Generalized Summation-by-Parts Operators for First and Second Derivatives

by

David César Del Rey Fernández

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Institute for Aerospace Studies
University of Toronto

Abstract

Generalized Summation-by-Parts Operators for First and Second Derivatives

David César Del Rey Fernández
Doctor of Philosophy
Graduate Department of Institute for Aerospace Studies
University of Toronto
2015

Higher-order methods represent an attractive means of efficiently solving partial differential equations (PDEs) for certain classes of problems. The theory of classical finite-difference (FD) summation-by-parts (SBP) operators with weak imposition of boundary conditions using simultaneous approximation terms (SATs) is extended in several new directions. The SBP-SAT approach is advantageous as it leads to provably consistent, conservative, and stable higher-order discretizations, gives a straightforward means to develop numerical boundary conditions and inter-block coupling, and results in efficient parallel schemes with constant communication overhead, regardless of the order of the scheme.

We develop a framework for generalized SBP (GSBP) operators that extends the theory of classical FD-SBP operators to operators with one or more of the following characteristics: i) non-repeating interior operator, ii) nonuniform nodal distribution in the computational domain, and iii) operators that exclude one or both boundary nodes. Necessary and sufficient conditions for the existence of first-derivative GSBP operators are derived. For diagonal-norm operators, we show that if a quadrature rule with positive weights exists, then an associated GSBP operator is guaranteed to exist. This reduces the search for diagonal-norm GSBP operators to the search for quadrature rules with positive weights. Furthermore, we prove that dense-norm GSBP operators on n distinct nodes of up to order $n-1$ always exist. The framework enables one to show that many known operators have the SBP property and gives a straightforward methodology for the construction of novel operators.

We extend the GSBP framework to include approximations to the second derivative with a variable coefficient and develop GSBP operators more accurate than the application of the first-derivative operator twice that lead to stable discretizations of PDEs with cross-

derivative terms. We propose a novel decomposition of classical FD-SBP operators that leads to efficient implementations and simplifies the construction of Jacobian matrices.

We show how to construct SATs, derive various novel operators, and demonstrate that they have preferential error characteristics relative to classical FD-SBP operators in the context of the linear convection and linear convection-diffusion equations.

Acknowledgements

A PhD is a journey that at times can feel very solitary, but in fact, it is far from solitary and depends on the support of many people. I would like to express my gratitude to my supervisor Professor Zingg for giving me much freedom to explore throughout my PhD, all the while keeping me from losing my way. His dedication and attention to detail has significantly improved my work. Furthermore, his untiring efforts have developed many exceptional opportunities to disseminate my research, present at international venues, and be exposed to some of my field's most eminent practitioners. Above all though, I would like to thank him for his earnest curiosity and dedication to research. Our many conversations on numerical methods have deeply impacted my thoughts and incalculably enriched both my work and my understanding.

I would like to thank my DEC committee, Professor Groth and Professor Damaren for reading numerous reports, sitting through multiple presentations, and giving me critical and useful feedback. Outside of the context of the DECs, it was a pleasure to have interesting conversations and I have been fortunate to have developed friendships with several of their excellent students. I also take this opportunity to express my gratitude to my external examiner Professor Gassner; it was a great pleasure to converse with him about numerical methods and share in the excitement of thinking about the future directions that the work in this thesis could take as well as the field, for example using nonlinearly stable schemes. I thank my internal appraiser, Professor Pugh, who encouraged me to not limit my interpretation of a numerical method as simply a mathematical trick but to look deeper and examine the physical meaning behind the various components. Finally, I thank the chair of my FOE, Professor Stephan, who was very welcoming and immediately put me at ease before the rest of the committee arrived for the examination.

To my fellow CFD colleagues: I thank all of you; in one way or another, each of you has touched an aspect of my PhD and my life at UTIAS. In no particular order, I single out David Boom for our numerous discussions and debates on SBP methods. It has been an incredible discovery process which I am very grateful to have shared with you. I have been very fortunate to write two excellent papers with Professor Hicken; his deep insight on numerical methods has not only challenged me, but also refined my understanding. My good friend John Gatsis has been my go-to matrix theory guy. I have also shared many insightful conversations with David Brown. Ramy Rashad has not only been a good friend but also one of the first people with whom I started discussing technical issues, which has not abated to this day. Finally, I have been lucky enough to befriend Hugo Gagnon, one of the nicest people I have had the pleasure of meeting.

We are not only supported by our supervisors and colleagues but also by the dedicated staff at UTIAS and I thank Jeff Cook, Pieter Miras, Gail Holliwell, Rossanna McGregor, and Nora Burnett for their help throughout my time at UTIAS. Last but definitively not least,

I would like to thank Joan DaCosta who besides her tireless help in numerous ways, I have had the absolute pleasure to have many conversations ranging from politics to sociology to economics, and you have added another dimension to my time here.

To my parents Marina and Martin Hug, I cannot thank you enough nor properly express my gratitude for your care, support, and encouragement. Last but not least, I would like to thank my significant other Jillian Look-Foe for her support, kind temperament, and encouragement throughout the last three years. Your patience has known no bounds, especially in your tireless help in reading and editing my funny english. You are my confidante and best friend, and as they say in Spanish eres la luz de mis ojos.

Contents

List of Tables

List of Figures

List of Symbols

Alphanumeric Symbols

$\mathbf{1}_j$, $\mathbf{1}_{\mathbf{u}_h}$, $\mathbf{1}_{\mathbf{v}_h}$ Vectors of ones of size $j \times 1$, $n_{\mathbf{u}_h} \times 1$, and $n_{\mathbf{v}_h} \times 1$, respectively, where $n_{\mathbf{u}_h}$ and $n_{\mathbf{u}_h}$ are the number of nodes in the left and right element to an interface

A Matrix composed of a_j (See Theorem 4.4)

A_v Matrix with smooth coefficients

a Constant from the linear convection equation

$\mathbf{a}_j$ Vector associated with the compatibility equations (See Theorem 4.4)

$\mathcal{B}$ Variable coefficient from the linear convection-diffusion equation

B, $\mathsf{B}_{\mathbf{u}_h}$, $\mathsf{B}_{\mathbf{v}_h}$ Diagonal matrix with the variable coefficient $\mathcal{B}$ projected along the diagonal, where the subscripts are used to denote matrices from different elements

$B_i(x)$ The i^{th} Bernoulli polynomial

$||\mathcal{U}||_{\mathcal{B}}$ Weighted L_2 norm (see Section 2.3)

C Compatibility matrix

$\mathsf{C}_i^{(p)}$ Matrix used in the construction of compatible and order-matched GSBP operators with a repeating interior operator (see Section 5.3)

$\mathsf{D}_1^{(p)}$, $\mathsf{D}_{1,\mathbf{u}_h}$, $\mathsf{D}_{1,\mathbf{v}_h}$ Operator approximating the first derivative of order p, where for GSBP operators we have the decomposition $\mathsf{D}_1^{(p)} = \mathsf{H}^{-1}\mathsf{Q}$, where the additional entry in the subscript is to denote matrices from different elements

$\mathsf{D}_{i,e}^{(a,b)}$ Operator approximating the i^{th} derivative with interior order of a and a minimum order of b at and near boundary nodes, where e specifies a particular instance of such an operator

D_2 Operator approximating the second derivative

$\mathsf{D}_2(\mathsf{B})$ Operator approximating the second derivative with a variable coefficient

D_b, $\mathsf{D}_{b,\mathbf{u}_h}$, $\mathsf{D}_{b,\mathbf{v}_h}$ Matrix approximating the first derivative arising from the decomposition of order-matched GSBP operators for the second derivative (see Definition 6), where the additional entries in the subscripts are used to denote matrices from different elements

$\tilde{\mathsf{D}}_i$ Undivided difference approximation to the i^{th} derivative

D_x Multi-dimensional GSBP operator for the x derivative with decomposition $\mathsf{D}_x = \mathsf{H}^{-1}\left(\mathsf{Q}_x^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}_x\right)$

D_y Multi-dimensional GSBP operator for the y derivative with decomposition $\mathsf{D}_y = \mathsf{H}^{-1}\left(\mathsf{Q}_y^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}_y\right)$

E, $\mathsf{E}_{\mathbf{u}_h}$, $\mathsf{E}_{\mathbf{v}_h}$ Constituent matrix of a GSBP operator where $\mathsf{E} = \mathsf{Q}+\mathsf{Q}^{\mathrm{T}}$, where the subscripts are to denote matrices from different elements

E_{c} Constituent matrix of CSBP operators, where $\mathsf{E}_{\mathrm{c}} = \mathsf{Q}+\mathsf{Q}^{\mathrm{T}} = \operatorname{diag}(-1, 0, \ldots, 0, 1)$

$\mathbf{e}_{p+1}$ Error vector for first-derivative GSBP operators (see 7.5)

$\mathbf{e}_{k,s}$ Error vector for GSBP operators approximating the second derivative with a variable coefficient (see 7.6)

$\tilde{\mathsf{E}}$ Matrix associated with E (See Theorem 4.4)

$\tilde{e}_{ij}$ Components of $\tilde{\mathsf{E}}$

$\mathsf{E}_{x_{\mathrm{L}}}$, $\mathsf{E}_{x_{\mathrm{R}}}$, $\mathsf{E}_{x_{\mathrm{L}},\mathbf{u}_h}$, $\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}$, $\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}$, $\mathsf{E}_{x_{\mathrm{R}},\mathbf{v}_h}$ Matrices arising from the decomposition $\mathsf{E} = \mathsf{E}_{x_{\mathrm{L}}} - \mathsf{E}_{x_{\mathrm{L}}}$, where the additional entries in the subscripts are to denote matrices from different elements

$\mathcal{F}$ Initial condition

$\mathcal{G}$, $\mathcal{G}_{x_{\mathrm{L}}}$, $\mathcal{G}_{x_{\mathrm{R}}}$ Value of the boundary condition, unspecified, left boundary, or right boundary, respectively

H, $\mathsf{H}_{\mathbf{u}_h}$ $\mathsf{H}_{\mathbf{v}_h}$ Norm matrix of a GSBP operator, where the additional subscripts are to denote matrices from different elements

h Mesh spacing

$\tilde{\mathsf{H}}$ Modified H for corner-corrected GSBP operators

I_j Identity matrix of size $j \times j$

$J_{\mathbf{e}}$	Objective function used to optimize first-derivative GSBP operators (see 7.4)
$J_{\mathbf{e},\mathsf{D}_2}$	Objective function used to optimize the second-derivative GSBP operator (see 7.7)
J_{Q}	Objective function used for corner-corrected operators to minimize the size of the entries
$\mathcal{L}_0$	Boundary differential operator
$L_i(x)$	The i^{th} Lagrange basis function (see (4.46))
M	Matrix that arises from the decomposition of order-matched GSBP operators (see Definition 6)
n	Number of nodes in an element or block
$\|\|\mathcal{U}\|\|$	L_2 norm of $\mathcal{U}$
n_x, n_y	The x and y components of the normal vector to a surface
$\mathbf{x} \odot \mathbf{y}$	Hadamard product of $\mathbf{x}$ and $\mathbf{y}$, which is the element-wise multiplication of the two vectors
Ω	Volume
$\mathcal{P}$	Differential operator
$\mathsf{P}_{i,e}$	The i^{th} exchange matrix of size $e \times e$
p	Order of an approximation to a derivative
$\partial\Omega$	Surface of volume Ω
Q, $\mathsf{Q}_{\mathbf{u}_h}$, $\mathsf{Q}_{\mathbf{v}_h}$	Constituent matrix of a GSBP operator with the property $\mathsf{Q} + \mathsf{Q}^{\mathrm{T}} = \mathsf{E}$, where the subscripts are to denote matrices from different elements
$\mathsf{Q}^{(\mathrm{A})}$, $\mathsf{Q}^{(\mathrm{S})}$	Antisymmetric and symmetric portions of the matrix Q
$R(x,t)$	Residual arising from the Galerkin approach (see Section 4.4)
$\mathsf{R}(\mathsf{B})$, $\mathsf{R}_{\mathbf{u}_h}(\mathsf{B}_{\mathbf{u}_h})$, $\mathsf{R}_{\mathbf{v}_h}(\mathsf{B}_{\mathbf{v}_h})$	Corrective term for compatible and order-matched GSBP operators for the second derivative with a variable coefficient(see Definition 7), where the additional subscripts are used to denote matrices from different elements
R_{c}	Corrective term for compatible and order-matched GSBP operators for the second derivative with a constant coefficient

$\mathbb{R}$	The set of real numbers
$\mathbb{R}^n$	The set of vectors with n real components
$\mathbb{R}^{n\times n}$	The set of $n \times n$ matrices with real components
$\mathcal{S}$	Source term
$\mathbf{t}_{x_\mathrm{L}}$, $\mathbf{t}_{x_\mathrm{R}}$, $\mathbf{t}_{x_\mathrm{L},\mathbf{u}_h}$, $\mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}$, $\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}$, $\mathbf{t}_{x_\mathrm{R},\mathbf{v}_h}$	Vectors that project the solution to the left or right boundary, where the additional entries in the subscripts are to denote vectors from different elements
$\mathbf{T}_j$	The j^{th} truncation error vector associated with the first-derivative operator
$\theta_i^{(p)}$	The i^{th} coefficient of the corrective term for an order p GSBP operator with a repeating interior operator (see Section 5.3)
$\tilde{a}$, $\tilde{j}$	Parameters used in the construction of the HGT and HGTL nodal distributions (see Section 7.2.2)
u_1, v_1, u_n, v_n	Value at the first and last node of the vectors $\mathbf{u}$ and $\mathbf{v}$
$\mathbf{u}_h$, $\mathbf{v}_h$	Solution to a semi-discrete or fully discrete system of equations
$\tilde{u}_{x_\mathrm{R}}$, $\tilde{v}_{x_\mathrm{L}}$	Projection of the solution in the left element to the interface and the projection of the solution in the right element to the interface
$\mathbf{u}, \mathbf{v}$	Restriction of the continuous functions $\mathcal{U}$ and $\mathcal{V}$ onto a nodal distribution
$\mathcal{U}$, $\mathcal{V}$, $\mathcal{W}$	Continuous functions
V	Matrix associated with the GSBP norm matrix H (see 4.56)
$\mathbf{x}^k$	Restriction of the monomial x^k onto the nodal distribution
$\tilde{\mathsf{X}}_{\mathsf{D}_2}$	Matrix with columns that are the projection of the second derivative of the monomials
x_L, x_R	Left and right end points, respectively, of the domain
x, t	Spatial and temporal coordinates
$\tilde{x}_i$, $\tilde{w}_i$	The i^{th} scaled nodal location used and associated quadrature weight used in determining the boundary node location for the HGT and HGTL nodal distributions

Greek Symbols

$\alpha_{x_{\mathrm{L}}}$, $\beta_{x_{\mathrm{L}}}$, $\alpha_{x_{\mathrm{R}}}$, $\beta_{x_{\mathrm{R}}}$ Constants from the boundary conditions for the linear convection-diffusion equation

ϵ Constant from the linear convection-diffusion equation

$\sigma_1^{\mathbf{u}_h}$, $\sigma_2^{\mathbf{u}_h}$, $\sigma_3^{\mathbf{u}_h}$, $\sigma_1^{\mathbf{v}_h}$, $\sigma_2^{\mathbf{v}_h}$, $\sigma_3^{\mathbf{v}_h}$ Penalty parameters for the Baumann and Oden interface SATs for the linear convection-diffusion equation, where the superscripts are to denote the SAT for the element to the left and right of the interface, respectively

τ_{E} Degree of the matrix E

τ_{Q} Degree of the matrix Q

$\tau_{\mathsf{Q}^{(\mathrm{A})}}$ Degree of the matrix $\mathsf{Q}^{(A)}$

$\tau_{\mathbf{u}_h}$, $\tau_{\mathbf{v}_h}$ Penalty coefficients for interface SATs for the linear convection equation

Abbreviations

CFD Computational fluid dynamics

CGL Chebyshev-Gauss-Lobatto

CSBP Classical finite-difference summation by parts

DG Discontinuous Galerkin

DOF Degrees of freedom

ES Equally spaced

FD Finite difference

GSBP Generalized summation by parts

HGT Hybrid Gauss-trapezoidal

HGTL Hybrid Gauss-trapezoidal-Lobatto

IBP Integration by parts

LHS Left-hand side

PDE Partial differential equation

RHS Right-hand side

SAT	Simultaneous approximation term
SBP	Summation by parts

Mathematical operators

$\lceil c \rceil$	Ceiling operator which gives the smallest integer greater than or equal to c
$\lfloor c \rfloor$	Floor operator which gives the largest integer less than or equal to c

Chapter 1

Introduction

> "*For all knowledge and wonder (which is the seed of knowledge) is an impression of pleasure in itself.*"
>
> —Francis Bacon, *The Advancement of Learning Book I, i, 3*

1.1 Introduction

This thesis is concerned with the solution of partial differential equations (PDEs) and resides within a larger research program aimed at multi-disciplinary optimization of aircraft. Thus we are motivated to search for efficient solution algorithms for the Euler and Navier-Stokes equations for the purpose of providing the optimizer with the relevant aerodynamic quantities.

In this chapter, justification is given for higher-order methods as an attractive approach for the solution of a certain class of problems. A brief review of the development of summation-by-parts (SBP) methods and some of their advantageous properties are highlighted. Finally, the outline and objectives of the thesis are presented.

1.2 High-order methods for computational fluid dynamics

In this section, we discuss the application of higher-order methods to the solution of problems in computational fluid dynamics (CFD). The focus is on CFD which features a rich history [48, 50, 69], as our own interests are in the numerical solution of the compressible Euler and Navier-Stokes equations. CFD can be seen as both an alternative and a complement to wind tunnel and in-flight testing [48, 50, 69], enabling the efficient and cost-effective solution to what would otherwise be time-consuming and costly wind tunnel and in-flight tests.

When a method for the solution of PDEs is said to be of order p, this means that the solution error e varies with the mesh size, h, as

$$e = \mathcal{O}\left(h^p\right), \tag{1.1}$$

for sufficiently smooth solutions, and we take higher order to mean $p > 2$ [95]. Relation (1.1) demonstrates that for smooth solutions and a sufficiently small error tolerance, higher-order methods can achieve the same error on coarser grids as lower-order methods. Therefore, despite the fact that higher-order methods are more computationally expensive per degree of freedom than lower-order methods, there exists some critical error tolerance where each higher-order method is necessarily more efficient than the corresponding lower-order method. However, despite these potential benefits, higher-order methods have not been widely adopted and the vast majority of codes used in academia and industry are low order [94, 95]. Some of the reasons for this are that they are seen as more complicated to code and less robust [96]. Furthermore, for a certain class of problems, higher-order methods are seen as costly, relative to lower-order methods, to reach engineering tolerances [95]. Conversely, there are problems for which lower-order methods are too inaccurate, and the required mesh densities result in prohibitively costly computations [96]. Examples of this class of problems are: vortex dominated flows, aeroacoustics, and direct numerical simulations [94].

There are numerous means of discretizing a PDE or associated integral form: examples include finite-volume [68, 93], finite-element [28, 95, 101], discontinuous-Galerkin (DG) [5, 6, 31, 85], discontinuous-Galerkin spectral element [37, 54], spectral-volume [41, 52], spectral-difference [58, 97], and finite-difference (FD) [59, 79], methods. For reviews of high-order methods for aerodynamics on unstructured and structured grids see Refs. 95 and 29 respectively. Our focus is on nodal methods, where the PDE is solved at discrete points in space, in contrast to modal methods, where one solves for the coefficients of the basis expansion of the solution. Note that some modal methods based on Lagrangian basis functions can be viewed as nodal methods. Typically, for DG methods, the PDE is recast in a variational formulation. However, recently, there have been a number of developments for nodal DG methods in the form of flux reconstruction [47] and correction procedures via reconstruction [36] methods, in addition to the nodal DG methods developed by Hesthaven and Warburton [42]. In these methods, rather than using the variational formulation, the PDE is solved in the differential form. Spectral-difference methods are also solved in the differential form. It was shown by Huynh [47] that in one dimension the flux-reconstruction approach can recover several well known collocation-based nodal DG and spectral-difference schemes [47] and is therefore a synthesis of such class of methods. The benefit of flux-reconstruction schemes is that they remove the need to implement and evaluate quadrature procedures [98]. Besides having po-

tential savings in computational costs, one of the attractive features of flux-reconstruction schemes is that they are substantially more intuitive to understand and therefore easier to implement than DG methods in variational form [94]. One of the drawbacks of higher-order methods is their relative complexity; by reducing this complexity it is possible that such methods will become more broadly appealing [94].

Our work on SBP operators with simultaneous approximation terms (SATs) for the imposition of boundary conditions and inter-element coupling is of the class of nodal-based methods. We take an FD perspective, where the PDE is left in differential form, and derivatives are approximated as a linear combination of the solution at nodal locations. Our development of SBP operators highlights the similarities between the above-mentioned nodal-based methods and we see a continuing synthesis of methods. By taking an FD viewpoint, this frees us from necessarily constructing element-type operators and provides a useful alternative set of operators from those thus far constructed for flux reconstruction and associated methodologies. In addition, the FD viewpoint further simplifies the presentation of this class of methods, doing away with expansions of solutions in basis functions, etc., thereby making the SBP-SAT approach easily accessible to a wide audience.

1.3 Summation-by-parts operators and simultaneous approximation terms

This section highlights some of the major advances in the theory of SBP-SAT methods. Nearly 40 years ago, Kreiss and Scherer [56] laid out the basic theory of first-derivative SBP operators. Their goal was to bring to higher-order FD methods a systematic means of proving stability through the energy method. The basic idea of an SBP operator is to mimic the integration-by-parts (IBP) property of the first derivative. By doing so, it is then possible to use the energy method to prove that the resultant discretization is stable. An SBP operator for the first derivative, D_1, has a generic decomposition $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$, where H is symmetric positive definite and referred to as the norm matrix. The norm matrix of an SBP operator is typically either diagonal or has square blocks in the upper left-hand and lower right-hand corners, denoted diagonal-norm or block-norm, respectively and is a discrete approximation to the L_2 inner product [21, 45]. SBP operators are constructed from centered-difference interior operators of order $2p$. Centered schemes are naturally SBP on periodic domains, but to retain the SBP property on finite domains, specific boundary operators at and near boundary nodes need to be constructed. Near the boundaries, the discretization error jumps to order p for diagonal-norm operators or $2p-1$ for block-norm operators. Consequently, the global order of accuracy is $p+1$ or $2p$ for diagonal-norm operators and block-norm operators, respectively [39]. Furthermore, in general Q and H for a given order of accuracy, are not

fully defined. This means that the remaining free parameters of an SBP operator can be used to optimize the operators, for example, to reduce the truncation error [27].

During the subsequent 20 years ($1974 - 1994$), the SBP method was predominantly developed by a small group of researchers at Uppsala University (see, for example, Refs. 57, 84, and 74). Strand [87] summarized much of the accumulated theory for SBP operators as of 1994. He proved the existence of restricted block-norm operators, meaning the first row and column of the norm matrix are all zeros except the first entry, resulting in a globally $2p$ order method [39]. Moreover, he analytically derived general solutions for diagonal-norm operators with order $p \in [2, 4]$, and block-norm and restricted block-norm operators with order $p = 3$ (also see Carpenter and Gottlieb [14] for construction of Padé-type SBP operators). Both diagonal-norm and block-norm operators contain free parameters after satisfying the accuracy constraints and the SBP property. Diener et al. [27] performed a systematic study examining various means of constructing optimized instances of SBP operators for the first derivative.

Many PDEs of practical importance, for example the Navier-Stokes equations, contain second-derivative terms. Such terms can always be approximated as the application of an SBP operator for the first derivative twice, while retaining the SBP property. However, this results in an operator that has a larger interior operator, is one order less accurate, and is less dissipative of under-resolved modes than alternatives. Carpenter, Nordström and Gottlieb [15] were the first to derive block-norm minimum-stencil SBP operators for the second derivative, such that the interior operators include the same number of nodes as the interior operators for the first derivative. Subsequently, Mattsson and Nordström [65] proposed a simpler form, and investigated constructing both block- and diagonal-norm minimum-stencil SBP operators for the second derivative. In that paper, the authors found that using minimum-stencil second-derivative SBP operators for parabolic problems resulted in a convergence rate of $p + 2$ rather than the anticipated $p + 1$. This superconvergence was then theoretically proven by Svärd and Nordström [90]. In a followup paper to Ref. 65, Mattsson, Svärd, and Shoeybi [67] outlined a systematic means of constructing SBP operators for the second derivative as the application of the first-derivative operator twice plus a corrective term. Moreover, they proved that for PDEs with cross-derivative terms, such as $\frac{\partial^2}{\partial x \partial y}$, the first-derivative operators used to construct approximations to the second derivative and the cross-derivative terms need to be the same for stability. Kamakoti and Pantano [51] investigated the construction of minimum-stencil interior operators for the second derivative with variable coefficients. Recently, Mattsson [62] extended the theory presented in Ref. 67 to SBP approximations of the second derivative with a variable coefficient.

For one-dimensional, constant-coefficient Cauchy or periodic problems in Cartesian coordinates, the SBP property is sufficient to prove stability. However, for problems where boundaries or block interfaces are present, the traditional method of using injection or strong

enforcement of boundary/interface conditions destroys the SBP property. In working on spectral methods, first Funaro [34] and then Funaro and Gottlieb [35] considered the idea of weakly imposing boundary conditions using penalty methods. In these methods, both the PDE and the boundary condition are combined at the boundary nodes. In a refinement of the concept, Carpenter, Gottlieb and Abarbanel [14] proposed the SAT method for imposing boundary conditions. Even though the boundary condition is enforced weakly, using SATs, the resultant solution has a value that is within the truncation error at boundary nodes. Around the same time, Olsson [75, 76] proposed enforcing boundary conditions using the projection method. In several papers, Carpenter, Nordström, and Gottlieb [15] and Nordström and Carpenter [71, 72] extended the SAT concept to handle various types of boundary conditions, as well as block-interface conditions in curvilinear coordinates for linear problems.

Mattsson [60] systematically compared SATs and projection methods on the linear convection-diffusion equation and a linear hyperbolic system of equations. He found that strict stability was lost using the projection method. After this paper, most of the development in the SBP community for imposition of boundary conditions and block interfaces has been within the SAT framework. Some additional important contributions to SATs in the context of the compressible Navier-Stokes equations include Svärd, Carpenter, and Nordström [89], who derived far-field SATs; Svärd and Nordström [91] for no-slip wall boundary SATs; Nordström et al. [73] for block-interface SATs; and Berg and Nordström [9] for Robin solid wall boundary SATs.

For nonlinear problems, some means of dissipating under-resolved high-frequency modes can be required. Mattsson, Svärd, and Nordström [66] and then Nordström [70] developed a method for constructing dissipation models for SBP schemes that do not destroy their stability properties, while maintaining the accuracy properties of the underlying scheme. Also of interest is the use of the skew-symmetric form and entropy splitting for nonlinear PDEs [32, 33, 77, 82, 99, 100].

Though block-norm operators present an improvement in the order of accuracy compared to diagonal-norm operators, Svärd [88] proved that the SBP property is lost for curvilinear coordinates if block norms are used. Consequently, much subsequent research has concentrated on diagonal-norm SBP operators. One of the drawbacks of diagonal-norm SBP operators is that although the interior operators are of order $2p$, the boundary operators are of order p, resulting in a $p+1$ accurate scheme. This loss of accuracy can be mitigated for functionals if the SBP-SAT discretization is dual consistent. Hicken and Zingg [44] proved that when a discretization is dual consistent, functionals computed with the norm of the SBP operator are superconvergent of order $2p$ (see also Ref. 46). In part, their analysis relies on the fact that the norm of the SBP operator represents a $2p$ order quadrature rule [45]. Berg and Nordström [10] subsequently extended the ideas in Ref. 44 to include temporal

dependence.

The classical FD-SBP-SAT approach has a number of advantageous properties. It leads to consistent, conservative, and provably stable higher-order discretizations and allows for a systematic derivation of numerical boundary conditions and inter-block coupling through the use of the energy method. Construction of higher-order schemes on multi-block topologies is greatly simplified relative to, for example, halo-node approaches. This results from the fact that at block interfaces the schemes discretize the equations in each block independently. Moreover, for the same reason, they lead to efficient parallel schemes with constant communication overhead, independent of the order of the scheme.

In this thesis, we present an extension to the classical definition of FD-SBP operators, denoted generalized SBP (GSBP) operators, that allows for the inclusion of a broader set of operators into the SBP-SAT framework, which could potentially lead to more efficient discretizations. We are not the first to attempt to extend the SBP concept to a broader range of operators: Carpenter and Gottlieb [13] realized that the definition of an SBP operator in [56] applies to a broader range of operators and developed a method for constructing the constituent matrices of a unique set of SBP operators on nearly arbitrary nodal distributions. Alternatively, Carpenter, Gottlieb, and Abarbanel [14] and then Abarbanel and Chertock [1] have proposed definitions of the SBP property that, while different in spirit from that of Kreiss and Scherer [56], allow for the use of the energy method to prove stability (see also Refs. 81, and 80).

1.4 Thesis objectives and outline

For certain problems, higher-order methods can be more efficient than lower-order methods. A number of higher-order discretizations methods are available; here, we concentrate on nodal methods with the SBP property. These methods are advantageous, as they lead to consistent, conservative, and provably stable discretizations that can be efficiently parallelized. Thus, the objectives of this thesis are as follows:

- extend the theory of classical FD-SBP operators to a broader set of operators with potentially improved error characteristics relative to classical FD-SBP operators;
- develop SBP methods for the solution of PDEs with a focus on CFD; this requires the development of first- and second-derivative operators and appropriate SATs for the weak imposition of boundary and inter-element coupling; and
- construct and characterize various novel operators for first and second derivatives.

This thesis is organized as follows: in Chapter 2, we discuss boundary and initial conditions that lead to stable and well-posed PDEs. Furthermore, the energy method is introduced

and it is shown how to use it to prove the stability of a PDE in combination with particular boundary and initial conditions. In Chapter 3, the SBP-SAT approach is introduced using classical FD-SBP operators. Chapter 4 details the derivation of the theory of GSBP operators for the first derivative, while Chapter 5 concentrates on the second derivative with a variable coefficient. In Chapter 6 we show how to derive SATs for element or block coupling, while construction of GSBP operators is the focus of Chapter 7. Furthermore, in Chapter 8 we solve the steady linear convection and convection diffusion equations and discuss the efficiency of various GSBP operators. Finally, conclusions, contributions, and recommendations are presented in Chapter 9.

Chapter 2

The Energy Method

"Stability is a fundamental concept for any type of PDE approximation."

—Bertil Gustafsson, *High Order Difference Methods for Time Dependent PDE*

2.1 Introduction

We are interested in the solution of PDEs that model physical systems. Typically, these PDEs do not have closed-form solutions and we therefore resort to numerical methods. However, before attempting to solve a PDE and associated boundary and initial conditions and forcing function—generically referred to as the data of the problem—it is fundamental to ensure that a given model of the physical world will give a reasonable answer. In this chapter, the energy method is reviewed as a means of, in part, answering this question. The energy method is used to check that a PDE and its data lead to a stable problem; in addition, if a unique solution exists, a stable problem is called well posed [40]. In later chapters, we turn our attention to the discretization of well-posed PDEs, and in a similar way as for the continuous problem, we are interested in knowing whether or not our numerical method gives reasonable answers, i.e., that it is stable. By forcing our discretization procedure to mimic certain properties of the continuous problem, the energy method can again be used to answer the question of stability. The purpose of this chapter is to review the concept of well posedness and the energy method as a means of determining stability. This review is done in the context of the linear convection equation and the linear convection-diffusion equation, which are later used as test problems to characterize the SBP-SATs procedure for the discretization of PDEs and the numerical imposition of data.

Hadamard is considered the first to have proposed the concept of well posedness as a set of criteria to answer the question "Is the mathematical model useful?" [40]—where it is assumed that the model accurately represents the physical system of interest. A mathematical model is said to be well posed if a solution exists, is unique, and depends continuously on the data

of the problem [40]. The first two criteria are minimum requirements. The third criterion "ensures that perturbations, such as errors in measurement, should not unduly affect the solution" (Gustafsson et al. 1996, 106).

The third criterion, referred to as stability, can be given a rigorous mathematical representation. There are various definitions of stability; here we follow the presentation in Ref. 40 (for more information regarding stability, also see Ref. 55). Consider the following system of differential equations:

$$\begin{aligned} &\tfrac{\partial \mathcal{U}}{\partial t} = \mathcal{P}\left(x, t, \tfrac{\partial}{\partial x}\right)\mathcal{U} + \mathcal{S},\ 0 \le x \le \infty,\ t \ge t_0, \\ &\mathcal{U}\left(x, t_0\right) = \mathcal{F}(x), \end{aligned} \tag{2.1}$$

where $\mathcal{S}$ is the source term and $\mathcal{F}$ is the initial condition. The solution $\mathcal{U} = [\mathcal{U}_1, \ldots, \mathcal{U}_m]^T$ is a vector function with m components, and the differential operator is of order p and has the following form:

$$\mathcal{P}\left(x, t, \frac{\partial}{\partial x}\right) = \sum_{v=0}^{p} \mathsf{A}_v\left(x, t\right) \frac{\partial^v}{\partial x^v}, \tag{2.2}$$

where the matrix coefficients A_v are assumed to be smooth. Finally, the boundary conditions are given as

$$\mathcal{L}_0\left(t, \frac{\partial}{\partial x}\right)\mathcal{U}\left(0, t\right) = \mathcal{G}\left(t\right), \tag{2.3}$$

where $\mathcal{L}_0$ is a differential operator of order r, and usually $r \le p - 1$ [40]. For homogeneous boundary conditions, stability and well posedness are defined as follows [40]:

Definition 1. Consider problem (2.1) with $\mathcal{S} = 0$ and $\mathcal{G} = 0$. The problem is called stable if there exists an estimate

$$\|\mathcal{U}\left(\cdot, t\right)\| \le K e^{\alpha(t - t_0)} \|\mathcal{U}\left(\cdot, t_0\right)\|, \tag{2.4}$$

where K and α do not depend on $\mathcal{F}$ and t_0. In addition, if a unique, smooth solution exists, then the problem is called well posed [40].

In Definition 1, we have used the L_2 inner product and norm, which on a volume Ω are defined as

$$(\mathcal{V}, \mathcal{U}) = \int_{\Omega} \mathcal{V}\mathcal{U} \mathrm{d}\Omega,\ \|\mathcal{U}\|^2 = \int_{\Omega} \mathcal{U}\mathcal{U} \mathrm{d}\Omega. \tag{2.5}$$

For nonhomogeneous data, stability and well posedness are defined as follows [40]:

Definition 2. Problem (2.1) is strongly stable if there exists an estimate

$$\|\mathcal{U}(\cdot,t)\|^2 \leq K(t,t_0)\left[\|\mathcal{U}(\cdot,t_0)\|^2 + \int_{t_0}^{t}\left(\|\mathcal{S}(\cdot,\tau)\|^2 + |\mathcal{G}(\tau)|^2\right)\mathrm{d}\tau\right], \tag{2.6}$$

where $K(t,t_0)$ does not depend on the data and is bounded in every finite time interval. In addition, if a unique, smooth solution exists, then the problem is called strongly well posed.

One method for constructing estimates of L_2 norm of the solution, in terms of the data, is the energy method, which comprises three main steps:

i) Multiply the PDE by the transpose of the solution and integrate in space;

ii) Use IBP to convert volume integrals to surface integrals to allow the insertion of the boundary conditions; and

iii) Integrate in time to get an estimate of the solution in terms of the data.

Stability is an equally important concept for numerical solutions to PDEs. One of the keys to the energy method is IBP; this motivates the construction of approximations to derivatives that discretely mimic IBP, i.e., methods with the SBP property. For both continuous and discrete problems, the energy method provides a straightforward means of determining stability. In addition, it can sometimes be used to prove uniqueness. To understand the energy method, it is best to apply it to a number of problems. In what follows, the stability and well posedness of two simple PDEs are analyzed using the energy method. These same problems will then be discretized later in the thesis.

2.2 Linear convection equation

In this section, we consider the stability of the linear convection equation

$$\frac{\partial \mathcal{U}}{\partial t} = -a\frac{\partial \mathcal{U}}{\partial x},\ x \in [x_{\mathrm{L}}, x_{\mathrm{R}}],\ t \geq 0,\ a > 0 \tag{2.7}$$

with an initial condition

$$\mathcal{U}(x,0) = \mathcal{F}(x). \tag{2.8}$$

The goal is to determine if the supplied boundary condition, to be specified later, and initial condition lead to a stable problem. The tool used to answer this question is the energy method. Multiplying (2.7) by the solution and integrating in space gives

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{U}\frac{\partial \mathcal{U}}{\partial t}\mathrm{d}x = -a\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{U}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x. \tag{2.9}$$

In one dimension, IBP is given as

$$\int_{x_\mathrm{L}}^{x_\mathrm{R}} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x = \mathcal{V}\mathcal{U}|_{x_\mathrm{L}}^{x_\mathrm{R}} - \int_{x_\mathrm{L}}^{x_\mathrm{R}} \mathcal{U}\frac{\partial \mathcal{V}}{\partial x}\mathrm{d}x. \tag{2.10}$$

Using IBP on the RHS to replace the volume integral with a surface integral, noting that $\mathcal{U}\frac{\partial \mathcal{U}}{\partial t} = \frac{1}{2}\frac{\partial \mathcal{U}^2}{\partial t}$, (2.9) becomes

$$\int_{x_\mathrm{L}}^{x_\mathrm{R}} \frac{\partial \mathcal{U}^2}{\partial t}\mathrm{d}x = -a\,\mathcal{U}^2\big|_{x_\mathrm{L}}^{x_\mathrm{R}}\,. \tag{2.11}$$

Applying Leibniz's rule to the LHS of (2.11) gives

$$\frac{\mathrm{d}\left\|\mathcal{U}\left(\cdot,t\right)\right\|^2}{\mathrm{d}t} = -a\,\mathcal{U}^2\big|_{x_\mathrm{L}}^{x_\mathrm{R}}\,. \tag{2.12}$$

Integrating (2.12) in time and applying the initial condition gives

$$\left\|\mathcal{U}\left(\cdot,t\right)\right\|^2 = \left\|\mathcal{F}\left(\cdot\right)\right\|^2 - a\int_0^t \left(\mathcal{U}^2(x_\mathrm{R},\tau) - \mathcal{U}^2(x_\mathrm{L},\tau)\right)\mathrm{d}\tau. \tag{2.13}$$

Equation (2.13) does not yet tell us if the solution is bounded by the data, since the integral on the RHS is indeterminate; i.e., its sign is not known. To proceed, it is necessary to specify a boundary condition. Consider

$$\mathcal{U}\left(x_\mathrm{L},t\right) = \mathcal{G}_{x_\mathrm{L}}\left(t\right). \tag{2.14}$$

Inserting (2.14) into (2.13) gives

$$\left\|\mathcal{U}(\cdot,t)\right\|^2 = \left\|\mathcal{F}(\cdot)\right\|^2 - a\int_0^t \mathcal{U}^2(x_\mathrm{R},\tau)\mathrm{d}\tau + a\int_0^t \mathcal{G}_{x_\mathrm{L}}^2(x_\mathrm{L},\tau)\mathrm{d}\tau. \tag{2.15}$$

Since the integral of the square of a function is greater than or equal to zero, (2.15) gives the following estimate of the solution in terms of the data (such an estimate is typically referred to as an energy estimate):

$$\left\|\mathcal{U}(\cdot,t)\right\|^2 \leq \left\|\mathcal{F}(\cdot)\right\|^2 + a\int_0^t \mathcal{G}_{x_\mathrm{L}}^2(x_\mathrm{L},\tau)\mathrm{d}\tau. \tag{2.16}$$

Instead, consider rearranging (2.15) as

$$\|\mathcal{U}(\cdot,t)\|^2 + a\int_0^t \mathcal{U}^2(x_{\mathrm{R}},\tau)\mathrm{d}\tau = \|\mathcal{F}\|^2 + a\int_0^t \mathcal{G}_{x_{\mathrm{L}}}^2(x_{\mathrm{L}},\tau)\mathrm{d}\tau. \tag{2.17}$$

The estimate (2.17) is a stronger statement than (2.16) since it bounds not only the solution at some time, but also the solution at the right boundary, in terms of the data. From (2.13) it can be seen that specifying $\mathcal{U}(x_{\mathrm{R}},t) = \mathcal{G}_{x_{\mathrm{R}}}(t)$ does not lead to an energy estimate.

Now we consider the periodic version of (2.7). In this case, $x_{\mathrm{L}} = 0$, $x_{\mathrm{R}} = 2\pi$ and the initial function is a 2π periodic continuous function. The boundary condition now becomes $\mathcal{U}(x_{\mathrm{L}},t) = \mathcal{U}(x_{\mathrm{L}}+2\pi,t)$; therefore, (2.13) reduces to

$$\|\mathcal{U}(\cdot,t)\|^2 = \|\mathcal{F}(\cdot)\|^2, \tag{2.18}$$

and the problem is stable.

It is possible to use the energy method to prove that if a solution to (2.7) exists, it is unique. To proceed, assume to the contrary that another solution to the non-periodic problem exists; thus,

$$\begin{gathered}\frac{\partial \mathcal{V}}{\partial t} = -a\frac{\partial \mathcal{V}}{\partial x},\ x\in[x_{\mathrm{L}},x_{\mathrm{R}}],\ t\geq 0,\ a>0,\\ \mathcal{V}(x_{\mathrm{L}},t) = \mathcal{G}_{x_{\mathrm{L}}}(x),\ \text{and}\ \mathcal{V}(x,0) = \mathcal{F}(x).\end{gathered} \tag{2.19}$$

By the linearity of both the PDE and the data, $\mathcal{W} = \mathcal{U} - \mathcal{V}$ is a solution to the homogeneous problem

$$\begin{gathered}\frac{\partial \mathcal{W}}{\partial t} = -a\frac{\partial \mathcal{W}}{\partial x},\ x\in[x_{\mathrm{L}},x_{\mathrm{R}}],\ t\geq 0,\ a>0,\\ \mathcal{W}(x_{\mathrm{L}},t) = 0,\ \text{and}\ \mathcal{W}(x,0) = 0.\end{gathered} \tag{2.20}$$

Applying the energy method to (2.20) results in the estimate

$$\|\mathcal{W}\|^2 + a\int_0^t \mathcal{W}^2(x_{\mathrm{R}},\tau)\mathrm{d}\tau = 0. \tag{2.21}$$

Equation (2.21) implies that $\|\mathcal{W}\|^2 = 0$, which means that $\mathcal{U} = \mathcal{V}$. This contradicts the assumption that $\mathcal{V}$ is different from $\mathcal{U}$, and it is concluded that if a solution to (2.7) exists, it is unique.

2.3 Linear convection-diffusion equation with a variable coefficient

Consider the linear convection-diffusion equation

$$\frac{\partial \mathcal{U}}{\partial t} = -a\frac{\partial \mathcal{U}}{\partial x} + \epsilon\frac{\partial}{\partial x}\left(\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\right), \quad x \in [x_{\mathrm{L}}, x_{\mathrm{R}}],\ t \geq 0, \quad \mathcal{B} > 0, \text{ and } a, \epsilon > 0, \tag{2.22}$$

where the boundary conditions and an initial condition are given as

$$\alpha_{x_{\mathrm{L}}}\mathcal{U}_{x_{\mathrm{L}}} + \beta_{x_{\mathrm{L}}}\mathcal{B}_{x_{\mathrm{L}}}\left.\frac{\partial \mathcal{U}}{\partial x}\right|_{x_{\mathrm{L}}} = \mathcal{G}_{x_{\mathrm{L}}},\ \alpha_{x_{\mathrm{R}}}\mathcal{U}_{x_{\mathrm{R}}} + \beta_{x_{\mathrm{R}}}\mathcal{B}_{x_{\mathrm{R}}}\left.\frac{\partial \mathcal{U}}{\partial x}\right|_{x_{\mathrm{R}}} = \mathcal{G}_{x_{\mathrm{R}}},\ \text{and } \mathcal{U}(x,0) = \mathcal{F}(x), \tag{2.23}$$

where we use the notation, for example $\mathcal{U}_{x_{\mathrm{L}}} = \mathcal{U}(x_{\mathrm{L}})$. The coefficients in the Robin boundary conditions are restricted to $\alpha_{x_{\mathrm{L}}}$, $\beta_{x_{\mathrm{L}}}$, $\beta_{x_{\mathrm{R}}} \neq 0$. It is possible to systematically consider other boundary conditions, for example, Dirichlet at both boundaries. However, here we limit the analysis to the Robin boundary conditions (2.23) as these are the boundary conditions that are implemented using the SAT procedure described Chapter 3.

