin{bmatrix} \frac{3}{2} & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & \frac{1}{2} \end{bmatrix}. \tag{3.7}$$

Here $\beta = \kappa = 1$ and

$$q_0 = \begin{bmatrix} -1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}, \qquad \tilde{Q} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{3.8}$$

The $2-4-2$ scheme that is expected to provide a global third order-of-accuracy [13, 14] is presented in Appendix B. The $3-4-3$ and $3-6-3$ schemes that provide global fourth order-of-accuracy are included in the supplementary material. Matlab scripts for each stencil are also included in the supplementary material.

Important attributes of these schemes are summarized in Table 1. Boundary blocks are of size $\kappa = 4$ in the high-order schemes. $1-2-1$ and $2-4-2$ schemes have diagonal norm matrix, while $3-4-3$ and $3-6-3$ schemes have full norm matrix. Symbolic computations with values of κ up to 8 did not yield diagonal-norm

Algorithm 3.1 Determine $B_u^{q,h}$ and $B_l^{q,h}$.

```
input : Boundary and interior order of accuracy (p_b, p_i)
input : Limiting value of κ (N_κ)

AccuracyConstraint, CondI, CondII ← false
M ← 0
κ ← p_b
Use κ and p_i to construct Q and H as given by (3.1)–(3.6)
while κ ≤ N_κ do
    AccuracyConstraint ← Can the free parameters in Q and H satisfy
                             the order-of-accuracy constraints?
    if (AccuracyConstraint) then
        Update B_u^{q,h} and B_l^{q,h} to satisfy the order-of-accuracy constraints
        CondI ← Can the remaining free parameters satisfy Condition I?
        if (CondI) then
            Update B_u^{q,h} and B_l^{q,h} to satisfy Condition I
            return B_u^{q,h} and B_l^{q,h}
        else
            Optimize entries of B_u^{q,h} and B_l^{q,h} to satisfy Condition II, while
                                                    maximizing ‖L‖ ‖R‖
            CondII ← Is an optimal solution found?
            if (CondII) then
                Update B_u^{q,h} and B_l^{q,h} if the found optimal ‖L‖ ‖R‖ > M
                M ← max (M, optimal ‖L‖ ‖R‖)
            end if
            κ++
        end if
    else
        κ++
    end if
end while

if(M > 0) return B_u^{q,h} and B_l^{q,h}
else return no solution found
```

$3-4-3$ and $3-6-3$ schemes that simultaneously satisfy the strong time-stability and conservation constraints. All the schemes listed in Table 1 are provably strongly time stable with strong (or exact) BCs for scalar convection problems as well as for the coupled hyperbolic systems with $\|L\|\,\|R\|$ values as listed in the table. To the best of our knowledge, strongly time-stable schemes with non-dissipative centered schemes in the interior and strong BCs have not been reported in literature for hyperbolic problems. The $1-2-1$ scheme satisfies Condition I, while the high-order schemes satisfy Condition II. Numerical tests to verify the accuracy and stability of these schemes are presented in the next section.

4. Numerical results. This section examines numerical results from application of the schemes discussed in the previous section. In all cases, time integration is performed using the classical fourth-order Runge-Kutta (RK4) method with a CFL of

Scheme	κ	norm	Strong time stability		Conservation	
			Scalar convection	Coupled system	Condition I	Condition II
$1-2-1$	1	diagonal	✓	$\|L\| \, \|R\| < 1$	✓	✓
$2-4-2$	4	diagonal	✓	$\|L\| \, \|R\| \leq 1/4$	X	✓
$3-4-3$	4	full	✓	$\|L\| \, \|R\| \leq 1/6$	X	✓
$3-6-3$	4	full	✓	$\|L\| \, \|R\| \leq 1/3$	X	✓

TABLE 1

Summary of the strong time stability and conservation properties of various schemes. ✓ *denotes that the scheme satisfies that condition, whereas* X *denotes that it does not.*

0.8, unless mentioned otherwise. For convergence studies, the time step is taken small enough such that the temporal errors are insignificant compared to the spatial truncation errors. The schemes discussed in Section 3 allow imposition of exact boundary conditions (EBC), therefore, for brevity, we will refer to them as EBC schemes in the following sections.

4.1. 1-D scalar advection equation. Consider the scalar hyperbolic equation (1.1) with the initial and the boundary condition given by

$$u(x,0) = \sin 2\pi x, \qquad u(0,t) = g(t) = \sin 2\pi\,(-t)\,. \tag{4.1}$$

The exact solution to the problem is $u(x,t) = \sin 2\pi\,(x-t)$. A semi-discretization to the problem, using strong BCs, the notation of (2.2), and the decomposition described in (2.9), is given by

$$\frac{d\tilde{\mathbf{u}}}{dt} = -D\mathbf{u} = -H^{-1}\tilde{Q}\tilde{\mathbf{u}} - H^{-1}\mathbf{q}_0 g. \tag{4.2}$$

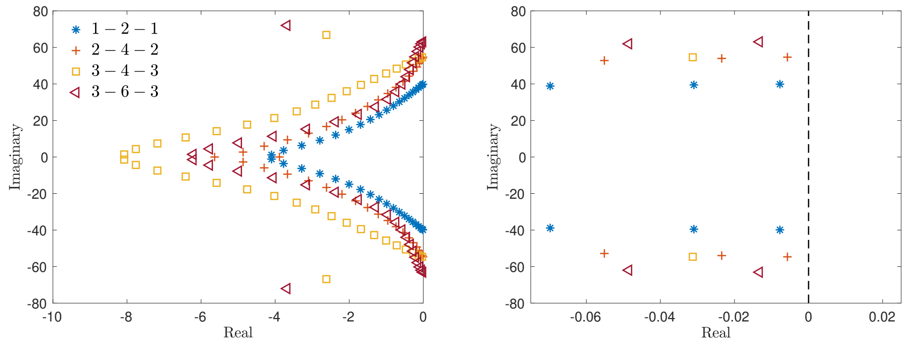
For a bounded boundary data $g(t)$, the stability of the semi-discretization depends on the properties of the matrix $M = -H^{-1}\tilde{Q}$, referred to as the system matrix [5]. If the semi-discretization (4.2) is time stable (as per Definition 1), then, from Theorem 1, the real part of all eigenvalues of the system matrix, M, must be non-positive. Figure 1 shows the eigenvalue spectrum of the system matrix using the EBC schemes with $n = 40$. All eigenvalues for all schemes lie in strict left half of the complex plane and, therefore, all the schemes show time stability for this problem, as expected from the theoretical proof.

Table 2 shows the L_2- and $L_\infty-$norm of the solution error, denoted by ε, and the respective convergence rates from the EBC schemes. As expected, all schemes converge with at least p_b+1 global order-of-accuracy, where p_b is the order-of-accuracy of the boundary stencils.

4.2. 1-D coupled hyperbolic system. This section examines the performance of the EBC schemes for a 2×2 system coupled by the boundary conditions. This system provides a severe test of numerical stability [6, 1] and, as noted by Carpenter *et al.* [6], no existing central difference scheme of order-of-accuracy greater than two is time stable for this system with strong BCs. Here, we evaluate the numerical stability and accuracy of boundary closures for various centered schemes with strong BCs.

The hyperbolic system, on domain $0 \leq x \leq 1$ and $t \geq 0$, is given by

$$\frac{\partial U}{\partial t} + \frac{\partial U}{\partial x} = 0, \tag{4.3}$$



(a) (b)

FIG. 1. *Eigenvalue spectrum of the system matrix to solve (1.1) with initial and boundary condition given by (4.1) using* $n = 40$ *and various schemes. (a) All eigenvalues and (b) magnified view near the imaginary axis. Legend is the same for both plots.*

n	1 − 2 − 1				2 − 4 − 2			
	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate
20	-1.442427		-1.234263		-1.828907		-1.541334	
40	-2.044080	1.999	-1.834978	1.996	-2.789357	3.215	-2.335029	2.637
80	-2.644558	1.995	-2.435158	1.994	-3.729319	3.298	-3.204515	2.888
160	-3.245543	1.996	-3.039630	2.008	-4.653197	3.110	-4.099137	2.972
320	-3.846993	1.998	-3.646874	2.017	-5.567189	3.046	-5.000487	2.994
640	-4.448730	1.999	-4.250385	2.005	-6.475805	3.027	-5.903084	2.998
n	3 − 4 − 3				3 − 6 − 3			
	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate
20	-1.946005		-1.801309		-1.681300		-1.160180	
40	-3.355294	4.682	-3.137755	4.440	-3.107325	4.737	-2.484588	4.399
80	-4.706377	4.488	-4.522492	4.599	-4.493823	4.606	-3.853948	4.549
160	-5.978925	4.227	-5.810145	4.277	-5.775224	4.257	-5.200389	4.473
320	-7.208962	4.086	-7.047904	4.112	-7.002522	4.077	-6.504670	4.333
640	-8.418204	4.017	-8.233410	3.938	-8.213472	4.023	-7.770852	4.206

TABLE 2

L_2- *and* $L_\infty-$*norm of the solution error and convergence rates from solving (1.1) using various schemes. Error calculations performed at* $t = 1.0$.

$$\frac{\partial V}{\partial t} - \frac{\partial V}{\partial x} = 0. \tag{4.4}$$

$$\text{Initial conditions}: \qquad U(x,0) = \sin 2\pi x, \qquad V(x,0) = -\sin 2\pi x. \tag{4.5}$$

$$\text{Boundary conditions}: \qquad U(0,t) = \alpha_1 V(0,t), \qquad V(1,t) = \alpha_2 U(1,t). \tag{4.6}$$

For $\alpha_1 = \alpha_2 = 1$, this system provides a strict test of numerical stability because it is neutrally stable, *i.e.*, the energy, $\int_0^1 \left[U(x,t)^2 + V(x,t)^2\right] \mathrm{d}x$, remains constant with time.

