

Interior Penalties for Summation-by-Parts Discretizations of Linear Second-Order Differential Equations

Jianfeng Yan,* Jared Crean,* and Jason E. Hicken†
Rensselaer Polytechnic Institute, Troy, New York, 12180

When discretizing equations with second-order derivatives, like the Navier-Stokes equations, the boundary and interior penalties play a critical role in terms of the conservation, consistency, energy stability, and adjoint consistency of summation-by-parts (SBP) methods. While interior penalties for tensor-product SBP operators have been studied, they have not been investigated in the context of multidimensional SBP operators. This paper presents a general discretization for analyzing interior penalties for multidimensional SBP operators that can be used to obtain several favorable properties. Under this discretization, we construct a stable and adjoint consistent high-order finite difference scheme for linear elliptic equations with a constant diffusion coefficient. Specifically, the equations are first discretized using multidimensional SBP operators with interior and boundary penalty matrices to be determined. Then, taking advantage of the properties of SBP operators, the analyses are generalized from those in the finite element literature and become entirely algebraic in nature. That is, the analyses give rise to explicit conditions on the penalty matrix coefficients and these conditions are free of integral. To validate these conditions numerically, several test cases are conducted.

I. Introduction

High-order methods are well suited for simulations with complex physics requiring high resolution, like turbulent flows and acoustics. High-order methods are also able to generate more accurate solutions for a given cost than low-order methods, at least for sufficiently smooth problems. Among high-order methods, finite element (FE) methods are recognized as possessing several desirable properties, including *a priori* and *a posteriori* estimates of discretization errors and convergence rate, modularity, compact stencils, and *hp* adaptation.

Summation-by-parts (SBP) finite-difference methods share many of these properties with FE methods, but they also offer their own unique traits. The stability and high-order accuracy of SBP discretizations make them attractive for simulating conservation laws over long time periods. Additionally, FE methods must employ cubature rules to evaluate non-linear terms, in general, which leads to an aliasing error. Aliasing errors may degrade the accuracy of the solution and, in the worst case, can lead to numerical instability if the physics are not adequately resolved.^{1,2} SBP methods, on the other hand, account for the inexact integration from the beginning and can be devised to avoid such problems. Indeed, it was recently shown how to construct entropy-conservative and entropy-stable SBP discretizations of the Euler and Navier-Stokes equations.³⁻⁵

The classic SBP operators⁶⁻⁸ are based on tensor products of one-dimensional SBP operators and can only be used with structured and multi-block grids. Hence, their application to problems with complex geometry has been limited. Recently, Hicken, Del Rey Fernández, and Zingg⁹ introduced multidimensional SBP operators by generalizing the SBP definition to arbitrary bounded domains. They showed that a degree p diagonal-norm SBP operator can be constructed on any given domain provided a degree $2p - 1$ cubature

*Graduate Research Assistant, Department of Mechanical, Aerospace and Nuclear Engineering, AIAA Student Member

†Assistant Professor, Department of Mechanical, Aerospace and Nuclear Engineering, AIAA Member

exists whose nodes produce a full-rank Vandermonde matrix. Multidimensional SBP operators inherit the advantages of FE methods mentioned above.

When discretizing second-order derivatives, like the viscous terms in the Navier-Stokes equations, the main difficulty for high-order SBP operators in discontinuous space is the stable and accurate enforcement of boundary conditions and inter-element coupling. A common approach is to introduce so-called simultaneous approximation terms (SATs), which are also known as interior penalties in the FE community. Generally speaking, sufficiently large penalties are able to ensure the stability of solutions. However, the conditioning of the resulting linear system significantly deteriorates as the penalties increase in magnitude. Hence, it is important to find the smallest possible (computable) bound for these penalties.

In addition to their critical role in stability, the SAT penalties make it possible to obtain adjoint consistency and achieve continuity toward C^0 or even C^1 on the discontinuous elements.¹⁰ Adjoint consistency is a favorable property first introduced and analyzed in the FE community, where the conditions necessary for adjoint consistency have been intensively studied; see the review [10] and the references therein.

Given the strong connection with FE discretizations, it should not be surprising that SBP discretizations also require careful treatment of SATs to ensure stability and accuracy. SATs for multidimensional SBP discretizations of the linear advection were recently studied in [11, 12], and SATs for tensor-product SBP discretizations of second-order PDEs have been investigated by a number of authors; see, e.g., [13] and the review [14]. Multidimensional SATs for discretizations of elliptic equations, to the best of our knowledge, have not been studied.

The current work presents the requirements on *dense* SAT penalties to obtain multidimensional SBP-SAT discretizations that are simultaneously consistent, conservative, adjoint consistent, and stable. The significance of this study is threefold: we introduce a general discretization with a linear combination of penalties; we show that the methodology facilitates and generalizes the analysis processes; we derive rigorous computable bounds on the penalties that yield robustness and optimal errors in L^2 norms.

The remainder of the paper is organized as follows. The multidimensional SBP discretization is given in Section II; in Section III we analyze adjoint consistency and present corresponding restrictions on the penalties; in Section IV, the conditions from adjoint analysis are utilized to simplify the energy-stability analysis which further constrains the SAT penalties. Numerical test cases are carried out in Section V, and conclusions are provided in Section VI.

II. Multi-dimensional SBP discretization of elliptic PDEs

A. Notation

Matrices are represented with an uppercase sans-serif type, for example $\mathbf{A} \in \mathbb{R}^{n \times m}$. Functions are denoted with capital letters in calligraphic font; for example $\mathcal{U} \in L^2(\Omega)$ is a square-integrable function on the domain Ω . The space of polynomials of total degree p in x and y on Ω is denoted by $\mathbb{P}_p(\Omega)$. A function evaluated on a node set is denoted by a lowercase letter in bold font. For example, the function \mathcal{U} evaluated at the nodes of $S = \{(x_i, y_i)\}_{i=1}^n$ is given by

$$\mathbf{u} = \left[\mathcal{U}(x_1, y_1) \quad \mathcal{U}(x_2, y_2) \quad \cdots \quad \mathcal{U}(x_n, y_n) \right]^T.$$

As with generic functions, a polynomial that is evaluated at the points of S will be represented using its corresponding lowercase letter in bold font; for example, for $\mathcal{P} \in \mathbb{P}_p(\Omega)$ we would have

$$\mathbf{p} \equiv \left[\mathcal{P}(x_1, y_1) \quad \mathcal{P}(x_2, y_2) \quad \cdots \quad \mathcal{P}(x_n, y_n) \right]^T.$$

B. SBP definition and face operators

The definition for an SBP operator approximating $\partial/\partial x$ on a two dimensional domain is provided below.

Definition 1. Two-dimensional summation-by-parts operator: Consider an open and bounded domain $\kappa \subset \mathbb{R}^2$ with a piecewise-smooth boundary $\partial\kappa$. The matrix \mathbf{D}_x is a degree p SBP approximation to the first derivative $\frac{\partial}{\partial x}$ on the nodes $S_\kappa = \{(x_i, y_i)\}_{i=1}^{n_\kappa}$ if

1. For all $\mathcal{P} \in \mathbb{P}_p(\kappa)$, the vector $\mathbf{D}_x \mathbf{p}$ is equal to $\partial \mathcal{P} / \partial x$ at the nodes S_κ ;

2. $D_x = H^{-1}Q_x$, where H is symmetric positive-definite, and;
3. $Q_x = S_x + \frac{1}{2}E_x$, where $S_x^T = -S_x$, $E_x^T = E_x$, and E_x satisfies

$$\mathbf{p}^T E_x \mathbf{q} = \oint_{\partial\kappa} \mathcal{P} Q n_x d\Gamma, \quad \forall \mathcal{P}, Q \in \mathbb{P}_r(\kappa),$$

where $r \geq p$, and n_x is the x component of $\mathbf{n} = [n_x, n_y]^T$, the outward pointing unit normal on $\partial\kappa$.

The definition for the SBP operator approximating $\partial/\partial y$ is analogous.

In this paper we consider only diagonal norm SBP operators, that is, SBP operators for which H is a diagonal matrix with positive entries. In [9] it was shown that, for a diagonal-norm SBP operator, the nodes S_κ and diagonal entries of H define a cubature rule that is exact for polynomials of total degree $2p - 1$ (at least).

To facilitate the definition of the SAT penalties, we follow [11] and introduce interpolation/extrapolation operators from the SBP element nodes to cubature nodes on the faces of the elements; in some cases the face cubature nodes may be a subset of the SBP volume nodes, in which case the interpolation/extrapolation operators become trivial. Consider an element κ with a piecewise smooth boundary $\partial\kappa$, and let $\gamma \subset \partial\kappa$ denote one of its faces. Let $S_\gamma = \{(x_j, y_j)\}_{j=1}^{n_\gamma} \subset \gamma$ be a set of cubature nodes, and let $\{b_j\}_{j=1}^{n_\gamma}$ be a corresponding set of positive cubature weights that is exact for polynomials of degree $2r$, where $r \geq p$. The matrix $R_{\gamma\kappa} \in \mathbb{R}^{n_\gamma \times n_\kappa}$ is a degree r interpolation/extrapolation operator from the SBP nodes S_κ to the face nodes S_γ if

$$(R_{\gamma\kappa} \mathbf{p}_\kappa)_j = \sum_{i=1}^{n_\kappa} (R_{\gamma\kappa})_{ji} \mathcal{P}(x_i, y_i) = \mathcal{P}(x_j, y_j), \quad \forall j = 1, 2, \dots, n_\gamma,$$

and for all $\mathcal{P} \in \mathbb{P}_r(\kappa)$.