For the second derivative, IBP has the form

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\frac{\partial}{\partial x}\left(\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\right)\mathrm{d}x = \mathcal{V}\mathcal{B}\left.\frac{\partial \mathcal{U}}{\partial x}\right|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} - \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \frac{\partial \mathcal{V}}{\partial x}\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x. \tag{2.24}$$

Applying the energy method to (2.22) results in

$$\frac{\mathrm{d}\|\mathcal{U}\|^2}{\mathrm{d}t} = -a\,\mathcal{U}^2\big|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} + 2\epsilon\mathcal{U}\mathcal{B}\left.\frac{\partial \mathcal{U}}{\partial x}\right|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} - 2\epsilon\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \frac{\partial \mathcal{U}}{\partial x}\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x. \tag{2.25}$$

Using the boundary conditions (2.23) to remove the derivative terms in (2.25) results in

$$\frac{\mathrm{d}\|\mathcal{U}\|^2}{\mathrm{d}t} + 2\epsilon\left\|\frac{\partial \mathcal{U}}{\partial x}\right\|_{\mathcal{B}}^2 = -\left(a + \frac{2\epsilon\alpha_{x_{\mathrm{R}}}}{\beta_{x_{\mathrm{R}}}}\right)\mathcal{U}_{x_{\mathrm{R}}}^2 + \left(a + \frac{2\epsilon\alpha_{x_{\mathrm{L}}}}{\beta_{x_{\mathrm{L}}}}\right)\mathcal{U}_{x_{\mathrm{L}}}^2 + 2\epsilon\left(\frac{\mathcal{U}_{x_{\mathrm{R}}}\mathcal{G}_{x_{\mathrm{R}}}}{\beta_{x_{\mathrm{R}}}} - \frac{\mathcal{U}_{x_{\mathrm{L}}}\mathcal{G}_{x_{\mathrm{L}}}}{\beta_{x_{\mathrm{L}}}}\right), \tag{2.26}$$

where the weighted norm $\|\mathcal{V}\|_{\mathcal{B}}^2 = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}^2\mathcal{B}\mathrm{d}x$ has been used to replace the integral. Completing the square in (2.26) gives

$$\frac{\mathrm{d}\|\mathcal{U}(\cdot,t)\|^2}{\mathrm{d}t} + 2\epsilon\left\|\frac{\partial \mathcal{U}}{\partial x}\right\|_{\mathcal{B}}^2 = \gamma_{x_{\mathrm{R}}}\left(\mathcal{U}_{x_{\mathrm{R}}} + \Gamma_{x_{\mathrm{R}}}\mathcal{G}_{x_{\mathrm{R}}}\right)^2 + \gamma_{x_{\mathrm{L}}}\left(\mathcal{U}_{x_{\mathrm{L}}} + \Gamma_{x_{\mathrm{L}}}\mathcal{G}_{x_{\mathrm{L}}}\right)^2 - \gamma_{x_{\mathrm{R}}}\Gamma_{x_{\mathrm{R}}}^2\mathcal{G}_{x_{\mathrm{R}}}^2 - \gamma_{x_{\mathrm{L}}}\Gamma_{x_{\mathrm{L}}}^2\mathcal{G}_{x_{\mathrm{L}}}^2, \tag{2.27}$$

where

$$\gamma_{x_{\mathrm{R}}} = -\left(a + \frac{2\epsilon\alpha_{x_{\mathrm{R}}}}{\beta_{x_{\mathrm{R}}}}\right),\ \Gamma_{x_{\mathrm{R}}} = -\frac{\epsilon}{a\beta_{x_{\mathrm{R}}} + 2\epsilon\alpha_{x_{\mathrm{R}}}},\ \gamma_{x_{\mathrm{L}}} = a + \frac{2\epsilon\alpha_{x_{\mathrm{L}}}}{\beta_{x_{\mathrm{L}}}},$$
$$\text{and } \Gamma_{x_{\mathrm{L}}} = \frac{\epsilon}{a\beta_{x_{\mathrm{L}}} + 2\epsilon\alpha_{x_{\mathrm{L}}}}. \tag{2.28}$$

Integrating in time and applying the initial condition results in

$$\begin{aligned}\|\mathcal{U}(\cdot,t)\|^2 + 2\epsilon\int_0^t \left\|\tfrac{\partial\mathcal{U}(\cdot,\tau)}{\partial x}\right\|_{\mathcal{B}}^2 \mathrm{d}\tau &= \|\mathcal{F}(\cdot)\|^2 \\ &\quad + \int_0^t \left(\gamma_{x_{\mathrm{R}}}\left(\mathcal{U}_{x_{\mathrm{R}}} + \Gamma_{x_{\mathrm{R}}}\mathcal{G}_{x_{\mathrm{R}}}\right)^2 + \gamma_{x_{\mathrm{L}}}\left(\mathcal{U}_{x_{\mathrm{L}}} + \Gamma_{x_{\mathrm{L}}}\mathcal{G}_{x_{\mathrm{L}}}\right)^2\right)\mathrm{d}\tau \\ &\quad - \int_0^t \left(\gamma_{x_{\mathrm{R}}}\Gamma_{x_{\mathrm{R}}}^2\mathcal{G}_{x_{\mathrm{R}}}^2 + \gamma_{x_{\mathrm{L}}}\Gamma_{x_{\mathrm{L}}}^2\mathcal{G}_{x_{\mathrm{L}}}^2\right)\mathrm{d}\tau.\end{aligned} \tag{2.29}$$

If $\gamma_{x_{\mathrm{R}}} < 0$ and $\gamma_{x_{\mathrm{L}}} < 0$, that is

$$a + \frac{2\epsilon\alpha_{x_{\mathrm{R}}}}{\beta_{x_{\mathrm{R}}}} > 0 \text{ and } a + \frac{2\epsilon\alpha_{x_{\mathrm{L}}}}{\beta_{x_{\mathrm{L}}}} < 0, \tag{2.30}$$

then it is possible to construct an estimate of the solution in terms of the data and we have shown that the problem is strongly stable. Furthermore, as with the linear convection equation, it can be proven that if a solution exists, it is unique.

2.4 Summary

In this chapter, we have discussed the concept of well-posed problems originating from mathematical models of physical systems. A well-posed problem has a unique solution that is stable. If a problem is not stable, then small changes in the data lead to large changes in the solution, making the solution of ill-posed problems difficult if not impossible with standard approaches.

The energy method was reviewed as a means of determining the stability of PDEs and associated data. In particular, the energy method was applied to the linear convection equation and the linear convection-diffusion equation to determine the conditions on the boundary conditions such that the resultant PDE and data are stable. In the energy method, the PDE is multiplied by the solution and integrated in space, then IBP is used to convert volume integrals into surface integrals so that the boundary conditions can be inserted. It is the IBP property of the first and second derivatives that is crucial to the application of the energy method. Our goal is to construct discretizations that mimic the IBP property

of the derivative operators so that the energy method can be utilized to prove that the semi-discrete and discrete equations that result from the application of the discretization procedure are stable. In the next chapter, we examine how to construct FD operators and numerical boundary conditions for which the energy method can be applied to prove that the semi-discrete equations are stable in an analogous way to the analysis of this chapter.

Chapter 3

Classical Finite-Difference Summation-by-Parts Operators

3.1 Introduction

Our interest is in the construction of consistent, conservative, and stable discretizations. In Chapter 2, the energy method was used to prove that a PDE and associated data are stable. We would like to construct a method for discretizing derivatives and implementing data such that the energy method can be used to prove the stability of the resulting discrete equations. In this chapter, the focus is on classical FD-SBP operators [22, 56, 87, 92] for the discretization of derivatives, and SATs [9,14,15,71–73,89,91] for the imposition of data. The analysis of Chapter 2 is replicated in the context of the semi-discrete equations. Here, we provide a brief exposition of the SBP-SAT method; this is done to motivate later work. In Chapter 2, we saw that the key ingredient in the energy method is IBP; this enables volume integrals to be converted to surface integrals and thereby the introduction of the data into the estimate on the solution. Similarly, we would like to define first-derivative and second-derivative operators that are mimetic of IBP. Before entering the analysis, we present the basic notation used in this thesis, some of which has already been implicitly used.

3.2 Notation

The basic notation used in this thesis is based on the notation in Refs. 44, 22, and 21. Vectors are denoted with small bold letters, for example, $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$, while matrices are presented using capital letters with sans-serif font, for example, M. Capital letters with script type are used to denote continuous functions on a specified domain $x \in [x_{\mathrm{L}}, x_{\mathrm{R}}]$. As an example, $\mathcal{U}(x) \in C^{\infty}[x_{\mathrm{L}}, x_{\mathrm{R}}]$ denotes an infinitely differentiable function on the domain $x \in [x_{\mathrm{L}}, x_{\mathrm{R}}]$. Lower-case bold font is used to denote the restriction of such functions onto a

grid; for example, the restriction of $\mathcal{U}$ onto the grid $\mathbf{x}$ is given by

$$\mathbf{u} = [\mathcal{U}(x_1), \ldots, \mathcal{U}(x_n)]^{\mathrm{T}}. \tag{3.1}$$

Vectors with a subscript h, for example, $\mathbf{u}_h \in \mathbb{R}^{n\times 1}$, represent the solution to a system of discrete or semi-discrete equations.

The restriction of monomials onto a set of n nodes is represented by $\mathbf{x}^k = \left[x_1^k, \ldots, x_n^k\right]^{\mathrm{T}}$, with the convention that $\mathbf{x}^k = 0$ if $k < 0$. A subscript is used to denote which derivative is being approximated and when necessary a superscript is used to denote the order of an operator. For example, $\mathsf{D}_1^{(p)}$ denotes an approximation to $\frac{\partial}{\partial x}$ of order p. We discuss the degree of SBP operators, that is, the degree of monomial for which they are exact, as well as the order of the operators. A matrix operator approximating a derivative has a leading truncation error term for each node proportional to some power of h. The order of the operator is taken as the smallest exponent of h in these truncation errors. The relation between degree and order for an operator approximating the m^{th} derivative is

$$\text{Order} = \text{degree} - m + 1. \tag{3.2}$$

For classical FD-SBP operators, the first and second derivatives are of different order on the interior and near the boundary. For later use, in order to differentiate between operators and the various orders, when necessary a superscript is appended to operators for the orders and a subscript is appended to denote which derivative is being approximated. For example, $\mathsf{D}_{i,e}^{(a,b)}$ denotes the operator for the i^{th} derivative with interior order of a and a minimum order of b at and near boundary nodes, while the additional subscript e is to differentiate among various versions of the operator. In some cases, one or several of the superscripts are not of interest and are replaced with colons; as an example, $\mathsf{D}_3^{(2,:)}$ denotes an approximation to the third derivative which is of order 2 on the interior, where the minimum order of nodes near and at the boundary is not specified.

In this thesis, we make a distinction between operators that have a diagonal norm H, which are referred to as diagonal-norm operators, and operators that have H that is not diagonal, which we denote as dense-norm operators. However, this does not imply that dense H are necessarily dense in the usual sense of that word. Furthermore, in keeping with the literature, for classical FD-SBP operators dense-norm operators are referred to as block-norm operators.

3.3 Summation-by-parts operators

To begin, the equations that must be satisfied by a first-derivative operator, D_1, of degree and order p, can be constructed using the restriction of monomials onto the grid. They are

referred to as the degree equations and given as [21, 25]

$$\mathsf{D}_1\mathbf{x}^k = k\mathbf{x}^{k-1}, \quad k \in [0, p]. \tag{3.3}$$

Consider the following definition of a first-derivative operator with the SBP property [22, 56, 87, 92]:

Definition 3. SBP operator for the first derivative: A matrix operator $\mathsf{D}_1 \in \mathbb{R}^{n\times n}$ is an approximation to $\frac{\partial}{\partial x}$, on the uniform nodal distribution $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$, of order and degree p with the SBP property if

1. $\mathsf{D}_1\mathbf{x}^k = \mathsf{H}^{-1}\mathsf{Q}\mathbf{x}^k = k\mathbf{x}^{k-1}$, $k \in [0, p]$;
2. H, denoted the norm matrix, is symmetric positive definite; and
3. $\mathsf{Q} + \mathsf{Q}^{\mathrm{T}} = \mathsf{E}_{\mathrm{c}} = \mathsf{diag}\,(-1, 0 \ldots, 0, 1)$.

For the derivation of SBP operators, a more convenient form of the degree equations (3.3) that avoids the introduction of nonlinear terms is

$$\mathsf{Q}\mathbf{x}^k = k\mathsf{H}\mathbf{x}^{k-1}, \ k \in [0, p]. \tag{3.4}$$

The interior operator of classical FD-SBP operators is of order $2p$, while the boundary operators at each boundary are typically of order p or $2p-1$ for diagonal- or block-norm operators (although some can be forced to be of higher order). Consider an operator with an interior operator of order 4 on 12 nodes. The block-norm matrix is given as

$$\mathsf{H} = h\begin{bmatrix}
h_{11} & h_{12} & h_{13} & h_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
h_{12} & h_{22} & h_{23} & h_{24} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
h_{13} & h_{23} & h_{33} & h_{34} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
h_{14} & h_{24} & h_{34} & h_{44} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & h_{44} & h_{34} & h_{24} & h_{14} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & h_{34} & h_{33} & h_{23} & h_{13} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & h_{24} & h_{23} & h_{22} & h_{12} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & h_{14} & h_{13} & h_{12} & h_{11}
\end{bmatrix}, \tag{3.5}$$

or if H is diagonal it has form

$$\mathsf{H} = h\ \mathsf{diag}\,[h_{11}, h_{22}, h_{33}, h_{44}, 1, \ldots, 1, h_{44}, h_{33}, h_{22}, h_{11}]\,. \tag{3.6}$$

where h is the mesh spacing. The matrix Q has the form

$$\mathsf{Q} = \begin{bmatrix} \mathsf{Q}(1:4,1:4) & & & & & & & & & & & \\ & & & & -\frac{1}{12} & & & & & & & \\ & & & & \frac{2}{3} & -\frac{1}{12} & & & & & & \\ & & & & 0 & \frac{2}{3} & -\frac{1}{12} & & & & & \\ & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & & & & \\ & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & & & \\ & & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & & \\ & & & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & \\ & & & & & & \frac{1}{12} & -\frac{2}{3} & & & & \\ & & & & & & & \frac{1}{12} & & & & \\ & & & & & & & & & & & \\ & & & & & & & & & & & \end{bmatrix}, \tag{3.7}$$

$$\begin{bmatrix} -\frac{1}{2} & q_{12} & q_{13} & q_{14} \\ -q_{12} & 0 & q_{23} & q_{24} \\ -q_{13} & -q_{23} & 0 & q_{34} \\ -q_{14} & -q_{24} & -q_{34} & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & q_{34} & q_{24} & q_{14} \\ -q_{34} & 0 & q_{23} & q_{13} \\ -q_{24} & -q_{23} & 0 & q_{12} \\ -q_{14} & -q_{13} & -q_{12} & \frac{1}{2} \end{bmatrix}$$

$\mathsf{Q}(1:4,1:4)$ $\qquad$ $-\mathsf{P}\mathsf{Q}(1:4,1:4)\mathsf{P}$

where the matrix $\mathsf{Q}(1:4,1:4)$ is nearly skew symmetric as shown, the matrix P is the exchange matrix, i.e., it is a permutation matrix with ones along the anti-diagonal, and the operation $-\mathsf{P}\mathsf{Q}(1:4,1:4)\mathsf{P}$ results in the matrix shown in (3.7). The repeating interior operator is highlighted in blue, while the green triangles have entries that originate from the repeating interior operator and ensure that the resultant Q is nearly skew symmetric. To apply this operator on a nodal distribution with more nodes, the matrix is expanded by inserting additional interior operators, which does not require a change to the operators near the boundaries. In this example, there are 4 boundary operators at either boundary, and this is the minimum required [22]. The number of boundary operators can be increased by increasing the size of $\mathsf{Q}(1:4,1:4)$ and $-\mathsf{P}\mathsf{Q}(1:4,1:4)\mathsf{P}$; this has the potential to result in more accurate operators [64]. The entries in Q and H are specified by satisfying the degree equations (3.3) and the constraint that H be positive definite.

For diagonal-norm classical FD-SBP operators up to $p \leq 4$, using $2p$ boundary operators at the first and last $2p$ nodes leads to unique H that are positive definite. For $p > 4$ increasing numbers of boundary nodes need to be used in order to introduce degrees of freedom such that a positive-definite H can be found. An interesting paper on this subject is Albin and Klarmann [3].

We will now show how Definition 3 results in operators that discretely mimic IBP, which in the continuous case, using the L_2 inner product, can be cast as (see (2.10) and (2.5) in

Chapter 2)

$$\left(\mathcal{V}, \frac{\partial \mathcal{U}}{\partial x}\right) = \mathcal{V}\mathcal{U}|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} - \left(\frac{\partial \mathcal{V}}{\partial x}, \mathcal{U}\right). \tag{3.8}$$

The norm matrix H is positive definite and can be used to define an inner product and norm as

$$(\mathbf{v}, \mathbf{u})_{\mathsf{H}} = \mathbf{v}^{\mathrm{T}}\mathsf{H}\mathbf{u}, \quad ||\mathbf{u}||_{\mathsf{H}}^2 = \mathbf{u}^{\mathrm{T}}\mathsf{H}\mathbf{u}. \tag{3.9}$$

As will be shown in Chapter 4, the norm matrix H of an SBP operator is a discrete approximation to the L_2 inner product. Taking $\mathbf{u}$ and $\mathbf{v}$ as the projection of the continuous functions $\mathcal{U}$ and $\mathcal{V}$ onto the nodal distribution $\mathbf{x}$, then the discrete counterpart to (3.8) is

$$\mathbf{v}^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathbf{u} = (v_n u_n - v_1 u_1) - (\mathsf{D}_1\mathbf{v})^{\mathrm{T}}\,\mathsf{H}\mathbf{u}. \tag{3.10}$$

Operators that satisfy Definition 3 satisfy (3.10) by definition. To see this, substitute $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$ into (3.10), which gives

$$\mathbf{v}^{\mathrm{T}}\mathsf{Q}\mathbf{u} = (v_n u_n - v_1 u_1) - \mathbf{v}^{\mathrm{T}}\mathsf{Q}^{\mathrm{T}}\mathbf{u}. \tag{3.11}$$

Using the SBP property $\mathsf{Q} + \mathsf{Q}^{\mathrm{T}} = \mathsf{E}_{\mathrm{c}}$ gives $\mathsf{Q} = \mathsf{E}_{\mathrm{c}} - \mathsf{Q}^{\mathrm{T}}$; therefore, (3.11) becomes

$$\mathbf{v}^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathbf{u} - \mathbf{v}^{\mathrm{T}}\mathsf{Q}^{\mathrm{T}}\mathbf{u} = (v_n u_n - v_1 u_1) - \mathbf{v}^{\mathrm{T}}\mathsf{Q}^{\mathrm{T}}\mathbf{u}. \tag{3.12}$$

However, $\mathbf{v}^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathbf{u} = (v_n u_n - v_1 u_1)$, and we obtain an identity. Hence, Definition 3 leads to operators that discretely mimic IBP, with respect to the norm H. This is not the only possibility; for example see Refs. 14, 19, 1, 2, 81, and 80. Alternative definitions of the SBP property are discussed in Chapter 4.

We are also interested in approximating the second derivative with a variable coefficient, and would like operators that are amenable to the energy method. In this chapter, for simplicity, the second derivative is approximated as the application of the first derivative-operator twice. Chapter 5 deals with alternative approximations. As for the first-derivative operator, the goal is to mimic IBP, which can be cast as (see (2.24) and (2.5) in Chapter 2)

$$\left(\mathcal{V}, \frac{\partial}{\partial x}\left(\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\right)\right) = \mathcal{V}\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\bigg|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} - \left(\frac{\partial \mathcal{V}}{\partial x}, \mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\right). \tag{3.13}$$

The discrete counterpart to (3.13), using the application of the first-derivative operator twice, is given as

$$\mathbf{v}^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathsf{B}\mathsf{D}_1\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathsf{B}\mathsf{D}_1\mathbf{u} - (\mathsf{D}_1\mathbf{v})^{\mathrm{T}}\,\mathsf{B}\mathsf{D}_1\mathbf{u}, \tag{3.14}$$

where $\mathsf{B} = \mathsf{diag}\left(\mathcal{B}\left(x_1\right), \ldots, \mathcal{B}\left(x_n\right)\right)$ and $\mathbf{v}^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathsf{B}\mathsf{D}_1\mathbf{u} = v_n\left(\mathsf{D}_1\mathbf{u}\right)_n - v_1\left(\mathsf{D}_1\mathbf{u}\right)_1$. Again, not all first-derivative operators satisfy (3.14). To show that the application of a first-derivative

SBP operator twice satisfies (3.14), we use the SBP property $\mathsf{Q} = \mathsf{E}_\mathrm{c} - \mathsf{Q}^\mathrm{T}$; therefore,

$$\mathsf{D}_1\mathsf{B}\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}\mathsf{B}\mathsf{D}_1 = \mathsf{H}^{-1}\left(\mathsf{E}_\mathrm{c}\mathsf{B}\mathsf{D}_1 - \mathsf{Q}^\mathrm{T}\mathsf{B}\mathsf{D}_1\right) = \mathsf{H}^{-1}\left(\mathsf{E}_\mathrm{c}\mathsf{B}\mathsf{D}_1 - \mathsf{D}_1^\mathrm{T}\mathsf{H}\mathsf{B}\mathsf{D}_1\right). \tag{3.15}$$

Multiplying (3.15) by $\mathbf{v}^\mathrm{T}\mathsf{H}$ on the left and by $\mathbf{u}$ on the right we get

$$\mathbf{v}^\mathrm{T}\mathsf{H}\mathsf{D}_1\mathsf{B}\mathsf{D}_1\mathbf{u} = \mathbf{v}^\mathrm{T}\mathsf{E}_\mathrm{c}\mathsf{B}\mathsf{D}_1\mathbf{u} - \mathbf{v}^\mathrm{T}\mathsf{D}_1^\mathrm{T}\mathsf{H}\mathsf{B}\mathsf{D}_1\mathbf{u}, \tag{3.16}$$

and we recover (3.14). The next two sections replicate the analysis of Chapter 2 for the semi-discrete case. We show how to apply the energy method to semi-discrete equations with derivatives discretized using SBP operators. Furthermore, we introduce the use of SATs for the imposition of the data from well-posed problems.

3.4 Linear convection equation

The semi-discrete form of the linear convection equation is

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{D}_1\mathbf{u}_h, \tag{3.17}$$

with an initial condition $\mathbf{u}_h(t=0) = \mathbf{f}$, where the boundary conditions have yet to be imposed. To apply the discrete energy method, (3.17) is multiplied by $\mathbf{u}_h^\mathrm{T}\mathsf{H}$, which is the discrete analogue of multiplying by the solution and integrating in space. Thus,

$$\mathbf{u}_h^\mathrm{T}\mathsf{H}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathbf{u}_h^\mathrm{T}\mathsf{Q}\mathbf{u}_h. \tag{3.18}$$

Adding the transpose of (3.18) to (3.18)

$$\mathbf{u}_h^\mathrm{T}\mathsf{H}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{u}_h^\mathrm{T}}{\mathrm{d}t}\mathsf{H}\mathbf{u}_h = -a\mathbf{u}_h^\mathrm{T}\left(\mathsf{Q} + \mathsf{Q}^\mathrm{T}\right)\mathbf{u}_h. \tag{3.19}$$

Using the SBP property and bringing the terms on the LHS of (3.19) within the time derivative results in

$$\frac{\mathrm{d}\|\mathbf{u}_h\|_\mathsf{H}^2}{\mathrm{d}t} = -a\left(u_n^2 - u_1^2\right), \tag{3.20}$$

which is analogous to the continuous case, before time integration. In the continuos case we determined that it is necessary to add a boundary condition at the left boundary. We therefore need a method for numerically imposing this boundary condition. Consider

$$\mathbf{SAT}_{x_\mathrm{L}} = \sigma_{\mathrm{x_L}}\mathsf{H}^{-1}\mathsf{E}_{x_\mathrm{L}}\left(\mathbf{u}_h - \mathcal{G}_{x_\mathrm{L}}\mathbf{1}\right), \tag{3.21}$$

where $\mathbf{1}$ is a vector of ones of size $n \times 1$. The matrix E_{c} is decomposed as $\mathsf{E}_{\mathrm{c}} = \mathsf{E}_{x_{\mathrm{R}}} - \mathsf{E}_{x_{\mathrm{L}}}$, where

$$\mathsf{E}_{x_{\mathrm{R}}} = \mathbf{t}_{x_{\mathrm{R}}}\mathbf{t}_{x_{\mathrm{R}}}^{\mathrm{T}} = \mathsf{diag}\left(0,\ldots,0,1\right), \mathsf{E}_{x_{\mathrm{L}}} = \mathbf{t}_{x_{\mathrm{L}}}\mathbf{t}_{x_{\mathrm{L}}}^{\mathrm{T}} = \mathsf{diag}\left(1,0,\ldots,0\right), \tag{3.22}$$

$$\mathbf{t}_{x_{\mathrm{R}}} = \left[1,0,\ldots,0\right]^{\mathrm{T}}, \text{ and } \mathbf{t}_{x_{\mathrm{L}}} = \left[0,0,\ldots,0,1\right]^{\mathrm{T}}.$$

Therefore,

$$\mathsf{E}_{x_{\mathrm{L}}}\left(\mathbf{u} - \mathcal{G}_{x_{\mathrm{L}}}\mathbf{1}\right) = \left(u_1 - \mathcal{G}_{x_{\mathrm{L}}}\right)\mathbf{t}_{x_{\mathrm{L}}} = \begin{bmatrix} \left(u_1 - \mathcal{G}_{x_{\mathrm{L}}}\right) & 0 & \ldots & 0 \end{bmatrix}^{\mathrm{T}}, \tag{3.23}$$

which adds an approximation to the continuous boundary condition at the first node. The SAT is therefore given as the following vector:

$$\mathbf{SAT}_{x_{\mathrm{L}}} = \begin{bmatrix} \frac{\sigma_{x_{\mathrm{L}}}}{\mathsf{H}(1,1)}\left(u_1 - \mathcal{G}_{x_{\mathrm{L}}}\right) & 0 & \ldots & 0 \end{bmatrix}^{\mathrm{T}}. \tag{3.24}$$

The term $\mathbf{SAT}_{x_{\mathrm{L}}}$ is added to (3.17) to enforce the boundary condition weakly; that is, at the boundary nodes, both the PDE and the boundary condition are enforced simultaneously.

As will be shown next, the added SAT leads to stable semi-discrete equations for only certain values of the free parameter $\sigma_{x_{\mathrm{L}}}$. In addition, for certain values of $\sigma_{x_{\mathrm{L}}}$ not only are the semi-discrete equations stable but they can satisfy additional properties, for example dual consistency [11, 12, 44, 46]. In Chapter 6, in the context of interface SATs, we shall also see that the free parameters are used to enforce discrete conservation. We now show how to determine the restriction on $\sigma_{x_{\mathrm{L}}}$, using the energy method, such that the resulting semi-discrete equations of the linear convection equation is stable. The discretization of the linear convection equation and the SAT is given as

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}x} = -a\mathsf{D}_1\mathbf{u}_h + \mathbf{SAT}_{x_{\mathrm{L}}}. \tag{3.25}$$

Applying the energy method to (3.25) results in

$$\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} = -a\left(u_n^2 - u_1^2\right) + 2\sigma_{\mathrm{x_L}}\left(u_1^2 - u_1\mathcal{G}_{x_{\mathrm{L}}}\right). \tag{3.26}$$

Completing the square gives

$$\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} = -au_n^2 + \left(a + 2\sigma_{\mathrm{x_L}}\right)\left(u_1 - \frac{\sigma_{\mathrm{x_L}}}{a + 2\sigma_{\mathrm{x_L}}}\mathcal{G}_{x_{\mathrm{L}}}\right)^2 - \frac{\sigma_{\mathrm{x_L}}^2}{a + 2\sigma_{\mathrm{x_L}}}\mathcal{G}_{x_{\mathrm{L}}}^2, \tag{3.27}$$

and if $a + 2\sigma_{\mathrm{x_L}} < 0$, then the problem is strongly stable. This results since if $a + 2\sigma_{\mathrm{x_L}} < 0$, we have the inequality

$$\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} \leq \left|\frac{\sigma_{\mathrm{x_L}}^2}{a + 2\sigma_{\mathrm{x_L}}}\right|\mathcal{G}_{x_{\mathrm{L}}}^2. \tag{3.28}$$

Integrating (3.28) in time and using the initial condition gives

$$\|\mathbf{u}_h(t)\|_{\mathsf{H}}^2 \leq \|\mathbf{f}\|_{\mathsf{H}}^2 + \int_0^t \left| \frac{\sigma_{\mathrm{x_L}}^2}{a + 2\sigma_{\mathrm{x_L}}} \right| \mathcal{G}_{x_{\mathrm{L}}}^2 \mathrm{d}\tau, \tag{3.29}$$

which shows that the solution is bounded in terms of all of the data of the problem and hence is strongly stable [40].

3.5 Linear convection-diffusion equation

The semi-discrete equation for the linear convection-diffusion equation is

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{D}_1\mathbf{u}_h + \epsilon\mathsf{D}_1\mathsf{B}\mathsf{D}_1\mathbf{u}_h, \tag{3.30}$$

with an initial condition $\mathbf{u}_h(t=0) = \mathbf{f}$, and again, the boundary conditions are ignored for now. To apply the energy method, we first multiply (3.30) by $\mathbf{u}_h^{\mathrm{T}}\mathsf{H}$, which results in

$$\mathbf{u}_h^{\mathrm{T}}\mathsf{H}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathbf{u}_h^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathbf{u}_h + \epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathsf{B}\mathsf{D}_1\mathbf{u}_h. \tag{3.31}$$

Adding the transpose of (3.31) to (3.31) results in

$$\begin{aligned} \mathbf{u}_h^{\mathrm{T}}\mathsf{H}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{u}_h^{\mathrm{T}}}{\mathrm{d}t}\mathsf{H}\mathbf{u}_h &= -a\mathbf{u}_h^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathbf{u}_h - a\left(\mathsf{D}_1\mathbf{u}_h\right)^{\mathrm{T}}\mathsf{H}\mathbf{u}_h \\ &\quad +\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathsf{B}\mathsf{D}_1\mathbf{u}_h + \epsilon\left(\mathbf{u}_h^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathsf{B}\mathsf{D}_1\mathbf{u}_h\right)^{\mathrm{T}}. \end{aligned} \tag{3.32}$$

To simplify (3.32), we use the fact that $\mathsf{D}_1\mathsf{B}\mathsf{D}_1 = \mathsf{H}^{-1}\left(-\mathsf{D}_1^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1 + \mathsf{E}_{\mathrm{c}}\mathsf{B}\mathsf{D}_1\right)$, which gives

$$\begin{aligned} \frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} &= -a\mathbf{u}_h^{\mathrm{T}}\left(\mathsf{Q} + \mathsf{Q}^{\mathrm{T}}\right)\mathbf{u}_h \\ &\quad -\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1\mathbf{u}_h - \epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{B}\mathsf{H}\mathsf{D}_1\mathbf{u}_h \\ &\quad +\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathsf{B}\mathsf{D}_1\mathbf{u}_h + \epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{B}\mathsf{E}_{\mathrm{c}}\mathbf{u}_h. \end{aligned} \tag{3.33}$$

Since the term $\mathbf{u}_h^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathsf{B}\mathsf{D}_1\mathbf{u}_h$ is a scalar, $\mathbf{u}_h^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathsf{B}\mathsf{D}_1\mathbf{u}_h = \mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{B}\mathsf{E}_{\mathrm{c}}\mathbf{u}_h$. Therefore, (3.33) reduces to

$$\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} = -a\mathbf{u}_h^{\mathrm{T}}\left(\mathsf{Q} + \mathsf{Q}^{\mathrm{T}}\right)\mathbf{u}_h - \epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\left(\mathsf{H}\mathsf{B} + \mathsf{B}\mathsf{H}\right)\mathsf{D}_1\mathbf{u}_h + 2\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{E}_{\mathrm{c}}\mathsf{B}\mathsf{D}_1\mathbf{u}_h. \tag{3.34}$$

Using the SBP property and the definition of E_{c}, (3.34) simplifies to

$$\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} = -a\left(u_n^2 - u_1^2\right) + 2\epsilon\left[u_n\left(\mathsf{B}\mathsf{D}_1\mathbf{u}_h\right)_n - u_1\left(\mathsf{B}\mathsf{D}_1\mathbf{u}_h\right)_1\right] - \epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\left(\mathsf{H}\mathsf{B} + \mathsf{B}\mathsf{H}\right)\mathsf{D}_1\mathbf{u}_h. \tag{3.35}$$

Since H is an approximation to the L_2 inner product,

$$\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\left(\mathsf{H}\mathsf{B} + \mathsf{B}\mathsf{H}\right)\mathsf{D}_1\mathbf{u}_h \approx 2\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \frac{\partial\mathcal{U}}{\partial x}\mathcal{B}\frac{\partial\mathcal{U}}{\partial x}\mathrm{d}x,$$

we see that (3.35) mimics the continuous case. However, there is an important difference: in the continuous case, the term $\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \frac{\partial\mathcal{U}}{\partial x}\mathcal{B}\frac{\partial\mathcal{U}}{\partial x}\mathrm{d}x$ is guaranteed to be greater than or equal to 0 for $\mathcal{B} \geq 0$. On the other hand, $\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\left(\mathsf{H}\mathsf{B} + \mathsf{B}\mathsf{H}\right)\mathsf{D}_1\mathbf{u}_h$ is not guaranteed to be positive semi-definite unless H is a diagonal matrix. We therefore limit the analysis to diagonal-norm operators, in which case $\mathsf{H}\mathsf{B} = \mathsf{B}\mathsf{H}$.

As for the linear convection equation, the SATs are modelled after the continuous boundary conditions and are given as

$$\begin{aligned}\mathbf{SAT}_{x_{\mathrm{L}}} &= \sigma_{x_{\mathrm{L}}}\mathsf{H}^{-1}\mathsf{E}_{x_{\mathrm{L}}}\left(\alpha_{x_{\mathrm{L}}}\mathbf{u}_h + \beta_{x_{\mathrm{L}}}\mathsf{B}\mathsf{D}_1\mathbf{u}_h - \mathcal{G}_{x_{\mathrm{L}}}\mathbf{1}\right), \text{ and} \\ \mathbf{SAT}_{x_{\mathrm{R}}} &= \sigma_{x_{\mathrm{R}}}\mathsf{H}^{-1}\mathsf{E}_{x_{\mathrm{R}}}\left(\alpha_{x_{\mathrm{R}}}\mathbf{u}_h + \beta_{x_{\mathrm{R}}}\mathsf{B}\mathsf{D}_1\mathbf{u}_h - \mathcal{G}_{x_{\mathrm{R}}}\mathbf{1}\right),\end{aligned} \tag{3.36}$$

where $\mathcal{G}_{x_{\mathrm{R}}}$ and $\mathcal{G}_{x_{\mathrm{L}}}$ are from the continuous boundary conditions and are functions of time. Again, the SATs weakly enforce the boundary condition. Similarly, for certain choice of $\sigma_{x_{\mathrm{L}}}$ and $\sigma_{x_{\mathrm{R}}}$ the scheme is stable and may satisfy other properties, like dual consistency. We now turn to the question of stability. The semi-discrete equations, with boundary conditions enforced using SATs, are given as

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{D}_1\mathbf{u}_h + \epsilon\mathsf{D}_1\mathsf{B}\mathsf{D}_1\mathbf{u}_h + \mathbf{SAT}_{x_{\mathrm{L}}} + \mathbf{SAT}_{x_{\mathrm{R}}}. \tag{3.37}$$

Applying the energy method to (3.37) results in

$$\begin{aligned}\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} &= -a\left(u_n^2 - u_1^2\right) + 2\epsilon\left[u_n\left(\mathsf{B}\mathsf{D}_1\mathbf{u}_h\right)_n - u_1\left(\mathsf{B}\mathsf{D}_1\mathbf{u}_h\right)_1\right] - 2\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1\mathbf{u}_h \\ &\quad + 2\sigma_{x_{\mathrm{L}}}\left[\alpha_{x_{\mathrm{L}}}u_1^2 + \beta_{x_{\mathrm{L}}}u_1\left(\mathsf{B}\mathsf{D}_1\mathbf{u}\right)_1 - u_1\mathcal{G}_{x_{\mathrm{L}}}\right] \\ &\quad + 2\sigma_{x_{\mathrm{R}}}\left[\alpha_{x_{\mathrm{R}}}u_n^2 + \beta_{x_{\mathrm{R}}}u_n\left(\mathsf{B}\mathsf{D}_1\mathbf{u}\right)_n - u_n\mathcal{G}_{x_{\mathrm{R}}}\right].\end{aligned} \tag{3.38}$$

To cancel the indeterminate terms in (3.38), we use $\sigma_{x_{\mathrm{L}}} = \frac{\epsilon}{\beta_{x_{\mathrm{L}}}}$ and $\sigma_{x_{\mathrm{R}}} = \frac{-\epsilon}{\beta_{x_{\mathrm{R}}}}$, which after

completing the square, result in

$$\begin{aligned}\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}}^2}{\mathrm{d}t} &= \gamma_{x_\mathrm{R}}\left(u_n+\Gamma_{x_\mathrm{R}}\mathcal{G}_{x_\mathrm{R}}\right)^2+\gamma_{x_\mathrm{L}}\left(u_1+\Gamma_{x_\mathrm{L}}\mathcal{G}_{x_\mathrm{L}}\right)^2 \\ &\quad -\gamma_{x_\mathrm{R}}\Gamma_{x_\mathrm{R}}^2\mathcal{G}_{x_\mathrm{R}}^2-\gamma_{x_\mathrm{L}}\Gamma_{x_\mathrm{L}}^2\mathcal{G}_{x_\mathrm{L}}^2-2\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1\mathbf{u}_h,\end{aligned} \tag{3.39}$$

where γ_{x_R}, γ_{x_L}, Γ_{x_R}, and Γ_{x_L} are the same as in the continuous case (2.28). Integrating (3.39), applying the initial condition, and rearranging gives

$$\begin{aligned}\|\mathbf{u}_h\|_{\mathsf{H}}^2 &= \|\mathbf{f}\|_{\mathsf{H}}^2+\int_0^t\left(\gamma_{x_\mathrm{R}}\left(u_n+\Gamma_{x_\mathrm{R}}\mathcal{G}_{x_\mathrm{R}}(\tau)\right)^2+\gamma_{x_\mathrm{L}}\left(u_1+\Gamma_{x_\mathrm{L}}\mathcal{G}_{x_\mathrm{L}}(\tau)\right)^2\right)\mathrm{d}\tau \\ &\quad -\int_0^t\left(\gamma_{x_\mathrm{R}}\Gamma_{x_\mathrm{R}}^2\mathcal{G}_{x_\mathrm{R}}^2+\gamma_{x_\mathrm{L}}\Gamma_{x_\mathrm{L}}^2\mathcal{G}_{x_\mathrm{L}}^2+2\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1\mathbf{u}_h\right)\mathrm{d}\tau.\end{aligned} \tag{3.40}$$

Finally, since H is diagonal and therefore $\mathbf{u}_h^{\mathrm{T}}\mathsf{D}_1^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1\mathbf{u}_h\geq 0$, if $\gamma_{x_\mathrm{R}}<0$ and $\gamma_{x_\mathrm{L}}<0$, we get the estimate

$$\|\mathbf{u}_h\|_{\mathsf{H}}^2 \leq \|\mathbf{f}\|_{\mathsf{H}}^2+\int_0^t|\gamma_{x_\mathrm{R}}|\,\Gamma_{x_\mathrm{R}}^2\mathcal{G}_{x_\mathrm{R}}^2+|\gamma_{x_\mathrm{L}}|\,\Gamma_{x_\mathrm{L}}^2\mathcal{G}_{x_\mathrm{L}}^2\mathrm{d}\tau, \tag{3.41}$$

which exactly mimics the continuous case. Hence, the semi-discrete equations are strongly stable.

3.6 Summary

In this chapter, we reviewed classical FD-SBP operators for the first and second derivatives that mimic the IBP property of the respective continuous cases. In combination with SATs for the imposition of boundary conditions, we showed that the energy method can be applied to prove that the resultant semi-discrete equations are stable. The energy method is used to guide the construction of not only the continuous problem, but also the discretization. The starting point is a PDE and associated data. The energy method is used to determine the conditions on the data that lead to a stable problem. The derivatives are then discretized using FD-SBP operators, and SATs are constructed based on the continuous boundary conditions. The energy method is then employed to specify the penalty parameters from the SATs such that the semi-discrete equations are stable. In the next two chapters, we examine how to extend the definition of classical FD-SBP operators to a broader set of operators.

Chapter 4

Generalized Summation-by-Parts Operators for the First Derivative

"Pseudospectral algorithms are simply $N-th$ order finite difference methods in disguise."

—John P. Boyd, *Chebyshev and Fourier Spectral Methods*

4.1 Introduction

In this chapter, we extend the theory of classical FD-SBP operators, first proposed by Kreiss and Scherer [56] and later developed by Strand [87], to a broader set of operators derived on nonuniform nodal distributions that may or may not include nodes on the boundaries. There have been a number of extensions to the SBP concept. For example, Carpenter and Gottlieb [13] realized that SBP operators could be constructed on nearly arbitrary grids by deriving operators from the Lagrangian interpolant. More recently, Gassner [37] used these same ideas to construct diagonal-norm SBP operators on the Legendre-Gauss-Lobatto quadrature nodes. In both cases, the nodal distributions contain the boundary nodes. Other developments include work by Hicken and Zingg [45], who showed that the norms of classical FD-SBP operators are associated with Gregory-type quadrature rules. As will be shown, the definition we choose for the SBP property gives rise to operators that have properties analogous to those which are generated using the definition of the SBP property originally proposed by Kreiss and Scherer [56]. What this means is that resultant operators lead to discrete approximations to the individual components of IBP with order related to the order of the first derivative. Alternative definitions that do not have these properties, but nevertheless lead to energy estimates, are possible, for example, see Refs. 14, 19, 1, 2, 81, and 80; in Section 4.5 we briefly discuss some of the alternatives.

In the next section, the definition of an SBP operator is extended to accommodate a

broader set of operators. The definition has consequences on the properties of the constituent matrices of an SBP operator, and we derive several theorems in this regard. Section 4.3 summarizes the theory of dense-norm GSBP operators, where the details are contained in Appendix A. In Section 4.4, we revisit the work of Carpenter and Gottlieb [13] from a GSBP perspective. Finally, in Section 4.5, the GSBP theoretical framework is presented from a multi-dimensional perspective.

The novelty of the GSBP framework is that it synthesizes and unifies many ideas in the SBP and high-order communities. It accommodates a host of discrete operators with the SBP property, many of which previously would not have been thought of as having the SBP property. Furthermore, it allows for the construction of novel GSBP operators and delineates the conditions under which such operators exist, as well as their coupling to quadrature rules and the degree attainable for a specific family of operators. More importantly, it provides a simple method for constructing SATs for the broader set of operators.

4.2 The theory of one-dimensional GSBP operators

We extend the theory of classical FD-SBP operators to a broader set of operators on general nodal distributions that have one or more of the following characteristics: i) non-repeating interior operator, ii) non-uniform nodal distribution, and iii) exclude one or both boundary nodes. It is necessary to extend the definition of an SBP operator to accommodate such operators. Consider the following extended definition of an SBP operator:

Definition 4.1. ***Generalized summation-by-parts operator:*** *A matrix operator,* D_1*, is an approximation to the first derivative,* $\frac{\partial}{\partial x}$*, on a nodal distribution* $\mathbf{x}$ *that may or may not contain the end points of the domain* $x_{\mathrm{L}} < x_{\mathrm{R}}$ *of degree* p *with the SBP property if*

i) $\mathsf{D}_1\mathbf{x}^j = \mathsf{H}^{-1}\mathsf{Q}\mathbf{x}^j = j\mathbf{x}^{j-1},\ j \in [0, p]$*;*

ii) H *is symmetric and positive definite;*

iii) $\left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{E}\mathbf{x}^j = x_{\mathrm{R}}^{i+j} - x_{\mathrm{L}}^{i+j},\ i, j \leq \tau_{\mathsf{E}},\ \tau_{\mathsf{E}} \geq p$*; and*

iv) $\mathsf{Q} = \mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}$*, where* $\mathsf{Q}^{(\mathrm{A})}$ *is antisymmetric and* E *is symmetric.*

Definition 4.1 looks deceptively similar to that given in Chapter 3. However, there are several key differences: first, $\mathbf{x}$ is not assumed to be uniform and does not need to have nodes that are on the boundary of the domain, and second, E is not necessarily equal to $\mathsf{diag}\,(-1, 0, \ldots, 0, 1)$. The definition of E is crucial to applying the SBP idea to nodal distributions that do not include nodes at the boundaries or have nodal distributions that overlap the domain. The reason E is chosen to satisfy the conditions in Definition 4.1 is that

by doing so, it approximates $\mathbf{u}^{\mathrm{T}}\mathsf{E}\mathbf{u} \approx u_n^2 - u_1^2$ to at least the degree of the operator. We will explore the consequence of this choice in more detail in Section 4.5.