Let $\mathbf{u}(t) = \begin{bmatrix} u_0(t) & \cdots & \cdots & u_n(t) \end{bmatrix}^T$ and $\mathbf{v}(t) = \begin{bmatrix} v_0(t) & \cdots & \cdots & v_n(t) \end{bmatrix}^T$ denote the grid function, assuming a spatial discretization of the above system with $n+1$ grid points. A semi-discretization of (4.3)-(4.6) using strong boundary conditions is given by

$$\frac{d\mathbf{w}}{dt} = -\mathcal{D}\mathbf{w}, \tag{4.7}$$

where $\mathbf{w}(t) = \begin{bmatrix} \tilde{\mathbf{u}}(t) & \tilde{\mathbf{v}}(t) \end{bmatrix}^T$ with $\tilde{\mathbf{u}}(t) = \begin{bmatrix} u_1(t) & \cdots & \cdots & u_n(t) \end{bmatrix}^T$ and $\tilde{\mathbf{v}}(t) = \begin{bmatrix} v_0(t) & \cdots & \cdots & v_{n-1}(t) \end{bmatrix}^T$. The derivative operator, $\mathcal{D}$, is given by

$$\mathcal{D} = \begin{bmatrix} H & 0 \\ 0 & H^{\#} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{Q} & \alpha_1 Q_0 \\ -\alpha_2 Q_0^{\#} & -\tilde{Q}^{\#} \end{bmatrix} = \mathcal{H}^{-1}\mathcal{Q},$$

where $\tilde{Q}$ and Q_0 are as described in (2.9) and (2.28), respectively, and the superscript $^{\#}$ denotes the matrix transformation (2.30). The off-diagonal entries of $\mathcal{Q}$, involving Q_0, apply the boundary conditions (4.6) strongly.

As mentioned earlier, existing high-order central difference schemes fail to be stable for this problem when solved with strong BCs. Figure 2(a) shows the eigenvalue spectrum of the system matrix, given by $-\mathcal{D}$ in (4.7), for the neutrally-stable problem with various high-order schemes from the literature. All schemes exhibit eigenvalues with positive real part, therefore, the numerical solution grows non-physically in a long-time simulation, as shown by the solution error (ε) plotted in Figure 2(b).

Figure 3 shows the the eigenvalue spectrum of the system matrix for the neutrally-stable problem from various EBC schemes discussed in Section 3. The eigenvalues lie in strict left half of the complex plane in all cases indicating time stability. Further, the eigenvalue spectrum for $\alpha_1 = \alpha_2 = 1/2$ from various EBC schemes is depicted in Figure 4. All derived schemes are also time stable for this problem, and larger negative real part of the eigenvalues compared to Figure 3 indicates the dissipative nature of the boundary conditions. Eigenvalue spectrum from various values of n (not presented here for brevity) showed similar time-stable behavior. Table 3 shows the L_2- and $L_\infty-$norm of the solution error, denoted by ε, and the respective convergence rates from the EBC schemes for this problem. All schemes converge with approximately $p_b + 1$ global order-of-accuracy.

4.3. Inviscid Burgers' equation. Consider the inviscid Burgers' equation with a source term,

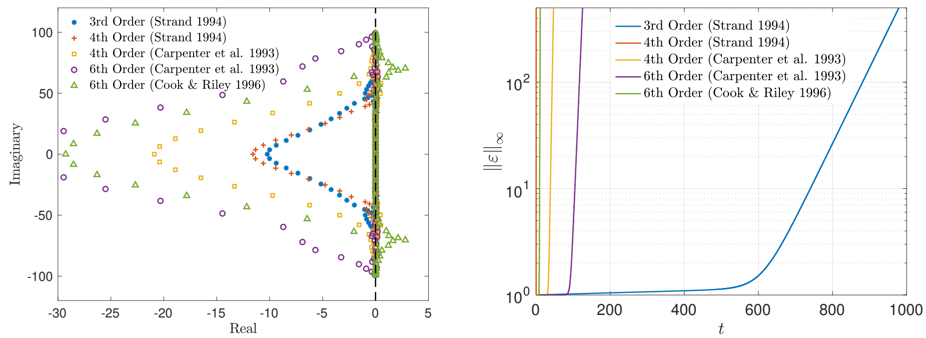
$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x}\left(\frac{U^2}{2}\right) = f_U. \qquad 0 \le x \le 1,\ t \ge 0, \tag{4.8}$$

The method of manufactured solutions [27] is employed to perform long-time simulations to assess the stability and the accuracy of the derived schemes. The source term prevents solution discontinuities. The solution is assumed to be

$$U(x,t) = \sin 2\pi (x-t) + C, \tag{4.9}$$

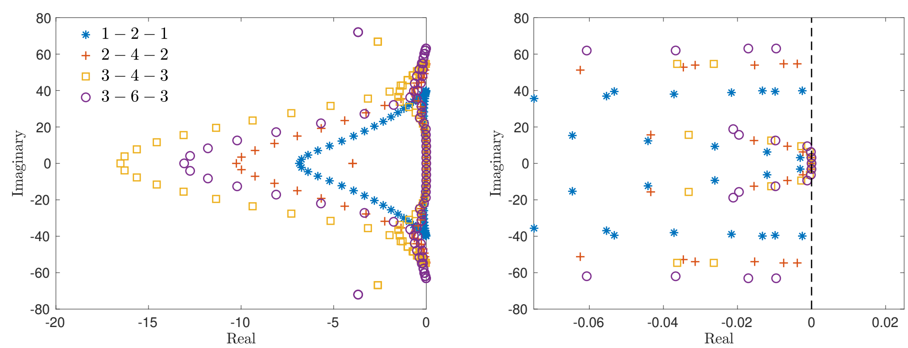
where $C = 1.0$ is a constant. (4.9) prescribes the initial and the boundary data, and the source term is given by

$$f_U(x,t) = \pi \sin 4\pi (x-t). \tag{4.10}$$



(a) (b)

FIG. 2. *(a) Eigenvalue spectrum of the system matrix near imaginary axis and (b) $L_\infty-$error from solving the coupled hyperbolic system (4.3)-(4.6) with $\alpha_1 = \alpha_2 = 1$ using various spatial schemes from literature with strong boundary conditions. Classical RK4 is used for time integration with a CFL of 0.25 and 40 grid points in the domain. 3rd Order (Strand 1994) denotes the diagonal-norm stencil in [32, Appendix A] that is second-order accurate at the boundary; 4th Order (Strand 1994) denotes the minimum-bandwidth full-norm stencil in [32, Appendix B] that is third-order accurate at the boundary; 4th Order (Carpenter et al. 1993) and 6th Order (Carpenter et al. 1993) denote the $4^3 - 4 - 4^3$ and $5^2, 5^2 - 6 - 5^2, 5^2$ stencil of [5], respectively; 6th Order (Cook & Riley 1996) denotes the sixth-order compact scheme of [9, Section 7.3].*

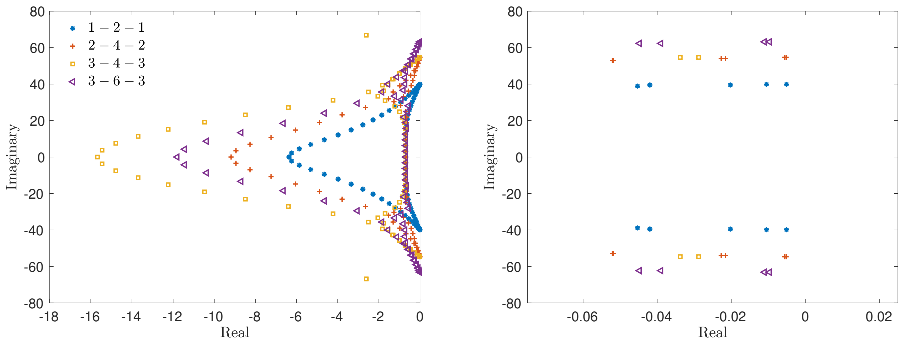


(a) (b)

FIG. 3. *Eigenvalue spectrum of the system matrix to solve (4.3)-(4.6) with $\alpha_1 = \alpha_2 = 1$ using $n = 40$ and various schemes. (a) All eigenvalues and (b) magnified view near the imaginary axis. Legend is the same for both plots..*

The solution (4.9) is non-negative in the domain at all times, therefore, the boundary condition $U(0,t) = \sin 2\pi(-t) + C$ makes the problem well-posed.