Using the above interpolation/extrapolation operators and face cubature rules, it was shown in [12] that there exists at least one SBP operator whose corresponding matrix E_x has the decomposition

$$E_x = \sum_{\gamma \subset \partial\kappa} R_{\gamma\kappa}^T N_{x,\gamma} B_\gamma R_{\gamma\kappa}, \quad (1)$$

where $B_\gamma = \text{diag}(b_1, b_2, \dots, b_{n_\gamma})$ is an $n_\gamma \times n_\gamma$ diagonal matrix holding cubature weights for γ along its diagonal, and $N_{x,\gamma} = \text{diag}(n_{x,1}, n_{x,2}, \dots, n_{x,n_\gamma})$ is an $n_\gamma \times n_\gamma$ diagonal matrix holding the x component of the outward unit normal with respect to κ at the cubature points of γ . The following analysis assumes that the SBP operators are such that E_x has the decomposition (1), and that the operators in the y direction have analogous decompositions.

C. The model PDE

Let Ω be a polygonal domain in \mathbb{R}^2 . The boundary of Ω , $\partial\Omega = \Gamma$, is partitioned into a Dirichlet boundary Γ^D and a Neumann boundary Γ^N , which are disjoint. Let $\hat{\mathbf{n}} = [n_x, n_y]^T$ be the outward pointing unit normal on $\partial\Omega$. The following linear parabolic problem is considered in this work:

$$\begin{aligned} \frac{\partial \mathcal{U}}{\partial t} - \lambda \nabla \cdot (\nabla \mathcal{U}) &= \mathcal{F}, & \text{in } \Omega, \\ \mathcal{U}(0, x, y) &= \mathcal{U}_0(x, y), & \text{in } \Omega, \\ \mathcal{U}(t, x, y) &= \mathcal{U}_D(t, x, y) & \text{on } \Gamma^D, \\ \hat{\mathbf{n}} \cdot (\lambda \nabla \mathcal{U}(t, x, y)) &= \mathcal{U}_N(t, x, y) & \text{on } \Gamma^N, \end{aligned} \quad (2)$$

where $\mathcal{F} \in L^2(\Omega \times [0, T])$ is a given source term; $\mathcal{U}_D \in L^2(\Gamma^D \times [0, T])$ and $\mathcal{U}_N \in L^2(\Gamma^N \times [0, T])$ are Dirichlet and Neumann boundary conditions, respectively; the diffusion coefficient λ is a positive constant. We assume that the Dirichlet boundary is nonempty, namely, $\Gamma^D \neq \emptyset$, so that the problem is well-posed.

D. Strong-form Discretization

By Ω_h we denote a polygonal approximation of the domain Ω ; let $\mathcal{T}_h = \bigcup_{\kappa=1}^K \kappa$ be a triangulation of the domain Ω_h into K SBP elements, where κ denotes the domain of an element. The discrete solution on κ is given by the vector $\mathbf{u}_\kappa \in \mathbb{R}^{n_\kappa}$ whose entries are the discrete solution at the SBP nodes S_κ . The global discrete solution, denoted \mathbf{u}_h , is the concatenation of all elementwise solutions.

The SBP-SAT discretization of (2) on element κ is given by

$$\frac{d\mathbf{u}_\kappa}{dt} = \mathbf{D}_\kappa \mathbf{u}_\kappa + \mathbf{f}_\kappa - \mathbf{H}_\kappa^{-1} \mathbf{s}_\kappa^{\mathcal{I}}(\mathbf{u}_h) - \mathbf{H}_\kappa^{-1} \mathbf{s}_\kappa^{\mathcal{B}}(\mathbf{u}_h, \mathbf{u}_{\mathcal{D}}, \mathbf{u}_{\mathcal{N}}), \quad (3)$$

where \mathbf{f}_κ is \mathcal{F} evaluated at the nodes of element κ and

$$\mathbf{D}_\kappa = \lambda(\mathbf{D}_x \mathbf{D}_x + \mathbf{D}_y \mathbf{D}_y) \quad (4)$$

is the SBP approximation of $\lambda \nabla \cdot \nabla$ on element κ . More generally, the subscript notation $(\cdot)_\kappa$ indicates a vector or operator on element κ .

The vectors $\mathbf{s}_\kappa^{\mathcal{I}}$ and $\mathbf{s}_\kappa^{\mathcal{B}}$ are the interface and boundary SAT penalties, respectively. For element κ these penalties are defined by

$$\mathbf{s}_\kappa^{\mathcal{I}}(\mathbf{u}_h) = \sum_{\gamma \in \Gamma_\kappa^{\mathcal{I}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa}^T & \mathbf{D}_{\gamma\kappa}^T \end{bmatrix} \begin{bmatrix} \Sigma_{\gamma\kappa}^{(1)} & \Sigma_{\gamma\kappa}^{(3)} \\ \Sigma_{\gamma\kappa}^{(2)} & \Sigma_{\gamma\kappa}^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{u}_\kappa - \mathbf{R}_{\gamma\nu} \mathbf{u}_\nu \\ \mathbf{D}_{\gamma\kappa} \mathbf{u}_\kappa + \mathbf{D}_{\gamma\nu} \mathbf{u}_\nu \end{bmatrix}$$

and

$$\mathbf{s}_\kappa^{\mathcal{B}}(\mathbf{u}_h, \mathbf{u}_{\mathcal{D}}, \mathbf{u}_{\mathcal{N}}) = \sum_{\gamma \in \Gamma_\kappa^{\mathcal{D}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa}^T & \mathbf{D}_{\gamma\kappa}^T \end{bmatrix} \begin{bmatrix} \Sigma_\gamma^{\mathcal{D}} \\ -\mathbf{B}_\gamma \end{bmatrix} (\mathbf{R}_{\gamma\kappa} \mathbf{u}_\kappa - \mathbf{u}_{\gamma\mathcal{D}}) + \sum_{\gamma \in \Gamma_\kappa^{\mathcal{N}}} \mathbf{R}_{\gamma\kappa}^T \mathbf{B}_\gamma (\mathbf{D}_{\gamma\kappa} \mathbf{u}_\kappa - \mathbf{u}_{\gamma\mathcal{N}}),$$

respectively. The discretization (3) is consistent, since both interface and boundary penalties vanish for the exact smooth solution. The set $\Gamma_\kappa = \partial\kappa$ represents the boundary of element κ , while the set of faces that coincide with Dirichlet and Neumann boundaries are denoted as $\Gamma_\kappa^{\mathcal{D}} = \Gamma_\kappa \cap \Gamma^{\mathcal{D}}$ and $\Gamma_\kappa^{\mathcal{N}} = \Gamma_\kappa \cap \Gamma^{\mathcal{N}}$, respectively. $\Gamma_\kappa^{\mathcal{I}}$ is the union of all interfaces of κ . The index ν is used to denote a generic element sharing face γ with the element κ , i.e., $\gamma = \kappa \cap \nu$. The vectors $\mathbf{u}_{\gamma\mathcal{D}}$ and $\mathbf{u}_{\gamma\mathcal{N}}$ in the boundary penalties denote the functions $\mathcal{U}_{\mathcal{D}}$ and $\mathcal{U}_{\mathcal{N}}$, respectively, evaluated at the cubature nodes of face γ .

For future use, we note that the normal derivative operators on face γ that discretize $\vec{n} \cdot (\lambda \nabla)$ for elements κ and ν , respectively, are given by

$$\begin{aligned} \mathbf{D}_{\gamma\kappa} &= \lambda(\mathbf{N}_{\gamma,x} \mathbf{R}_{\gamma\kappa} \mathbf{D}_{x,\kappa} + \mathbf{N}_{\gamma,y} \mathbf{R}_{\gamma\kappa} \mathbf{D}_{y,\kappa}), \\ \mathbf{D}_{\gamma\nu} &= -\lambda(\mathbf{N}_{\gamma,x} \mathbf{R}_{\gamma\nu} \mathbf{D}_{x,\nu} + \mathbf{N}_{\gamma,y} \mathbf{R}_{\gamma\nu} \mathbf{D}_{y,\nu}). \end{aligned}$$

Recall that $\mathbf{N}_{\gamma,x}$ (resp. $\mathbf{N}_{\gamma,y}$) is a diagonal matrix holding the x (resp. y) component of the unit outward normal, with respect to κ , at the cubature nodes of face γ . Thus, the matrix $\mathbf{D}_{\gamma\nu}$ must be negated.

The objective of the subsequent analysis is to determine the matrices $\Sigma_{\gamma\kappa}^{(i)} = \left(\Sigma_{\gamma\kappa}^{(i)}\right)^T \in \mathbb{R}^{n_\gamma \times n_\gamma}$, $i = 1, 2, 3, 4$, which denote the symmetric SAT coefficient matrices for element κ on face γ . Similarly, $\Sigma_\gamma^{\mathcal{D}}$ is the coefficient matrix for the SAT on a boundary face of κ . Note that $\Sigma_{\gamma\kappa}^{(i)} \neq \Sigma_{\gamma\nu}^{(i)}$ in general; that is, we do not assume that the coefficient matrices of two adjacent elements are necessarily equal.

E. Weak forms of the discretization

The discretization (3) is the element-based strong form. For the subsequent analysis, two equivalent face-based weak forms will prove more useful. Before deriving these weak formulations, we introduce two identities that will be helpful.

