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

Jianfeng Yan · Jared Crean · Jason E. Hicken

Abstract This work focuses on multidimensional summation-by-parts (SBP) discretizations of linear elliptic operators with variable coefficients. We consider a general SBP discretization with dense simultaneous approximation terms (SATs), which serve as interior penalties to enforce boundary conditions and inter-element coupling in a weak sense. Through the analysis of adjoint consistency and stability, we present several conditions on the SAT penalties. Based on these conditions, we generalize the modified scheme of Bassi and Rebay (BR2) and the symmetric interior penalty Galerkin (SIPG) method to SBP-SAT discretizations. Numerical experiments are carried out on unstructured grids with triangular elements to verify the theoretical results.

1 Introduction

Summation-by-parts (SBP) operators \([1, 2]\) are high-order finite-difference operators that can be used to construct time-stable discretizations. The stability and high-order accuracy of SBP discretizations makes them attractive for simulating conservation laws over long periods of time. Diagonal-norm SBP operators are especially appealing for compressible fluid flow simulations, because they can be used to construct discretizations that are entropy stable without relying on exact integration \([3, 4]\).

A drawback of classical SBP operators is that, like all one-dimensional operators, they require multiblock tensor-product grids in order to be applied to complex geometries; generating high-quality, multiblock, hexahedral grids is difficult to automate and time consuming when done manually. Motivated by this...
drawback, Hicken, Del Rey Fernández, and Zingg [5] generalized the SBP definition to arbitrary bounded domains. They showed that a degree \( p \) diagonal-norm SBP operator can be constructed on a given domain provided a degree \( 2p - 1 \) cubature exists whose nodes produce a full-rank Vandermonde matrix.

Ultimately, we are interested in using diagonal-norm, multidimensional SBP operators to construct high-order, entropy-stable discretizations of the Navier-Stokes equations on unstructured grids. We believe this approach will combine the stability and accuracy advantages of classical SBP operators with the flexibility of tetrahedral grids; however, there are several developments necessary to bring multidimensional SBP methods to the level of maturity of classical SBP methods. The development considered in this work is the discretization of linear elliptic operators with variable coefficients, which are a prototype for the viscous terms in the Navier-Stokes equations.

In this paper we do not address the construction of specialized second-derivative SBP operators [6–8]. Instead, we adopt the so-called “first-derivative twice” approach. For classical finite-difference methods, applying the first-derivative twice approximately doubles the stencil size and is typically less accurate [7]; however, the multidimensional SBP operators that we intend to use are dense on each element, so the stencil size is not changed by applying the first-derivative operator twice. That said, we believe that more general multidimensional SBP operators, particularly those that are sparse on each element, would benefit from a generalization of the works in [7] and [8].

Rather than SBP operators themselves, the focus of this paper is the analysis of penalty terms to enforce boundary conditions and inter-element coupling for multidimensional SBP discretizations of elliptic and parabolic partial differential equations (PDEs). In the SBP literature, such penalty terms are called simultaneous approximation terms (SATs) and were popularized by [9]. SATs for multidimensional SBP discretizations of the linear advection equation were recently studied in [10], and SATs for tensor-product SBP discretizations of second-order PDEs have been investigated by a number of authors; see, e.g., [11] and the review [12].

SATs are analogous to interior penalties (IPs) used in the finite-element (FE) community; see the review [13] and the references therein. Indeed, there are strong connections between FE and SBP discretizations, and the present work draws heavily from the discontinuous Galerkin (DG) literature.

Despite the extensive prior work in this area, the present work makes the following contributions.

1. We present the requirements on dense SAT coefficient matrices to obtain multidimensional SBP-SAT discretizations that are simultaneously consistent, conservative, adjoint consistent, and stable.
2. We generalize the modified scheme of Bassi and Rebay (BR2) [14] and the symmetric interior penalty Galerkin (SIPG) method [13, 15, 16] to multidimensional SBP discretizations. In addition, we show how, in the SBP-SAT context, SIPG can be derived from BR2 using matrix analysis.
3. We show that a particular SBP-SAT implementation of the BR2 penalty has a computational cost of \( O(p) \) in 2D and \( O(p^2) \) in 3D, in contrast with \( O(p^2) \) and \( O(p^3) \), respectively, for DG implementations with Lagrange basis functions.
The above results do not rely on exact integration, so they are valid for fully discretized elliptic problems and semi-discretized parabolic problems.

The remaining sections are organized as follows. After introducing notation and reviewing SBP operators, Section 2 presents the model parabolic PDE and its SBP-SAT discretization. Section 3 investigates the adjoint consistency of the discretization and delineates the necessary adjoint-consistency conditions on the SAT penalties. The penalties are further constrained by the energy-stability analysis in Section 4. The resulting conditions are used to generalize the BR2 and SIPG methods to multidimensional SBP discretizations in Section 5. Verification studies are provided in Section 6, and a summary is provided in Section 7.

2 Multi-dimensional SBP discretization of parabolic PDEs

2.1 Notation

Functions are denoted with capital letters in calligraphic font; for example $U \in L^2(\Omega)$ is a square-integrable function on the domain $\Omega$. A function evaluated on a node set is denoted by a lowercase letter in bold font. For example, the function $U$ evaluated at the nodes of $S = \{(x_i, y_i)\}^{n_{\kappa}}_{i=1}$ is given by

$$u = [U(x_1, y_1) \ U(x_2, y_2) \ldots U(x_n, y_n)]^T.$$

The space of polynomials of total degree $p$ in $x$ and $y$ on $\Omega$ is denoted by $\mathbb{P}_p(\Omega)$. 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 $P \in \mathbb{P}_p(\Omega)$ we would have

$$p \equiv [P(x_1, y_1) \ P(x_2, y_2) \ldots P(x_n, y_n)]^T.$$

Matrices are represented with an uppercase sans-serif type, for example $A \in \mathbb{R}^{n \times m}$.

2.2 SBP definition and face operators

We adopt the definition of multidimensional SBP operators proposed in [5]. To keep the presentation self-contained, the definition for an operator approximating $\partial / \partial x$ on a two dimensional domain is provided below. The definition for the SBP operator approximating $\partial / \partial y$ is analogous.

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

1. For all $P \in \mathbb{P}_p(\Omega_\kappa)$, the vector $D_x p$ is equal to $\partial P / \partial x$ at the nodes $S_\kappa$;
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
\[
p^T E_x q = \int_{\Gamma} P Q n_x d\Gamma, \quad \forall P, Q \in \mathbb{P}_r(\Omega),
\]
where \( r \geq p \), and \( n_x \) is the \( x \) component of \( n = [n_x, n_y]^T \), the outward pointing unit normal on \( \partial\Omega \).

The subsequent analysis is restricted to so-called diagonal-norm SBP operators, that is, SBP operators for which \( H \) is a diagonal matrix with positive entries. In this case, it was shown in [5] that the nodes \( S_\kappa \) and diagonal entries of \( H \) define a cubature rule that is exact for polynomials of total degree \( 2p - 1 \).

In order to define SATs for multidimensional SBP operators, we follow References [10, 17] and introduce interpolation/extrapolation operators from the SBP element nodes to cubature nodes on the faces of the elements. For example, consider an element \( \Omega_\kappa \) with a piecewise smooth boundary \( \partial\Omega_\kappa \), and let \( \gamma \subset \partial\Omega_\kappa \) denote one of its faces. Let \( S_\gamma = \{(x_j, y_j)\}_{j=1}^{n_\gamma} \subset \gamma \) be a set of cubature nodes with corresponding positive weights \( \{b_j\}_{j=1}^{n_\gamma} \) that is exact for polynomials of degree \( 2r \), where \( r \geq p \). The matrix \( R_{\gamma\kappa} \in \mathbb{R}^{n_\gamma \times n_\gamma} \) is a degree \( r \) interpolation/extrapolation operator from the SBP nodes \( S_\kappa \) to the face nodes \( S_\gamma \) if, for all \( P \in \mathbb{P}_r(\Omega_\kappa) \),
\[
(R_{\gamma\kappa} p)_j = \sum_{i=1}^{n_\gamma} (R_{\gamma\kappa})_{ji} P(x_i, y_i) = P(x_j, y_j), \quad \forall j = 1, 2, \ldots, n_\gamma.
\]

For a given (strong) cubature rule of degree \( 2p - 1 \) defined on \( \Omega_\kappa \), it was shown in [10] that there exists at least one SBP operator whose corresponding matrix \( E_x \) has the decomposition
\[
E_x = \sum_{\gamma \subset \partial\Omega_\kappa} R_{\gamma\kappa}^T N_{x,\gamma} B_\gamma R_{\gamma\kappa}, \quad (1)
\]
where \( B_\gamma = \text{diag}(b_1, b_2, \ldots, b_{n_\gamma}) \) is an \( n_\gamma \times n_\gamma \) diagonal matrix holding the cubature weights for \( \gamma \) along its diagonal, and \( N_{x,\gamma} = \text{diag}(n_{x,1}, n_{x,2}, \ldots, n_{x,n_\gamma}) \) is an \( n_\gamma \times n_\gamma \) diagonal matrix holding the \( x \) component of the outward unit normal with respect to \( \Omega_\kappa \) at the cubature points of \( \gamma \). We will assume in the following analysis that the SBP operators are such that \( E_x \) has the decomposition (1), and that the operators in the \( y \) direction have analogous decompositions.

2.3 The model PDE

We consider the following linear parabolic PDE — or the corresponding steady Poisson PDE — defined on the compact domain \( \Omega \subset \mathbb{R}^2 \):
\[
\frac{\partial U}{\partial t} - \nabla \cdot (\Lambda \nabla U) = F, \quad \forall (x, y) \in \Omega,
\]
where \( F \in L^2(\Omega \times [0, T]) \) is a given source term and
\[
\Lambda = \begin{bmatrix}
\lambda_{xx} & \lambda_{xy} \\
\lambda_{yx} & \lambda_{yy}
\end{bmatrix}
\]
is a symmetric, positive-definite tensor. The parabolic PDE is provided with the initial condition
\[ U(0, x, y) = U_0(x, y), \quad \forall (x, y) \in \Omega, \quad (3) \]
where \( U_0 \in L^2(\Omega) \). Finally, the PDE is supplied with the Dirichlet and Neumann boundary conditions,
\[ U(t, x, y) = U_D(t, x, y), \quad \forall (x, y) \in \Gamma^D, \]
\[ \hat{n} \cdot (\Lambda \nabla U(t, x, y)) = U_N(t, x, y), \quad \forall (x, y) \in \Gamma^N, \quad (4) \]
where \( U_D \in L^2(\Gamma^D \times [0, T]) \) and \( U_N \in L^2(\Gamma^N \times [0, T]) \). The vector \( \hat{n} = [n_x, n_y]^T \) is the outward pointing unit normal on the boundary \( \partial \Omega \). We assume that the Dirichlet boundary is nonempty, \( \Gamma^D \neq \emptyset \), so that the solution is unique. Furthermore, \( \Gamma = \Gamma^D \cup \Gamma^N \) and \( \Gamma \setminus \Gamma^D = \Gamma^N \).