At this point, it is useful to introduce some additional terminology to differentiate among various SBP operators. As mentioned, the operators originally developed by Kreiss and Scherer [56] and Strand [87] are referred to as classical FD-SBP operators. These operators are characterized by a uniform nodal distribution that includes both boundary nodes and a repeating interior operator. SBP operators with a repeating interior operator are referred to as block operators. They are normally implemented in a traditional FD manner where mesh refinement is accomplished by increasing the number of mesh nodes where the interior operator is applied. Conversely, a GSBP operator with no repeating interior operator must be implemented as an element-type operator where h-refinement is carried out by increasing the number of elements while maintaining the element size. Note that a block operator can also be implemented as an element-type operator.

To prove stability, it is necessary to further decompose E. The decomposition used in this thesis is as follows [21]:

$$\mathsf{E} = \mathsf{E}_{x_{\mathrm{R}}} - \mathsf{E}_{x_{\mathrm{L}}} = \mathbf{t}_{x_{\mathrm{R}}}\mathbf{t}_{x_{\mathrm{R}}}^{\mathrm{T}} - \mathbf{t}_{x_{\mathrm{L}}}\mathbf{t}_{x_{\mathrm{L}}}^{\mathrm{T}}, \tag{4.1}$$

where

$$\mathbf{t}_{x_{\mathrm{L}}}^{\mathrm{T}}\mathbf{x}^k = x_{\mathrm{L}}^k,\ \mathbf{t}_{x_{\mathrm{R}}}^{\mathrm{T}}\mathbf{x}^k = x_{\mathrm{R}}^k,\ k \in [0, \tau_{\mathsf{E}}]. \tag{4.2}$$

By (4.2), the vectors $\mathbf{t}_{x_{\mathrm{L}}}$ and $\mathbf{t}_{x_{\mathrm{R}}}$ are degree τ_{E} approximations to $\mathcal{U}(x_{\mathrm{L}})$ and $\mathcal{U}(x_{\mathrm{R}})$. The next theorem relates the degree of $\mathbf{t}$ to its order:

Theorem 4.1. *The projection vectors $\mathbf{t}_{x_{\mathrm{L}}}$ and $\mathbf{t}_{x_{\mathrm{L}}}$ are order $\tau_{\mathsf{E}}+1$ approximations to $\mathcal{U}(x_{\mathrm{L}})$ and $\mathcal{U}(x_{\mathrm{R}})$, respectively.*

Proof. This can be shown by considering the Taylor expansion of $\mathcal{U}$ and the binomial theorem. □

The maximum degree attainable by D_1 is given by the following lemma:

Lemma 1. *The maximum attainable degree for a first-derivative operator, D_1, on a nodal distribution with n unique nodes is $n-1$; the resultant operator exists and is unique.*

Proof. Consider the degree equations for $p = n-1$, which can be compactly written as

$$\mathsf{D}_1\mathsf{X} = \tilde{\mathsf{X}}_{\mathsf{D}}, \tag{4.3}$$

where $\mathsf{X} = [\mathbf{x}^0, \ldots, \mathbf{x}^{n-1}]$ is the Vandermonde matrix and is invertible, and

$$\tilde{\mathsf{X}}_{\mathsf{D}} = \left[\mathbf{0}^0, \mathbf{x}^0 \ldots, n\mathbf{x}^{n-2}\right].$$

Since X is invertible, we can immediately write the solution as

$$\mathsf{D}_1 = \tilde{\mathsf{X}}_{\mathsf{D}} \mathsf{X}^{-1}. \tag{4.4}$$

We have shown that D_1 that is of at least degree $n-1$ exists and is unique. Furthermore, we see that D_1 maps all of $\mathbb{R}^n$ into $\mathbb{R}^{n-1}$ since X is a basis for the former and $\tilde{\mathsf{X}}_{\mathsf{D}}$ spans the latter. Since $\mathbf{x}^n$ is not in $\text{span}(\tilde{\mathsf{X}}_{\mathsf{D}})$, the degree of D_1 is $n-1$, i.e., we do not serendipitously satisfy additional degree equations. □

The form of E has been specified by assumption. However, it is not yet clear what the implications are on the remaining three matrices, H, Q, and $\mathsf{Q}^{(\mathrm{A})}$. We first characterize H by developing a set of relations between H and E. To do so, we start by left multiplying the degree equations (3.3) by H:

$$\left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right) \mathbf{x}^j = j\mathsf{H}\mathbf{x}^{j-1},\ j \in [0, p]. \tag{4.5}$$

Left multiplying (4.5) by $\left(\mathbf{x}^i\right)^{\mathrm{T}}$ gives the system of equations

$$\left(\mathbf{x}^i\right)^{\mathrm{T}} \left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right) \mathbf{x}^j = j \left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{j-1},\ i, j \in [0, p]. \tag{4.6}$$

Swapping indices in (4.6) gives

$$\left(\mathbf{x}^j\right)^{\mathrm{T}} \left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right) \mathbf{x}^i = i \left(\mathbf{x}^j\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{i-1},\ i, j \in [0, p]. \tag{4.7}$$

Adding (4.6) and (4.7) results in

$$\left(\mathbf{x}^i\right)^{\mathrm{T}} \left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right) \mathbf{x}^j + \left(\mathbf{x}^j\right)^{\mathrm{T}} \left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right) \mathbf{x}^i = j \left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{j-1} + i \left(\mathbf{x}^j\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{i-1},\ i, j \in [0, p]. \tag{4.8}$$

Using the fact that $\left(\mathbf{x}^j\right)^{\mathrm{T}} \mathsf{Q}^{(\mathrm{A})} \mathbf{x}^i$ is a scalar and $\mathsf{Q}^{(\mathrm{A})}$ is antisymmetric gives

$$\left(\mathbf{x}^j\right)^{\mathrm{T}} \mathsf{Q}^{(\mathrm{A})} \mathbf{x}^i = \left(\left(\mathbf{x}^j\right)^{\mathrm{T}} \mathsf{Q}^{(\mathrm{A})} \mathbf{x}^i\right)^{\mathrm{T}} = -\left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{Q}^{(\mathrm{A})} \mathbf{x}^j. \tag{4.9}$$

Using (4.9) and the fact that E is symmetric, (4.8) reduces to

$$j \left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{j-1} + i \left(\mathbf{x}^j\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{i-1} - \left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{E}\mathbf{x}^j = 0,\ i, j \in [0, p], \tag{4.10}$$

which are referred to as the compatibility equations [56] that H must satisfy.

In what follows, the degree of the various components of D_1 is derived. These matrices are approximations to various continuous bilinear forms and the following definition is used

in analyzing these approximations:

Definition 4. Consider the continuous bilinear form

$$(\mathcal{V},\mathcal{U}) \tag{4.11}$$

and the discrete approximation

$$(\mathbf{v},\mathbf{u})_d\,, \tag{4.12}$$

where $\mathbf{v}$ and $\mathbf{u}$ are the restrictions of the continuous functions $\mathcal{V}$ and $\mathcal{U}$ onto the grid. Approximation (4.12) of (4.11) is said to be of degree p if it is exact for the monomials $\mathcal{V} = x^i$, $\mathcal{U} = x^j$ for $i + j \leq p$; that is,

$$\left(\mathbf{x}^i,\mathbf{x}^j\right)_d = \left(x^i,x^j\right),\ i+j \leq p, \tag{4.13}$$

where i and j are integers and $i, j \in [0, p]$.

Efforts to characterize the components of D_1 have been pursued by other authors. For example, Carpenter and Gottlieb [13] characterized H and Q for a subset of what we would call GSBP operators of maximum degree that include boundary nodes (see also Ref. 37), Kitson et al. examined operators with the SBP property on periodic domains [53], and Hicken and Zingg [45] characterized H and Q for classical FD-SBP operators.

4.2.1 Diagonal-norm GSBP operators

For a diagonal-norm GSBP operator, the following theorem characterizes the norm H:

Theorem 4.2. *Given a diagonal-norm GSBP operator,* $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$ *of degree* p*, then*

$$\mathbf{v}^{\mathrm{T}}\mathsf{H}\mathbf{u} \tag{4.14}$$

is a degree $\tau_{\mathsf{H}} \geq 2p-1$ *approximation to the* L_2 *inner product; that is,*

$$\mathbf{v}^{\mathrm{T}}\mathsf{H}\mathsf{v} \approx \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\mathcal{U}\mathrm{d}x. \tag{4.15}$$

Proof. For a diagonal-norm GSBP operator, the compatibility equations (4.10) reduce to

$$\begin{aligned}\sum_{k=1}^{n}\left(jx_k^i x_k^{j-1}+ix_k^j x_k^{i-1}\right)\mathsf{H}(k,k) &= x_{\mathrm{R}}^{i+j}-x_{\mathrm{L}}^{i+j}\\ \sum_{k=1}^{n}(i+j)\,x_k^{i+j-1}\mathsf{H}(k,k) &= x_{\mathrm{R}}^{i+j}-x_{\mathrm{L}}^{i+j}\\ \sum_{k=1}^{n}x_k^{i+j-1}\mathsf{H}(k,k) &= \frac{x_{\mathrm{R}}^{i+j}-x_{\mathrm{L}}^{i+j}}{i+j},\ i,j\in[0,p],\ i\neq j\neq 0.\end{aligned} \tag{4.16}$$

Note that the condition $i=j=0$ is automatically fulfilled. The equations (4.16) are equivalent to

$$\sum_{k=1}^{n}x_k^{i+j}\mathsf{H}(k,k)=\frac{x_{\mathrm{R}}^{i+j+1}-x_{\mathrm{L}}^{i+j+1}}{i+j+1}=\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}x^i x^j\mathrm{d}x,\ i+j\in[0,2p-1], \tag{4.17}$$

and the theorem is proven. □

An immediate consequence of Theorem 4.2 is

Corollary 1. *The diagonal norm of a GSBP operator, D_1 of degree p, is associated with a degree $\tau_{\mathsf{H}}\geq 2p-1$ quadrature rule.*

Proof. By Theorem 4.2,

$$\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathsf{H}\mathbf{x}^j=\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}x^{i+j}\mathrm{d}x,\ i+j\in[0,2p-1]. \tag{4.18}$$

Taking $i=0$ gives

$$\mathbf{1}^{\mathrm{T}}\mathsf{H}\mathbf{x}^j=\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}x^j\mathrm{d}x,\ j\in[0,2p-1]. \tag{4.19}$$

These are the conditions on a degree $2p-1$ quadrature rule, and the quadrature weights are simply the diagonal elements of H. □

Theorem 4.3. *A necessary condition for the existence of a degree p GSBP operator with a diagonal norm is the existence of a quadrature rule of degree $\tau\geq 2p-1$ with positive weights.*

Proof. The theorem follows directly from Corollary 1. □

Hicken and Zingg [45] were the first to prove that the norms of classical FD-SBP operators are associated with Gregory-type quadrature rules, and Theorem 4.2 and Corollary 1 are natural extensions of that work.

The question that remains is: under what conditions do diagonal-norm GSBP operators exist? This is the subject of the next theorem.

Theorem 4.4. *A quadrature rule of degree τ with positive weights for a nodal distribution $\mathbf{x}$ is necessary and sufficient for the existence of a diagonal-norm GSBP approximation to the first derivative, $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$, that is exact for polynomials of degree $p \leq \min\left(\lceil \frac{\tau}{2} \rceil, n-1\right)$, where $n \geq 3$ is the size of D_1.*

Proof. By Theorem 4.3, a necessary condition on a diagonal-norm GSBP operator is that the diagonal entries of H be the positive weights of a quadrature rule of degree of at least $2p-1$. Therefore, $p \leq \frac{\tau+1}{2}$, but since p must be an integer, $p \leq \lceil \frac{\tau}{2} \rceil$, where $\lceil \cdot \rceil$ is the ceiling operator which gives the smallest integer greater than or equal to the argument. Moreover, by Theorem 1, the maximum degree possible is $n-1$. It is now necessary to prove that there exist Q matrices that lead to first-derivative operators of degree p. Using Definition 4.1, the degree equations (3.3) can be cast in matrix form as

$$\begin{aligned} \mathsf{D}_1\mathsf{X} = \mathsf{H}^{-1}\left(\mathsf{Q}^{(\mathrm{A})} + \tfrac{1}{2}\mathsf{E}\right)\mathsf{X} &= [\mathbf{0}, \mathbf{x}^0, \ldots, p\mathbf{x}^{p-1}, \mathbf{T}_1, \ldots, \mathbf{T}_{n-p-1}] \\ &= \tilde{\mathsf{X}}_{\mathsf{D}}, \end{aligned} \tag{4.20}$$

where the vectors $\mathbf{T}_j$ have yet to be determined. Solving for $\mathsf{Q}^{(\mathrm{A})}$ in (4.20) gives

$$\begin{aligned} \mathsf{Q}^{(\mathrm{A})} &= \left(\mathsf{H}\tilde{\mathsf{X}}_{\mathsf{D}} - \tfrac{1}{2}\mathsf{E}\mathsf{X}\right)\mathsf{X}^{-1} \\ &= \left(\mathsf{X}^{\mathrm{T}}\right)^{-1}\mathsf{X}^{\mathrm{T}}\left(\mathsf{H}\tilde{\mathsf{X}}_{\mathsf{D}} - \tfrac{1}{2}\mathsf{E}\mathsf{X}\right)\mathsf{X}^{-1} \\ &= \left(\mathsf{X}^{-1}\right)^{\mathrm{T}}\mathsf{X}^{\mathrm{T}}\left(\mathsf{H}\tilde{\mathsf{X}}_{\mathsf{D}} - \tfrac{1}{2}\mathsf{E}\mathsf{X}\right)\mathsf{X}^{-1}. \end{aligned} \tag{4.21}$$

We must now prove that the undetermined coefficients in E and the vectors $\mathbf{T}_j$ can be chosen such that the RHS of (4.21) is antisymmetric. This is equivalent to showing that

$$\mathsf{X}^{\mathrm{T}}\mathsf{H}\tilde{\mathsf{X}}_{\mathsf{D}} - \frac{1}{2}\mathsf{X}^{\mathrm{T}}\mathsf{E}\mathsf{X}, \tag{4.22}$$

is antisymmetric. We can recast (4.22) as

$$\mathsf{X}^{\mathrm{T}}\mathsf{A}, \tag{4.23}$$

where $\mathsf{A} = [\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}]$, with

$$\mathbf{a}_j = j\mathsf{H}\mathbf{x}^{j-1} - \frac{1}{2}\mathsf{E}\mathbf{x}^j. \tag{4.24}$$

To prove that antisymmetric (4.23) exist, we first show that the $(p+1) \times (p+1)$ upper

left-hand submatrix is antisymmetric and then we show that there are sufficient degrees of freedom in E and the $\mathbf{T}_j$ so that the remaining entries of (4.23) are antisymmetric.

The compatibility equations (4.10) can be rearranged as

$$j\left(\mathbf{x}^i\right)\mathsf{H}\mathbf{x}^{j-1} - \tfrac{1}{2}\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathsf{E}\mathbf{x}^j \;=\; -\left[i\left(\mathbf{x}^j\right)\mathsf{H}\mathbf{x}^{i-1} - \tfrac{1}{2}\left(\mathbf{x}^j\right)^{\mathrm{T}}\mathsf{E}\mathbf{x}^i\right] \tag{4.25}$$

and using (4.24) reduce to

$$\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathbf{a}_j = -\left(\mathbf{x}^j\right)^{\mathrm{T}}\mathbf{a}_i. \tag{4.26}$$

Note that for $i = j$, (4.25) implies that $\left(\mathbf{x}^j\right)^{\mathrm{T}}\mathbf{a}_j = 0$ and we can see that $\left(\mathsf{X}^{\mathrm{T}}\mathsf{A}\right)_{1:p+1,1:p+1}$ must be antisymmetric. To see this, consider the case of $p = 3$, then

$$\begin{aligned}
\mathsf{X}^{\mathrm{T}}\mathsf{A}\left(1:4,1:4\right) &= \begin{bmatrix}
\left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_0 & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_1 & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_2 & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_3 \\
\left(\mathbf{x}^1\right)^{\mathrm{T}}\mathbf{a}_0 & \left(\mathbf{x}^1\right)^{\mathrm{T}}\mathbf{a}_1 & \left(\mathbf{x}^1\right)^{\mathrm{T}}\mathbf{a}_2 & \left(\mathbf{x}^1\right)^{\mathrm{T}}\mathbf{a}_3 \\
\left(\mathbf{x}^2\right)^{\mathrm{T}}\mathbf{a}_0 & \left(\mathbf{x}^2\right)^{\mathrm{T}}\mathbf{a}_1 & \left(\mathbf{x}^2\right)^{\mathrm{T}}\mathbf{a}_2 & \left(\mathbf{x}^2\right)^{\mathrm{T}}\mathbf{a}_3 \\
\left(\mathbf{x}^3\right)^{\mathrm{T}}\mathbf{a}_0 & \left(\mathbf{x}^3\right)^{\mathrm{T}}\mathbf{a}_1 & \left(\mathbf{x}^3\right)^{\mathrm{T}}\mathbf{a}_2 & \left(\mathbf{x}^3\right)^{\mathrm{T}}\mathbf{a}_3
\end{bmatrix} \\
&= \begin{bmatrix}
0 & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_1 & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_2 & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_3 \\
-\left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_1 & 0 & \left(\mathbf{x}^1\right)^{\mathrm{T}}\mathbf{a}_2 & \left(\mathbf{x}^2\right)^{\mathrm{T}}\mathbf{a}_3 \\
-\left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_2 & -\left(\mathbf{x}^1\right)^{\mathrm{T}}\mathbf{a}_2 & 0 & \left(\mathbf{x}^2\right)^{\mathrm{T}}\mathbf{a}_3 \\
-\left(\mathbf{x}^0\right)^{\mathrm{T}}\mathbf{a}_3 & -\left(\mathbf{x}^1\right)^{\mathrm{T}}\mathbf{a}_3 & -\left(\mathbf{x}^2\right)^{\mathrm{T}}\mathbf{a}_3 & 0
\end{bmatrix}.
\end{aligned} \tag{4.27}$$

To show that there are sufficient degrees of freedom in E and the vectors $\mathbf{T}_j$ to make the remaining entries in (4.23) result in an antisymmetric matrix, consider constructing E from $\mathsf{X}^{\mathrm{T}}\mathsf{E}\mathsf{X} = \tilde{\mathsf{E}}$, where

$$\tilde{\mathsf{E}}(i+1,j+1) = \begin{cases} x_{\mathrm{R}}^{i+j} - x_{\mathrm{L}}^{i+j},\ i,j \in [0,p] \\ \\ \tilde{e}_{i+1,j+1} \end{cases}, \tag{4.28}$$

which results in an E that satisfies Definition 4.1.

To see that we can always choose the degrees of freedom in either $\tilde{\mathsf{E}}$ or the $\mathbf{T}_j$, consider

the case of $p = 2$ on $n = 5$ nodes, then

$$\mathsf{X}^{\mathrm{T}}\mathsf{A} =$$

$$\begin{bmatrix} x & x & x & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_1 - \frac{1}{2}\tilde{e}_{14} & \left(\mathbf{x}^0\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_2 - \frac{1}{2}\tilde{e}_{15} \\ x & x & x & \left(\mathbf{x}^1\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_1 - \frac{1}{2}\tilde{e}_{24} & \left(\mathbf{x}^1\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_2 - \frac{1}{2}\tilde{e}_{25} \\ x & x & x & \left(\mathbf{x}^2\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_1 - \frac{1}{2}\tilde{e}_{34} & \left(\mathbf{x}^2\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_2 - \frac{1}{2}\tilde{e}_{35} \\ -\frac{1}{2}\tilde{e}_{14} & (\mathbf{x}^3)^{\mathrm{T}}\mathsf{H}\mathbf{x}^0 - \frac{1}{2}\tilde{e}_{24} & 2(\mathbf{x}^3)^{\mathrm{T}}\mathsf{H}\mathbf{x}^1 - \frac{1}{2}\tilde{e}_{34} & \left(\mathbf{x}^3\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_1 - \frac{1}{2}\tilde{e}_{44} & \left(\mathbf{x}^3\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_2 - \frac{1}{2}\tilde{e}_{45} \\ -\frac{1}{2}\tilde{e}_{15} & \left(\mathbf{x}^4\right)^{\mathrm{T}}\mathsf{H}\mathbf{x}^0 - \frac{1}{2}\tilde{e}_{25} & 2\left(\mathbf{x}^4\right)^{\mathrm{T}}\mathsf{H}\mathbf{x}^1 - \frac{1}{2}\tilde{e}_{35} & \left(\mathbf{x}^4\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_1 - \frac{1}{2}\tilde{e}_{45} & \left(\mathbf{x}^4\right)^{\mathrm{T}}\mathsf{H}\mathbf{T}_1 - \frac{1}{2}\tilde{e}_{55} \end{bmatrix}, \tag{4.29}$$

where the x represent the known entries in the submatrix which we have proven is anti-symmetric. For the case where $\mathbf{T}_j$ are specified, it is immediately clear that we can choose $\tilde{e}_{ij}$ such that $\mathsf{X}^{\mathrm{T}}\mathsf{A}$ is antisymmetric. Now consider the alternative, where E is fully defined, which is mainly the case considered in this thesis. Now we show, with this example, that it is possible to sequentially solve for the $\mathbf{T}_j$ starting with $\mathbf{T}_1$, specifying the remaining degrees of freedom, such that $\mathsf{X}^{\mathrm{T}}\mathsf{A}$ is antisymmetric. We can see that the equations for $\mathbf{T}_1$ can be arranged as

$$\mathsf{X}^{\mathrm{T}}\mathsf{H}\mathbf{T}_1 = \mathbf{b}_1, \tag{4.30}$$

however, X^{T} and H are invertible, therefore $\mathbf{T}_1 = \mathsf{H}^{-1}\left(\mathsf{X}^{\mathrm{T}}\right)^{-1}\mathbf{b}_1$. Having solved for $\mathbf{T}_1$, we can proceed in a similar fashion for $\mathbf{T}_2$, which has a solution $\mathbf{T}_2 = \mathsf{H}^{-1}\left(\mathsf{X}^{\mathrm{T}}\right)^{-1}\mathbf{b}_2$. □

Theorem 4.4 implies that the search for diagonal-norm SBP operators reduces to the search for quadrature rules with positive weights. In Chapter 6, SATs are constructed for the imposition of inter-element coupling. To prove that the resulting scheme is conservative requires certain properties of Q, which are summarized in the following theorem:

Theorem 4.5. *A consistent GSBP operator,* $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$*, approximating the first derivative, has a* Q *that satisfies*

$$\mathsf{Q}\mathbf{1} = \mathbf{0} \tag{4.31}$$

and

$$\mathbf{1}^{\mathrm{T}}\mathsf{Q} = \mathbf{1}^{\mathrm{T}}\mathsf{H}\mathsf{D}_1 = \mathbf{1}^{\mathrm{T}}\mathsf{E}. \tag{4.32}$$

Proof. The above properties are well known [22] and can be derived by considering the degree equations for $\mathsf{D}_1\mathbf{0} = \mathbf{0}$ and $\mathsf{D}_1\mathbf{1} = \mathbf{0}$. □

The following theorem characterizes Q:

Theorem 4.6. *Given a GSBP approximation to the first derivative,*

$$\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q} = \mathsf{H}^{-1}\left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right)$$

of degree p with $\tau_{\mathsf{E}} \geq p$, then

$$\mathbf{v}^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathsf{Q}\mathbf{u} \tag{4.33}$$

is a degree $\tau_{\mathsf{Q}} = \min(\tau_{\mathsf{E}}, 2p)$ approximation to

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x. \tag{4.34}$$

Proof. The proof follows straightforwardly from that given by Hicken and Zingg [45] for classical FD-SBP operators. □

The estimate in Theorem 4.6 for $\mathsf{E} = \mathsf{E}_{\mathrm{c}} = \mathrm{diag}(-1, 0, \ldots, 0, 1)$ is $\tau_{\mathsf{Q}} = 2p$, since E_{c} is exact; in other words, $\tau_{\mathsf{E}_{\mathrm{c}}}$ is of infinite degree. This bound only appears to be different from that given in Hicken and Zingg [45] as a result of Definition 4 being in terms of degree rather than order.

It is also possible to characterize $\mathsf{Q}^{(\mathrm{A})}$, as given in the following theorem:

Theorem 4.7. *Given a degree p diagonal-norm GSBP operator to the first derivative, $\mathsf{D}_1 = \mathsf{H}^{-1}\left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right)$, then*

$$\mathbf{v}^{\mathrm{T}}\mathsf{Q}^{(\mathrm{A})}\mathbf{u} \tag{4.35}$$

is a degree $\tau_{\mathsf{Q}^{(\mathrm{A})}} = \min(\tau_{\mathsf{E}}, 2p)$ approximation to

$$\int_{\Omega} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}\Omega - \frac{1}{2}\oint_{\partial\Omega} \mathcal{V}\mathcal{U}n_x\mathrm{d}s, \tag{4.36}$$

where $\mathbf{n} = [n_x, n_y]^{\mathrm{T}}$ is the unit normal to the surface and Ω is the volume under consideration, with surface $\partial\Omega$. In one dimension, (4.36) reduces to

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x - \frac{1}{2}\,\mathcal{V}\mathcal{U}|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}\,. \tag{4.37}$$

Proof. Theorem 4.6 gives that

$$\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathsf{Q}\mathbf{x}^j = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^i\frac{\partial x^j}{\partial x}\mathrm{d}x,\ i + j \leq \min(\tau_{\mathsf{E}}, 2p)\,. \tag{4.38}$$

Expanding $\mathsf{Q} = \mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}$ in (4.38) and rearranging results in

$$\left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{Q}^{(\mathrm{A})}\mathbf{x}^j = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^i \frac{\partial x^j}{\partial x}\mathrm{d}x - \frac{1}{2}\left(x^i\right)^{\mathrm{T}} \mathsf{E}\mathbf{x}^j = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^i \frac{\partial x^j}{\partial x}\mathrm{d}x - \frac{1}{2}\, x^i x^j\big|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}, \; i+j \leq \min\left(\tau_{\mathsf{E}}, 2p\right); \tag{4.39}$$

thus, (4.35) is a degree $\tau_{\mathsf{Q}^{(\mathrm{A})}}$ approximation of (4.37). □

As in the case of Q, assuming that $\mathsf{E} = \mathsf{E}_{\mathrm{c}} = \mathrm{diag}\left(-1, 0, \ldots, 0, 1\right)$ gives that $\tau_{\mathsf{Q}^{(\mathrm{A})}} = 2p$. Theorems 4.6 and 4.7 show that the Definition 4.1 leads to operators that discretely approximate other bilinear forms in addition to IBP. These theorems are also useful in delineating the degree of terms that typically arise from the application of the energy method.

From Theorem 4.2 through Theorem 4.6, it can be seen that the individual components of the discrete analogue of IBP are higher-order approximations to their continuous counterparts; that is,

$$\overbrace{\mathbf{v}^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathbf{u}}^{\approx \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x} = \overbrace{\mathbf{v}^{\mathrm{T}}\mathsf{E}\mathbf{u}}^{\approx \mathcal{V}\mathcal{U}|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}} \overbrace{-\mathbf{u}^{\mathrm{T}}\mathsf{H}\mathsf{D}_1\mathbf{v}}^{\approx -\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{U}\frac{\partial \mathcal{V}}{\partial x}\mathrm{d}x}. \tag{4.40}$$

Equation (4.40) succinctly demonstrates that our particular choice for defining the SBP property leads to operators that not only mimic IBP but also lead to discrete approximations of the individual terms of IBP that are of order related to the order of the first-derivative operator. In contrast, alternative definitions of SBP, for example Refs. 14 and 19, while mimicking IBP, do not have such properties.

4.3 Dense-norm GSBP operators

The same type of theory that we have developed for diagonal-norm GSBP operators in the previous section can be developed for dense-norm GSBP operators, meaning here that the norm is not diagonal. In this section, rather than go through the details, which are presented in Appendix A, we summarize the main theorems for such operators.

Similar to diagonal-norm GSBP operators, the constituent matrices of the dense-norm GSBP operators are discrete approximations to various bilinear forms. To begin, the norm matrix is an approximation to the L_2 inner product as given in the following theorem:

Theorem 4.8. *A dense-norm GSBP operator, $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$, of degree p, has a norm H that is a degree $\tau_{\mathsf{H}} \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$ approximation to the L_2 inner product*

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\mathcal{U}\mathrm{d}x. \tag{4.41}$$

Proof. See Appendix A. □

Theorem 4.8 leads to the following corollary:

Corollary 2. *The norm* H *of a dense-norm GSBP operator is associated with a degree* $\tau \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$ *quadrature rule.*

Proof. See Appendix A. □

As with diagonal-norm GSBP operators, Q can be characterized as follows:

Theorem 4.9. *A dense-norm GSBP operator,* $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$*, of degree* p*, has a* Q *that is a degree* $\tau_{\mathsf{Q}} \geq \min\left(\tau_{\mathsf{E}}, 2\left\lfloor \frac{p-1}{2} \right\rfloor + 2\right)$ *approximation to the bilinear form*

$$(\mathcal{V},\mathcal{U}) = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x. \tag{4.42}$$

Proof. See Appendix A. □

Finally, $\mathsf{Q}^{(\mathrm{A})}$ is characterized in the following corollary:

Corollary 3. *A dense-norm GSBP operator,* $\mathsf{D}_1 = \mathsf{H}^{-1}\left(\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}\right)$*, of degree* p*, and* E *of degree* $\tau_{\mathsf{E}} \geq p$*, has a* $\mathsf{Q}^{(\mathrm{A})}$ *that is a degree* $\tau_{\mathsf{Q}^{(\mathrm{A})}} \geq \min\left(\tau_{\mathsf{E}}, 2\left\lfloor \frac{p-1}{2} \right\rfloor + 2\right)$ *approximation to the bilinear form*

$$(\mathcal{V},\mathcal{U}) = \int_{\Omega} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}\Omega - \frac{1}{2}\oint \Omega \mathcal{V}\mathcal{U}n_x \mathrm{d}s, \tag{4.43}$$

where $\mathbf{n} = [n_x, n_y]^{\mathrm{T}}$ *is the unit normal to the surface and* Ω *is the volume under consideration, with surface* $\partial\Omega$*. In one dimension, (4.43) reduces to*

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x - \frac{1}{2}\,\mathcal{V}\mathcal{U}|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}\,. \tag{4.44}$$

Proof. See Appendix A. □

We have the following result for the existence of dense-norm GSBP operators:

Theorem 4.10. *Given a nodal distribution,* $\mathbf{x}$*, then there exist dense-norm GSBP operators with degree* $p \in [0, n-1]$ *with a dense-norm* H *that is an approximation to the* L_2 *inner product of degree* $\tau_{\mathsf{H}} \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$*.*

In addition to the above theorems, we prove that we can always construct dense-norm GSBP operators of a certain degree starting from a known quadrature rule, even if the weights are negative (Theorem A.6 in Appendix A). These theorems show that there is

a rich array of possible operators that exist for each nodal distribution and that one has great flexibility in designing operators on a given nodal distribution. The downside of this flexibility is that it can be difficult to solve the nonlinear equations necessary to ensure that H is positive definite. Alternatively, one can simply construct dense-norm GSBP operators of degree $n-1$, which is the topic of the next section.

4.4 GSBP operators for the first derivative of degree $n-1$

In this section, we examine how to construct GSBP operators of degree $n-1$, and the relationship between these operators and the set of operators associated with Lagrangian basis functions. Carpenter and Gottlieb [13] were the first to realize that by dropping the requirement for an interior operator, the definition of classical FD-SBP operators encompassed a wider set of operators. They showed that the set of operators which results from application of the Galerkin approach, for Lagrange basis functions on nodal distributions that contain the boundary nodes, leads to operators that have the SBP property. To motivate their discovery, we examine the linear convection equation with unit wave speed:

$$\frac{\partial \mathcal{U}}{\partial t} + \frac{\partial \mathcal{U}}{\partial x} = 0. \tag{4.45}$$

Consider approximating $\mathcal{U}$ by the interpolant through the nodal distribution $\mathbf{x}$, given as $\mathcal{U} \approx \sum_{i=0}^{n-1} u_i(t) L_i(x)$, where $u_i(t) = \mathcal{U}(x_i, t)$, and L_i is the i^{th} Lagrangian basis function, given as

$$L_i(x) = \prod_{\substack{1 \le m \le n \\ m \neq i}} \frac{x - x_m}{x_{i+1} - x_m}. \tag{4.46}$$

Inserting the expansion into (4.45) gives

$$R(x,t) = \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=0}^{n-1} u_i(t) L_i(x) + \sum_{i=0}^{n-1} u_i(t) L_i'(x), \tag{4.47}$$

where $L_i'(x) = \frac{\partial L_i(x)}{\partial x}$. In the Galerkin approach, the residual is forced to be orthogonal to the space spanned by the basis functions used in the expansion of the solution, relative to a weighted L_2 inner product [42]. Taking the weight to be unity means that we require that [42]

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \left(\frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=0}^{n-1} u_i(t) L_i(x) + \sum_{i=0}^{n-1} u_i(t) L_i'(x) \right) L_j(x) \mathrm{d}x = 0, \ i, \ j \in [0, n-1]. \tag{4.48}$$

Equation (4.48) can be recast in matrix form as [42]

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathsf{H}\mathbf{u} + \mathsf{Q}\mathbf{u} = 0. \tag{4.49}$$

Thus,

$$\mathsf{H}(i+1, j+1) = \int_{x_\mathrm{L}}^{x_\mathrm{R}} L_j L_i \mathrm{d}x,\ i,\ j \in [0, n-1], \tag{4.50}$$

and

$$\mathsf{Q}(i+1, j+1) = \int_{x_\mathrm{L}}^{x_\mathrm{R}} L_j L_i' \mathrm{d}x,\ i,\ j \in [0, n-1]. \tag{4.51}$$

If H is invertible, we can see from (4.49) that $\mathsf{H}^{-1}\mathsf{Q}$ is necessarily a matrix operator approximating the first derivative. Carpenter and Gottlieb [13] showed that the matrix operators that result from the application of the Galerkin procedure can be used to construct derivative operators with the SBP property. In particular, they proved that H as defined by (4.50) is symmetric positive definite and that Q has the property that $\mathsf{Q} + \mathsf{Q}^\mathrm{T} = \mathsf{diag}(-1, 0, \ldots, 0, 1)$.

We see that these operators represent a subset of GSBP operators. However, we would like to prove that this same Galerkin procedure results in GSBP operators on nodal distributions that exclude one or both boundary nodes. What we need to show is that the E that results satisfies Definition 4.1. Consider

$$\begin{aligned} \left(\mathsf{Q} + \mathsf{Q}^\mathrm{T}\right)_{i,j} &= \int_{x_\mathrm{L}}^{x_\mathrm{R}} L_j L_i' + L_i L_j' \mathrm{dx} \\ &= \int_{x_\mathrm{L}}^{x_\mathrm{R}} \tfrac{\partial L_j L_i}{\partial x} = L_j L_i|_{x_\mathrm{L}}^{x_\mathrm{R}}. \end{aligned} \tag{4.52}$$

Thus, we have

$$\mathsf{E}(i+1, j+1) = L_j L_i|_{x_\mathrm{L}}^{x_\mathrm{R}},\ i,\ j \in [0, n-1]. \tag{4.53}$$

The matrix version of (4.53) is

$$\mathsf{E} = \mathbf{t}_{x_\mathrm{R}} \mathbf{t}_{x_\mathrm{R}}^\mathrm{T} - \mathbf{t}_{x_\mathrm{L}} \mathbf{t}_{x_\mathrm{L}}^\mathrm{T}, \tag{4.54}$$

where $\mathbf{t}_{x_\mathrm{R}} = [L_0(x_\mathrm{R}), \ldots, L_{n-1}(x_\mathrm{R})]$ and $\mathbf{t}_{x_\mathrm{L}} = [L_0(x_\mathrm{L}), \ldots, L_{n-1}(x_\mathrm{L})]$. The resultant $\mathbf{t}_{x_\mathrm{R}}$ and $\mathbf{t}_{x_\mathrm{L}}$ have the required properties since the Lagrangian interpolant interpolates polynomials up to degree $n-1$ exactly. We have therefore shown that the Galerkin procedure using Lagrangian basis functions always leads to operators with the SBP property, as defined in the GSBP framework.

The above set of GSBP operators need not be derived by utilizing the Galerkin procedure and can instead be constructed directly (for more details see Appendix A). The norm matrix

can be constructed as

$$\mathsf{H} = \left(\mathsf{X}^{-1}\right)^{\mathrm{T}} \mathsf{V}\mathsf{X}^{-1}, \tag{4.55}$$

where

$$\mathsf{V}(i+1, j+1) = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^{i+j} \mathrm{d}x. \tag{4.56}$$

To construct Q, note that from Lemma 1,

$$\mathsf{D}_1 = \tilde{\mathsf{X}}_{\mathsf{D}} \mathsf{X}^{-1}; \tag{4.57}$$

therefore,

$$\mathsf{Q} = \mathsf{H}\tilde{\mathsf{X}}_{\mathsf{D}} \mathsf{X}^{-1}. \tag{4.58}$$

Furthermore,

$$\mathsf{E} = \left(\mathsf{X}^{\mathrm{T}}\right)^{-1} \tilde{\mathsf{E}} \mathsf{X}^{-1}, \tag{4.59}$$

where

$$\tilde{\mathsf{E}}(i+1, j+1) = x_{\mathrm{R}}^{i+j} - x_{\mathrm{L}}^{i+j}, \ i, \ j \in [0, n-1], \tag{4.60}$$

and by definition we get

$$\mathsf{Q}^{(\mathrm{A})} = \mathsf{Q} - \frac{1}{2}\mathsf{E}. \tag{4.61}$$

We note here that both the Galerkin procedure and the direct construction procedure typically results in a dense norm matrix. However, the norm matrix is not necessarily unique, though the operator is, and can sometimes be replaced with a diagonal matrix, for example see Gassner [37].

4.5 A multi-dimensional perspective

We have seen how extending the definition of E allows for the derivation of SBP operators on more general nodal distributions. Moreover, we have shown that this particular choice of E has a number of consequences on the properties of the constituent matrices of a GSBP operator. In this section, we take a multi-dimensional perspective and show that our choice of definition of E can be interpreted as a bilinear form approximating a surface integral over the surface of the element or block associated with the derivative operator. However, the choice we make is not the only possibility, and we discuss more general definitions of SBP operators. For simplicity, the presentation is limited to two dimensions. However the extension to three dimensions is straightforward.

The multi-dimensional form of IBP over a volume Ω enclosed by a surface $\partial\Omega$ for the x

derivative is given as

$$\int_{\Omega} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}\Omega = \oint_{\partial\Omega} \mathcal{V}\mathcal{U}n_x \mathrm{d}s - \int_{\Omega} \mathcal{U}\frac{\partial \mathcal{V}}{\partial x}\mathrm{d}\Omega, \tag{4.62}$$

where the unit normal to the surface is $\mathbf{n} = [n_x, n_y]^{\mathrm{T}}$. The discrete representation of (4.62) is given as

$$\mathbf{v}^{\mathrm{T}}\mathsf{H}\mathsf{D}_x\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathsf{E}_x\mathbf{u} - \mathbf{u}^{\mathrm{T}}\mathsf{H}\mathsf{D}_x\mathbf{v},\ \forall\ \mathbf{v},\mathbf{u} \in \mathbb{R}^n, \tag{4.63}$$

where $\mathbf{u}$ and $\mathbf{v}$ are the projections of $\mathcal{U}$ and $\mathcal{V}$ onto a set of n nodes, $S = \{(x_{i,}, y_i)|i \in [1, n]\}$, and H is symmetric positive definite. Furthermore, $\mathsf{D}_x \in \mathbb{R}^{n\times n}$ is a matrix operator approximating the x derivative. Analogous relations can be constructed for the operator D_y approximating the y derivative. Note that both operators share the same norm matrix H.

Not all matrices H, D_x, and E_x satisfy (4.63) and the symmetric positive definite requirement on H. The following lemma summarizes the basic requirements:

Lemma 2. *A trio of matrices* H, D_x, *and* E_x *satisfy the SBP property, (4.63), if* H *is symmetric positive definite,* $\mathsf{D}_x = \mathsf{H}^{-1}\left(\mathsf{Q}_x^{(A)} + \frac{1}{2}\mathsf{E}_x\right)$, *where* $\mathsf{Q}_x^{(\mathrm{A})}$ *is antisymmetric, and* E_x *is a symmetric matrix.*

Proof. Defining $\mathsf{Q}_x = \mathsf{H}\mathsf{D}_x$, (4.63) becomes

$$\mathbf{v}^{\mathrm{T}}\mathsf{Q}_x\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathsf{E}_x\mathbf{u} - \mathbf{u}^{\mathrm{T}}\mathsf{Q}_x\mathbf{v},\ \forall\ \mathbf{v},\mathbf{u} \in \mathbb{R}^n. \tag{4.64}$$

Since (4.64) is a scalar equation, $\mathbf{u}^{\mathrm{T}}\mathsf{Q}_x\mathbf{v} = (\mathbf{u}^{\mathrm{T}}\mathsf{Q}_x\mathbf{v})^{\mathrm{T}} = \mathbf{v}^{\mathrm{T}}\mathsf{Q}_x^{\mathrm{T}}\mathbf{u}$. Using this fact, (4.64) becomes

$$\mathbf{v}^{\mathrm{T}}\left(\mathsf{Q}_x + \mathsf{Q}_x^{\mathrm{T}}\right)\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathsf{E}_x\mathbf{u},\ \forall\ \mathbf{v},\mathbf{u} \in \mathbb{R}^n. \tag{4.65}$$

Decomposing $\mathsf{Q}_x = \mathsf{Q}_x^{(\mathrm{A})} + \mathsf{Q}_x^{(\mathrm{S})}$, where $\mathsf{Q}_x^{(\mathrm{A})}$ is antisymmetric and $\mathsf{Q}_x^{(\mathrm{S})}$ is symmetric, and inserting this decomposition into (4.65) results in

$$2\mathbf{v}^{\mathrm{T}}\mathsf{Q}_x^{(\mathrm{S})}\mathbf{u} = \mathbf{v}^{\mathrm{T}}\mathsf{E}_x\mathbf{u},\ \forall\ \mathbf{v},\mathbf{u} \in \mathbb{R}^n. \tag{4.66}$$

Thus $\mathsf{Q}_x^{(\mathrm{S})} = \frac{1}{2}\mathsf{E}_x$, leading to the conclusion that E_x must be symmetric and

$$\mathsf{Q}_x = \mathsf{Q}_x^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}_x. \tag{4.67}$$

Finally, since H is invertible, we have that

$$\mathsf{D}_x = \mathsf{H}^{-1}\left(\mathsf{Q}_x^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}_x\right), \tag{4.68}$$

which proves the lemma. □

Lemma 2 applies to the trio of matrices H, D_y, and E_y by changing x to y. Before

proceeding, it is necessary to specify how to construct the conditions on D_x and D_y so that they approximate the first derivative. These matrix operators are to be exact for polynomials of up to degree p on the set of nodes $S = \{(x_i,, y_i) | i \in [1, n]\}$, which means that they need to be exact for all combinations $x^{a_1} y^{a_2}$, such that $a_1 + a_2 \leq p$. The required combinations, in two dimensions, can be deduced from Pascal's triangle for monomials [8]:

$$\begin{array}{cccc} & 1 & & p = 0 \\ & x \quad y & & p = 1 \\ & x^2 \quad xy \quad y^2 & & p = 2 \\ & x^3 \quad x^2 y \quad xy^2 \quad y^3 & & p = 3 \\ x^4 & x^3 y \quad x^2 y^2 \quad xy^3 & y^4 & p = 4 \\ & \vdots & & \vdots \end{array}$$

where for each p it is necessary to satisfy the combinations for the preceding values of p. Thus, an operator D_x of degree p must satisfy the following degree equations:

$$\mathsf{D}_x \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = a_1 \mathbf{x}^{a_1 - 1} \odot \mathbf{y}^{a_2}, \forall\ a_1 + a_2 \leq p, \tag{4.69}$$

where $\odot$ is the Hadamard product, i.e., element-wise multiplication, $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$, and $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$. Similarly, for D_y of degree p, we have the following degree equations:

$$\mathsf{D}_y \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = a_2 \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2 - 1}, \forall\ a_1 + a_2 \leq p. \tag{4.70}$$

In two dimensions, the minimum number of required nodes is

$$n = \frac{(p+1)(p+2)}{2}, \tag{4.71}$$

and more generally speaking, for m dimensions it is necessary to have at least $\binom{p+m}{m}$ nodes.