Let \mathbf{D}_κ be defined as in (4). Then, $\forall \mathbf{u}_\kappa, \mathbf{v}_\kappa \in \mathbb{R}^{n_\kappa}$,

$$\mathbf{v}_\kappa^T \mathbf{H}_\kappa \mathbf{D}_\kappa \mathbf{u}_\kappa = -\mathbf{v}_\kappa^T \mathbf{M}_\kappa \mathbf{u}_\kappa + \sum_{\gamma \in \Gamma_\kappa} \mathbf{v}_\kappa^T \mathbf{R}_{\gamma\kappa}^T \mathbf{B}_\gamma \mathbf{D}_{\gamma\kappa} \mathbf{u}_\kappa, \quad (5)$$

$$\text{and} \quad -\mathbf{v}_\kappa^T \mathbf{M}_\kappa \mathbf{u}_\kappa = \mathbf{v}_\kappa^T \mathbf{D}_\kappa^T \mathbf{H}_\kappa \mathbf{u}_\kappa - \sum_{\gamma \in \Gamma_\kappa} \mathbf{v}_\kappa^T \mathbf{D}_{\gamma\kappa}^T \mathbf{B}_\gamma \mathbf{R}_{\gamma\kappa} \mathbf{u}_\kappa, \quad (6)$$

where M_κ is the symmetric semi-definite matrix

$$M_\kappa = \lambda(D_x^T H D_x + D_y^T H D_y), \quad (7)$$

Identities (5) and (6) follow from straightforward application of the properties of SBP operators. They are the SBP analogs of applying integration by parts to $\int_\kappa \mathcal{V} \nabla \cdot (\lambda \nabla \mathcal{U}) d\Omega$ once (the first identity) and twice (the second identity).

To obtain the element-based weak formulation, we first left multiply (3) by $\mathbf{v}_\kappa^T H_\kappa$, where $\mathbf{v}_\kappa \in \mathbb{R}^{n_\kappa}$ is an arbitrary vector, and then apply (5). This produces the following weak form of the discretization: for all $\kappa = 1, 2, \dots, K$, find $\mathbf{u}_\kappa \in \mathbb{R}^{n_\kappa}$ such that, $\forall \mathbf{v}_\kappa \in \mathbb{R}^{n_\kappa}$,

$$\mathbf{v}_\kappa^T H_\kappa \frac{d\mathbf{u}_\kappa}{dt} = -\mathbf{v}_\kappa^T M_\kappa \mathbf{u}_\kappa + \sum_{\gamma \subset \Gamma_\kappa} \mathbf{v}_\kappa^T R_{\gamma\kappa}^T B_\gamma D_{\gamma\kappa} \mathbf{u}_\kappa + \mathbf{v}_\kappa^T H_\kappa \mathbf{f}_\kappa - \mathbf{v}_\kappa^T \mathbf{s}_\kappa^I(\mathbf{u}_h) - \mathbf{v}_\kappa^T \mathbf{s}_\kappa^B(\mathbf{u}_h, \mathbf{u}_D, \mathbf{u}_N).$$

Next, we sum the element-based weak form over all elements κ . After rearrangement, this gives the first of two face-based weak formulations: find $\mathbf{u}_h \in \mathbb{R}^{\sum n_\kappa}$ such that

$$\sum_{\kappa \in \mathcal{T}_h} \mathbf{v}_\kappa^T H_\kappa \frac{d\mathbf{u}_\kappa}{dt} = B_h(\mathbf{u}_h, \mathbf{v}_h), \quad \forall \mathbf{v}_h \in \mathbb{R}^{\sum n_\kappa},$$

where the bilinear form B_h is defined by

$$\begin{aligned} B_h(\mathbf{u}_h, \mathbf{v}_h) := & - \sum_{\kappa \in \mathcal{T}_h} \mathbf{v}_\kappa^T M_\kappa \mathbf{u}_\kappa + \sum_{\kappa \in \mathcal{T}_h} \mathbf{v}_\kappa^T H_\kappa \mathbf{f}_\kappa \\ & - \sum_{\gamma \subset \Gamma^I} \begin{bmatrix} R_{\gamma\kappa} \mathbf{v}_\kappa \\ R_{\gamma\nu} \mathbf{v}_\nu \\ D_{\gamma\kappa} \mathbf{v}_\kappa \\ D_{\gamma\nu} \mathbf{v}_\nu \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma\kappa}^{(1)} & -\Sigma_{\gamma\kappa}^{(1)} & \Sigma_{\gamma\kappa}^{(3)} - B_\gamma & \Sigma_{\gamma\kappa}^{(3)} \\ -\Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\nu}^{(3)} & \Sigma_{\gamma\nu}^{(3)} - B_\gamma \\ \Sigma_{\gamma\kappa}^{(2)} & -\Sigma_{\gamma\kappa}^{(2)} & \Sigma_{\gamma\kappa}^{(4)} & \Sigma_{\gamma\kappa}^{(4)} \\ -\Sigma_{\gamma\nu}^{(2)} & \Sigma_{\gamma\nu}^{(2)} & \Sigma_{\gamma\nu}^{(4)} & \Sigma_{\gamma\nu}^{(4)} \end{bmatrix} \begin{bmatrix} R_{\gamma\kappa} \mathbf{u}_\kappa \\ R_{\gamma\nu} \mathbf{u}_\nu \\ D_{\gamma\kappa} \mathbf{u}_\kappa \\ D_{\gamma\nu} \mathbf{u}_\nu \end{bmatrix} \\ & - \sum_{\gamma \subset \Gamma^D} \begin{bmatrix} R_{\gamma\kappa} \mathbf{v}_\kappa \\ D_{\gamma\kappa} \mathbf{v}_\kappa \end{bmatrix}^T \begin{bmatrix} \Sigma_\gamma^D & -B_\gamma \\ -B_\gamma & 0 \end{bmatrix} \begin{bmatrix} R_{\gamma\kappa} \mathbf{u}_\kappa - \mathbf{u}_{\gamma D} \\ D_{\gamma\kappa} \mathbf{u}_\kappa \end{bmatrix} + \sum_{\gamma \subset \Gamma^N} \mathbf{v}_\kappa^T R_{\gamma\kappa}^T B_\gamma \mathbf{u}_{\gamma N}. \end{aligned} \quad (8)$$

The bilinear form (8) will be useful in the energy stability analysis presented later.

A second, equivalent face-based bilinear form is obtained by using (6) in (8). This produces

$$\begin{aligned} B_h(\mathbf{u}_h, \mathbf{v}_h) \equiv & \sum_{\kappa \in \mathcal{T}_h} \mathbf{v}_\kappa^T D_\kappa^T H_\kappa \mathbf{u}_\kappa + \sum_{\kappa \in \mathcal{T}_h} \mathbf{v}_\kappa^T H_\kappa \mathbf{f}_\kappa + \sum_{\gamma \subset \Gamma^D} \mathbf{v}_\kappa^T D_{\gamma\kappa}^T B_\gamma \mathbf{u}_{\gamma D} \\ & - \sum_{\gamma \subset \Gamma^I} \begin{bmatrix} R_{\gamma\kappa} \mathbf{v}_\kappa \\ R_{\gamma\nu} \mathbf{v}_\nu \\ D_{\gamma\kappa} \mathbf{v}_\kappa \\ D_{\gamma\nu} \mathbf{v}_\nu \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma\kappa}^{(1)} & -\Sigma_{\gamma\kappa}^{(1)} & \Sigma_{\gamma\kappa}^{(3)} - B_\gamma & \Sigma_{\gamma\kappa}^{(3)} \\ -\Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\nu}^{(3)} & \Sigma_{\gamma\nu}^{(3)} - B_\gamma \\ \Sigma_{\gamma\kappa}^{(2)} + B_\gamma & -\Sigma_{\gamma\kappa}^{(2)} & \Sigma_{\gamma\kappa}^{(4)} & \Sigma_{\gamma\kappa}^{(4)} \\ -\Sigma_{\gamma\nu}^{(2)} & \Sigma_{\gamma\nu}^{(2)} + B_\gamma & \Sigma_{\gamma\nu}^{(4)} & \Sigma_{\gamma\nu}^{(4)} \end{bmatrix} \begin{bmatrix} R_{\gamma\kappa} \mathbf{u}_\kappa \\ R_{\gamma\nu} \mathbf{u}_\nu \\ D_{\gamma\kappa} \mathbf{u}_\kappa \\ D_{\gamma\nu} \mathbf{u}_\nu \end{bmatrix} \\ & - \sum_{\gamma \subset \Gamma^D} \begin{bmatrix} R_{\gamma\kappa} \mathbf{v}_\kappa \end{bmatrix}^T \begin{bmatrix} \Sigma_\gamma^D & -B_\gamma \end{bmatrix} \begin{bmatrix} R_{\gamma\kappa} \mathbf{u}_\kappa - \mathbf{u}_{\gamma D} \\ D_{\gamma\kappa} \mathbf{u}_\kappa \end{bmatrix} + \sum_{\gamma \subset \Gamma^N} \begin{bmatrix} R_{\gamma\kappa} \mathbf{v}_\kappa \\ D_{\gamma\kappa} \mathbf{v}_\kappa \end{bmatrix}^T \begin{bmatrix} B_\gamma \mathbf{u}_{\gamma N} \\ -B_\gamma R_{\gamma\kappa} \mathbf{u}_\kappa \end{bmatrix}. \end{aligned} \quad (9)$$

The bilinear form (9) will be useful for the adjoint analysis, which we present in the next section.

III. Adjoint consistency analysis

Adjoint consistency is a desirable property that we would like our multi-dimensional SBP discretizations to satisfy. It is well known in the finite-element community that adjoint, or dual, consistency of discretizations ensures optimal error rates in the L^2 -norm of $\mathcal{O}(h^{p+1})$ ¹⁵ while inconsistent discretizations result in a suboptimal $\mathcal{O}(h^p)$ measured in L^2 . More generally, adjoint consistency leads to superconvergent (integral) functional estimates,¹⁶ which can significantly improve the accuracy of outputs like lift and drag when

using high-order methods. Given the close connection between SBP finite-difference methods and the FE methods, it is perhaps not surprising that SBP discretizations also exhibit functional superconvergence when discretized in a dual consistent manner.¹⁷

In the following section, we investigate the constraints on the SAT penalties in (3) that guarantee adjoint consistency. We begin by briefly reviewing the dual problem associated with the steady version of (2).