Figure 5 shows the $L_\infty-$errors with time in long-time simulations using various schemes. A constant error profile indicates time-stable behavior. Figure 5(a) shows the errors from the EBC schemes and, for comparison, figure 5(b) shows the errors from the schemes (from literature) used in Figure 2. While all the schemes of figure 5(b) were unstable with strong BC implementation for the neutrally-stable coupled system of Section 4.2, the diagonal-norm 3rd-order scheme of [32] and the 4th-order compact scheme of [5] show time stability for this problem. The other schemes diverge early in time. Table 4 shows the L_2- and $L_\infty-$norm of the solution error and the respective convergence rates from the EBC schemes. All schemes show approximately



(a) (b)

Fig. 4. *Eigenvalue spectrum of the system matrix to solve (4.3)-(4.6) with $\alpha_1 = \alpha_2 = 1/2$ using $n = 40$ and various schemes. (a) All eigenvalues and (b) magnified view near the imaginary axis. Legend is the same for both plots..*

n	1 − 2 − 1				2 − 4 − 2			
	$\log_{10}\Vert\varepsilon\Vert_2$	Rate	$\log_{10}\Vert\varepsilon\Vert_\infty$	Rate	$\log_{10}\Vert\varepsilon\Vert_2$	Rate	$\log_{10}\Vert\varepsilon\Vert_\infty$	Rate
20	-1.217223		-1.225890		-1.676188		-1.508359	
40	-1.803716	1.948	-1.770808	1.810	-2.643277	3.215	-2.351858	2.802
80	-2.398761	1.977	-2.353810	1.937	-3.582599	3.120	-3.206750	2.840
160	-2.997715	1.990	-2.955241	1.998	-4.505004	3.064	-4.099017	2.964
320	-3.598344	1.995	-3.555882	1.995	-5.417936	3.035	-5.000116	2.993
640	-4.199721	1.998	-4.157098	1.997	-6.325949	3.016	-5.902821	2.999
n	3 − 4 − 3				3 − 6 − 3			
	$\log_{10}\Vert\varepsilon\Vert_2$	Rate	$\log_{10}\Vert\varepsilon\Vert_\infty$	Rate	$\log_{10}\Vert\varepsilon\Vert_2$	Rate	$\log_{10}\Vert\varepsilon\Vert_\infty$	Rate
20	-1.771245		-1.777406		-1.992622		-1.811695	
40	-3.162562	4.622	-3.117156	4.451	-3.419956	4.742	-3.201140	4.616
80	-4.494335	4.424	-4.487070	4.551	-4.765353	4.469	-4.638031	4.773
160	-5.760716	4.207	-5.754529	4.210	-6.020126	4.168	-5.970614	4.427
320	-6.983890	4.063	-6.979341	4.069	-7.237780	4.045	-7.187086	4.041
640	-8.083748	3.654	-8.059036	3.587	-8.445220	4.011	-8.394934	4.012

Table 3

L_2- and $L_\infty-$norm of the solution error and convergence rates from solving (4.3)-(4.6) using various schemes. Error calculations performed at $t = 1.0$.

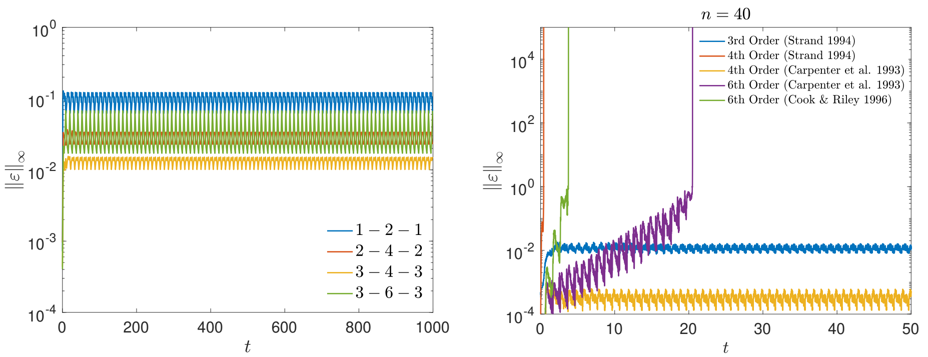
$p_b + 1$ global order-of-accuracy.

4.4. 2-D variable-coefficient advection equation . Consider the scalar problem

$$\frac{\partial \phi}{\partial t} + u\frac{\partial \phi}{\partial x} + v\frac{\partial \phi}{\partial y} = 0, \qquad 0 \le x, y \le L \qquad t \ge 0,$$

$$u(x,y) = \frac{\partial r}{\partial x}, \;\; v(x,y) = \frac{\partial r}{\partial y}, \tag{4.11}$$

$$r\,(x,y) = \sqrt{\left(x - x_0\right)^2 + \left(y - y_0\right)^2},$$



(a) (b)

Fig. 5. *$L_\infty-$error from long-time simulations of (4.8) using $n = 40$ with (a) EBC schemes and (b) schemes from literature referenced in Figure 2. Note the difference in axis scales between the two subfigures.*

n	$1-2-1$				$2-4-2$			
	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate
20	-1.176427		-0.738973		-1.987098		-1.528051	
40	-1.781249	2.009	-1.280048	1.797	-2.833219	2.811	-2.230067	2.332
80	-2.414563	2.104	-1.740013	1.528	-3.677578	2.805	-3.068572	2.785
160	-3.042268	2.085	-2.244332	1.675	-4.547528	2.890	-3.765753	2.316
320	-3.658045	2.046	-2.793888	1.826	-5.397666	2.824	-4.496024	2.426
640	-4.268635	2.028	-3.375075	1.931	-6.297327	2.989	-5.369695	2.902
n	$3-4-3$				$3-6-3$			
	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate
20	-2.222065		-1.804324		-2.437228		-2.136384	
40	-3.540079	4.378	-2.998817	3.968	-3.463385	3.409	-2.876606	2.459
80	-4.765848	4.072	-4.154203	3.838	-4.686090	4.062	-4.306489	4.750
160	-6.170204	4.665	-5.414077	4.185	-5.876633	3.955	-5.540050	4.098
320	-7.502206	4.425	-6.640841	4.075	-7.068063	3.958	-6.705179	3.870
640	-8.748263	4.139	-7.871633	4.089	-8.264577	3.975	-7.884398	3.917

Table 4

L_2- and $L_\infty-$norm of the solution error and convergence rates from solving (4.8) using various schemes. Error calculations performed at $t = 1.0$.

where $L = \sqrt{2}$, $x_0 = -0.25$ and $y_0 = -0.25$. The initial and the boundary conditions are given by

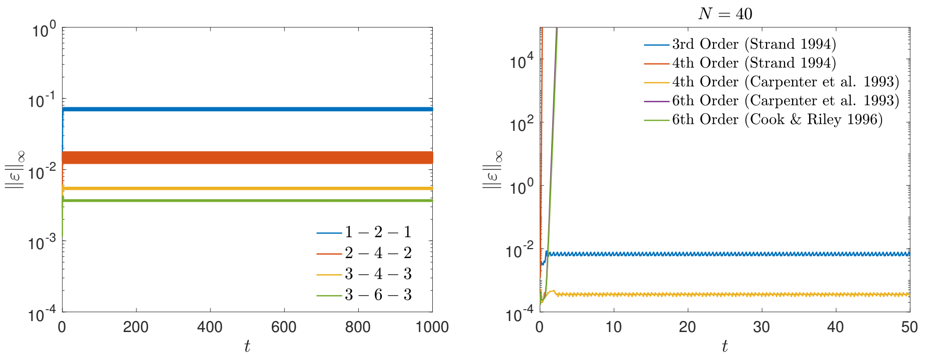
$$\phi(x, y, 0) = \sin 2\pi r, \tag{4.12}$$

and

$$\phi(0, y, t) = \sin 2\pi \left(r\left(0, y\right) - t\right), \qquad \phi(x, 0, t) = \sin 2\pi \left(r\left(x, 0\right) - t\right), \tag{4.13}$$

respectively. The exact solution to the problem is $\phi(x, y, t) = \sin 2\pi\left(r - t\right)$.

Figure 6 shows the $L_\infty-$errors from long-time simulations of (4.11)-(4.13) using various schemes with $N \times N$ grid points. To highlight the efficacy of the derived



(a) (b)

Fig. 6. *$L_\infty-$error from long-time simulations of (4.11)-(4.13) using $N = 40$ with (a) EBC schemes and (b) schemes from literature referenced in Figure 2. Note the difference in axis scales between the two subfigures.*

N	$1-2-1$				$2-4-2$			
	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate
30	-1.404196		-1.037912		-2.118175		-1.439252	
60	-2.018962	1.993	-1.615432	1.872	-3.092760	3.160	-2.445756	3.263
120	-2.626948	1.995	-2.207732	1.944	-4.035679	3.095	-3.435688	3.249
240	-3.232256	1.999	-2.801850	1.962	-4.954626	3.034	-4.343758	2.998
N	$3-4-3$				$3-6-3$			
	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate	$\log_{10}\|\varepsilon\|_2$	Rate	$\log_{10}\|\varepsilon\|_\infty$	Rate
30	-2.337262		-1.760082		-2.513605		-1.942708	
60	-3.646564	4.245	-2.929111	3.790	-3.838707	4.296	-3.070683	3.657
120	-4.956536	4.299	-4.123419	3.920	-5.147756	4.296	-4.249164	3.868
240	-6.268282	4.331	-5.324285	3.965	-6.456819	4.322	-5.447414	3.957

Table 5

L_2- and $L_\infty-$norm of the solution error and convergence rates from solving (4.11)-(4.13) on a $N \times N$ grid using various schemes. Error calculations performed at $t = 1.0$.

schemes, a CFL of 1.5, calculated from

$$\mathrm{CFL} = \Delta t \left(\frac{|u|}{\Delta x} + \frac{|v|}{\Delta y} \right),$$

is used for the results of this figure. Figure 6(a) shows the errors from the EBC schemes and figure 6(b) shows the errors from the schemes used in Figure 2. As in the case of inviscid Burgers' equation in the previous section, the diagonal-norm 3rd-order scheme of [32] and the 4th-order compact scheme of [5] show time stability. The other schemes tend to diverge very early in time. Table 5 shows the L_2- and $L_\infty-$norm of the solution error and the respective convergence rates from the EBC schemes. All schemes show approximately $p_b + 1$ global order-of-accuracy.