From the foregoing discussion, necessary but not sufficient conditions for the construction of a multi-dimensional SBP operator can be delineated as follows: Consider an enclosed volume $\Omega \subset \mathbb{R}^2$ and associated surface $\partial\Omega \subset \mathbb{R}$ and set of points $S = \{(x_i, y_i) | i \in [1, n]\}$ defined by the vectors $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$ and $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$, then the matrices D_x and D_y are approximations to the first derivatives $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, respectively, of degree p with the SBP property if

i) $\mathsf{D}_x \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = \mathsf{H}^{-1} \mathsf{Q}_x \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = \mathsf{H}^{-1} \left(\mathsf{Q}_x^{(\mathrm{A})} + \frac{1}{2} \mathsf{E}_x \right) \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = a_1 \mathbf{x}^{a_1 - 1} \odot \mathbf{y}^{a_2}$,

$\forall a_1 + a_2 \leq p$, where $\mathsf{Q}_x^{(\mathrm{A})} = \frac{\mathsf{Q}_x - \mathsf{Q}_x^{\mathrm{T}}}{2}$, and E_x is symmetric;

ii) $\mathsf{D}_y \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = \mathsf{H}^{-1} \mathsf{Q}_y \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = \mathsf{H}^{-1} \left(\mathsf{Q}_y^{(\mathrm{A})} + \frac{1}{2} \mathsf{E}_y \right) \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = a_2 \mathbf{x}^{a_1} \odot \mathbf{y}^{a_2 - 1}$,

$\forall\ a_1 + a_2 \leq p$, where $\mathsf{Q}_y^{(\mathrm{A})} = \frac{\mathsf{Q}_y - \mathsf{Q}_y^{\mathrm{T}}}{2}$, and E_y is symmetric; and

iii) H is symmetric positive-definite.

The above conditions do not impose any restriction on the relation between the nodal distribution and the volume under consideration. That is, nodes can be contained within and outside of the volume, and nodes need not lie on the surface. These conditions are very general, and it is necessary to determine additional constraints so that energy estimates can be constructed. For example, in one dimension, one can proceed by defining $\mathsf{E} = \mathsf{E}_{x_{\mathrm{R}}} - \mathsf{E}_{x_{\mathrm{L}}}$ such that $\mathsf{E}_{x_{\mathrm{R}}}$ and $\mathsf{E}_{x_{\mathrm{L}}}$ are symmetric positive semi-definite. Using this restriction, it is possible to construct SATs that lead to stable discretizations, which is the approach taken in Refs. 14, 19, 1, 2 for one-dimensional FD operators on equispaced nodal distributions; also see the work in Refs. 81 and 80 on overlapping SBP operators. This approach presents an interesting possibility for multi-dimensional GSBP operators worth further consideration.

Alternatively, we look to construct SBP operators such that the individual components of (4.63) and the analogue for the y derivative are higher-order approximations to the continuous analogues, much in the same way as classical FD-SBP operators originally proposed by Kreiss and Scherer [56] and the GSBP operators presented in this chapter. Here, we constrain the definition of E_x such that $\mathbf{v}^{\mathrm{T}}\mathsf{E}_x\mathbf{u}$ is an approximation of the surface integral in (4.62), and similarly for E_y. These ideas lead to the following definition:

Definition 5. Two-dimensional generalized summation-by-parts operators: Consider an enclosed volume $\Omega \subset \mathbb{R}^2$ and associated surface $\partial\Omega \subset \mathbb{R}$ and set of points $S = \{(x_i, y_i) | i \in [1, n]\}$ defined by the vectors $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$ and $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$, then the matrices D_x and D_y are degree p approximations to the first derivatives $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$, respectively, with the SBP property if

i) $\mathsf{D}_x\mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = \mathsf{H}^{-1}\left(\mathsf{Q}_x^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}_x\right)\mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = a_1\mathbf{x}^{a_1-1} \odot \mathbf{y}^{a_2},\ \forall a_1 + a_2 \leq p;$

ii) $\mathsf{D}_y\mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = \mathsf{H}^{-1}\left(\mathsf{Q}_y^{(\mathrm{A})} + \frac{1}{2}\mathsf{E}_y\right)\mathbf{x}^{a_1} \odot \mathbf{y}^{a_2} = a_2\mathbf{x}^{a_1} \odot \mathbf{y}^{a_2-1}, \forall\ a_1 + a_2 \leq p;$

iii) H is symmetric positive definite;

iv)
$$\left(\mathbf{x}^{a_1} \odot \mathbf{y}^{a_2}\right)^{\mathrm{T}} \mathsf{E}_x\mathbf{x}^{b_1} \odot \mathbf{y}^{b_2} = \oint_{\partial\Omega} x^{a_1+b_1}y^{a_2+b_2}n_x \mathrm{d}s\ \forall\ a_1 + a_2, b_1 + b_2 \leq \tau_{\mathsf{E}},$$

$$\left(\mathbf{x}^{a_1} \odot \mathbf{y}^{a_2}\right)^{\mathrm{T}} \mathsf{E}_y\mathbf{x}^{b_1} \odot \mathbf{y}^{b_2} = \oint_{\partial\Omega} x^{a_1+b_1}y^{a_2+b_2}n_y \mathrm{d}s\ \forall\ a_1 + a_2, b_1 + b_2 \leq \tau_{\mathsf{E}}$$

where $\tau_{\mathsf{E}} \geq p$ and $\mathbf{n} = [n_x, n_y]^{\mathrm{T}}$ is the outward pointing unit normal to the surface $\partial\Omega$; and

v) $\mathsf{Q}_x^{(\mathrm{A})}$ and $\mathsf{Q}_y^{(\mathrm{A})}$ are antisymmetric, and E_x and E_y are symmetric.

Definition 5 implies that E_x and E_y are degree τ_{E} approximations to the surface integrals

$$\oint_{\partial\Omega} \mathcal{V}\mathcal{U}n_x \mathrm{d}s, \tag{4.72}$$

and

$$\oint_{\partial\Omega} \mathcal{V}\mathcal{U}n_y \mathrm{d}s, \tag{4.73}$$

respectively.

The multi-dimensional viewpoint makes explicit the implicit assumption in both the classical FD-SBP theory as well as the GSBP theory that E is an approximation to the surface integral arising from IBP. Moreover, from this viewpoint, the result that H is an approximation to the L_2 inner product $(\mathcal{V},\mathcal{U}) = \int_{\Omega} \mathcal{V}\mathcal{U}\mathrm{d}\Omega$ seems a natural consequence of assuming that E is an approximation to a surface integral.

4.6 Summary

In this chapter, we proposed an extension to the definition of classical FD-SBP operators given by Kreiss and Scherer [56] which allows the construction of a broader array of operators on nodal distributions that are nonuniform and may not include nodes on the boundaries. The consequence of this definition is that the constituent matrices of GSBP operators retain the properties associated with classical FD-SBP operators, for example, norm matrices that are approximations to the L_2 inner product. We developed the theory of diagonal-norm GSBP operators; this theory helps guide the construction of such operators, as the search for diagonal-norm GSBP operators can be reduced to the search for quadrature rules with positive weights. The theory of dense-norm GSBP operators was also developed. We also discussed how the Galerkin procedure with Lagrangian basis functions leads to GSBP operators of degree $n-1$ and how to construct such operators directly. Finally, we presented an extension of the SBP concept to multi-dimensional operators. We showed that from this vantage point, the choice of E in our definition of GSBP operators is an approximation to the surface integral that arises in IBP. In the next chapter, we extend the GSBP theory for the first derivative to the second derivative with a variable coefficient.

Chapter 5

Generalized Summation-by-Parts Operators for the Second Derivative

5.1 Introduction

In this chapter, the definition of classical FD-SBP operators approximating the second derivative, given by Refs. 65, 67, and 62, is extended to accommodate the derivation of GSBP operators. The form we propose combines ideas from Refs. 65, 67, and 62, as well as our extension of the ideas of Kamakoti and Pantano [51] on the interior operator of classical FD-SBP operators (see Section 5.3).

As discussed in Section 3.3, the motivation for the form of the operators comes in part from the integration by parts property of the second derivative with variable coefficients. For example, consider the variable coefficient heat equation

$$\frac{\partial \mathcal{U}}{\partial t} = \frac{\partial}{\partial x}\left(\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\right). \tag{5.1}$$

Applying the energy method to (5.1), i.e., multiplying by the solution, integrating in space, and using integration by parts, results in

$$\frac{\mathrm{d}\|\mathcal{U}\|^2}{\mathrm{d}t} = 2\,\mathcal{B}\mathcal{U}\frac{\partial \mathcal{U}}{\partial x}\bigg|_{x_\mathrm{L}}^{x_\mathrm{R}} - 2\int_{x_\mathrm{L}}^{x_\mathrm{R}} \frac{\partial \mathcal{U}}{\partial x}\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x. \tag{5.2}$$

It was shown in Section 3.5 that applying a first-derivative classical FD-SBP operator twice mimics integrations by parts. This was then used to show that, with appropriate SATs, energy estimates can be constructed for the semi-discrete form of the linear convection-diffusion equation. Similarly, we show here that the application of the first-derivative operator twice for both classical FD-SBP and GSBP operator leads to energy estimates mimetic of (5.2). To see this, we first note that the application of the first-derivative operator twice can be

decomposed as

$$\mathsf{D}_1^{(p)}\mathsf{B}\mathsf{D}_1^{(p)} = \mathsf{H}^{-1}\left[-\left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1^{(p)} + \mathsf{E}\mathsf{B}\mathsf{D}_1^{(p)}\right]. \tag{5.3}$$

The semi-discrete version of (5.1), using (5.3), is

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = \mathsf{D}_1^{(p)}\mathsf{B}\mathsf{D}_1^{(p)}\mathbf{u}_h = \mathsf{H}^{-1}\left[-\left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1^{(p)} + \mathsf{E}\mathsf{B}\mathsf{D}_1^{(p)}\right]\mathbf{u}_h. \tag{5.4}$$

Applying the energy method to (5.4) gives

$$\frac{\mathrm{d}\|\mathbf{u}\|_{\mathsf{H}}^2}{\mathrm{d}t} = \overbrace{2\mathbf{u}_h^{\mathrm{T}}\mathsf{E}\mathsf{B}\mathsf{D}_1^{(p)}\mathbf{u}_h}^{\approx 2\,\mathcal{B}\mathcal{U}\frac{\partial\mathcal{U}}{\partial x}\big|_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}} - \overbrace{2\left(\mathsf{D}_1^{(p)}\mathbf{u}_h\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1^{(p)}\mathbf{u}_h}^{\approx 2\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}}\frac{\partial\mathcal{U}}{\partial x}\mathcal{B}\frac{\partial\mathcal{U}}{\partial x}\mathrm{d}x}, \tag{5.5}$$

and we see that the application of the first-derivative operator twice is mimetic of (5.2). Our goal is to retain the ability to construct energy estimates in a similar fashion as the application of the first-derivative operator twice, but with operators that are more accurate.

The equations that an operator must satisfy in order to approximate the second derivative with a variable coefficient, denoted the degree equations, are based on monomials restricted onto the nodes of the grid. Given that the operator must approximate $\frac{\partial}{\partial x}\left(\mathcal{B}\frac{\partial\mathcal{U}}{\partial x}\right)$, it is necessary to determine what degree monomial to insert for $\mathcal{B}$ and $\mathcal{U}$ in constructing the degree equations. Taking $\mathcal{B} = x^k$ and $\mathcal{U} = x^s$ and inserting into the second derivative gives

$$\frac{\partial}{\partial x}\left(x^k\frac{\partial x^s}{\partial x}\right) = s(k+s-1)x^{k+s-2}. \tag{5.6}$$

To be of order p, second-derivative operators must be of degree $p+1$. This implies that all combinations of $k+s \leq p+1$ must be satisfied. Thus, the degree equations have the following form:

$$\mathsf{D}_2^{(p)}\left(\mathsf{diag}\left(\mathbf{x}^k\right)\right)\mathbf{x}^s = s(k+s-1)\mathbf{x}^{k+s-2},\ k+s \leq p+1, \tag{5.7}$$

where $\mathsf{diag}\left(\mathbf{x}^k\right)$ is a diagonal matrix such that the i^{th} diagonal entry is the i^{th} entry of $\mathbf{x}^k$. If there are n nodes in the nodal distribution, then each combination of $k+s$ in (5.7) returns a vector of n equations.

The maximum attainable order and degree of an operator approximating the second derivative are stated in the following lemma:

Lemma 3. *An operator $\mathsf{D}_2 \in \mathbb{R}^{n\times n}$ approximating the second derivative is at most of order $p = n-2$ and degree $n-1$. Furthermore, such an operator of maximum degree exists and is unique.*

Proof. Consider the degree equations for a constant-coefficient second-derivative operator:

$$\mathsf{D}_2\mathbf{x}^k = k\,(k-1)\,\mathbf{x}^{k-2},\ k \in [0, p+1]. \tag{5.8}$$

Taking $p = n-2$, the degree equations can be recast as

$$\mathsf{D}_2\mathsf{X} = \tilde{\mathsf{X}}_{\mathsf{D}_2}. \tag{5.9}$$

Since the Vandermonde matrix $\tilde{\mathsf{X}}$ is invertible, a unique solution exists, given as $\mathsf{D}_2 = \tilde{\mathsf{X}}_{\mathsf{D}_2}\mathsf{X}^{-1}$, and by examining the range of the operator D_2, i.e., $\tilde{\mathsf{X}}_{\mathsf{D}_2}$, it is clear that D_2 is at most of degree $n-1$ and hence order $p = n-2$. □

An immediate consequence of Lemma 3 is the following corollary:

Corollary 4. *An operator $\mathsf{D}_2(\mathsf{B}) \in \mathbb{R}^{n\times n}$ approximating the second derivative with a variable coefficient is at most of order $p = n-2$ and degree $n-1$. Furthermore, such an operator always exists.*

Proof. The set of equations for the constant-coefficient case is a subset of the equations for the variable-coefficient operator, and therefore, by Lemma 3, $\mathsf{D}_2^{(p)}(\mathsf{B})$ is at best of order $n-2$ and degree $n-1$. To show that an operator that satisfies (5.7) always exists, consider constructing the operator as

$$\mathsf{D}_2(\mathsf{B}) = \sum_{i=1}^{n} \mathsf{B}(i,i)\mathsf{F}_i. \tag{5.10}$$

Rather than the degree equations, consider

$$\mathsf{D}_2\left(\mathsf{diag}\left(\mathbf{x}^k\right)\right)\mathbf{x}^s = s(k+s-1)\mathbf{x}^{k+s-2},\ k, s \in [0, n-1], \tag{5.11}$$

which contain the degree equations (5.7) as a subset. The equations (5.11) can be compactly written as

$$\mathsf{X}^{\mathrm{T}} \otimes \mathsf{I}_n \overbrace{\begin{bmatrix} \mathsf{F}_1 \\ \vdots \\ \mathsf{F}_n \end{bmatrix}}^{\tilde{\mathsf{F}}} = \overbrace{\begin{bmatrix} \tilde{\mathsf{X}}^{(0)}_{\mathsf{D}_2(\mathsf{B})}\mathsf{X}^{-1} \\ \vdots \\ \tilde{\mathsf{X}}^{(n-1)}_{\mathsf{D}_2(\mathsf{B})}\mathsf{X}^{-1} \end{bmatrix}}^{\mathsf{G}}, \tag{5.12}$$

where $\otimes$ is the Kronecker product (for more information see Ref. 83), I_n is an $n \times n$ identity matrix, and

$$\tilde{\mathsf{X}}^{(k)}_{\mathsf{D}_2(\mathsf{B})}(:, s+1) = s(k+s-1)\mathbf{x}^{k+s-2},\ s \in [0, n-1]. \tag{5.13}$$

The Kronecker product of two invertible matrices is invertible [83], therefore

$$\tilde{\mathsf{F}} = \left(\mathsf{X}^{\mathrm{T}} \otimes \mathsf{I}_n\right)^{-1}\mathsf{G}.$$

We have shown that a unique operator that satisfies (5.11) always exists. □

Even though the operator in Corollary 4 is unique, the application of the first-derivative operator twice of degree $n-1$ also satisfies the degree equations (5.7) for $k+s \leq n-1$, as we shall soon prove. This implies that an operator satisfying the degree equations is not unique.

5.2 GSBP operators for the second derivative

For classical FD-SBP operators, the drawback to the application of the first-derivative operator twice is that the interior operator uses nearly twice as many nodes and is less dissipative of under-resolved modes as compared to minimum-stencil operators. For GSBP operators that can only be applied using an element approach, the concept of minimum stencil does not exist. Regardless, the application of the first-derivative operator twice results in an approximation that is of lower order than the first-derivative operator. Thus, in general, we search for GSBP operators approximating the second derivative that match the order of the first-derivative operator; such operators are denoted order matched. These ideas lead to the following definition:

Definition 6. Order-matched second-derivative GSBP operator: The operator $\mathsf{D}_2^{(p)}(\mathsf{B}) \in \mathbb{R}^{n\times n}$ is a GSBP operator approximating the second derivative, $\frac{\partial}{\partial x}\left(\mathcal{B}\frac{\partial \mathcal{U}}{\partial x}\right)$, of degree $p+1$ and order p that is order matched to the GSBP operator $\mathsf{D}_1^{(p)} = \mathsf{H}^{-1}\mathsf{Q}$, on a nodal distribution $\mathbf{x}$, if if satisfies the equations

$$\mathsf{D}_2^{(p)}\left(\mathsf{diag}\left(\mathbf{x}^k\right)\right)\mathbf{x}^s = s(k+s-1)\mathbf{x}^{k+s-2},\ k+s \leq p+1, \tag{5.14}$$

and is of the form

$$\mathsf{D}_2^{(p)}(\mathsf{B}) = \mathsf{H}^{-1}\left\{-\mathsf{M}\left(\mathsf{B}\right) + \mathsf{E}\mathsf{B}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\right\}, \tag{5.15}$$

where

$$\mathsf{M}(\mathsf{B}) = \sum_{i=1}^{n} \mathsf{B}(i,i)\mathsf{M}_i. \tag{5.16}$$

The matrices M_i, B, and $\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}$ are $\in \mathbb{R}^{n\times n}$, M_i is symmetric positive semi-definite,

$$\mathsf{B} = \mathsf{diag}(\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n)),$$

and $\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}$ is an approximation to the first derivative of degree and order $\geq p+1$.

If one takes B to be the identity matrix, then Definition 6 collapses onto that given by Mattsson and Nordström [65] for classical FD-SBP operators—defining the relevant matrix in their definition as the sum of the M_i—where we do not specify further restrictions on

the form of the M_i in order to allow for GSBP operators. The extension to variable coefficients, by taking the sum of matrices multiplied by the variable coefficient, is an extension and simplification of the work by Kamakoti and Pantano [51], who decompose the interior operator of FD approximations to the second derivative with a variable coefficient as the sum of the variable coefficient multiplying a third-order tensor. Definition 6 can be applied to dense-norm GSBP operators, though we do not pursue this further in this thesis.

Definition 6 leads to energy estimates similar to the application of the first-derivative operator twice. Consider discretizing (5.1) using a GSBP operator, as given in Definition 6, which results in the following semi-discrete equations:

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = \mathsf{H}^{-1}\left[-\mathsf{M}\left(\mathsf{B}\right) + \mathsf{E}\mathsf{B}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\right]. \tag{5.17}$$

Applying the energy method to (5.17) results in

$$\frac{\mathrm{d}\|\mathbf{u}\|_{\mathsf{H}}^2}{\mathrm{d}t} = 2\mathbf{u}_h^{\mathrm{T}}\mathsf{E}\mathsf{B}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\mathbf{u}_h - 2\mathbf{u}_h^{\mathrm{T}}\mathsf{M}\left(\mathsf{B}\right)\mathbf{u}_h. \tag{5.18}$$

We see that Definition 6 leads to an estimate that is similar to the continuous case in that the estimate has the correct boundary terms and a negative semi-definite term, i.e. $-2\mathbf{u}_h^{\mathrm{T}}\mathsf{M}\left(\mathsf{B}\right)\mathbf{u}_h$. However, it is not fully mimetic of the continuous case in that the negative semi-definite term is not, in general, an approximation to the continuous counterpart, i.e., the volume integral in (5.2). Nevertheless, with appropriate SATs Definition 6 is sufficient to derive energy estimates for PDEs that do not contain cross-derivative terms. Without additional constraints, however, it does not guarantee that an energy estimate exists for PDEs with cross-derivative terms. One possibility is what is referred to as compatible operators [67]. With appropriate SATs, these operators are guaranteed to produce energy estimates for PDEs with cross-derivative terms. These ideas lead to the following definition for diagonal-norm compatible and order-matched operator:

Definition 7. Order-matched compatible second-derivative GSBP operator: A diagonal-norm order-matched GSBP operator, $\mathsf{D}_2^{(p)}(\mathsf{B}) \in \mathbb{R}^{n\times n}$, for the second derivative is compatible with the first-derivative GSBP operator, $\mathsf{D}_1^{(p)}$, if in addition to the requirements of Definition 6,

$$\mathsf{M}\left(\mathsf{B}\right) = \left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1^{(p)} - \mathsf{R}\left(\mathsf{B}\right), \tag{5.19}$$

where

$$\mathsf{R}\left(\mathsf{B}\right) = \sum_{i=1}^{n}\mathsf{B}(i,i)\mathsf{R}_i, \tag{5.20}$$

and the matrices $\mathsf{R}_i \in \mathbb{R}^{n\times n}$ are symmetric negative semi-definite.

The decomposition in (7) is inspired by the observation that the application of the first-

derivative classical FD-SBP and GSBP operators twice can be reformulated as (5.3). The idea of constructing FD-SBP approximations to the second derivative as the application of the first-derivative operator twice plus corrective terms was first proposed by Mattsson et al. [67] and later used by Mattsson [62] to construct classical FD-SBP operators to approximate the second derivative with a variable coefficient. We say corrective terms since not only has the term $\mathsf{H}^{-1}\mathsf{R}\left(\mathsf{B}\right)$ been added, but $\mathsf{EBD}_1^{(p)}$ has been replaced by $\mathsf{EBD}_{1,\mathrm{b}}^{(\geq p+1)}$, which can always be construed as adding a corrective term to $\mathsf{EBD}_1^{(p)}$. The definition of compatible operators is limited to diagonal-norm operators; for the variable-coefficient case, it is unclear how to derive energy estimates for dense-norm operators (see Mattsson and Almquist [63] for a discussion and potential solution).

Definition 7 also leads to energy estimates, and in this case the estimate mimics the continuous case. Discretizing (5.1) with a GSBP operator as given by Definition 7 results in the following semi-discrete equations:

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = \mathsf{H}^{-1}\left[-\left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{HBD}_1^{(p)} + \mathsf{R}\left(\mathsf{B}\right) + \mathsf{EBD}_{1,\mathrm{b}}^{(\geq p+1)}\right]. \tag{5.21}$$

Applying the energy method to (5.21) results in

$$\frac{\mathrm{d}\|\mathbf{u}\|_{\mathsf{H}}^2}{\mathrm{d}t} = 2\mathbf{u}_h^{\mathrm{T}}\mathsf{EBD}_{1,\mathrm{b}}^{(\geq p+1)}\mathbf{u}_h - 2\mathbf{u}_h^{\mathrm{T}}\left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{HBD}_1^{(p)}\mathbf{u}_h + 2\mathbf{u}_h^{\mathrm{T}}\mathsf{R}\left(\mathsf{B}\right)\mathbf{u}_h. \tag{5.22}$$

Now the RHS of (5.21) mimics the continuous energy estimate, where $2\mathbf{u}_h^{\mathrm{T}}\mathsf{R}\left(\mathsf{B}\right)\mathbf{u}_h$ adds a term of the order of the discretization error.

The compatibility that is necessary is between the first-derivative operators approximating the mixed derivatives and the second-derivative operator. In addition, an energy estimate is guaranteed to exist, with appropriate SATs, if the norms of all operators are the same. In practice, this means that all first-derivative terms are typically approximated using the same GSBP operator.

An order-matched and compatible $\mathsf{D}_2\left(\mathsf{B}\right)$ SBP operator, as given in Definition 7, is the application of the first-derivative operator twice plus corrective terms to increase the degree of the resultant operator. The application of the first-derivative operator twice already satisfies a number of the degree equations (5.7), and the corrective terms are added such that the remaining degree equations in order for the operator to be of order p are satisfied, while continuing to satisfy those degree equations satisfied by the application of the first-derivative operator twice. Applying an order p SBP operator for the first derivative twice results in

$$\begin{aligned}\mathsf{D}_1\mathsf{diag}\left(\mathbf{x}^k\right)\mathsf{D}_1\mathbf{x}^s = s\mathsf{D}_1\mathbf{x}^{s+k-1} \quad &= \quad s(s+k-1)\mathbf{x}^{s+k-2}, \\ &\qquad s+k \leq p+1,\ s \leq p.\end{aligned} \tag{5.23}$$

Equations (5.23) show that only the equations for $s = p+1,\ k = 0$ are not satisfied by the application of the first-derivative operator twice. We now use this observation to propose a construction of compatible and order-matched GSBP operators for the second derivative with a variable coefficient as the application of the first-derivative operator twice plus corrective terms modelled after the constant-coefficient operator, as described in the following theorem:

Theorem 5.1. *The existence of a diagonal-norm compatible and order-matched GSBP operator* $\mathsf{D}_2^{(p)}$ *for the constant-coefficient case of order* p *and degree* $p+1$ *is sufficient for the existence of a compatible and order-matched GSBP operator* $\mathsf{D}_2^{(p)}(\mathsf{B})$*, for* $p+1 \leq n-1$ *and* $n \geq 3$.

Proof. Consider constructing the operator as

$$\mathsf{H}^{-1}\left[-\left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1^{(p)} + \frac{\sum_{i=1}^{n}\mathsf{B}(i,i)}{n}\mathsf{R}_{\mathrm{c}} + \mathsf{E}\mathsf{B}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\right], \tag{5.24}$$

where R_{c} and $\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}$ are from the constant-coefficient operator. As has been argued, the additional equations that must be satisfied are for $(k,s) = (0,p+1)$. Since (5.24) collapses onto the constant-coefficient operator for this condition, it automatically satisfies these additional equations. What remains to be shown is that the remaining degree equations are still satisfied.

Now $\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)} = \mathsf{D}_1^{(p)} + \mathsf{K}$, where K is a corrective term such that $\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}$ is at least one order more accurate than the first-derivative operator. The application of the first-derivative operator twice can decomposed into

$$\mathsf{D}_1^{(p)}\mathsf{B}\mathsf{D}_1^{(p)} = \mathsf{H}^{-1}\left[-\left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1^{(p)} + \mathsf{E}\mathsf{B}\mathsf{D}_1^{(p)}\right]. \tag{5.25}$$

Therefore, (5.24) can be recast as

$$\mathsf{D}_1^{(p)}\mathsf{B}\mathsf{D}_1^{(p)} + \mathsf{H}^{-1}\left\{\frac{\sum_{i=1}^{n} b_i}{n}\mathsf{R}_{\mathrm{c}} + \mathsf{E}\mathsf{B}\mathsf{K}\right\}. \tag{5.26}$$

Examining the constant-coefficient version of (5.24), it can be seen that both K and R_{c} must be zero for $\mathbf{x}^s$ for $s \leq p$. Therefore, we have proven that (5.24) leads to a compatible and order-matched GSBP operator for the second derivative with a variable coefficient. □

The implication of Theorem 5.1 is that the search for compatible and order-matched GSBP operators for the variable-coefficient case reduces to the search for compatible and

order-matched GSBP operators for the constant-coefficient case. This substantially simplifies both the proof that compatible and order-matched GSBP operators exist for a given nodal distribution as well as their construction. The conditions to enforce $\mathsf{R}(\mathsf{B})$ to be positive semi-definite lead to n eigenvalue problems of size $n \times n$, the solution of which is a highly nontrivial task. In practice, one solves the constant-coefficient problem and if such operators exist Theorem 5.1 guarantees that compatible and order-matched GSBP operators exist for the variable-coefficient case. Moreover, Theorem 5.1 gives a simple means of constructing compatible and order-matched GSBP operators from the constant-coefficient compatible and order-matched GSBP operators.

5.3 GSBP operators with a repeating interior operator

The focus of this section is on compatible order-matched operators with a repeating interior operator. The repeating interior operator requires satisfying additional constraints and thus further specifies the form of $\mathsf{R}(\mathsf{B})$. Here we present two ways of constructing such operators: one based on ideas in Refs. 65, 67, 62, 23, 24, and 22 and another based on a simplification of the ideas of Kamakoti and Pantano [51] for the construction of the interior operator that allows for a simple extension to operators that include nodes at and near boundaries. The first form, which corresponds to classical minimum-stencil FD-SBP operators, is given as

$$\begin{aligned} \mathsf{D}_2^{(2p,p)}(\mathsf{B}) &= \mathsf{H}^{-1}\left[-\left(\mathsf{D}_1^{(p)}\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1^{(p)} + \mathsf{R}(\mathsf{B}) + \mathsf{E}\mathsf{B}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\right], \\ \mathsf{R}(\mathsf{B}) &= \sum_{i=p+1}^{2p} \theta_i^{(p)} h^{2i-1}\left(\mathsf{D}_{i,p}^{(2,1)}\right)^{\mathrm{T}}\mathsf{C}_i^{(p)}\mathsf{B}\mathsf{D}_{i,p}^{(2,1)}. \end{aligned} \tag{5.27}$$

The form of $\mathsf{R}(\mathsf{B})$ was first proposed for constant-coefficient operators by Mattsson et al. [67] and has its origin in the observation that minimum-stencil centred operators can be reformulated as the application of centered approximations of the first derivative twice plus second-order approximations to even derivatives. Now we list some important properties of the various matrices used in (5.27).

- The first-derivative operators, $\mathsf{D}_1^{(2p,p)}$, with a repeating interior operator have different orders at interior nodes and at boundary nodes. In particular for those considered in this thesis, the first and last $2p$ nodes are of order p while interior nodes are of order $2p$;

- the matrix operators $\mathsf{D}_{i,p}^{(2,1)}$ have a centered interior stencil that spans $2p+1$ entries, while the boundary stencils at the first and last $2p$ rows have $3p$ entries starting from the left or right, respectively. Application of the interior operator results in a second-order

centered-difference approximation to the i^{th} derivative, while the boundary operators result in first-order approximations;

- $\mathsf{C}_i^{(p)}$ are diagonal matrices of the form

$$\mathsf{C}_i^{(p)} = \operatorname{diag}\left(c_{i,1}^{(p)}, \ldots, c_{i,2p}^{(p)}, 1, \ldots, 1, c_{i,2p}^{(p)}, \ldots, c_{i,1}^{(p)}\right)$$

  and are positive semi-definite;

- the operator $\mathsf{D}_{1,\text{b}}^{(\geq p+1)}$ is an approximation to the first derivative of at least order $p+1$.

The $\theta_i^{(p)}$ coefficients are

- $p = 1$: $\theta_2^{(1)} = \frac{1}{4}$;
- $p = 2$: $\theta_3^{(2)} = \frac{1}{18}$, $\theta_4^{(2)} = \frac{1}{48}$;
- $p = 3$: $\theta_4^{(3)} = \frac{1}{80}$, $\theta_5^{(3)} = \frac{1}{100}$, $\theta_6^{(3)} = \frac{1}{720}$; and
- $p = 4$: $\theta_5^{(4)} = \frac{1}{350}$, $\theta_6^{(4)} = \frac{1}{252}$, $\theta_7^{(4)} = \frac{1}{980}$, $\theta_8^{(4)} = \frac{1}{11200}$.

The θ coefficients are determined by zeroing the contributions to the interior operator from (5.27) that are not within $\pm p$ of the diagonal entry. We note that for $p > 4$, the first-derivative operator requires more than $2p$ nodes that do not have the interior operator (see Ref. 3), so the above definitions would need to be changed accordingly.

Using the above θ coefficients, (5.27) fully specifies the interior operator. However, the proposed form of the $\mathsf{D}_{i,p}^{(2,1)}$ and C operators does not automatically satisfy the degree equations (5.7) at and near boundary nodes. It is therefore necessary to solve the degree equations at and near boundary nodes. Nevertheless, typically, free parameters remain and need to be specified, for example by optimizing the operator in some way (these issues are discussed in more detail in Chapter 7). The form of the various constituent matrices is different from that proposed by Mattsson [62]. Regardless, both formulations lead to the same interior operator, and the analysis in this section applies to both formulations.

The second form is given as

$$\mathsf{D}_2^{(2p,p)}(\mathsf{B}) = \mathsf{H}^{-1}\left[\sum_{i=1}^{n} -\mathsf{B}\,(i,i)\,(\mathsf{M}_i - \mathsf{R}_i) + \mathsf{E}\mathsf{B}\mathsf{D}_{1,\text{b}}^{(\geq p+1)}\right], \tag{5.28}$$

where

$$\left(\mathsf{D}_1^{(2p,p)}\right)^{\mathrm{T}} \mathsf{H}\mathsf{B}\mathsf{D}_1^{(2p,p)} = \sum_{i=1}^{n} \mathsf{B}\,(i,i)\,\mathsf{M}_i, \tag{5.29}$$

and all of the matrices are $\in \mathbb{R}^{n\times n}$. Clearly, the first form can be recast in the second form. Here we concentrate on GSBP operators that have the same interior operator as

(5.27). However, form (5.28) easily allows contemplating other types of interior operators, for example, those constructed by Kamakoti and Pantano [51]. Form (5.28) can be further collapsed by retaining only the nonzero blocks, such that the operators, applied to a vector $\mathbf{u}$, can be constructed as

$$\begin{aligned} \mathsf{D}_2^{(2p,p)}(\mathsf{B})\,\mathbf{u} &= \sum_{i=1}^{g} \mathsf{B}(i,i)\left(\mathsf{F}_i\mathbf{u}_i + \tilde{\mathsf{F}}_i\mathbf{u}_{n-i+1}\right) \\ &\quad + \sum_{i=g+1}^{n-g} \mathsf{B}(i,i)\mathsf{F}_{\text{INT}}\mathbf{u}(i-p:i+p), \end{aligned} \tag{5.30}$$

where the $\tilde{\mathsf{F}}$ are the permutation of the rows and columns of the F_i. The $\mathbf{u}_i$ are portions of the vector $\mathbf{u}$, and F_{INT} is a matrix that originates from the coefficients of the interior point operator. Form (5.30) is to be understood as constructing the vector $\mathsf{D}_2(\mathsf{B})\mathbf{u}$ using the sequence implied by the right-hand side of (5.30). To make the idea more transparent, consider the case for $p = 2$ with 13 nodes, which is the minimum required to have one node where the internal operator is applied. Then we have that

- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(1:4,1) = \mathsf{B}(1,1)\mathsf{F}_1\mathbf{u}\,(1:4)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(1:3) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(1:3) + \mathsf{B}(2,2)\mathsf{F}_2\mathbf{u}\,(1:3)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(1:5) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(1:5) + \mathsf{B}(3,3)\mathsf{F}_3\mathbf{u}\,(1:5)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(1:6) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(1:6) + \mathsf{B}(4,4)\mathsf{F}_4\mathbf{u}\,(1:6)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(3:7) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(3:7) + \mathsf{B}(5,5)\mathsf{F}_5\mathbf{u}\,(3:7)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(4:8) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(4:8) + \mathsf{B}(6,6)\mathsf{F}_6\mathbf{u}\,(4:8)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(5:9) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(5:9) + \mathsf{B}(7,7)\mathsf{F}_{\text{INT}}\mathbf{u}\,(5:9)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(6:10) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(6:10) + \mathsf{B}(8,8)\tilde{\mathsf{F}}_6\mathbf{u}\,(6:10)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(7:10) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(7:10) + \mathsf{B}(9,9)\tilde{\mathsf{F}}_5\mathbf{u}\,(7:10)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(8:13) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(8:13) + \mathsf{B}(10,10)\tilde{\mathsf{F}}_4\mathbf{u}\,(8:13)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(9:13) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(9:13) + \mathsf{B}(11,11)\tilde{\mathsf{F}}_3\mathbf{u}\,(9:13)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(11:13) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(11:13) + \mathsf{B}(12,12)\tilde{\mathsf{F}}_2\mathbf{u}\,(11:13)$
- $(\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(10:13) = (\mathsf{D}_2^{(4,2)}(\mathsf{B})\,\mathbf{u})(10:13) + \mathsf{B}(13,13)\tilde{\mathsf{F}}_1\mathbf{u}\,(10:13),$

where

$$\mathsf{F}_1 = \begin{bmatrix} \frac{920}{289} & -\frac{1740}{289} & \frac{1128}{289} & -\frac{308}{289} \\ \frac{12}{17} & -\frac{59}{68} & \frac{2}{17} & \frac{3}{68} \\ -\frac{96}{731} & \frac{118}{731} & -\frac{16}{731} & -\frac{6}{731} \\ -\frac{36}{833} & \frac{177}{3332} & -\frac{6}{833} & -\frac{9}{3332} \end{bmatrix}, \mathsf{F}_2 = \begin{bmatrix} -\frac{59}{68} & 0 & \frac{59}{68} \\ 0 & 0 & 0 \\ \frac{59}{172} & 0 & -\frac{59}{172} \end{bmatrix}, \tag{5.31}$$

$$\mathsf{F}_3 = \begin{bmatrix} -\frac{16}{731} & \frac{118}{731} & 0 & -\frac{118}{731} & \frac{16}{731} \\ \frac{2}{43} & -\frac{59}{172} & 0 & \frac{59}{172} & -\frac{2}{43} \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{118}{2107} & \frac{3481}{8428} & 0 & -\frac{3481}{8428} & \frac{118}{2107} \\ \frac{1}{129} & -\frac{59}{1032} & 0 & \frac{59}{1032} & -\frac{1}{129} \end{bmatrix}, \tag{5.32}$$

$$\mathsf{F}_4 = \begin{bmatrix} -\frac{2735483983}{17072162892} & \frac{524736737}{1422680241} & -\frac{57572339}{397027044} & -\frac{184955633}{4268040723} & -\frac{1792646}{27895691} & \frac{256900}{5838633} \\ \frac{8893843}{83687073} & -\frac{6941291}{27895691} & \frac{260141}{1946211} & \frac{2446619}{83687073} & \frac{7350}{1640923} & -\frac{2842}{114483} \\ -\frac{57572339}{1004244876} & \frac{15348319}{83687073} & -\frac{513718045}{1004244876} & -\frac{5409871}{251061219} & \frac{729766}{1640923} & -\frac{556864}{14768307} \\ -\frac{184955633}{12301999731} & \frac{2945929}{83687073} & -\frac{5409871}{286093017} & -\frac{1038361}{251061219} & -\frac{50950}{80405227} & \frac{59102}{16829001} \\ -\frac{896323}{39382152} & \frac{72275}{13127384} & \frac{364883}{915864} & -\frac{25475}{39382152} & -\frac{8574937}{19691076} & \frac{8391}{152644} \\ \frac{64225}{4121388} & -\frac{83839}{2747592} & -\frac{34804}{1030347} & \frac{29551}{8242776} & \frac{8391}{152644} & -\frac{40583}{4121388} \end{bmatrix}, \tag{5.33}$$

$$\mathsf{F}_5 = \begin{bmatrix} -\frac{2}{43} & \frac{8}{43} & -\frac{6}{43} & 0 & 0 \\ \frac{8}{49} & -\frac{40}{49} & \frac{24}{49} & \frac{8}{49} & 0 \\ -\frac{1}{8} & \frac{1}{2} & -\frac{3}{4} & \frac{1}{2} & -\frac{1}{8} \\ 0 & \frac{1}{6} & \frac{1}{2} & -\frac{5}{6} & \frac{1}{6} \\ 0 & 0 & -\frac{1}{8} & \frac{1}{6} & -\frac{1}{24} \end{bmatrix}, \mathsf{F}_6 = \begin{bmatrix} -\frac{2}{49} & \frac{8}{49} & -\frac{6}{49} & 0 & 0 \\ \frac{1}{6} & -\frac{5}{6} & \frac{1}{2} & \frac{1}{6} & 0 \\ -\frac{1}{8} & \frac{1}{2} & -\frac{3}{4} & \frac{1}{2} & -\frac{1}{8} \\ 0 & \frac{1}{6} & \frac{1}{2} & -\frac{5}{6} & \frac{1}{6} \\ 0 & 0 & -\frac{1}{8} & \frac{1}{6} & -\frac{1}{24} \end{bmatrix}, \tag{5.34}$$

and

$$\mathsf{F}_{\text{INT}} = \begin{bmatrix} -\frac{1}{24} & \frac{1}{6} & -\frac{1}{8} & 0 & 0 \\ \frac{1}{6} & -\frac{5}{6} & \frac{1}{2} & \frac{1}{6} & 0 \\ -\frac{1}{8} & \frac{1}{2} & -\frac{3}{4} & \frac{1}{2} & -\frac{1}{8} \\ 0 & \frac{1}{6} & \frac{1}{2} & -\frac{5}{6} & \frac{1}{6} \\ 0 & 0 & -\frac{1}{8} & \frac{1}{6} & -\frac{1}{24} \end{bmatrix}. \tag{5.35}$$

The form (5.28) of the operator is not only convenient for analysis, but also from an implementation standpoint. It reduces the application of the operator to one loop. This form is also advantageous for the construction of implicit methods that require the linearization of

order-matched GSBP operators, since the linearization is completely transparent. Moreover, it is a convenient formalism for presenting particular instances of operators.

We discuss the difficulties presented by (5.30) in deriving compatible and order-matched GSBP and classical minimum-stencil FD-SBP operators in Section 7.2.3. Here we are interested in the internal stencil. First, some properties of F_{INT} from (5.27) are summarized in the following proposition:

Proposition 1. *The matrix F_{INT} of a minimum-stencil compatible and order-matched GSBP operator, $\mathsf{D}_2^{(2p,p)}(\mathsf{B})$, constructed from (5.27), has the following properties:*

- *it is of size $(2p+1)\times(2p+1)$;*
- *the j^{th} coefficient of the internal stencil is given as the sum of the j^{th} off diagonal multiplied by the corresponding variable coefficient, with the convention that the main diagonal is zero, those to the right are enumerated using positive numbers, and those to the left are enumerated using negative numbers;*
- *it is bisymmetric, that is, $\mathsf{F}_{\text{INT}} = \mathsf{F}_{\text{INT}}^{\text{T}}$, and $\mathsf{F}_{\text{INT}}\mathsf{P} = \mathsf{P}\mathsf{F}_{\text{INT}}$, where P is the exchange matrix, i.e., a matrix with ones along the antidiagonal;*
- *the entries in the upper right-hand triangle with corner entries $(1, 2p+1)$, $(1, p+2)$, and $(2p+1, p)$, and the associated lower left-hand triangle, have entries that are zero; and*
- *all rows and columns sum to zero.*

The properties listed in Proposition 1 can easily be observed by expanding (5.27). As an example, consider F_{INT} for the classical minimum-stencil FD-SBP operator with $p = 2$ using (5.27), see (5.35), compared to the application of the first-derivative operator twice, which has

$$\mathsf{F}_{\text{INT}} = \begin{bmatrix} -\frac{1}{144} & \frac{1}{18} & 0 & -\frac{1}{18} & \frac{1}{144} \\ \frac{1}{18} & -\frac{4}{9} & 0 & \frac{4}{9} & -\frac{1}{18} \\ 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{18} & \frac{4}{9} & 0 & -\frac{4}{9} & \frac{1}{18} \\ \frac{1}{144} & -\frac{1}{18} & 0 & \frac{1}{18} & -\frac{1}{144} \end{bmatrix}. \tag{5.36}$$

To obtain the above matrices one sets B to be a zero matrix except with one diagonal entry, associated with an interior node, as 1 in $\mathsf{D}_1^{(2p,p)}\mathsf{B}\mathsf{D}_1^{(2p,p)}$. A similar procedure can be used on form (5.27) to construct the matrices in the example.