2.4 Strong-form Discretization

Let \( T_h = \bigcup_{\kappa=1}^K \Omega_\kappa \) denote a partition of the domain \( \Omega \) into \( K \) SBP elements, where \( \Omega_\kappa \) denotes the domain of the \( \kappa \)th element. The discrete solution on element \( \Omega_\kappa \) will be represented by the vector \( u_\kappa \in \mathbb{R}^{n_\kappa} \) whose entries are the discrete solution at the SBP nodes \( S_\kappa \). The global discrete solution, denoted \( u_h \in \mathbb{R}^{\sum n_\kappa} \), is the concatenation of all elementwise solutions.

A consistent SBP-SAT semi-discretization of (2) on element \( \kappa \) is given by
\[ \frac{d u_\kappa}{dt} = D_\kappa u_\kappa + f_\kappa - H^{-1}_\kappa s_I^\kappa(u_h) - H^{-1}_\kappa s_B^\kappa(u_h, u_D, u_N), \quad (5) \]
where \( f_\kappa \) is \( F \) evaluated at the nodes of element \( \Omega_\kappa \), and
\[ D_\kappa = \begin{bmatrix} D_x & D_y \\ \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{bmatrix} \]
\[ \left( \begin{array}{c} D_x \\ D_y \end{array} \right) \]
\[ = \begin{bmatrix} [\Lambda_{xx} & \Lambda_{xy}] \\ [\Lambda_{yx} & \Lambda_{yy}] \end{bmatrix} \begin{pmatrix} D_x \\ D_y \end{pmatrix} \quad (6) \]
is the SBP approximation of \( \nabla \cdot (\Lambda \nabla) \) on element \( \Omega_\kappa \), with \( D_x \in \mathbb{R}^{n_\kappa \times n_\kappa} \) and \( D_y \in \mathbb{R}^{n_\kappa \times n_\kappa} \) the first-derivative SBP operators in the \( x \) and \( y \) directions, respectively. The Cartesian elements of the tensor \( \Lambda \) are evaluated at the SBP nodes and stored in the diagonal matrices \( \Lambda_{xx}, \Lambda_{xy}, \Lambda_{yx} \text{ and } \Lambda_{yy} \). For example,
\[ \Lambda_{xx} = \text{diag} (\lambda_{xx}(x_1, y_1), \lambda_{xx}(x_2, y_2), \ldots, \lambda_{xx}(x_n, y_n)). \]

The vectors \( s_I^\kappa \) and \( s_B^\kappa \) on the right-hand side of (5) are the interface and boundary SAT penalties, respectively. For element \( \kappa \) these penalties are defined by
\[ s_I^\kappa(u_h) = \sum_{\gamma \subset \Gamma^I_\kappa} \begin{bmatrix} R_{\gamma \kappa}^T D_{\gamma \kappa}^T \\ \sum_{\gamma \kappa}^{(1)} \\ \sum_{\gamma \kappa}^{(2)} \\ \sum_{\gamma \kappa}^{(3)} \\ \sum_{\gamma \kappa}^{(4)} \end{bmatrix} \begin{bmatrix} \Sigma_{\gamma \kappa}^{(1)} \\ \Sigma_{\gamma \kappa}^{(2)} \\ \Sigma_{\gamma \kappa}^{(3)} \\ \Sigma_{\gamma \kappa}^{(4)} \\ \Sigma_{\gamma \kappa}^{(5)} \end{bmatrix} \begin{bmatrix} R_{\gamma \kappa} u_\kappa - R_{\gamma \nu} u_\nu \\ D_{\gamma \kappa} u_\kappa + D_{\gamma \nu} u_\nu \end{bmatrix} \]
\[ s_B^\kappa(u_h, u_D, u_N) \]
and

\[
\begin{align*}
\mathbf{s}_\gamma^D(\mathbf{u}_h, \mathbf{u}_D, \mathbf{u}_N) &= \sum_{\gamma \in I_\gamma^D} \left[ R_{\gamma\gamma}^{T} D_{\gamma\gamma}^{T} \right] \left[ \Sigma_{\gamma_{\gamma}}^D \right] \left( R_{\gamma\gamma} \mathbf{u}_\kappa - \mathbf{u}_{\gamma\gamma} \right) \\
&\quad + \sum_{\gamma \in I_N} R_{\gamma\gamma}^{T} \mathbf{B}_\gamma \left( D_{\gamma\gamma} \mathbf{u}_\kappa - \mathbf{u}_{\gamma\gamma} \right),
\end{align*}
\]

respectively. The set \( I_\kappa = \partial \Omega_\kappa \) is the boundary of element \( \Omega_\kappa \), while \( I_\kappa^D = I_\kappa \cap I_\kappa^D \) and \( I_\kappa^N = I_\kappa \cap I_\kappa^N \). We use \( \nu \) as the generic index of the element sharing face \( \gamma \) with the \( \kappa \)th element, i.e., \( \gamma = \Omega_\kappa \cap \Omega_{\nu} \). The vectors \( \mathbf{u}_{\gamma\gamma} \) and \( \mathbf{u}_{\gamma\gamma} \) in the boundary penalties denote the functions \( \mathcal{U}_D \) and \( \mathcal{U}_N \), respectively, evaluated at the cubature nodes of face \( \gamma \).

Recall that the matrix \( \mathbf{R}_{\gamma\kappa} \in \mathbb{R}^{n_x \times n_x} \) is the interpolation/extrapolation operator from the nodes of \( \Omega_\kappa \) to the nodes of face \( \gamma \subset I_\kappa \), while \( \mathbf{R}_{\gamma\nu} \in \mathbb{R}^{n_x \times n_x} \) is the interpolation/extrapolation operator from the nodes of the neighbor to the nodes of \( \gamma \). Therefore, the normal derivative operators that discretize \( \mathbf{n} \cdot (\mathbf{A} \mathbf{U}) \) on face \( \gamma \) are given by

\[
\begin{align*}
\mathbf{D}_{\gamma\kappa} &= \mathbf{N}_x, \mathbf{r}_{\gamma, \kappa} \left( \Lambda_{xx} \mathbf{D}_x + \Lambda_{xy} \mathbf{D}_y \right)_\kappa + \mathbf{N}_y, \mathbf{r}_{\gamma, \kappa} \left( \Lambda_{yx} \mathbf{D}_x + \Lambda_{yy} \mathbf{D}_y \right)_\kappa, \\
\mathbf{D}_{\gamma\nu} &= -\mathbf{N}_x, \mathbf{r}_{\gamma, \nu} \left( \Lambda_{xx} \mathbf{D}_x + \Lambda_{xy} \mathbf{D}_y \right)_\nu + \mathbf{N}_y, \mathbf{r}_{\gamma, \nu} \left( \Lambda_{yx} \mathbf{D}_x + \Lambda_{yy} \mathbf{D}_y \right)_\nu.
\end{align*}
\]

In addition, recall that \( \mathbf{N}_x, \gamma \) (resp. \( \mathbf{N}_y, \gamma \)) is a diagonal matrix holding the \( x \) (resp. \( y \)) component of the unit outward normal, with respect to \( \kappa \), at the cubature nodes of face \( \gamma \). Thus, the sign of this matrix must be reversed for \( \mathbf{D}_{\gamma\nu} \).

Finally, the matrices \( \Sigma_{\gamma_{\gamma}}^{(i)} = \left( \Sigma_{\gamma_{\gamma}}^{(i)} \right)^T \in \mathbb{R}^{n_x \times n_x}, i = 1, 2, 3, 4 \) denote the symmetric SAT coefficient matrices for element \( \kappa \) on face \( \gamma \). Similarly, \( \Sigma_{\gamma}^{D} \) is the coefficient matrix for the SAT on a Dirichlet boundary face of \( \kappa \). These coefficients are to be determined in the following analysis. Note that \( \Sigma_{\gamma_{\gamma}}^{(i)} \neq \Sigma_{\gamma_{\gamma}}^{(i)} \) in general; that is, we do not assume \textit{ab initio} that the coefficient matrices of two adjacent elements are necessarily equal.

2.5 Face-based weak forms of the discretization

The discretization (5) 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.

**Proposition 1** Let \( \mathbf{D}_\kappa \) be defined as in (6). Then, \( \forall \mathbf{u}_\kappa, \mathbf{v}_\kappa \in \mathbb{R}^{n_x} \),

\[
\begin{align*}
\mathbf{v}_\kappa^T \mathbf{H}_\kappa \mathbf{D}_\kappa \mathbf{u}_\kappa &= -\mathbf{v}_\kappa^T \mathbf{M}_\kappa \mathbf{u}_\kappa + \sum_{\gamma \subset I_\kappa} \mathbf{v}_\kappa^T \mathbf{R}_{\gamma\kappa}^T \mathbf{B}_\gamma \mathbf{D}_{\gamma\kappa} \mathbf{u}_\kappa, \quad (7) \\
\end{align*}
\]

and

\[
\begin{align*}
-\mathbf{v}_\kappa^T \mathbf{M}_\kappa \mathbf{u}_\kappa &= \mathbf{v}_\kappa^T \mathbf{D}_\kappa^T \mathbf{H}_\kappa \mathbf{u}_\kappa - \sum_{\gamma \subset I_\kappa} \mathbf{v}_\kappa^T \mathbf{D}_\kappa^T \mathbf{B}_\gamma \mathbf{R}_{\gamma\kappa} \mathbf{u}_\kappa, \quad (8)
\end{align*}
\]

where \( \mathbf{M}_\kappa \) is the symmetric semi-definite matrix

\[
\mathbf{M}_\kappa = \begin{bmatrix}
\mathbf{D}_x^T & \begin{bmatrix} \mathbf{H}_{\Lambda xx} & \mathbf{H}_{\Lambda xy} \\
\mathbf{H}_{\Lambda ex} & \mathbf{H}_{\Lambda ey} \end{bmatrix} \mathbf{D}_x \\
\mathbf{D}_y^T & \begin{bmatrix} \mathbf{H}_{\Lambda yx} & \mathbf{H}_{\Lambda yy} \end{bmatrix} \mathbf{D}_y
\end{bmatrix}.
\]
The proof of Proposition 1 is a straightforward application of the properties of SBP operators and is omitted.

**Remark 1** The identities in Proposition 1 are the SBP analogs of applying integration by parts to \( \int_{\Omega}\nabla \cdot (A\nabla u) \, d\Omega \) once (the first identity) and twice (the second identity).

