S©MR

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ
Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru
ISSN 1813-3304

Vol. 20, No. 2, pp. 144–145 (2023)
https://doi.org/10.33048/semi.2023.20.???

УДК 519.63
MSC 65M20

SUMMATION-BY-PARTS SCHEMES FOR SYMMETRIC HYPERBOLIC SYSTEMS

ALEXANDER MALYSHEV

Communicated by P.P. Petrov

Abstract: We apply the method of lines to numerically solve general initial-boundary value problems for symmetric hyperbolic systems of linear differential equations with variable coefficients. Semi-discretization of symmetric hyperbolic systems is performed using classical summation-by-parts difference operators. Strictly dissipative boundary conditions are weakly enforced using the so-called simultaneous approximation terms. All theoretical constructions are provided with full proofs. The stability of explicit Runge-Kutta methods for semi-bounded operators is proved using recent results on strong stability for semi-dissipative operators.

Keywords: symmetric hyperbolic system, dissipative boundary conditions, summation-by-parts scheme, simultaneous approximation terms, strong stability of explicit Runge-Kutta methods.

1 Introduction

We begin our study of the Summation-By-Parts (SBP) difference operators and enforcement of boundary conditions with the Simultaneous Approximation Terms (SAT) with an example demonstrating how the SBP-SAT method extends energy estimates obtained for partial differential equations to semi-discrete approximations of the differential equations with little effort.

Malyshev, A.N., Summation-by-parts schemes for symmetric hyperbolic systems.

Received January, 1, 2023, Published December, 31, 2023.

1.1. The energy identity for the linear advection equation. Solution $u(t,x)$ of the initial-boundary value problem

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}=f(t,x),\quad 0<x<1,\ 0<t<T, \tag{1}$$

$$u(0,x)=u_0(x),\quad u(t,0)=g(t), \tag{2}$$

satisfies the equations $(u^2)_t=u_tu+uu_t=-u_xu+fu-uu_x+uf$ and

$$\frac{d}{dt}\int_0^1 u^2dx=-\int_0^1(u_xu+uu_x)dx+2\int_0^1 fudx.$$

Integration by parts ensures the identity

$$\int_0^1(u_xu+uu_x)dx=u^2(1)-u^2(0), \tag{3}$$

which allows us to derive the *energy identity*

$$\frac{d}{dt}\int_0^1 u^2(t,x)dx=u^2(t,0)-u^2(t,1)+2\int_0^1 f(t,x)u(t,x)dx. \tag{4}$$

We split the solution $u(t,x)$ into the sum $u=u_1+u_2$, where $u_1(t,x)$ satisfies (1), (2) with $g(t)\equiv 0$ while $u_2(t,x)$ satisfies the system (1), (2) with $f(t,x)\equiv 0$ and $u_0(x)\equiv 0$. It follows from (4) that

$$\begin{aligned}\frac{d}{dt}\int_0^1 u_1^2(t,x)dx&\le 2\int_0^1 f(t,x)u_1(t,x)dx\\&\le 2\left(\int_0^1 f^2(t,x)dx\right)^{\frac12}\left(\int_0^1 u_1^2(t,x)dx\right)^{\frac12}\end{aligned}$$

and, for $0\le t\le T$,

$$\left(\int_0^1 u_1^2(t,x)dx\right)^{\frac12}\le\left(\int_0^1 u_1^2(0,x)dx\right)^{\frac12}+t\max_{0\le s\le T}\left(\int_0^1 f^2(s,x)dx\right)^{\frac12}. \tag{5}$$

Similarly, (4) implies the estimates

$$\frac{d}{dt}\int_0^1 u_2^2(t,x)dx\le g^2(t),\quad \int_0^1 u_2^2(t,x)dx\le\int_0^t g^2(t)dt. \tag{6}$$

Combining the estimates (5) and (6), we arrive at the *energy estimate* for the solution of (1), (2),

$$\begin{aligned}\left(\int_0^1 u^2(t,x)dx\right)^{\frac12}&\le\left(\int_0^1 u_1^2(t,x)dx\right)^{\frac12}+\left(\int_0^1 u_2^2(t,x)dx\right)^{\frac12}\\&\le\left(\int_0^1 u_0^2(x)dx\right)^{\frac12}\\&\quad+\left(\int_0^t g^2(t)dt\right)^{\frac12}+t\max_{0\le s\le T}\left(\int_0^1 f^2(s,x)dx\right)^{\frac12}.\end{aligned} \tag{7}$$

1.2. Semi-discretization of the advection equation. Our goal is to reproduce all the derivations of Section 1.1 for an appropriate discrete analogue of (1). Discretization of (1) is carried out only along the x axis, and here we do not use discretization along the variable t. The resulting system of ordinary differential equations in t is referred to as a *semi-discretization*. The combination of semi-discretization with a subsequent time integration by Runge-Kutta methods or by multi-step methods is often called the *method of lines*. In what follows, we consider semi-discrete functions $u(t, x_j) = u_j(t)$ defined at the grid nodes $x_j = jh$, $j = 0, 1, \dots, N$, $h = 1/N$.

Finite difference approximations of the first derivative of a discrete function $V = \begin{bmatrix} v_0 & v_1 & \dots & v_N \end{bmatrix}^T$ are conveniently represented by the product $\mathsf{D}V$ with a square matrix D, which is often called the *differentiation matrix*. We choose the differentiation matrix in the factorized form $\mathsf{D} = P^{-1}Q$ with a diagonal matrix P and an "almost" skew-symmetric matrix Q as

$$P = \begin{bmatrix} \frac{h}{2} & & & & & \\ & h & & & & \\ & & h & & & \\ & & & \ddots & & \\ & & & & h & \\ & & & & & \frac{h}{2} \end{bmatrix}, \quad Q = \begin{bmatrix} -1/2 & 1/2 & & & & \\ -1/2 & 0 & 1/2 & & & \\ & -1/2 & 0 & 1/2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1/2 & 0 & 1/2 \\ & & & & -1/2 & 1/2 \end{bmatrix}. \tag{8}$$

The semi-discretization of (1) with initial value u_0 but without boundary conditions is as follows:

$$\frac{d}{dt}V(t) + P^{-1}QV(t) = f_h(t), \quad V(0) = (u_0)_h. \tag{9}$$

The subscript h indicates the discretization of the functions f and u_0 of the continuous argument x at the grid points x_j.

The boundary condition $u(t, 0) = g(t)$ is imposed by means of the so-called *Simultaneous Approximation Terms (SAT)* method [3], which enforces boundary conditions in the semi-discretized differential equations as penalty terms. The final semi-discretization is

$$\frac{d}{dt}U(t) + P^{-1}QU(t) = f_h(t) + \mathbb{SAT}, \quad U(0) = (u_0)_h \tag{10}$$

with the SAT term, depending on a suitable scalar σ,

$$\mathbb{SAT} = -\frac{\sigma}{2}P^{-1}\begin{bmatrix} U_0(t) - g(t) \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{11}$$

1.3. Convergence analysis for the semi-discrete approximation of the advection equation. Let us denote the semi-discretization of the exact solution $u(t, x)$ of (1), (2), and its derivative $u_x(t, x)$ by the vector functions u_h and $(u_x)_h$ respectively. Then, obviously, $(u_t)_h = (u_h)_t$ and

$(u_h)_t + (u_x)_h = f_h$. The standard argument based on Taylor expansions proves the consistency equation

$$P^{-1}Qu_h = (u_x)_h + T_h \tag{12}$$

with the truncation error

$$T_h = [O(h), O(h^2), \ldots, O(h^2), O(h)]^T. \tag{13}$$

As a result, we arrive at the equation

$$\frac{d}{dt}u_h + P^{-1}Qu_h = f_h + T_h. \tag{14}$$

In addition, $(u_h)_0 = g$ owing to the boundary condition.

The error function $\mathscr{E}(t) = U(t) - u_h(t) = \begin{bmatrix}\mathscr{E}_0 & \mathscr{E}_1 & \ldots & \mathscr{E}_N\end{bmatrix}^T$ for $U(t)$ satisfying the semi-discretization (10), (11) solves the system

$$\frac{d}{dt}\mathscr{E} + P^{-1}Q\mathscr{E} = -T_h - \frac{\sigma}{2}P^{-1}\begin{bmatrix}\mathscr{E}_0 & 0 & \ldots & 0\end{bmatrix}^T. \tag{15}$$

In the formulas below we use the weighted inner product $(U,V)_P = U^*PV$, the Euclidean inner product $(U,V) = U^*V$ and the weighted norm $\|U\|_P = \sqrt{(U,U)_P}$. We derive from (15) that

$$\begin{aligned}
\frac{d}{dt}(\mathscr{E},\mathscr{E})_P &= (\mathscr{E}_t,\mathscr{E})_P + (\mathscr{E},\mathscr{E}_t)_P \\
&= -(P^{-1}Q\mathscr{E},\mathscr{E})_P - (\mathscr{E},P^{-1}Q\mathscr{E})_P - 2(T_h,\mathscr{E})_P - \sigma\mathscr{E}_0^2 \\
&= -(Q\mathscr{E},\mathscr{E}) - (\mathscr{E},Q\mathscr{E}) - 2(T_h,\mathscr{E})_P - \sigma\mathscr{E}_0^2 \\
&= -((Q+Q^T)\mathscr{E},\mathscr{E}) - 2(T_h,\mathscr{E})_P - \sigma\mathscr{E}_0^2 \\
&= \mathscr{E}_0^2 - \mathscr{E}_N^2 - 2(T_h,\mathscr{E})_P - \sigma\mathscr{E}_0^2 \\
&= \mathscr{E}_0^2(1-\sigma) - \mathscr{E}_N^2 - 2(T_h,\mathscr{E})_P.
\end{aligned}$$

Thus we arrive at the energy equality for the error $\mathscr{E}(t)$,

$$\frac{d}{dt}(\mathscr{E},\mathscr{E})_P = \mathscr{E}_0^2(1-\sigma) - \mathscr{E}_N^2 - 2(T_h,\mathscr{E})_P. \tag{16}$$

It follows from (16) that $\frac{d}{dt}(\mathscr{E},\mathscr{E})_P \le 2|(T_h,\mathscr{E})_P|$, when $\sigma \ge 1$. Furthermore, $\frac{d}{dt}\|\mathscr{E}\|_P^2 \le 2\|T_h\|_P\|\mathscr{E}\|_P$, $\frac{d}{dt}\|\mathscr{E}\|_P \le \|T_h\|_P$ and $\|\mathscr{E}(t)\|_P \le \|\mathscr{E}(0)\|_P + t\|T_h\|_P$. Note that $\|T_h\|_P = O(h^{3/2})$. Taking into account $\mathscr{E}(0) = 0$, we obtain the estimate of the convergence rate for the scheme (10), (11):

$$\|\mathscr{E}(t)\|_P \le O(h^{3/2}) \text{ if } \sigma \ge 1. \tag{17}$$

Remark 1. *A more refined proof shows that the error $\|\mathscr{E}(t)\|_P$ is bounded from above by $O(h^2)$; see e.g.* [26].

1.4. Summation-By-Parts (SBP) difference operator. The key component in deriving the energy identity (4) is the integration-by-parts rule (3). A convenient discrete analogue of (3) would allow us to convert the derivations of the energy identities for partial differential equations into the semi-discrete case with relative ease. For a differentiation matrix D and discrete functions $V = \begin{bmatrix} v_0 & v_1 & \dots & v_N \end{bmatrix}^T$, the most convenient discrete analogue of (3) might be the identity $(\mathsf{D}V, V)+(V, \mathsf{D}V) = (\mathsf{E}V, V)$, where $\mathsf{E} = \mathrm{diag}[-1, 0, 0, \dots, 0, 1]$ and $(\cdot, \cdot)$ are suitable inner products. This approach indeed works, and we describe it in detail in the present paper.

A differentiation matrix for approximation of the first derivative is sought in the form $\mathsf{D} = P^{-1}Q$, where the matrix P is symmetric positive definite and the matrix Q satisfies the SBP property

$$Q + Q^T = \mathsf{E} = \mathrm{diag}[-1, 0, 0, \dots, 0, 1]. \tag{18}$$

Such a differentiation matrix D is referred to as the *Summation-By-Parts (SBP)* operator. The corresponding discrete analogue of the integration-by-parts has the form

$$(P^{-1}QU, V)_P + (U, P^{-1}QV)_P = (\mathsf{E}U, V), \tag{19}$$

where $(\cdot, \cdot)$ is the Euclidean inner product and $(\cdot, \cdot)_P$ is the energy inner product $(U, V)_P = (PU, V)$.

Remark 2. *Suppose that $a = x_0 < x_1 < \dots < x_{s_x} = b$ is an arbitrary grid and consider the vectors $\mathbf{x}^j = \begin{bmatrix} x_0^j & x_1^j & \dots & x_{s_x}^j \end{bmatrix}^T$. The system of equations $P^{-1}Q\mathbf{x}^j = j\mathbf{x}^{j-1}$ for $0 \le j \le \tau$ defines accuracy of order τ for the SBP operator $P^{-1}Q$. It is equivalent to $Q\mathbf{x}^j = jP\mathbf{x}^{j-1}$. Multiplying the latter equation from the left by $(\mathbf{x}^i)^T$ leads to the system $(\mathbf{x}^i)^T Q\mathbf{x}^j = j(\mathbf{x}^i)^T P\mathbf{x}^{j-1}$ for $0 \le i, j \le \tau$. Now we derive that*

$$(\mathbf{x}^i)^T E\mathbf{x}^j = (\mathbf{x}^i)^T Q\mathbf{x}^j + (\mathbf{x}^i)^T Q^T\mathbf{x}^j = j(\mathbf{x}^i)^T P\mathbf{x}^{j-1} + i(\mathbf{x}^j)^T P\mathbf{x}^{i-1}.$$

Let $P = \mathrm{diag}\,[p_0, p_1, \dots, p_{s_x}]$. We arrive at the identity

$$\sum_{k=0}^{s_x} p_k x_k^{i+j-1} = \frac{b^{i+j} - a^{i+j}}{i+j} \tag{20}$$

valid for all i, j such that $0 \le i + j \le 2\tau$. The identity (20) proves that the diagonals p_k are the weights of a quadrature rule of accuracy $2\tau - 1$ on the interval $[a, b]$ at the given nodes x_k. More about SBP on general grids is found in [5].

We use only SBP operators with diagonal matrices P because those with full matrices P are not applicable to differential operators with variable coefficients. Note that the diagonal matrix P in the SBP operator for the first derivative on an equidistant grid with step size h is equal to a constant diagonal matrix multiplied by h.