For classical FD-SBP operators approximating the second derivative with constant coefficients, one of the motivations for minimum-stencil operators [65,67] is that besides increasing

the degree of the operator, the resulting interior stencil has smaller bandwidth than the application of the first-derivative operator twice. For example, taking $p = 2$, the application of the first-derivative operator twice has an operator stencil

$$\frac{1}{h^2}\begin{bmatrix} \frac{1}{144} & -\frac{1}{9} & \frac{4}{9} & \frac{1}{9} & -\frac{65}{72} & \frac{1}{9} & \frac{4}{9} & -\frac{1}{9} & \frac{1}{144} \end{bmatrix}, \tag{5.37}$$

and in general has $4p+1$ coefficients. On the other hand, the minimum-stencil operator has an interior operator given as

$$\frac{1}{h^2}\begin{bmatrix} -\frac{1}{12} & \frac{4}{3} & -\frac{5}{2} & \frac{4}{3} & -\frac{1}{12} \end{bmatrix}, \tag{5.38}$$

and in general has $2p+1$ coefficients. Thus, for constant-coefficient operators, the minimum-stencil requires fewer floating-point operations for interior nodes, as compared to the application of the first-derivative operator twice.

The form of compatible classical FD-SBP operators for the second derivative with constant coefficients motivated the construction of the variable-coefficient operator. Thus, the stencil width has the same properties as the constant-coefficient case; that is, the application of the first-derivative operator twice uses $4p + 1$ nodes, while the minimum-stencil operator uses $2p + 1$ nodes. However, the number of nodes hides a paradox: for $p < 4$, the interior operator of the compatible classical minimum-stencil FD-SBP operator, $\mathsf{D}_2^{(2p,p)}(\mathsf{B})$, requires a greater number of operations to construct as compared to the application of the first-derivative operator twice. Table 5.1 summarizes the number of nonzero entries in F_{INT} and is, therefore, reflective of the number of floating-point operations necessary for constructing the interior operator (the last row is constructed from the observable pattern in $p \in [1, 4]$ and suggests that for $p > 3$, the minimum-stencil operator has fewer nonzero entries). Nevertheless, for $p \leq 3$, minimum-stencil operators are more accurate than the application of the first-derivative operator twice. Moreover, they can be advantageous when used for problems where the Jacobian matrix must be constructed, for example, in optimization. For operators with a repeating interior operator, this results from the fact that the minimum-stencil operators have smaller bandwidth than the application of the first-derivative operator twice, and therefore leads to Jacobian matrices that require less storage and fewer floating-point operations to invert—this issue does not apply to element-type operators.

5.4 Summary

In this chapter, we investigated both classical FD-SBP and GSBP operators approximating the second derivative with a variable coefficient. The proposed operators, called order-matched, are more accurate than the application of the first-derivative operator twice and

Table 5.1: The number of nonzero entries in F_{INT}

p	$\mathsf{D}_2(\mathsf{B})$	DBD
1	7	4
2	19	16
3	37	36
4	61	64
p	$(2p+1)+2\sum_{i=1}^{p} p+i$	$4p^2$

more dissipative of under-resolved modes. For the discretization of PDEs with cross-derivative terms, we proposed using compatible and order-matched operators for the approximation of the second derivative, which lead to provably stable discretizations of such PDEs. Furthermore, we proposed a simple means of constructing order-matched GSBP operators for the second derivative, utilizing the constant-coefficient operator, that greatly simplifies the construction of these operators. Based on the ideas of Kamakoti and Pantano [51], we propose a novel decomposition of compatible and order-matched classical FD-SBP operators for the second derivative that leads to efficient implementations and simplifies construction of Jacobian matrices. We also found that compatible and order-matched classical FD-SBP operators for order $p \leq 3$ require more floating-point operations to apply the interior operator than the application of the first-derivative operator twice. In Chapter 7, we examine the construction of GSBP operators, and propose a number of novel GSBP operators, while in the next chapter we detail the construction of SATs for inter-element or block coupling.

Chapter 6

Simultaneous Approximation Terms at Element Interfaces for GSBP Methods

> "*The construction of spatial difference approximations on structured grids is quite straightforward as long as we stay away from the boundaries*"
>
> —Bertil Gustafsson, *High Order Difference Methods for Time Dependent PDE*

6.1 Introduction

In Chapter 3 we discussed how to construct SATs for the imposition of boundary conditions that lead to stable semi-discrete forms. In this chapter, we continue this work and derive SATs for element or block coupling that leads to stable and conservative schemes. To avoid repeating element and block, element is used to refer to both (see Section 4.2 for a discussion of GSBP operators and the different means of implementing them).

To motivate the conditions imposed on the discrete equations to enforce conservation, consider a conservation law

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega}\mathcal{U}\mathrm{d}\Omega = -\oint_{\partial\Omega}\mathcal{F}\left(\mathcal{U}\right)\cdot\mathbf{n}\mathrm{d}s, \tag{6.1}$$

where $\mathbf{n}$ is an outward-pointing normal (for more details see, for example, Refs. 59 and 20). Equation (6.1) shows that the time rate of change of the integral of the solution depends only on the surface integral of the flux $\mathcal{F}$. In implementing a multi-element approach it is necessary to couple the elements. We do so with SATs, which, in addition to maintaining stability, are constructed to mimic the conservation property of the underlying PDE, i.e., (6.1).

6.2 Linear convection equation

To discretize the linear convection equation with a multi-element approach necessitates SATs to couple the elements. These SATs can then be used to enforce periodic boundary conditions. The required SATs need to lead to conservative and stable schemes. By conservative, we mean that the interface SATs lead to semi-discrete equations that discretely mimic the conservation property of the PDE. This implies that when we integrate, the only contributions from the spatial terms occur at the boundaries. As discussed in Chapter 4, the norm matrix H is an approximation to the L_2 inner product. Here, we prove conservation relative to this norm.

Consider two abutting elements, with solution $\mathbf{u}_h$ in the left element and solution $\mathbf{v}_h$ in the right element. The semi-discrete equations in the left element, with an interface SAT for the interface shared between the two elements, are given by

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{D}_{1,\mathbf{u}_h}\mathbf{u}_h + \tau_{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{E}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}^\mathrm{T}\mathbf{v}_h\right), \tag{6.2}$$

and the contribution from the right element is

$$\frac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t} = -a\mathsf{D}_{1,\mathbf{v}_h}\mathbf{v}_h + \tau_{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{E}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}^\mathrm{T}\mathbf{u}_h\right), \tag{6.3}$$

where the terms multiplied by the τs are the SATs. The subscripts, for example $\mathbf{u}_h$ and $\mathbf{v}_h$, mean that the matrix operators are for the left or right element and are of size $n_{\mathbf{u}_h} \times n_{\mathbf{u}_h}$ and $n_{\mathbf{v}_h} \times n_{\mathbf{v}_h}$, where $n_{\mathbf{u}_h}$ and $n_{\mathbf{v}_h}$ are the number of nodes in the left and right element, respectively. We recall from Section 4.2 that

$$\mathsf{Q}_{\mathbf{u}_h} + \mathsf{Q}_{\mathbf{u}_h}^\mathrm{T} = \mathsf{E}_{\mathbf{u}_h} = \mathsf{E}_{x_\mathrm{R},\mathbf{u}_h} - \mathsf{E}_{x_\mathrm{L},\mathbf{u}_h} = \mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}^\mathrm{T} - \mathbf{t}_{x_\mathrm{L},\mathbf{u}_h}\mathbf{t}_{x_\mathrm{L},\mathbf{u}_h}^\mathrm{T}, \tag{6.4}$$

and

$$\mathsf{Q}_{\mathbf{v}_h} + \mathsf{Q}_{\mathbf{v}_h}^\mathrm{T} = \mathsf{E}_{\mathbf{v}_h} = \mathsf{E}_{x_\mathrm{R},\mathbf{v}_h} - \mathsf{E}_{x_\mathrm{L},\mathbf{v}_h} = \mathbf{t}_{x_\mathrm{R},\mathbf{v}_h}\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}^\mathrm{T} - \mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}^\mathrm{T}. \tag{6.5}$$

Now we determine the restriction on the penalty parameters $\tau_{\mathbf{u}_h}$ and $\tau_{\mathbf{v}_h}$, such that the scheme is conservative. Integration over the volume of the left element is performed by left multiplying (6.2) by $\mathbf{1}_{\mathbf{u}_h}^\mathrm{T}\mathsf{H}_{\mathbf{u}_h}$, where $\mathbf{1}_{\mathbf{u}_h}$ is a vector of ones of size $n_{\mathbf{u}_h} \times 1$, which gives

$$\frac{\mathrm{d}\mathbf{1}_{\mathbf{u}_h}^\mathrm{T}\mathsf{H}_{\mathbf{u}_h}\mathbf{u}_h}{\mathrm{d}t} = -a\mathbf{1}_{\mathbf{u}_h}^\mathrm{T}\mathsf{Q}_{\mathbf{u}_h}\mathbf{u}_h + \tau_{\mathbf{u}_h}\mathbf{1}_{\mathbf{u}_h}^\mathrm{T}\left(\mathsf{E}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}^\mathrm{T}\mathbf{v}_h\right). \tag{6.6}$$

However, by Theorem 4.5, $\mathbf{1}_{\mathbf{u}_h}^\mathrm{T}\mathsf{Q}_{\mathbf{u}_h} = \mathbf{1}_{\mathbf{u}_h}^\mathrm{T}\mathsf{E}_{x,\mathbf{u}_h}$. Furthermore, since only the contributions at

the interface are of interest, we drop all other terms and (6.6) reduces to

$$\frac{\mathrm{d}\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\mathbf{u}_h}{\mathrm{d}t} = -a\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h + \tau_{\mathbf{u}_h}\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\left(\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathbf{v}_h\right). \tag{6.7}$$

Defining

$$\tilde{u}_{\mathrm{x}_R} := \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}^{\mathrm{T}}\mathbf{u}_h, \text{ and } \tilde{v}_{x_{\mathrm{L}}} := \mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathbf{v}_h, \tag{6.8}$$

then

$$\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{E}_{x_{\mathrm{R}}}\mathbf{u}_h = \mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}^{\mathrm{T}}\mathbf{u}_h = \tilde{u}_{\mathrm{x}_R}, \text{ and } \mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathbf{t}_{\mathbf{u}_h,\mathrm{x}_{\mathrm{R}}}\mathbf{t}_{\mathbf{v}_h,\mathrm{x}_{\mathrm{L}}}^{\mathrm{T}}\mathbf{v}_h = \tilde{v}_{x_{\mathrm{L}}}. \tag{6.9}$$

Therefore, (6.7) becomes

$$\frac{\mathrm{d}\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\mathbf{u}_h}{\mathrm{d}t} = -a\tilde{u}_{\mathrm{x}_R} + \tau_{\mathbf{u}_h}\left(\tilde{u}_{\mathrm{x}_R} - \tilde{v}_{x_{\mathrm{L}}}\right), \tag{6.10}$$

and in a similar fashion, (6.3) becomes

$$\frac{\mathrm{d}\mathbf{1}_{\mathbf{v}_h}^{\mathrm{T}}\mathsf{H}_{\mathbf{v}_h}\mathbf{v}_h}{\mathrm{d}t} = a\tilde{v}_{x_{\mathrm{L}}} + \tau_{\mathbf{v}_h}\left(\tilde{v}_{x_{\mathrm{L}}} - \tilde{u}_{\mathrm{x}_R}\right). \tag{6.11}$$

In order to be conservative, the sum of (6.10) and (6.11) must be zero; that is,

$$\frac{\mathrm{d}\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\mathbf{u}_h}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{1}_{\mathbf{v}_h}^{\mathrm{T}}\mathsf{H}_{\mathbf{v}_h}\mathbf{v}_h}{\mathrm{d}t} = \left(-a + \tau_{\mathbf{u}_h} - \tau_{\mathbf{v}_h}\right)\tilde{u}_{\mathrm{x}_R} - \left(-a + \tau_{\mathbf{u}_h} - \tau_{\mathbf{v}_h}\right)\tilde{v}_{x_{\mathrm{L}}} = 0, \tag{6.12}$$

where $\mathbf{1}_{\mathbf{v}_h}$ is a vector of ones of size $n_{\mathbf{v}_h} \times 1$. Therefore, for conservation, it is necessary that $\tau_{\mathbf{v}_h} = \tau_{\mathbf{u}_h} - a$, which is consistent with the classical FD-SBP operators [38], and the result in Ref. 21. We have shown that with appropriate choice of penalty parameters the interface SATs make no contribution to the integration of the semi-discrete equations. Therefore, the only contributions can come from the boundary SATs, as desired.

For stability, we must show that the solution can be bounded by the data of the problem and we do so by applying the energy method. The goal is to further restrict the penalty parameters such that an energy estimate exists. To apply the energy method, we first multiply (6.2) by $\mathbf{u}_h^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}$, which gives

$$\mathbf{u}_h^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathbf{u}_h^{\mathrm{T}}\mathsf{Q}_{\mathbf{u}_h}\mathbf{u}_h + \tau_{\mathbf{u}_h}\mathbf{u}_h^{\mathrm{T}}\left(\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathbf{v}_h\right). \tag{6.13}$$

Adding the transpose of (6.13) to itself, using the SBP property, and simplifying results in

$$\frac{\mathrm{d}\|\mathbf{u}_h\|_{\mathsf{H}_{\mathbf{u}_h}}^2}{\mathrm{d}t} = -a\mathbf{u}_h\mathsf{E}_{\mathbf{u}_h}\mathbf{u}_h + 2\tau_{\mathbf{u}_h}\left(\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathbf{v}_h\right). \tag{6.14}$$

Finally, ignoring the contributions from the left boundary, (6.14) reduces to

$$\frac{\mathrm{d}\|\mathbf{u}_h\|^2_{\mathsf{H}_{\mathbf{u}_h}}}{\mathrm{d}t} = -a\tilde{u}^2_{\mathrm{x}_R} + 2\tau_{\mathbf{u}_h}\left(\tilde{u}^2_{\mathrm{x}_R} - \tilde{u}_{\mathrm{x}_R}\tilde{v}_{x_\mathrm{L}}\right). \tag{6.15}$$

Similarly, (6.3) gives

$$\frac{\mathrm{d}\|\mathbf{v}_h\|^2_{\mathsf{H}_{\mathbf{v}_h}}}{\mathrm{d}t} = a\tilde{v}^2_{x_\mathrm{L}} + 2\tau_{\mathbf{v}_h}\left(\tilde{v}^2_{x_\mathrm{L}} - \tilde{v}_{x_\mathrm{L}}\tilde{u}_{\mathrm{x}_R}\right). \tag{6.16}$$

Adding (6.15) to (6.17), applying the conservation condition $\tau_{\mathbf{v}_h} = \tau_{\mathbf{u}_h} - a$, and simplifying results in

$$\frac{\mathrm{d}\|\mathbf{u}_h\|^2_{\mathsf{H}_{\mathbf{u}_h}}}{\mathrm{d}t} + \frac{\mathrm{d}\|\mathbf{v}_h\|^2_{\mathsf{H}_{\mathbf{v}_h}}}{\mathrm{d}t} = \left(-a + 2\tau_{\mathbf{u}_h}\right)\left(\tilde{u}_{\mathrm{x}_R} - \tilde{v}_{x_\mathrm{L}}\right), \tag{6.17}$$

therefore, for stability, $-a+2\tau_{\mathbf{u}_h} \leq 0$, which is consistent with the classical FD-SBP operators [38] and the result in Ref. 21.

In Appendix B, we detail the construction of boundary SATs for periodic problems and show that with a specific set of penalty parameters the multi-element approach leads to spatial operators with eigenvalues that have zero real parts. The Fourier transform of the first derivative is an imaginary number, and in a similar way the eigenvalues of the circulant matrix that results from the discretization of the first derivative for a periodic problem using centred FD operators have zero real parts (because the matrix is skew symmetric). Thus, with this specific set of penalty parameters, the GSBP-SAT approach can be forced to mimic the properties of the continuous operator. However, typically we do not use such a choice for the penalty parameters, instead opting for a set that leads a dual-consistent discretization. In this case, the eigenvalues are no longer purely imaginary. This highlights the idea that the choice of penalty parameter can achieve additional objectives besides stability and conservation.

6.3 Linear convection-diffusion equation with a variable coefficient

This section is concerned with the construction of SATs to couple elements for the discretization of PDEs that have a second derivative. Gong and Nordström [38] have previously derived SATs for classical FD-SBP methods for the linear convection-diffusion equation with constant coefficients. Here, their derivation is extended to GSBP operators and the linear convection-diffusion equation with a variable coefficient. The analysis presented here is limited to diagonal-norm operators, and we use this fact throughout to make simplifications, typically, $\mathsf{HB} + \mathsf{BH} = 2\mathsf{HB}$.

There are various forms of the interface SATs available. Here the Baumann-Oden variant [38] is constructed (an alternative has been derived by Carpenter, Nordström and Gottlieb

[17]). As for the linear convection equation we specify the penalty parameters from the SATs such that we discretely mimic the conservative property of the PDE. This means that when we integrate the PDE using the SBP norm, the contribution from the interface SATs must be zero. Furthermore, the SATs should be such that an energy estimate exists and therefore the semi-discrete form is stable.

Consider two abutting elements: the semi-discrete equations for the left element are

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} &= -a\mathsf{D}_{1,\mathbf{u}_h}\mathbf{u}_h + \epsilon\mathsf{H}_{\mathbf{u}_h}^{-1}\left[-\left(\mathsf{D}_{1,\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{1,\mathbf{u}_h} - \mathsf{R}_{\mathbf{u}_h}(\mathsf{B}_{\mathbf{u}_h}) + \mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right]\mathbf{u}_h \\
&\quad +\sigma_1^{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathbf{v}_h\right) \\
&\quad +\sigma_2^{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h\right) \\
&\quad +\sigma_3^{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{u}_h}\left(\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathbf{v}_h\right),
\end{aligned} \tag{6.18}$$

and for the right element are

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t} &= -a\mathsf{D}_{1,\mathbf{v}_h}\mathbf{v}_h + \epsilon\mathsf{H}_{\mathbf{v}_h}^{-1}\left[-\left(\mathsf{D}_{1,\mathbf{v}_h}\right)^{\mathrm{T}}\mathsf{H}_{\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{1,\mathbf{v}_h} - \mathsf{R}_{\mathbf{v}_h}(\mathsf{B}_{\mathbf{v}_h}) + \mathsf{E}_{x_{L,\mathbf{v}_h}}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathsf{b},\mathbf{v}_\mathsf{h}}\right]\mathbf{v}_h \\
&\quad +\sigma_1^{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}^{\mathrm{T}}\mathbf{u}_h\right) \\
&\quad +\sigma_2^{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}^{\mathrm{T}}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h\right) \\
&\quad +\sigma_3^{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{v}_h}\left(\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}^{\mathrm{T}}\mathbf{u}_h\right).
\end{aligned} \tag{6.19}$$

The terms multiplied by the σs are the interface SATs that couple the right boundary of the left element to the left boundary of the right element. We need to determine the conditions on the penalty parameters such that the resultant scheme is conservative and stable. We start by deriving the conditions for stability and then move on to conservation.

For stability, we apply the energy method to each set of equations by multiplying by the transpose of the solution and the norm matrix for that element. Our interest is only in the contribution from the interface SATs, therefore, when we expand $\mathsf{D}_2(\mathsf{B}_{\mathbf{u}_h})$ and $\mathsf{D}_2(\mathsf{B}_{\mathbf{v}_h})$, the contribution to the left or right boundary for the left and right elements, respectively, is removed. Therefore, $\mathsf{E}_{x,\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}$ is replaced by $\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}$ and $\mathsf{E}_{x,\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}$ is replaced by

$\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}$. Applying the energy method to (6.18) gives

$$\begin{aligned}
\frac{\mathrm{d}\|\mathbf{u}_h\|^2_{\mathsf{H}_{\mathbf{u}_h}}}{\mathrm{d}t} &= -a\mathbf{u}_h\left(\mathsf{Q}_{\mathbf{u}_h}+\mathsf{Q}^{\mathrm{T}}_{\mathbf{u}_h}\right)\mathbf{u}_h \underbrace{-2\epsilon\mathbf{u}_h^{\mathrm{T}}\left(\mathsf{D}_{1,\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{1,\mathbf{u}_h}\mathbf{u}_h - 2\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{R}_{\mathbf{u}_h}\left(\mathsf{B}_{\mathbf{u}_h}\right)\mathbf{u}_h}_{\mathrm{Diss}_{\mathbf{u}_h}} \\
&+2\epsilon\mathbf{u}_h^{\mathrm{T}}\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h \\
&+\sigma_1^{\mathbf{u}_h}(2\mathbf{u}_h^{\mathrm{T}}\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h^{\mathrm{T}} - \mathbf{u}_h^{\mathrm{T}}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h - \mathbf{v}_h^{\mathrm{T}}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h) \\
&+\sigma_2^{\mathbf{u}_h}(2\mathbf{u}_h^{\mathrm{T}}\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h - \mathbf{u}_h^{\mathrm{T}}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h - \mathbf{v}_h^{\mathrm{T}}\mathsf{D}^{\mathrm{T}}_{\mathrm{b},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h + \mathbf{u}_h) \\
&+\sigma_3^{\mathbf{u}_h}(2\mathbf{u}_h^{\mathrm{T}}\left(\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{u}_h}\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h - \mathbf{u}_h^{\mathrm{T}}\left(\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h \\
&-\mathbf{v}_h^{\mathrm{T}}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h).
\end{aligned} \tag{6.20}$$

A similar expression can be developed for (6.19). Since we are only interested in the interface between the elements, $\mathsf{Q}_{\mathbf{u}_h}+\mathsf{Q}^{\mathrm{T}}_{\mathbf{u}_h}$ is replaced by $\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}$ and $\mathsf{Q}_{\mathbf{v}_h}+\mathsf{Q}^{\mathrm{T}}_{\mathbf{v}_h}$ is replaced by $\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}$, such that only the influence of terms corresponding to the interface is analyzed. Adding (6.20) to the analogue for the right domain results in

$$\begin{aligned}
\frac{\mathrm{d}\left(\|\mathbf{u}_h\|^2_{\mathsf{H}_{\mathbf{u}_h}}+\|\mathbf{v}_h\|^2_{\mathsf{H}_{\mathbf{v}_h}}\right)}{\mathrm{d}t} &= \tilde{u}^2_{\mathrm{x}_R}\left(-a+2\sigma_1^{\mathbf{u}_h}\right)+\tilde{v}^2_{x_{\mathrm{L}}}\left(a+2\sigma_1^{\mathbf{v}_h}\right)+2\tilde{u}_{\mathrm{x}_R}\tilde{v}_{x_{\mathrm{L}}}\left(\sigma_1^{\mathbf{u}_h}-\sigma_1^{\mathbf{v}_h}\right) \\
&+\tilde{u}_{\mathrm{x}_R}\left(\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h\right)\left(2\epsilon+2\sigma_2^{\mathbf{u}_h}\right)+\tilde{v}_{x_{\mathrm{L}}}\left(\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h\right)\left(-2\epsilon+2\sigma_2^{\mathbf{v}_h}\right) \\
&\tilde{u}_{\mathrm{x}_R}\left(\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h\right)\left(-\sigma_2^{\mathbf{u}_h}-\sigma_3^{\mathbf{v}_h}\right)+\left(\mathbf{v}_h^{\mathrm{T}}\mathsf{D}_{x,b,\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\right)\tilde{u}_{\mathrm{x}_R}\left(-\sigma_2^{\mathbf{u}_h}-\sigma_3^{\mathbf{v}_h}\right) \\
&+2\sigma_3^{\mathbf{u}_h}\left(\mathbf{u}_h^{\mathrm{T}}\left(\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\right)\tilde{u}_{\mathrm{x}_R}+2\sigma_3^{\mathbf{v}_h}\left(\mathbf{v}_h^{\mathrm{T}}\left(\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{v}_h}\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\right)\tilde{v}_{x_{\mathrm{L}}} \\
&+\left(\mathbf{u}_h\left(\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{u}_h}\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\right)\tilde{v}_{x_{\mathrm{L}}}\left(-\sigma_3^{\mathbf{u}_h}-\sigma_2^{\mathbf{v}_h}\right)+\tilde{v}_{x_{\mathrm{L}}}\left(\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h\right)\left(-\sigma_3^{\mathbf{u}_h}-\sigma_2^{\mathbf{v}_h}\right) \\
&+\mathrm{Diss}_{\mathbf{u}_h}+\mathrm{Diss}_{\mathbf{v}_h},
\end{aligned} \tag{6.21}$$

where

$$\tilde{u}_{\mathrm{x}_R}=\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h \text{ and } \tilde{v}_{x_{\mathrm{L}}}=\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h.$$

By defining the vector

$$\mathbf{w}=\left[\tilde{u}_{\mathrm{x}_R},\left(\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h\right)^{\mathrm{T}},\tilde{v}_{x_{\mathrm{L}}},\left(\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h\right)^{\mathrm{T}}\right]^{\mathrm{T}},$$

(6.21) can be written compactly as

$$\frac{\mathrm{d}\left(\|\mathbf{u}_h\|^2_{\mathsf{H}_{\mathbf{u}_h}} + \|\mathbf{v}_h\|^2_{\mathsf{H}_{\mathbf{v}_h}}\right)}{\mathrm{d}t} = \mathrm{Diss}_{\mathrm{L}} + \mathrm{Diss}_{\mathrm{R}} + \mathbf{w}^{\mathrm{T}}\mathsf{A}\mathbf{w}, \tag{6.22}$$

where

$$\mathsf{A} = \begin{bmatrix} (-a + 2\sigma_1^{\mathbf{u}_h}) & (\epsilon + \sigma_2^{\mathbf{u}_h} + \sigma_3^{\mathbf{u}_h}) & (-\sigma_1^{\mathbf{u}_h} - \sigma_1^{\mathbf{v}_h}) & (-\sigma_2^{\mathbf{u}_h} - \sigma_3^{\mathbf{v}_h}) \\ (\epsilon + \sigma_2^{\mathbf{u}_h} + \sigma_3^{\mathbf{u}_h}) & 0 & (-\sigma_3^{\mathbf{u}_h} - \sigma_2^{\mathbf{v}_h}) & 0 \\ (-\sigma_1^{\mathbf{u}_h} - \sigma_1^{\mathbf{v}_h}) & (-\sigma_3^{\mathbf{u}_h} - \sigma_2^{\mathbf{v}_h}) & (a + 2\sigma_1^{\mathbf{v}_h}) & (-\epsilon + \sigma_2^{\mathbf{v}_h} + \sigma_3^{\mathbf{v}_h}) \\ (-\sigma_2^{\mathbf{u}_h} - \sigma_3^{\mathbf{v}_h}) & 0 & (-\epsilon + \sigma_2^{\mathbf{v}_h} + \sigma_3^{\mathbf{v}_h}) & 0 \end{bmatrix}. \tag{6.23}$$

For an energy estimate to exist, we require $\mathsf{A} \leq 0$, since $\mathrm{Diss}_{\mathrm{L}} \leq 0$ and $\mathrm{Diss}_{\mathrm{R}} \leq 0$. This result is identical to that found by [38], and it can be shown that $\mathsf{A} \leq 0$ if

$$\begin{gathered} \sigma_1^{\mathbf{u}_h} \leq \tfrac{a}{2}, \qquad \sigma_1^{\mathbf{v}_h} = \sigma_1^{\mathbf{u}_h} - a, \\ \sigma_2^{\mathbf{v}_h} = \epsilon + \sigma_2^{\mathbf{u}_h} \quad \sigma_3^{\mathbf{u}_h} = -\epsilon - \sigma_2^{\mathbf{u}_h}, \quad \sigma_3^{\mathbf{v}_h} = -\sigma_2^{\mathbf{u}_h}. \end{gathered} \tag{6.24}$$

To show that the proposed interface coupling procedure is conservative, it is necessary to show that $\mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} + \mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}\frac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t} = 0$. Multiplying (6.18) and (6.19) by $\mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}$ and $\mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}$, respectively, yields

$$\begin{aligned} \mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}\tfrac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} &= -a\mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\mathsf{Q}_{\mathbf{u}_h}\mathbf{u}_h + \epsilon\mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\left[-\left(\mathsf{D}_{1,\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{1,\mathbf{u}_h} - \mathsf{R}_{\mathbf{u}_h}(\mathsf{B}_{\mathbf{u}_h}) + \mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right]\mathbf{u}_h \\ &+\sigma_1^{\mathbf{u}_h}\mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\left(\mathsf{E}_{x_{R,\mathbf{u}_h}}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}_{\mathrm{x_L},\mathbf{u}_h}\mathbf{v}_h\right) + \sigma_2^{\mathbf{u}_h}\mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\left(\mathsf{E}_{x_{\mathrm{R}},mbu}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h\right) \\ &+\sigma_3^{\mathbf{u}_h}\mathbf{1}^{\mathrm{T}}_{\mathbf{u}_h}\left(\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{u}_h}\left(\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h\right), \end{aligned} \tag{6.25}$$

and

$$\begin{aligned} \mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}\tfrac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t} &= -a\mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\mathsf{Q}_{\mathbf{v}_h}\mathbf{v}_h + \epsilon\mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\left[-\left(\mathsf{D}_{1,\mathbf{v}_h}\right)^{\mathrm{T}}\mathsf{H}_{\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{1,\mathbf{v}_h} - \mathsf{R}_{\mathbf{v}_h}(\mathsf{B}_{\mathbf{v}_h}) - \mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathsf{b},\mathbf{v}_h}\right]\mathbf{v}_h \\ &+\sigma_1^{\mathbf{v}_h}\mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\left(\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}^{\mathrm{T}}_{\mathrm{x_R},\mathbf{v}_h}\mathbf{u}_h\right) + \sigma_2^{\mathbf{v}_h}\mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\left(\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h\right) \\ &+\sigma_3^{\mathbf{v}_h}\mathbf{1}^{\mathrm{T}}_{\mathbf{v}_h}\left(\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\right)^{\mathrm{T}}\mathsf{B}_{\mathbf{v}_h}\left(\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}\mathbf{t}^{\mathrm{T}}_{x_{\mathrm{R}},\mathbf{u}_h}\mathbf{u}_h\right). \end{aligned} \tag{6.26}$$

To determine the conditions on the σ parameters such that the interface SATs lead to a conservative scheme, (6.25) is added to (6.26). To simplify the resultant equation, recall that $\mathsf{Q}+\mathsf{Q}^{\mathrm{T}}=\mathsf{E}$, and thus $\mathsf{Q}=\mathsf{E}-\mathsf{Q}^{\mathrm{T}}$, and from Chapter 4 Theorem 4.5 recall that $\mathsf{Q}\mathbf{1}=\mathbf{0}$ and $\mathbf{1}^{\mathrm{T}}\mathsf{Q}=\mathbf{1}^{\mathrm{T}}\mathsf{E}$. Since only the interface is of concern, this means that $\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{Q}_{\mathbf{u}_h}$ is replaced by $\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{E}_{x_{\mathrm{R}},\mathbf{u}_h}$ and similarly, $\mathbf{1}_{\mathbf{v}_h}^{\mathrm{T}}\mathsf{Q}_{\mathbf{v}_h}$ is replaced by $\mathbf{1}_{\mathbf{v}_h}^{\mathrm{T}}\mathsf{E}_{x_{\mathrm{L}},\mathbf{v}_h}$. Furthermore,

$$\begin{aligned}\mathsf{D}_2(\mathsf{B})\mathbf{1}&=0,\\ \mathsf{H}^{-1}\left[-\left(\mathsf{D}_1\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1-\mathsf{R}(\mathsf{B})+\mathsf{E}\mathsf{B}\mathsf{D}_{\mathrm{b}}\right]\mathbf{1}&=0.\end{aligned}\tag{6.27}$$

However $\mathsf{D}_{\mathrm{b}}\mathbf{1}=0$, thus,

$$\left[-\left(\mathsf{D}_1\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1-\mathsf{R}(\mathsf{B})\right]\mathbf{1} \;=\; 0,\tag{6.28}$$

which implies that

$$\mathbf{1}^{\mathrm{T}}\left[-\left(\mathsf{D}_1\right)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1-\mathsf{R}(\mathsf{B})\right] \;=\; 0.\tag{6.29}$$

With these relations and using the properties of the extrapolation operators, the addition of (6.25) and (6.26) reduces to

$$\begin{aligned}&\mathbf{1}_{\mathbf{u}_h}^{\mathrm{T}}\mathsf{H}_{\mathbf{u}_h}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t}+\mathbf{1}_{\mathbf{v}_h}^{\mathrm{T}}\mathsf{H}_{\mathbf{v}_h}\frac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t}=\tilde{u}_{\mathrm{x}_R}\left(-a+\sigma_1^{\mathbf{u}_h}-\sigma_1^{\mathbf{v}_h}\right)+\tilde{v}_{x_{\mathrm{L}}}\left(a-\sigma_1^{\mathbf{u}_h}+\sigma_1^{\mathbf{v}_h}\right)\\ &+\mathbf{t}_{x_{\mathrm{R}},\mathbf{u}_h}^{\mathrm{T}}\mathsf{B}_{\mathbf{u}_h}\mathsf{D}_{\mathrm{b},\mathbf{u}_h}\mathbf{u}_h\left(\epsilon+\sigma_2^{\mathbf{u}_h}-\sigma_2^{\mathbf{v}_h}\right)+\mathbf{t}_{x_{\mathrm{L}},\mathbf{v}_h}^{\mathrm{T}}\mathsf{B}_{\mathbf{v}_h}\mathsf{D}_{\mathrm{b},\mathbf{v}_h}\mathbf{v}_h\left(-\epsilon-\sigma_2^{\mathbf{u}_h}+\sigma_2^{\mathbf{v}_h}\right).\end{aligned}\tag{6.30}$$

Therefore, if $-a+\sigma_1^{\mathbf{u}_h}-\sigma_1^{\mathbf{v}_h}=0$ and $\epsilon+\sigma_2^{\mathbf{u}_h}-\sigma_2^{\mathbf{v}_h}=0$, the proposed SATs lead to a conservative interface treatment. However, these conditions are automatically satisfied by (6.24). It is concluded that SATs satisfying (6.24) lead to a stable and conservative interface treatment and the following proposition has been proven:

Proposition 2. *The semi-discrete equations (6.18) and (6.19), with appropriate SATs for the remaining interface and boundary conditions, and with conditions (6.24) satisfied, lead to a conservative and stable semi-discrete form.*

6.4 Summary

In this chapter, we continued the work started in Chapter 3 of constructing SATs that lead to conservative and stable semi-discrete forms. In particular, we examined the construction of SATs for the coupling between elements for the linear convection and the linear convection-diffusion equations. We derived the restrictions on the penalty parameters from the SATs such that the resultant semi-discrete forms are both conservative and stable. In Appendix B, for the linear convection equation, we prove that with a certain set of penalty parameters, the spatial operator has eigenvalues with zero real part.

Chapter 7

Construction of Generalized Summation-by-Parts Operators

7.1 Introduction

In Chapters 4 and 5, the theory of GSBP operators for first and second derivatives was developed. In this chapter, more detail is given on how to construct particular families of such operators. These families are associated with specific sets of nodal distributions. These operators will be used for the solution of two PDEs in Chapter 8.

Our main interest is in the solution of the NS equations. These equations have second-derivative terms with variable coefficients, and we typically apply our discretization procedure to the equations in curvilinear coordinates. This means that we must necessarily use diagonal-norm operators such that the resultant semi-discrete form is stable [25, 88], and therefore our focus is on diagonal-norm operators.

In Section 7.2, we discuss the construction of operators with a repeating interior operator. On uniform nodal distributions, we describe classical FD-SBP operators and a modified version of such operators. These modified operators are provably stable on curvilinear coordinates while retaining the order of the interior operator at all nodes, i.e., $2p$. In contrast, only diagonal-norm classical FD-SBP operators can be used on curvilinear coordinates and these operators are globally of order p. In addition, in Section 7.2.2, we examine operators with a repeating interior operator on nodal distributions that have uniform spacing on the interior with a finite number of boundary nodes with unequal spacings. By allowing a variation in the nodal spacing at boundary nodes, it is possible to significantly reduce the truncation error coefficients of the resultant operators.

In Section 7.3, we discuss the construction of element-type operators on nodal distributions associated with classical quadrature rules. For such nodal distributions, we construct pseudo-spectral operator matrices, which typically have dense norms, using the GSBP frame-

work, as well as a number of novel GSBP operators with diagonal norms. The latter represent an attractive option since they can be used on curvilinear coordinates while retaining stability. We start by reiterating the degree equations for the first- and second-derivative operators and delineating optimization criteria that are used for the derivation of the various families of GSBP operators.

The form of the degree equations that we use for a degree p GSBP operator for the first derivative is

$$\mathsf{Q}\mathbf{x}^k = k\mathsf{H}\mathbf{x}^{k-1}, \quad k \in [0, p]. \tag{7.1}$$

In this thesis, we focus exclusively on the construction of compatible and order-matched operators. From Chapter 5, the generic form of such operators is given by

$$\mathsf{D}_2\left(\mathsf{B}\right) = \mathsf{H}^{-1}\left\{-\mathsf{D}_1^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1 - \mathsf{R}\left(\mathsf{B}\right) + \mathsf{E}\mathsf{B}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\right\}, \tag{7.2}$$

and the degree equations that the operator must satisfy to be of order p are

$$\mathsf{D}_2\left(\mathsf{diag}\left(\mathbf{x}^k\right)\right)\mathbf{x}^s = s(k+s-1)\mathbf{x}^{k+s-2},\ k+s \leq p+1. \tag{7.3}$$

The solution to the degree equations (7.1) and (7.3) typically results in free parameters that must be specified. This naturally leads to the concept of optimization. The GSBP norm matrix is an approximation to the L_2 inner product and is used to compute the error in simulations, as well as functionals of the solution. Therefore, the discrete GSBP L_2 inner product of the error vector is used as the objective function, which for the first-derivative operator, D_1, is given as

$$J_{\mathbf{e}} = \mathbf{e}_{p+1}^{\mathrm{T}}\mathsf{H}\mathbf{e}_{p+1}, \tag{7.4}$$

where the error vector is given as

$$\mathbf{e}_{p+1} = \mathsf{D}_1\mathbf{x}^{p+1} - (p+1)\,\mathbf{x}^p. \tag{7.5}$$

For GSBP operators approximating the second derivative with a variable coefficient, there are several error vectors, each of which is given by

$$\mathbf{e}_{k,s} = \mathsf{D}_2\left(\mathsf{diag}\left(\mathbf{x}^k\right)\right)\mathbf{x}^s - s(s+k-1)\mathbf{x}^{s+k-2}, \tag{7.6}$$

and the objective function is constructed as

$$J_{\mathbf{e},\mathsf{D}_2} = \sum_{i=0}^{p+2}\left(\mathbf{e}_{i,p+2-i}\right)^{\mathrm{T}}\mathsf{H}\mathbf{e}_{i,p+2-i}. \tag{7.7}$$

In the remainder of this chapter we discuss specific families of operators and how to specify

free parameters that remain after satisfying the degree equations with optimization using the above objective functions.

7.2 Operators with a repeating interior operator

7.2.1 Classical and modified FD-SBP operators for the first derivative

In this section, we discuss the construction of classical FD-SBP operators with a repeating interior operator on uniform nodal distributions as well as a modified version of such operators that have a diagonal norm but retain the order of the interior operator. The steps to constructing both diagonal-norm and block-norm classical FD-SBP operators are

- solve the degree equations (7.1) (for $p \leq 4$ this uniquely specifies H); and
- optimize J_{e}, defined in (7.4) and set any remaining free parameters to 0.

In this thesis, we do not construct classical FD-SBP operators with $p > 4$. However, for such operators, we note that it is necessary to increase the number of boundary operators to greater than $2p$ at each boundary in order for symmetric positive-definite H to exist [3].

One of the disadvantages of diagonal-norm classical FD-SBP operators is that, while the interior operator is of order $2p$, some of the boundary operators are of order p. This renders the operator order p, leading to a solution of order $p+1$ for the hyperbolic case [39]. If a dual-consistent discretization is utilized, functionals converge with the order of the interior operator [44]. We now propose a modification that allows the construction of diagonal-norm operators that retain the order of the interior operator at all nodes. To discuss the modification, consider the case of an operator with an interior operator of order 4. We now construct the derivative operator as $\mathsf{D}_1^{(4)} = \tilde{\mathsf{H}}^{-1}\tilde{\mathsf{Q}}$, where

$$\tilde{\mathsf{H}} = h\mathsf{diag}\left(\tilde{h}_{11}, \tilde{h}_{22}, \tilde{h}_{33}, \tilde{h}_{44}, 1, \ldots, 1, \tilde{h}_{44}, \tilde{h}_{33}, \tilde{h}_{22}, \tilde{h}_{11}\right), \tag{7.8}$$

and

$$\tilde{\mathsf{Q}} = \begin{bmatrix} \tilde{\mathsf{Q}}(1:4,1:4) & & & & & & & & \tilde{\mathsf{Q}}(n-3:n,n-3:n) \\ -\frac{1}{2} & \tilde{q}_{12} & \tilde{q}_{13} & \tilde{q}_{14} & 0 & 0 & 0 & 0 & c_{11} & c_{12} & c_{13} & c_{14} \\ -\tilde{q}_{12} & 0 & \tilde{q}_{23} & \tilde{q}_{24} & 0 & 0 & 0 & 0 & c_{21} & c_{22} & c_{23} & c_{13} \\ -\tilde{q}_{13} & -\tilde{q}_{23} & 0 & \tilde{q}_{34} & -\frac{1}{12} & 0 & 0 & 0 & c_{31} & c_{32} & c_{22} & c_{12} \\ -\tilde{q}_{14} & -\tilde{q}_{24} & -\tilde{q}_{34} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 & c_{41} & c_{31} & c_{21} & c_{11} \\ 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & 0 & 0 \\ -c_{11} & -c_{21} & -c_{31} & -c_{41} & 0 & 0 & \frac{1}{12} & -\frac{2}{3} & 0 & \tilde{q}_{34} & \tilde{q}_{24} & \tilde{q}_{14} \\ -c_{12} & -c_{22} & -c_{32} & -c_{31} & 0 & 0 & 0 & \frac{1}{12} & -\tilde{q}_{34} & 0 & \tilde{q}_{23} & \tilde{q}_{13} \\ -c_{13} & -c_{23} & -c_{22} & -c_{21} & 0 & 0 & 0 & 0 & -\tilde{q}_{24} & -\tilde{q}_{23} & 0 & \tilde{q}_{12} \\ -c_{14} & -c_{13} & -c_{12} & -c_{11} & 0 & 0 & 0 & 0 & -\tilde{q}_{14} & -\tilde{q}_{13} & -\tilde{q}_{12} & \frac{1}{2} \\ -\tilde{\mathsf{Q}}^{\mathrm{T}}(n-3:n,n-3:n) & & & & & & & & -\mathsf{P}\tilde{\mathsf{Q}}(1:4,1:4)\mathsf{P} & & & \end{bmatrix}, \tag{7.9}$$

where the sub-matrix $\tilde{\mathsf{Q}}(n-3:n,n-3:n)$ has the property that $\mathsf{P}\tilde{\mathsf{Q}}(n-3:n,n-3:n) = \tilde{\mathsf{Q}}(n-3:n,n-3:n)^{\mathrm{T}}\mathsf{P}$ such that the resultant operator satisfies the asymmetry of the first derivative under a reflection of the x-axis, that is, $\tilde{x} = -x$ leads to $\frac{\partial}{\partial x} = -\frac{\partial}{\partial \tilde{x}}$. The entries in $\tilde{\mathsf{Q}}(1:4,1:4)$ are not, in general, equal to those in $\mathsf{Q}(1:4,1:4)$ and similarly the entries in $\tilde{\mathsf{H}}$ are not the same as in H. The addition of the $\tilde{\mathsf{Q}}(n-3:n,n-3:n)$ matrix allows the construction of operators that are of order $2p$ everywhere, where the interior operator is of order $2p$. The entries in $\tilde{\mathsf{Q}}(n-3:n,n-3:n)$ are dependent on the number of nodes in the block. Therefore, such an operator must be implemented as an element-type operator. This means that operators must be constructed for each block size and we denote such operators as corner-corrected operators.