To obtain the element-based weak formulation, we first left multiply (5) by \( v_\kappa^T H_\kappa \), where \( v_\kappa \in \mathbb{R}^{n_\kappa} \) is an arbitrary vector, and then apply (7). This produces the following form of the discretization: for all \( \kappa = 1, 2, \ldots, K \), find \( u_\kappa \in \mathbb{R}^{n_\kappa} \) such that, \( \forall v_\kappa \in \mathbb{R}^{n_\kappa} \),

\[
v_\kappa^T H_\kappa \frac{du_\kappa}{dt} = -v_\kappa^T M_\kappa u_\kappa + \sum_{\gamma \in \Gamma_\kappa} v_\kappa^T R_{\gamma \kappa}^T B_{\gamma} u_\kappa + v_\kappa^T H_\kappa f_\kappa - v_\kappa^T s_\kappa^T (u_h) - v_\kappa^T s_\kappa^T (u_h, u_D, u_N).
\]

To obtain the first of two face-based weak formulations, we sum the element-based weak form over all \( \kappa \). After rearrangement, this gives the statement: find \( u_h \in \mathbb{R}^{\sum n_\kappa} \) such that

\[
\sum_{\kappa \in T_h} v_\kappa^T H_\kappa \frac{du_\kappa}{dt} = B_h(u_h, v_h), \quad \forall v_h \in \mathbb{R}^{\sum n_\kappa},
\]

where the bilinear form on the right is defined by

\[
B_h(u_h, v_h) := -\sum_{\kappa \in T_h} v_\kappa^T M_\kappa u_\kappa + \sum_{\kappa \in T_h} v_\kappa^T H_\kappa f_\kappa - \sum_{\gamma \in \Gamma^T} \begin{bmatrix} R_{\gamma \kappa} v_\kappa \\ D_{\gamma \kappa} v_\kappa \end{bmatrix}^T \begin{bmatrix} \Sigma^{(1)} & -\Sigma^{(1)} & \Sigma^{(3)} & -B_{\gamma} & \Sigma^{(3)} \\ -\Sigma^{(1)} & \Sigma^{(1)} & \Sigma^{(3)} & \Sigma^{(3)} & -B_{\gamma} \\ \Sigma^{(2)} & -\Sigma^{(2)} & \Sigma^{(4)} & \Sigma^{(4)} & -B_{\gamma} \\ -\Sigma^{(2)} & \Sigma^{(2)} & \Sigma^{(4)} & \Sigma^{(4)} & -B_{\gamma} \end{bmatrix} \begin{bmatrix} R_{\gamma \kappa} u_\kappa \\ R_{\gamma \kappa} u_{D} \\ D_{\gamma \kappa} u_\kappa \\ D_{\gamma \kappa} u_{D} \end{bmatrix} - \sum_{\gamma \in \Gamma^F} \begin{bmatrix} R_{\gamma \kappa} v_\kappa \\ D_{\gamma \kappa} v_\kappa \end{bmatrix}^T \begin{bmatrix} \Sigma^{(1)} & -B_{\gamma} & 0 \\ -B_{\gamma} & 0 \end{bmatrix} \begin{bmatrix} R_{\gamma \kappa} u_\kappa - u_{\gamma D} \\ D_{\gamma \kappa} u_\kappa \end{bmatrix} + \sum_{\gamma \in \Gamma^N} v_\kappa^T R_{\gamma \kappa}^T B_{\gamma} u_{\gamma N}. \tag{9}
\]

The bilinear form (9) will be our starting point in the energy stability analysis presented later.

An equivalent face-based bilinear form, which will be useful for the adjoint analysis, is obtained by using (8) in (9). This produces

\[
B_h(u_h, v_h) \equiv \sum_{\kappa \in T_h} v_\kappa^T D_{\gamma \kappa}^T H_\kappa u_\kappa + \sum_{\kappa \in T_h} v_\kappa^T H_\kappa f_\kappa + \sum_{\gamma \in \Gamma^T} \begin{bmatrix} R_{\gamma \kappa} v_\kappa \\ D_{\gamma \kappa} v_\kappa \end{bmatrix}^T \begin{bmatrix} \Sigma^{(1)} & -\Sigma^{(1)} & \Sigma^{(3)} & -B_{\gamma} & \Sigma^{(3)} \\ -\Sigma^{(1)} & \Sigma^{(1)} & \Sigma^{(3)} & \Sigma^{(3)} & -B_{\gamma} \\ \Sigma^{(2)} & -\Sigma^{(2)} & \Sigma^{(4)} & \Sigma^{(4)} & -B_{\gamma} \\ -\Sigma^{(2)} & \Sigma^{(2)} & \Sigma^{(4)} & \Sigma^{(4)} & -B_{\gamma} \end{bmatrix} \begin{bmatrix} R_{\gamma \kappa} u_\kappa \\ R_{\gamma \kappa} u_{D} \\ D_{\gamma \kappa} u_\kappa \\ D_{\gamma \kappa} u_{D} \end{bmatrix} - \sum_{\gamma \in \Gamma^F} \begin{bmatrix} R_{\gamma \kappa} v_\kappa \\ D_{\gamma \kappa} v_\kappa \end{bmatrix}^T \begin{bmatrix} \Sigma^{(1)} & -B_{\gamma} & 0 \\ -B_{\gamma} & 0 \end{bmatrix} \begin{bmatrix} R_{\gamma \kappa} u_\kappa - u_{\gamma D} \\ D_{\gamma \kappa} u_\kappa \end{bmatrix} + \sum_{\gamma \in \Gamma^N} v_\kappa^T R_{\gamma \kappa}^T B_{\gamma} u_{\gamma N}. \tag{10}
\]
3 Adjoint consistency analysis

It is well known in the finite-element community that adjoint, or dual, consistency is necessary for obtaining optimal error rates in the $L^2$ norm [13]. More generally, adjoint consistency leads to superconvergent (integral) functional estimates [18–24], 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 classical (i.e. tensor-product) SBP discretizations also exhibit superconvergent functionals when discretized in a dual consistent manner [25,26].

For the reasons listed above, adjoint consistency is a property that we would like our multi-dimensional SBP discretizations to satisfy. Therefore, in the following sections, we investigate the constraints on the SAT penalties in (5) that guarantee adjoint consistency. We begin by briefly reviewing the dual problem associated with steady version of (2).

3.1 A generic adjoint PDE

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

$$J(u) = \int_{\Omega} G u d\Omega + \int_{\Gamma^N} V_N u d\Gamma - \int_{\Gamma^D} V_D \hat{n} \cdot (A\nabla u) d\Gamma,$$

(11)

where $G \in L^2(\Omega)$, $V_D \in L^2(\Gamma^D)$ and $V_N \in L^2(\Gamma^N)$. One can show that the adjoint PDE corresponding to the steady form of (2) and (11) is [27]

$$-\nabla \cdot (A\nabla V) = G, \quad \forall \, x \in \Omega$$

$$V = V_D, \quad \forall \, x \in \Gamma^D$$

$$\hat{n} \cdot (A\nabla V) = V_N, \quad \forall \, x \in \Gamma^N.$$  (12)

3.2 Functional and adjoint discretization

We discretize the functional (11) as

$$J_h(u_h) := \sum_{\kappa \in T_h} g^T_h H_\kappa u_\kappa + \sum_{\gamma \in \Gamma^N} v^T_{\gamma N} B_\gamma R_{\gamma \kappa} u_\kappa - \sum_{\gamma \in \Gamma^D} v^T_{\gamma D} B_\gamma D_{\gamma \kappa} u_\kappa$$

$$+ \sum_{\gamma \in \Gamma^D} v^T_{\gamma D} \Sigma^D_\gamma (R_{\gamma \kappa} u_\kappa - u_{\gamma D}),$$

(13)

where $v_{\gamma N}$ and $v_{\gamma D}$ denote $V_N$ and $V_D$, respectively, evaluated at the cubature nodes of the generic face $\gamma$, and

$$g^T_h = [G(x_0) \ G(x_1) \ldots \ G(x_n)].$$  (14)

Remark 2 The first three terms in (13) are direct discretizations of the first three terms in (11). The fourth term in (13) is an order $h^{r+1}$ term; the interpolation/extrapolation operators are exact for degree $r \geq p$ polynomials, so $R_{\gamma \kappa} u_\kappa = u_{\gamma D} + O(h^{r+1})$. This last term in $J_h$ is necessary for adjoint consistency.
The discrete adjoint equation is defined implicitly based on \( J_h \) and the discretization (5). Specifically, to find the adjoint equation we form the discrete Lagrangian, take its first variation with respect to \( u_h \), and set the result to zero. To this end, we add the face-based weak form (10) to \( J_h \) and, after some algebraic manipulation\(^1\), we get the Lagrangian

\[
L_h(u_h, v_h) = J_h(u_h) + B_h(u_h, v_h) = J^*_h(v_h) + B^*_h(v_h, u_h),
\]

where the dual form of the functional is defined by

\[
J^*_h(v_h) = \sum_{\kappa \in T_h} v_h^T H_\kappa f_\kappa + \sum_{\gamma \subset F^N} u_{\gamma h}^T B_\gamma R_{\gamma h} v_\kappa + \sum_{\gamma \subset F^D} u_{\gamma h}^T B_\gamma D_{\gamma h} v_\kappa + \sum_{\gamma \subset F^D} u_{\gamma h}^T \Sigma^D_\gamma (R_{\gamma h} v_\kappa - v_{\gamma D}),
\]

and the adjoint bilinear form is given by

\[
B^*_h(v_h, u_h) = \sum_{\kappa \in T_h} u_{\kappa h}^T H_\kappa D_{\kappa h} v_\kappa + \sum_{\Omega_e \in T_h} u_{\kappa h}^T H_\kappa g_\kappa
- \sum_{\gamma \subset F^D} \left[ \begin{array}{c} R_{\gamma h} u_\kappa \\ D_{\gamma h} u_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ R_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ \end{array} \right]^T \left[ \begin{array}{cccc} \Sigma^{(1)}_\gamma & -\Sigma^{(1)}_\gamma & \Sigma^{(2)}_\gamma & -\Sigma^{(2)}_\gamma \\ -\Sigma^{(2)}_\gamma & \Sigma^{(2)}_\gamma & -\Sigma^{(2)}_\gamma & -\Sigma^{(2)}_\gamma \\ \Sigma^{(3)}_\gamma & -\Sigma^{(3)}_\gamma & \Sigma^{(3)}_\gamma & -\Sigma^{(3)}_\gamma \\ & & & \\ \Sigma^{(4)}_\gamma & -\Sigma^{(4)}_\gamma & \Sigma^{(4)}_\gamma & -\Sigma^{(4)}_\gamma \\ & & & \\ \end{array} \right] \left[ \begin{array}{c} R_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ R_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ \end{array} \right]
- \sum_{\gamma \subset F^D} \left[ \begin{array}{c} \Sigma^D_\gamma \\ -B_\gamma \\ \end{array} \right] (R_{\gamma h} v_\kappa - v_{\gamma D}) - \sum_{\gamma \subset F^N} u_{\kappa h}^T R_{\kappa h}^T B_\gamma (D_{\gamma h} v_\kappa - v_{\gamma N}).
\]

As already mentioned, the discrete adjoint equation is found by setting the first variation of \( L_h(u_h, v_h) \) with respect to \( u_h \) 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 \( u_h \). Furthermore, we see that \( J^*_h(v_h) \) does not depend on \( u_h \), so we only need to consider the gradient of \( B^*_h(v_h, u_h) \).