SBP operators on equidistant grids are called the classical SBP operators. They were introduced and constructed in [14, 15, 21]. More recent descriptions of classical SBP operators can be found in [23, 10]. Modern reviews of the SBP-SAT theory and applications with extensive bibliography are [25, 6]. SBP operators for general function spaces are proposed in [8].

2 Classical SBP operators

Theorem 1. *The skew-symmetric difference operator on the equidistant mesh $x_\nu = \nu h$, $\nu \in \mathbb{N}$,*

$$v(x_n) = \frac{1}{h}\left[\sum_{\nu=1}^{s} \alpha_\nu u(x_{n+\nu}) - \sum_{\nu=1}^{s} \alpha_\nu u(x_{n-\nu})\right] \tag{21}$$

approximates the first derivative $u'(x_n)$ with order $2s$ if and only if

$$\sum_{\nu=1}^{s} \alpha_\nu \nu^{2k+1} = \begin{cases} \frac{1}{2}, & k = 0, \\ 0, & k = 1, 2, \dots, s-1. \end{cases} \tag{22}$$

The coefficients α_ν determined by (22) equal

$$\alpha_\nu = \frac{(-1)^{\nu-1}(s!)^2}{\nu(s+\nu)!(s-\nu)!}, \quad \nu = 1, 2, \dots, s. \tag{23}$$

Proof. Using the Taylor expansion $u(x_{n+\nu}) = \sum_{j=0}^{\infty} u^{(j)}(x_n)\frac{(\nu h)^j}{j!}$ we obtain the equalities

$$\begin{aligned} v(x_n) &= \frac{1}{h}\left[\sum_{\nu=1}^{s} \alpha_\nu \sum_{j=0}^{\infty} u^{(j)}(x_n)\frac{(\nu h)^j}{j!} - \sum_{\nu=1}^{s} \alpha_\nu \sum_{j=0}^{\infty} u^{(j)}(x_n)\frac{(-\nu h)^j}{j!}\right] \\ &= \frac{1}{h}\sum_{j=0}^{\infty} u^{(j)}(x_n)\frac{h^j}{j!}[1-(-1)^j]\sum_{\nu=1}^{s} \alpha_\nu \nu^j = u^{(1)}(x_n) + O(h^{2s}), \end{aligned}$$

where the last equality holds because $v(x_n)$ approximates $u'(x_n)$ with order $2s$. Hence the coefficients α_ν satisfy the system (22).

The system (22) is the Vandermonde system

$$V\begin{bmatrix} \alpha_1 \\ 2\alpha_2 \\ \vdots \\ s\alpha_s \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1^2 & 2^2 & \dots & s^2 \\ \vdots & \vdots & \ddots & \vdots \\ (1^2)^{s-1} & (2^2)^{s-1} & \dots & (s^2)^{s-1} \end{bmatrix}\begin{bmatrix} \alpha_1 \\ 2\alpha_2 \\ \vdots \\ s\alpha_s \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Cramer's rule gives the solution $\nu\alpha_\nu = \frac{1}{2}(-1)^{\nu-1}\det V_\nu / \det V$, where the corresponding submatrix

$$V_\nu = \begin{bmatrix} 1^2 & \dots & (\nu-1)^2 & (\nu+1)^2 & \dots & s^2 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 1^{2(s-1)} & \dots & (\nu-1)^{2(s-1)} & (\nu+1)^{2(s-1)} & \dots & s^{2(s-1)} \end{bmatrix}$$

is again of the Vandermonde type. The formulas of the Vandermonde determinants [12] are $\det V = \Pi_{1\le i<j\le s}(j^2-i^2)$ and

$$\begin{aligned}\det V_\nu &= \Pi_{k=1,k\ne\nu}^{s} k^2 \Pi_{1\le i<j\le s,\, i\ne\nu, j\ne\nu}(j^2-i^2)\\ &= \frac{(s!)^2}{\nu^2}\frac{\det V}{\Pi_{1\le i<\nu}(\nu^2-i^2)\Pi_{\nu<j\le s}(j^2-\nu^2)} = 2\frac{(s!)^2\det V}{(s-\nu)!(s+\nu)!}.\end{aligned}$$

□

We look for a finite difference operator approximating the first derivative on a semi-infinite mesh x_ν, $\nu = 0, 1, \ldots$, as a block-partitioned semi-infinite matrix

$$\frac{1}{h}\begin{bmatrix} H^{-1}B & H^{-1}C \\ -C^T & D \end{bmatrix} \tag{24}$$

having blocks

$$C = \begin{bmatrix} C_0 & 0 & \cdots \\ C_s & 0 & \cdots \end{bmatrix}, \quad C_s = \begin{bmatrix} \alpha_s & 0 & \cdots & \cdots & 0 \\ \alpha_{s-1} & \alpha_s & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \alpha_1 & \cdots & \cdots & \alpha_{s-1} & \alpha_s \end{bmatrix},$$

$$D = \begin{bmatrix} 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots & & & 0 & \cdots \\ -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots & & & \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & & & \\ -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & \ddots & & \vdots \\ 0 & -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots \\ \vdots & \ddots & \ddots & & \ddots & \ddots & \ddots & & \ddots & \ddots \end{bmatrix}.$$

The matrix C_0 is the zero $(r-s)\times s$ matrix. The coefficients α_ν in C_s and in the semi-infinite matrix D are defined in Theorem 1. The block D provides $2s$-order accurate approximation of the first derivative at the inner grid points x_ν, $\nu \ge r$.

We assume that $r \ge s$ and define the structure of the real $r\times r$ matrix B by the following conditions:

$$B = B_1 + B_2, \quad B_1 = \mathrm{diag}(-1/2, 0, \ldots, 0), \quad B_2^T = -B_2. \tag{25}$$

We want to find a symmetric positive definite $r\times r$ matrix H and an "almost" skew-symmetric matrix B of the form (25) such that for arbitrary smooth functions $u(x)$,

$$H^{-1}B\begin{bmatrix} u(x_0) \\ u(x_1) \\ \vdots \\ u(x_{r-1}) \end{bmatrix} + H^{-1}\begin{bmatrix} C_0 \\ C_s \end{bmatrix}\begin{bmatrix} u(x_r) \\ u(x_{r+1}) \\ \vdots \\ u(x_{r+s-1}) \end{bmatrix} = h\begin{bmatrix} u_x(x_0) \\ u_x(x_1) \\ \vdots \\ u_x(x_{r-1}) \end{bmatrix} + \mathcal{O}(h^{\tau+1}). \tag{26}$$

Condition (26) provides approximation of order τ for the first derivative at the grid nodes x_ν, $0 \le \nu < r$, near the boundary. To satisfy (26), it suffices to check it only for monomial functions $u(x) = (x - x_r)^j$. It follows that condition (26) is equivalent to the algebraic system

$$(-1)^j H^{-1} B \mathbf{x}^j + (-1)^j H^{-1} \begin{bmatrix} C_0 \\ C_s \end{bmatrix} \mathbf{y}^j = (-1)^{j-1} j \mathbf{x}^{j-1}, \quad j = 0, 1, \ldots, \tau, \tag{27}$$

where we denote

$$\mathbf{x}^j = \begin{bmatrix} r^j & (r-1)^j & \ldots & 1^j \end{bmatrix}^T, \quad \mathbf{y}^j = \begin{bmatrix} 0^j & (-1)^j & \ldots & (1-s)^j \end{bmatrix}^T$$

and agree on that $0^0 = 1$ and $\mathbf{x}^{-1} = 0$. Multiplying (27) from the left by $H(-1)^j$ and using (25) gives the system

$$B_2 \mathbf{x}^j + B_1 \mathbf{x}^j + \begin{bmatrix} C_0 \\ C_s \end{bmatrix} \mathbf{y}^j = -jH\mathbf{x}^{j-1}, \quad j = 0, 1, \ldots, \tau.$$

As a result, we arrive at the following equations for $j = 0, 1, \ldots, \tau$

$$B_2 \mathbf{x}^j = \mathbf{g}_j, \text{ where } \mathbf{g}_j = -B_1 \mathbf{x}^j - \begin{bmatrix} C_0 \\ C_s \end{bmatrix} \mathbf{y}^j - jH\mathbf{x}^{j-1}, \tag{28}$$

or in matrix form

$$B_2 X = G, \tag{29}$$

where $X = \begin{bmatrix} \mathbf{x}^0 & \mathbf{x}^1 & \cdots & \mathbf{v}^\tau \end{bmatrix}$, $G = \begin{bmatrix} \mathbf{g}_0 & \mathbf{g}_1 & \cdots & \mathbf{g}_\tau \end{bmatrix}$.

2.1. System of linear equations defining the diagonal matrix H.

Theorem 2. *Under the restriction $\tau + 1 \le r$, a necessary and sufficient condition for existence of a skew-symmetric solution B_2 of (29) is the skew-symmetry of the matrix $X^T G$.*

Proof. The restriction $\tau + 1 \le r$ is needed for existence of a QR factorization $X = QR$, where $R = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}$ and $\hat{R}$ is a square non-singular matrix. The matrix X is a Vandermonde matrix, therefore, $\hat{R}$ is non-singular. Multiplying the system (29) by Q^T from the left and by $\hat{R}^{-1}$ from the right gives the equation

$$(Q^T B_2 Q) \begin{bmatrix} I \\ 0 \end{bmatrix} = Q^T G \hat{R}^{-1}. \tag{30}$$

Let us partition the skew-symmetric matrix $\Theta = Q^T B_2 Q$ as

$$\Theta = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix},$$

where the block Θ_{11} is of size $(\tau+1) \times (\tau+1)$. The block of columns $0, 1, \ldots, \tau$ in Θ equals

$$\begin{bmatrix} \Theta_{11} \\ \Theta_{12} \end{bmatrix} = Q^T G \hat{R}^{-1} \tag{31}$$

due to equation (30). The upper square block $\Theta_{11} = \begin{bmatrix} I & 0 \end{bmatrix} Q^T G \hat{R}^{-1}$ of the skew-symmetric matrix Θ must be skew-symmetric. Hence the product $\hat{R}^T \Theta_{11} \hat{R} = X^T G$ is also skew-symmetric, and the necessity is proved.

Sufficiency is proved by using the QR factorization of X again. The block matrix $\begin{bmatrix} \Theta_{11} \\ \Theta_{12} \end{bmatrix}$ is constructed as in (31), where the skew-symmetry of Θ_{11} is guaranteed by the skew-symmetry of $X^T G$. The block Θ_{12} is constructed skew-symmetrically as $\Theta_{12} = -\Theta_{21}^T$. The remaining block Θ_{22} is chosen arbitrarily but it must be skew-symmetric. Therefore, Θ_{22} is not unique if its order is larger than 1. □

Skew-symmetry of $X^T G$ is equivalent to the system of equations

$$(\mathbf{x}^i)^T g_j + g_i^T \mathbf{x}^j = 0, \quad i, j = 0, 1, \ldots, \tau. \tag{32}$$

Due to the structure of the vectors g_i in (28) and structure of B_1, we have

$$(\mathbf{x}^i)^T \mathbf{g}_j = \frac{1}{2} r^{i+j} - (\mathbf{x}^i)^T \begin{bmatrix} C_0 \\ C_s \end{bmatrix} \mathbf{y}^j - j (\mathbf{x}^i)^T H \mathbf{x}^{j-1}.$$

Thus, the system (32) is equivalent to the system

$$j (\mathbf{x}^i)^T H \mathbf{x}^{j-1} + i (\mathbf{x}^j)^T H \mathbf{x}^{i-1} = r^{i+j} - \xi_{ij} - \xi_{ji}, \tag{33}$$

where ξ_{ij} are the entries of the matrix Ξ of order $\tau+1$ defined by the formula

$$\xi_{ij} = \begin{bmatrix} s^i & (s-1)^i & \ldots & 1^i \end{bmatrix} C_s \begin{bmatrix} 0^j & (-1)^j & \ldots & (1-s)^j \end{bmatrix}^T. \tag{34}$$

Recall that we are interested only in the diagonal matrices H. Let us denote the diagonal entries of H by the column vector $\mathbf{h}$. In this case, $(\mathbf{x}^i)^T H \mathbf{x}^{j-1} = (\mathbf{x}^{i+j-1})^T \mathbf{h}$ and $(\mathbf{x}^j)^T H \mathbf{x}^{i-1} = (\mathbf{x}^{i+j-1})^T \mathbf{h}$ in (33). The following linear system defining the desired vector $\mathbf{h}$ is obtained from (33) and is equivalent to (32):

$$(i+j)(\mathbf{x}^T)^{i+j-1} \mathbf{h} = r^{i+j} - \xi_{ij} - \xi_{ji}, \quad i, j = 0, 1, \ldots, \tau. \tag{35}$$

The system (35) is overdetermined. The left-hand side of (35) is a function of $i+j$, so the right-hand side of (35) must be the same. In other words, the sums $\xi_{ij} + \xi_{ji}$ must be equal for the same $i+j$. A matrix is said to have the Hankel structure if the entries on each anti-diagonal are equal. Thus, for (35) to be consistent, it is necessary that the matrix $\Xi + \Xi^T$ be a Hankel matrix. The equation for $i = j = 0$ in (35) is $0 = 1 - 2\xi_{00}$. Since this equation does not define $\mathbf{h}$, the equality $\xi_{00} = 1/2$ is also a consistency condition.

The number of anti-diagonals in $\Xi + \Xi^T$, excluding the one with $i+j = 0$, is 2τ. It is equal to the number of independent equations in (35) provided that all consistency conditions are satisfied. The set of independent equations obviously has a Vandermonde structure and determines r entries of the vector $\mathbf{h}$. Therefore, the maximum τ in (26) satisfies the equality $2\tau = r$, i.e. the maximum approximation order of the constructed SBP operator is $\tau = s$ provided that the elements of the vector $\mathbf{h}$ satisfying (35) are positive.