Without additional constraints, some of the operators that result have very large coefficients. These operators can be highly susceptible to round-off error. Therefore, in addition to (7.4) for $2p+1$, we use a second objective function, J_{Q}, which is the sum of the squares of the entries of Q, given by

$$J_{\mathsf{Q}} = \mathbf{1}^{\mathrm{T}}\mathsf{Q} \odot \mathsf{Q}\mathbf{1}. \tag{7.10}$$

Maple's© minimize function is used to determine the minimum of objective function (7.4). Free parameters that do not affect J_{e} are used to optimize J_{Q}.

One of the difficulties in deriving SBP operators for $p > 4$ is satisfying the positive-definite constraint on H. For corner-corrected SBP operators, we have found that restricting

the number of non-unity weights in $\frac{1}{h}\tilde{\mathsf{H}}$ to $2p$ at either boundary and first solving for $\tilde{\mathsf{H}}$ results in a positive-definite $\tilde{\mathsf{H}}$ for the operators presented in this paper. The degree equations that $\tilde{\mathsf{H}}$ must satisfy are

$$\mathbf{1}^{\mathrm{T}}\tilde{\mathsf{H}}\mathbf{x}^k - \frac{\left(x_{\mathrm{R}}^{k+1} - x_{\mathrm{L}}^{k+1}\right)}{k+1} = 0,\ k \in [0, 4p-1]; \tag{7.11}$$

therefore, the diagonal entries in $\tilde{\mathsf{H}}$ are quadrature weights. The steps to construct a corner-corrected SBP operator of order $2p$ are:

- specify the number of nodes;
- solve for the quadrature rule using (7.11) with $\tilde{\mathsf{H}}$ constructed to have $2p$ non-unity weights at the first and last $2p$ nodes; it is necessary to check that the resulting $\tilde{\mathsf{H}}$ is positive definite; if it is not then one increases the number of boundary nodes and restarts the process using the additional degrees of freedom (DOFs) to insure that $\tilde{\mathsf{H}}$ is positive definite;
- construct Q using form (7.9) and solve the degree equations (7.1); and
- optimize the free parameters using objective $J_{\mathbf{e}}$, (7.4), and specify the remaining free parameters by optimizing using objective J_{Q}, (7.10).

The above steps are sufficient for the operators considered in this thesis. However, for even higher-order operators, it may be the case that $\frac{1}{h}\tilde{\mathsf{H}}$ will require greater than $2p$ non-unity weights at the first and last number of nodes. Similarly, the number of boundary operators and therefore the number of entries in the corner correction in $\tilde{\mathsf{Q}}$ may need to be expanded. Moreover, we have not yet investigated the use of such operators for approximating the second derivative.

7.2.2 GSBP operators with a repeating interior operator for the first derivative

In this section, we discuss the construction of GSBP operators with a repeating interior operator that have a number of nodes at the boundaries that are not uniformly spaced. The idea of allowing the nodes near the boundary to have variable spacing was first proposed in Ref. 64; however, the nodal distributions we investigate are not the same. By allowing the nodal distribution to vary near the boundary, it is possible to construct operators with reduced error but with no effect on the order of accuracy [25, 26, 64].

Deriving the optimal nodal locations beyond two or three nodes while at the same time ensuring that a positive-definite norm matrix can be found is difficult [64]. Instead, for diagonal-norm SBP operators as discussed in Chapter 4, it is possible to start with a quadrature rule with positive weights and then construct the norm matrix by injecting the weights

of the quadrature rule along the diagonal. Quadrature rules on symmetric nodal distributions that have a number of unequally spaced nodes at and near boundaries with equally spaced interior nodes were proposed by Alpert [4] and have been successfully used to construct GSBP operators [25, 26]. The nodal locations and quadrature weights are determined from the solution to

$$\sum_{i=1}^{j} \tilde{w}_i \tilde{x}_i^r = \tfrac{B_{r+1}(\tilde{a})}{r+1}, \quad r = 0, 1, \ldots, 2\tilde{j} - 2, \tag{7.12}$$

where $B_i(x)$ is the i^{th} Bernoulli polynomial, and $B_0(x) = 1$, and the parameters $\tilde{a}$ and $\tilde{j}$ are chosen so that a particular degree is attained. If they are chosen such that $\tilde{a} = \tilde{j}$, which is the approach taken here, then it is possible to show that the resultant quadrature rule has positive weights up to degree 20 [4]. To enforce a node at the left boundary the equations (7.12) are constrained by

$$\tilde{x}_1 = 0. \tag{7.13}$$

Since the resultant nodal distribution is symmetric it thus includes both boundary nodes. Thus, we consider two nodal distributions: 1) hybrid Gauss-trapezoidal (HGT), which does not include the boundary nodes, and 2) hybrid Gauss-trapezoidal-Lobatto (HGTL), which does include the boundary nodes. To construct a nodal distribution on $x \in [0, 1]$, the following relations are used:

$$\begin{aligned} x_i &= h\tilde{x}_i, \quad x_{n-(i-1)} = 1 - h\tilde{x}_i, \quad i \in [1, \tilde{j}], \\ x_{i+\tilde{j}+1} &= h(\tilde{a} + i), \quad i \in [0, \tilde{n} - 1], \end{aligned} \tag{7.14}$$

where $h = \frac{1}{\tilde{n}+2\tilde{a}-1}$, $\tilde{n}$ is the number of uniformly distributed nodes, and the total number of nodes is given as $n = \tilde{n} + 2\tilde{j}$.

Rather than using the quadrature rules given by Alpert [4], we use only his nodal distributions. It is possible to combine the various ideas presented for operators with a repeating interior operator, namely, using the classical form of Q or the modified form $\tilde{\mathsf{Q}}$ and the two nodal distributions. In this thesis, we construct GSBP operators with a repeating interior operator on the HGT nodal distribution with the classical form of Q and GSBP operators on the HGTL nodal distribution with either the classical form of Q or the form for corner-corrected operators, $\tilde{\mathsf{Q}}$.

7.2.3 Diagonal-norm classical FD-SBP and GSBP operators with a repeating interior operator for the second derivative with a variable coefficient

In Chapter 5, a very general form for classical FD-SBP or GSBP operators approximating the second derivative with a variable coefficient with a repeating interior operator was proposed, i.e., (5.28). The generality of form (5.28) results in a large system of nonlinear equations for the positive semi-definite requirement on $\mathsf{R}(\mathsf{B})$ (see (7.2) and Section 5.3). Without further simplification, Maple© is unable to find solutions; facilitating the solution of the more general form is an ongoing area of research. Thus in this thesis we solve form (5.27) from Chapter 5. However, form (5.28) is extremely convenient both for implementing the operators in a computer code, as well as for presenting them.

The steps taken to construct both classical FD-SBP operators and the HGTL and HGT operators for the second derivative are summarized as follows:

- solve the degree equations (7.1) for the first-derivative GSBP operator;
- if there are free parameters, optimize using (7.4);
- if any free parameters remain, set them to zero;
- construct the GSBP operator for the second derivative using (5.27);
- the degree equations (7.3) are formed and the first $2p$ are solved;
- typically, this results in families of solutions, each with free parameters;
- free parameters are specified through optimization, using the objective function (7.7), with the constraint that the C matrices in (5.27) are positive semi-definite; then
- the remaining free parameters are set to zero.

The form (5.27) leads to nonlinear equations and hence multiple families of solutions. Each one of these families can be optimized with the constraint that the C matrices are positive semi-definite. Some of these families are more difficult to optimize than others, particularly for the HGTL and HGT operators. Here we take the path of least resistance and choose one family for each operator that is easily optimized by Maple©.

7.3 Element-type GSBP operators for the first and second derivatives

We construct a number of GSBP operators on nodal distributions for classical quadrature rules often associated with pseudo-spectral methods. These operators show the flexibility

afforded by the GSBP framework. If one is limited to a Galerkin type procedure, than for each nodal distribution one is forced to use the associated unique operator. From the GSBP perspective, given a nodal distribution, one has the ability to demand additional characteristics of the constructed operators, for example, a diagonal-norm. In this thesis, we only touch on a few families of element-type operators, for more examples see Ref. 21.

To construct the diagonal-norm operators for the first derivative we use the following steps:

- solve the degree equations (7.1) for $p = n - 1$;
- if no solution is found, lower the degree by one and return to the first step;
- if a solution is found, check that H is positive definite; if H is not uniquely positive definite use any degrees of freedom to force H to be positive definite, otherwise lower the degree by one and return to the first step;
- if free parameters remain, optimize using J_{e} (7.4); then
- set any remaining free parameters to zero.

For dense-norm operators, we use the direct solution method from Section 4.4 to construct operators of degree $n - 1$.

From Definition 7 in Chapter 5, compatible and order-matched operators require the construction of $\mathsf{R}(\mathsf{B})$, (see (7.2)). In the most general case, $\mathsf{R}(\mathsf{B})$ can be constructed as

$$\mathsf{R}(\mathsf{B}) = \sum_{i=1}^{N} \mathsf{B}(i,i)\mathsf{R}_i, \tag{7.15}$$

with the restriction that R_i is symmetric negative semi-definite. This formulation leads to linear degree equations (5.7), but nonlinear constraints for R_i to be symmetric negative semi-definite. Alternatively, R_i is constructed to be symmetric negative semi-definite as follows:

$$\mathsf{R}_i = \mathsf{L}_i^{\mathrm{T}} \mathsf{\Lambda}_i \mathsf{L}_i, \tag{7.16}$$

where L_i is lower unitriangular, and $\mathsf{\Lambda}_i$ is a diagonal matrix. Now the constraints that R_i be symmetric negative semi-definite reduce to the constraints that $\mathsf{\Lambda}_i$ be negative semi-definite; however, the degree equations become nonlinear. Although (7.16) is guaranteed to result in compatible order-matched operators, if solutions can be found, the resultant system of equations is very difficult to solve, particularly for operators with many nodes. This motivates the search for simplifications of $\mathsf{R}(\mathsf{B})$ as have been found for classical FD-SBP operators. This is a current area of research.

We seek a construction of $\mathsf{R}(\mathsf{B})$ so that it is of the form (7.15) and satisfies the restriction that the R_i be symmetric negative semi-definite, but avoids solving a large system of nonlinear equations. This is accomplished by taking advantage of Theorem 5.1. First, the constant-coefficient order-matched GSBP operator for the second derivative is constructed as

$$\mathsf{D}_2(\mathsf{B}) = \mathsf{H}^{-1}\left[-(\mathsf{D}_1)^{\mathrm{T}}\mathsf{H}\mathsf{D}_1 + \mathsf{R}_{\mathrm{c}} + \mathsf{E}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\right], \tag{7.17}$$

which has degree equations

$$\mathsf{D}_2\mathbf{x}^k = k(k-1)\mathbf{x}^{k-2}, \quad j \in [0,p]. \tag{7.18}$$

By Theorem 5.1, if R_{c} is symmetric negative semi-definite, then a compatible order-matched GSBP operator is given by

$$\mathsf{D}_2(\mathsf{B}) = \mathsf{H}^{-1}\left[-(\mathsf{D}_1)^{\mathrm{T}}\mathsf{H}\mathsf{B}\mathsf{D}_1 + \sum_{i=1}^{n}\frac{\mathsf{B}(i,i)}{n}\mathsf{R}_{\mathrm{c}} + \mathsf{E}\mathsf{B}\mathsf{D}_{1,\mathrm{b}}^{(\geq p+1)}\right]. \tag{7.19}$$

The general steps to construct order-matched diagonal-norm GSBP operators are as follows:

- solve the degree equations (7.1) for the first-derivative GSBP operator;
- if there are free parameters, optimize using (7.4);
- if any free parameters remain, set them to zero;
- solve the degree equations for the constant-coefficient second derivative (7.18);
- use free parameters to ensure that R_{c} is negative semi-definite; then
- if there are free parameters, the operator is optimized using $J_{\mathrm{e},\mathsf{D}_2}$(7.7), and any remaining free parameters are set to zero.

As examples, we construct a number of element-type GSBP operators on the following nodal distributions:

- Equally spaced
- Chebyshev-Gauss

$$x_k = -\cos\left(\frac{(2k+1)\,\pi}{1\,(N-1)+2}\right), \quad k \in [0, n-1] \tag{7.20}$$

- Chebyshev-Lobatto

$$x_k = -\cos\left(\frac{k\pi}{N-1}\right), \quad k \in [0, n-1] \tag{7.21}$$

- Legendre-Gauss-Lobatto: where the x are the solutions to

$$\frac{\mathrm{d}P_{n-1}}{\mathrm{d}x} = 0, \tag{7.22}$$

where the Legendre polynomial, P_n, has the explicit representation

$$P_{n-1} = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k}^2 (x-1)^{n-k} (x+1)^k . \tag{7.23}$$

Even though the first-derivative diagonal-norm GSBP operators are constructed on pseudo-spectral nodal distributions, the operators obtained are not the classical pseudo-spectral operators associated with those nodal distributions, which have dense norms [86], while on the other hand the dense-norm operators of order $n-1$ are the same.

7.4 Summary of operators studied in Chapter 8

Table 7.1 lists the abbreviations used to refer to the various GSBP operators used in Sections 8.2 and 8.3. For PDEs with second-derivative terms we use the additional argument app, such that app = 1 means the application of the first-derivative operator twice and app = 2 refers to compatible and order-matched operators. As an example, a discretization using a compatible and order-matched operator on a nodal distribution with 5 equally spaced nodes is referred to as ES5(2) for the linear convection-diffusion equation, while for the linear convection equation, only the first derivative is used; therefore, reference is made to ES5. Furthermore, for dense-norm operators, we append a prefix "dense". Thus, continuing the example, a dense-norm operator on a nodal distribution with 5 equally spaced nodes would be denoted denseNC5. Some operators with a repeating interior operator can be implemented in the traditional FD manner where mesh refinement is accomplished by increasing the number of mesh nodes where the interior operator is applied. Alternatively, these same operators can be implemented as elements whereby mesh refinement is carried out by increasing the number of elements. To demarcate an operator as having been implemented in traditional FD manner, we append a prefix "trad", for example tradCSBP is a classical FD-SBP operator with a diagonal norm implemented in the traditional FD manner. Finally, the corner-corrected operators are demarcated by appending a prefix "corr".

The degree and order of the various operators, on Chebyshev-Gauss-Lobatto and Chebyshev-Gauss nodal distributions with n nodes, for the first derivative are given as

$$\text{degree} = \text{order} = \lceil \tfrac{n}{2} \rceil, \tag{7.24}$$

Table 7.1: Abbreviations for GSBP operators

Abbreviation	Operator
ES[n](app)	Diagonal-norm element-type GSBP operators constructed on n equally spaced nodes
denseES[n](app)	Dense-norm element-type GSBP operators constructed on n equally spaced nodes
CGL[n](app)	Diagonal-norm element-type GSBP operators constructed on the Chebyshev-Gauss-Lobatto nodal distribution with n nodes
denseCGL[n](app)	Dense-norm element-type GSBP operators constructed on the Chebyshev-Gauss-Lobatto nodal distribution with n nodes
CG[n](app)	Diagonal-norm element-type GSBP operators constructed on the Chebyshev-Gauss nodal distribution with n nodes
denseCG[n](app)	Dense-norm element-type GSBP operators constructed on the Chebyshev-Gauss nodal distribution with n nodes
LGL[n]	Diagonal-norm element-type GSBP operators constructed on the Legendre-Gauss-Lobatto nodal distribution with n nodes
CSBP[p](app)	Diagonal-norm classical FD-SBP operator
denseCSBP[p]	Dense-norm classical FD-SBP operator
corrCSBP[n]_d	element-type corner-corrected FD-SBP operator for the first derivative on n nodes where d is the degree of the operator
HGTL[p](app)	Diagonal-norm GSBP operators on the hybrid Gauss-trapezoidal-Lobatto nodal distribution with n nodes
denseHGTL[p]	Dense-norm GSBP operators on the hybrid Gauss-trapezoidal-Lobatto nodal distribution with n nodes
corrHGTL[n]_d	element-type corner-corrected GSBP operator constructed on the HGTL nodal distribution with n nodes where d is the degree of the operator
HGT[p](app)	Diagonal-norm GSBP operator on the hybrid Gauss-Trapezoidal nodal distribution with n nodes
denseHGT[p]	Dense-norm GSBP operator on the hybrid Gauss-Trapezoidal nodal distribution with n nodes

and for the application of the first-derivative operator twice as

$$\text{degree} = \lceil \tfrac{n}{2} \rceil, \text{ and } \quad \text{order} = \lceil \tfrac{n}{2} \rceil - 1. \tag{7.25}$$

For order-matched operators for the second derivative, the relationship is given as

$$\text{degree} = \lceil \tfrac{n}{2} \rceil + 1, \text{ and } \quad \text{order} = \lceil \tfrac{n}{2} \rceil. \tag{7.26}$$

For operators constructed on the Legendre-Gauss nodal distributions, the first-derivative operator is of degree

$$\text{degree } = n - 1, \tag{7.27}$$

while the application of the first-derivative operator twice is of degree and order

$$\text{degree} = n - 1, \text{ and } \quad \text{order} = n - 2. \tag{7.28}$$

For operators with a repeating interior stencil, the following relations hold for the application of the first-derivative operator twice:

$$\text{degree} = p, \text{ and } \quad \text{order} = p - 1, \tag{7.29}$$

while for order-matched operators, the relationship is given as

$$\text{degree} = p + 1, \text{ and } \quad \text{order} = p. \tag{7.30}$$

For corner-corrected operators the degree, and hence order, of the first-derivative operator is specified when discussing such operators.

Here, we discuss only the degree and order of the operators themselves. However, it is the order of the solution which is mainly of interest. In some contexts, for example optimization, functionals of the solution are important. For such cases, constructing discretizations that are dual consistent could be advantageous as functionals computed with the norm matrix H can converge at rates greater than the convergence rate of the solution [11, 44].

7.5 Summary

In this chapter, we presented various novel operators for the first derivative with a repeating interior operator including operators constructed on the HGT and HGTL nodal distributions, as well as corner-corrected operators. For the second derivative, we discussed the construction of compatible and order-matched operators with a repeating interior operator, and element-type operators. We presented strategies to construct specific instances of the

various operators. The basic steps are to solve the accuracy equations and optimize the resultant operator using the H norm of the error in approximating polynomials one degree higher than the degree of the operator. In Chapter 8, we use the various operators described here to solve the steady linear convection and linear convection-diffusion equations with a source term.

Chapter 8

Numerical Results

"A computation is a temptation that should be resisted as long as possible"

—John P. Boyd, *Chebyshev and Fourier Spectral Methods*

8.1 Introduction

In this chapter, we apply the various families of GSBP operators constructed in Chapter 7 to solve the steady linear convection and convection-diffusion equations, with three objectives in mind. First, we are interested in determining the order of convergence of the solution error. Second, we investigate the efficiency of operators having the same order of the solution error. Third, we determine the effect of using the traditional FD manner of grid convergence versus implementing operators as elements. We have tested a wide range of operators on both the linear convection and linear convection-diffusion equation. The full set of results is contained in Appendix C; here, we present a subset of these results to highlight the observed trends.

8.2 Linear convection equation

We solve the steady linear convection equation, which has been previously used for numerical validation and characterization (see for example Refs. 30 and 21), given as

$$-\frac{\mathrm{d}\mathcal{U}}{\mathrm{d}x}+\mathcal{S}=0, x\in[0,1], \tag{8.1}$$

where the source term $\mathcal{S}$ and the boundary condition $\mathcal{U}(x=0)=\mathcal{G}_{x_{\mathrm{L}}}$ are constructed such that

$$\mathcal{U}(x)=1+\left(\left(-32\,x+16\right)\sin\left(10\,\pi\,x\right)+10\,\cos\left(10\,\pi\,x\right)\pi\right)\mathrm{e}^{-4\,(2\,x-1)^{2}} \tag{8.2}$$

is the solution to (8.1). The discretization is given as

$$-\mathsf{D}_1\mathbf{u}_h + \mathbf{s} + \mathbf{SAT} = 0, \tag{8.3}$$

where the SAT for the left boundary condition is given as

$$\mathbf{SAT}_{x_\mathrm{L}} = -a\mathsf{H}^{-1}\left(\mathbf{u}_h - \mathcal{G}_{x_\mathrm{L}}\mathbf{1}\right). \tag{8.4}$$

The SATs to the left and right of an interface are given as

$$\begin{aligned}
\mathbf{SAT}_{\mathbf{u}_h} &= \tau_{\mathbf{u}_h}\mathsf{H}_{\mathbf{u}_h}\left(\mathsf{E}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{t}_{\mathbf{v}_h,x_\mathrm{L}}\mathbf{v}_h\right), \\
\mathbf{SAT}_{\mathbf{v}_h} &= \tau_{\mathbf{v}_h}\mathsf{H}_{\mathbf{v}_h}\left(\mathsf{E}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{t}_{\mathbf{u}_h,x_\mathrm{R}}\mathbf{u}_h\right),
\end{aligned} \tag{8.5}$$

where $\mathbf{u}_h$ is the solution in the left element, $\mathbf{v}_h$ is the solution in the right element, $\tau_{\mathbf{u}_h} = 0$ and $\tau_{\mathbf{v}_h} = -a$.

Figures 8.1 through 8.3 depict the convergence of $\|\mathbf{e}\|_\mathsf{H}$ versus $\frac{1}{\mathrm{DOF}}$ or $\frac{1}{\mathrm{NNZE}}$ for a subset of the various GSBP operators considered in this thesis (see Appendix C for the rest of the operators). Table 8.1 gives the convergence rates, computed by determining the slope of the line of best fit through the points $(x, y) = (\log(h), \log(\|\mathbf{e}\|_\mathsf{H}))$ associated with the filled-in markers in the figures. For operators implemented in the traditional FD manner, h is taken as the average spacing between nodes, i.e., $\frac{x_\mathrm{R}-x_\mathrm{L}}{n-1}$, while for elements, h is computed as the size of the element. The acronym NNZE is the number of non-zero entries in the spatial operator. It is reflective of the computational cost of computing the RHS for the ODE that results in general from the application of the spatial discretization to a time-varying PDE. DOF stands for the degrees of freedom in the spatial operator. For operators applied in the traditional FD manner, this is simply the number of nodes, while for element-type operators, it is the number of nodes in each element multiplied by the number of elements used to tesselate the domain. The table shows that the convergence rates are greater than or equal to $p+1$ for both operators with a repeating interior operator, whether implemented as elements or in the traditional FD manner, and element-type operators.

For operators with a repeating interior operator, there are numerous trends in the data; however, here we limit the discussion to a few of the more important observations. For a given interior operator, by construction, the corner-corrected operators have order greater than the alternatives and therefore have the largest rates of convergence, as can be seen in Figures 8.1 and 8.2. This increased rate comes at the price of increased truncation error coefficients, particularly as the number of nodes of such operators increases. Nevertheless, relative to diagonal-norm operators with a classical Q structure for sufficiently small error tolerance, the corner-corrected operators are more efficient (this can be seen by examining

the plots of solution error versus $\frac{1}{\text{NNZE}}$ in Figures 8.1 and 8.2). Given that the corner-corrected operators have the highest rates of convergence, they must necessarily be more efficient than dense-norm operators beyond some error tolerance, but for this particular problem, the dense-norm operators are more efficient for most error levels. There are several caveats besides the fact that dense-norm operators are not provably stable when used for curvilinear coordinates [88]. For the NS equations, we need approximations to the second-derivative terms with variable coefficients, and again, dense-norm operators are not provably stable (see Mattsson and Almquist [63] for a potential solution). More importantly, in the flux-reconstruction community, it has been found that using dense-norm operators leads to stability issues for nonlinear problems [18, 49].

From a DG perspective, this appears to be a natural consequence of the fact that the H of dense-norm operators is typically a much lower order approximation to the L_2 inner product and this leads to an aliasing problem for such methods [49]. It is hypothesized that the same problem persists for flux-reconstruction methods [18, 49], and we believe that the GSBP-SAT approach will similarly suffer for dense-norm operators. Nevertheless, it would be interesting to see if in fact dense-norm operators do suffer from these issues for the GSBP-SAT approach.

Another important observation is that the traditional FD manner of performing grid refinement leads to substantially lower global error. This highlights the benefit of removing the error introduced by the lower-order point operators near the interfaces.

Examining Figure 8.3, it becomes evident that within a family of operators, the higher the order the more efficient the discretization procedure. However, given the simplicity and linearity of the current test problem, it is likely the case that these trends will not hold for nonlinear problems (for example, the stiffness of the resultant nonlinear equations may make going to arbitrarily higher order more expensive). It is instructive to compare the various families of operators holding the order of the solution error constant. Figure 8.4 compares operators with solution error of order 4 and we see that the majority of methods are clustered together. Nevertheless, for order 4, the diagonal-norm HGT and HGTL $p = 3$ operators applied in a traditional FD manner are significantly more efficient than the alternatives. These results corroborate our previous observation that the traditional FD manner of grid refinement substantially improves the performance of GSBP operators with a repeating interior operator. Furthermore, we see that one can construct diagonal-norm operators that are more efficient than operators associated with pseudo-spectral methods. For example, for solution error of order $3 - 5$ the diagonal-norm HGTL operators applied in the traditional FD manner are significantly more efficient than the LGL operators (see Appendix C for orders 3 and 5). We also see the penalty for increasing the order of accuracy incurred by the corner-corrected operators is that they are less efficient than operators of the same order. Nevertheless, the measure of efficiency used here, which is akin to the

computational effort for a residual evaluation, is not necessarily pertinent to all situations. For example, in optimization, it is necessary to construct the Jacobian matrix and the bandwidth of this matrix is of primary importance to the efficiency of the resulting numerical method.

We summarize the main observations:

- The corner-corrected operators have an increased order of convergence and are more efficient than diagonal-norm operators with a classical Q structure.
- The traditional FD manner of performing grid refinement significantly reduces the global error relative to the element approach.
- It is possible to construct GSBP operators that are more efficient than pseudo-spectral methods.

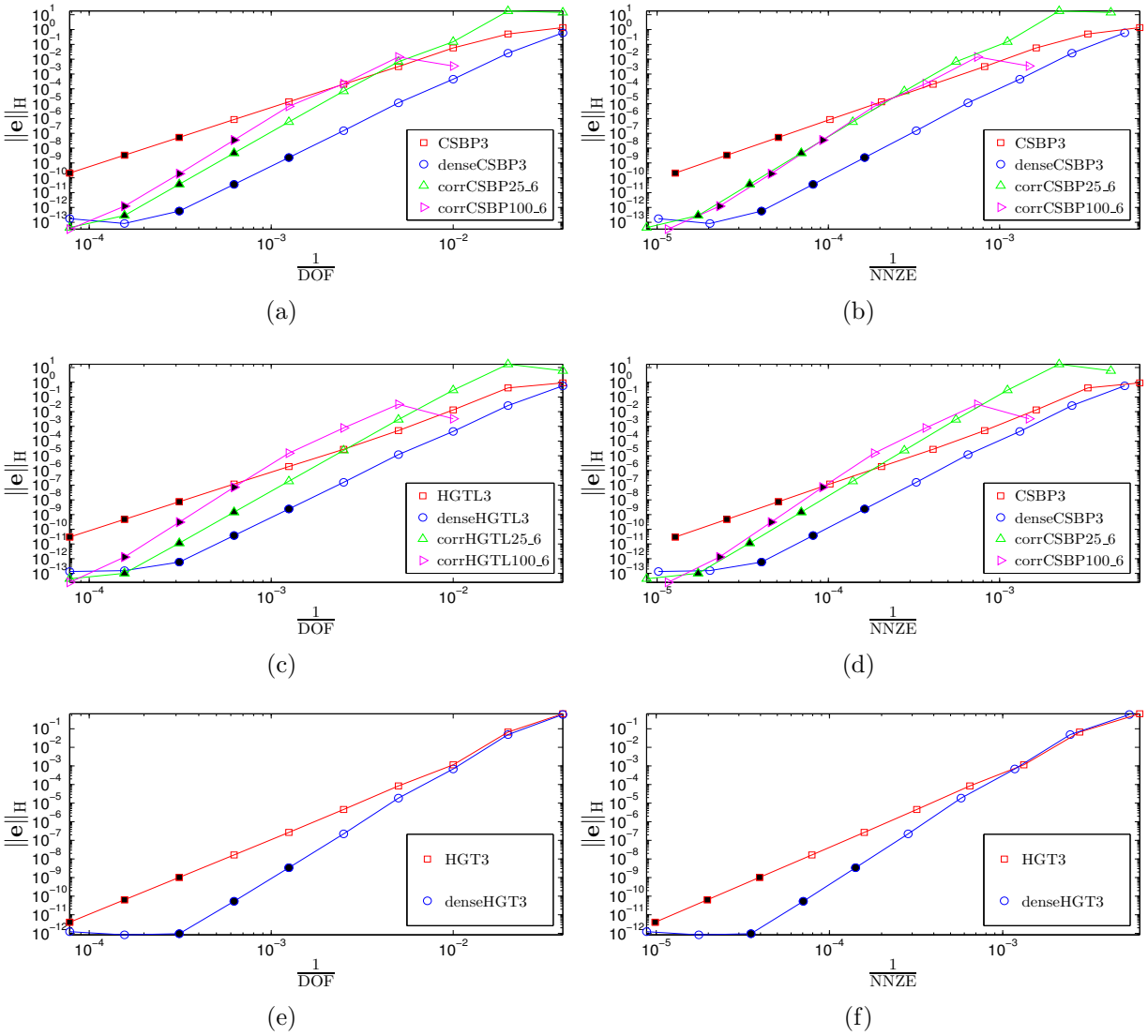


Figure 8.1: Operators with a repeating interior operator of order 6 implemented as elements with 25 nodes, with the exception of the corner-corrected operator with 100 nodes. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$, (a), (c), and (e) or versus $\frac{1}{\text{NNZE}}$, (b), (d) and (f). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 3$.

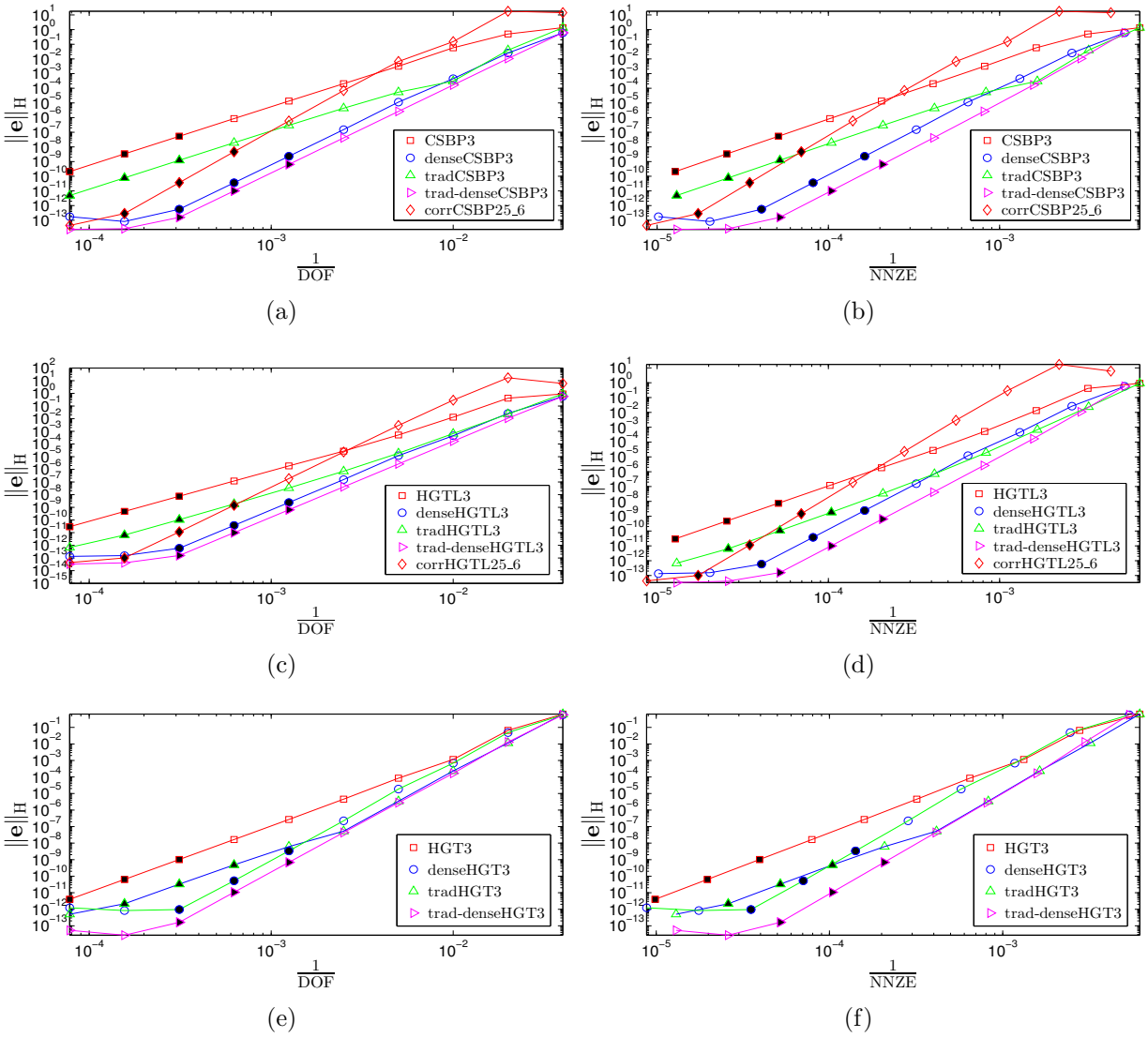
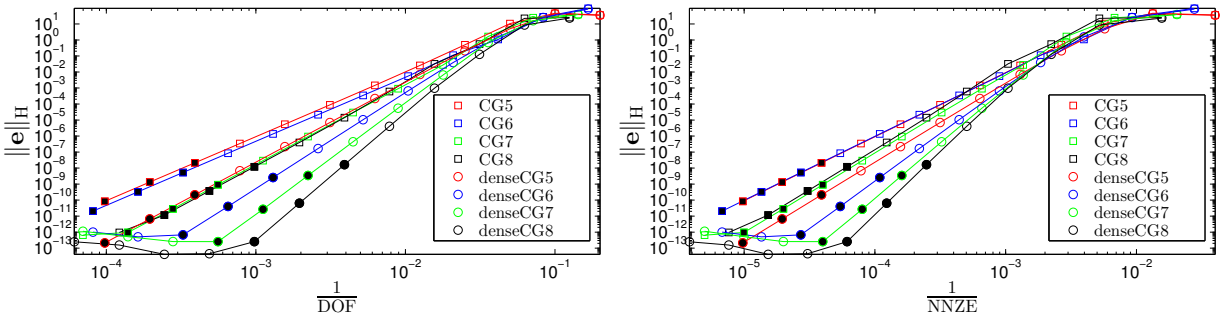
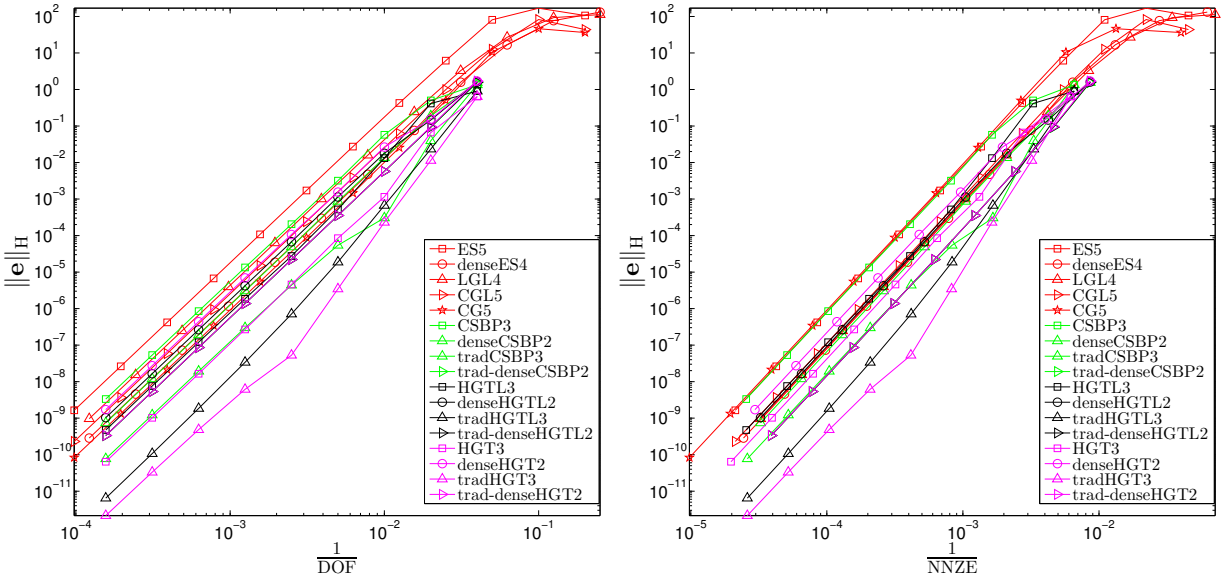


Figure 8.2: Operators with a repeating interior operator of order 6 implemented in a traditional FD manner. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$ (a), (c) and (e) or versus $\frac{1}{\text{NNZE}}$, (b), (d) and (f). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 3$.



(a) (b)

Figure 8.3: Element-type GSBP operators. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\mathrm{DOF}}$, (a) or versus $\frac{1}{\mathrm{NNZE}}$ (b).



(a) (b)

Figure 8.4: H norm of the error in the solution to problem (8.1), for operators with solution error of order 4, versus $\frac{1}{\mathrm{DOF}}$, (a) or versus $\frac{1}{\mathrm{NNZE}}$, (d).

Table 8.1: Convergence of the H norm of the error in the solution of problem (8.1)

Operator	CSBP2	denseCSBP2	corrCSBP25_4	corrCSBP100_4	tradCSBP2	trad-denseCSBP2		
Order	2.9997	3.9999	5.0033	5.2698	3.0038	3.9999		
Operator	HGTL2	denseHGTL2	corrHGTL25_4	corrHGTL100_4	tradHGTL2	trad-denseHGTL2		
Order	2.9996	3.9999	5.0079	5.0128	3.0033	3.9998		
Operator	HGT2	denseHGT2	tradHGT2	trad-denseHGT2				
Order	3	3.9988	3.0238	3.9998				
Operator	CSBP3	denseCSBP3	corrCSBP25_6	corrCSBP100_6	tradDSBP3	trad-denseCSBP3		
Order	3.9991	5.9990	6.9972	7.4099	3.9891	5.9881		
Operator	HGTL3	denseHGTL3	corrHGTL25_6	corrHGTL100_6	tradHGTL3	trad-denseHGTL3		
Order	3.9981	5.9968	6.9087	7.8829	4.0775	5.9925		
Operator	HGT3	denseHGT3	tradHGT3	trad-denseHGT3				
Order	4.0008	5.8948	3.9067	6.0138				
Operator	CSBP4	denseCSBP4	tradCSBP4	trad-denseCSBP4				
Order	4.9956	8.3056	5.2159	8.0448				
Operator	HGTL4	denseHGTL4	corrHGTL25_8	corrHGTL100_8	tradHGTL4	trad-denseHGTL4		
Order	5.0209	8.3439	8.9681	9.3727	4.8768	8.0207		
Operator	ES4	ES5	ES6	denseES4	denseES5	denseES6		
Order	3	4	4	4	4.9392	5.9957		
Operator	CGL5	CGL7	CGL9	denseCGL5	denseCGL7	denseCGL9		
Order	4	4.9979	6.0124	4.9701	6.9418	8.9881		
Operator	CG5	CG6	CG7	CG8	denseCG5	denseCG6	denseCG7	denseCG8
Order	4.0001	3.9999	4.9479	5.0066	4.9796	5.9506	6.8556	7.9830
Operator	LGL2	LGL3	LGL4	LGL5	LGL6	LGL7	LGL13	
Order	1.9999	3	4	4.9697	5.9844	6.9993	12.7811	

8.3 Linear convection-diffusion equation

We solve the steady linear convection-diffusion equation given as

$$-\frac{\mathrm{d}\mathcal{U}}{\mathrm{d}x}+\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathcal{B}\frac{\mathrm{d}\mathcal{U}}{\mathrm{d}x}\right)+\mathcal{S}=0,\ x\in[0,1], \tag{8.6}$$

where the variable coefficient is given as

$$\mathcal{B}=\tanh(x)+\sinh(x). \tag{8.7}$$

The source term $\mathcal{S}$ and the boundary conditions

$$\alpha_{x_\mathrm{L}}\mathcal{U}_{x_\mathrm{L}}+\beta_{x_\mathrm{L}}\mathcal{B}_{x_\mathrm{L}}\left.\frac{\mathrm{d}\mathcal{U}}{\mathrm{d}x}\right|_{x_\mathrm{L}}=\mathcal{G}_{x_\mathrm{L}},\ \alpha_{x_\mathrm{R}}\mathcal{U}_{x_\mathrm{R}}+\beta_{x_\mathrm{R}}\mathcal{B}_{x_\mathrm{R}}\left.\frac{\mathrm{d}\mathcal{U}}{\mathrm{d}x}\right|_{x_\mathrm{R}}=\mathcal{G}_{x_\mathrm{R}} \tag{8.8}$$

are constructed such that (8.2) is the solution to (8.6). The discrete equations are given as

$$-\mathsf{D}_1\mathbf{u}_h+\mathsf{D}_2\left(\mathsf{B}\right)\mathbf{u}_h+\mathbf{s}+\mathbf{SAT}=0, \tag{8.9}$$

where $\mathsf{D}_2\left(\mathsf{B}\right)$ generically represents the application of the first-derivative operator twice or a compatible and order-matched operator. The boundary conditions are implemented using the following SATs [38]:

$$\begin{aligned}\mathbf{SAT}_{x_\mathrm{L}} &= \mathsf{H}^{-1}\mathsf{E}_{x_\mathrm{L}}\left(\alpha_{x_\mathrm{L}}\mathbf{u}_h+\beta_{x_\mathrm{L}}\mathsf{B}\mathsf{D}_b\mathbf{u}_h-\mathcal{G}_{x_\mathrm{L}}\mathbf{1}\right),\\ \mathbf{SAT}_{x_\mathrm{R}} &= \mathsf{H}^{-1}\mathsf{E}_{x_\mathrm{R}}\left(\alpha_{x_\mathrm{R}}\mathbf{u}_h+\beta_{x_\mathrm{R}}\mathsf{B}\mathsf{D}_b\mathbf{u}_h-\mathcal{G}_{x_\mathrm{L}}\mathbf{1}\right).\end{aligned} \tag{8.10}$$

The interface SATs are modelled after the Baumann and Oden [7] type SATs given in Section 6.3 (see Refs. 16, 17, and 38 for more details). We use a set of penalty coefficients similar to those used in Ref. 38 and given as

$$\begin{aligned}&\alpha_{x_L}=1 \quad \beta_{x_L}=-1 \quad \alpha_{x_R}=0 \quad \beta_{x_L}=1\\ &\sigma_{x_L}=-1 \quad \sigma_{x_R}=-1\\ &\sigma_1^{(\mathbf{u})_h}=\tfrac{1}{2} \quad \sigma_2^{(\mathbf{u})_h}=1 \quad \sigma_3^{(\mathbf{u})_h}=-2\\ &\sigma_1^{(\mathbf{v})_h}=-\tfrac{1}{2} \quad \sigma_2^{(\mathbf{v})_h}=2 \quad \sigma_3^{(\mathbf{v})_h}=-1.\end{aligned} \tag{8.11}$$

Figure 8.5 presents the convergence of the H norm of the error of the solution for operators

with a repeating interior operator of order 6. Examining Table 8.2, we see that both the application of the first-derivative operator twice and the order-matched operators converge at a rate greater than or equal to $p+1$ and $p+2$, respectively, where p is the order of the first-derivative operator. For operators implemented in a traditional FD manner, this result is consistent with the theory in Ref. 90.

Figure 8.6 displays the convergence of the H norm of the error of the solution for element-type GSBP operators. For these operators, below a certain error tolerance, the order-matched operators have lower global error. However, in contrast to operators with a repeating interior operator, the convergence rates do not attain the same superconvergence. The ES42 operator only attains an order of $p+1$, while ES61 displays an order of $p+3$. Furthermore, all of the LGL operators have convergence rates of p. Previous studies of the Baumann and Oden [7] SATs have also shown suboptimal convergence rates for pseudo-spectral operators [16, 17], and it is possible that the same mechanisms are at play here.