A. A generic adjoint PDE

An adjoint PDE is defined by the primal PDE and a particular functional of interest. For the following adjoint consistency analysis, we consider the linear functional

$$\mathcal{J}(\mathcal{U}) = \int_{\Omega} \mathcal{G}\mathcal{U} \, d\Omega + \int_{\Gamma^{\mathcal{N}}} \mathcal{V}_{\mathcal{N}}\mathcal{U} \, d\Gamma - \int_{\Gamma^{\mathcal{D}}} \mathcal{V}_{\mathcal{D}}\hat{\mathbf{n}} \cdot (\lambda\nabla\mathcal{U}) \, d\Gamma, \quad (10)$$

where $\mathcal{G} \in L^2(\Omega)$, $\mathcal{V}_{\mathcal{D}} \in L^2(\Gamma^{\mathcal{D}})$ and $\mathcal{V}_{\mathcal{N}} \in L^2(\Gamma^{\mathcal{N}})$. One can show that the adjoint PDE corresponding to the primal problem (2) and (10) is¹⁸

$$\begin{aligned} -\nabla(\lambda\nabla\mathcal{V}) &= \mathcal{G}, & \forall x \in \Omega, \\ \mathcal{V} &= \mathcal{V}_{\mathcal{D}}, & \forall x \in \Gamma^{\mathcal{D}}, \\ \hat{\mathbf{n}} \cdot (\lambda\nabla\mathcal{V}) &= \mathcal{V}_{\mathcal{N}} & \forall x \in \Gamma^{\mathcal{N}}. \end{aligned} \quad (11)$$

B. Functional and adjoint discretization

We discretize the functional (10) as

$$J_h(\mathbf{u}_h) := \sum_{\kappa \in \mathcal{T}_h} \mathbf{g}_{\kappa}^T \mathbf{M}_{\kappa} \mathbf{u}_{\kappa} + \sum_{\gamma \in \Gamma^{\mathcal{N}}} \mathbf{v}_{\gamma\mathcal{N}}^T \mathbf{B}_{\gamma} \mathbf{R}_{\gamma\kappa} \mathbf{u}_{\kappa} - \sum_{\gamma \in \Gamma^{\mathcal{D}}} \mathbf{v}_{\gamma\mathcal{D}}^T \mathbf{B}_{\gamma} \mathbf{D}_{\gamma\kappa} \mathbf{u}_{\kappa} + \sum_{\gamma \in \Gamma^{\mathcal{D}}} \mathbf{v}_{\gamma\mathcal{D}}^T \Sigma_{\gamma}^{\mathcal{D}} (\mathbf{R}_{\gamma\kappa} \mathbf{u}_{\kappa} - \mathbf{u}_{\gamma\mathcal{D}}), \quad (12)$$

where, similar to the primal problem, $\mathbf{v}_{\gamma\mathcal{N}}$ and $\mathbf{v}_{\gamma\mathcal{D}}$ denote the function value of $\mathcal{V}_{\mathcal{N}}$ and $\mathcal{V}_{\mathcal{D}}$, respectively, evaluated at the cubature nodes of the generic face γ , while \mathbf{g}_{κ} is the value of \mathcal{G} evaluated at the cubature nodes of the element κ .

The first three terms in (12) are direct discretizations of the first three terms in (10), whereas the last term is necessary for recovering adjoint consistency on the Dirichlet boundary.¹⁹ Given that the interpolation/extrapolation operators are exact for degree $r \geq p$ polynomials, namely, $\mathbf{R}_{\gamma\kappa} \mathbf{u}_{\kappa} = \mathbf{u}_{\gamma\mathcal{D}} + \mathcal{O}(h^{r+1})$, the fourth term in (12) is of order h^{r+1} .

We employ the discrete Lagrangian to find the discrete adjoint equation. Specifically, we first add the face-based weak form (9) to J_h and, after some algebraic manipulation^a, we get the Lagrangian

$$L_h(\mathbf{u}_h, \mathbf{v}_h) = J_h(\mathbf{u}_h) + B_h(\mathbf{u}_h, \mathbf{v}_h) = J_h^*(\mathbf{v}_h) + B_h^*(\mathbf{v}_h, \mathbf{u}_h),$$

where the dual form of the functional is defined by

$$J_h^*(\mathbf{v}_h) = \sum_{\kappa \in \mathcal{T}_h} \mathbf{v}_{\kappa}^T \mathbf{H}_{\kappa} \mathbf{f}_{\kappa} + \sum_{\gamma \in \Gamma^{\mathcal{N}}} \mathbf{u}_{\gamma\mathcal{N}}^T \mathbf{B}_{\gamma} \mathbf{R}_{\gamma\kappa} \mathbf{v}_{\kappa} + \sum_{\gamma \in \Gamma^{\mathcal{D}}} \mathbf{u}_{\gamma\mathcal{D}}^T \mathbf{B}_{\gamma} \mathbf{D}_{\gamma\kappa} \mathbf{v}_{\kappa} + \sum_{\gamma \in \Gamma^{\mathcal{D}}} \mathbf{u}_{\gamma\mathcal{D}}^T \Sigma_{\gamma}^{\mathcal{D}} (\mathbf{R}_{\gamma\kappa} \mathbf{v}_{\kappa} - \mathbf{v}_{\gamma\mathcal{D}}),$$

and the adjoint bilinear form is given by

$$\begin{aligned} B_h^*(\mathbf{v}_h, \mathbf{u}_h) &= \sum_{\kappa \in \mathcal{T}_h} \mathbf{u}_{\kappa}^T \mathbf{H}_{\kappa} \mathbf{D}_{\kappa} \mathbf{v}_{\kappa} + \sum_{\kappa \in \mathcal{T}_h} \mathbf{u}_{\kappa}^T \mathbf{H}_{\kappa} \mathbf{g}_{\kappa} \\ &- \sum_{\gamma \in \Gamma^{\mathcal{I}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{u}_{\kappa} \\ \mathbf{R}_{\gamma\nu} \mathbf{u}_{\nu} \\ \mathbf{D}_{\gamma\kappa} \mathbf{u}_{\kappa} \\ \mathbf{D}_{\gamma\nu} \mathbf{u}_{\nu} \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma\kappa}^{(1)} & -\Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\kappa}^{(2)} + \mathbf{B}_{\gamma} & -\Sigma_{\gamma\nu}^{(2)} \\ -\Sigma_{\gamma\kappa}^{(1)} & \Sigma_{\gamma\nu}^{(1)} & -\Sigma_{\gamma\kappa}^{(2)} & \Sigma_{\gamma\nu}^{(2)} + \mathbf{B}_{\gamma} \\ \Sigma_{\gamma\kappa}^{(3)} - \mathbf{B}_{\gamma} & \Sigma_{\gamma\nu}^{(3)} & \Sigma_{\gamma\kappa}^{(4)} & \Sigma_{\gamma\nu}^{(4)} \\ \Sigma_{\gamma\kappa}^{(3)} & \Sigma_{\gamma\nu}^{(3)} - \mathbf{B}_{\gamma} & \Sigma_{\gamma\kappa}^{(4)} & \Sigma_{\gamma\nu}^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{v}_{\kappa} \\ \mathbf{R}_{\gamma\nu} \mathbf{v}_{\nu} \\ \mathbf{D}_{\gamma\kappa} \mathbf{v}_{\kappa} \\ \mathbf{D}_{\gamma\nu} \mathbf{v}_{\nu} \end{bmatrix} \\ &- \sum_{\gamma \in \Gamma^{\mathcal{D}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{u}_{\kappa} \\ \mathbf{D}_{\gamma\kappa} \mathbf{u}_{\kappa} \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma}^{\mathcal{D}} \\ -\mathbf{B}_{\gamma} \end{bmatrix} (\mathbf{R}_{\gamma\kappa} \mathbf{v}_{\kappa} - \mathbf{v}_{\gamma\mathcal{D}}) - \sum_{\gamma \in \Gamma^{\mathcal{N}}} \mathbf{u}_{\kappa}^T \mathbf{R}_{\gamma\kappa}^T \mathbf{B}_{\gamma} (\mathbf{D}_{\gamma\kappa} \mathbf{v}_{\kappa} - \mathbf{v}_{\gamma\mathcal{N}}). \end{aligned}$$

^aIn particular, note that the functional and bilinear form are scalars, so $J_h(\mathbf{u}_h)^T = J_h(\mathbf{u}_h)$ and $B_h(\mathbf{u}_h, \mathbf{v}_h)^T = B_h(\mathbf{u}_h, \mathbf{v}_h)$.