5. Application to the Euler equations using characteristic boundary conditions. This section discusses the application of the schemes derived in Section 3 to solve the two-dimensional Euler equations. The extension to three-dimensions follows a similar approach. The primary interest of this study is in high-fidelity fluid-flow simulations, and hence the performance of the derived schemes is analyzed for the Euler equations. Theoretical stability and convergence analysis of finite-difference and pseudo-spectral schemes for other non-linear hyperbolic PDEs can be found in [37, 3, 7, 38], for example.

The two-dimensional Euler equations, assuming a calorically perfect gas, in generalized coordinates are given by

$$\frac{\partial \mathbf{Q}}{\partial \tau} + \frac{\partial \mathbf{F}}{\partial \xi} + \frac{\partial \mathbf{G}}{\partial \eta} = 0, \tag{5.1}$$

$$\mathbf{Q} = \frac{1}{J}\begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{bmatrix}, \qquad \mathbf{F} = \frac{1}{J}\begin{bmatrix} \rho U \\ \rho u U + \xi_x p \\ \rho v U + \xi_y p \\ \rho E U + \xi_{x_i} u_i p \end{bmatrix}, \qquad \mathbf{G} = \frac{1}{J}\begin{bmatrix} \rho V \\ \rho u V + \eta_x p \\ \rho v V + \eta_y p \\ \rho E V + \eta_{x_i} u_i p \end{bmatrix}, \tag{5.2}$$

$$U = \xi_t + \xi_x u + \xi_y v, \qquad V = \eta_t + \eta_x u + \eta_y v,$$

$$\rho E = \frac{p}{\gamma - 1} + \rho\left(\frac{u^2 + v^2}{2}\right).$$

The coordinate transformation between the physical domain $\mathbf{x} = (x, y)$ and the computational domain $\boldsymbol{\xi} = (\xi, \eta)$ is $\boldsymbol{\xi} = \boldsymbol{\Xi}(\mathbf{x}, t)$ with the inverse transformation $\mathbf{x} = \mathbf{X}(\boldsymbol{\xi}, \tau)$ and the metric Jacobian $J = \det(\partial\boldsymbol{\xi}/\partial\mathbf{x}) = (x_\xi y_\eta - x_\eta y_\xi)^{-1}$. We assume the time to be invariant, therefore, $\tau = t$. u, v are the Cartesian velocity components, ρ denotes the density, p is the pressure, and E is the total energy per unit mass.

Let i and j denote the grid indices in ξ and η direction, respectively, where $0 \le i \le N_\xi$ and $0 \le j \le N_\eta$ for a $(N_\xi + 1) \times (N_\eta + 1)$ computational grid. To simplify the discussion, let us consider the boundary located at $i = 0$, which has a constant ξ value. The flux-derivative in the ξ-direction in (5.1), then, has to be modified to account for the physical boundary condition. (5.1) can be transformed to a characteristic form in the direction normal to the $i = 0$ boundary by using a similarity transformation $A = \partial\mathbf{F}/\partial\mathbf{Q} = T_\xi \Lambda_\xi T_\xi^{-1}$, where the columns of T_ξ contain the right eigenvectors of A and Λ_ξ is a diagonal matrix containing the eigenvalues of A. The expressions for Λ_ξ and T_ξ can be found in [24]. The resulting characteristic equations are given by (*e.g.* [17])

$$\frac{\partial \mathbf{R}}{\partial t} + \mathbf{L} = \mathbf{S}_C, \tag{5.3}$$

where $\mathbf{R}$ is the vector of characteristic variables,

$$\mathbf{L} = J T_\xi^{-1}\left\{\frac{\partial \mathbf{F}}{\partial \xi} - \left[\mathbf{F}\frac{\partial}{\partial \xi}\left(\frac{\xi_x}{J}\right) + \mathbf{G}\frac{\partial}{\partial \xi}\left(\frac{\xi_y}{J}\right)\right]\right\} \tag{5.4}$$

and

$$\mathbf{S}_C = -J T_\xi^{-1}\left\{\frac{\partial \mathbf{G}}{\partial \eta} + \left[\mathbf{F}\frac{\partial}{\partial \xi}\left(\frac{\xi_x}{J}\right) + \mathbf{G}\frac{\partial}{\partial \xi}\left(\frac{\xi_y}{J}\right)\right]\right\}. \tag{5.5}$$

The square brackets in (5.4)–(5.5) preserve the conservative form of the equation [17].

Following the one-dimensional discretization described in Section 2, a semi-discretization of (5.1) at grid points within the boundary-stencil depth from the $i = 0$ boundary, *i.e.*, $0 \leq i \leq \kappa$, can be written as

$$\frac{dq_{ij}}{dt} = -\left(\frac{1}{J} S_\xi \mathbf{L}^* + \left[\mathbf{F}\frac{\partial}{\partial\xi}\left(\frac{\xi_x}{J}\right) + \mathbf{G}\frac{\partial}{\partial\xi}\left(\frac{\xi_y}{J}\right)\right]\right)_{ij} - (D_\eta \mathbf{g})_{ij}\,, \tag{5.6}$$

where q_{ij} and $(D_\eta \mathbf{g})_{ij}$ are the discrete approximations of $\mathbf{Q}$ and $\partial\mathbf{G}/\partial\eta$ at i, j grid point and $\mathbf{L}^*$ denotes the modified characteristic convection term in ξ-direction given by

$$\mathbf{L}^* = \mathbf{L}^*_{\mathrm{SBP}} + \mathbf{L}^*_{\mathrm{EBC}}. \tag{5.7}$$

The derivative operators derived in Section 3 to satisfy the stability and conservation constraints of Section 2 are non-square and use different stencils at inflow (where physical boundary condition is applied) and outflow boundaries. The outflow boundary uses an SBP stencil, whereas stencils for the inflow boundary, derived in Section 3, that impose the exact boundary conditions (EBCs) will be referred to as the EBC stencils. $\mathbf{L}^*_{\mathrm{SBP}}$ denotes the convection terms for the outgoing waves calculated using the SBP stencil. The outgoing characteristics correspond to the negative entries of Λ_ξ at the $i = 0$ boundary, therefore, the elements of $\mathbf{L}^*_{\mathrm{SBP}}$ can be obtained from

$$(\mathbf{L}^*_{\mathrm{SBP}})_k = \frac{|\lambda_k| - \lambda_k}{2\,|\lambda_k|}\,(\mathbf{L}_{\mathrm{SBP}})_k\,, \tag{5.8}$$

where $(\bullet)_k$ denotes the k-th entry of the vector, λ_k is the k-th diagonal entry of Λ_ξ and $\mathbf{L}_{\mathrm{SBP}}$ is $\mathbf{L}$ in (5.4) calculated using the SBP derivative approximation. The prefactor $\frac{|\lambda_k|-\lambda_k}{2|\lambda_k|}$ ensures that the SBP stencil is applied only to the outgoing characteristic calculations. $\mathbf{L}^*_{\mathrm{EBC}}$ denotes the incoming characteristic convection terms that at $i = 0$ are calculated using the physical boundary data and at $0 < i \leq \kappa$ calculated using the EBC derivative stencils from

$$(\mathbf{L}^*_{\mathrm{EBC}})_k = \frac{|\lambda_k| + \lambda_k}{2\,|\lambda_k|}\,(\mathbf{L}_{\mathrm{EBC}})_k\,, \tag{5.9}$$

where the expressions are as described for (5.8).

Next, we describe the application of the above discretization to solve problems where the exact or target boundary data for all conservative variables may or may not be known. The metric terms are calculated using the SBP derivative approximation and time integration is performed using the classical fourth-order Runge-Kutta (RK4) method with a CFL of 0.6 for all results discussed in the following sections. For convergence studies, the time step is taken small enough such that the temporal errors are insignificant compared to the spatial truncation errors.