2.2. $\Xi + \Xi^T$ is a Hankel matrix. For convenience, we introduce the following function of j and σ, $j \le \sigma$,

$$J_{\sigma,j} = \xi_{\sigma-j,j} + \xi_{j,\sigma-j}, \tag{36}$$

and prove that $J_{\sigma,j}$ is independent of j. Formula (34) allows us to derive the equality

$$\xi_{ij} = \sum_{\nu=1}^{s} \alpha_\nu \sum_{\mu=1}^{\nu} \mu^i (\mu-\nu)^j. \tag{37}$$

A simple rearrangement of (37) gives the representations

$$\xi_{ji} = \sum_{\nu=1}^{s} \alpha_\nu \sum_{\mu=1}^{\nu} \mu^j (\mu-\nu)^i = (-1)^{i+j} \sum_{\nu=1}^{s} \alpha_\nu \sum_{\mu=0}^{\nu-1} \mu^i (\mu-\nu)^j \tag{38}$$

and

$$\xi_{ij} + \xi_{ji} = \sum_{\nu=1}^{s} \alpha_\nu \left[\sum_{\mu=1}^{\nu} \mu^i (\mu-\nu)^j + (-1)^\sigma \sum_{\mu=0}^{\nu-1} \mu^i (\mu-\nu)^j \right]. \tag{39}$$

Hence

$$J_{\sigma,j} = \sum_{\nu=1}^{s} \alpha_\nu \left[\sum_{\mu=1}^{\nu} \mu^{\sigma-j} (\mu-\nu)^j + (-1)^\sigma \sum_{\mu=0}^{\nu-1} \mu^{\sigma-j} (\mu-\nu)^j \right]. \tag{40}$$

2.2.1. σ is odd. Formula (40) reduces to the equality

$$J_{\sigma,j} = \sum_{\nu=1}^{s} \alpha_\nu [\nu^i 0^j - 0^i (-\nu)^j], \quad i = \sigma - j.$$

Therefore,

$$J_{\sigma,j} = \begin{cases} 0, & i > 0,\ j > 0, \\ -\sum_{\nu=1}^{s} \alpha_\nu (-\nu)^j, & i = 0,\ j > 0, \\ \sum_{\nu=1}^{s} \alpha_\nu \nu^i, & i > 0,\ j = 0. \end{cases}$$

Owing to (22),

$$J_{\sigma,j} = \begin{cases} \frac{1}{2}, & \sigma = 1, \\ 0, & \sigma = 3, 5, \ldots, 2s-1. \end{cases} \tag{41}$$

2.2.2. σ is even. First we consider $\sigma = 0$. In this case, $i = j = 0$ and

$$\xi_{ij} + \xi_{ji} = 2 \sum_{\nu=1}^{s} \alpha_\nu \sum_{\mu=0}^{\nu-1} \mu^0 (\mu-\nu)^0 = 2 \sum_{\nu=1}^{s} \alpha_\nu \nu = 1.$$

Thus, we have proved the consistency condition $\xi_{00} = 1/2$.

Let us rewrite (40) in the form

$$J_{\sigma,j} = \sum_{\nu=1}^{s} \alpha_\nu N_{\sigma.j}(\nu), \tag{42}$$

where

$$N_{\sigma,j}(\nu) = \sum_{\mu=0}^{\nu-1} \mu^{\sigma-j}(\mu-\nu)^j + \sum_{\mu=1}^{\nu} \mu^{\sigma-j}(\mu-\nu)^j. \tag{43}$$

It is easy to verify that the polynomial $N_{\sigma,j}(\nu)$ of degree j satisfies the recursion

$$N_{\sigma,j}(\nu) = N_{\sigma,j-1}(\nu) - \nu N_{\sigma-1,j-1}(\nu). \tag{44}$$

The recursion is solved by the binomial formula

$$N_{\sigma,j}(\nu) = \sum_{n=0}^{j} \binom{j}{n} N_{\sigma-n,0}(\nu)(-\nu)^n. \tag{45}$$

Theorem 3.

$$N_{\sigma,0}(\nu) = 2\begin{cases} \frac{\nu^{\sigma+1}}{\sigma+1} + \frac{1}{2}B_2\binom{\sigma}{1}\nu^{\sigma-1} + \frac{1}{4}B_4\binom{\sigma}{3}\nu^{\sigma-3} + \cdots + B_\sigma\nu, & \sigma \text{ even} \\ \frac{\nu^{\sigma+1}}{\sigma+1} + \frac{1}{2}B_2\binom{\sigma}{1}\nu^{\sigma-1} + \frac{1}{4}B_4\binom{\sigma}{3}\nu^{\sigma-3} + \cdots + \frac{\sigma}{2}B_{\sigma-1}\nu^2, & \sigma \text{ odd,} \end{cases}$$

where the coefficients B_σ are Bernoulli's numbers.

Proof. We set $j = 0$ in (43) and obtain that

$$N_{\sigma,0}(\nu) = \sum_{\mu=0}^{\nu-1} \mu^\sigma + \sum_{\mu=1}^{\nu} \mu^\sigma = 2\sum_{\mu=1}^{\nu} \mu^\sigma - \nu^\sigma.$$

Then we apply the Bernoulli formula

$$\sum_{\mu=1}^{\nu} \mu^\sigma = \frac{\nu^{\sigma+1}}{\sigma+1} + \frac{B_1}{1}\binom{\sigma}{0}\nu^\sigma + \frac{B_2}{2}\binom{\sigma}{1}\nu^{\sigma-1} + \cdots + \frac{B_{\sigma-1}}{\sigma-1}\binom{\sigma}{\sigma-2}\nu^2 + B_\sigma\nu,$$

where $B_1 = 1/2$ and $B_{2n+1} = 0$ for $n > 0$. □

Corollary 1.

$$N_{\sigma,j}(\nu) = \gamma_1^{\sigma,j}\nu^{\sigma+1} + \gamma_3^{\sigma,j}\nu^{\sigma-1} + \cdots + \gamma_{3+2k}^{\sigma,j}\nu^{\sigma-1-2k} + \cdots,\ k = 0, 1, 2, \ldots.$$

where the last term is proportional to ν or ν^2 depending on whether σ is even or odd. When σ is even, the last term is $2B_\sigma\nu$.

Proof. Using (45) and Theorem 3 we derive the representation

$$N_{\sigma,j}(\nu) = \sum_{n=0}^{j} \binom{j}{n}(-1)^n 2\left\{\frac{\nu^{\sigma+1}}{\sigma-n+1} + \frac{1}{2}B_2\binom{\sigma-n}{1}\nu^{\sigma-1} + \cdots\right\}.$$

The last term appears in the sum only when $n = 0$. □

Since σ is even, all monomials in the polynomial $N_{\sigma,j}(\nu)$ have odd degree. Applying Theorem 1 to (42) when $\sigma + 1 \le 2s - 1$ we obtain $J_{\sigma,j} = B_\sigma$ for $\sigma = 2, 4, \ldots, 2s-2$.

The case $\sigma = 2s$ is not covered by Theorem 1 because $\sigma + 1 > 2s - 1$. It is treated by formula (37) for $i = j = s$:

$$\xi_{ij} + \xi_{ji} = 2 \sum_{\nu=1}^{s} \alpha_\nu \sum_{\mu=0}^{\nu-1} \mu^s (\mu - \nu)^s.$$

2.3. Construction of a diagonal positive definite matrix H and skew-symmetric matrix B_2. The diagonal matrix satisfying (35) is positive definite only if $1 \le s \le 4$. The corresponding SBP operator has interior accuracy of order $2s$ and is formally accurate of order s near boundaries. But the overall accuracy is of order $s + 1$; see e.g. [26].

Let $X = Q \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}$ be a QR factorization of X in (29). Then the skew-symmetric matrix B_2 have the structure $B_2 = Q \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} Q^T$, where $\begin{bmatrix} \Theta_{11} \\ \Theta_{21} \end{bmatrix} = Q^T G \hat{R}^{-1}$, $\Theta_{12} = -\Theta_{21}^T$ and Θ_{22} is an arbitrary skew-symmetric matrix. The skew-symmetric matrix Θ_{22} is of size $(s-1) \times (s-1)$ and therefore depends on $(s-2)(s-1)/2$ free parameters.

The following MATLAB function requires the Symbolic Math Toolbox. The output parameters `alpha,h,B` denote the vector α in Theorem 1, the diagonal of H and the matrix B, respectively. The computed matrix B has $\Theta_{22} = 0$; see line 14 of the function.

```
function [alpha,h,B] = SBP(s)
s = sym(s); r = 2*s;
alpha = -(-1).^(1:s).*cumprod((s:-1:1)./(s+1:2*s))./(1:s);
Cs = toeplitz(alpha(s:-1:1), ...
     [alpha(s) sym(zeros(1,double(s-1)))]);
C = [sym(zeros(double(r-s))); Cs];
X = (r:-1:1)'.^(0:r-1); Y = (0:-1:1-s)'.^(0:s);
F = -C*Y; F(1,:) = F(1,:) + r.^(0:s)/2;
Hankel = X(:,1:s+1)'*F + F'*X(:,1:s+1);
h = Hankel([2:s+1, (s+1)*s+2:(s+1)^2])./(1:r)/X;
G = F - diag(h)*X(:,[1 1:s])*diag(0:s);
[Q,R] = qr(X(:,1:s+1));
T1 = Q'*G/R(1:s+1,:);
T = [T1 [-T1(s+2:r,:)'; sym(zeros(double(r-s-1)))]];
B = Q*T*Q'; B(1,1) = -1/2;
```

Here is the output for $s = 1, 2$:

$$s = 1\text{:} \quad \alpha_1 = 1/2,\ H = \mathrm{diag}[1/2, 1],\ B = \begin{bmatrix} -1/2, & 1/2 \\ -1/2, & 0 \end{bmatrix};$$

$$s=2: \qquad \alpha_1=2/3,\ \alpha_2=-1/12,\ H=\mathrm{diag}[17,59,43,49]/18,$$

$$B=\begin{bmatrix} -1/2, & 59/96, & -1/12, & -1/32\\ -59/96, & 0, & 59/96, & 0\\ 1/12, & -59/96, & 0, & 59/96\\ 1/32, & 0, & -59/96, & 0 \end{bmatrix}.$$

2.4. Mirror classical SBP operator. We build an SBP operator that approximates the first derivative on the semi-infinite grid $x_\nu=-\nu h$, $\nu=0,1,\dots$, in the form of a block-partitioned semi-infinite matrix

$$h^{-1}\begin{bmatrix} \hat D & \hat C\\ -\hat H^{-1}\hat C^T & \hat H^{-1}\hat B \end{bmatrix}, \tag{46}$$

where

$$\hat C=\begin{bmatrix} \vdots & \vdots\\ 0 & 0\\ C_s & C_0^T \end{bmatrix}, \quad C_s=\begin{bmatrix} \alpha_s & 0 & \cdots & \cdots & 0\\ \alpha_{s-1} & \alpha_s & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & \ddots & 0\\ \alpha_1 & \cdots & \cdots & \alpha_{s-1} & \alpha_s \end{bmatrix},$$

$$\hat D=\begin{bmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots\\ 0 & -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0\\ & \ddots & \ddots & -\alpha_2 & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s\\ & & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots\\ & & & \ddots & -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1\\ & & & & 0 & -\alpha_s & \cdots & -\alpha_1 & 0 \end{bmatrix}.$$

The matrix C_0 is the zero $(r-s)\times s$ matrix. The coefficients α_ν in C_s and in the semi-infinite matrix $\hat D$ are defined in Theorem 1. The semi-infinite block $\hat D$ provides $2s$-order accurate approximation of the first derivative at the inner grid points x_ν, $\nu\le -r$.

Let us introduce the $r\times r$ permutation matrix

$$\Pi=\begin{bmatrix} & & & 1\\ & & \iddots & \\ & 1 & & \\ 1 & & & \end{bmatrix}$$

which flips a vector v such that $(\Pi v)_i=r+1-i$, $i=1,2,\dots,r$. We assume that $r\ge s$ and choose the $r\times r$ matrix $\hat B$ from (46) in the form

$$\hat B=-\Pi B_1\Pi+\hat B_2, \quad B_1=\mathrm{diag}(-1/2,0,\dots,0), \quad \hat B_2^T=-\hat B_2. \tag{47}$$

The algebraic system for approximation of order τ near the boundary is

$$\hat{H}^{-1}\hat{B}\Pi\mathbf{x}^j - \hat{H}^{-1}\begin{bmatrix} C_s^T \\ C_0 \end{bmatrix}\hat{\mathbf{y}}^j = j\Pi\mathbf{x}^{j-1}, \quad j = 0, 1, \ldots, \tau, \tag{48}$$

where $\mathbf{x} = \begin{bmatrix} r & r-1 & \ldots & 1 \end{bmatrix}^T$ and $\hat{\mathbf{y}} = \begin{bmatrix} 1-s & \ldots & -1 & 0 \end{bmatrix}^T$ is the flipped vector $\mathbf{y} = \begin{bmatrix} 0 & -1 & \ldots & 1-s \end{bmatrix}^T$. Multiplying (48) by $\Pi\hat{H}$ from the left gives the system

$$\Pi\hat{B}_2\Pi\mathbf{x}^j - B_1\mathbf{x}^j - \Pi\begin{bmatrix} C_s^T \\ C_0 \end{bmatrix}\hat{\mathbf{y}}^j = j\Pi\hat{H}\Pi\mathbf{x}^{j-1}, \quad j = 0, 1, \ldots, \tau.$$

As a result, we arrive at the following equations for $j = 0, 1, \ldots, \tau$

$$\Pi\hat{B}_2\Pi\mathbf{x}^j = -\mathbf{g}_j, \text{ where } \mathbf{g}_j = -B_1\mathbf{x}^j - \begin{bmatrix} C_0 \\ C_s \end{bmatrix}\mathbf{y}^j - j\Pi\hat{H}\Pi\mathbf{x}^{j-1}. \tag{49}$$

Comparing (49) with (28) gives the desired formulas

$$\hat{B}_2 = -\Pi B_2\Pi, \quad \hat{H} = \Pi H\Pi. \tag{50}$$

3 Symmetric hyperbolic systems with dissipative boundary conditions

Consider a symmetric hyperbolic system [7, 9]

$$A(t,x,y)\frac{\partial}{\partial t}u + B(t,x,y)\frac{\partial}{\partial x}u + C(t,x,y)\frac{\partial}{\partial y}u + D(t,x,y)u = f(t,x,y) \tag{51}$$

for $0 \le t \le T$, $a \le x \le b$, $c \le y \le d$. The matrices A, B and C are Hermitian that is $A = A^*$, $B = B^*$, $C = C^*$. Moreover, the matrix A is positive definite. The matrices A, B, C, D and the vector functions $u(t,x,y)$ and $f(t,x,y)$ may have complex entries. The real part of a complex number z is denoted by $\mathsf{Re}\, z$.

Theorem 4. *Solution of (51) satisfies the matrix identity*

$$\frac{\partial}{\partial t}(Au,u) + \frac{\partial}{\partial x}(Bu,u) + \frac{\partial}{\partial y}(Cu,u) + 2\mathsf{Re}(Du,u) = ((A_t + B_x + C_y)u,u) + 2\mathsf{Re}(f,u), \tag{52}$$

where $(u,v) = u^*v$ is the Euclidean inner product of vectors u and v.