Next, we set the first variation of $L_h(\mathbf{u}_h, \mathbf{v}_h)$ with respect \mathbf{u}_h to zero. Since the primal variable is finite dimensional here, taking the first variation is equivalent to finding the gradient of L_h with respect to \mathbf{u}_h . Furthermore, we see that $J_h^*(\mathbf{v}_h)$ does not depend on \mathbf{u}_h , so we only need to consider the gradient of $B_h^*(\mathbf{v}_h, \mathbf{u}_h)$. Taking the gradient of $B_h^*(\mathbf{v}_h, \mathbf{u}_h)$ with respect to \mathbf{u}_κ , multiplying by \mathbf{H}_κ^{-1} , and setting the result to zero (i.e. setting the first variation to zero), gives the following element-based strong form of the adjoint equation:

$$\mathbf{H}_\kappa^{-1} \frac{\partial B_h^*}{\partial \mathbf{u}_\kappa} = \mathbf{D}_\kappa \mathbf{v}_\kappa + \mathbf{g}_\kappa - \mathbf{H}_\kappa^{-1} (\mathbf{s}_\kappa^{\mathcal{I}})^*(\mathbf{v}_h) - \mathbf{H}_\kappa^{-1} (\mathbf{s}_\kappa^{\mathcal{B}})^*(\mathbf{v}_h, \mathbf{v}_\mathcal{D}, \mathbf{v}_\mathcal{N}) = \mathbf{0}, \quad (13)$$

where the adjoint SAT penalties for the interfaces are

$$(\mathbf{s}_\kappa^{\mathcal{I}})^*(\mathbf{v}_h) = \sum_{\gamma \subset \Gamma_\kappa^{\mathcal{I}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa}^T & \mathbf{D}_{\gamma\kappa}^T \end{bmatrix} \begin{bmatrix} \Sigma_{\gamma\kappa}^{(1)} & -\Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\kappa}^{(2)} + \mathbf{B}_\gamma & -\Sigma_{\gamma\nu}^{(2)} \\ \Sigma_{\gamma\kappa}^{(3)} - \mathbf{B}_\gamma & \Sigma_{\gamma\nu}^{(3)} & \Sigma_{\gamma\kappa}^{(4)} & \Sigma_{\gamma\nu}^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{v}_\kappa \\ \mathbf{R}_{\gamma\nu} \mathbf{v}_\nu \\ \mathbf{D}_{\gamma\kappa} \mathbf{v}_\kappa \\ \mathbf{D}_{\gamma\nu} \mathbf{v}_\nu \end{bmatrix}.$$

and the penalties for the boundaries are

$$(\mathbf{s}_\kappa^{\mathcal{B}})^*(\mathbf{u}_h, \mathbf{u}_\mathcal{D}, \mathbf{u}_\mathcal{N}) = \sum_{\gamma \subset \Gamma_\kappa^{\mathcal{D}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa}^T & \mathbf{D}_{\gamma\kappa}^T \end{bmatrix} \begin{bmatrix} \Sigma_\gamma^{\mathcal{D}} \\ -\mathbf{B}_\gamma \end{bmatrix} (\mathbf{R}_{\gamma\kappa} \mathbf{v}_\kappa - \mathbf{v}_{\gamma\mathcal{D}}) + \sum_{\gamma \subset \Gamma_\kappa^{\mathcal{N}}} \mathbf{R}_{\gamma\kappa}^T \mathbf{B}_\gamma (\mathbf{D}_{\gamma\kappa} \mathbf{v}_\kappa - \mathbf{v}_{\gamma\mathcal{N}})$$

C. Adjoint consistency

The spatial derivatives and the boundary SATs in (13) are consistent with the continuous dual problem (11). To see this, note that the sum $\mathbf{D}_\kappa \mathbf{v}_\kappa + \mathbf{g}_\kappa$ in (13) is an order h^{p+1} discretization of the continuous adjoint problem (11) on Ω . Indeed, the operator $\nabla \cdot (\lambda \nabla)$ is self-adjoint, so \mathbf{D}_κ is the same operator used in the primal discretization. Similarly, the boundary SAT, $(\mathbf{s}_\kappa^{\mathcal{B}})^*$, also introduces an error $O(h^{p+1})$. To see this, recall that $\mathbf{R}_{\gamma\kappa}$ and $\mathbf{D}_{\gamma\kappa}$ are exact for polynomials of degree p , and $\mathbf{v}_{\gamma\mathcal{D}}$ and $\mathbf{v}_{\gamma\mathcal{N}}$ are the exact boundary values evaluated at the nodes of γ . Thus, the differences $\mathbf{R}_{\gamma\kappa} \mathbf{v}_\kappa - \mathbf{v}_{\gamma\mathcal{D}}$ and $\mathbf{D}_{\gamma\kappa} \mathbf{v}_\kappa - \mathbf{v}_{\gamma\mathcal{N}}$ vanish for polynomial solutions of degree p or less. Only the interface SATs require further scrutiny to determine adjoint consistency.

Theorem 1. *The primal discretization (3) and functional discretization (12) are adjoint consistent of order h^{p+1} provided the exact solution \mathcal{V} is C^{p+1} continuous on Ω , and the SAT penalty matrices satisfy*

$$\begin{aligned} \Sigma_{\gamma\kappa}^{(1)} &= \Sigma_{\gamma\nu}^{(1)}, & \Sigma_{\gamma\kappa}^{(2)} + \Sigma_{\gamma\nu}^{(2)} &= -\mathbf{B}_\gamma, \\ \Sigma_{\gamma\kappa}^{(4)} &= \Sigma_{\gamma\nu}^{(4)}, & \Sigma_{\gamma\kappa}^{(3)} + \Sigma_{\gamma\nu}^{(3)} &= \mathbf{B}_\gamma. \end{aligned} \quad (14)$$

The proof of Theorem 1 is given in Ref. [20].

The conditions (14) automatically give rise to an elementwise conservative discretization. To see this, first we define a union of any collection of elements $\mathcal{S}_h \subset \mathcal{T}_h$. Let $\mathbf{v}_\kappa = \mathbf{1}$ for all $\kappa \in \mathcal{S}_h$. By the SBP operator properties we have $\mathbf{R}_{\gamma\kappa} \mathbf{v}_\kappa = \mathbf{1}$ and $\mathbf{D}_{\gamma\kappa} \mathbf{v}_\kappa = \mathbf{0}$. Then, making use of these two identities and (14), the bilinear form (9) without source terms is reduced to

$$B_h(\mathbf{u}_h, \mathbf{1}) = - \sum_{\gamma \subset \partial \mathcal{S}_h} \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma\kappa}^{(1)} & -\Sigma_{\gamma\kappa}^{(1)} & \Sigma_{\gamma\kappa}^{(3)} - \mathbf{B}_\gamma & \Sigma_{\gamma\kappa}^{(3)} \\ -\Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\nu}^{(1)} & \Sigma_{\gamma\nu}^{(3)} & \Sigma_{\gamma\nu}^{(3)} - \mathbf{B}_\gamma \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{u}_\kappa \\ \mathbf{R}_{\gamma\nu} \mathbf{u}_\nu \\ \mathbf{D}_{\gamma\kappa} \mathbf{u}_\kappa \\ \mathbf{D}_{\gamma\nu} \mathbf{u}_\nu \end{bmatrix}$$

The above equation only includes a sum over the boundary of \mathcal{S}_h , and it is independent of interface terms interior to \mathcal{S}_h . Therefore, the conservation property is satisfied in that no artificial source is introduced into the computational domain \mathcal{S}_h through interface SATs.

IV. Energy analysis

In this section we further constrain the SAT penalty matrices based on the conditions for discrete energy stability. Before presenting the conditions for energy stability, we simplify the penalty matrices based on the adjoint consistency conditions (14). First, we will drop the dependence of the $\Sigma^{(1)}$ and $\Sigma^{(4)}$ matrices on the elements:

$$\Sigma_{\gamma\kappa}^{(1)} = \Sigma_{\gamma\nu}^{(1)} \equiv \Sigma_{\gamma}^{(1)}, \quad \text{and} \quad \Sigma_{\gamma\kappa}^{(4)} = \Sigma_{\gamma\nu}^{(4)} \equiv \Sigma_{\gamma}^{(4)}.$$

Second, we will also assume that $\Sigma_{\gamma\kappa}^{(2)} = \Sigma_{\gamma\nu}^{(2)}$ and $\Sigma_{\gamma\kappa}^{(3)} = \Sigma_{\gamma\nu}^{(3)}$, although this is not strictly required by the adjoint-consistency analysis; see [20] for a more general analysis. This assumption together with the conditions in (14) gives

$$\Sigma_{\gamma\kappa}^{(2)} = \Sigma_{\gamma\nu}^{(2)} = -\frac{1}{2}\mathbf{B}_{\gamma}, \quad \text{and} \quad \Sigma_{\gamma\kappa}^{(3)} = \Sigma_{\gamma\nu}^{(3)} = \frac{1}{2}\mathbf{B}_{\gamma}.$$

In addition to simplifying the stability analysis, our motivation for this assumption is that $\Sigma^{(2)}$ and $\Sigma^{(3)}$ do not help control the coercivity of the bilinear form, that is, the positive definiteness of the system matrix.

We will need the following lemma for the stability analysis. The purpose of the lemma is to shift the volume terms in the bilinear form B_h to the faces, so that these terms can contribute to the semi-definiteness of the interface terms.