Taking the gradient of \( B^*_h(v_h, u_h) \) with respect to \( u_h \), multiplying by \( H^{-1}_h \), 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:

\[
H^{-1}_h \frac{\partial B^*_h}{\partial u_h} = \partial_x g_h - H^{-1}_h (s^*_h)^T (v_h) - H^{-1}_h (s^*_h)^T (v_h, v_D, v_N) = 0,
\]

where the adjoint SAT penalties for the interfaces are

\[
(s^*_h)^T (v_h) = \sum_{\gamma \subset F^D} [ R_{\gamma h}^T D_{\gamma h}^T \left[ \begin{array}{cc} \Sigma^{(1)}_\gamma & -\Sigma^{(1)}_\gamma \\ \Sigma^{(2)}_\gamma & -\Sigma^{(2)}_\gamma \\ \Sigma^{(3)}_\gamma & -\Sigma^{(3)}_\gamma \\ \Sigma^{(4)}_\gamma & -\Sigma^{(4)}_\gamma \\ \end{array} \right] \left[ \begin{array}{c} R_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ D_{\gamma h} v_\kappa \\ \end{array} \right] - \sum_{\gamma \subset F^N} u_{\kappa h}^T R_{\kappa h}^T B_\gamma (D_{\gamma h} v_\kappa - v_{\gamma N}).
\]

\(^1\) In particular, note that the functional and bilinear form are scalars, so \( J_h(u_h)^T = J_h(u_h) \) and \( B_h(u_h, v_h)^T = B_h(u_h, v_h) \).
and the penalties for the boundaries are
\[
(s^B_\gamma)^* (u_h, u_D, u_N) = \sum_{\gamma \subset \Gamma_p^*} \left[ R_{\gamma \kappa}^T D_{\gamma \kappa}^T \right] \begin{bmatrix} \Sigma_{\gamma \kappa}^D \\ \Sigma_{\gamma \kappa}^p \end{bmatrix} (R_{\gamma \kappa} v_\kappa - v_{\gamma D})
+ \sum_{\gamma \subset \Gamma_N^*} R_{\gamma \kappa}^T B_{\gamma} (D_{\gamma \kappa} v_\kappa - v_{\gamma N}).
\]

### 3.3 Adjoint consistency of the interface SAT

The sum \(D_\kappa v_\kappa + g_\kappa\) in (15) is an order \(h^{p+1}\) discretization of the adjoint PDE in (12). Indeed, \(D_\kappa\) is the same operator used in the primal discretization. Furthermore, the boundary SAT, \((s^B_\gamma)^*\), introduces an error that is also \(O(h^{p+1})\). To see this, recall that \(R_{\gamma \kappa}\) and \(D_{\gamma \kappa}\) are exact for polynomials of degree \(p\), and \(v_{\gamma D}\) and \(v_{\gamma N}\) are the exact boundary values evaluated at the nodes of \(\gamma\). Thus, the differences \(R_{\gamma \kappa} v_\kappa - v_{\gamma D}\) and \(D_{\gamma \kappa} v_\kappa - v_{\gamma N}\) 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 (5) and functional discretization (13) are adjoint consistent of order \(h^{p+1}\) provided the exact adjoint \(V\) is sufficiently smooth on \(\Omega\) and the SAT penalty matrices satisfy
\[
\Sigma_{\gamma \kappa}^{(1)} = \Sigma_{\gamma \nu}^{(1)}, \quad \Sigma_{\gamma \kappa}^{(2)} + \Sigma_{\gamma \nu}^{(2)} = -B_{\gamma},
\Sigma_{\gamma \kappa}^{(3)} + \Sigma_{\gamma \nu}^{(3)} = B_{\gamma}, \quad \Sigma_{\gamma \kappa}^{(4)} = \Sigma_{\gamma \nu}^{(4)}.
\]

**Proof** We have already considered the discretization of the spatial derivatives and the boundary SATs. To show that the interface SAT is order \(h^{p+1}\), it is sufficient to show that \((s^B_\kappa)^* (v_h) = 0\) for polynomial solutions \(V \in P_p(\Omega_h)\). For these polynomials, the interpolation/extrapolation and normal-derivative operators are exact and we have
\[
R_{\gamma \kappa} v_\kappa = R_{\gamma \nu} v_\nu \equiv v_{\gamma}, \quad \text{and} \quad D_{\gamma \kappa} v_\kappa = -D_{\gamma \nu} v_\nu \equiv v'_{\gamma}.
\]
Substituting these identities into the adjoint interface SATs (16) gives
\[
(s^B_\kappa)^* (v_h) = \sum_{\gamma \subset \Gamma_h^*} \left[ R_{\gamma \kappa}^T D_{\gamma \kappa}^T \right] \begin{bmatrix} \Sigma_{\gamma \kappa}^{(1)} - \Sigma_{\gamma \nu}^{(1)} \\ \Sigma_{\gamma \kappa}^{(3)} + \Sigma_{\gamma \nu}^{(3)} - B_{\gamma} \\ \Sigma_{\gamma \kappa}^{(2)} + \Sigma_{\gamma \nu}^{(2)} + B_{\gamma} \\ \Sigma_{\gamma \kappa}^{(4)} - \Sigma_{\gamma \nu}^{(4)} \end{bmatrix} \begin{bmatrix} v_{\gamma} \\ v'_{\gamma} \end{bmatrix}.
\]

Thus, the adjoint interface penalties vanishes under the conditions (17). \(\Box\)

**Remark 3** It is straightforward to show that the conditions (17) also imply that the SBP-SAT discretization is locally conservative, in the sense that \(\sum_{\kappa \in T_h'} \int_{T_h'} \partial_t u_h \, dt\) depends only on the boundary faces of \(T_h'\) when \(f_h = 0\), for any subset of elements \(T_h' \subset T_h\).
4 Energy analysis

4.1 Energy analysis of the PDE

The difference between two solutions of (2), \( W = U - V \), satisfies the homogeneous PDE with \( F = 0 \), \( U_D = 0 \), \( U_N = 0 \), and \( U_0 = 0 \). It follows that \( W \) satisfies

\[
\frac{1}{2} \frac{d}{dt} \int_{\Omega} W^2 \, d\Omega = - \int_{\Omega} (\nabla W) \cdot A (\nabla W) \, d\Omega \leq 0,
\]

and, since \( W = 0 \) initially, \( W = 0 \) for all time. This well-known proof shows that solutions to (2) are unique. We would like to mimic this proof in the case of the discretization (5). The following section investigates the conditions on the SAT penalties that make this possible.

4.2 Energy analysis of the discrete homogeneous problem

The objective of this section is to further constrain the SAT penalty matrices based on the conditions for discrete energy stability. Before presenting the conditions for energy stability, we first simplify the penalty matrices based on the adjoint consistency conditions (17). In particular, we will drop the dependence of the \( \Sigma^{(1)} \) and \( \Sigma^{(4)} \) matrices on the elements:

\[
\Sigma^{(1)}_{\gamma\kappa} = \Sigma^{(1)}_{\gamma\nu} \equiv \Sigma^{(1)}_{\gamma}, \quad \text{and} \quad \Sigma^{(4)}_{\gamma\kappa} = \Sigma^{(4)}_{\gamma\nu} \equiv \Sigma^{(4)}_{\gamma}.
\]

In addition, we will also assume that

\[
\Sigma^{(3)}_{\gamma\kappa} - \Sigma^{(2)}_{\gamma\kappa} = B_{\gamma}.
\]  

This is not strictly required by the adjoint-consistency analysis, but by the desire to make the 4x4 block matrix in (9) symmetric, which significantly simplifies the energy analysis; note that the above condition together with the conditions of Theorem 1 imply that \( \Sigma^{(3)}_{\gamma\kappa} = -\Sigma^{(2)}_{\gamma\nu} \), \( \Sigma^{(3)}_{\gamma\nu} = -\Sigma^{(2)}_{\gamma\kappa} \), and \( \Sigma^{(3)}_{\gamma\nu} - \Sigma^{(2)}_{\gamma\nu} = B_{\gamma} \), which, in turn, imply the symmetry of the 4x4 block matrix in (9).

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 \( \gamma \) of element \( \kappa \), let \( \alpha_{\gamma\kappa} \geq 0 \) 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) can be written as

\[
B_h(u_h, v_h) = - \sum_{\gamma \in \Gamma} \begin{bmatrix} R_{\gamma \kappa} v_\kappa \\ R_{\gamma \nu} v_\nu \end{bmatrix}^T \begin{bmatrix} \Sigma_\gamma^{(1)} & -\Sigma_\gamma^{(1)} \\ -\Sigma_\gamma^{(1)} & \Sigma_\gamma^{(1)} \end{bmatrix} \begin{bmatrix} \Sigma_\gamma^{(2)} C_{\gamma \kappa} & -\Sigma_\gamma^{(2)} C_{\gamma \nu} \\ -\Sigma_\gamma^{(2)} C_{\gamma \kappa} & \Sigma_\gamma^{(2)} C_{\gamma \nu} \end{bmatrix} \begin{bmatrix} R_{\gamma \kappa} u_\kappa \\ R_{\gamma \nu} u_\nu \end{bmatrix} \\
- \sum_{\gamma \in \Gamma} \begin{bmatrix} D_{\gamma \kappa} v_\kappa \\ D_{\gamma \nu} v_\nu \end{bmatrix}^T \begin{bmatrix} \Sigma_\gamma^{(4)} \Sigma_\gamma^{(4)} \\ \Sigma_\gamma^{(4)} \Sigma_\gamma^{(4)} \end{bmatrix} \begin{bmatrix} D_{\gamma \kappa} u_\kappa \\ D_{\gamma \nu} u_\nu \end{bmatrix} \\
- \sum_{\gamma \in \Gamma} \begin{bmatrix} F_{\kappa} v_\kappa \\ F_{\nu} v_\nu \end{bmatrix}^T \begin{bmatrix} \Sigma_\gamma^D & -B_{\gamma} C_{\gamma \kappa} \\ -B_{\gamma} C_{\gamma \nu} & \alpha_{\gamma \kappa} A_{\gamma \kappa}^* \end{bmatrix} \begin{bmatrix} F_{\kappa} u_\kappa \\ F_{\nu} u_\nu \end{bmatrix},
\]

where we have introduced the matrices

\[
F_\kappa = \begin{bmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{bmatrix}, \quad C_{\gamma \kappa} = \begin{bmatrix} N_{x,\gamma} R_{\gamma \kappa} & N_{y,\gamma} R_{\gamma \kappa} \end{bmatrix}, \\
F_\nu = \begin{bmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{bmatrix}, \quad C_{\gamma \nu} = -\begin{bmatrix} N_{x,\gamma} R_{\gamma \nu} & N_{y,\gamma} R_{\gamma \nu} \end{bmatrix},
\]

and

\[
A_{\gamma \kappa}^* = \begin{bmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{bmatrix}^{-1}, \quad A_{\gamma \nu}^* = \begin{bmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{bmatrix}^{-1}.
\]

**Proof** The full proof follows from straightforward algebra and is omitted; however, we will highlight two observations that make the connection between (19) and (9) clearer. First, we note that

\[
C_{\gamma \kappa} F_\kappa = D_{\gamma \kappa} \quad \text{and} \quad C_{\gamma \nu} F_\nu = D_{\gamma \nu}.
\]

Second, the elemental matrix \( M_{\kappa} \) can be decomposed as

\[
M_{\kappa} = \begin{bmatrix} D_{x,\kappa}^T & H \Lambda_{xx} & H \Lambda_{xy} \\ D_{y,\kappa}^T & H \Lambda_{yx} & H \Lambda_{yy} \end{bmatrix} \begin{bmatrix} D_{x,\kappa} \\ D_{y,\kappa} \end{bmatrix} = \sum_{\gamma \in \Gamma} \alpha_{\gamma \kappa} F^T_{\kappa} A_{\gamma \kappa}^* F_{\kappa}.
\]

\[\square\]

We will now state and prove the main energy-stability result.