The proof of Theorem 4 is direct.

Let us introduce the L^2 inner product of vector functions $u(t,x,y)$ and $v(t,x,y)$

$$\langle u, v\rangle = \int_a^b \int_c^d u^* v dx dy$$

and the L^2 inner products on the boundaries:

$$\langle u,v\rangle|_{x=a}=\int_c^d (u^*v)|_{x=a}dy,\quad \langle u,v\rangle|_{x=b}=\int_c^d (u^*v)|_{x=b}dy,$$
$$\langle u,v\rangle|_{x=a}^{x=b}=\langle u,v\rangle|_{x=b}-\langle u,v\rangle|_{x=a}.$$

Theorem 5. *Solution of (51) satisfies the energy identity*

$$\frac{d}{dt}\langle Au,u\rangle+\langle Bu,u\rangle|_{x=a}^{x=b}+\langle Cu,u\rangle|_{y=c}^{y=d}+2\mathsf{Re}\langle Du,u\rangle \\ =\langle (A_t+B_x+C_y)u,u\rangle+2\mathsf{Re}\langle f,u\rangle. \quad (53)$$

The proof of Theorem 5 is direct.

3.1. Boundary conditions. Assume that boundary conditions imposed on the solution of the symmetric hyperbolic system (51) are of the form

$$\begin{aligned} &Z_a(t,y)u=g_a(t,y) \text{ for } x=a, \qquad &&Z_b(t,y)u=g_b(t,y) \text{ for } x=b,\\ &Z_c(t,x)u=g_c(t,x) \text{ for } y=c, \qquad &&Z_d(t,x)u=g_d(t,x) \text{ for } y=d, \end{aligned} \quad (54)$$

where the matrices $Z_a(t,y)$, $Z_b(t,y)$, $Z_c(t,x)$ and $Z_d(t,x)$ have full row rank. We also assume that $B(t,x,y)$ is nonsingular when $x=a$ or $x=b$, and $C(t,x,y)$ is nonsingular when $y=c$ or $x=d$. The theory of hyperbolic equations requires the following necessary conditions on the boundaries:

- $x=a$**:** the number of rows in $Z_a(t,y)$ is equal to the number of positive eigenvalues of the matrix $B(t,a,y)$;
- $x=b$**:** the number of rows in $Z_b(t,y)$ is equal to the number of negative eigenvalues of the matrix $B(t,a,y)$;
- $y=c$**:** the number of rows in $Z_c(t,y)$ is equal to the number of positive eigenvalues of the matrix $B(t,x,c)$;
- $y=d$**:** the number of rows in $Z_d(t,y)$ is equal to the number of negative eigenvalues of the matrix $B(t,x,d)$;

The boundary conditions (54) are called *dissipative* if

$$\begin{aligned} &(B(t,a,y)v,v)\le 0 \text{ whenever } Z_a(t,y)v=0,\\ &(B(t,b,y)v,v)\ge 0 \text{ whenever } Z_b(t,y)v=0,\\ &(C(t,x,c)v,v)\le 0 \text{ whenever } Z_c(t,y)v=0,\\ &(C(t,x,d)v,v)\ge 0 \text{ whenever } Z_d(t,y)v=0. \end{aligned} \quad (55)$$

Let the boundary conditions (54) be homogeneous i.e. $g_a\equiv 0$, $g_b\equiv 0$, $g_c\equiv 0$, $g_d\equiv 0$. When the homogeneous boundary conditions are dissipative the following energy inequality holds:

$$\frac{d}{dt}\langle Au,u\rangle\le\langle (A_t+B_x+C_y)u,u\rangle-2\mathsf{Re}\langle Du,u\rangle+2\mathsf{Re}\langle f,u\rangle. \quad (56)$$

A symmetric hyperbolic system (51) with dissipative boundary conditions (54), (55) has a unique solution; see e.g. [18, 9, 2].

4 Tensor operations

We are given integer indices i, j, k in the range $1 \le i \le s_1$, $0 \le j \le s_2$, $0 \le k \le s_3$. The vector V_i is a tensor of order 1, the matrix M_{ij} is a tensor of order 2. We are mainly interested in tensors of order 3 given by the values $U = U_{ijk}$, which are generally complex. The indices j and k indicate the grid nodes along the x and y axes, respectively. The index i is usually used for multiplication of U by a matrix coefficient.

4.1. Vectorization. A tensor U of order 3 can be unfolded into a vector $\vec{U} = \mathrm{vec}(U)$ by using the lexicographic ordering

$$\vec{U}_{i+js_1+ks_1(s_2+1)} = U_{ijk}. \tag{57}$$

Tensors of order 2 are vectorized similarly, i.e. $\vec{U}_{i+js_1} = U_{ij}$.

4.2. Inner product. The inner product of tensors U and V of order 3 is the number

$$(U, V) = (\vec{U}, \vec{V}) = \sum_{ijk} \bar{U}_{ijk} V_{ijk} \tag{58}$$

where the bar over U_{ijk} denotes complex conjugation.

4.3. Multiplication by a matrix in a certain index mode. The product of the tensor U_{ijk} with a $s_1' \times s_1$ matrix A in mode 1, that is, with respect to index i, is the tensor $V = A \times_1 U$ with entries

$$V_{i'jk} = \sum_i A_{i'i} U_{ijk}. \tag{59}$$

The product of a tensor with a matrix in modes 2 or 3 is defined analogously.

4.4. Kronecker product of matrices. The Kronecker product of an $m \times n$ matrix A and a matrix B is the block matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}.$$

The Kronecker product satisfies the identity $(A \otimes B) \otimes C = A \otimes (B \otimes C)$. The Kronecker product of three matrices $A \otimes B \otimes C$ is $(A \otimes B) \otimes C$ by definition. Moreover, $A_1A_2 \otimes B_1B_2 = (A_1 \otimes B_1)(A_2 \otimes B_2)$ and

$$A_1A_2 \otimes B_1B_2 \otimes C_1C_2 = (A_1 \otimes B_1 \otimes C_1)(A_2 \otimes B_2 \otimes C_2).$$

The vectorized equivalent for the operation $V = A \times_1 U$ is

$$\vec{V} = (I \otimes I \otimes A)\vec{U},$$

where I denotes the identity matrix of suitable size. Similar expressions hold for multiplication along modes 2 or 3.

4.5. Multiplication by variable matrix coefficients. In what follows, we will often operate with sets of matrices $A(t,j,k)$ that depend on a continuous parameter t and integer indices j,k. Then the product $V = AU$ is defined as

$$V_{i'jk}(t) = \sum_i (A(t,j,k))_{i'i} U_{ijk}(t). \tag{60}$$

Note that this type of multiplication is always performed in the first dimension, i, of the tensor U.

To rewrite (60) in a vectorized form, we form a block-diagonal matrix

$$\mathcal{A}(t) = \text{blockdiag}\{A(t,j,k)\colon 0 \le j \le s_2, 0 \le k \le s_3\}, \tag{61}$$

where the diagonal blocks $A(t,j,k)$ are ordered in ascending order with respect to the linear index $j + k(s_2+1)$. Then the product (60) is given by the vectorized expression $\vec{V}(t) = \mathcal{A}(t)\vec{U}(t)$.

4.6. Boundary slices. Given tensor values U_{ijk} on a rectangular grid (x_j, y_k), the boundary slices of U are the tensors $U^{[x_0]}$, $U^{[x_{\text{end}}]}$, $U^{[y_0]}$, $U^{[y_{\text{end}}]}$ of order 2 defined by the entries

$$U^{[x_0]}_{ik} = U_{i,0,k}, \quad U^{[x_{\text{end}}]}_{ik} = U_{i,s_2,k}, \quad U^{[y_0]}_{ij} = U_{i,j,0}, \quad U^{[y_{\text{end}}]}_{ij} = U_{i,j,s_3}.$$

For a set of matrix coefficients $A(t,j,k)$, we also introduce the boundary subsets $A^{[x_0]}$, $A^{[x_{\text{end}}]}$, $A^{[y_0]}$, $A^{[y_{\text{end}}]}$ such that

$$A^{[x_0]}(t,k) = A(t,0,k), \quad A^{[x_{\text{end}}]}(t,k) = A(t,s_2,k),$$
$$A^{[y_0]}(t,j) = A(t,j,0), \quad A^{[y_{\text{end}}]}(t,j) = A(t,j,s_3).$$

5 Energy estimates for semi-discretized symmetric hyperbolic systems

The main advantage of SBP operators is that they allow one to replicate with little effort the derivation of the energy identities for partial differential equations when performing semi-discrete approximation.

We perform semi-discretization of (51) on the 2-dimensional grid (x_j, y_k) composed from the equidistant 1-dimensional grids,

$$a = x_0 < x_1 < \cdots < x_{s_x} = b \quad \text{and} \quad c = y_0 < y_1 < \cdots < y_{s_y} = d.$$

The tensor components $U_{ijk}(t)$ are approximations to the values $u_i(t, x_j, y_k)$.

Recall that the SBP differentiation matrices that approximate the first derivative on a 1-dimensional grid have the form $\mathsf{D} = P^{-1}Q$, where P is a positive definite diagonal matrix and Q satisfies the SBP property

$$Q + Q^T = \mathsf{E} = \text{diag}[-1, 0, 0, \ldots, 0, 1].$$

We have two SBP operators, $\mathsf{D}_x = P_x^{-1}Q_x$ and $\mathsf{D}_y = P_y^{-1}Q_y$. For brevity, we will denote

$$\mathbb{D}_x V = P_x^{-1}Q_x \times_2 V \quad \text{and} \quad \mathbb{D}_y V = P_y^{-1}Q_y \times_3 V.$$

The vectorized form of $V(t) = P_x^{-1}Q_x \times_2 U(t)$ is

$$\vec{V}(t) = (I \otimes P_x^{-1}Q_x \otimes I)\vec{U}(t).$$

Similarly, the vectorized form of $V(t) = P_y^{-1}Q_y \times_2 U(t)$ is given by

$$\vec{V}(t) = (P_y^{-1}Q_y \otimes I \otimes I)\vec{U}(t).$$

Remark 3. *When the matrix coefficient B varies with respect to x, the discretization of the term $B(t,x,y)\frac{\partial}{\partial x}u$ from (51) in the form $B\mathbb{D}_x U$ does not lead to a discrete analog of the energy identity.*

A working trick is to use the equality $Bu_x = \frac{1}{2}[(Bu)_x + Bu_x - B_x u]$ and discretize $\frac{1}{2}[(Bu)_x + Bu_x - B_x u]$ instead of Bu_x. Thus, the SBP semi-discretization is applied to the equivalent system

$$\begin{aligned} A(t,x,y)\frac{\partial}{\partial t}u + \frac{1}{2}\frac{\partial}{\partial x}\left[B(t,x,y)u\right] + \frac{1}{2}B(t,x,y)\frac{\partial}{\partial x}u - \frac{1}{2}\frac{\partial}{\partial x}B(t,x,y)u \\ + \frac{1}{2}\frac{\partial}{\partial y}\left[C(t,x,y)u\right] + \frac{1}{2}C(t,x,y)\frac{\partial}{\partial y}u - \frac{1}{2}\frac{\partial}{\partial y}C(t,x,y)u \\ + D(t,x,y)u = f(t,x,y) \end{aligned} \quad (62)$$

instead of (51). We do not discretize the derivatives A_t, B_x and C_y and assume for brevity that these matrices are given.

The system (62) is semi-discretized as follows:

$$\begin{aligned} A(t,j,k)U_t + \frac{1}{2}\mathbb{D}_x\left[B(t,j,k)U\right] + \frac{1}{2}B(t,j,k)\mathbb{D}_x U - \frac{1}{2}B_x(t,j,k)U \\ + \frac{1}{2}\mathbb{D}_y\left[C(t,j,k)U\right] + \frac{1}{2}C(t,j,k)\mathbb{D}_y U - \frac{1}{2}C_y(t,j,k)U + D(t,j,k)U \\ = F(t,j,k) + \mathbb{SAT}. \end{aligned} \quad (63)$$

Dissipative boundary conditions are discretized by the SAT terms in $\mathbb{SAT}$. The exact expression for $\mathbb{SAT}$ is given later.

Let us introduce the block-diagonal matrices

$$\begin{aligned} \mathcal{A} &= \mathrm{blockdiag}(A(t,j,k)), \quad & \mathcal{A}_t &= \mathrm{blockdiag}(A_t(t,j,k)), \\ \mathcal{B} &= \mathrm{blockdiag}(B(t,j,k)), \quad & \mathcal{B}_x &= \mathrm{blockdiag}(B_x(t,j,k)), \\ \mathcal{C} &= \mathrm{blockdiag}(C(t,j,k)), \quad & \mathcal{C}_y &= \mathrm{blockdiag}(C_y(t,j,k)), \\ \mathcal{D} &= \mathrm{blockdiag}(D(t,j,k)), \quad & \mathcal{P} &= P_y \otimes P_x \otimes I. \end{aligned} \quad (64)$$

The matrices $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{A}_t$, $\mathcal{B}_x$, $\mathcal{C}_y$ are hermitian. The matrix $\mathcal{P}$ is diagonal.

Lemma 1. *Owing to the special diagonal structure of $\mathcal{P}$ and block structure of matrices in (64),*

$$\begin{aligned} \mathcal{P}\mathcal{A} = \mathcal{A}\mathcal{P}, \quad \mathcal{P}\mathcal{B} = \mathcal{B}\mathcal{P}, \quad \mathcal{P}\mathcal{C} = \mathcal{C}\mathcal{P}, \quad \mathcal{P}\mathcal{D} = \mathcal{D}\mathcal{P}, \\ \mathcal{P}\mathcal{A}_t = \mathcal{A}_t\mathcal{P}, \quad \mathcal{P}\mathcal{B}_x = \mathcal{B}_x\mathcal{P}, \quad \mathcal{P}\mathcal{C}_y = \mathcal{C}_y\mathcal{P}. \end{aligned}$$

The commutativity also holds for the diagonal matrices $P_y \otimes \mathsf{E}_x \otimes I$ and $\mathsf{E}_y \otimes P_x \otimes I$ instead of $\mathcal{P}$.