Lemma 1. *For each face γ of element κ , let a face-weight coefficient $\alpha_{\gamma\kappa} > 0$ be given such that $\sum_{\gamma \subset \Gamma_{\kappa}} \alpha_{\gamma\kappa} = 1$. Then the bilinear form corresponding to the SBP-SAT discretization of the homogeneous version of the PDE (2) can be written as*

$$\begin{aligned} B_h(\mathbf{u}_h, \mathbf{v}_h) = & - \sum_{\gamma \subset \Gamma^{\mathcal{I}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{v}_{\kappa} \\ \mathbf{R}_{\gamma\nu} \mathbf{v}_{\nu} \\ \mathbf{G}_{\kappa} \mathbf{v}_{\kappa} \\ \mathbf{G}_{\nu} \mathbf{v}_{\nu} \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma}^{(1)} & -\Sigma_{\gamma}^{(1)} & -\mathbf{C}_{\gamma\kappa} & \mathbf{C}_{\gamma\nu} \\ -\Sigma_{\gamma}^{(1)} & \Sigma_{\gamma}^{(1)} & \mathbf{C}_{\gamma\kappa} & -\mathbf{C}_{\gamma\nu} \\ -\mathbf{C}_{\gamma\kappa}^T & \mathbf{C}_{\gamma\kappa}^T & \lambda^{-1}\alpha_{\gamma\kappa}\tilde{\mathbf{H}}_{\kappa} & \\ \mathbf{C}_{\gamma\nu}^T & -\mathbf{C}_{\gamma\nu}^T & & \lambda^{-1}\alpha_{\gamma\nu}\tilde{\mathbf{H}}_{\nu} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{u}_{\kappa} \\ \mathbf{R}_{\gamma\nu} \mathbf{u}_{\nu} \\ \mathbf{G}_{\kappa} \mathbf{u}_{\kappa} \\ \mathbf{G}_{\nu} \mathbf{u}_{\nu} \end{bmatrix} \\ & - \sum_{\gamma \subset \Gamma^{\mathcal{I}}} \begin{bmatrix} \mathbf{D}_{\gamma\kappa} \mathbf{v}_{\kappa} \\ \mathbf{D}_{\gamma\nu} \mathbf{v}_{\nu} \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma}^{(4)} & \Sigma_{\gamma}^{(4)} \\ \Sigma_{\gamma}^{(4)} & \Sigma_{\gamma}^{(4)} \end{bmatrix} \begin{bmatrix} \mathbf{D}_{\gamma\kappa} \mathbf{u}_{\kappa} \\ \mathbf{D}_{\gamma\nu} \mathbf{u}_{\nu} \end{bmatrix} \\ & - \sum_{\gamma \subset \Gamma^{\mathcal{D}}} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{v}_{\kappa} \\ \mathbf{G}_{\kappa} \mathbf{v}_{\kappa} \end{bmatrix}^T \begin{bmatrix} \Sigma_{\gamma}^{\mathcal{D}} & -2\mathbf{C}_{\gamma\kappa} \\ -2\mathbf{C}_{\gamma\kappa}^T & \lambda^{-1}\alpha_{\gamma\kappa}\tilde{\mathbf{H}}_{\kappa} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\gamma\kappa} \mathbf{u}_{\kappa} \\ \mathbf{G}_{\kappa} \mathbf{u}_{\kappa} \end{bmatrix} \end{aligned} \quad (15)$$

where we have introduced the matrices

$$\begin{aligned} \mathbf{G}_{\kappa} &= \lambda \begin{bmatrix} \mathbf{D}_x \\ \mathbf{D}_y \end{bmatrix}_{\kappa}, & \mathbf{G}_{\nu} &= \lambda \begin{bmatrix} \mathbf{D}_x \\ \mathbf{D}_y \end{bmatrix}_{\nu}, \\ \mathbf{C}_{\gamma\kappa} &= \frac{1}{2}\mathbf{B}_{\gamma} \begin{bmatrix} \mathbf{N}_{x,\gamma} \mathbf{R}_{\gamma\kappa} & \mathbf{N}_{y,\gamma} \mathbf{R}_{\gamma\kappa} \end{bmatrix}, & \mathbf{C}_{\gamma\nu} &= -\frac{1}{2}\mathbf{B}_{\gamma} \begin{bmatrix} \mathbf{N}_{x,\gamma} \mathbf{R}_{\gamma\nu} & \mathbf{N}_{y,\gamma} \mathbf{R}_{\gamma\nu} \end{bmatrix}, \end{aligned}$$

and

$$\tilde{\mathbf{H}}_{\kappa} = \begin{bmatrix} \mathbf{H}_{\kappa} & \\ & \mathbf{H}_{\kappa} \end{bmatrix}, \quad \tilde{\mathbf{H}}_{\nu} = \begin{bmatrix} \mathbf{H}_{\nu} & \\ & \mathbf{H}_{\nu} \end{bmatrix}.$$

The full proof follows from straightforward algebra and is omitted; see [20].

Next, we introduce the conditions that ensure energy stability.

Theorem 2. *For each face γ of element κ , let face-weight coefficient $\alpha_{\gamma\kappa} \geq 0$ be given such that $\sum_{\gamma \subset \Gamma_{\kappa}} \alpha_{\gamma\kappa} = 1$. Then, the bilinear form corresponding to the SBP-SAT discretization of the homogeneous version of the PDE (2) is energy stable provided*

$$\Sigma_{\gamma}^{(1)} - \lambda(\alpha_{\gamma\kappa}^{-1}\mathbf{C}_{\gamma\kappa}\tilde{\mathbf{H}}_{\kappa}\mathbf{C}_{\gamma\kappa}^T + \alpha_{\gamma\nu}^{-1}\mathbf{C}_{\gamma\nu}\tilde{\mathbf{H}}_{\nu}^{-1}\mathbf{C}_{\gamma\nu}^T) \succeq 0 \quad (16)$$

$$\Sigma_{\gamma}^{\mathcal{D}} - 4\lambda\alpha_{\gamma\kappa}^{-1}\mathbf{C}_{\gamma\kappa}\tilde{\mathbf{H}}_{\kappa}^{-1}\mathbf{C}_{\gamma\kappa}^T \succeq 0 \quad (17)$$

and $\Sigma_{\gamma}^{(4)} \succeq 0$, where $\mathbf{A} \succeq 0$ indicates \mathbf{A} is positive semi-definite.

Proof. The proof is provided in the Appendix. □

V. Test cases and results

In this section several test cases are presented in order to verify the theory developed in Sections III and IV. For all test cases we employ SBP operators defined on simplex elements, specifically the SBP- Γ operators introduced in [11]. In addition, although many choices of the penalty matrices satisfy the requirements of Theorem 1 and 2, a straightforward choice that is adopted for the following experiments is

$$\begin{aligned}\Sigma_{\gamma}^{(1)} &= \lambda(\alpha_{\gamma\kappa}^{-1}\mathbf{C}_{\gamma\kappa}\tilde{\mathbf{H}}_{\kappa}\mathbf{C}_{\gamma\kappa}^T + \alpha_{\gamma\nu}^{-1}\mathbf{C}_{\gamma\nu}\tilde{\mathbf{H}}_{\nu}^{-1}\mathbf{C}_{\gamma\nu}^T), \\ \Sigma_{\gamma}^{(4)} &= 0, \\ \Sigma_{\gamma}^{\mathcal{D}} &= 4\lambda\alpha_{\gamma\kappa}^{-1}\mathbf{C}_{\gamma\kappa}\tilde{\mathbf{H}}_{\kappa}^{-1}\mathbf{C}_{\gamma\kappa}^T,\end{aligned}$$

with face-weight coefficients based on the face area:

$$\alpha_{\gamma\kappa} = \begin{cases} \frac{\mathcal{A}(\gamma)}{\mathcal{A}(\Gamma_{\kappa}^{\mathcal{I}}) + 2\mathcal{A}(\Gamma_{\kappa}^{\mathcal{D}})}, & \gamma \in \Gamma^{\mathcal{I}} \\ \frac{2\mathcal{A}(\gamma)}{\mathcal{A}(\Gamma_{\kappa}^{\mathcal{I}}) + 2\mathcal{A}(\Gamma_{\kappa}^{\mathcal{D}})}, & \gamma \in \Gamma^{\mathcal{D}} \end{cases},$$

where the function $\mathcal{A}(\gamma)$ denotes the size of face γ .

A. Accuracy study

The first test is to verify primal and adjoint consistency by examining the convergence rates of a discrete solution and an associated functional. We use a manufactured solution on the unit square $\Omega = [0, 1]^2$ with $\lambda = 10$ and the exact solution given by

$$u = e^{x+y} \sin(4\pi x) \sin(4\pi y), \quad (18)$$

to derive the source term \mathcal{F} . The functional is defined as

$$\mathcal{J} = \int_{\Omega} \mathcal{U} d\Omega, \quad (19)$$

which is a special case of (10) with $\mathcal{G} = 1$, $\mathcal{V}_{\mathcal{N}} = 0$ and $\mathcal{V}_{\mathcal{D}} = 0$. We use a sequence of uniformly refined meshes consisting of $K = 128, 512, 2048$, and 8192 triangular elements in order to estimate the asymptotic convergence rates. The coarsest mesh is shown in Figure 1a. The nominal element size is given by $h \equiv 1/\sqrt{K/2}$, which is the element edge length along the domain boundaries.

Figure 2a shows the solution error measured in terms of the L^2 -norm, which in this paper is approximated using the cubature defined by the SBP operator. We see that, under the uniform mesh refinement, the solution errors behave asymptotically like $O(h^{p+1})$, which is in agreement with design accuracy.

Figure 2b shows the error in the functional value given in (19). A convergence rate of approximately $2p$ is achieved for all the degrees, which is also in agreement with the theoretical order of convergence.^{16,17}

B. Tightness of the stability bound

In the second test we evaluate the tightness of the stability conditions. Since the stability conditions in Theorem 2 are sufficient but not necessary, a relaxation factor $\alpha \in (0, 1]$ acting on $\Sigma^{(1)}$ and $\Sigma^{\mathcal{D}}$ may still yield a stable bilinear form. We use this relaxation factor as a measure of the tightness of the bound on the penalties. Overly conservative SAT penalties will allow for a relaxation factor that is much smaller than 1, while a necessary and sufficient stability conditions would only permit $\alpha \geq 1$.

Energy stability is equivalent to a negative definite bilinear form, so we consider the effect of the relaxation factor on the largest eigenvalue of the linear system, i.e., the eigenvalue with the smallest magnitude. Once this largest eigenvalue becomes positive, the corresponding relaxation factor value is referred as the ‘‘allowable relaxation factor’’; finding this allowable relaxation factor is the objective of this experiment. Due to the expense of computing the smallest magnitude eigenvalue of large matrices, the solutions are solved on the coarse 8×8 randomly perturbed mesh shown in Figure 1b. The manufactured solution given in (18) is used again.