5.1. Isentropic convecting vortex. The two-dimensional Euler equations are solved for a compressible isentropic vortex propagation. Initial and boundary conditions are applied using the exact solution given by (*e.g.* [28])

$$\begin{aligned}
&\rho = \left(1 - \frac{\varpi^2(\gamma-1)}{8\pi^2 c_0^2} e^{1-\varphi^2 r^2}\right)^{\frac{1}{\gamma-1}}, && u = u_0 - \frac{\varpi}{2\pi}\varphi(y - y_0 - v_0 t)e^{\frac{1-\varphi^2 r^2}{2}},\\
&v = v_0 + \frac{\varpi}{2\pi}\varphi(x - x_0 - u_0 t)e^{\frac{1-\varphi^2 r^2}{2}}, && E = \frac{p}{\gamma-1} + \frac{1}{2}\rho(u^2+v^2),\\
&p = \rho^\gamma, && r^2 = (x - x_0 - u_0 t)^2 + (y - y_0 - v_0 t)^2,
\end{aligned} \tag{5.10}$$

where $(x_0,\, y_0)$ denotes the initial position of the vortex, (u_0, v_0) denotes the vortex convective velocity, φ is a scaling factor and ϖ denotes the non-dimensional circulation. $\gamma = 1.4$, $\varphi = 11$ and $\varpi = 1$ is used for all simulations. All quantities in (5.10) are non-dimensional, obtained from the density scale $= \rho_0^*$, velocity scale $u_0^* = \frac{c_0^*}{\sqrt{\gamma}}$, unit length scale and pressure scale $= \rho_0^* u_0^{*2}$, where $*$ denotes the dimensional quantities. The non-dimensional ambient speed of sound is $c_0 = \sqrt{\gamma}$.

Figure 8 shows the $L_\infty-$errors of velocity magnitude and density from simulations using $(x_0,\, y_0) = (-1.5,\, 0)$ on the domain shown in Figure 7, *i.e.*, the vortex is initially located outside the computational domain. A subsonic $(u_0 = 1.0,\, v_0 = 0)$ and a supersonic $(u_0 = 2.0,\, v_0 = 0)$ convective velocity is used to examine the robustness of the boundary implementation. In the subsonic case, the left/right boundary has three/one incoming and one/three outgoing characteristics. As per the characteristic eigenvalue/eigenvector matrices of [24], for the subsonic left boundary, the outgoing wave $(\mathbf{L}^*)_4 = (\mathbf{L}_{\mathrm{SBP}})_4$, the incoming waves $(\mathbf{L}^*)_{1,2,3}$ are calculated directly from the exact solution at $i = 0$ and $(\mathbf{L}^*)_{1,2,3} = (\mathbf{L}_{\mathrm{EBC}})_{1,2,3}$ at $0 < i \le \kappa$. For the subsonic right boundary, the outgoing waves $(\mathbf{L}^*)_{1,2,3} = (\mathbf{L}_{\mathrm{SBP}})_{1,2,3}$, the incoming wave $(\mathbf{L}^*)_4$ is calculated directly from the exact solution at $i = N_x = N_\xi$ and $(\mathbf{L}^*)_4 = (\mathbf{L}_{\mathrm{EBC}})_4$ at $N_\xi - \kappa \le i < N_\xi$. A similar characteristic treatment is used for the boundaries normal to the y-direction, where the incoming/outgoing waves are determined by the entries of Λ_η, obtained from the similarity transformation $B = \partial\mathbf{G}/\partial\mathbf{Q} = T_\eta \Lambda_\eta T_\eta^{-1}$ [24]. The supersonic case has characteristic velocities of the same sign at each x-boundary, therefore, theoretically, no similarity transformation is required to impose the boundary conditions. However, the code implementation performs a decomposition and assigns $\mathbf{L}^* = \mathbf{L}_{\mathrm{SBP}}$ at the right boundary and, at the left boundary, $\mathbf{L}^*$ is calculated directly from the exact solution at $i = 0$ and $\mathbf{L}^* = \mathbf{L}_{\mathrm{EBC}}$ at $0 < i \le \kappa$.

In the simulation duration shown in Figure 8, the vortex enters and exits the domain through the left and the right boundary, respectively. The two spikes in the plots of Figure 8 mark the time of vortex entry and exit. The time interval between the entry and the exit is longer for the subsonic case, as expected. The vortex entry/exit triggers numerical reflections from the inflow/outflow boundary, which can be a source of instability and, therefore, the simulation is setup to examine if the errors grow with time. All schemes of Section 3 are stable for this problem. The error decay rate is higher in the supersonic cases, likely, because of the simpler boundary treatment where all characteristic eigenvalues have the same sign.

The extent/magnitude of numerical reflections at the outflow boundary may depend on the flow direction at the boundary [2]. To examine the robustness of the developed schemes, several numerical tests were performed with vortex traveling in a direction that is oblique to the boundary. Figure 9 shows the velocity magnitude and density errors with time for a subsonic vortex traveling through the top-right corner of computational domain. Initial vortex location $(x_0,\, y_0) = (0,\, 0)$ with convective velocity $(u_0 = 0.8,\, v_0 = 0.4)$ allow the vortex to exit the domain in $t \lesssim 2$, allowing an assessment of error growth with time. Figure 9 shows the results from the EBC schemes of Section 3. All schemes produce stable results without any ad hoc stabilization measures indicating the suitability of these schemes for high-fidelity turbulent flow calculations [19, 29].

L_2- and $L_\infty-$norm of the solution error and respective convergence rates from the stable schemes for this problem are given in Table 6. The errors are calculated at $t = 1$ using $(x_0,\, y_0) = (-0.5,\, 0)$ for the subsonic $(u_0 = 1.0)$ case. All schemes exhibit a global order-of-accuracy approaching $p_b + 1$ or higher.

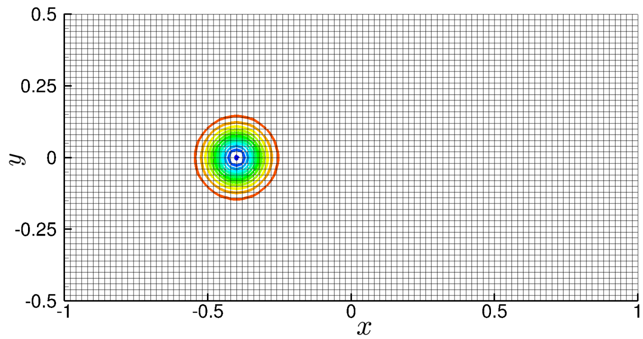
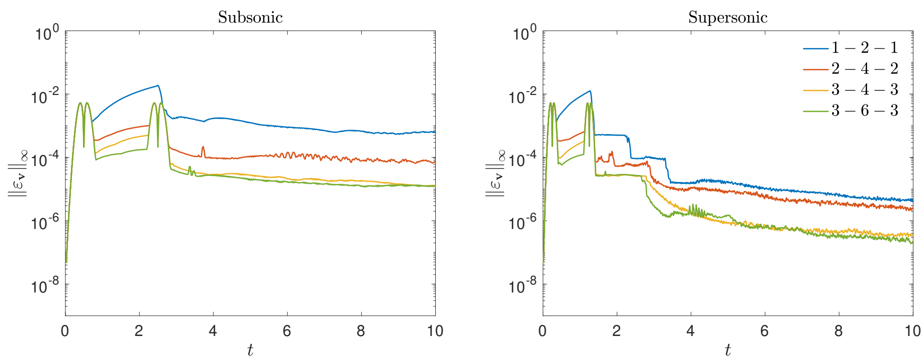
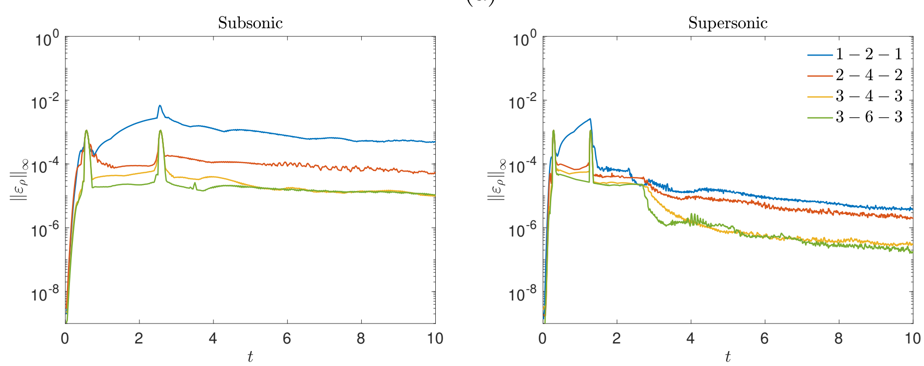


FIG. 7. *Computational domain for isentropic convective vortex simulations.*



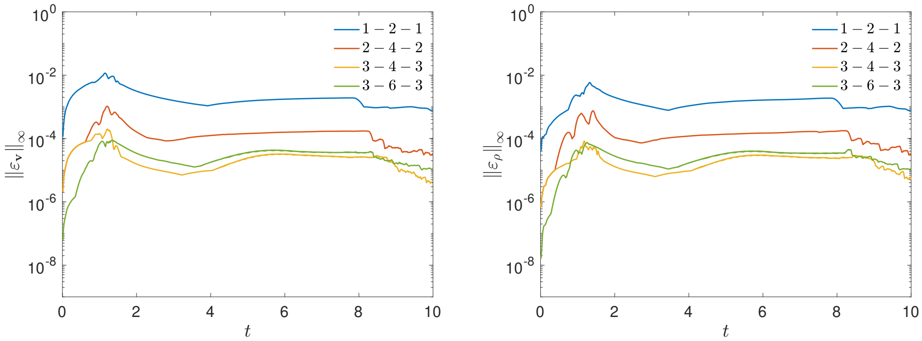
(a)