Proof. Consider, for example, the identity $\mathcal{PB} = \mathcal{BP}$. The block with indices j, k in both parts of the identity is the product of $B(t, j, k)$ with the scalar $P_x(j, j)P_y(k, k)$, where $P_x(j, j)$ and $P_y(k, k)$ are the corresponding diagonal elements in the diagonal matrices P_x and P_y. The remaining identities are proved similarly. □

For convenience, we transform (63) equivalently into vectorized form

$$\mathcal{A}\vec{U}_t + \frac{1}{2}(I \otimes \mathsf{D}_x \otimes I)\mathcal{B}\vec{U} + \frac{1}{2}\mathcal{B}(I \otimes \mathsf{D}_x \otimes I)\vec{U} - \frac{1}{2}\mathcal{B}_x\vec{U} + \frac{1}{2}(\mathsf{D}_y \otimes I \otimes I)\mathcal{C}\vec{U} + \frac{1}{2}\mathcal{C}(\mathsf{D}_y \otimes I \otimes I)\vec{U} - \frac{1}{2}\mathcal{C}_y\vec{U} + \mathcal{D}\vec{U} = \vec{F} + \mathbb{S}\vec{\mathbb{A}}\mathbb{T}. \quad (65)$$

Let us introduce the inner products on the boundaries $x = a$ and $x = b$:

$$(\mathcal{B}\vec{U}, \vec{U})_{P_y}^{[x_0]} = \sum_k P_y(k, k) \sum_{i,i';j=0} B_{ii'}(t, j, k)\bar{U}_{i'jk}U_{ijk} \quad (66)$$

$$(\mathcal{B}\vec{U}, \vec{U})_{P_y}^{[x_{\text{end}}]} = \sum_k P_y(k, k) \sum_{i,i';j=s_x} B_{ii'}(t, j, k)\bar{U}_{i'jk}U_{ijk} \quad (67)$$

The inner products on the other boundaries, $(\mathcal{C}\vec{U}, \vec{U})_{P_x}^{[y_0]}$ and $(\mathcal{C}\vec{U}, \vec{U})_{P_x}^{[y_{\text{end}}]}$, are defined accordingly.

Theorem 6. *The following energy identity holds for a solution of (65):*

$$\frac{d}{dt}(\mathcal{A}\vec{U}, \vec{U})_{\mathcal{P}} + (\mathcal{B}\vec{U}, \vec{U})_{P_y}^{[x_{\text{end}}]} - (\mathcal{B}\vec{U}, \vec{U})_{P_y}^{[x_0]} + (\mathcal{C}\vec{U}, \vec{U})_{P_x}^{[y_{\text{end}}]} - (\mathcal{C}\vec{U}, \vec{U})_{P_x}^{[y_0]} = ([\mathcal{A}_t + \mathcal{B}_x + \mathcal{C}_y]\vec{U}, \vec{U})_{\mathcal{P}} - 2\mathsf{Re}(\mathcal{D}\vec{U}, \vec{U})_{\mathcal{P}} + 2\mathsf{Re}(\vec{F}, \vec{U})_{\mathcal{P}} + (\mathbb{S}\vec{\mathbb{A}}\mathbb{T}, \vec{U})_{\mathcal{P}} + (\vec{U}, \mathbb{S}\vec{\mathbb{A}}\mathbb{T})_{\mathcal{P}}. \quad (68)$$

Proof. Introduce the auxiliary matrices

$$\Phi = \frac{1}{2}(I \otimes \mathsf{D}_x \otimes I)\mathcal{B} + \frac{1}{2}\mathcal{B}(I \otimes \mathsf{D}_x \otimes I) - \frac{1}{2}\mathcal{B}_x$$

$$\Psi = \frac{1}{2}(\mathsf{D}_y \otimes I \otimes I)\mathcal{C} + \frac{1}{2}\mathcal{C}(\mathsf{D}_y \otimes I \otimes I) - \frac{1}{2}\mathcal{C}_y$$

and rewrite the system in a compact form as $\mathcal{A}\vec{U}_t + \Phi\vec{U} + \Psi\vec{U} + \mathcal{D}\vec{U} = \vec{F} + \mathbb{S}\vec{\mathbb{A}}\mathbb{T}$. Then

$$\begin{aligned}
\frac{d}{dt}(\mathcal{A}\vec{U}, \vec{U})_{\mathcal{P}} &= (\mathcal{A}\vec{U}_t, \vec{U})_{\mathcal{P}} + (\mathcal{A}\vec{U}, \vec{U}_t)_{\mathcal{P}} + (\mathcal{A}_t\vec{U}, \vec{U})_{\mathcal{P}} \\
&= (-\Phi\vec{U} - \Psi\vec{U} - \mathcal{D}\vec{U} + \vec{F} + \mathbb{S}\vec{\mathbb{A}}\mathbb{T}, \vec{U})_{\mathcal{P}} \\
&\quad + (\vec{U}, -\Phi\vec{U} - \Psi\vec{U} - \mathcal{D}\vec{U} + \vec{F} + \mathbb{S}\vec{\mathbb{A}}\mathbb{T})_{\mathcal{P}} \\
&= -((\mathcal{P}\Phi + \Phi^*\mathcal{P})\vec{U}, \vec{U}) - ((\mathcal{P}\Psi + \Psi^*\mathcal{P})\vec{U}, \vec{U})) \\
&\quad + (\mathcal{A}_t\vec{U}, \vec{U})_{\mathcal{P}} - 2\mathsf{Re}(\mathcal{D}\vec{U}, \vec{U})_{\mathcal{P}} + 2\mathsf{Re}(\vec{F}, \vec{U})_{\mathcal{P}} \\
&\quad + (\mathbb{S}\vec{\mathbb{A}}\mathbb{T}, \vec{U})_{\mathcal{P}} + (\vec{U}, \mathbb{S}\vec{\mathbb{A}}\mathbb{T})_{\mathcal{P}}.
\end{aligned}$$

Using Lemma 1 we derive the equalities

$$\begin{aligned}\mathcal{P}\Phi+\Phi^T\mathcal{P}&=\frac{1}{2}(P_y\otimes Q_x\otimes I)\mathcal{B}+\frac{1}{2}\mathcal{B}(P_y\otimes Q_x\otimes I)-\frac{1}{2}\mathcal{P}\mathcal{B}_x\\&\quad+\frac{1}{2}(P_y\otimes Q_x^T\otimes I)\mathcal{B}+\frac{1}{2}\mathcal{B}(P_y\otimes Q_x^T\otimes I)-\frac{1}{2}\mathcal{B}_x\mathcal{P}\\&=\frac{1}{2}(P_y\otimes(Q_x+Q_x^T)\otimes I)\mathcal{B}+\frac{1}{2}\mathcal{B}(P_y\otimes(Q_x+Q_x^T)\otimes I)-\mathcal{P}\mathcal{B}_x.\end{aligned}$$

Since $Q_x+Q_x^T=\mathsf{E}_x$ is diagonal, we obtain

$$\mathcal{P}\Phi+\Phi^T\mathcal{P}=(P_y\otimes\mathsf{E}_x\otimes I)\mathcal{B}-\mathcal{P}\mathcal{B}_x. \tag{69}$$

Analogously, $Q_y+Q_y^T=\mathsf{E}_y$ is diagonal and

$$\mathcal{P}\Psi+\Psi^T\mathcal{P}=(\mathsf{E}_y\otimes P_x\otimes I)\mathcal{C}-\mathcal{P}\mathcal{C}_y. \tag{70}$$

As a result,

$$\begin{aligned}\frac{d}{dt}(\mathcal{A}\vec{U},\vec{U})_{\mathcal{P}}+((P_y\otimes\mathsf{E}_x\otimes I)\mathcal{B}\vec{U},\vec{U})+((\mathsf{E}_y\otimes P_x\otimes I)\mathcal{C}\vec{U},\vec{U})=\\([\mathcal{A}_t+\mathcal{B}_x+\mathcal{C}_y]\vec{U},\vec{U})_{\mathcal{P}}-2\mathsf{Re}(\mathcal{D}\vec{U},\vec{U})_{\mathcal{P}}+2\mathsf{Re}(\vec{F},\vec{U})_{\mathcal{P}}\\+(\mathbb{S}\vec{\mathbb{A}}\mathbb{T},\vec{U})_{\mathcal{P}}+(\vec{U},\mathbb{S}\vec{\mathbb{A}}\mathbb{T})_{\mathcal{P}}.\end{aligned}$$

Since $\mathsf{E}_x=\mathrm{diag}[-1,0,\dots,1]$, we obtain the equality

$$\begin{aligned}(P_y\otimes\mathsf{E}_x\otimes I)\mathcal{B}=&-\mathrm{blockdiag}[P_y(k,k)B(j,k)\delta_j^0]\\&+\mathrm{blockdiag}[P_y(k,k)B(j,k)\delta_j^{s_x}],\end{aligned}$$

where $\delta_j^n=1$ if $j=n$ and $\delta_j^n=0$ if $j\neq n$. Therefore

$$((P_y\otimes\mathsf{E}_x\otimes I)\mathcal{B}\vec{U},\vec{U})=-(\mathcal{B}\vec{U},\vec{U})_{P_y}^{[x_0]}+(\mathcal{B}\vec{U},\vec{U})_{P_y}^{[x_{\mathrm{end}}]}.$$

Similarly,

$$((\mathsf{E}_y\otimes P_x\otimes I)\mathcal{C}\vec{U},\vec{U})=-(\mathcal{C}\vec{U},\vec{U})_{P_x}^{[y_0]}+(\mathcal{C}\vec{U},\vec{U})_{P_x}^{[x_{\mathrm{end}}]}.$$

□

5.1. SAT terms on the boundaries. Let us construct the following penalty terms corresponding to the boundary conditions (54):

$$\begin{aligned}\mathsf{B}_a&=-\frac{\sigma_a}{2}Z_a^*\Lambda_a(Z_aU^{[x_0]}-G_a),\\\mathsf{B}_b&=-\frac{\sigma_b}{2}Z_b^*\Lambda_b(Z_bU^{[x_{\mathrm{end}}]}-G_b),\\\mathsf{B}_c&=-\frac{\sigma_c}{2}Z_c^*\Lambda_c(Z_cU^{[y_0]}-G_c),\\\mathsf{B}_d&=-\frac{\sigma_d}{2}Z_d^*\Lambda_d(Z_dU^{[y_{\mathrm{end}}]}-G_d).\end{aligned} \tag{71}$$

These functions depend on positive constants σ_a, σ_b, σ_c, σ_d and on positive definite hermitian matrices Λ_a, Λ_b, Λ_c, Λ_d. The matrices Λ_a, Λ_b, Λ_c, Λ_d are

chosen freely based convenience. In this paper, we choose

$$\Lambda_a = (Z_a Z_a^*)^{-1}, \quad \Lambda_b = (Z_b Z_b^*)^{-1}, \quad \Lambda_c = (Z_c Z_c^*)^{-1}, \quad \Lambda_d = (Z_d Z_d^*)^{-1}.$$

The functions B_a, B_b, B_c, B_d are extended by zero from the boundaries to the whole domain:

$$F_a(t,j,k) = \begin{cases} \mathsf{B}_a(t,k), & j=0, \\ 0, & j>0, \end{cases} \qquad F_b(t,j,k) = \begin{cases} \mathsf{B}_b(t,k), & j=s_2, \\ 0, & j<s_2, \end{cases}$$

$$F_a(t,j,k) = \begin{cases} \mathsf{B}_c(t,j), & k=0, \\ 0, & k>0, \end{cases} \qquad F_b(t,j,k) = \begin{cases} \mathsf{B}_d(t,j), & k=s_3, \\ 0, & j<s_3. \end{cases}$$

Such an extension is done solely to comply with formula (63).

The SAT terms take the form

$$\mathbb{SAT}_a = P_x^{-1} \times_2 F_a(t,j,k), \quad \mathbb{SAT}_b = P_x^{-1} \times_2 F_b(t,j,k), \tag{72}$$

$$\mathbb{SAT}_c = P_y^{-1} \times_3 F_c(t,j,k), \quad \mathbb{SAT}_d = P_y^{-1} \times_3 F_d(t,j,k). \tag{73}$$

Boundary conditions (54) are enforced in (63) by a penalty using the combination of SAT terms:

$$\mathbb{SAT} = \mathbb{SAT}_a + \mathbb{SAT}_b + \mathbb{SAT}_c + \mathbb{SAT}_d. \tag{74}$$

Since the matrix P_y is diagonal, we use Lemma 1 and derive the identities

$$\begin{aligned}(\mathbb{SAT}_a, U)_{\mathcal{P}} + (U, \mathbb{SAT}_a)_{\mathcal{P}} = &-\sigma_a (Z_a^* \Lambda_a Z_a U^{[x_0]}, U^{[x_0]})_{P_y} \\ &+ \frac{\sigma_a}{2}(Z_a^* \Lambda_a G_a, U^{[x_0]})_{P_y} + \frac{\sigma_a}{2}(U^{[x_0]}, Z_a^* \Lambda_a G_a)_{P_y}.\end{aligned}$$

Other SAT terms are treated similarly.

We arrive at the following theorem.

Theorem 7. *The following energy identity holds for a solution of (63):*

$$\begin{aligned}\frac{d}{dt}(\mathcal{A}\vec{U}, \vec{U})_{\mathcal{P}} &+ (S^{[x_0]} U^{[x_0]}, U^{[x_0]})_{P_y} + (S^{[x_{\text{end}}]} U^{[x_{\text{end}}]}, U^{[x_{\text{end}}]})_{P_y} \\ &+ (S^{[y_0]} U^{[y_0]}, U^{[y_0]})_{P_x} + (S^{[y_{\text{end}}]} U^{[y_{\text{end}}]}, U^{[y_{\text{end}}]})_{P_x} \\ &= ([\mathcal{A}_t + \mathcal{B}_x + \mathcal{C}_y]\vec{U}, \vec{U})_{\mathcal{P}} - 2\mathsf{Re}(\mathcal{D}\vec{U}, \vec{U})_{\mathcal{P}} + 2\mathsf{Re}(\vec{F}, \vec{U})_{\mathcal{P}} \\ &+ \sigma_a \mathsf{Re}(Z_a^* \Lambda_a G_a, U^{[x_0]})_{P_y} + \sigma_b \mathsf{Re}(Z_b^* \Lambda_b G_b, U^{[x_{\text{end}}]})_{P_y} \\ &+ \sigma_c \mathsf{Re}(Z_c^* \Lambda_c G_c, U^{[y_0]})_{P_x} + \sigma_d \mathsf{Re}(Z_d^* \Lambda_d G_d, U^{[y_{\text{end}}]})_{P_x},\end{aligned} \tag{75}$$

where we use the hermitian matrices

$$S^{[x_0]}(t,k) = -B^{[x_0]} + \sigma_a Z_a^* \Lambda_a Z_a, \;\; S^{[x_{\text{end}}]}(t,k) = B^{[x_{\text{end}}]} + \sigma_b Z_b^* \Lambda_b Z_b, \tag{76}$$

$$S^{[y_0]}(t,j) = -B^{[y_0]} + \sigma_c Z_c^* \Lambda_c Z_c, \;\; S^{[y_{\text{end}}]}(t,j) = B^{[y_{\text{end}}]} + \sigma_d Z_d^* \Lambda_d Z_d. \tag{77}$$

We will use *strictly dissipative* boundary conditions instead of non-strictly dissipative conditions as in (55). Strict dissipation assumes the inequalities

$$(-B^{[x_0]}v, v) > 0 \text{ for all nonzero vectors } v \text{ such that } Z_a v = 0,$$
$$(B^{[x_{\text{end}}]}v, v) > 0 \text{ for all nonzero vectors } v \text{ such that } Z_b v = 0,$$
$$(-B^{[y_0]}v, v) > 0 \text{ for all nonzero vectors } v \text{ such that } Z_c v = 0,$$
$$(B^{[y_{\text{end}}]}v, v) > 0 \text{ for all nonzero vectors } v \text{ such that } Z_d v = 0,$$

One can choose the parameters σ_a, σ_b, σ_c and σ_d in (71) such that all matrices $S^{[x_0]}$, $S^{[x_{\text{end}}]}$, $S^{[y_0]}$, $S^{[y_{\text{end}}]}$ in (76) and (77) are non-negative definite.