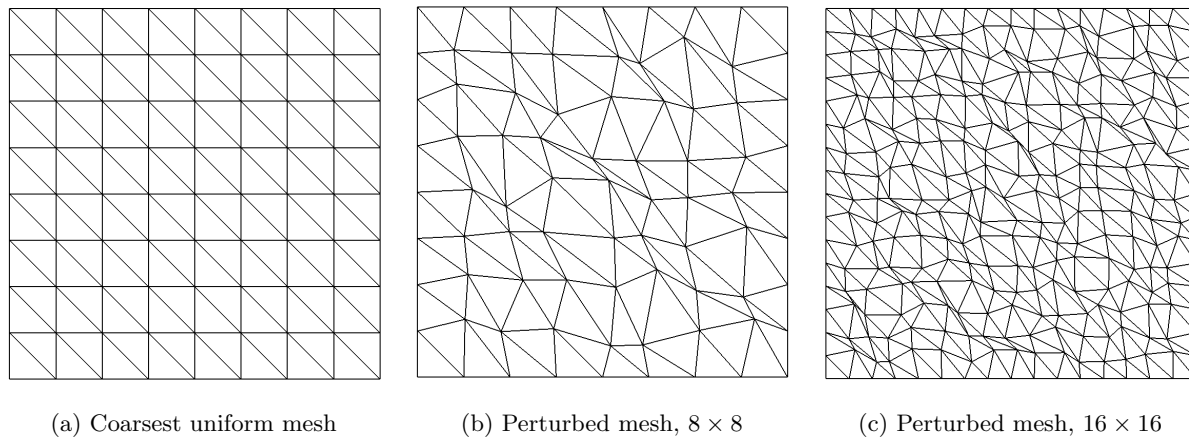


Figure 1: Different meshes used for test cases

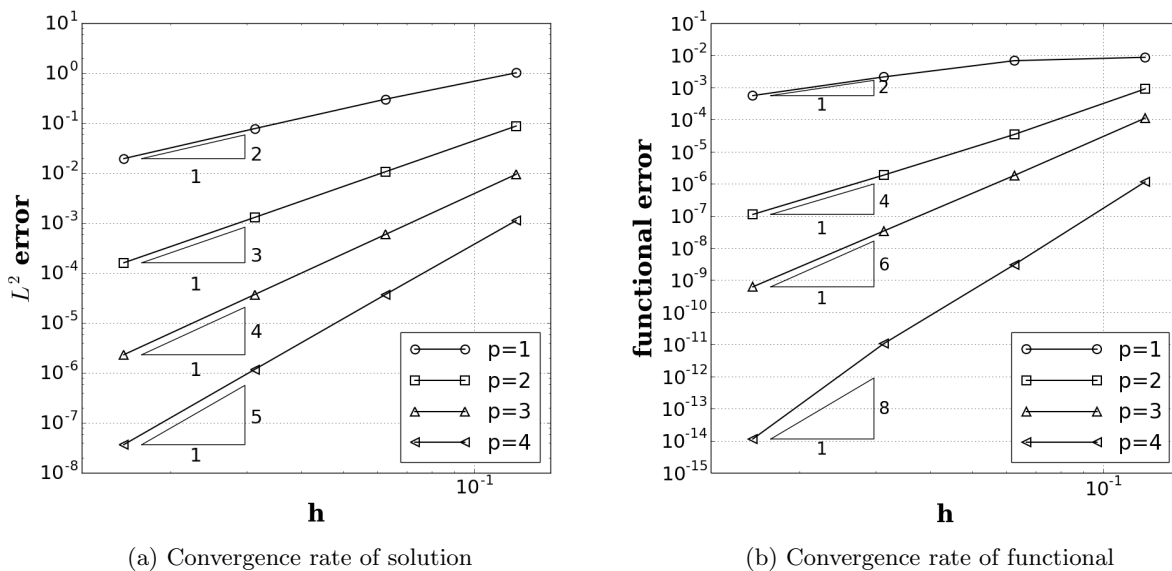


Figure 2: Convergence rate study

The eigenvalues with the smallest magnitude are plotted in Figure 3. The allowable relaxation factors are less than one for all degrees, which verifies Theorem 2. Furthermore, the smallest allowable relaxation factor is between 0.45 and 0.6, which suggests that the bound is relatively tight (i.e., the allowable relaxation factor is not $\ll 1$).

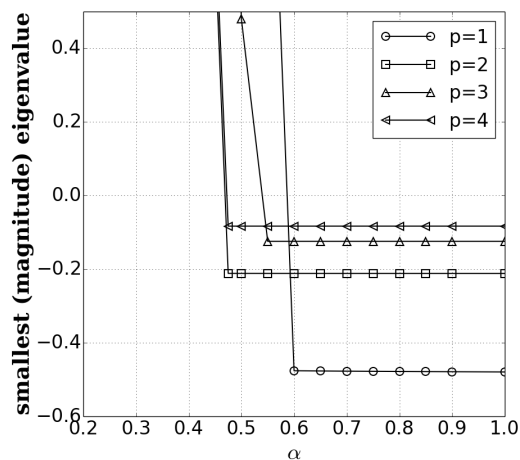


Figure 3: Relaxation effect on SATs

C. Energy stability

To complement the preceding investigation, we solve an unsteady problem with homogeneous Dirichlet boundary conditions and no source term using two different relaxation factors: $\alpha = 1$ and $\alpha = 0.45$. The choice $\alpha = 1$ produces a stable solution while $\alpha = 0.45$ produces an unstable solution (based on the results in Figure 3). The PDE solution “energy” should be monotonically decreasing as time evolves. The solution of the SBP-SAT discretization will also have a decreasing energy, provided the stability conditions of Theorem 2 are satisfied.

For this study, the time derivative is discretized using the second-order backward differentiation formula (BDF2) with a time step $\Delta t = 1.0 \times 10^{-4}$. Since the penalties are mesh-dependent, this experiment is performed on a structured triangular mesh that is randomly perturbed, as shown in Figure 1c. As can be seen, the mesh is extremely nonsmooth and almost tangled; indeed, the largest angle in the mesh is 179.90° .

Figure 4 shows the energy evolution. As can be seen, the energy for solutions of the unscaled discretizations (i.e., $\alpha = 1$) is monotonically decreasing, as expected. In contrast, all solutions based on scaled penalties ($\alpha = 0.45$) diverge after a short period of time; note the logarithmic time scale. The results further verify that conditions in Theorem 2 are not only valid but also quite tight.

VI. Conclusion

We described a general framework to facilitate the analysis of interior penalty methods arising in multidimensional SBP discretizations of second-order linear PDEs. In this framework, we considered a general form of SAT that uses dense penalty coefficient matrices on each face of the SBP elements. We then derived the conditions upon which the discretizations are simultaneously conservative, consistent, adjoint consistent and energy stable. These conditions are entirely algebraic and do not depend on exact integration.

Finally, several test cases were carried out to verify the analysis using a particular SBP operator. Specifically, the convergence rate study confirmed that the discretizations achieved design order for both solution and functional. Furthermore, the numerical results on extremely skewed mesh suggested that our stability bound is relatively tight in the sense that a scaling factor applied to one of the SATs could not be reduced below one order of magnitude without causing instability.

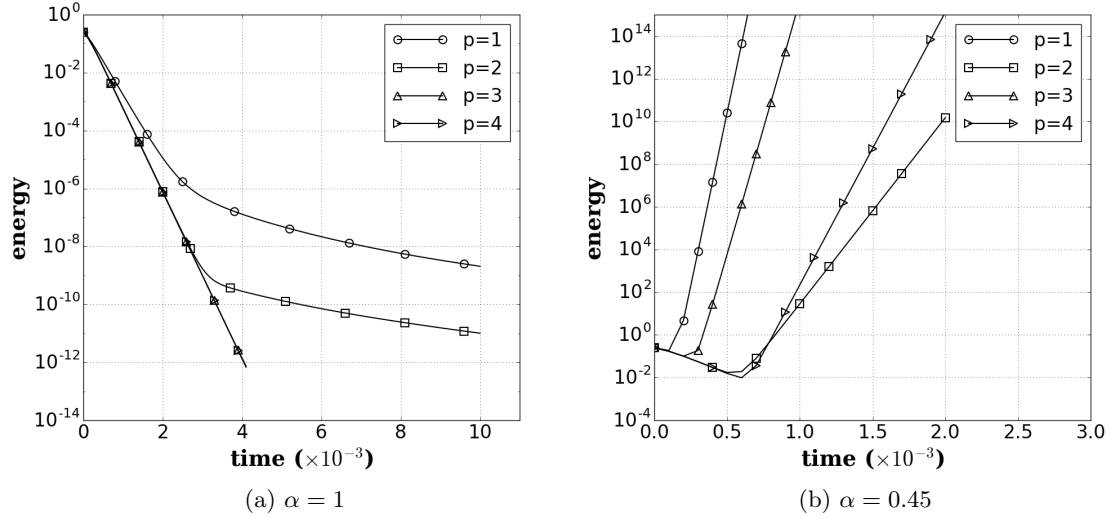


Figure 4: Energy history of homogeneous problem

Appendix: Proof of Theorem 2

In this section we give the proof of energy stability, i.e., Theorem 2. For convenience, Theorem 2 is reprinted here:

Theorem. For each face γ of element κ , let face-weight coefficient $\alpha_{\gamma\kappa} \geq 0$ be given such that $\sum_{\gamma \in \Gamma_\kappa} \alpha_{\gamma\kappa} = 1$. Then, the bilinear form corresponding to the SBP-SAT discretization of the homogeneous version of the PDE (2) is energy stable provided

$$\Sigma_\gamma^{(1)} - \lambda(\alpha_{\gamma\kappa}^{-1} C_{\gamma\kappa} \tilde{H}_\kappa C_{\gamma\kappa}^T + \alpha_{\gamma\nu}^{-1} C_{\gamma\nu} \tilde{H}_\nu^{-1} C_{\gamma\nu}^T) \succeq 0 \quad (20)$$

$$\Sigma_\gamma^{\mathcal{D}} - \lambda \alpha_{\gamma\kappa}^{-1} C_{\gamma\kappa} \tilde{H}_\kappa^{-1} C_{\gamma\kappa}^T \succeq 0 \quad (21)$$

and $\Sigma_\gamma^{(4)} \succeq 0$, where $A \succeq 0$ indicates A is positive semi-definite.