(b)

FIG. 8. *L_∞−norm of (a) velocity magnitude error and (b) density error from solving the Euler equations for isentropic convecting vortex using various EBC schemes with 201×101 grid points. Left and right columns show errors from a subsonic ($u_0 = 1.0$, $v_0 = 0$) and supersonic ($u_0 = 2.0$, $v_0 = 0$) convective velocity, respectively. Initial vortex location is $(x_0, y_0) = (-1.5, 0)$ for all simulations. Legend is the same for all plots.*

5.2. Acoustic scatter by a rigid cylinder. This section examines the performance of the EBC schemes on curvilinear grid to solve problems where the exact



(a) (b)

Fig. 9. *$L_\infty-$norm of (a) velocity magnitude error and (b) density error from solving the Euler equations for isentropic convecting vortex using various EBC schemes with 201×101 grid points. Convective velocity ($u_0 = 0.8$, $v_0 = 0.4$) with initial vortex location (x_0, y_0) = (0, 0) is used to simulate a subsonic vortex traveling through the top-right corner of computational domain.*

N	$1-2-1$				$2-4-2$			
	$\log_{10}\|\varepsilon_\rho\|_2$	Rate	$\log_{10}\|\varepsilon_\rho\|_\infty$	Rate	$\log_{10}\|\varepsilon_\rho\|_2$	Rate	$\log_{10}\|\varepsilon_\rho\|_\infty$	Rate
50	-2.97741		-1.99754		-3.55779		-2.68874	
100	-3.5008	1.714	-2.50662	1.667	-4.34167	2.566	-3.46567	2.544
150	-3.85544	1.997	-2.86293	2.007	-4.83949	2.804	-3.99159	2.962
200	-4.10382	1.977	-3.11973	2.043	-5.21171	2.962	-4.38812	3.155
250	-4.29947	2.010	-3.32405	2.099	-5.50531	3.016	-4.70008	3.204
N	$3-4-3$				$3-6-3$			
	$\log_{10}\|\varepsilon_\rho\|_2$	Rate	$\log_{10}\|\varepsilon_\rho\|_\infty$	Rate	$\log_{10}\|\varepsilon_\rho\|_2$	Rate	$\log_{10}\|\varepsilon_\rho\|_\infty$	Rate
50	-3.0783		-2.14384		-3.19404		-2.28162	
100	-4.06836	3.242	-3.14236	3.269	-4.25752	3.482	-3.32069	3.402
150	-4.70384	3.579	-3.73467	3.336	-4.94213	3.856	-3.97785	3.701
200	-5.22411	4.140	-4.22631	3.912	-5.50794	4.502	-4.51338	4.262
250	-5.65743	4.451	-4.64963	4.349	-5.96407	4.686	-4.95957	4.584

Table 6

L_2- and $L_\infty-$norm of the density error and convergence rates from solving the Euler equations for isentropic vortex convection on a $N \times N$ grid using various schemes. Error calculations are performed at $t = 1.0$.

(or target) values of all conservative variables are not known at the boundary, as is often the case in practical flow simulations. The strong BC implementations, unlike the weak enforcement, does not require target values for all conservative variables.

The Euler equations (5.1) are solved for scattering of an initial pressure pulse by a cylinder [34], as shown in Figure 10. The initial condition is given by

$$(5.11) \qquad p = \frac{1}{\gamma} + \varepsilon \exp\left[-(\ln 2)\,\frac{(x-4)^2 + y^2}{0.2^2}\right], \quad \rho = \left(1 - \frac{1}{\gamma}\right) + p, \quad u = v = 0,$$

where a small value of $\varepsilon = 10^{-4}$ is considered to trigger a linear response allowing comparison with the linearized Euler equations solution. The pressure disturbance is centered at $(x_s, y_s) = (4, 0)$. All quantities in (5.11) are non-dimensional, obtained from the density scale $= \rho_\infty^*$, velocity scale $= c_\infty^*$, length scale $= r_0$ (cylinder radius) and pressure scale $= \rho_\infty^* c_\infty^{*2}$, where $*$ denotes the dimensional quantities, subscript ∞ denotes the ambient values and c is the speed of sound.

Figure 10(a) shows the computational grid and the boundary conditions for the problem. The inviscid wall imposes the no-penetration condition normal to the wall and slip condition in the tangential direction. The no-penetration condition makes the contravariant velocity U in (5.2) zero and, therefore, $(\mathbf{L}^*)_1 = (\mathbf{L}^*)_2 = 0$ in (5.7), based on the eigenvalue arrangement of the characteristic matrices of [24]. $(\mathbf{L}^*)_4$ corresponds to the outgoing wave, therefore, $(\mathbf{L}^*)_4 = (\mathbf{L}_{\mathrm{SBP}})_4$ and the incoming wave $(\mathbf{L}^*)_3 = (\mathbf{L}_{\mathrm{SBP}})_4 + (\mathbf{S}_C)_3 - (\mathbf{S}_C)_4$, see [17]. The outflow has three outgoing and one incoming wave. $(\mathbf{L}^*)_{1,2,3}$ are the convection terms of outgoing waves, therefore, $(\mathbf{L}^*)_{1,2,3} = (\mathbf{L}_{\mathrm{SBP}})_{1,2,3}$ and $(\mathbf{L}^*)_4$ is specified using a pressure relaxation term, as in [23].

Figures 10(b) to (d) show the pressure fluctuation contours at various times. The solution consists of the incident pulse and the pulse reflected by the cylinder. The exact solution of pressure fluctuation is given by (see [34])

$$p'(x,y,t) = \mathrm{Re}\left\{ \int_0^\infty \left(A_i\left(x,y,\omega\right) + A_r\left(x,y,\omega\right)\right)\omega e^{-i\omega t}\mathrm{d}\omega \right\}.$$

The contribution of the incident pulse is estimated from

$$A_i\left(x,y,\omega\right) = \frac{1}{2b} e^{-i\omega^2/2b} J_0\left(\omega r_s\right),$$

where $r_s = \sqrt{\left(x-4\right)^2 + y^2}$ and J_0 is the Bessel function of order zero. The reflected pulse contribution is calculated from

$$A_r\left(x,y,\omega\right) = \sum_{k=0}^{\infty} C_k(\omega) H_k^{(1)}(r\omega) cos(k\theta),$$

where $H_k^{(1)}$ is the Hankel function of the first kind of order k, $r = \sqrt{x^2+y^2}$, $\theta = \mathrm{atan2}\left(y,x\right)$, and

$$C_k\left(\omega\right) = \frac{\omega}{2b} e^{-i\omega^2/2b} \frac{\varepsilon_k}{\pi\omega H_k^{(1)}(\omega)} \int_0^\pi J_1(\omega r_{s0}) \frac{1 - 4\cos\theta}{r_{s0}} \cos(k\theta)\, d\theta,$$

where $r_{s0} = r_s|_{r=r_0=1} = \sqrt{\left(\cos\theta - 4\right)^2 + \sin^2\theta}$, $\varepsilon_0 = 1$ and $\varepsilon_k = 2$ for $k \neq 0$.

A comparison of the exact solution with the numerical results from various schemes at different spatial locations is shown in Figure 11. The subfigures in the left column show the time history of pressure fluctuation and the right column shows the respective errors. The spatial locations span different regions of the domain; $x = 2$, $y = 0$ (top subfigures) lies in between the cylinder and the acoustic source, $x = 0$, $y = 5$ (middle subfigures) lies above the cylinder, and $x = -5$, $y = 0$ (bottom figures) lies behind the cylinder with respect to the source. The polar grid shown in Figure 10(a) with an outer radius of 12 and 251 grid points uniformly distributed in the radial and azimuthal directions is used for all simulations. The two peaks in the top and the middle subfigure of Figure 11(a) correspond to the incident and the reflected pulse.