The SAT method for a weak enforcement of boundary conditions in the method of lines was proposed in [3].

5.2. How to choose σ_a, σ_b, σ_c and σ_d. The matrices in (76) and (77) have the general form

$$S = -\hat{B} + \sigma Z^* \Lambda Z,$$

where $\hat{B}$ denotes $B^{[x_0]}$, $-B^{[x_{\text{end}}]}$, $B^{[y_0]}$, or $-B^{[y_{\text{end}}]}$. Strict dissipation for a boundary condition $ZU = G$ means that $(\hat{B}v, v) < 0$ for all nonzero vectors v satisfying the equation $Zv = 0$.

Let $Z^* = Q\hat{R}$ be a QR factorization, where Q is unitary, $\hat{R} = \begin{bmatrix} R \\ 0 \end{bmatrix}$ and R is nonsingular. Using the matrix Q we introduce the following block matrices:

$$Q^*\hat{B}Q = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \quad Q^*Z^*\Lambda ZQ = \begin{bmatrix} R\Lambda R^* & 0 \\ 0 & 0 \end{bmatrix}.$$

The block B_{22} is negative definite, that is $B_{22} < 0$, due to the strict dissipative boundary condition $ZU = G$. Since $\Lambda = (ZZ^*)^{-1} = (R^*R)^{-1}$, we have $R\Lambda R^* = I$. As a result,

$$S = -\hat{B} + \sigma Z^*\Lambda Z = Q \begin{bmatrix} \sigma I - B_{11} & -B_{12} \\ -B_{21} & -B_{22} \end{bmatrix} Q^*. \tag{78}$$

Since the matrix $-B_{22}$ is positive definite we require non-negativity of the Schur complement $\sigma I - B_{11} + B_{12}B_{22}^{-1}B_{21}$ to $-B_{22}$. The Schur complement is non-negative definite if

$$\sigma \geq \lambda_{\max}(B_{11} - B_{12}B_{22}^{-1}B_{21}), \tag{79}$$

where $\lambda_{\max}$ denotes the largest eigenvalue of a matrix.

The inequality (79) guarantees that the matrix S is non-negative definite.

6 SAT terms for the interfaces between spatial domains

Consider now two adjacent rectangular domains $\Omega = [a, b] \times [c, d]$ and $\hat{\Omega} = [\hat{a}, \hat{b}] \times [c, d]$, where $b = \hat{a}$. The common boundary consists of the points (x, y) such that $x = b = \hat{a}$, $y \in [c, d]$. In the continuous case, we set the

following interface boundary conditions:

$$u(t,b,y) = \hat{u}(t,\hat{a},y) \text{ in domain } \Omega,$$
$$\hat{u}(t,\hat{a},y) = u(t,b,y) \text{ in domain } \hat{\Omega}.$$

Assume that semi-discretization of a hyperbolic symmetric system with SBP operators is done separately in Ω and $\hat{\Omega}$. We also assume dissipative boundary conditions on the external boundaries of Ω and $\hat{\Omega}$ and discretize these conditions with appropriate SAT terms. Solution of the semi-discretized problem in Ω is denoted by U and in $\hat{\Omega}$ by $\hat{U}$.

Following [4], we construct penalty terms at the interface in the form

$$\mathsf{B}_b(t,k) = \frac{1}{2}B(U^{[x_{\text{end}}]} - \hat{U}^{[\hat{x}_0]}), \quad \hat{\mathsf{B}}_{\hat{a}}(t,k) = -\frac{1}{2}B(\hat{U}^{[\hat{x}_0]} - U^{[x_{\text{end}}]}), \tag{80}$$

where the matrix B is taken at the corresponding interface nodes. The penalty functions are extended by zero from the interface boundary to the interior grid points. The extensions $F_b(t,j,k)$ and $\hat{F}_{\hat{a}}(t,\hat{j},k)$ allow us to form the corresponding SAT terms

$$\mathbb{SAT}_b = P_x^{-1} \times_2 F_b(t,j,k), \qquad \mathbb{SAT}_{\hat{a}} = P_{\hat{x}}^{-1} \times_2 \hat{F}_{\hat{a}}(t,\hat{j},k). \tag{81}$$

We add the energy identities provided by Theorem 7 separately for each of the two domains:

$$\frac{d}{dt}[(\mathcal{A}\vec{U},\vec{U})_{\mathcal{P}} + (\hat{\mathcal{A}}\vec{\hat{U}},\vec{\hat{U}})_{\hat{\mathcal{P}}})] + (BU,U)_{P_y}^{[x_{\text{end}}]} - (B\hat{U},\hat{U})_{P_y}^{[\hat{x}_0]} + \ldots = \Upsilon + \ldots. \tag{82}$$

The contribution of $\mathbb{SAT}_b$ and $\mathbb{SAT}_{\hat{a}}$ to the right-hand side of (82) is

$$\begin{aligned}
\Upsilon &= (\mathbb{SAT}_b, U)_{\mathcal{P}} + (\mathbb{SAT}_{\hat{a}}, \hat{U})_{\hat{\mathcal{P}}} + (U, \mathbb{SAT}_b)_{\mathcal{P}} + (\hat{U}, \mathbb{SAT}_{\hat{a}})_{\hat{\mathcal{P}}} \\
&= \frac{1}{2}(B(U^{[x_{\text{end}}]} - \hat{U}^{[\hat{x}_0]}), U^{[x_{\text{end}}]})_{P_y} + \frac{1}{2}(BU^{[x_{\text{end}}]}, U^{[x_{\text{end}}]} - \hat{U}^{[\hat{x}_0]})_{P_y} \\
&\quad - \frac{1}{2}(B(\hat{U}^{[\hat{x}_0]} - U^{[x_{\text{end}}]}), \hat{U}^{[\hat{x}_0]})_{P_y} - \frac{1}{2}(B\hat{U}^{[\hat{x}_0]}, \hat{U}^{[\hat{x}_0]} - U^{[x_{\text{end}}]})_{P_y} \\
&= (BU^{[x_{\text{end}}]}, U^{[x_{\text{end}}]})_{P_y} - (B\hat{U}^{[\hat{x}_0]}, \hat{U}^{[\hat{x}_0]})_{P_y}.
\end{aligned}$$

Thus, the boundary terms $(BU^{[x_{\text{end}}]}, U^{[x_{\text{end}}]})_{P_y} - (B\hat{U}^{[\hat{x}_0]}, \hat{U}^{[\hat{x}_0]})_{P_y}$ are completely compensated by the SAT terms on the interface.

7 Method of lines with explicit Runge-Kutta methods

To integrate the semi-discrete system (63) in time, we use explicit Runge-Kutta (RK) methods. Recall that an s-stage explicit Runge-Kutta method is applied to the system of ordinary differential equations $\frac{dy}{dt} = f(t,y)$, and the

integration step from t_n to $t_{n+1} = t_n + \tau$ consists of the following s stages:

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_i &= f\left(t_n + c_i\tau, y_n + \tau \sum_{j=1}^{i-1} a_{ij} k_j\right), \quad i = 2, \dots s, \\ y_{n+1} &= y_n + \tau \sum_{i=1}^{s} b_i k_i. \end{aligned}$$

The coefficients of the Runge-Kutta methods are conveniently represented by the *Butcher tableau*

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array}$$

Note that the entries a_{ij} for $j \geq i$ in the tableau for explicit Runge-Kutta methods are zero and are usually not displayed. A survey of *low-storage* implementations of explicit Runge-Kutta methods is given in [13].

If an explicit Runge-Kutta method with s stages is applied to a linear system $\dot{y} = \mathcal{L}y + b$ with constant matrix $\mathcal{L}$ and constant vector b, then it is equivalent to the one-step iteration $y_{n+1} = \mathscr{P}_s(\tau\mathcal{L})y_n + \check{\mathscr{P}}_{s-1}(\tau\mathcal{L})\tau b$, where $\mathscr{P}_s(z) = \sum_{i=0}^{s} p_i z^i$ is a polynomial with real coefficients, called the stability function in [11, Definition 2.1]. For a matrix A and a vector b from the Butcher tableau, [11, Proposition 3.1] yields the formula

$$\mathcal{P}_s(z) = 1 + zb^T(I - zA)^{-1}\mathbf{1},$$

where $\mathbf{1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$ is a 1-vector. The polynomial $\check{\mathscr{P}}_{s-1}(z)$ is related with $\mathscr{P}_s(z)$ by the formula $\check{\mathscr{P}}_{s-1}(z) = (\mathscr{P}_s(z) - 1)/z$.

An explicit Runge-Kutta method with s stages is accurate of order r if

$$\mathscr{P}_s(z) = \sum_{i=0}^{r} \frac{z^i}{i!} + \sum_{i=r+1}^{s} p_i z^i. \tag{83}$$

All 3-stage explicit Runge-Kutta methods of order 3 have the polynomials

$$\mathscr{P}_3(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6}, \quad \check{\mathscr{P}}_2(z) = 1 + \frac{z}{2} + \frac{z^2}{6}.$$

All 4-stage explicit Runge-Kutta methods of order 4 have the polynomials

$$\mathscr{P}_4(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}, \quad \check{\mathscr{P}}_3(z) = 1 + \frac{z}{2} + \frac{z^2}{6} + \frac{z^3}{24}.$$

7.1. Semi-bounded dynamic systems. The semi-discretized hyperbolic systems depend on step sizes h in the spatial domain. We therefore pay particular attention to the question of stability, which is uniform even for arbitrarily small h.

Following [27, 28], a dynamic system $\dot{u} = \mathcal{L}u$, which depends on parameters h, is called *semi-bounded* if there exist a constant η independent of h and a positive-definite hermitian matrix $\mathbb{H}$ such that

$$\mathbb{H}\mathcal{L} + \mathcal{L}^*\mathbb{H} \leq 2\eta\mathbb{H}.$$

Solution of $\dot{u} = \mathcal{L}u$, subject to arbitrary initial data $u(0) = u_0$, satisfies the estimate

$$\|u(t)\|_{\mathbb{H}} \leq e^{\eta t}\|u_0\|_{\mathbb{H}},$$

where $(u, v)_{\mathbb{H}} = u^*\mathbb{H}v$ denotes the inner product with the weight matrix $\mathbb{H}$ and $\|u\|_{\mathbb{H}} = \sqrt{(u,u)_{\mathbb{H}}}$. Indeed, $\frac{d}{dt}(u,u)_{\mathbb{H}} = (\mathcal{L}u, u)_{\mathbb{H}} + (u, \mathcal{L}u)_{\mathbb{H}} = ([\mathbb{H}\mathcal{L} + \mathcal{L}^*\mathbb{H}]u, u) \leq 2\eta(u,u)_{\mathbb{H}}$ and $\|u(t)\|^2_{\mathbb{H}} \leq e^{2\eta t}\|u_0\|^2_{\mathbb{H}}$.

When

$$\mathbb{H}\mathcal{L} + \mathcal{L}^*\mathbb{H} \leq 0, \tag{84}$$

the operator $\mathcal{L}$ is called *semi-dissipative*; see e.g. [1]. Solution of $\dot{u} = \mathcal{L}u$ governed by a semi-dissipative operator $\mathcal{L}$ satisfy the monotonic estimate

$$\|u(t)\|_{\mathbb{H}} \leq \|u_0\|_{\mathbb{H}}. \tag{85}$$

7.2. Local stability on the imaginary axis.

Definition 1. [16] *An explicit Runge-Kutta method is called locally stable on the imaginary axis if there exists a constant $R_s > 0$ such that*

$$|\mathscr{P}_s(i\sigma)| \leq 1, \quad -R_s \leq \sigma \leq R_s. \tag{86}$$

In other words, the region of absolute stability $\mathcal{A}_s = \{z \in \mathbb{C}\colon |\mathscr{P}_s| \leq 1\}$ of a locally stable Runge-Kutta method on the imaginary axis must contain a nontrivial interval $[-iR_s, iR_s]$. Condition (86) ensures uniform stability of the RK method applied to the scalar semi-dissipative equation $\dot{u} = i\frac{a}{h}u$, where $i = \sqrt{-1}$, a is a non-zero real number and the parameter $h > 0$ can be arbitrarily small.

A precise characterization of local stability on the imaginary axis is given in [16]. Namely, a general s-stage RK method satisfies Definition 1 if and only if

$$\begin{cases} (-1)^{\frac{r+1}{2}}(p_{r+1} - 1) < 0, & r \text{ is odd}, \\ (-1)^{\frac{r+2}{2}}[p_{r+2} - (r+2)p_{r+1} + r + 1] < 0, & r \text{ is even}. \end{cases}$$

The condition (86) rules out the forward Euler method as well as the 2-stage Heun method of order 2 (also known as the modified Euler method).

In the particular cases $s = r = 3$ and $s = r = 4$ the explicit Runge-Kutta method satisfies Definition 1.