**Theorem** The SBP-SAT discretization corresponding to the homogeneous version of (2) has a non-increasing solution norm, with respect to the \( H \) matrix, provided

\[
\Sigma_\gamma^{(1)} - \Sigma_\gamma^{(2)} C_{\gamma \kappa} (\alpha_{\gamma \kappa} A_{\gamma \kappa}^*)^{-1} C_{\gamma \kappa}^T \Sigma_\gamma^{(2)} C_{\gamma \nu} (\alpha_{\gamma \nu} A_{\gamma \nu}^*)^{-1} C_{\gamma \nu}^T \Sigma_\gamma^{(2)} \preceq 0, \quad \Sigma_\gamma^{(5)} - B_{\gamma} C_{\gamma \kappa} (\alpha_{\gamma \kappa} A_{\gamma \kappa}^*)^{-1} C_{\gamma \kappa}^T B_{\gamma} \preceq 0,
\]

and \( \Sigma_\gamma^{(5)} \succeq 0 \), where \( A \succeq 0 \) indicates that \( A \) is positive semi-definite.
Proof. The SBP-SAT discretization of the homogeneous equation is given by

\[ \sum_{\kappa \in T_h} v_{\kappa}^T H_{\kappa} \frac{dw_{\kappa}}{dt} = B_h(w_h, v_h), \]

where \( B_h(w_h, v_h) \) is defined in (19). If we can show that \( B_h(w_h, w_h) \leq 0 \) for all \( w_h \), then we will have \( \sum_{\kappa \in T_h} w_{\kappa}^T H_{\kappa} \frac{dw_{\kappa}}{dt} \leq 0 \) and the desired result will follow.

The scalar \( B_h(w_h, w_h) \) is nonpositive if the symmetric matrices in the three sums of (19) are positive semi-definite. We begin by considering the matrix that appears in the sum over the faces of the Dirichlet boundary:

\[
\begin{bmatrix}
\Sigma^D_{\gamma} & -B_{\gamma} C_{\gamma \kappa} \\
-C_{\gamma \kappa}^T B_{\gamma} & \alpha_{\gamma \kappa}^* A_{\kappa}^*
\end{bmatrix} \succeq 0.
\]

Since, \( \alpha_{\gamma \kappa} A_{\kappa}^* \) is positive definite, the above matrix is positive semi-definite if the associated Schur complement is positive semi-definite:

\[
\Sigma^D_{\gamma} - B_{\gamma} C_{\gamma \kappa} (\alpha_{\gamma \kappa} A_{\kappa}^*)^{-1} C_{\gamma \kappa}^T B_{\gamma} \succeq 0.
\]

which is precisely the condition (21).

Next, consider the matrix involving \( \Sigma_{\gamma}^{(4)} \) in (19):

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

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 twice the eigenvalues of \( \Sigma_{\gamma}^{(4)} \) and \( n_{\gamma} \) 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 \( \alpha_{\gamma \kappa} A_{\kappa}^* \) and \( \alpha_{\gamma \nu} A_{\nu}^* \) are positive definite to conclude that the \( 4 \times 4 \) block matrix is positive semi-definite if the Schur complement is also positive semi-definite, i.e.

\[
\begin{bmatrix}
1 & -1 \\
-1 & 1
\end{bmatrix} \otimes \begin{bmatrix}
\Sigma_{\gamma}^{(1)} & -\Sigma_{\gamma}^{(2)} C_{\gamma \kappa} C_{\gamma \nu} \Sigma_{\gamma}^{(2)} \\
-\Sigma_{\gamma}^{(2)} C_{\gamma \kappa} C_{\gamma \nu} & \Sigma_{\gamma}^{(2)}
\end{bmatrix} \begin{bmatrix}
(\alpha_{\gamma \kappa} A_{\kappa}^*)^{-1} \\
(\alpha_{\gamma \nu} A_{\nu}^*)^{-1}
\end{bmatrix} \begin{bmatrix}
C_{\gamma \kappa}^T \Sigma_{\gamma}^{(2)} & C_{\gamma \nu}^T \Sigma_{\gamma}^{(2)}
\end{bmatrix} \succeq 0.
\]

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

\[
\Sigma_{\gamma}^{(1)} - \Sigma_{\gamma}^{(2)} C_{\gamma \kappa} (\alpha_{\gamma \kappa} A_{\kappa}^*)^{-1} C_{\gamma \kappa}^T \Sigma_{\gamma}^{(2)} - \Sigma_{\gamma}^{(2)} C_{\gamma \nu} (\alpha_{\gamma \nu} A_{\nu}^*)^{-1} C_{\gamma \nu}^T \Sigma_{\gamma}^{(2)} \succeq 0,
\]

which is condition (20).  \qed
5 Generalization of existing methods

In Sections 3 and 4 we obtained sufficient conditions that ensure adjoint consistency and energy stability. In this section we show that these conditions can be used to recover two popular interior penalty methods used in FE methods, namely, the modified scheme of Bassi and Rebay (BR2) [14] and the symmetric interior penalty method (SIPG) [23,28].

While the stability conditions of Theorem 2 depend on $\Sigma^{(2)}_{\gamma\kappa}$ and $\Sigma^{(2)}_{\gamma\nu}$, there remains considerable flexibility in the values adopted for these matrices, provided they satisfy (18) and the conditions in Theorem 1. Additionally, although a positive semi-definite $\Sigma^{(4)}_{\gamma\kappa}$ may influence the accuracy and continuity of solutions, it is not necessary nor is it sufficient to guarantee coercivity of the bilinear form. Accordingly, a simple and effective choice for the penalty matrices is

$$
\Sigma^{(3)}_{\gamma\kappa} = -\Sigma^{(2)}_{\gamma\kappa} = \frac{1}{2} B_{\gamma},
$$

$$
\Sigma^{(4)}_{\gamma\kappa} = \Sigma^{(4)}_{\gamma\nu} = 0,
$$

which are the values used for the remainder of the paper. Note that other choices are possible that lead to asymmetric or one-sided schemes, such as the local discontinuous Galerkin scheme [29], but these are not considered in this paper.

We now investigate two specific expressions for $\Sigma^{(1)}_{\gamma}$ and $\Sigma^{D}_{\gamma}$ and show how these are related to BR2 and SIPG.

5.1 The modified scheme of Bassi and Rebay (BR2)

Based on the stability analysis in Section 4, specifically Theorem 2, a straightforward choice for the SAT penalties is

$$
\Sigma^{(1)}_{\gamma} = \frac{1}{4} B_{\gamma} \left[ C_{\gamma\kappa} (\alpha_{\gamma\kappa} \Lambda_{\kappa})^{-1} C^{T}_{\gamma\kappa} + C_{\gamma\nu} (\alpha_{\gamma\nu} \Lambda_{\nu})^{-1} C^{T}_{\gamma\nu} \right] B_{\gamma},
$$

$$
\Sigma^{D}_{\gamma} = B_{\gamma} C_{\gamma\kappa} (\alpha_{\gamma\kappa} \Lambda_{\kappa})^{-1} C^{T}_{\gamma\kappa} B_{\gamma},
$$

We now show that the above penalty matrices generalize the modified scheme of Bassi and Rebay [14] to multidimensional SBP discretizations. For ease of exposition, we will consider the scalar constant-coefficient diffusion case, that is

$$
\Lambda = \begin{bmatrix} \lambda_{xx} & \lambda_{xy} \\ \lambda_{yx} & \lambda_{yy} \end{bmatrix} = \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
$$

A similar analysis of problems with a spatially varying tensor $\Lambda$ gives the same conclusion. In addition, we will only focus on the interface penalty of BR2, since the relationship to $\Sigma^{D}_{\gamma}$ is similar.

The penalties in the BR2 method that correspond with the matrix $\Sigma^{(1)}_{\gamma}$ are of the form

$$
C_{BR2} \int_{\partial \Omega_{\kappa} \cap \Gamma} V_{\gamma} \left[ n_{x} \left( \mathcal{L}_{x,\kappa}^{+} + \mathcal{L}_{x,\nu}^{+} \right) + n_{y} \left( \mathcal{L}_{y,\kappa}^{+} + \mathcal{L}_{y,\nu}^{+} \right) \right] d\Gamma,
$$

where $C_{BR2}$ is a positive constant, $U_{\kappa}$ and $U_{\nu}$ denote the finite-dimensional solution on the elements $\Omega_{\kappa}$ and $\Omega_{\nu}$, respectively, and $V_{\kappa}$ denotes the test function on $\Omega_{\kappa}$.
defined by the variational statements

\[ \int_{\Omega} \frac{1}{2} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega = \int_{\gamma} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Gamma, \quad \forall \mathcal{V}_k \in \mathcal{P}_p(\Omega_\gamma), \]

and

\[ \int_{\Omega} \frac{1}{2} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega = \int_{\gamma} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Gamma, \quad \forall \mathcal{V}_k \in \mathcal{P}_p(\Omega_\gamma). \]

The multidimensional SBP discretization of the two penalty expressions is

\[ v_k^T H_k \int_{\Omega} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega = \frac{1}{2} V_k \lambda (U_k - U_\nu), \quad \forall v_k \in \mathbb{R}^{n_k}, \]

and

\[ v_k^T H_k \int_{\Omega} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega = \frac{1}{2} V_k \lambda (U_k - U_\nu), \quad \forall v_k \in \mathbb{R}^{n_k}, \]

where \( H_k \in \mathbb{R}^{n_k \times n_k} \) and \( H_k \in \mathbb{R}^{n_k \times n_k} \) are the discrete lifting operators. Choosing \( v \) appropriately, we obtain the explicit expressions

\[ I_{x,\kappa} = \frac{1}{2} \nabla_k H_k^{-1} \int_{\Omega} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega = \frac{1}{2} \nabla_k H_k^{-1} \int_{\Omega} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega, \]

and

\[ I_{y,\kappa} = \frac{1}{2} \nabla_k H_k^{-1} \int_{\Omega} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega = \frac{1}{2} \nabla_k H_k^{-1} \int_{\Omega} \nabla_k \mathcal{L}_{\gamma,\kappa} \, d\Omega. \]