Proof. The SBP-SAT discretization of the homogeneous equation is given by

$$\sum_{\kappa \in \mathcal{T}_h} \mathbf{v}_\kappa^T \mathbf{H}_\kappa \frac{d\mathbf{w}_\kappa}{dt} = B_h(\mathbf{w}_h, \mathbf{v}_h),$$

where $B_h(\mathbf{w}_h, \mathbf{v}_h)$ is defined in (15). If $B_h(\mathbf{v}_h, \mathbf{v}_h)$ in (15) is guaranteed to be nonpositive for arbitrary \mathbf{v}_h , the discretization is energy stable. This can be realized if the symmetric matrices in the three sums of (15) are positive semi-definite. We begin by considering the matrix that appears in the sum over the Dirichlet boundary faces:

$$\begin{bmatrix} \Sigma_\gamma^{\mathcal{D}} & -C_{\gamma\kappa} \\ -C_{\gamma\kappa}^T & \lambda^{-1} \alpha_{\gamma\kappa} \tilde{H}_\kappa \end{bmatrix} \succeq 0.$$

Since, $\lambda \alpha_{\gamma\kappa}^{-1} \tilde{H}_\kappa$ is positive definite, the above matrix is positive semi-definite if the associated Schur complement is positive semi-definite:

$$\Sigma_\gamma^{\mathcal{D}} - \lambda \alpha_{\gamma\kappa}^{-1} C_{\gamma\kappa} \tilde{H}_\kappa^{-1} C_{\gamma\kappa}^T \succeq 0,$$

which is precisely the condition (21).

Next, consider the matrix involving $\Sigma_\gamma^{(4)}$ in (15):

$$\begin{bmatrix} \Sigma_\gamma^{(4)} & \Sigma_\gamma^{(4)} \\ \Sigma_\gamma^{(4)} & \Sigma_\gamma^{(4)} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes \Sigma_\gamma^{(4)},$$

where \otimes denotes the Kronecker product. Since the eigenvalues of $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ are zero and two, it follows from the spectral theory of Kronecker products that the eigenvalues of the above matrix are 2 times the eigenvalues of $\Sigma_\gamma^{(4)}$ and n_γ zeros. Thus, we require that $\Sigma_\gamma^{(4)} \succeq 0$.

Finally, we analyze the matrix containing $\Sigma_\gamma^{(1)}$. Similar to the matrix in the boundary-face sum, we make use of the fact that $\lambda^{-1}\alpha_{\gamma\kappa}\tilde{H}_\kappa$ and $\lambda^{-1}\alpha_{\gamma\nu}\tilde{H}_\nu$ are positive definite to conclude that the 4×4 block matrix is positive semi-definite if the Schur complement is also positive semi-definite, i.e.

$$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \left\{ \Sigma_\gamma^{(1)} - \lambda(\alpha_{\gamma\kappa}^{-1}C_{\gamma\kappa}\tilde{H}_\kappa C_{\gamma\kappa}^T + \alpha_{\gamma\nu}^{-1}C_{\gamma\nu}\tilde{H}_\nu^{-1}C_{\gamma\nu}^T) \right\} \succeq 0.$$

The eigenvalues of $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ are zero and two; thus, to ensure that the above Kronecker product is positive semi-definite, we must require that

$$\Sigma_\gamma^{(1)} - \lambda(\alpha_{\gamma\kappa}^{-1}C_{\gamma\kappa}\tilde{H}_\kappa C_{\gamma\kappa}^T + \alpha_{\gamma\nu}^{-1}C_{\gamma\nu}\tilde{H}_\nu^{-1}C_{\gamma\nu}^T) \succeq 0,$$

which is condition (20). □

Acknowledgements

J. Hicken was partially funded by the Air Force Office of Scientific Research Award FA9550-15-1-0242, and J. Crean was supported by the National Science Foundation under Grant No 1554253. The authors gratefully acknowledge this support. We also thank RPI's Scientific Computation Research Center for the use of computer facilities.

References

- ¹Kirby, R. M. and Karniadakis, G. E., "De-aliasing on non-uniform grids: algorithms and applications," *Journal of Computational Physics*, Vol. 191, No. 1, 2003, pp. 249–264.
- ²Kirby, R. M. and Sherwin, S. J., "Aliasing errors due to quadratic nonlinearities on triangular spectral /hp element discretisations," *Journal of Engineering Mathematics*, Vol. 56, No. 3, 2006, pp. 273–288.
- ³Fisher, T. C. and Carpenter, M. H., "High-order entropy stable finite difference schemes for nonlinear conservation laws: Finite domains," *Journal of Computational Physics*, Vol. 252, No. 1, 2013, pp. 518–557.
- ⁴Carpenter, M. H., Fisher, T. C., Nielsen, E. J., and Frankel, S. H., "Entropy Stable Spectral Collocation Schemes for the Navier–Stokes Equations: Discontinuous Interfaces," *SIAM Journal on Scientific Computing*, Vol. 36, No. 5, 2014, pp. B835B867.
- ⁵Crean, J., Hicken, J. E., Del Rey Fernández, D. C., Zingg, D. W., and Carpenter, M. H., "Entropy-Stable Summation-By-Parts Discretization of the Euler Equations on General Curved Elements," submitted to *Journal of Scientific Computing*, 2017.
- ⁶Kreiss, H.-O. and Scherer, G., "Finite element and finite difference methods for hyperbolic partial differential equations." *Mathematical aspects of finite elements in partial differential equations*, Academic Press, New York/London, 1974, pp. 195–212.
- ⁷Hicken, J. E. and Zingg, D. W., "A parallel Newton-Krylov solver for the Euler equations discretized using simultaneous approximation terms," *AIAA Journal*, Vol. 46, No. 11, Nov. 2008, pp. 2773–2786.
- ⁸Nordström, J., Gong, J., van der Weide, E., and Svärd, M., "A stable and conservative high order multi-block method for the compressible Navier-Stokes equations," *Journal of Computational Physics*, Vol. 228, No. 24, 2009, pp. 9020–9035.
- ⁹Hicken, J. E., Del Rey Fernández, D. C., and Zingg, D. W., "Multi-dimensional Summation-By-Parts Operators: General Theory and Application to Simplex Elements," *SIAM Journal on Scientific Computing*, Vol. 38, No. 4, 2016, pp. A1935–A1958.
- ¹⁰Arnold, D. N., Brezzi, F., Cockburn, B., and Marini, L. D., "Unified analysis of discontinuous Galerkin methods for elliptic problems," *SIAM journal on numerical analysis*, Vol. 39, No. 5, 2002, pp. 1749–1779.
- ¹¹Hicken, J. E., Del Rey Fernández, D. C., and Zingg, D. W., "Simultaneous approximation terms for multi-dimensional summation-by-parts operators," *46th AIAA Fluid Dynamics Conference*, Washington, DC, June 2016, AIAA–2016–3971.
- ¹²Del Rey Fernández, D. C., Hicken, J. E., and Zingg, D. W., "Simultaneous approximation terms for multi-dimensional summation-by-parts operators," submitted to *Journal of Scientific Computing*, 2016, in revision.
- ¹³Gong, J. and Nordström, J., "Interface procedures for finite difference approximations of the advection-diffusion equation," *Journal of Computational and Applied Mathematics*, Vol. 236, No. 5, 2011, pp. 602–620.
- ¹⁴Carpenter, M. H., Nordström, J., and Gottlieb, D., "Revisiting and extending interface penalties for multi-domain summation-by-parts operators," *Journal of Scientific Computing*, Vol. 45, No. 1, June 2010, pp. 118–150.
- ¹⁵Hartmann, R. and Houston, P., "Symmetric interior penalty DG methods for the compressible Navier-Stokes equations I: Method formulation," *International Journal of Numerical Analysis & Modeling*, Vol. 3, No. 1, 2005, pp. 1–20.
- ¹⁶Hartmann, R. and Houston, P., "An optimal order interior penalty discontinuous Galerkin discretization of the compressible Navier–Stokes equations," *Journal of Computational Physics*, Vol. 227, No. 22, 2008, pp. 9670 – 9685.
- ¹⁷Hicken, J. E. and Zingg, D. W., "Superconvergent functional estimates from summation-by-parts finite-difference discretizations," *SIAM Journal on Scientific Computing*, Vol. 33, No. 2, 2011, pp. 893–922.

¹⁸Lanczos, C., *Linear Differential Operators*, D. Van Nostrand Company, Limited, London, England, 1961.

¹⁹Hartmann, R., “Adjoint Consistency Analysis of Discontinuous Galerkin Discretizations,” *SIAM Journal on Numerical Analysis*, Vol. 45, No. 6, 2007, pp. 2671–2696.

²⁰Yan, J., Crean, J., and Hicken, J. E., “Interior Penalties for Summation-by-Parts Discretizations of Linear Second-Order Differential Equations,” *Journal of Scientific Computing*, 2016, submitted.