Finally, Figure 8.7 compares the various operators roughly based on the order of solution error, for order 4. We say roughly since some of the operators with a repeating interior operator exhibit an unexpected increase in their convergence rates when implemented in the traditional FD manner. However, if the element implementation has a solution with error of order r, then the traditional FD implementation is compared to operators with solutions that have error of the same order. For all orders of accuracy considered, we see that one of the various operators with a repeating interior operator leads to the most efficient method. Furthermore, we see that for those orders where they are available, the order-matched operators are nearly as efficient as the application of the first-derivative operator twice. This likely results since, for a given order, the order-matched operators have much larger truncation error coefficients on the interior and likely at the boundaries. They are therefore competitive relative to the application of the first-derivative operator twice because of the reduced number of floating-point operations.

We summarize the main trends:

- The traditional FD manner of performing grid refinement significantly reduces the global error.
- Order-matched operators are more accurate and therefore more efficient than the application of the first-derivative operator twice.
- Order-matched operators with a repeating interior operator are more efficient then the element-based operators examined.

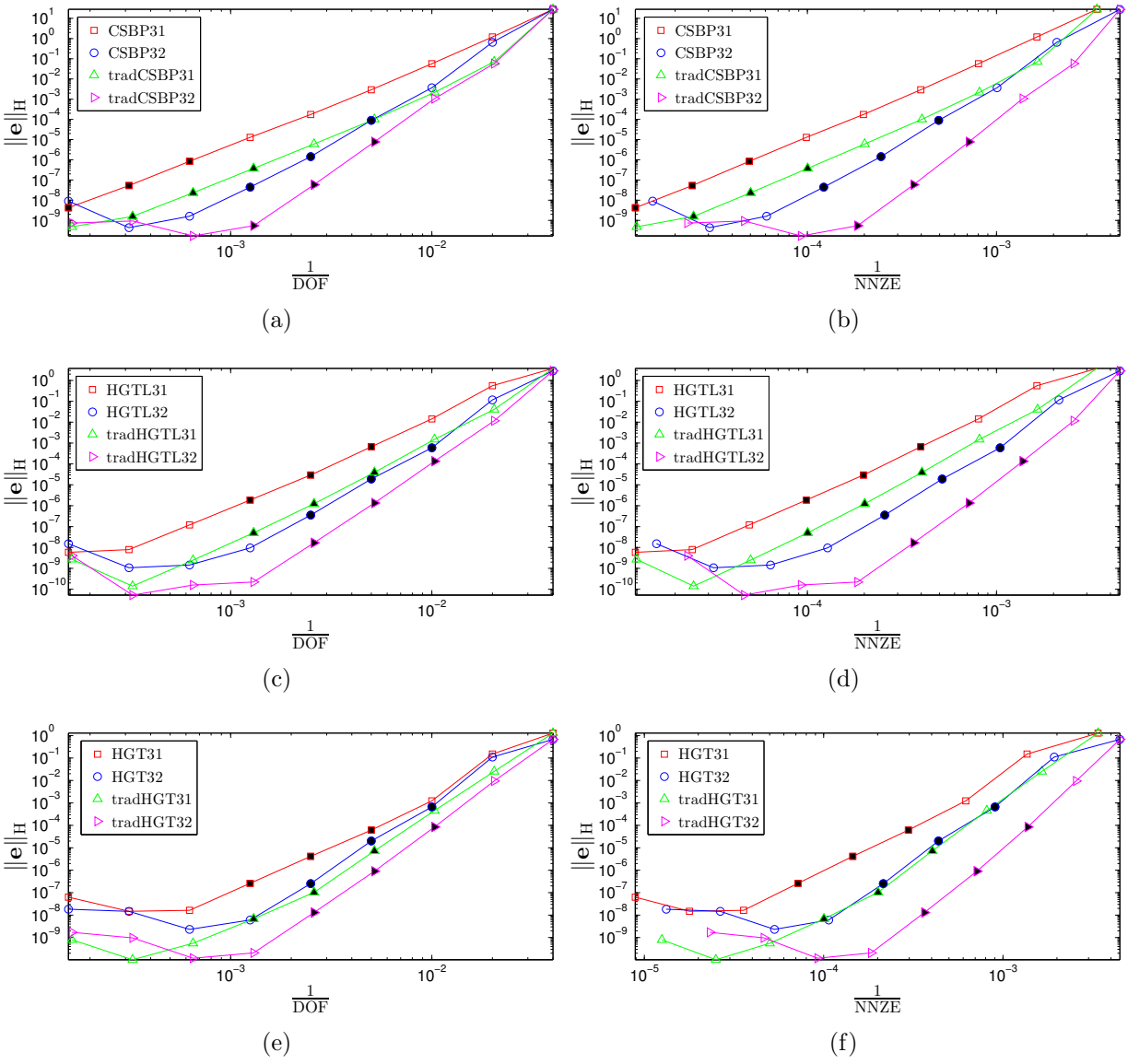


Figure 8.5: Operators with a repeating interior operator of order 6 implemented as elements with 25 nodes or in a traditional FD manner. H norm of the error in the solution to problem (8.6) versus $\frac{1}{\text{DOF}}$, (a), (c), and (e) or versus $\frac{1}{\text{NNZE}}$, (b), (d), and (f). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 3$.

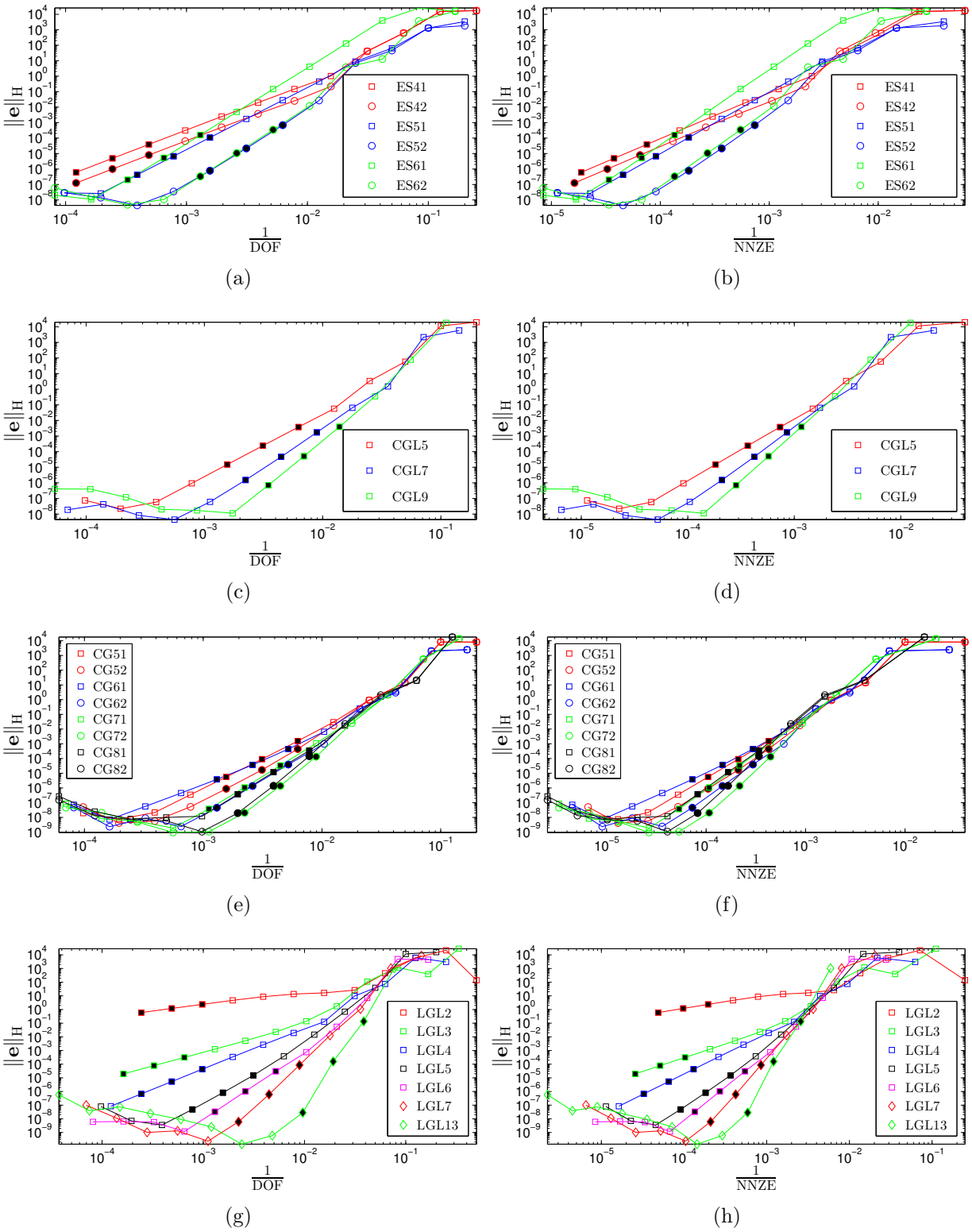
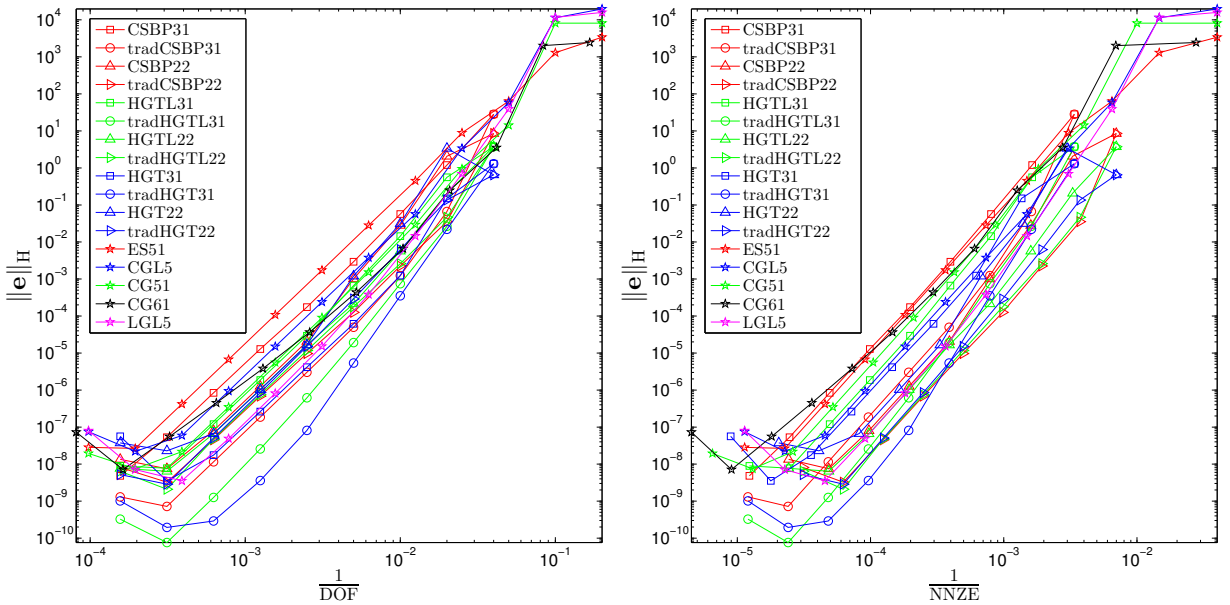


Figure 8.6: Element-type GSBP operators. H norm of the error in the solution to problem (8.6) versus $\frac{1}{\mathrm{DOF}}$, (a), (c), and (g) or versus $\frac{1}{\mathrm{NNZE}}$, (b), (d), (f), and (h).



(a) (b)

Figure 8.7: H norm of the error in the solution to problem (8.6), for operators with solution error of order 4, versus $\frac{1}{\text{DOF}}$, (a) or versus $\frac{1}{\text{NNZE}}$, (b).

Table 8.2: Convergence of the H norm of the error of the solution to problem (8.6)

Operator	CSBP21	CSBP22	tradCSBP21	tradCSBP22				
Order	3.0082	4.0032	3.0187	3.8642				
Operator	HGTL21	HGTL22	tradHGTL21	tradHGTL22				
Order	3.0112	3.9828	3.0206	3.9686				
Operator	HGT21	HGT22	tradHGT21	tradHGT22				
Order	2.9948	3.9765	3.3336	4.1045				
Operator	CSBP31	CSBP32	tradCSBP31	tradCSBP32				
Order	3.7223	5.4929	4.0044	6.8103				
Operator	HGTL31	HGTL32	tradHGTL31	tradHGTL32				
Order	4.2375	5.3504	4.7588	6.4776				
Operator	HGT31	HGT32	tradHGT31	tradHGT32				
Order	3.9386	5.6799	5.2579	6.3148				
Operator	CSBP41	CSBP42	tradCSBP41	tradCSBP42				
Order	5.9131	6.5185	5.8951	7.8534				
Operator	HGTL41	HGTL42	tradHGTL41	tradHGTL42				
Order	4.5654	—	7.0775	—				
Operator	ES41	ES42	ES51	ES52	ES61	ES62		
Order	2.9998	2.9919	4.0011	4.8845	4.8087	5.0064		
Operator	CGL5	CGL7	CGL9					
Order	3.9835	5.0722	6.2078					
Operator	CG51	CG52	CG61	CG62	CG71	CG72	CG81	CG82
Order	4.0488	4.4845	3.4246	4.8738	4.9086	6.3371	4.9168	6.4207
Operator	LGL2	LGL3	LGL4	LGL5	LGL6	LGL7	LGL13	
Order	0.9968	2.0011	2.9984	4.1411	4.9265	6.8913	11.0887	

8.4 Summary

In this chapter, we applied various families of GSBP operators to the solution of the steady linear convection and convection-diffusion equations. It was found that several of the novel families of operators presented in this thesis are more efficient than CSBP operators and pseudo-spectral operators such as on the LGL nodes. Furthermore, we saw that diagonal-norm operators with a repeating interior operator can be constructed that have solutions with error of higher order than dense-norm operators with a repeating interior operator. Finally, we demonstrated the clear advantage of using the traditional FD manner of grid refinement in reducing the global error of the solution.

Chapter 9

Conclusions, Contributions, and Recommendations

9.1 Conclusions and contributions

9.1.1 Theory of GSBP operators for the first derivative

In Chapter 4, the theory of GSBP operators was developed for both diagonal-norm and dense-norm operators. The GSBP framework extends the theory of classical FD-SBP operators of Kreiss and Scherer [56] and Strand [87] to a broader set of operators that exist on nonuniform nodal distributions that may not include nodes at the boundaries. By doing so, it is possible to derive families of operators that have preferential error characteristics, relative to diagonal-norm classical FD-SBP operators.

At a more fundamental level, the GSBP framework synthesizes a number of seemingly unrelated numerical methods. As FD practitioners, our starting point is local approximations based on Taylor expansions. Foreign to us is the idea of using basis function expansions and the variational formulation of PDEs. We perform grid refinement by increasing the number of nodes on the interior of the operator, where the added nodes are discretized by the same interior operator. However, through the GSBP framework, we see that these same FD operators can be treated as elements in the same way that, for example, pseudo-spectral or DG methods lead to difference operators. This revelation highlights the simple fact that regardless of the chosen path for deriving a GSBP operator, the end result is the same and the important point is the properties that these operators are equipped with.

Conclusions:

- The SBP property can be extended to a much broader class of operators than classical FD-SBP operators. The idea that the SBP property extends to other operators is not by itself surprising, considering that the original

intent of Kreiss and Scherer [56] was to extend to FD methods the SBP property inherent in some finite-element methods. Moreover, the work of Carpenter and Gottlieb [13] and more recently Gassner [37] makes clear that there are other known operators that also possess the SBP property. What is novel in our work is that we not only delineate a theory for understanding such operators, but we show that the SBP property applies to a broader class of methods than previously recognized or exploited.

- GSBP operators are tightly coupled to quadrature rules, and in particular the existence of diagonal-norm operators depends on the existence of quadrature rules of at least degree $2p-1$ with positive weights, Theorem 4.4.

- It was proven that dense-norm operators of orders $[0, n-1]$ exist on a nodal distribution having n nodes, Theorem 4.10.

- The numerical results show that there exist operators with the SBP property that are more efficient than classical FD-SBP operators and that applying operators with a repeating interior operator in a traditional FD manner is substantially more efficient than applying these same operators as elements.

- Theorems 4.2, 4.6, 4.7, 4.8, and 4.9, and Corollary 3, show that the constituent matrices of GSBP operators are discrete approximations to various bilinear forms.

Contributions:

- An algorithm to construct GSBP operators using the degree equations was developed that avoids the solution of nonlinear equations.

- The theory of GSBP operators was developed.

- The work of Carpenter and Gottlieb [13] was extended to include GSBP operators of degree $n-1$ that may not include one or both boundary nodes.

- Numerical experiments were conducted which demonstrated the efficiency of GSBP operators for the first derivative.

- A framework for extending the one-dimensional GSBP theory to multi-dimensional operators was proposed. These ideas highlight the importance of the chosen definition of the SBP property and the link between Definition 4.1 and the approximation of surface integrals.

9.1.2 Theory of GSBP operators for the second derivative

The NS equations require approximations for the second derivative with a variable coefficient. The easiest means of constructing such approximations is by applying a first-derivative operator twice. However, for operators with a repeating interior operator, this has the disadvantage of having a substantially wider interior operator, and the resultant operator is less dissipative of under-resolved modes. Element-type GSBP operators are usually dense matrices and therefore the application of the first-derivative operator twice does not significantly increase the size of any of the point operators. Nevertheless, the application of the first-derivative operator twice results in an operator that is one order less accurate than the alternatives. In this thesis, the theory of GSBP approximations to the second derivative was developed. We proposed operators, called order matched, that have preferential error characteristics relative to the application of the first-derivative operator twice for parabolic problems. Furthermore, we extended the idea of compatible operators to the GSBP framework. These operators lead to stable discretizations of PDEs with cross-derivative terms.

For classical FD-SBP operators with $p \leq 3$, it was shown that despite having a smaller interior stencil width, the application of the interior operator is computationally more expensive than the interior operator from the application of the first-derivative operator twice. Notwithstanding, for $p \leq 3$, compatible and order-matched operators are still advantageous from our viewpoint given that 1) our experience has shown that the solution of the nonlinear system of equations requires less computational effort, and 2) for high-order optimization based on the discrete adjoint method, compatible and order-matched operators are preferred if the Jacobian matrix is explicitly constructed.

This thesis project started with the development of compatible and order-matched classical FD-SBP operators for the second derivative—this was based on the work of Mattsson [61, 65, 67] but independent of Ref. 62, which appeared subsequently. Having developed the theory of GSBP operators for the first derivative, it was natural to extend these ideas to the second derivative (see Chapter 5). However, the construction of such operators becomes very difficult, as a result of the number of nonlinear equations that must be solved such that provably stable discretizations can be constructed. Instead, we propose a much simpler construction of GSBP operators for the second derivative with a variable coefficient by borrowing matrices from the constant-coefficient operators. This greatly simplifies the construction of the variable-coefficient operator, as a substantially reduced system of nonlinear equations must be solved.

Conclusions:

- The GSBP framework can be extended to the construction of approximations to the second derivative and in particular approximations that are more accurate than the application of the first-derivative operator twice,

Definitions 6 and 7.

- It was proven that compatible and order-matched GSBP operators for the second derivative with a variable coefficient are guaranteed to exist if compatible and order-matched GSBP operators for the second derivative with a constant coefficient exist, Theorem 5.1.
- The numerical results show that compatible and order-matched GSBP operators are more efficient than the application of the first-derivative operator twice for parabolic equations. For other types of equations, more study is necessary.

Contributions:

- The GSBP framework was extended to include approximations to the second derivative with a variable or constant coefficient.
- The definition of compatible and order-matched operators was extended to the GSBP framework.
- A novel decomposition of GSBP and classical FD-SBP operators with a repeating interior operator was proposed for the second derivative with a variable coefficient that 1) leads to more efficient implementations, and 2) can be used with more general interior operators.
- Numerical experiments were conducted that demonstrate the efficiency of GSBP operators for the second derivative.

9.1.3 Simultaneous approximation terms

In Chapter 6, we developed SATs for the coupling of elements or blocks for the linear convection and linear convection-diffusion equations. The construction of and proof that the resultant SATs lead to conservative and stable semi-discrete forms follow directly from classical FD-SBP theory and therefore achieve one of our main objectives, namely to extend this theory to a broader set of operators. In Appendix B, we prove that for a certain set of penalty parameters, the spatial operator resulting from the SBP-SAT approach has purely imaginary eigenvalues. While this is likely known to specialists in classical FD-SBP methods, we have not seen it in writing.

Conclusion: SATs can be constructed for GSBP operators that lead to consistent, conservative, and stable discrete forms.

Contributions:

- It was demonstrated that the SAT method applies to GSBP operators.

- Penalty parameters were derived that lead to discrete operators with purely imaginary eigenvalues for the periodic linear convection equation.

9.1.4 Construction of GSBP operators

One of the main challenges in constructing classical FD-SBP and GSBP operators is in exploiting the free variables that result. In Chapter 7, a systematic means of constructing and optimizing both first- and second-derivative operators was developed. The proposed objective functions were constructed as the H norm of the leading truncation error terms in approximating the derivative of monomials. Furthermore, several novel operators with a repeating interior operator were proposed, including operators on the hybrid Gauss-trapezoidal and hybrid Gauss-trapezoidal-Lobatto nodes—these operators have characteristics similar to those proposed by Mattsson et al. [64]. Diagonal-norm classical FD-SBP, hybrid Gauss-trapezoidal, and hybrid Gauss-trapezoidal-Lobatto families of operators are of an order that is half the order of the interior operator. To increase the order of the resultant operators, a modification of the classical FD-SBP and hybrid Gauss-trapezoidal-Lobatto families of operators was proposed that can attain the order of the interior operator everywhere.

Conclusions: A number of GSBP operators exist that have preferential error characteristics, including hybrid Gauss-trapezoidal-Lobatto, hybrid Gauss-trapezoidal, the corner-corrected operators. It is critical to correctly optimize these operators.

Contributions:

- A procedure was developed for the construction and optimization of GSBP and classical FD-SBP operators for first and second derivatives.
- A number of novel GSBP operators of element-type for the first and second derivatives were developed on nodal distributions associated with classical quadrature rules including the equally-spaced, Chebyshev-Gauss, and Chebyshev-Gauss-Lobatto nodal distributions.
- Several novel GSBP operators with a repeating interior operator were proposed. These were constructed on the hybrid Gauss-trapezoidal and hybrid Gauss-trapezoidal-Lobatto nodal distributions, and result in operators that have lower global error and hence are more efficient than classical FD-SBP operators for both first and second derivatives.
- A modification was proposed to the form of Q of classical FD-SBP operators with a repeating interior operator that allows diagonal-norm operators to be constructed that are of order $2p$ everywhere for the first derivative.

9.2 Recommendations for future work

For nonlinear problems, after applying the spatial discretization it is necessary to solve a system of nonlinear equations, and the benefits of going higher order may be outweighed by an increase in the difficulty of obtaining a solution to the nonlinear system of equations. Moreover, for problems where a discontinuity in the flow exists, for example shocks, it is well known that in the vicinity of such a feature, one must necessarily reduce to first-order accuracy. This immediately raises the question, "For such problems, are high-order methods still useful?" That is to say, the formal accuracy of the method reduces to second order, but is it still beneficial to resolve the problem in smooth regions using high-order operators? These are basic but important questions that must be investigated before high-order methods are more broadly adopted.

In the context of GSBP operators, a first step is to solve smooth nonlinear problems and characterize the efficiency of the various families of operators. Ultimately, we are interested in the solution of the compressible NS equations and this necessitates some form of additional stabilization, for example artificial dissipation. For operators with a repeating interior operator, there are known models that can be used (see Refs. 66, 43, and 78). However, it is not clear how such models translate to element-type GSBP operators.

In this thesis, we sketch an extension of the GSBP framework to multi-dimensional operators. The benefit of such operators is that one can use unstructured grids. Such grids allow for an easier means of capturing geometric complexity. Nevertheless, for low to moderate geometric complexity, Kronecker product operators are the natural choice. This results from the fact that truly multi-dimensional operators couple all nodes in an element for the approximation of derivative terms at a point. Going forward, for multi-dimensional SBP operators, the number of degrees of freedom will substantially increase. This will make choosing a procedure for constructing specific instances of operators even more important.

In this thesis, we did not delve into the theory of dual-consistent discretizations [11, 12, 44, 46]. Such discretizations offer substantial accuracy gains in integrated quantities, i.e., functionals. In this regard, SATs that lead to dual-consistent discretizations of the NS equations, using compatible and order-matched operators, have yet to be derived.

In Chapter 7, we saw that by modifying the form of SBP and GSBP operators with a repeating interior operator, diagonal-norm operators can be constructed that are of the order of the interior operator everywhere. These ideas naturally lead to the question of finding the most efficient operator. Consider, for example, constructing sparse operators on nonuniform nodal distributions.

REFERENCES

[1] Saul S. Abarbanel, Alina E. Chertock, and Amir Yefet. Strict stability of high-order compact implicit finite-difference schemes: The role of boundary conditions for hyperbolic PDEs, I. *Journal of Computational Physics*, 160:42–66, 2000.

[2] Saul S. Abarbanel, Alina E. Chertock, and Amir Yefet. Strict stability of high-order compact implicit finite-difference schemes: the role of boundary conditions for hyperbolic PDEs II. *Journal of Computational Physics*, 160:67–86, 2000.

[3] Nathan Albin and Joshua Klarmann. Existence of SBP operators with diagonal norm. *arXiv:1403.5750v1*, 2014.

[4] Bradley K. Alpert. Hybrid Gauss-trapezoidal quadrature rules. *SIAM Journal on Scientific Computing*, 5:1551–1584, 1999.

[5] Francesco Bassi, Alessandro Colombo, Nicoletta Franchina, Antonio Ghidoni, and Stefano Rebay. High-order accurate p-multigrid discontinuous Galerkin solution of the RANS and k ω turbulence equations. In *V. European Conference on Computational Fluid Dynamics*, 2010.

[6] Francesco Bassi, Andrea Crivellini, Stefano Rebay, and Marco Savini. Discontinuous Galerkin solution of the Reynolds-averaged Navier-Stokes and k-ω turbulence model equations. *Computers & Fluids*, 34(4-5):507–540, 2005.

[7] Carlos Erik Baumann and J. Tinsley Oden. A discontinuous hp finite element method for convection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 175(3-4):311–341, 1999.

[8] Eric E. Becker, Grahm F. Carcy, and J. Tinsley Oden. *Finite elements an introduction volume I.* Princeton-Hall, Inc, 1981.

[9] Jens Berg and Jan Nordström. Stable robin solid wall boundary conditions for the Navier-Stokes equations. *Journal of Computational Physics*, 230(19):7519–7532, 2011.

[10] Jens Berg and Jan Nordström. Superconvergent functional output for time-dependent problems using finite differences on summation-by-parts. *Journal of Computational Physics*, 231(20):6846–6860, 2012.

[11] Jens Berg and Jan Nordström. On the impact of boundary conditions on dual consistent finite difference discretizations. *Journal of Computational Physics*, 236:41–55, 2013.

[12] Jens Berg and Jan Nordström. Duality based boundary conditions and dual consistent finite difference discretizations of the Navier–Stokes and Euler equations. *Journal of Computational Physics*, 259(135–153), 2014.

[13] Mark H. Carpenter and David Gottlieb. Spectral methods on arbitrary grids. *Journal of Computational Physics*, 129(1):74–86, 1996.

[14] Mark H. Carpenter, David Gottlieb, and Saul Abarbanel. Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes. *Journal of Computational Physics*, 111(2):220–236, 1994.

[15] Mark H. Carpenter, Jan Nordström, and David Gottlieb. A stable and conservative interface treatment of arbitrary spatial accuracy. *Journal of Computational Physics*, 148(2):341–365, 1999.

[16] Mark H. Carpenter, Jan Nordström, and David Gottlieb. Revisiting and extending interface penalties for multi-domain summation-by-parts operators. Technical report, NASA Langley Research Center, 2007.

[17] Mark H. Carpenter, Jan Nordström, and David Gottlieb. Revisiting and extending interface penalties for multi-domain summation-by-parts operators. *Journal of Scientific Computing*, 45(1-3):118–150, 2010.

[18] Patrice Castonguay, Peter Vincent, and Antony Jameson. Application of high-order energy stable flux reconstruction schemes to the Euler equations. *AIAA paper 2011-686*, 2011.

[19] Alina E. Chertock. *Strict stability of high-order compact implicit finite-difference schemes: the role of boundary conditions for hyperbolic PDEs*. PhD thesis, Tel-Aviv Univsersity, 1998.

[20] Ian G. Currie. *Fundamental Mechanics of Fluids Third Edition*. Marcel Dekker, 2003.

[21] David C. Del Rey Fernández, Pieter D. Boom, and David W. Zingg. A generalized framework for nodal first derivative summation-by-parts operators. *Journal of Computational Physics*, 266(1):214–239, 2014.

[22] David C. Del Rey Fernández, Jason E. Hicken, and David W. Zingg. Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations. *Computers & Fluids*, 95(22):171–196, 2014.

[23] David C. Del Rey Fernández and David W. Zingg. High-order compact-stencil summation-by-parts operators for the second derivative with variable coefficients. *ICCFD7-2803*, 2012.

[24] David C. Del Rey Fernández and David W. Zingg. High-order compact-stencil summation-by-parts operators for the compressible Navier-Stokes equations. *AIAA Paper 2013-2570*, page , 2013.

[25] David C. Del Rey Fernández and David W. Zingg. Generalized summation-by-parts operators for the second derivative with a variable coefficient. *Submitted to SIAM Journal on Scientific Computing, (see arXiv:1410.0201[Math.NA])*, 2014.

[26] David C. Del Rey Fernández and David W. Zingg. New diagonal-norm summation-by-parts operators for the first derivative with increased order of accuracy. *AIAA aviation 2015*, 2014.

[27] Peter Diener, Ernst Nils Dorband, Erik Schnetter, and Manuel Tiglio. Optimized high-order derivative and dissipation operators satisfying summation by parts, and applications in three-dimensional multi-block evolutions. *Journal of Scientific Computing*, 32(1):109–145, 2007.

[28] Jean Donea and Antonio Huerta. *Finite Element Methods for Flow Problems*.

[29] John A. Ekaterinaris. High-order accurate, low numerical diffusion methods for aerodynamics. *Progress in Aerospace Sciences*, 3-4(2005):192–300, 41.

[30] Sofia Eriksson and Jan Nordström. Analysis of the order of accuracy for node-centered finite volume schemes. *Applied Numerical Mathematics*, 59(10):2659–2676, 2009.

[31] Krzysztof J. Fidkowski, Todd A. Oliver, James Lu, and David L. Darmofal. p-multigrid solution of high-order discontinuous Galkerin discretizations of the compressible Navier-Stokes equations. *Journal of Computational Physics*, 207(1):92–113, 2005.

[32] Travis C. Fisher and Mark H. Carpenter. High-order entropy stable finite difference schemes for nonlinear conservation laws: Finite domains. *Journal of Computational Physics*, 252(1):518–557, 2013.

[33] Travis C. Fisher, Mark H. Carpenter, Jan Nordström, and Nail K. Yamaleev. Discretely conservative finite-difference formulations for nonlinear conservation laws in split form: Theory and boundary conditions. *Journal of Computational Physics*, 234(1):353–375, 2013.

[34] Daniel Funaro. Domain decomposition methods for pseudo spectral approximations part I. Second order equations in one dimension. *Numerische Mathematik*, 52(3):329–344, 1987.

[35] Daniele Funaro and David Gottlieb. A new method of imposing boundary conditions in pseudospectral approximations of hyperbolic equations. *Mathematics of Computation*, 51(184):599–613, 1988.

[36] Haiyang Gao, Z. J. Wang, and H. T. Huynh. Differential formulation of discontinuous Galerkin and related methods for the Navier-Stokes equations. *Communications in Computational Physics*, 13(4):1013–1044, 2013.

[37] Gregor J. Gassner. A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods. *SIAM Journal on Scientific Computing*, 35(3):A1233–A1253, 2013.

[38] Jing Gong and Jan Nordström. Interface procedures for finite difference approximations of the advection-diffusion equation. *Journal of Computational and Applied Mathematics*, 236:602–620, 2011.

[39] Bertil Gustafsson. The convergence rate for difference approximations to mixed initial boundary value problems. *Mathematics of Computation*, 29(130):396–406, 1975.

[40] Bertil Gustafsson, Heinz-Otto Kreiss, and Joseph Oliger. *Time-Dependent Problems and Difference Methods*. Pure and Applied Mathematics. Wiley, second edition, 2013.

[41] Robert E. Harris and Z. J. Wang. High-order adaptive quadrature-free spectral volume method on unstructured grids. *Computers & Fluids*, 38(10):2006–2025, 2009.

[42] Jan S. Hesthaven and Tim Warburton. *Nodal Discontinuous Galerkin Methods Algorithms, Analysis, and Applications*. Springer, 2008.

[43] Jason E. Hicken and David W. Zingg. A parallel Newton-Krylov solver for the Euler equations discretized using simultaneous approximation terms. *AIAA Journal*, 46(11):2773–2786, 2008.

[44] Jason E. Hicken and David W. Zingg. Superconvergent functional estimates from summation-by-parts finite-difference discretizations. *SIAM Journal on Scientific Computing*, 33(2):893–922, 2011.

[45] Jason E. Hicken and David W. Zingg. Summation-by-parts operators and high-order quadrature. *Journal of Computational and Applied Mathematics*, 237(1):111–125, 2013.

[46] Jason E. Hicken and David W. Zingg. Dual consistency and functional accuracy: A finite-difference perspective. *Journal of Computational Physics*, 256(1):161–182, 2014.

[47] H. T. Huynh. A flux reconstruction approach to high-order schemes including discontinuous Galerkin methods. *AIAA paper 2007-4079*, 2007.

[48] Antony Jameson. CFD for aerodynamic design and optimization: Its evolution over the last three decades. *AIAA paper 2003-3438*, 2003.

[49] Antony Jameson, Peter E. Vincent, and Patrice Castonguay. On the non-linera stability of flux reconstruction schemes. *Journal of Scientific Computing*, 50(2):434–445, 2012.

[50] Forrester T. Johnson, Edward N. Tinoco, and N. Jong Yu. Thirty years of development and application of CFD at Boeing Commercial Airplanes, Seattle. *Computers & Fluids*, 34(10):1115–1151, 2005.

[51] Ramji Kamakoti and Carlos Pantano. High-order narrow stencil finite-difference approximations of second-derivatives involving variable coefficients. *SIAM Journal on Scientific Computing*, 31(6):4222–4243, 2009.

[52] Ravi Kannan and Z. J. Wang. A study of viscous flux formulations for a p-multigrid spectral volume Navier Stokes solver. *Journal of Scientific Computing*, 41(2):165–199, 2009.

[53] Adrian Kitson, Robert I. McLachlan, and Nicolas Robidoux. Skew-adjoint finite difference methods on nouniform grids. *New Zealand Journal of Mathematics*, 32(2):139–159, 2003.

[54] David A. Kopriva and Gregor J. Gassner. An energy stable discontinuous Galerkin spectral element discretization for variable coefficient advection problems. *SIAM Journal on Scientific Computing*, 4(36):A2076–A2099, 2014.

[55] Heinz-Otto Kreiss and Jens Lorenz. *Initial-Boundary Value Problems and the Navier-Stokes Equations*, volume 47 of *Classics in Applied Mathematics*. SIAM, 2004.

[56] Heinz-Otto Kreiss and Godela Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. In *Mathematical aspects of finite elements in partial differential equations*, pages 195–212. Academic Press, New York/London, 1974.

[57] Heinz-Otto Kreiss and Godela Scherer. On the existence of energy estimates for difference approximations for hyperbolic system. Technical report, Department of Information Technology Uppsala University, 1977.

[58] Chunlei Liang, Antony Jameson, and Z. J. Wang. Spectral difference method for compressible flow on unstructured grids with mixed elements. *Journal of Computational Physics*, 228(8):2847–2858, 2009.

[59] Harvard Lomax, Thomas H. Pulliam, and David W. Zingg. *Fundamentals of Computational Fluid Dynamics*. Springer-Verlag, 2003.

[60] Ken Mattsson. Boundary procedures for summation-by-parts operators. *Journal of Scientific Computing*, 18(1):133–153, 2003.

[61] Ken Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. Technical report, Department of Information Technology, Uppsala University, October 2010.

[62] Ken Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *Journal of Scientific Computing*, 51(3):650–682, 2012.

[63] Ken Mattsson and Martin Almquist. A solution to the stability issues with block norm summation by parts operators. *Journal of Computational Physics*, 15:418–442, 2013.

[64] Ken Mattsson, Martin Almquist, and Mark H. Carpenter. Optimal diagonal-norm SBP operators. *Journal of Computational Physics*, 264(1):91–111, 2014.

[65] Ken Mattsson and Jan Nordström. Summation by parts operators for finite difference approximations of second derivatives. *Journal of Computational Physics*, 199:503–540, 2004.

[66] Ken Mattsson, Magnus Svärd, and Jan Nordström. Stable and accurate artificial dissipation. *Journal of Scientific Computing*, 21(1):57–79, 2004.

[67] Ken Mattsson, Magnus Svärd, and Mohammad Shoeybi. Stable and accurate schemes for the compressible Navier-Stokes equations. *Journal of Computational Physics*, 227(4):2293–2316, 2008.

[68] Amir Nejat and Carl Ollivier-Gooch. A high-order accurate unstructured finite volume Newton-Krylov algorithm for inviscid compressible flows. *Journal of Computational Physics*, 227(4):2582–2609, 2008.

[69] Thomas E. Nelson and David W. Zingg. Fifty years of aerodynamics: Successes, challenges, and opportunities. *CASI*, 50(1):61–84, 2004.

[70] Jan Nordström. Conservative finite difference formulations, variable coefficients, energy estimates and artificial dissipation. *Journal of Scientific Computing*, 29(3):375–404, 2006.

[71] Jan Nordström and Mark H. Carpenter. Boundary and interface conditions for high-order finite-difference methods applied to the Euler and Navier-Stokes equations. *Journal of Computational Physics*, 148(2):621–645, 1999.

[72] Jan Nordström and Mark H. Carpenter. High-order finite-difference methods, multi-dimensional linear problems, and curvilinear coordinates. *Journal of Computational Physics*, 173(1):149–174, 2001.

[73] Jan Nordström, Jing Gong, Edwin van der Weide, and Magnus Svärd. A stable and conservative high order multi-block method for the compressible Navier-Stokes equations. *Journal of Computational Physics*, 228(24):9020–9035, 2009.

[74] Pelle Olsson. *High-order difference methods and dataparallel implementation*. PhD thesis, Uppsala Univsersity, 1992.

[75] Pelle Olsson. Summation by parts, projections, and stability. I. *Mathematics of Computation*, 64(211):1035–1065, 1995.

[76] Pelle Olsson. Summation by parts, projections, and stability. II. *Mathematics of Computation*, 64(212):1473–1493, 1995.

[77] Pelle Olsson and Joseph Oliger. Energy and maximum norm estimates for nonlinear conservation laws. Technical Report 94–01, The Research Institute of Advanced Computer Science, 1994.

[78] Michal Osusky and David W. Zingg. A parallel Newton-Krylov flow solver for the Navier-Stokes equations discretized using summation-by-parts operators. *AIAA Journal*, 51(12):2833–2851, 2013.

[79] Thomas H. Pulliam and David W. Zingg. *Fundamental Algorithms in Computational Fluid Dynamics*. Springer, 2014.

[80] Adam Reichert, Michel T. Heath, and Daniel J. Bodony. Energy stable numerical method for hyperbolic partial differential equations using overlapping domain decomposition. *Journal of Computational Physics*, 231:5243–5265, 2012.

[81] Adam Harold Reichert. *Stable numerical methods for hyperbolic partial differential equations using overlapping domain decomposition*. PhD thesis, University of Illinois at Urbane-Champaign, 2011.

[82] Neil D. Sandham, Q. Li, and Helen C. Yee. Entropy splitting for high-order numerical simulation of compressible turbulence. *Journal of Computational Physics*, 178(2):307–322, 2002.

[83] Kathrin Schäcke. On the kronecker product. Technical report, University of Waterloo, 2013.

[84] Godela Scherer. *On energy estimates for difference approximations to hyperbolic partial differential equations*. PhD thesis, Uppsala Univsersity, October 1977.

[85] Khosro Shahbazi, Dimitri J. Mavriplis, and Nicholas K. Burgess. Multigrid algorithms for high-order discontinuous Galerkin discretizations of the compressible Navier-Stokes equations. *Journal of Computational Physics*, 228(21):7917–7940, 2009.

[86] Jien Shen, Tao Tang, and Li-Lian Wang. *Spectral methods algorithms, analysis and applications*. Springer, 2011.

[87] Bo Strand. Summation by parts for finite difference approximations for d/dx. *Journal of Computational Physics*, 110(1):47–67, 1994.

[88] Magnus Svärd. On coordinate transformations for summation-by-parts operators. *Journal of Scientific Computing*, 20(1):29–42, 2004.

[89] Magnus Svärd, Mark H. Carpenter, and Jan Nordström. A stable high-order finite difference scheme for the compressible Navier-Stokes equations, far-field boundary conditions. *Journal of Computational Physics*, 225(1):1020–1038, 2007.

[90] Magnus Svärd and Jan Nordström. On the order of accuracy for difference approximation of initial-boundary value problems. *Journal of Computational Physics*, 218(1):333–352, 2006.

[91] Magnus Svärd and Jan Nordström. A stable high-order finite difference scheme for the compressible Navier-Stokes equations: No-slip wall boundary conditions. *Journal of Computational Physics*, 227(10):4805–4824, 2008.

[92] Magnus Svärd and Jan Nordström. Review of summation-by-parts schemes for initial-boundary-value-problems. *Journal of Computational Physics*, 268(1):17–38, 2014.

[93] Jean-Marc Vaassen, Didier Vigneron, and Jean-André Essers. An implicit high order finite volume scheme for the solution of 3D Navier-Stokes equations with new discretization of diffusive terms. *Journal of Computational and Applied Mathematics*, 215(2):595–601, 2008.

[94] Pieter E. Vincent and Antony Jameson. Facilitating the adoption of undstructured high-order methods amongst a wider community of fluid dynamicists. *Mathematical Modelling of Natural Phenomena*, 6(3):97–140, 2011.

[95] Z. J. Wang. High-order methods for the Euler and Navier-Stokes equations on unstructured grids. *Progress in Aerospace Sciences*, 43(1-3):1–41, 2007.

[96] Z. J. Wang, Krzysztof J. Fidkowski, Rémi Abgrall, Francesco Bassi, Doru Caraeni, Andrew Cary, Herman Deconinck, Ralf Hartmann, Koen Hillewaert, H. T. Huynh, Norbert Kroll, Georg Mayer, Per-Olof Persson, Bram van Leer, and Miguel Visbal. High-order CFD methods: Current status and perspective. *International journal for numerical methods in fluids*, 72(8):811–845, 2013.

[97] Z. J. Wang, Yuzhi Sun, C. Liang, and Yen Liu. Extension of the SD method to viscous flow on unstructured grids. *Computational Fluid Dynamics*, Part 2:119–124, 2006.

[98] David M. Williams, Patrice Castonguay, Pieter E. Vincent, and Antony Jameson. Energy stable flux reconstruction schemes for advection-diffusion problems on triangles. *Journal of Computational Physics*, 250(1):53–76, 2013.

[99] Helen C. Yee and Björn Sjögreen. *Designing Adaptive Low-dissipative High Order Schemes for Long-time Integrations*, volume 66 of *Turbulent Flow Computation Fluid Mechanics and Its Applications*, chapter 5, pages 141–198. Springer, 2004.

[100] Helen C. Yee, Marcel Vinokur, and M. Jahed Djomehri. Entropy splitting and numerical dissipation. *Journal of Computational Physics*, 162(1):33–81, 2000.

[101] Olgierd C. Zienkiewicz, Robert L. Taylor, and Perumal Nithiarasu. *The finite element method for fluid dynamics*. Butterworth-Heinemann, 6 edition, 2005.