Next, we turn to the SBP discretization of the BR2 penalty (24). Using the above expressions for \( I_{x,\kappa} \) and \( I_{y,\kappa} \), and the analogous ones for \( I_{x,\nu} \) and \( I_{y,\nu} \), we obtain the discretization

\[ \frac{\mathcal{C}_{\text{BR2}}}{2} v_k^T R_{\gamma,\kappa} H_k \left[ N_{\gamma,\kappa} \left( R_{\gamma,\kappa} I_{x,\kappa} + R_{\gamma,\nu} I_{x,\nu} \right) + N_{\gamma,\nu} \left( R_{\gamma,\kappa} I_{y,\kappa} + R_{\gamma,\nu} I_{y,\nu} \right) \right] \]

\[ = \frac{\mathcal{C}_{\text{BR2}}}{4} v_k^T R_{\gamma,\kappa} H_k \left[ \left( N_{\gamma,\kappa} R_{\gamma,\kappa} R_{\gamma,\kappa}^{-1} R_{\gamma,\kappa} N_{\gamma,\kappa} + N_{\gamma,\nu} R_{\gamma,\kappa} R_{\gamma,\kappa}^{-1} R_{\gamma,\nu} N_{\gamma,\kappa} N_{\gamma,\nu} \right) \right] \]

\[ + \left( N_{\gamma,\kappa} R_{\gamma,\nu} R_{\gamma,\nu}^{-1} R_{\gamma,\nu} N_{\gamma,\kappa} + N_{\gamma,\nu} R_{\gamma,\nu} R_{\gamma,\nu}^{-1} R_{\gamma,\nu} N_{\gamma,\nu} \right) \right] \]

\[ = v_k^T R_{\gamma,\kappa} \Sigma_{\text{BR2}} \left( R_{\gamma,\kappa} N_{\gamma,\kappa} - R_{\gamma,\nu} N_{\gamma,\nu} \right), \]

where

\[ \Sigma_{\text{BR2}} = \frac{\mathcal{C}_{\text{BR2}}}{4} B_{\gamma} \left[ N_{\gamma,\kappa} R_{\gamma,\kappa} R_{\gamma,\kappa}^{-1} R_{\gamma,\kappa} N_{\gamma,\kappa} + N_{\gamma,\nu} R_{\gamma,\kappa} R_{\gamma,\kappa}^{-1} R_{\gamma,\nu} N_{\gamma,\nu} \right] \]

\[ + \left( N_{\gamma,\kappa} R_{\gamma,\nu} R_{\gamma,\nu}^{-1} R_{\gamma,\nu} N_{\gamma,\kappa} + N_{\gamma,\nu} R_{\gamma,\nu} R_{\gamma,\nu}^{-1} R_{\gamma,\nu} N_{\gamma,\nu} \right) \right] \]

\[ = \frac{\mathcal{C}_{\text{BR2}}}{4} B_{\gamma} \left[ \left( N_{\gamma,\kappa} R_{\gamma,\kappa} N_{\gamma,\kappa} R_{\gamma,\kappa}^{-1} \right) \left[ \left( \Lambda_{\gamma,\kappa}^{-1} \right) \right] \left[ \left( \Lambda_{\gamma,\kappa}^{-1} \right) \right] \right] \]

\[ + \left( N_{\gamma,\kappa} R_{\gamma,\nu} N_{\gamma,\nu} R_{\gamma,\nu}^{-1} \right) \left[ \left( \Lambda_{\gamma,\kappa}^{-1} \right) \right] \left[ \left( \Lambda_{\gamma,\kappa}^{-1} \right) \right] \]

\[ = \frac{\mathcal{C}_{\text{BR2}}}{4} B_{\gamma} C_{\gamma,\kappa} \left( \Lambda_{\gamma,\kappa}^{-1} \right) C_{\gamma,\kappa}^{-1} + C_{\gamma,\nu} \left( \Lambda_{\gamma,\nu}^{-1} \right) C_{\gamma,\nu}^{-1} \left( \Lambda_{\gamma,\nu}^{-1} \right) B_{\gamma}. \]

In the above derivation, we reversed the direction of \( N_{\gamma,\kappa} \) and \( N_{\gamma,\nu} \) for element \( \Omega_\nu \), and we used \( B_{\gamma} N_{\gamma,\kappa} = N_{\gamma,\kappa} B_{\gamma} \) and \( B_{\gamma} N_{\gamma,\nu} = N_{\gamma,\nu} B_{\gamma} \).

From the above expression for \( \Sigma_{\text{BR2}} \), we see that (22) is indeed the SBP generalization of the BR2 penalty (24) with \( C_{\text{BR2}} = \alpha_{\gamma,\kappa}^{-1} \). In the results section, we will refer to this scheme as SAT-BR2.
Remark 4 To the best of our knowledge, this is the first time the SBP generalization of the BR2 scheme has been presented. This is significant, because it provides a means of implementing the popular BR2 scheme with multidimensional SBP operators that do not have underlying basis functions.

5.2 The symmetric interior penalty method (SIPG)

A disadvantage of the SAT-BR2 penalties is that their $\Sigma_\gamma^{(1)}$ and $\Sigma_\gamma^D$ matrices can be computationally expensive to evaluate. This is not an issue for linear problems — since these matrices can be pre-computed and stored if sufficient memory is available — but it can be an issue in nonlinear problems when the diffusion coefficient(s) depend on the state.

In contrast to dense penalty matrices, the symmetric interior penalty method (SIPG) [16,23,28,30] uses diagonal (or block diagonal) $\Sigma_\gamma^{(1)}$ and $\Sigma_\gamma^D$ with a single parameter that is chosen to be sufficiently large to ensure stability. In this section, we demonstrate how the multidimensional SBP-SAT generalization of SIPG can be derived from the conditions in Theorem 2. First, we need the following lemma.

Lemma 2 Let $(\lambda_{\text{max}})_\gamma$ be the largest eigenvalue of $[A_{\rho\gamma} A_{\rho\gamma}]$ and let $\|A\|_2 = \sqrt{\rho(\text{tr}(A^T A))}$ denote the matrix 2-norm. Then

$$B_\gamma C_{\gamma\gamma} (A_{\gamma\gamma} A_{\gamma\gamma})^{-1} C_{\gamma\gamma}^T B_\gamma \leq \frac{(\lambda_{\text{max}})_\gamma}{\alpha_{\gamma\gamma}} \|B_\gamma \|_{R\gamma, R\gamma} H^{-\frac{1}{2}} \|B_\gamma\|_{R\gamma, R\gamma}^2.$$  \(25\)

Proof We recall a few facts that will be useful. The matrices $B_\gamma, N_{x,\gamma},$ and $N_{y,\gamma}$ are diagonal; therefore, they commute with one another. Furthermore, the diagonal matrix $B_\gamma$ holds positive cubature weights on its diagonal, so it can be factored as $B_\gamma = B_\gamma^x B_\gamma^y$. For similar reasons we can write $H_\gamma = H_\gamma^x H_\gamma^y$. Finally, $N_{x,\gamma} N_{x,\gamma}^T + N_{y,\gamma} N_{y,\gamma}^T = 1$, since $N_{x,\gamma}$ and $N_{y,\gamma}$ hold the $x$ and $y$ components of the unit normal along $\gamma$.

Now, let $u_\gamma \in \mathbb{R}^{n_{\gamma}}$ be an arbitrary solution on the nodes of the face $\gamma$. Then products with $u_\gamma$ and the matrix on the left of (25) can be bounded as follows:

$$u_\gamma^T B_\gamma C_{\gamma\gamma} (A_{\gamma\gamma} A_{\gamma\gamma})^{-1} C_{\gamma\gamma}^T B_\gamma u_\gamma$$

$$= \frac{1}{\alpha_{\gamma\gamma}} u_\gamma^T B_\gamma \left[ N_{x,\gamma} R_{\gamma\gamma} N_{y,\gamma} R_{\gamma\gamma} \right] \left[ H^{-1}_\gamma \right] \left[ A_{zx} A_{zy} \right] \left[ A_{yx} A_{yy} \right] \left[ R_{\gamma\gamma}^y R_{\gamma\gamma}^y \right] B_\gamma u_\gamma$$

$$\leq \frac{(\lambda_{\text{max}})_\gamma}{\alpha_{\gamma\gamma}} \|B_\gamma \|_{R\gamma, R\gamma} H^{-\frac{1}{2}} \|u_\gamma\|_{R\gamma, R\gamma} \left[ N_{x,\gamma} N_{y,\gamma} \right] B_\gamma^y \|u_\gamma\|_{R\gamma, R\gamma}^2$$

$$\leq \frac{(\lambda_{\text{max}})_\gamma}{\alpha_{\gamma\gamma}} \|B_\gamma \|_{R\gamma, R\gamma} H^{-\frac{1}{2}} \|u_\gamma\|_{R\gamma, R\gamma} \left[ N_{x,\gamma} N_{y,\gamma} \right] B_\gamma^y \|u_\gamma\|_{R\gamma, R\gamma}^2$$

The desired result follows from the above inequality, since $u_\gamma$ is arbitrary. □

We can now state the SAT-SIPG penalties that lead to energy stability.
Theorem 3 The discretization (5) is energy stable if

\[ \Sigma^{(1)}_\gamma = \delta^{(1)}_\gamma B_\gamma, \quad \text{and} \quad \Sigma^D_\gamma = \delta^D_\gamma B_\gamma, \]

where

\[ \delta^{(1)}_\gamma = \frac{(\lambda_{\text{max}})_\kappa \|B^{\frac{1}{2}}_\gamma R_\kappa H^{\frac{1}{2}}_{\kappa}\|_2^2}{4\alpha_{\gamma\kappa}} + \frac{(\lambda_{\text{max}})_\nu \|B^{\frac{1}{2}}_\gamma R_\nu H^{\frac{1}{2}}_{\nu}\|_2^2}{4\alpha_{\gamma\nu}}, \]

\[ \delta^D_\gamma = \frac{(\lambda_{\text{max}})_\kappa \|B^{\frac{1}{2}}_\gamma R_\kappa H^{\frac{1}{2}}_{\kappa}\|_2^2}{\alpha_{\gamma\kappa}}. \]

Proof The proof follows from Lemma 2, the conditions in Theorem 2, and the aforementioned choice \( \Sigma^{(3)}_\gamma = -\delta^{(2)}_\gamma = 1/2B_\gamma. \)

Remark 5 The SAT-SIPG penalties require that we compute \((\lambda_{\text{max}})_\kappa\) on each element. For nonlinear problems, we recommend replacing this value with an estimate for the upper bound of the spectral radius of the tensor \( \Lambda \) over all nodes of \( \kappa \); otherwise the computational advantage of SAT-SIPG over SAT-BR2 will be compromised.

Remark 6 To the best of our knowledge, this is the first time the SIPG penalty has been related to BR2 using straightforward matrix analysis. The approximations that lead to SIPG produce a more conservative bound, on the one hand, but a cheaper penalty, on the other hand.