8 Strong stability of the 3-stage third-order Runge-Kutta methods for time-invariant semi-dissipative systems

We begin with time-invariant systems $\frac{d}{dt}u = \mathcal{L}u$ and semi-dissipative matrices $\mathcal{L}$. To overcome difficulties of extending stability analysis to time-dependent systems and including lower-order terms, we restrict ourselves to the so called *strong stability* which requires u_n to decrease monotonically in an appropriate norm as

$$\|u_{n+1}\| \le \|u_n\|, \quad n \in \mathbb{N}.$$

Remark 4. *The semi-discretized symmetric hyperbolic system from (63) and the SAT terms, which are defined in (72) and (73), yield the system*

$$U_t = -\frac{1}{2}A^{-1}\left[\mathbb{D}_x BU + B\mathbb{D}_x U + \mathbb{D}_y CU + C\mathbb{D}_y U\right] + \frac{1}{2}A^{-1}\left[B_x U + C_y U - DU\right] + A^{-1}F + A^{-1}\mathbb{S}\mathbb{A}\mathbb{T}. \quad (87)$$

The system (87) satisfies the energy identity in terms of the weighted inner product $(\mathcal{A}\vec{U}, \vec{V})_{\mathcal{P}}$, as in Theorem 7. This inner product can be inherited into the method of lines via the following notation:

$$\langle u, v\rangle = (\mathcal{A}u, v)_{\mathcal{P}}, \quad \|u\| = \sqrt{\langle u, u\rangle}. \quad (88)$$

Under the strict dissipative boundary conditions, the operator $\mathcal{L}$ in (87) is semi-bounded in the inner product (88).

The stability analysis below is valid for arbitrary inner product $\langle\cdot,\cdot\rangle$ and for arbitrary $\mathcal{L}$ which is semi-dissipative in this inner product, i.e.

$$\langle \mathcal{L}v, v\rangle + \langle v, \mathcal{L}v\rangle \le 0 \text{ for all vectors } v. \quad (89)$$

One step of size τ with an explicit Runge-Kutta method applied to the system $\frac{d}{dt}u = \mathcal{L}u$ computes a vector u_+ such that

$$u_+ = \mathscr{P}(L)u, \quad L = \tau\mathcal{L}. \quad (90)$$

The polynomial $\mathscr{P}(z)$ in (90) is the stability function of the RK method. The square of the norm after one step fulfils

$$\|u_+\|^2 - \|u\|^2 = \|\mathscr{P}(L)u\|^2 - \|u\|^2 = \mathsf{Re}\,\langle(\mathscr{P}(L) - I)u, (\mathscr{P}(L) + I)u\rangle. \quad (91)$$

The following procedure is proposed in [20] as a generalization of the proof given by E. Tadmor and D. Levermore in [27].

Procedure 1. *In order to estimate the squared norm ("energy") after one step (90), proceed as follows:*

1. *If there is only one term of lowest order in L on either the left or the right hand side of the scalar product estimating $\|u_+\|^2 - \|u\|^2$, remove this lowest order term using the semi-boundedness of L.*
2. *If the lowest order terms on the left and right hand side are of equal order, estimate the remaining terms with respect to $\|L\| = \tau\|\mathcal{L}\|$.*

Theorem 8. [27, 20] *Consider the stability function of any 3-stage Runge-Kutta method of order 3,*

$$\mathscr{P}_3(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3.$$

Suppose that $\mathcal{L}$ is a semi-dissipative matrix in some inner product $\langle\cdot,\cdot\rangle$, i.e.

$$\mathsf{Re}\,\langle \mathcal{L}v, v\rangle \le 0 \text{ for all vectors } v. \tag{92}$$

Let the vector norm $\|u\| = \langle u, u\rangle^{1/2}$ be defined by the inner product used in (92), and $\|\mathcal{L}\|$ be the matrix norm induced by this vector norm. The Runge-Kutta solution of the time-invariant system $\dot{u} = \mathcal{L}u$ after one step of size τ is $u_+ = \mathscr{P}_3(\tau\mathcal{L})u$. Then $\|u_+\| \le \|u\|$ for arbitrary u if $\tau\|\mathcal{L}\| \le 1$.

Proof.

Lemma 2. *Let $\mathscr{P}(L) = \mathscr{P}_3(L) = I + L + \frac{1}{2}L^2 + \frac{1}{6}L^3$. Then for all u*

$$\begin{aligned}&\langle(\mathscr{P}(L) - I)u, (\mathscr{P}(L) + I)u\rangle\\ &\qquad = 2\langle Lv_1, v_1\rangle + 6\langle v_2, Lv_2\rangle + \langle(-\frac{1}{2}I + \frac{1}{6}L)v_3, \frac{1}{6}(I + L)v_3\rangle,\end{aligned}$$

where $v_1 = (I + \frac{1}{2}L + \frac{1}{6}L^2)u$, $v_2 = \frac{1}{6}(I + L)Lu$, $v_3 = L^2u$.

Proof. We apply Procedure 1.

$$\begin{aligned}&\langle(\mathscr{P}(L) - I)u,(\mathscr{P}(L) + I)u\rangle\\ &\qquad = \langle L(I + \frac{1}{2}L + \frac{1}{6}L^2)u, 2(I + \frac{1}{2}L + \frac{1}{4}L^2 + \frac{1}{12}L^3)u\rangle\\ &\qquad = 2\langle L(I + \frac{1}{2}L + \frac{1}{6}L^2)u, (I + \frac{1}{2}L + \frac{1}{6}L^2)u\rangle\\ &\qquad\quad + \langle L(I + \frac{1}{2}L + \frac{1}{6}L^2)u, (\frac{1}{6}L^2 + \frac{1}{6}L^3)u\rangle\end{aligned}$$

and

$$\begin{aligned}&\langle L(I + \frac{1}{2}L + \frac{1}{6}L^2)u, (\frac{1}{6}L^2 + \frac{1}{6}L^3)u\rangle = \langle(L + \frac{1}{2}L^2 + \frac{1}{6}L^3)u, L\frac{1}{6}(L + L^2)u\rangle\\ &\quad = 6\langle\frac{1}{6}(L + L^2)u, L\frac{1}{6}(L + L^2)u\rangle + \langle(-\frac{1}{2}L^2 + \frac{1}{6}L^3)u, \frac{1}{6}L(L + L^2)u\rangle.\end{aligned}$$

Finally,

$$\langle(-\frac{1}{2}L^2 + \frac{1}{6}L^3)u, \frac{1}{6}L(L + L^2)u\rangle = \langle(-\frac{1}{2}I + \frac{1}{6}L)L^2u, \frac{1}{6}(I + L)L^2u\rangle.$$

□

Lemma 2 and (91) imply the equality

$$\|u_+\|^2 - \|u\|^2 = 2\mathsf{Re}\langle Lv_1, v_1\rangle + 6\mathsf{Re}\langle v_2, Lv_2\rangle + \mathsf{Re}\langle(-\frac{1}{2}I + \frac{1}{6}L)v_3, \frac{1}{6}(I + L)v_3\rangle.$$

Since $\mathsf{Re}\langle Lv, v\rangle \le 0$ for all v, we have $2\mathsf{Re}\langle Lv_1, v_1\rangle + 6\mathsf{Re}\langle v_2, Lv_2\rangle \le 0$. It follows that

$$\begin{aligned}\|u_+\|^2 - \|u\|^2 &\le \mathsf{Re}\langle(-\frac{1}{2}I + \frac{1}{6}L)L^2u, \frac{1}{6}(I+L)L^2u\rangle \\ &\le \frac{1}{12}\left[-1 + \frac{2}{3}\|L\| + \frac{1}{3}\|L\|^2\right]\|L^2u\|^2.\end{aligned}$$

The polynomial $\frac{1}{3}x^2 + \frac{2}{3}x - 1 = (x-1)(\frac{1}{3}x+1)$ is not positive for $0 \le x \le 1$. Hence $\|u_+\|^2 - \|u\|^2 \le 0$ when $\|L\| \le 1$. □

For completeness, we prove an estimate of the contribution from the vector b in the system $\frac{d}{dt}u = \mathcal{L}u + b$.

Theorem 9. *Consider the polynomial*

$$\check{\mathscr{P}}_2(z) = 1 + \frac{1}{2}z + \frac{1}{6}z^2.$$

Under the assumptions of Theorem 8, $\|\check{\mathscr{P}}_2(z)u\| \le \|u\|$ for arbitrary u when $\tau\|\mathcal{L}\| \le 1$.

Proof.

Lemma 3. *Let $\mathscr{P}(L) = \check{\mathscr{P}}_2(L) = I + \frac{1}{2}L + \frac{1}{6}L^2$. Then for all vectors u*

$$\langle(\mathscr{P}(L) - I)u, (\mathscr{P}(L) + I)u\rangle = \langle Lv_1, v_1\rangle + \frac{1}{12}\langle(I + \frac{1}{3}L)v_2, (-I + L)v_2\rangle,$$

where $v_1 = (I + \frac{1}{3}L)u$, $v_2 = Lu$.

Proof. By Procedure 1,

$$\begin{aligned}\langle(\mathscr{P}(L) - I)u, (\mathscr{P}(L) + I)u\rangle &= \langle(\frac{1}{2}L + \frac{1}{6}L^2)u, (2I + \frac{1}{2}L + \frac{1}{6}L^2)u\rangle \\ &= \langle\frac{1}{2}L(I + \frac{1}{3}L)u, 2(I + \frac{1}{4}L + \frac{1}{12}L^2)u\rangle \\ &= \langle L(I + \frac{1}{3}L)u, (I + \frac{1}{3}L)u\rangle + \langle L(I + \frac{1}{3}L)u, (-\frac{1}{12}L + \frac{1}{12}L^2)u\rangle,\end{aligned}$$

$$\langle L(I + \frac{1}{3}L)u, (-\frac{1}{12}L + \frac{1}{12}L^2)u\rangle = \frac{1}{12}\langle(I + \frac{1}{3}L)Lu, (-I + L)Lu\rangle.$$

□

Lemma 3 and (91) give the equality

$$\|\check{\mathcal{P}}_2(L)u\|^2 - \|u\|^2 = 2\mathsf{Re}\langle Lv_1, v_1\rangle + \frac{1}{12}\mathsf{Re}\langle(I + \frac{1}{3}L)v_2, (-I + L)v_2\rangle.$$

Since $\mathsf{Re}\langle Lv, v\rangle \le 0$ for all v, we have $2\mathsf{Re}\langle Lv_1, v_1\rangle \le 0$. It follows that

$$\begin{aligned}\|\check{\mathscr{P}}_2(L)u\|^2 - \|u\|^2 &\le \frac{1}{12}\mathsf{Re}\langle(I + \frac{1}{3}L)L^2u, (-I + L)L^2u\rangle \\ &\le \frac{1}{12}\left[-1 + \frac{2}{3}\|L\| + \frac{1}{3}\|L\|^2\right]\|L^2u\|^2.\end{aligned}$$

The polynomial $\frac{1}{3}x^2 + \frac{2}{3}x - 1$ is not positive for $0 \le x \le 1$. Therefore, $\|\check{\mathscr{P}}_2(L)u\|^2 - \|u\|^2 \le 0$ when $\|L\| \le 1$. □

When applying a 3-stage Runge-Kutta method of order 3 to the system $\frac{d}{dt}y = \mathcal{L}y + b$ with semi-dissipative $\mathcal{L}$, we have the strong stability estimate

$$\|u_{n+1}\| \le \|u_n\| + \tau\|b\|, \quad n \in \mathbb{N}. \tag{93}$$

9 Stability of SSPRK3 for time-dependent systems

When the matrix $\mathcal{L}$ depends on time t, the Runge-Kutta scheme can no longer be written as the iteration $y_{n+1} = \mathscr{P}_s(\tau\mathcal{L})y_n$, where $\mathscr{P}_s(z)$ is a truncated series for the exponential. Moreover, for example, the various 3-stage third-order Runge-Kutta schemes are no longer equivalent.

We demonstrate how to investigate stability of time-dependent systems using the SSPRK3 Runge-Kutta method [22] as an example. Other RK methods can be treated similarly; see, e.g., [24]. The SSPRK3 method is defined by the stages

$$\begin{aligned} k_1 &= f(t_n, y_n), \quad k_2 = f(t_n + \tau, y_n + \tau k_1), \\ k_3 &= f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{4}(k_1 + k_2)), \\ y_{n+1} &= y_n + \frac{\tau}{6}\left[k_1 + k_2 + 4k_3\right]. \end{aligned} \tag{94}$$

One step of SSPRK3 reduces to the iteration $y_{n+1} = \mathcal{R}(t_n)y_n$ with

$$\begin{aligned} \mathcal{R}(t_n) = I &+ \frac{\tau}{6}\left[\mathcal{L}(t_n) + \mathcal{L}(t_{n+1}) + 4\mathcal{L}(t_{n+1/2})\right] \\ &+ \frac{\tau^2}{6}\left[\mathcal{L}(t_{n+1})\mathcal{L}(t_n) + \mathcal{L}(t_{n+1/2})\mathcal{L}(t_n) + \mathcal{L}(t_{n+1/2})\mathcal{L}(t_{n+1})\right] \\ &+ \frac{\tau^3}{6}\mathcal{L}(t_{n+1/2})\mathcal{L}(t_{n+1})\mathcal{L}(t_n). \end{aligned} \tag{95}$$

Proposition 1. *Suppose that $\mathcal{L}$ satisfies the Lipschitz condition, that is, there exists positive ζ such that for all t', t''*

$$\|\mathcal{L}(t') - \mathcal{L}(t''\| \le \zeta|t' - t''|. \tag{96}$$

Then

$$\begin{aligned} \|\mathcal{R}(t_n) - \mathscr{P}_3(\tau\mathcal{L}(t_n))\| \le \tau\frac{\tau\zeta}{4}\left[1 + (1 + \tau\|\mathcal{L}(t_n)\|)^2\right] \\ + \tau^2\frac{(\tau\zeta)^2}{12}(1 + \tau\|\mathcal{L}(t_n)\|). \end{aligned}$$

Proof. If we denote $\delta_1 = \mathcal{L}(t_{n+1}) - \mathcal{L}(t_n)$ and $\delta_{1/2} = \mathcal{L}(t_{n+1/2}) - \mathcal{L}(t_n)$, then (95) allows us to derive that

$$\begin{aligned}\mathcal{R}(t_n) = \mathscr{P}_3(\tau\mathcal{L}(t_n)) &+ \frac{\tau}{6}\left(\delta_1 + 4\delta_{1/2}\right)\\ &+ \frac{\tau^2}{6}\left[\delta_1\mathcal{L}(t_n) + \delta_{1/2}\mathcal{L}(t_n) + \delta_{1/2}\mathcal{L}(t_n) + \mathcal{L}(t_n)\delta_1 + \delta_{1/2}\delta_1\right]\\ &+ \frac{\tau^3}{6}[\delta_{1/2}\mathcal{L}(t_n) + \mathcal{L}(t_n)\delta_1 + \delta_{1/2}\delta_1]\mathcal{L}(t_n).\end{aligned}$$