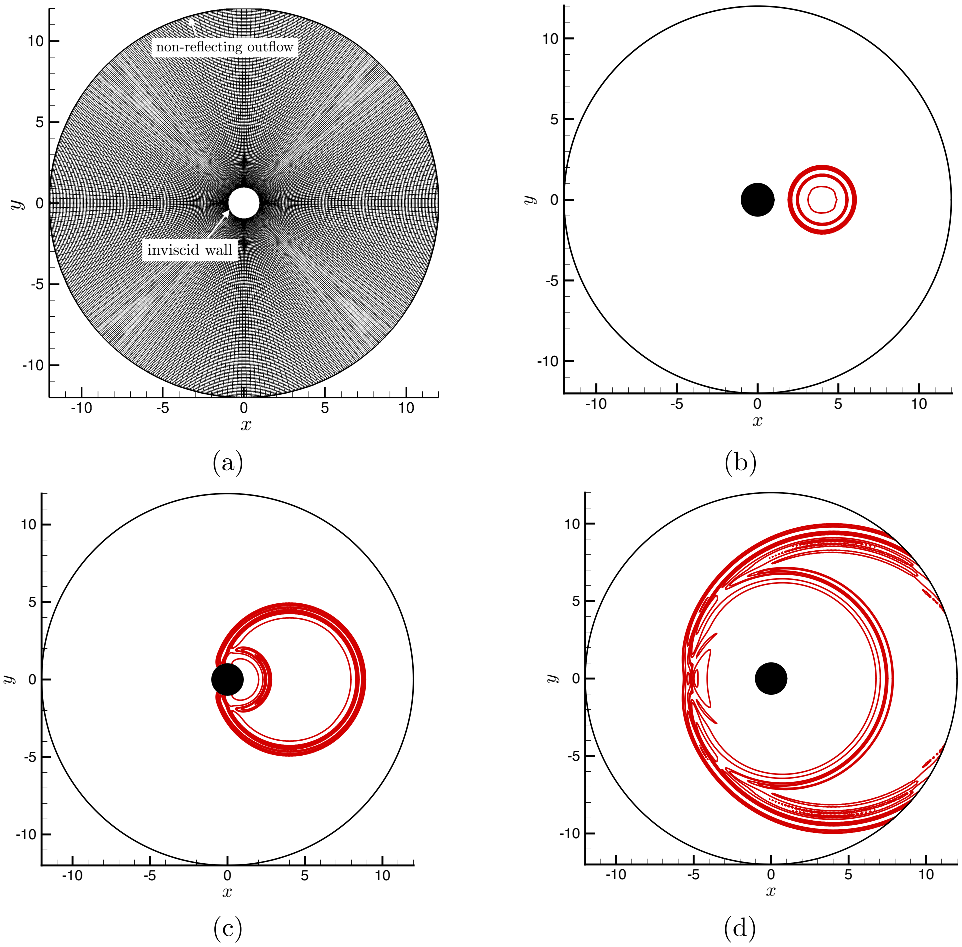


FIG. 10. *Initial pressure-pulse problem: (a) computational grid and boundary conditions, and pressure fluctuation contours at (b) $t \approx 1.5$, (c) $t \approx 4.5$, and (d) $t \approx 9.5$. The contour lines show* 10 *levels in the range* $[-5,\, 5] \times 10^{-6}$.

All EBC schemes of Section 3 are stable for this problem. The error plots show the significance of high-order schemes for acoustic (wave propagation) problems. The second-order scheme has poor dispersion properties and, as a result, highest error among all schemes. The error decreases with increase in order-of-accuracy of the interior scheme, as expected.

6. Conclusions. A systematic approach is developed to derive strongly time-stable high-order finite-difference schemes that enforce boundary conditions strongly for hyperbolic systems. Time-stability and conservation constraints are derived for non-square first-derivative operators that, by construction, exclude calculations at grid points where physical boundary condition is imposed. Schemes of global order-of-accuracy up to fourth-order are derived that show time stability for problems that previously could not be solved for long times with high-order schemes and strong

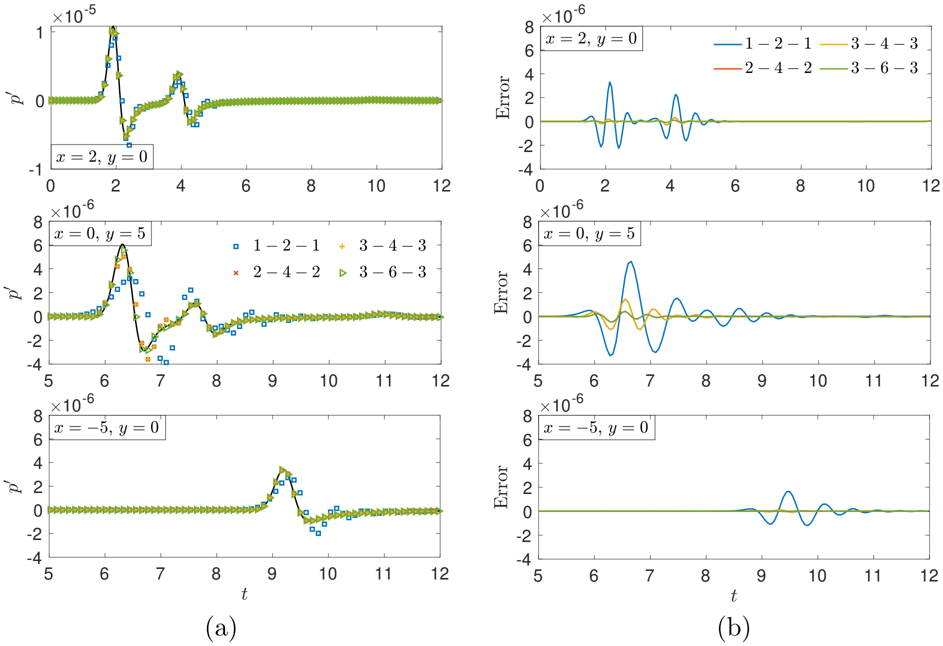


FIG. 11. *Numerical results from various schemes showing time history of (a) pressure fluctuation and (b) pressure-fluctuation error at $x = 2$, $y = 0$ (top), $x = 0$, $y = 5$ (middle) and $x = -5$, $y = 0$ (bottom). The black solid line in subfigures of (a) shows the exact solution. Note the difference in axis scales of the top subfigure in each column. Legend is the same for each subfigure of a column. In subfigures of column (b), the absolute of the maximum error is less than 1.5×10^{-6} for the $2-4-2$ and $3-4-3$ schemes, and less than 4.5×10^{-7} for the $3-6-3$ scheme at all times.*

boundary conditions without additional stability measures, *e.g.* artificial dissipation/filters. The robustness of the proposed method is verified for various problems solving: (a) 1-D scalar advection equation, (b) 1-D coupled hyperbolic system, (c) 1-D inviscid Burgers' equation, (d) 2-D variable-coefficient advection equation, and (e) 2-D Euler equations in curvilinear coordinates.

Appendix A. Additional proofs.

LEMMA A.1. *A square derivative operator, $\hat{D}$, that ensures discrete conservation in solving (1.1) is not conservative after the row omission for strong BC enforcement.*

Proof. Consider the grid function $\mathbf{u}(t) = \begin{bmatrix} u_0(t) & \cdots & u_n(t) \end{bmatrix}^T$ for solving (1.1) over the domain $0 \le x \le 1$ with $n+1$ equidistant grid points. A square $(n+1) \times (n+1)$ derivative operator, $\hat{D}$, typically satisfies the discrete analogue of (1.2) given by

$$\frac{d}{dt}\int_0^1 U\,dx \approx \frac{d}{dt}\sum_{i=0}^{n}\left(\hat{H}\mathbf{u}\right)_i = -\sum_{i=0}^{n}\left(\hat{H}\hat{D}\mathbf{u}\right)_i = u_0(t) - u_n(t), \tag{A.1}$$

where $(\mathbf{v})_i$ denotes the i-th component of a vector $\mathbf{v} = \begin{bmatrix} v_0 & \cdots & v_n \end{bmatrix}^T$ and $\hat{H}$ is a $(n+1) \times (n+1)$ matrix that constitutes a quadrature for the spatial domain. (A.1) implies for the entries $\hat{q}_{ij}$ of $\hat{Q} = \hat{H}\hat{D}$ that

$$\sum_{i=0}^{n} \hat{q}_{ij} = \sum_{i=0}^{n}\sum_{k=0}^{n} \hat{h}_{ik}\hat{d}_{kj} = \begin{cases} -1 & j=0 \\ 1 & j=n \\ 0 & \text{otherwise} \end{cases}, \tag{A.2}$$

where $\hat{h}_{ik}$ and $\hat{d}_{kj}$ denote the entries of $\hat{H}$ and $\hat{D}$, respectively.

To enforce BC strongly, if the first row of $\hat{D}$ is omitted, *i.e.* if $\hat{d}_{kj} = 0$ is assumed for $k = 0$, then (A.2) holds only if $\hat{d}_{0j} = 0$ for all $0 \le j \le n$ in $\hat{D}$. But, if $\hat{D}$ is a valid derivative operator at all grid points, including the boundary points, then $\hat{d}_{0j} \neq 0$ for some values of j. The omission of the first row of $\hat{D}$, therefore, introduces a conservation error at the j-th grid point of $\sum\limits_{i=0}^{n} \hat{h}_{i0}\hat{d}_{0j}$, which is $\mathcal{O}(1)$ at some grid points.

LEMMA A.2. *The rows of a derivative operator D sum to zero, and hence the rows of $Q = HD$ should also sum to zero.*

Proof. The rows of a derivative operator D sum to zero, *i.e.* $D\mathbf{1} = 0$, where $\mathbf{1}$ denotes a vector whose all entries are one. For a symmetric positive definite H, $D = H^{-1}Q$. Hence, $D\mathbf{1} = 0$ implies $H^{-1}Q\mathbf{1} = 0$. Multiplying both sides of $H^{-1}Q\mathbf{1} = 0$ by H, yields $Q\mathbf{1} = 0$ or that the rows of Q sum to zero.