The SIPG penalty parameters \( \delta^{(1)}_\gamma \) and \( \delta^D_\gamma \) are similar to those given by Shahbazi [30]. Indeed, we have verified that they are identical for degree \( p \) operators on simplex elements with constant-coefficient scalar diffusion, provided

1. the SBP matrices \( H_\kappa \) and \( B_\gamma \) and their corresponding nodes define cubature rules that are exact for polynomials of degree \( 2p \), and;
2. the number of SBP nodes is equal to the number of basis functions for \( P_p \).

When these two conditions are satisfied the SBP cubatures reproduce the \( L^2 \) norm on the volume and face exactly, so the inverse trace inequalities of Warburton and Hesthaven apply [31]. However, in general, \( H_\kappa \) is only exact for polynomials of degree \( 2p - 1 \) and there are more SBP nodes than basis functions in \( P_p \), so the penalties given here differ from [30]. Furthermore, the penalties \( \delta^{(1)}_\gamma \) and \( \delta^D_\gamma \) are more general than those provided in [30], because they are applicable to spatially varying tensor diffusion and elements other than simplices.

In practice, we define SBP operators on a reference element and employ a coordinate transformation for each element in the physical domain. Therefore, some remarks are warranted regarding the implementation of the SAT-SIPG penalties when coordinate transformations are used. Let \( x(\xi) \) be an affine and bijective coordinate transformation from reference space, \( \xi = [\xi, \eta]^T \in \Omega_\xi \), to physical space. When such a coordinate mapping is used with SAT-SIPG penalties, \((\lambda_{\text{max}})_\kappa\) corresponds to the largest eigenvalue of

\[ \begin{bmatrix} JA_{xx} & JA_{xy} \\ JA_{yx} & JA_{yy} \end{bmatrix}_\kappa, \]
where $J$ is a diagonal matrix holding the determinant of the mapping Jacobian at each node of $\kappa$. In addition, the penalties matrices $\Sigma_\gamma^{(1)}$ and $\Sigma_\gamma^D$ in (26) must be multiplied by the squared norm of the scaled contravariant basis vectors at the face nodes, i.e., the diagonal matrix whose $j$th entry is

$$
\| [J(n_\xi \nabla_\xi + n_\eta \nabla_\eta)]_j \|^2, \quad \forall j = 1, 2, \ldots, n_\gamma,
$$

where $J = \text{det}(\partial x/\partial \xi)$ is the determinant of the mapping Jacobian, $\nabla_\xi$ and $\nabla_\eta$ are the contravariant basis vectors, and $n_\xi$ and $n_\eta$ are the components of the unit normal on face $\gamma$ in reference space. Note that the squared norm $\| B_\gamma^2 \|_2^2$ can be pre-computed in reference space.

6 Numerical experiments

This section presents some numerical experiments to verify the theory developed in Sections 3, 4, and 5. For the verifications, we consider two families of SBP operators developed for simplex elements.

SBP-$\Omega$: These operators have strictly internal nodes, and the number of nodes is equal to the number of basis functions in $P^p(\Omega_\kappa)$; therefore, the SBP-SAT discretizations based on these operators are equivalent to collocation discontinuous-Galerkin finite-element methods. For the degree $p = 1$ and $p = 2$ operators, the SBP norm is a $2^p$ degree cubature, while for $p = 3$ and $p = 4$, the norm is a degree $2^p - 1$ cubature. Thus, for constant coefficient-diffusion, the SIPG penalty is identical to Shahbazi’s for $p = 1$ and $p = 2$, while it is different for $p = 3$ and $p = 4$ (see the discussion in Section 5.2). The SBP-$\Omega$ operators were first presented in [10]; see also [17].

SBP-$\Gamma$: These operators were designed to have $p + 1$ nodes on each face; consequently, the interpolation operator $R_{\gamma\kappa}$ uses only those nodes that lie on face $\gamma$. With the exception of $p = 1$, the SBP-$\Gamma$ operators have more SBP nodes than basis functions in $P^p(\Omega_\kappa)$. For this reason, there are no (known) basis functions associated with these operators for $p > 1$; they are finite-difference operators but not finite-element operators. The SBP-$\Gamma$ operators were presented in [5].

Table 1 summarizes the SBP-$\Omega$ and SBP-$\Gamma$ operators considered in this work. For further details on the construction of these operators, please see [5] and [10].

The SAT-SIPG and SAT-BR2 generalizations in Section 5 are implemented with face-weight parameters, $\alpha_{\gamma\kappa}$, computed using the face area as follows.

$$
\alpha_{\gamma\kappa} = \begin{cases} 
\frac{A(\gamma)}{A(T_\kappa^\xi) + 2A(T_\kappa^D)} & \gamma \in T^2, \\
\frac{2A(\gamma)}{A(T_\kappa^\xi) + 2A(T_\kappa^D)} & \gamma \in T,
\end{cases}
$$

where the function $A(\gamma)$ computes the size of face $\gamma$, i.e., length in 2D and area in 3D. The condition $\sum_{\gamma \subset T_\kappa} \alpha_{\gamma\kappa} = 1$ required in Lemma 1 is clearly satisfied by the above definition.

As noted above, the interpolation operator $R_{\gamma\kappa}$ for the SBP-$\Gamma$ discretizations depends only on the nodes on the boundary of $\gamma$, so the cost of applying $R_{\gamma\kappa}$ or...
Table 1: The SBP-Ω and SBP-Γ operators for the triangle. The open circles denote the locations of the SBP nodes, while the black squares denote the locations of the face cubature points (for a given degree $p$ SBP operator, the face cubatures are the same for both families).

| family   | degree | $p = 1$ | $p = 2$ | $p = 3$ | $p = 4$ |
|----------|--------|---------|---------|---------|---------|
| SBP-Ω    | 3 nodes| 6 nodes | 10 nodes| 15 nodes|
| SBP-Γ    | 3 nodes| 7 nodes | 12 nodes| 18 nodes|

$R^T_{\gamma \kappa}$ is $O(p)$ in two-dimensions. In three dimensions, the SBP-Γ operators have $(p + 1)(p + 2)/2$ nodes on each face and the cost of applying the interpolation operator and its transpose is $O(p^2)$. A consequence of this sparsity structure of $R_{\gamma \kappa}$, as well as the norm $H_\kappa$ being diagonal, is that the SAT-BR2 penalty matrix $\Sigma_{BR2}$ has an asymptotic cost of $O(p)$ and $O(p^2)$ in two and three dimensions, respectively, when SBP-Γ operators are used.

In contrast, the SBP-Ω operators require dense extrapolation operators, so the cost of applying $R_{\gamma \kappa}$ and $R^T_{\gamma \kappa}$ — and, consequently, the cost of $\Sigma_{BR2}$ — scales as $O(p^2)$ in two dimensions and $O(p^3)$ in three dimensions.

6.1 Accuracy study

The first experiment is intended to verify primal and adjoint consistency by examining the convergence rates of a discrete solution and an associated functional. To this end, we use a manufactured solution on the unit square $\Omega = [0, 1]^2$. Specifically, we adopt a second-order polynomial function for $\Lambda$ and a trigonometric solution for $U$ defined by, respectively,

$$\Lambda = \begin{bmatrix} x^2 + 1 & xy \\ xy & y^2 + 1 \end{bmatrix},$$

and

$$U(x, y) = \sin(2\pi x) \sin(2\pi y).$$

(27)

The source term $F$ is found by substituting $\Lambda$ and $U$ into (2). Homogeneous Dirichlet boundary conditions are applied along the boundaries of $\Omega$. In addition, the time derivative in (2) is dropped, i.e., we consider the Poisson PDE, since we are
interested in the accuracy of the spatial discretization. The functional is defined as
\[ J = \int_{\Omega} u d\Omega, \tag{28} \]
which is a special case of (11) with \( V_N = 0 \) and \( V_D = 0 \). To estimate the asymptotic convergence rates, we use a sequence of uniformly refined meshes consisting of \( K = 128, 512, 2048, \) and \( 8192 \) triangular elements. The coarsest mesh is shown in Figure 1a. The nominal element size is given by \( h \equiv 1/\sqrt{K/2} \).

Figure 2a shows the solution error measured in terms of the \( L^2 \)-norm; the integral in the \( L^2 \) norm is approximated using the SBP matrices \( H_k \) scaled appropriately by the mapping Jacobian. We see that, under mesh refinement, the solution errors are asymptotically \( O(h^{p+1}) \), which is in agreement with the design accuracy.

Figure 2b plots the errors in the SBP-SAT approximation of the functional (28). A convergence rate of \( 2p \) is achieved for both operators, which is also in agreement with the theoretical order of convergence \([23,25]\).

For this particular problem, the results show that for lower order approximations \((p = 1 \text{ and } p = 2)\), the SBP-\( \Omega \) operators produce more accurate solutions, typically by 25%. However, for the higher order discretizations \((p = 3 \text{ and } p = 4)\) the solutions based on SBP-\( \Gamma \) are more accurate. The SBP-\( \Gamma \) operators produce more accurate estimates of the functional for all orders of accuracy; however, we caution against generalizing any of these results based on this one case.

6.2 Comparison between SAT-BR2 and SAT-SIPG

Interior penalties can have a significant influence on the conditioning of the discretized systems; therefore, in this experiment we assess the conditioning imparted by the SAT-BR2 and SAT-SIPG penalties. We expect that SAT-BR2 will produce lower condition numbers, because it provides a tighter stability bound than SAT-SIPG. To numerically verify this, we compare the condition numbers of the linear systems produced by the SAT-BR2 and SAT-SIPG penalties. We also compare the corresponding solution errors, because there is often a trade-off between conditioning and accuracy.
The penalties are mesh dependent, so this experiment is performed on a structured triangular mesh that is randomly perturbed, as shown in Figure 1c. The mesh is extremely nonsmooth and almost tangled; indeed, the largest angle in the mesh is 179.90°.

The $L^2$ errors and condition numbers are listed in Tables 2 and 3, respectively, for different orders of approximations and SBP families. Table 2 shows that the SAT-BR2 and SAT-SIPG schemes produce comparable errors in the $L^2$ norm for both SBP families. This is also observable in the results of the mesh refinement study presented earlier. While the BR2 variant is somewhat more accurate for the SBP-$\Omega$ family, no obvious trend is apparent for the SBP-$\Gamma$ family.

In contrast, Table 3 shows that the BR2 variant consistently produces a smaller condition number for both families, as expected. Furthermore, at least for this particular experiment, SBP-$\Gamma$ always produces considerably lower condition numbers than SBP-$\Omega$.