Due to (96), $\|\delta_1\| \le \tau\zeta$ and $\|\delta_{1/2}\| \le \frac{\tau}{2}\zeta$. Therefore,

$$\begin{aligned}&\|\mathcal{R}(t_n) - \mathscr{P}_3(\tau\mathcal{L}(t_n))\|\\ &\quad\le \frac{\tau}{6}3\tau\zeta + \frac{\tau^2}{6}3\tau\zeta\|\mathcal{L}(t_n)\| + \frac{\tau^2}{6}\frac{1}{2}\tau^2\|\mathcal{L}(t_n)\|^2\\ &\qquad + \frac{\tau^3}{6}\frac{3}{2}\tau\zeta\eta^2 + \frac{\tau^3}{6}\frac{1}{2}\tau^2\zeta^2\|\mathcal{L}(t_n)\|\\ &\quad= \tau\frac{\tau\zeta}{4}\left[1 + (1 + \tau\|\mathcal{L}(t_n)\|)^2\right] + \tau^2\frac{(\tau\zeta)^2}{12}(1 + \tau\|\mathcal{L}(t_n)\|).\end{aligned}$$

□

Suppose that $\tau\|\mathcal{L}(t_n)\| \le 1$ and the matrix $\mathcal{L}(t_n)$ is semi-dissipative. Then $\|\mathscr{P}_3(\tau\mathcal{L}(t_n))\| \le 1$ by Theorem 8. By Proposition 1

$$\|\mathcal{R}(t_n)\| \le 1 + \tau\frac{5\tau\zeta}{4} + \tau^2\frac{(\tau\zeta)^2}{6}.$$

The structure of (87) implies that

$$\zeta = \frac{\zeta_1}{h_x} + \frac{\zeta_2}{h_y} + \zeta_3$$

with positive constants ζ_1, ζ_2 and ζ_3. The step-sizes h_x and h_y can be arbitrarily small. Hence $\tau\zeta = \frac{\tau}{h_x}\zeta_1 + \frac{\tau}{h_y}\zeta_2 + \tau\zeta_3$. Since $\tau\|\mathcal{L}(t_n)\| \le 1$, there is a constant $C > 0$ such that following inequality holds:

$$\tau \le C\min(h_x, h_y).$$

It follows that $\tau\zeta = C(\zeta_1 + \zeta_2) + \tau\zeta_3$ and

$$\|\mathcal{R}(t_n)\| \le 1 + \tau\left\{\frac{5}{4}[C(\zeta_1 + \zeta_2) + \tau\zeta_3] + \frac{1}{6}\tau[C(\zeta_1 + \zeta_2) + \tau\zeta_3]^2\right\} \tag{97}$$

$$= 1 + \tau C(\zeta_1 + \zeta_2) + O(\tau^2) = 1 + O(\tau). \tag{98}$$

The bound (98) ensures Lax-Richtmyer stability [19].

9.1. Lower order terms. The above derivation of bound (98) does not take into account the lower order terms

$$\mathcal{T}(t) = \frac{1}{2}A^{-1}\left[B_x U + C_y U - DU\right]$$

in (87). They are included by adding $\mathcal{T}(t)$ to all matrices $\mathcal{L}(t)$ in (95):

$$\begin{aligned}\tilde{\mathcal{R}}(t_n) = I &+ \frac{\tau}{6}\left[\mathcal{L}(t_n) + \mathcal{T}(t_n) + \mathcal{L}(t_{n+1}) + \mathcal{T}(t_{n+1}) + \ldots\right] \\ &+ \frac{\tau^2}{6}\left[(\mathcal{L}(t_{n+1}) + \mathcal{T}(t_{n+1}))(\mathcal{L}(t_n) + \mathcal{T}(t_n)) + \ldots\right] \\ &+ \frac{\tau^3}{6}(\mathcal{L}(t_{n+1/2}) + \mathcal{T}(t_{n+1/2}))(\mathcal{L}(t_{n+1}) + \mathcal{T}(t_{n+1}))(\mathcal{L}(t_n) + \mathcal{T}(t_n)). \quad (99)\end{aligned}$$

For convenience, we denote $\theta = \max\{\|\mathcal{L}(t)\| \colon t \in \{t_n, t_{n+1/2}, t_{n+1}\}\}$. Let us suppose that

$$\tau\theta \le 1 \text{ and } \|\mathcal{T}(t)\| \le \eta \text{ for } t \in \{t_n, t_{n+1/2}, t_{n+1}\}.$$

Then

$$\begin{aligned}\|\tilde{\mathcal{R}}(t_n) - \mathcal{R}(t_n)\| &\le \frac{\tau}{6}3\eta + \frac{\tau^2}{6}3[2\theta\eta + \eta^2] + \frac{\tau^3}{6}3[3\theta^2\eta + 3\theta\eta^2 + \eta^3] \\ &\le \frac{\tau}{2}\left[\eta + 2\eta + \tau\eta^2 + 3\eta + 3\tau\eta^2 + \tau^2\eta^3\right] \\ &= \frac{\tau}{2}\left[6\eta + 4\tau\eta^2 + \tau^2\eta^3\right] = 3\tau\eta + O(\tau^2). \quad (100)\end{aligned}$$

Combining (100) with (98) gives the stability bound $\|\tilde{\mathcal{R}}(t_n)\| \le 1 + O(\tau)$.

10 On the stability of the classical Runge-Kutta method

The classical fourth-order Runge-Kutta method [17] has 4 stages:

$$\begin{aligned}k_1 &= f(t_n, y_n), & k_2 &= f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}k_1), \\ k_3 &= f(t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2}k_2), & k_4 &= f(t_n + \tau, y_n + \tau k_3), \quad (101) \\ y_{n+1} &= y_n + \frac{\tau}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right].\end{aligned}$$

The method has the stability function $\mathscr{P}_4(z) = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!}$. In [24] it is shown that $\|\mathscr{P}_4(\tau\mathcal{L})\| > 1$, in the 2-norm, for the semi-dissipative matrix

$$\mathcal{L} = -\begin{bmatrix}1 & 2 & 2\\ 0 & 1 & 2\\ 0 & 0 & 1\end{bmatrix}$$

and for arbitrarily small $\tau > 0$.

However, $\|\mathscr{P}_4^2(\tau\mathcal{L})\| \le 1$ for arbitrary semi-dissipative $\mathcal{L}$ and sufficiently small $\tau > 0$. This fact was first proved in [24]. Below we give an alternative proof which is based on Procedure 1. Note that the proof of Lemma 4 is relatively simple when using a computer program for symbolic computation.

Lemma 4. *Let* $\mathscr{P}(L) = I + L + \frac{1}{2}L^2 + \frac{1}{6}L^3 + \frac{1}{24}L^4$. *Then for all vectors* u

$$\langle(\mathscr{P}^2(L) - I)u, (\mathscr{P}^2(L) + I)u\rangle = \langle Lv_1, v_1\rangle + 3\langle v_2, Lv_2\rangle + 8\langle Lv_3, v_3\rangle + \langle(p_5(L)L^3u, q_5(L)L^3u\rangle,$$

where $\langle\cdot,\cdot\rangle$ *is any inner product and*

$$\begin{aligned}
v_1 &= (2I + 2L + \frac{4}{3}L^2 + \frac{2}{3}L^3 + \frac{1}{4}L^4 + \frac{5}{72}L^5 + \frac{1}{72}L^6 + \frac{1}{576}L^7)u,\\
v_2 &= (\frac{2}{3}I + \frac{2}{3}L + \frac{5}{12}L^2 + \frac{13}{72}L^3 + \frac{1}{18}L^4 + \frac{7}{576}L^5 + \frac{1}{576}L^6)Lu,\\
v_3 &= (\frac{1}{12}I + \frac{1}{8}L + \frac{1}{12}L^2 + \frac{19}{576}L^3 + \frac{5}{576}L^4 + \frac{1}{576}L^5)L^2u,\\
p_5(z) &= \frac{1}{12} + \frac{1}{8}z + \frac{1}{12}z^2 + \frac{19}{576}z^3 + \frac{5}{576}z^4 + \frac{1}{576}z^5,\\
q_5(z) &= -\frac{1}{3} - \frac{1}{4}z - \frac{1}{12}z^2 - \frac{1}{72}z^3 - \frac{1}{576}z^4 + \frac{1}{576}z^5.
\end{aligned}$$

Since the matrix L is semi-dissipative, we obtain the inequality

$$\mathsf{Re}\langle Lv_1, v_1\rangle + 3\langle v_2, Lv_2\rangle + 8\langle Lv_3, v_3\rangle \le 0.$$

Then we have $\langle(p_5(L)v, q_5(L)v\rangle \le r(\|L\|)\|v\|^2$, where the polynomial $r(z)$ is obtained from the product $p_5(z)q_5(z)$ by replacing its coefficients with their absolute values, but excluding the coefficient $-1/36$ for zero degree:

$$r(z) = -\frac{1}{36} + \frac{z}{16} + \frac{19z^2}{288} + \frac{25z^3}{576} + \frac{23z^4}{1152} + \frac{31z^5}{4608} + \frac{z^6}{648} + \frac{59z^7}{331776} + \frac{z^8}{55296} + \frac{z^9}{82944} + \frac{z^{10}}{331776}.$$

Since $r(x) < 0$ for $x \in [0, 0.3147]$, we arrive at the desired stability bound

$$\|\mathscr{P}_4^2(\tau\mathcal{L})\| < 1 \text{ when } \|\tau\mathcal{L}\| \le 0.3147. \tag{102}$$

References

[1] F. Achleitner, A. Arnold, A. Jüngel, *Necessary and sufficient conditions for strong stability of explicit Runge-Kutta methods*, In Carlen, E., Gonçalves, P., Soares, A.J. (eds.) From Particle Systems to Partial Differential Equations. PSPDE 2022, Springer Proc. in Math. & Statistics, vol. **465**, Springer, Cham, (2024), 1–21.

[2] S. Benzoni-Gavage, D. Serre, *Multidimensional Hyperbolic Partial Differential Equations. First-order Systems and Applications*, Oxford Univ. Press, NY, 2007.

[3] M.H. Carpenter, D. Gottlieb, S. Abarbanel, *Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes*, J. Comput. Phys., **111** (1994), 220–236.

[4] M.H. Carpenter, J. Nordström, D. Gottlieb, *A stable and conservative interface treatment of arbitrary spatial accuracy*, J. Comput. Phys., **148** (1999), 341–365.

[5] D.C. Del Rey Fernandez, P.D. Boom, D.W. Zingg, *A generalized framework for nodal first derivative summation-by-parts operators*, J. Comput. Phys., **266** (2014), 214–239.

[6] D.C. Del Rey Fernandez, J.E. Hicken, 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), 171–196.
[7] K.O. Friedrichs, *Symmetric hyperbolic linear differential equations*, Comm. Pure Appl. Math., **7** (1954), 345–392.
[8] J. Glaubitz, J. Nordström, P. Öffner, *Summation-by-parts operators for general function spaces*, SIAM J. Numer. Anal., **61**:2 (2023), 733–754.
[9] S.K. Godunov, *Equations of Mathematical Physics*, 2nd ed., Nauka, Moscow, 1979.
[10] B. Gustafsson, H.-O. Kreiss, J. Oliger, *Time Dependent Problems and Difference Methods*, John Wiley & Sons, Inc., 1995.
[11] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, 2nd ed., Springer, Berlin, 1996.
[12] R.A. Horn, C.R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, NY, 2013.
[13] C.A. Kennedy, M.H. Carpenter, R.M. Lewis, *Low-storage, explicit Runge-Kutta schemes for the compressible Navier-Stokes equations*, Appl. Numer. Math., **35** (2000), 177–219.
[14] H.-O. Kreiss, G. Scherer, *Finite element and finite difference methods for hyperbolic partial differential equations*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, NY (1974), 195–212.
[15] H.-O. Kreiss, G. Scherer, *On the existence of energy estimates for difference approximations for hyperbolic systems*, manuscript, 1977.
[16] H.-O. Kreiss, G. Scherer, *Method of lines for hyperbolic differential equations*, SIAM J. Numer. Anal., **29**:3 (1992), 640–646.
[17] M. Kutta, *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*, Zeitschrift für Mathematik und Physik, **46** (1901), 435–453.
[18] P.D. Lax, R.S. Phillips, *Local boundary conditions for dissipative symmetric linear differential operators*, Comm. Pure Appl. Math., **13** (1960), 427–455.
[19] P.D. Lax, R.D. Richtmyer, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math., **9**:2 (1956), 267–293.
[20] H. Ranocha, P. Öffner, *L_2 stability of explicit Runge-Kutta schemes*, J. Sci. Comput., **75** (2018), 1040–1056.
[21] G. Scherer, *On Energy Estimates for Difference Approximations to Hyperbolic Partial Differential Equations*, PhD Thesis, Uppsala Univ., Uppsala, Sweden, 1977.
[22] Chi-Wang Shu, S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., **77**, (1988), 439–471.
[23] B. Strand, *Summation by parts for finite difference approximations for d/dx*, J. Comput. Phys., **110**:1 (1994), 47–67.
[24] Z. Sun, Chi-Wang Shu, *Stability of the fourth order Runge-Kutta method for time-dependent partial differential equations*, Annals Math. Sci. Appl., **2**:2 (2017), 255–284.
[25] M. Svärd, J. Nordström, *Review of summation-by-parts schemes for initial-boundary value problems*, J. Comput. Phys., **268** (2014), 17–38.
[26] M. Svärd, J. Nordström, *On the convergence rates of energy-stable finite-difference schemes*, J. Comput. Physics, **397**, (2019), 108819.
[27] E. Tadmor, *From semi-discrete to fully discrete: stability of Runge-Kutta schemes by the energy method. II* In "Collected Lectures on the Preservation of Stability under Discretization", Lecture Notes from Colorado State Univ. Conf., Fort Collins, CO, 2001 (D. Estep and S. Tavener, eds.) Proc. Appl. Math. **109**, SIAM (2002), 25–49.
[28] E. Tadmor, *Runge-Kutta methods are stable*, arXiv:2312.15546v1, 2023.

Aleksandr Malyshev
University of Bergen,
Department of Mathematics,
Postbox 7803,
5020, Bergen, Norway
Email address: alexander.malyshev@uib.no