Appendix B. $2-4-2$ stencil.

$$H = \Delta x \,\mathrm{diag}\left(h_{11},\, h_{22},\, h_{33},\, h_{44},\, 1,\, \cdots,1,\, \frac{49}{48},\, \frac{43}{48},\, \frac{59}{48},\, \frac{17}{48}\right),$$

$$D = \frac{1}{\Delta x}\begin{bmatrix} d_{10} & d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & & & & & & \\ d_{20} & d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & & & & & & \\ d_{30} & d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & & & & & & \\ d_{40} & d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & & & & & & \\ & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & & & & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & & & & \\ & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & \\ & & & & & & & \frac{4}{49} & -\frac{32}{49} & 0 & \frac{59}{98} & 0 & -\frac{3}{98} \\ & & & & & & & 0 & \frac{4}{43} & -\frac{59}{86} & 0 & \frac{59}{86} & -\frac{4}{43} \\ & & & & & & & 0 & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & & & & & 0 & 0 & \frac{3}{34} & \frac{4}{17} & -\frac{59}{34} & \frac{24}{17} \end{bmatrix}. \tag{B.1}$$

$h_{11} = 1.117853598033634$	$h_{22} = 1.734954607723689$	$h_{33} = 0.493492831348563$	$h_{44} = 1.153698962894113$
$d_{10} = -0.558055563977424$	$d_{20} = -0.177806646597481$	$d_{30} = 0.197577181565075$	$d_{40} = 0.053103321910167$
$d_{11} = 0.206193447640676$	$d_{21} = -0.148032843241780$	$d_{31} = -0.349146497048670$	$d_{41} = 0.031031686127352$
$d_{12} = 0.229753040942520$	$d_{22} = 0.010938409310223$	$d_{32} = -0.469159274307636$	$d_{42} = -0.272872172147738$
$d_{13} = 0.154135831102631$	$d_{23} = 0.133448297494816$	$d_{33} = 0.026584989564182$	$d_{43} = -0.326375382961636$
$d_{14} = -0.032026755708402$	$d_{24} = 0.181452783034222$	$d_{34} = 0.763007924163851$	$d_{44} = 0.009492491845307$
$d_{15} = 0$	$d_{25} = 0$	$d_{35} = -0.168864323936802$	$d_{45} = 0.577851491687484$
$d_{16} = 0$	$d_{26} = 0$	$d_{36} = 0$	$d_{46} = -0.072231436460936$

Acknowledgments. We gratefully acknowledge discussions with Dr. Harsha Nagarajan on optimization problems and tools.

REFERENCES

[1] S. S. Abarbanel, A. E. Chertock, and A. Yefet, *Strict stability of high-order compact implicit finite-difference schemes: the role of boundary conditions for hyperbolic PDEs, II*, Journal of Computational Physics, 160 (2000), pp. 67–87.

[2] E. Albin, Y. d'Angelo, and L. Vervisch, *Flow streamline based navier–stokes characteristic boundary conditions: modeling for transverse and corner outflows*, Computers & Fluids, 51 (2011), pp. 115–126.

[3] A. Baskaran, J. S. Lowengrub, C. Wang, and S. M. Wise, *Convergence analysis of a second order convex splitting scheme for the modified phase field crystal equation*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 2851–2873.

[4] P. T. Brady and D. Livescu, *High-order, stable, and conservative boundary schemes for central and compact finite differences*, Computers & Fluids, 183 (2019), pp. 84–101.

[5] M. H. Carpenter, D. Gottlieb, and S. Abarbanel, *The stability of numerical boundary treatments for compact high-order finite-difference schemes*, Journal of Computational Physics, 108 (1993), pp. 272–295.

[6] M. H. Carpenter, D. Gottlieb, and S. Abarbanel, *Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes*, Journal of Computational Physics, 111 (1994), pp. 220–236.

[7] K. Cheng, W. Feng, S. Gottlieb, and C. Wang, *A fourier pseudospectral method for the "good" boussinesq equation with second-order temporal accuracy*, Numerical Methods for Partial Differential Equations, 31 (2015), pp. 202–224.

[8] T. Colonius and S. K. Lele, *Computational aeroacoustics: progress on nonlinear problems of sound generation*, Progress in Aerospace Sciences, 40 (2004), pp. 345–416.

[9] A. W. Cook and J. J. Riley, *Direct numerical simulation of a turbulent reactive plume on a parallel computer*, Journal of Computational Physics, 129 (1996), pp. 263–283.

[10] M. J. Corless and A. Frazho, *Linear systems and control: an operator perspective*, CRC Press, 2003.

[11] W. De Roeck, W. Desmet, M. Baelmans, and P. Sas, *An overview of high-order finite difference schemes for computational aeroacoustics*, in Proceedings of the International Conference on Noise and Vibration Engineering, Citeseer, 2004, pp. 353–368.

[12] D. C. D. R. Fernández, J. E. Hicken, and D. W. Zingg, *Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations*, Computers & Fluids, 95 (2014), pp. 171–196.

[13] B. Gustafsson, *The convergence rate for difference approximations to mixed initial boundary value problems*, Mathematics of Computation, 29 (1975), pp. 396–406.

[14] B. Gustafsson, *The convergence rate for difference approximations to general mixed initial-boundary value problems*, SIAM Journal on Numerical Analysis, 18 (1981), pp. 179–190.

[15] B. Gustafsson, *High order difference methods for time dependent PDE*, vol. 38, Springer Science & Business Media, 2007.

[16] W. R. Inc., *Mathematica, Version 12.2*, https://www.wolfram.com/mathematica. Champaign, IL, 2020.

[17] J. W. Kim and D. J. Lee, *Generalized characteristic boundary conditions for computational aeroacoustics, part 2*, AIAA Journal, 42 (2004), pp. 47–55.

[18] H.-O. Kreiss and G. Scherer, *On the existence of energy estimates for difference approximations for hyperbolic systems*, tech. report, Technical report, Dept. of Scientific Computing, Uppsala University, 1977.

[19] M. Mohammadi-Aragh, K. Klingbeil, N. Brüggemann, C. Eden, and H. Burchard, *The impact of advection schemes on restratifiction due to lateral shear and baroclinic instabilities*, Ocean Modelling, 94 (2015), pp. 112–127.

[20] Y. Morinishi, T. S. Lund, O. V. Vasilyev, and P. Moin, *Fully conservative higher order finite difference schemes for incompressible flow*, Journal of Computational Physics, 143 (1998), pp. 90–124.

[21] P. Olsson, *Summation by parts, projections, and stability. i*, Mathematics of Computation, 64 (1995), pp. 1035–1065.

[22] P. Olsson, *Summation by parts, projections, and stability. ii*, Mathematics of Computation, 64 (1995), pp. 1473–1493.

[23] T. J. Poinsot and S. Lele, *Boundary conditions for direct simulations of compressible viscous flows*, Journal of Computational Physics, 101 (1992), pp. 104–129.

[24] T. H. Pulliam and D. Chaussee, *A diagonal form of an implicit approximate-factorization algorithm*, Journal of Computational Physics, 39 (1981), pp. 347–363.

[25] M. M. Rai and P. Moin, *Direct simulations of turbulent flow using finite-difference schemes*, Journal of Computational Physics, 96 (1991), pp. 15–53.

[26] J. Ryu and D. Livescu, *Turbulence structure behind the shock in canonical shock–vortical*

turbulence interaction, Journal of Fluid Mechanics, 756 (2014).

[27] K. Salari and P. Knupp, *Code verification by the method of manufactured solutions*, tech. report, Sandia National Labs., Albuquerque, NM (US), 2000.

[28] N. Sharan, *Time-stable high-order finite difference methods for overset grids*, PhD thesis, University of Illinois at Urbana-Champaign, 2016.

[29] N. Sharan, G. Matheou, and P. E. Dimotakis, *Mixing, scalar boundedness, and numerical dissipation in large-eddy simulations*, Journal of Computational Physics, 369 (2018), pp. 148–172.

[30] N. Sharan, G. Matheou, and P. E. Dimotakis, *Turbulent shear-layer mixing: initial conditions, and direct-numerical and large-eddy simulations*, Journal of Fluid Mechanics, 877 (2019), pp. 35–81.

[31] N. Sharan, C. Pantano, and D. J. Bodony, *Time-stable overset grid method for hyperbolic problems using summation-by-parts operators*, Journal of Computational Physics, 361 (2018), pp. 199–230.

[32] B. Strand, *Summation by parts for finite difference approximations for d/dx*, Journal of Computational Physics, 110 (1994), pp. 47–67.

[33] C. K. Tam and Z. Dong, *Wall boundary conditions for high-order finite-difference schemes in computational aeroacoustics*, Theoretical and Computational Fluid Dynamics, 6 (1994), pp. 303–322.

[34] C. K. Tam and F. Q. Hu, *An optimized multi-dimensional interpolation scheme for computational aeroacoustics applications using overset grids*, AIAA Paper, 2812 (2004), p. 2004.

[35] L. N. Trefethen, *Stability of finite-difference models containing two boundaries or interfaces*, Mathematics of Computation, 45 (1985), pp. 279–300.

[36] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Mathematical Programming, 106 (2006), pp. 25–57.

[37] C. Wang and S. M. Wise, *An energy stable and convergent finite-difference scheme for the modified phase field crystal equation*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 945–969.

[38] C. Zhang, H. Wang, J. Huang, C. Wang, and X. Yue, *A second order operator splitting numerical scheme for the "good" boussinesq equation*, Applied Numerical Mathematics, 119 (2017), pp. 179–193.

[39] M. Zhuang and R. Chen, *Optimized upwind dispersion-relation-preserving finite difference scheme for computational aeroacoustics*, AIAA Journal, 36 (1998), pp. 2146–2148.