Table 2: $L^2$-norm of error on the perturbed $16 \times 16$ mesh

| SBP-$\Gamma$    | SBP-$\Omega$ |
|-----------------|--------------|
| **SAT-BR2**     | **SAT-SIPG** | **SAT-BR2** | **SAT-SIPG** |
| $p = 1$         | 1.160e-1     | 8.030e-2    | 7.802e-2     | 9.045e-2     |
| $p = 2$         | 5.274e-3     | 4.720e-3    | 4.969e-3     | 5.697e-3     |
| $p = 3$         | 4.103e-4     | 4.021e-4    | 5.896e-4     | 6.274e-4     |
| $p = 4$         | 3.132e-5     | 3.163e-5    | 5.016e-5     | 5.298e-5     |

(a) Convergence rate of solution  
(b) Convergence rate of functional

Fig. 2: Convergence rate study
Table 3: Condition number on the perturbed 16 × 16 mesh

| SBP-Γ | SAT-BR2  | SAT-SIPG | SBP-Ω  | SAT-BR2  | SAT-SIPG |
|-------|----------|----------|--------|----------|----------|
| p = 1 | 4.734e5  | 4.852e5  | 1.705e6| 1.572e6  | 1.705e6  |
| p = 2 | 2.843e6  | 2.917e6  | 6.354e6| 5.514e6  | 6.354e6  |
| p = 3 | 6.977e6  | 7.032e6  | 2.008e7| 1.789e7  | 2.008e7  |
| p = 4 | 1.337e7  | 1.440e7  | 4.201e7| 3.794e7  | 4.201e7  |

6.3 Tightness of the stability bound

Since the stability conditions for both SAT-BR2 and SAT-SIPG are sufficient but not necessary, a relaxation factor $\alpha \in (0, 1]$ acting on $\Sigma^{(1)}$ may still yield a stable bilinear form. To some degree, such a relaxation factor can serve as a measure of the tightness of the stability conditions. For example, overly conservative SAT penalties will allow for a relaxation factor that is much smaller than 1; on the other hand, a necessary and sufficient stability condition would only permit $\alpha \geq 1$.

To study the tightness of the stability conditions, we consider the effect of scaling the penalties on the largest eigenvalue of the linear system, i.e. the eigenvalue with the smallest magnitude. In particular, we form the discretization for the problem defined by (27) but scale $\Sigma^{(1)}$ by $\alpha$, then evaluate the linear system’s largest eigenvalue and plot it versus the relaxation factor. Due to the expense of computing the smallest magnitude eigenvalue of large matrices, the matrices here are evaluated on the coarser 8 × 8 randomly perturbed mesh shown in Figure 1b.

The eigenvalues of smallest magnitude are plotted in Figure 3. As can be seen, the allowable relaxation factors are less than one for both penalties and all SBP operators considered, which verifies the conditions in Theorem 2. Furthermore, the smallest allowable relaxation factor is between 0.4 and 0.6 for SBP-Γ and between 0.25 and 0.4 for SBP-Ω, which suggests that the bound is relatively tight in the sense that it is not orders of magnitude larger than necessary.

Finally, the comparison between SAT-BR2 and SAT-SIPG is in agreement with the theoretical analysis in Section 5: SAT-BR2 represents a tighter stability bound and cannot tolerate as small a relaxation factor. As $\alpha$ is reduced in Figure 3, the eigenvalue corresponding to BR2 becomes positive before that of SIPG for the same SBP operator. Additionally, for SBP-Γ, a higher order approximation allows a lower relaxation factor, while for SBP-Ω the opposite trend is observed.

6.4 Unsteady energy stability

In this section, to complement the preceding investigation, we solve an unsteady problem with homogeneous Dirichlet boundary conditions and no source term using a SBP-Γ discretization with two different relaxation factors: a stable value of $\alpha = 1$ and an unstable value of $\alpha = 0.3$ (based on the results in Figure 3a). As mentioned in Section 4, 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 $\Delta t = 10^{-3}$. Once again, the $16 \times 16$ perturbed mesh shown in Figure 1c is utilized. The same random solution is prescribed as the initial condition for the same order of approximations.

Figure 4 shows the energy evolution using different order SBP-\(\Gamma\) operators with SAT-BR2 penalties. The results using SBP-\(\Omega\) and/or SAT-SIPG are qualitatively the same and are not shown here. As can be seen, all solutions based on $\alpha = 0.3$ diverge after a short period of time; note the logarithmic time scale. In contrast, the energy for solutions of the unscaled SAT-BR2 discretizations (i.e., $\alpha = 1$) is monotonically decreasing, as expected. Additionally, we see that the $p = 2$, $p = 3$, and $p = 4$ discretizations produce indistinguishable results, while $p = 1$ produces a slightly different energy history. This suggests that operators above degree one are necessary to resolve the energy history of the homogeneous model problem on this particular mesh.

7 Summary and Conclusions

We generalized the SAT methodology to accommodate multi-dimensional SBP discretizations of second-order PDEs, including SBP operators whose volume nodes do not coincide with a boundary cubature. We considered a general form of SAT that uses dense penalty coefficient matrices on each face of the SBP elements. Starting with this general framework, we carried out analyses of adjoint consistency and energy stability, and, based on these analyses, we determined conditions on the coefficient matrices that guarantee a conservative, energy-stable, primal-consistent, and adjoint-consistent discretization.

In contrast with previous finite-element analyses of interior penalties, the SAT conditions given here apply to general (tensor) diffusion coefficients and arbitrary elements. Furthermore, the conditions are entirely algebraic. Using the properties of SBP operators, our analysis accounts for inexact integration explicitly from the beginning such that numerical instability caused by aliasing errors is avoided.

Two popular interior penalty methods used in the FE community, BR2 and SIPG, were generalized to multi-dimensional SBP-SAT discretizations. We demon-
Fig. 4: Energy history of homogeneous problem using BDF2

It was demonstrated that the SIPG penalty can be obtained from BR2 using straightforward matrix analysis; to the best of our knowledge, this algebraic connection has not been previously reported.

Several numerical test cases were carried out to verify the analysis and compare the performance of SAT-BR2 and SAT-SIPG when applied in conjunction with two families of SBP operators: the so-called SBP-Γ and SBP-Ω operators. Mesh refinement studies confirmed that the discretizations achieve design order and that they produce superconvergent functionals. Comparisons between different discretizations were carried out on an extremely skewed mesh. The results suggest that the SBP-Γ operators produce better conditioned systems than the SBP-Ω operators. Our stability bound was shown to be 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.

References

1. Kreiss, 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.
2. Strand, B., “Summation by parts for finite difference approximations for d/dx,” Journal of Computational Physics, Vol. 110, No. 1, 1994, pp. 47-67.
3. Fisher, 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.
4. Crean, J., Hicken, J. E., Del Rey Fernández, D. C., Zingg, D. W., and Carpenter, M. H., “Entropy-conservative, multidimensional summation-by-parts discretization of the Euler equations,” (in preparation), 2016.
5. Hicken, 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. A1905–A1938.
6. Mattsson, K. and Nordström, J., “Summation by parts operators for finite difference approximations of second derivatives,” *Journal of Computational Physics*, Vol. 199, No. 2, 2004, pp. 503–540.
7. Mattsson, K., “Summation by Parts Operators for Finite Difference Approximations of Second-Derivatives with Variable Coefficients,” *Journal of Scientific Computing*, Vol. 51, No. 3, June 2012, pp. 650–682.
8. Del Rey Fernández, D. C. and Zingg, D. W., “Generalized summation-by-parts operators for the second derivative with a variable coefficient,” *SIAM Journal on Scientific Computing*, Vol. 37, No. 6, 2015, pp. A2840–A2864.
9. Carpenter, M. H., Nordström, J., and Gottlieb, D., “A stable and conservative interface treatment of arbitrary spatial accuracy,” *Journal of Computational Physics*, Vol. 148, No. 2, 1999, pp. 341–365.
10. Del Rey Fernández, D. C., Hicken, J. E., and Zingg, D. W., “Simultaneous approximation terms for multi-dimensional summation-by-parts operators,” *Journal of Scientific Computing*, 2016, submitted.
11. Gong, 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.
12. Carpenter, 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.
13. Arnold, 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.
14. Bassi, F., Crivellini, A., Rebay, S., and Savini, M., “Discontinuous Galerkin solution of the Reynolds-averaged Navier-Stokes and k turbulence model equations,” *Computers & Fluids*, Vol. 34, No. 45, 2005, pp. 507–540.
15. Douglas, J. and Dupont, T., “Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods Computing Methods in Applied Sciences,” *Computing Methods in Applied Sciences*, edited by R. Glowinski and J. L. Lions, Vol. 58 of Lecture Notes in Physics, chap. 6, Springer Berlin / Heidelberg, Berlin, Heidelberg, 1976, pp. 207–216.
16. Hartmann, 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.
17. Hicken, 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.
18. Babuška, I. and Miller, A., “The post-processing approach in the finite element method part 1: Calculation of displacements, stresses and other higher derivatives of the displacements,” *International Journal for Numerical Methods in Engineering*, Vol. 20, No. 6, June 1984, pp. 1085–1109.
19. Babuška, I. and Miller, A., “The post-processing approach in the finite element method–Part 2: The calculation of stress intensity factors,” *International Journal for Numerical Methods in Engineering*, Vol. 20, No. 6, June 1984, pp. 1111–1129.
20. Pierce, N. A. and Giles, M. B., “Adjoint recovery of superconvergent functionals from PDE approximations,” *SIAM Review*, Vol. 42, No. 2, 2000, pp. 247–264.
21. Giles, M. B. and Süli, E., “Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality,” *Acta Numerica*, Vol. 11, 2002, pp. 145–236.
22. Lu, J. C., *An a posteriori error control framework for adaptive precision optimization using discontinuous Galerkin finite element method*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2005.
23. Hartmann, R., “Adjoint Consistency Analysis of Discontinuous Galerkin Discretizations,” *SIAM Journal on Numerical Analysis*, Vol. 45, No. 6, 2007, pp. 2671–2696.
24. Fidkowski, K. J. and Darmofal, D. L., “Review of output-based error estimation and mesh adaptation in computational fluid dynamics,” *AIAA Journal*, Vol. 49, No. 4, April 2011, pp. 673–694.
25. Hicken, 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.
26. Hicken, J. E. and Zingg, D. W., “Dual consistency and functional accuracy: a finite-difference perspective,” *Journal of Computational Physics*, Vol. 256, Jan. 2014, pp. 161–182.
27. Lanczos, C., *Linear Differential Operators*, D. Van Nostrand Company, Limited, London, England, 1961.

28. Shahbazi, K., Mavriplis, D. J., and Burgess, N. K., “Multigrid algorithms for high-order discontinuous Galerkin discretizations of the compressible Navier-Stokes equations,” *Journal of Computational Physics*, Vol. 228, No. 21, Nov. 2009, pp. 7917–7940.

29. Cockburn, B. and Shu, C.-W., “The Local Discontinuous Galerkin Method for Time-Dependent Convection-Diffusion Systems,” *SIAM Journal on Numerical Analysis*, Vol. 35, No. 6, 1998, pp. 2440–2463.

30. Shahbazi, K., “An explicit expression for the penalty parameter of the interior penalty method,” *Journal of Computational Physics*, Vol. 205, No. 2, May 2005, pp. 401–407.

31. Warburton, T. and Hesthaven, J., “On the constants in hp-finite element trace inverse inequalities,” *Computer Methods in Applied Mechanics and Engineering*, Vol. 192, No. 25, 2003, pp. 2765 – 2773.