APPENDICES

Appendix A

Dense-norm GSBP operators for the first derivative

A.1 Introduction

In this appendix, we prove the various theorems for dense-norm operators presented in Section 4.3. What is presented here is a further develop the theory of dense-norm GSBP operators presented in Ref. 21. The theory of dense-norm CSBP operators was previously developed by Strand [87], while Carpenter and Gottlieb [13] have shown how to construct dense-norm GSBP operators of maximum degree on nearly arbitrary nodal distributions that have nodes at the boundaries (see Section 4.4). Our approach is to show that symmetric matrices can be constructed that satisfy the compatibility equations. Next, we characterize the constituent matrices of dense-norm GSBP operators as approximations to various bilinear forms. Finally, we prove that given a nodal distribution with n unique nodes, dense-norm GSBP operators of order $[0, n-1]$ always exist and derive the conditions under which such operators can be constructed such that the GSBP norm is associated with a given quadrature rule.

A.2 Theory of dense-norm GSBP operators

Consider the $n \times n$ compatibility matrix, C, defined by

$$\mathsf{C}_{ij} = j \left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{j-1} + i \left(\mathbf{x}^j\right)^{\mathrm{T}} \mathsf{H}\mathbf{x}^{i-1} - \left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{E}\mathbf{x}^j, \ i, j \in [0, n-1], \tag{A.1}$$

where we use the notation $\mathsf{C}_{ij} = \mathsf{C}(i, j)$. If a particular (i, j) entry in C is zero, then the SBP norm, H, satisfies that particular compatibility equation. The compatibility matrix can be constructed directly as

$$\mathsf{C} = \mathsf{X}^{\mathrm{T}}\mathsf{H}\tilde{\mathsf{X}} + \tilde{\mathsf{X}}^{\mathrm{T}}\mathsf{H}\mathsf{X} - \mathsf{X}^{\mathrm{T}}\mathsf{E}\mathsf{X}, \tag{A.2}$$

since $i\left(\mathbf{x}^j\right)^{\mathrm{T}}\mathsf{H}\mathbf{x}^{i-1}$ is a scalar; therefore, $i\left(\mathbf{x}^j\right)^{\mathrm{T}}\mathsf{H}\mathbf{x}^{i-1} = i\left(\mathbf{x}^{i-1}\right)^{\mathrm{T}}\mathsf{H}\mathbf{x}^j$. The following definitions are useful:

Definition 8. The compatibility sub-matrix of a GSBP operator for the first derivative of degree p refers to the sub-matrix of the compatibility matrix defined by $i, j \in [0, p]$.

Definition 9. The norm matrix is defined as

$$\mathsf{N} = \mathsf{X}^{\mathrm{T}}\mathsf{H}\mathsf{X}. \tag{A.3}$$

Using (A.3), the compatibility matrix can be redefined as

$$\begin{aligned} \mathsf{C} &= \mathsf{X}^{\mathrm{T}}\left(\mathsf{X}^{\mathrm{T}}\right)^{-1}\mathsf{N}\mathsf{X}^{-1}\tilde{\mathsf{X}} + \tilde{\mathsf{X}}^{\mathrm{T}}\left(\mathsf{X}^{\mathrm{T}}\right)^{-1}\mathsf{N}\mathsf{X}^{-1}\mathsf{X} - \mathsf{X}^{\mathrm{T}}\mathsf{E}\mathsf{X} \\ &= \mathsf{N}\mathsf{X}^{-1}\tilde{\mathsf{X}} + \left(\mathsf{X}^{-1}\tilde{\mathsf{X}}\right)^{\mathrm{T}}\mathsf{N} - \mathsf{X}^{\mathrm{T}}\mathsf{E}\mathsf{X}. \end{aligned} \tag{A.4}$$

The following proposition is necessary in what follows:

Proposition 3. *The product* $\mathsf{X}^{-1}\tilde{\mathsf{X}} =: \mathsf{S}$ *has the following structure:*

$$\mathsf{X}^{-1}\tilde{\mathsf{X}} = \begin{bmatrix} 0 & 1 & & & & \\ & 0 & 2 & & & \\ & & \ddots & \ddots & & \\ & & & 0 & (n-2) & \\ & & & & 0 & \end{bmatrix}. \tag{A.5}$$

Proof. From the definition of the inverse $\mathsf{X}^{-1}\mathsf{X} = \mathsf{I}$, where I is an identity matrix, the product of the i^{th} row of X^{-1}, defined here as $\mathbf{x}_i^{-1}$, with the j^{th} column of X satisfies

$$\mathbf{x}_i^{-1}\mathbf{x}^j = \delta_{ij} = \begin{cases} 1, \text{ if } i = j, \\ 0, \text{ if } i \neq j \end{cases}. \tag{A.6}$$

Thus,

$$\mathsf{S}_{ij} = \left(\mathsf{X}^{-1}\tilde{\mathsf{X}}\right)_{ij} = j\mathbf{x}_i^{-1}\mathbf{x}^{j-1} = j\delta_{i(j-1)} = \begin{cases} j, \text{ if } i = j-1, \\ 0, \text{ if } i \neq j-1 \end{cases}, \; i, j \in [0, n-1]. \tag{A.7}$$

Note that for $j < 0$, $\mathbf{x}^j$ is defined as $\mathbf{x}^j = 0$. □

Using the definition of S, the compatibility matrix can now be recast as:

$$\mathsf{C} = \mathsf{N}\mathsf{S} + \mathsf{S}^{\mathrm{T}}\mathsf{N} - \mathsf{X}^{\mathrm{T}}\mathsf{E}\mathsf{X}. \tag{A.8}$$

The matrix S has some useful properties given in the following proposition:

Proposition 4. *The matrices* S *and* S^{T} *have the properties*

$$(\mathsf{WS})_{ij} = \begin{cases} j\mathsf{W}_{i(j-1)}, \ \textit{if } j > 0 \\ \\ 0 \ \textit{if } j = 0 \end{cases}, \ \textit{and} \ \left(\mathsf{S}^{\mathrm{T}}\mathsf{W}\right)_{ij} = \begin{cases} i\mathsf{W}_{(i-1)j}, \ \textit{if } i > 0 \\ \\ 0 \ \textit{if } i = 0 \end{cases}. \tag{A.9}$$

Proof. By Proposition 3, consider that $\mathsf{S}_{ij} = j\delta_{i(j-1)}$; thus,

$$(\mathsf{WS})_{ij} = \sum_{k=1}^{n} \mathsf{W}_{ik}\mathsf{S}_{kj} = \sum_{k=1}^{n} \mathsf{W}_{ik} j \delta_{k(j-1)} = j\mathsf{W}_{i(j-1)}. \tag{A.10}$$

For the second property, $\left(\mathsf{S}^{\mathrm{T}}\right)_{ij} = i\delta_{(i-1)j}$; thus,

$$\left(\mathsf{S}^{\mathrm{T}}\mathsf{W}\right)_{ij} = \sum_{k=1}^{n} \mathsf{S}^{\mathrm{T}}_{ik}\mathsf{W}_{kj} = \sum_{k=1}^{n} i\delta_{(i-1)k}\mathsf{W}_{kj} = i\mathsf{W}_{(i-1)j}. \tag{A.11}$$

□

By definition, the matrix E satisfies

$$\left(\mathbf{x}^i\right)^{\mathrm{T}} \mathsf{E}\mathbf{x}^j = x_{\mathrm{R}}^{i+j} - x_{\mathrm{L}}^{i+j} = (i+j)\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^{i+j-1}\mathrm{d}x \tag{A.12}$$

for $i, j \in [0, \tau_{\mathsf{E}} \geq p]$. The integration matrix, V, is defined as

$$\mathsf{V}_{ij} = \frac{x_{\mathrm{R}}^{i+j+1} - x_{\mathrm{L}}^{i+j+1}}{i+j+1} = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^{i+j}\mathrm{d}x. \tag{A.13}$$

The relationship between E and V is given in the following theorem:

Theorem A.1. *If* E *satisfies (A.12) for* $i, j \in [0, \tau_{\mathsf{E}}]$*, then*

$$\left(\mathsf{VS} + \mathsf{S}^{\mathrm{T}}\mathsf{V}\right)_{ij} = \left(\mathsf{X}^{\mathrm{T}}\mathsf{EX}\right)_{ij}, \ i, j \in [0, \tau_{\mathsf{E}}]. \tag{A.14}$$

Proof. For $i \neq j \neq 0$ by Proposition 4 and the definition of V, the LHS of (A.14) is

$$\begin{aligned} \left(\mathsf{VS} + \mathsf{S}^{\mathrm{T}}\mathsf{V}\right)_{ij} &= j\mathsf{V}_{i(j-1)} + i\mathsf{V}_{(i-1)j} \\ &= j\frac{\left(x_{\mathrm{R}}^{i+j} - x_{\mathrm{L}}^{i+j}\right)}{i+j} + i\frac{\left(x_{\mathrm{R}}^{i+j} - x_{\mathrm{L}}^{i+j}\right)}{i+j} = x_{\mathrm{R}}^{i+j} - x_{\mathrm{L}}^{i+j} = \left(\mathsf{X}^{\mathrm{T}}\mathsf{EX}\right)_{ij}. \end{aligned} \tag{A.15}$$

Finally, for $i = 0, j = 0$ by Proposition 4 and the definition of E,

$$\left(\mathsf{VS} + \mathsf{S}^{\mathrm{T}}\mathsf{V}\right)_{00} = 0 = \left(\mathsf{X}^{\mathrm{T}}\mathsf{EX}\right)_{00}. \tag{A.16}$$

□

An immediate consequence of Theorem A.1 is

Corollary 5. *If* E *satisfies (A.12) for* $i, j \in [0, n-1]$*, then by Theorem A.1*

$$\mathsf{VS} + \mathsf{S}^{\mathrm{T}}\mathsf{V} = \mathsf{X}^{\mathrm{T}}\mathsf{EX}, \tag{A.17}$$

and the compatibility matrix can, therefore, be recast as

$$\mathsf{C} = (\mathsf{N} - \mathsf{V})\,\mathsf{S} + \mathsf{S}^{\mathrm{T}}\,(\mathsf{N} - \mathsf{V})\,. \tag{A.18}$$

To construct a GSBP operator of degree p, the compatibility matrix must satisfy $C_{ij} = 0$ for $i, j \in [0, p]$; that is, the compatibility sub-matrix must be a matrix of zeros. The system of equations can be developed by using the properties of S given in Proposition 4, the form of the compatibility matrix given by (A.8), and Theorem A.1, which results in

$$\mathsf{C}_{ij} = \begin{cases} 0 & \text{if } i = j = 0 \\ j\mathsf{N}_{i(j-1)} - j\mathsf{V}_{i(j-1)} & \text{if } i = 0,\ j > 0 \\ i\mathsf{N}_{(i-1)j} - i\mathsf{V}_{(i-1)j} & \text{if } i = 0,\ i > 0 \\ j\mathsf{N}_{i(j-1)} - j\mathsf{V}_{i(j-1)} + i\mathsf{N}_{(i-1)j} - i\mathsf{V}_{(i-1)j} & \text{otherwise} \end{cases} \tag{A.19}$$

$$i, j \in [0, p].$$

An equivalent expression to (A.19) is

$$\mathsf{C}_{ij} = j\mathsf{N}_{i(j-1)} - j\mathsf{V}_{i(j-1)} + i\mathsf{N}_{(i-1)j} - i\mathsf{V}_{(i-1)j},\ i, j \in [0, p], \tag{A.20}$$

where the components of N and V are allowed to vary outside of $[0, n-1]$.

The goal is to determine similar relations for dense-norm operators as in the case of diagonal-norm GSBP operators. Before proceeding, it is necessary to discuss some terminology that simplifies the ensuing analysis.

Definition 10. The k^{th} anti-diagonal of a $p \times p$ symmetric matrix W is defined as those elements (we take 0 to be an even number)

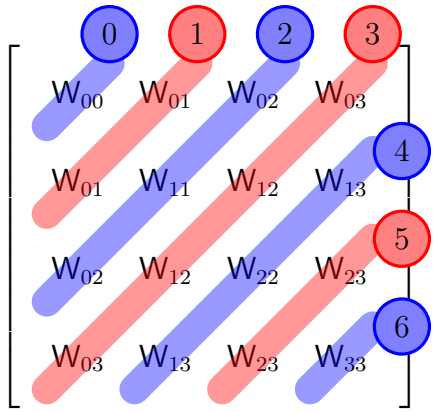


Figure A.1: Anti-diagonal numbering convention.

- Odd k:

$$\mathsf{W}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} = \mathsf{W}_{\left(\frac{k+1+2m}{2}\right)\left(\frac{k-1-2m}{2}\right)},$$

$$\text{for } 1 \leq k \leq p,\ m \in \left[0, \tfrac{k-1}{2}\right], \tag{A.21}$$

$$\text{for } p < k \leq 2p-1,\ m \in \left[0, p - \tfrac{(k+1)}{2}\right].$$

- Even k:

$$\mathsf{W}_{\left(\frac{k-2m}{2}\right)\left(\frac{k+2m}{2}\right)} = \mathsf{W}_{\left(\frac{k+2m}{2}\right)\left(\frac{k-2m}{2}\right)},$$

$$\text{for } 0 \leq k \leq p,\ m \in \left[0, \tfrac{k}{2}\right], \tag{A.22}$$

$$\text{for } p < k \leq 2p,\ m \in \left[0, p - \tfrac{k}{2}\right].$$

As an example, consider Figure A.1. Some important observations regarding anti-diagonals, for later use, are summarized in the following proposition:

Proposition 5. *Consider the k^{th} anti-diagonal of an $n \times n$ matrix, then*

- *$k \leq 2p-1$ for odd anti-diagonals that intersect the compatibility sub-matrix.*
- *$k \leq 2p$ for even anti-diagonals that intersect the compatibility sub-matrix.*
- *The k^{th} even anti-diagonal contains the element $\mathsf{C}_{\left(\frac{k}{2}\right)\left(\frac{k}{2}\right)}$.*

The analysis is based on examining the solution of the system of equations developed by equating the elements of the compatibility sub-matrix to zero for odd and even anti-diagonals. For an odd anti-diagonal, the following lemma gives an explicit form to the solution of this system of equations:

Lemma 4. *The solution of the system of equations developed by equating the elements of the compatibility matrix to zero along the k^{th} odd anti-diagonal for $1 \leq k \leq p$ is given as*

$$\mathsf{N}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)} = \mathsf{V}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)},\ m \in \left[0, \frac{k-1}{2}\right]. \tag{A.23}$$

and for $p < k \leq 2p-1$ is given as

$$\begin{aligned} \mathsf{N}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} &= \mathsf{V}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} \\ &\quad + (-1)^{m+1} \prod_{r=0}^{m} \tfrac{(k+1+2r)}{(k-1-2r)} \left(\mathsf{N}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)} - \mathsf{V}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)}\right), \\ &\quad m \in \left[0, p - \tfrac{(k+1)}{2}\right]. \end{aligned} \tag{A.24}$$

Proof. The proof is based on constructing a recurrence relation and then repeatedly applying it. The compatibility matrix is symmetric and therefore it is only necessary to consider the upper triangular portion of the compatibility sub-matrix. Moreover, by Proposition 5, odd anti-diagonals do not contain elements from the diagonal. Thus, the subset of the equations of interest are given by

$$\begin{gathered} \mathsf{C}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} = 0, \\ 1 \leq k \leq p,\ m \in \left[0, \tfrac{k-1}{2}\right],\ p < k \leq 2p-1,\ m \in \left[0, p - \tfrac{(k+1)}{2}\right], \end{gathered} \tag{A.25}$$

which by (A.20) becomes

$$\begin{gathered} 0 = \left(\tfrac{k+1+2m}{2}\right) \tilde{\mathsf{N}}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)} + \left(\tfrac{k-1-2m}{2}\right) \tilde{\mathsf{N}}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)}, \\ 1 \leq k \leq p,\ m \in \left[0, \tfrac{k-1}{2}\right],\ p < k \leq 2p-1,\ m \in \left[0, p - \tfrac{(k+1)}{2}\right], \end{gathered} \tag{A.26}$$

where $\tilde{\mathsf{N}}_{ij} = \mathsf{N}_{ij} - \mathsf{V}_{ij}$. Solving for $\tilde{\mathsf{N}}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)}$ in (A.26) gives

$$\begin{gathered} S_m = \tilde{\mathsf{N}}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} = -\tfrac{(k+1+2m)}{(k-1-2m)} \tilde{\mathsf{N}}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)}, \\ 3 \leq k \leq p,\ m \in \left[0, \tfrac{k-1}{2} - 1\right],\ p < k \leq 2p-1,\ m \in \left[0, p - \tfrac{(k+1)}{2}\right], \end{gathered} \tag{A.27}$$

where the first limit has been changed to avoid a division by zero; the equations that are dropped by doing so will be revisited shortly. By manipulating the subscripts of the RHS

term, (A.27) can be converted into the following recursion formula:

$$S_m \;=\; -\tfrac{(k+1+2m)}{(k-1-2m)} S_{m-1},\; 3 \le k \le p,\; m \in \left[0, \tfrac{k-1}{2} - 1\right],\; p < k \le 2p-1,\; m \in \left[0, p - \tfrac{(k+1)}{2}\right]. \tag{A.28}$$

Repeated application of (A.28) and solving for S_0 using (A.26) gives

$$S_m \;=\; -\tfrac{(k+1+2m)}{(k-1-2m)} \left[-\tfrac{(k+1+2(m-1))}{(k-1-2(m-1))}\right] \cdots \left[-\tfrac{(k+1+2)}{(k-1-2)}\right] S_0. \tag{A.29}$$

However, evaluating (A.26) for $m = 0$ gives that

$$S_0 = \tilde{\mathsf{N}}_{\left(\frac{k-3}{2}\right)\left(\frac{k+1}{2}\right)} = -\frac{(k+1)}{(k-1)} \tilde{\mathsf{N}}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)}, \tag{A.30}$$

thus,

$$\begin{aligned} \mathsf{N}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} &= \mathsf{V}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} \\ &+ (-1)^{m+1} \prod_{r=0}^{m} \tfrac{(k+1+2r)}{(k-1-2r)} \left(\mathsf{N}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)} - \mathsf{V}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)}\right), \end{aligned} \tag{A.31}$$

$$3 \le k \le p,\; m \in \left[0, \tfrac{k-1}{2} - 1\right],\; p < k \le 2p-1,\; m \in \left[0, p - \tfrac{(k+1)}{2}\right].$$

For $1 \le k \le p$, C_{0k} is within the compatibility sub-matrix; thus, setting $m = \frac{k-1}{2}$ in (A.26) gives

$$\mathsf{C}_{0k} = 0 = k\tilde{\mathsf{N}}_{0(k-1)},\; 1 \le k \le p \tag{A.32}$$

and therefore

$$\mathsf{N}_{0(k-1)} = \mathsf{V}_{0(k-1)},\; 1 \le k \le p. \tag{A.33}$$

On the other hand, setting $m = \frac{k-1}{2} - 1$ in (A.31) gives

$$\begin{aligned} \mathsf{N}_{0(k-1)} &= \mathsf{V}_{0(k-1)} + (-1)^{\frac{k-1}{2}} \prod_{r=0}^{\frac{k-1}{2}-1} \tfrac{(k+1+2r)}{(k-1-2r)} \left(\mathsf{N}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)} - \mathsf{V}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)}\right) \\ &3 \le k \le p, \end{aligned} \tag{A.34}$$

which, when combined with (A.33), implies that

$$\mathsf{N}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)} = \mathsf{V}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)},\; 1 \le k \le p. \tag{A.35}$$

However, by (A.31), (A.35) implies that

$$\mathsf{N}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} = \mathsf{V}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)},\ 3 \leq k \leq p,\ m \in \left[0, \frac{k-1}{2} - 1\right]. \tag{A.36}$$

Shifting the m index in (A.36) and combining the result with (A.35) gives

$$\mathsf{N}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)} = \mathsf{V}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)},\ 1 \leq k \leq p,\ m \in \left[0, \frac{k-1}{2}\right]. \tag{A.37}$$

Finally, for $p < k \leq 2p-1$, $\mathsf{C}_{0(k-1)}$ is not within the compatibility sub-matrix and the solution is given by (A.31); in other words, (A.24) has been proven. □

The following lemma is now proven:

Lemma 5. *The solution of the equations developed, by equating the elements of the i^{th} even anti-diagonal that are contained within the compatibility sub-matrix to zero, is unique and given by*

$$\begin{gathered}\mathsf{N}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)} = \mathsf{V}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)}, \\ 1 \leq k \leq p,\ m \in \left[0, \tfrac{k}{2} - 1\right],\ p < k \leq 2p,\ m \in \left[0, p - \tfrac{k}{2}\right]\end{gathered} \tag{A.38}$$

Proof. The proof follows the same path as for odd anti-diagonals. The equations that need to be satisfied are given as

$$\begin{aligned}\mathsf{C}_{\left(\frac{k-2m}{2}\right)\left(\frac{k+2m}{2}\right)} &= 0 \\ &= \left(\tfrac{k+2m}{2}\right)\tilde{\mathsf{N}}_{\left(\frac{k-2m}{2}\right)\left(\frac{k-2+2m}{2}\right)} + \left(\tfrac{k-2m}{2}\right)\tilde{\mathsf{N}}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)}, \\ &\quad 1 \leq k \leq p,\ m \in \left[0, \tfrac{k}{2}\right],\ p < k \leq 2p,\ m \in \left[0, p - \tfrac{k}{2}\right].\end{aligned} \tag{A.39}$$

Solving for $\tilde{\mathsf{N}}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)}$ in (A.39) results in

$$\begin{gathered}S_m = \tilde{\mathsf{N}}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)} = -\tfrac{(k+2m)}{(k-2m)}\tilde{\mathsf{N}}_{\left(\frac{k-2m}{2}\right)\left(\frac{k-2+2m}{2}\right)}, \\ 1 \leq k \leq p,\ m \in \left[0, \tfrac{k}{2} - 1\right],\ p < k \leq 2p,\ m \in \left[0, p - \tfrac{k}{2}\right],\end{gathered} \tag{A.40}$$

where the change in one of the limits has again been made to avoid a division by zero; the neglected equation is revisited later. Manipulating the subscript in the term on the RHS of

(A.40) gives the following recursion relation:

$$S_m = \tilde{\mathsf{N}}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)} = -\tfrac{(k+2m)}{(k-2m)} S_{m-1}, \qquad 1 \le k \le p,\ m \in \left[0, \tfrac{k}{2} - 1\right],\ p < k \le 2p,\ m \in \left[0, p - \tfrac{k}{2}\right], \tag{A.41}$$

Repeated application of (A.41) gives

$$S_m = -\tfrac{(k+2m)}{(k-2m)} \left[-\tfrac{(k+2(m-1))}{(k-2(m-1))}\right] \cdots \left[-\tfrac{(k+2)}{(k-2)}\right] S_0 \qquad 1 \le k \le p,\ m \in \left[0, \tfrac{k}{2} - 1\right],\ p < k \le 2p,\ m \in \left[0, p - \tfrac{k}{2}\right]. \tag{A.42}$$

Evaluating (A.39) for $m = 0$ gives

$$0 = \tfrac{k}{2}\tilde{\mathsf{N}}_{\left(\frac{k}{2}\right)\left(\frac{k-2}{2}\right)} + \tfrac{k}{2}\tilde{\mathsf{N}}_{\left(\frac{k-2}{2}\right)\left(\frac{k}{2}\right)} \qquad 1 \le k \le p,\ m \in \left[0, \tfrac{k}{2}\right],\ p < k \le 2p,\ m \in \left[0, p - \tfrac{k}{2}\right]. \tag{A.43}$$

Solving for $\tilde{\mathsf{N}}_{\left(\frac{k}{2}\right)\left(\frac{k-2}{2}\right)}$ in (A.43), we have that

$$S_0 = \tfrac{k}{2}\tilde{\mathsf{N}}_{\left(\frac{k}{2}\right)\left(\frac{k-2}{2}\right)} = -\tfrac{k}{k}\tilde{\mathsf{N}}_{\left(\frac{k-2}{2}\right)\left(\frac{k}{2}\right)} \qquad 1 \le k \le p,\ m \in \left[0, \tfrac{k}{2}\right],\ p < k \le 2p,\ m \in \left[0, p - \tfrac{k}{2}\right]. \tag{A.44}$$

Therefore, (A.41) with rearrangement gives

$$\begin{aligned} \mathsf{N}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)} \;=\; & \mathsf{V}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)} \\ & -\tfrac{(k+2m)}{(k-2m)} \left[-\tfrac{(k+2(m-1))}{(k-2(m-1))}\right] \cdots \left[-\tfrac{(k)}{(k)}\right] \left(\mathsf{N}_{\left(\frac{k-2}{2}\right)\left(\frac{k}{2}\right)} - \mathsf{V}_{\left(\frac{k-2}{2}\right)\left(\frac{k}{2}\right)}\right) \end{aligned} \tag{A.45}$$

$$1 \le k \le p,\ m \in \left[0, \tfrac{k}{2} - 1\right],\ p < k \le 2p,\ m \in \left[0, p - \tfrac{k}{2}\right].$$

For $m = 0$, (A.45) implies that $\mathsf{N}_{\left(\frac{k-2}{2}\right)\left(\frac{k}{2}\right)} = \mathsf{V}_{\left(\frac{k-2}{2}\right)\left(\frac{k}{2}\right)}$; this implies that

$$\mathsf{N}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)} = \mathsf{V}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)}, \qquad 1 \le k \le p,\ m \in \left[0, \tfrac{k}{2} - 1\right],\ p < k \le 2p,\ m \in \left[0, p - \tfrac{k}{2}\right], \tag{A.46}$$

which is the desired result. Furthermore, setting $m = \frac{k}{2}$ in (A.39) and rearranging results in

$$\mathsf{N}_{0(k-1)} = \mathsf{V}_{0(k-1)}, \tag{A.47}$$

which does not give an additional equation. Therefore, (A.38) has been proven. □

The consequences of Lemma 4 and Lemma 5 are summarized in the following lemma:

Lemma 6. *The system of equations developed, by equating the entries of the CMS to zero has the following solution:*

- *It is fully determined for the k^{th} odd or even anti-diagonal contained within the compatibility sub-matrix, i.e., $k \leq p$, and has solution*
  - *Odd:*

$$\mathsf{N}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)} = \mathsf{V}_{\left(\frac{k-1-2m}{2}\right)\left(\frac{k-1+2m}{2}\right)},\ m \in \left[0, \frac{k-1}{2}\right]. \tag{A.48}$$

  - *Even:*

$$\begin{gathered} \mathsf{N}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)} = \mathsf{V}_{\left(\frac{k-2-2m}{2}\right)\left(\frac{k+2m}{2}\right)}, \\ 1 \leq k \leq p,\ m \in \left[0, \tfrac{k}{2} - 1\right]; \end{gathered} \tag{A.49}$$

- *Any equations developed from elements on an even anti-diagonal have a unique solution, as given above, where for $p < k \leq 2p$, the limits are given as $m \in \left[0, p - \frac{k}{2}\right]$; and*
- *The system of equations developed from the elements of the k^{th} odd anti-diagonal not fully contained within the compatibility sub-matrix is under-determined. Moreover, a solution exists and for $p < k \leq 2p - 1$ is given as*

$$\begin{aligned} \mathsf{N}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} \ &= \ \mathsf{V}_{\left(\frac{k-3-2m}{2}\right)\left(\frac{k+1+2m}{2}\right)} \\ &\quad + (-1)^{m+1} \prod_{r=0}^{m} \tfrac{(k+1+2r)}{(k-1-2r)} \left(\mathsf{N}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)} - \mathsf{V}_{\left(\frac{k-1}{2}\right)\left(\frac{k-1}{2}\right)}\right), \\ &\quad m \in \left[0, p - \tfrac{(k+1)}{2}\right]. \end{aligned} \tag{A.50}$$

Proof. We have already proven the above result for the odd and even anti-diagonals taken individually. What remains is to prove that the systems of equations from each anti-diagonal are independent. This is easily seen by adding the subscripts of the elements of N given by the above solution and noting that the resultant number is unique, therefore, each system of equations has an independent set of elements from N. □

Lemma 6 gives explicit formulas for the solution to the compatibility equations and thereby proves that it is always possible to construct a symmetric matrix that satisfies the compatibility equations. It is still necessary to prove that the resultant set of matrices contains members that are positive definite. However, first we explore the properties of the symmetric matrices that satisfy the compatibility equations.

Theorem A.2. *A dense-norm GSBP operator,* $\mathsf{D} = \mathsf{H}^{-1}\mathsf{Q}$*, of degree* p*, has a norm* H *that is a degree* $\tau_{\mathsf{H}} \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$ *approximation to the* L_2 *inner product*

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathcal{V}\mathcal{U}\mathrm{d}x. \tag{A.51}$$

Proof. The norm matrix of a GSBP operator must satisfy the compatibility equations, therefore, the entries of H are given by Lemma 6. First, let us consider when p is even. The first p anti-diagonals of the CMS are uniquely determined by equations (A.48) and (A.49). By adding the subscripts in N_{ij} in both of these we see that $\mathsf{N}_{ij} = \mathsf{V}_{ij}$ for $i + j \leq p - 1$. Since $\mathsf{X}^T\mathsf{H}\mathsf{X} = \mathsf{N}$ and

$$\mathsf{V}_{ij} = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^{i+j}\mathrm{d}x \tag{A.52}$$

this implies that

$$\mathbf{x}_i^{\mathrm{T}}\mathsf{H}\mathbf{x}_j = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^{i+j}\mathrm{d}x,\ i + j \leq p - 1, \tag{A.53}$$

which means that H is a degree $p - 1$ approximation to the L_2 inner product.

Now consider odd p: As before, the first p anti-diagonals of the CMS have a unique solution and so H is at least a degree $p-1$ approximation to the L_2 inner product. However, the first anti-diagonal not fully contained within the CMS is even and it has a unique solution, as given in Lemma 6. Furthermore, we can see that this anti-diagonal equates $\mathsf{N}_{ij} = \mathsf{V}_{ij}$ for $i + j = p$, this means that the p anti-diagonal of N and V are equal and therefore, H is a degree p approximation to the L_2 inner product. □

Theorem A.2 leads to the following corollary:

Corollary 6. *The norm* H *of a dense-norm GSBP operator is associated with a degree* $\tau \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$ *quadrature rule.*

Proof. By Theorem A.2,

$$\mathbf{1}^{\mathrm{T}}\mathsf{H}\mathbf{u} \tag{A.54}$$

is a degree $\tau \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$ approximation to

$$\int_{x_\mathrm{L}}^{x_\mathrm{R}} \mathcal{U} \mathrm{d}x. \tag{A.55}$$

By taking $w_i = \mathbf{1}^\mathrm{T} \mathsf{H}_{:i}$, the following quadrature rule of degree $\tau \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$ is developed:

$$\sum_{i=0}^{n-1} w_i \mathcal{U}(x_i). \tag{A.56}$$

□

As with diagonal-GSBP operators, Q can be characterized as follows:

Theorem A.3. *A dense-norm GSBP operator, $\mathsf{D}_1 = \mathsf{H}^{-1}\mathsf{Q}$, of degree p, has Q that is a degree $\tau_\mathsf{Q} \geq \min\left(\tau_\mathsf{E}, 2\left\lfloor \frac{p-1}{2} \right\rfloor + 2\right)$ approximation to the bilinear form*

$$(\mathcal{V}, \mathcal{U}) = \int_{x_\mathrm{L}}^{x_\mathrm{R}} \mathcal{V} \frac{\partial \mathcal{U}}{\partial x} \mathrm{d}x. \tag{A.57}$$

Proof. Multiplying the degree equations (3.3) by $\left(\mathbf{x}^i\right)^\mathrm{T} \mathsf{H}$ gives

$$\left(\mathbf{x}^i\right)^\mathrm{T} \mathsf{Q}\mathbf{x}^j = j \left(\mathbf{x}^i\right)^\mathrm{T} \mathsf{H}\mathbf{x}^{j-1}, \ i, j \in [0, p]. \tag{A.58}$$

By Theorem A.2, if p is even, H is a $p-1$ approximation to the L_2 inner product; therefore, (A.58) becomes

$$\left(\mathbf{x}^i\right)^\mathrm{T} \mathsf{Q}\mathbf{x}^j = j \int_{x_\mathrm{L}}^{x_\mathrm{R}} x^i x^{j-1} \mathrm{d}x = \int_{x_\mathrm{L}}^{x_\mathrm{R}} x^i \frac{\partial x^j}{\partial x} \mathrm{d}x, \ i + j - 1 \leq p - 1, \ j \neq 0. \tag{A.59}$$

For $j = 0$ we have that

$$\left(\mathbf{x}^p\right)^\mathrm{T} \mathsf{Q}\mathbf{1} = 0, \tag{A.60}$$

therefore, for even p, Q is a degree p approximation to (A.57). Similarly, for odd p, by Theorem A.2, H is a degree p approximation to the L_2 inner product; therefore, (A.58) becomes

$$\left(\mathbf{x}^i\right)^\mathrm{T} \mathsf{Q}\mathbf{x}^j = j \int_{x_\mathrm{L}}^{x_\mathrm{R}} x^i x^{j-1} \mathrm{d}x = \int_{x_\mathrm{L}}^{x_\mathrm{R}} x^i \frac{\partial x^j}{\partial x} \mathrm{d}x, \ i + j - 1 \leq p, \ j \neq 0. \tag{A.61}$$

Relation (A.61) shows that Q is almost at least degree $p+1$ approximation to (A.57). The missing conditions are $i = p + 1$ and $j = 0$, which is automatically satisfied by consistency,

and $i = 0$ and $j = p + 1$, which gives

$$\mathbf{1}^{\mathrm{T}}\mathsf{Q}\mathbf{x}^{p+1} = \mathbf{1}^{\mathrm{T}}\left(\mathsf{E} - \mathsf{Q}^{\mathrm{T}}\right)\mathbf{x}^{p+1} = \mathbf{1}^{\mathrm{T}}\mathsf{E}\mathbf{x}^{p+1}. \tag{A.62}$$

If $\tau_{\mathsf{E}} \geq p + 1$ then we have that

$$\mathbf{1}^{\mathrm{T}}\mathsf{Q}\mathbf{x}^{p+1} = \mathbf{1}^{\mathrm{T}}\mathsf{E}\mathbf{x}^{p+1} = x_{\mathrm{R}}^{p+1} - x_{\mathrm{L}}^{p+1} = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} x^0 \frac{\partial x^{p+1}}{\partial x}\mathrm{d}x, \tag{A.63}$$

and Q is at least a degree $p + 1$ approximation to (A.57). □

Finally, $\mathsf{Q}^{(A)}$ is characterized in the following corollary:

Corollary 7. *A dense-norm GSBP operator, $\mathsf{D}_1 = \mathsf{H}^{-1}\left(\mathsf{Q}^A + \frac{1}{2}\mathsf{E}\right)$, of degree p, and E of degree $\tau_{\mathsf{E}} \geq p$, has $\mathsf{Q}^{(A)}$ that is a degree $\tau_{\mathsf{Q}^{(A)}} \geq \min\left(\tau_{\mathsf{E}}, 2\left\lfloor\frac{p-1}{2}\right\rfloor + 2\right)$ approximation to the bilinear form*

$$(\mathcal{V},\mathcal{U}) = \int_{x_L}^{x_R} \mathcal{V}\frac{\partial \mathcal{U}}{\partial x}\mathrm{d}x - \frac{1}{2}\oint \mathcal{V}\mathcal{U}\mathbf{n}\mathrm{d}s. \tag{A.64}$$

Proof. By Theorem A.3,

$$\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathsf{Q}\mathbf{x}^j = \int_{x_L}^{x_R} x^i\frac{\partial x^j}{\partial x}\mathrm{d}x,\ i + j \leq \min\left(\tau_{\mathsf{E}}, 2\left\lfloor\frac{p-1}{2}\right\rfloor + 2\right). \tag{A.65}$$

Expanding Q and solving for $\mathsf{Q}^{(A)}$ in (A.65) gives, by Theorem A.3,

$$\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathsf{Q}^{(A)}\mathbf{x}^j = \int_{x_L}^{x_R} x^i\frac{\partial x^j}{\partial x}\mathrm{d}x - \frac{1}{2}\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathsf{E}\mathbf{x}^j,\ i + j \leq \min\left(\tau_{\mathsf{E}}, 2\left\lfloor\frac{p-1}{2}\right\rfloor + 2\right). \tag{A.66}$$

Therefore,

$$\left(\mathbf{x}^i\right)^{\mathrm{T}}\mathsf{Q}^{(A)}\mathbf{x}^j = \int_{x_L}^{x_R} x^i\frac{\partial x^j}{\partial x}\mathrm{d}x - \frac{1}{2}\oint x^i x^j\mathbf{n}\mathrm{d}s,\ i + j \leq \min\left(\tau_{\mathsf{E}}, 2\left\lfloor\frac{p-1}{2}\right\rfloor + 2\right). \tag{A.67}$$

□

As with diagonal-norm operators, we can see that the individual components of the discrete version of IBP are higher-order approximations to the continuous counter parts.

Thus far, the discussion has centred around the properties of the constituent matrices of dense-norm GSBP operators. However, it is still not known whether such operators exist. To construct an existence proof requires showing that the sub-matrices of V are symmetric positive definite:

Theorem A.4. *The matrix defined by* $(\mathsf{W})_{ij} = (\mathsf{V})_{ij}$, $i, j \in [0, r]$, $r \leq n-1$ *is symmetric positive definite.*

Proof. By definition, since V is symmetric, so too is W. The proof that W is symmetric positive definite is based on the classical proof that a Hilbert matrix is symmetric positive definite and is taken from [87]; it is necessary to show that

$$\mathbf{v}^{\mathrm{T}}\mathsf{W}\mathbf{v} > 0, \ \forall \mathbf{v} \neq \mathbf{0}. \tag{A.68}$$

Expanding the LHS of (A.68) gives

$$\sum_{k=1}^{n}\sum_{m=1}^{n} v_k v_m \frac{x_{\mathrm{R}}^{k+m-1} - x_{\mathrm{L}}^{k+m-1}}{k+m-1} > 0. \tag{A.69}$$

Moreover,

$$\frac{x_{\mathrm{R}}^{k+m-1} - x_{\mathrm{L}}^{k+m-1}}{k+m-1} = \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} y^{k+m-2}\mathrm{d}y. \tag{A.70}$$

Substituting (A.70) into (A.69) gives

$$\sum_{k=1}^{n}\sum_{m=1}^{n} v_k v_m \int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} y^{k+m-2}\mathrm{d}y > 0; \tag{A.71}$$

thus,

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \sum_{k=1}^{n}\sum_{m=1}^{n} v_k v_m y^{k+m-2}\mathrm{d}y > 0. \tag{A.72}$$

This can be recast as

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} \mathbf{v}^T \mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{v}\mathrm{d}y > 0, \tag{A.73}$$

where $\mathbf{y}^{\mathrm{T}} = [y^0, \ldots, y^{n-1}]$. Making the substitution $p(y) = \mathbf{y}^{\mathrm{T}}\mathbf{v}$ results in

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{R}}} p^2(y)\mathrm{d}y \geq 0. \tag{A.74}$$

The equality in (A.74) implies that $p(y) = 0$, which cannot be the case, unless the monomials, $[y^0, \ldots, y^{n-1}]$ are linearly dependent. However, on a finite interval, the monomials are linearly independent, which results in the conclusion that $p(y) > 0$ if $\mathbf{v} \neq 0$, and finally, that W is symmetric positive definite. □

We are now in a position to prove that given a nodal distribution with n distinct nodes,

dense-norm GSBP operators of orders $[0, n-1]$ always exist.

Theorem A.5. *Given a nodal distribution,* $\mathbf{x}$*, then there exist dense-norm GSBP operators with degree* $p \in [0, n-1]$ *with a dense-norm* H *that is an approximation to the* L_2 *inner product of degree* $\tau_{\mathsf{H}} \geq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$*.*

Proof. By Theorem A.2, H exists such that the required compatibility equations are satisfied and H is an approximation of the L_2 inner product. Moreover, by the same arguments used in Theorem 4.4, there exists Q such that the accuracy equations are satisfied. What remains to be shown is that it is possible to make H symmetric positive definite at the same time.

We recall that $\mathsf{N} = \mathsf{X}^{\mathrm{T}}\mathsf{H}\mathsf{X}$; as proven in Theorem A.2, the entries of N are equal to those of V for $i+j \leq 2\left\lfloor \frac{p-1}{2} \right\rfloor + 1$. Furthermore, we note that the only free variables in the solution of the compatibility equations are some of the diagonal entries of the matrix N (see Lemma 6). Thus, the norm matrix N has the form

$$\mathsf{N} = \mathsf{X}^{\mathrm{T}}\mathsf{H}\mathsf{X} = \begin{bmatrix} \mathsf{V}\,(0:\gamma, 0:\gamma) & \mathbf{b}_{(\gamma+1)} & \dots & \mathbf{b}_p & \mathbf{z}_{p+1} & \dots & \mathbf{z}_{n-1} \\ \mathbf{b}_{(\gamma+1)}^{\mathrm{T}} & \mathsf{N}_{(\gamma+1)(\gamma+1)} & & & & & \\ \vdots & & \ddots & & & & \\ \mathbf{b}_p^{\mathrm{T}} & & & \mathsf{N}_{pp} & & & \\ \mathbf{z}_{p+1}^{\mathrm{T}} & & & & \mathsf{N}_{(p+1)(p+1)} & & \\ \vdots & & & & & \ddots & \\ \mathbf{z}_{n-1}^{\mathrm{T}} & & & & & & \mathsf{N}_{(n-1)(n-1)} \end{bmatrix}, \tag{A.75}$$

where $\gamma = \left\lfloor \frac{p-1}{2} \right\rfloor$. The vectors $\mathbf{b}$ are not arbitrary, but can contain elements of V and the diagonal elements N_{ii}; more will be said about this shortly. Furthermore, $\mathbf{b}_i$ are vectors with $i-1$ components and $\mathbf{z}_i = \left[\mathsf{N}_{0,i}, \dots, \mathsf{N}_{(i-1)i}\right]^{\mathrm{T}}$, where the entries are free variables.

In order to be symmetric positive definite, $\mathbf{v}^{\mathrm{T}}\mathsf{N}\mathbf{v} > 0$ for all $\mathbf{v}$ that are not identically zero. The procedure to enforce the symmetric positive definite condition on N is to sequentially choose the N_{ii} free parameters such that the sub-matrix with lower right-hand corner element N_{ii} is symmetric positive definite. The first step is demonstrated for (A.75), which makes it clear how to proceed. It is necessary that

$$\left[\mathbf{v}^{\mathrm{T}}, u\right] \begin{bmatrix} \mathsf{V}\,(0:\gamma, 0:\gamma) & \mathbf{b}_{(\gamma+1)} \\ \mathbf{b}(\gamma+1)^{\mathrm{T}} & \mathsf{N}_{(\gamma+1)(\gamma+1)} \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ u \end{bmatrix} > 0 \tag{A.76}$$

for all $\mathbf{v}$ and u, such that both are not simultaneously zero. By Theorem A.4, any submatrix of V of the form $\mathsf{V}(0:j, 0:j)$ is symmetric positive definite; thus, $\mathsf{V}(0:\gamma, 0:\gamma)$ is symmetric positive definite and has decomposition

$$\mathsf{V}(0:\gamma, 0:\gamma) = \mathsf{L}^{\mathrm{T}}\Lambda\mathsf{L}, \tag{A.77}$$

where L is unitriangular and is, therefore, invertible, and $\Lambda > 0$ and is diagonal. Expanding (A.76) gives

$$\mathbf{v}^{\mathrm{T}}\mathsf{L}^{\mathrm{T}}\Lambda\mathsf{L}\mathbf{v} + 2u\mathbf{b}^{\mathrm{T}}\mathbf{v} + \mathsf{N}_{(\gamma+1)(\gamma+1)}u^2. \tag{A.78}$$

Defining $\tilde{\mathbf{v}} = \mathsf{L}\mathbf{v}$ and, therefore, $\mathbf{v} = \mathsf{L}^{-1}\mathbf{v}$, and $\tilde{\mathbf{b}} = \mathsf{L}^{\mathrm{T}}\mathbf{b}$ (A.78) becomes

$$\tilde{\mathbf{v}}^{\mathrm{T}}\Lambda\tilde{\mathbf{v}} + 2u\tilde{\mathbf{b}}^{\mathrm{T}}\tilde{\mathbf{v}} + \mathsf{N}_{(\gamma+1)(\gamma+1)}u^2. \tag{A.79}$$

Expanding (A.79) into components results in

$$\sum_{i=0}^{\gamma} \lambda_i\tilde{v}_i^2 + 2u\tilde{b}_i\tilde{v}_i + \mathsf{N}_{(\gamma+1)(\gamma+1)}u^2. \tag{A.80}$$

Completing the square on (A.80), with $\Gamma_i = \frac{\tilde{b}_i}{\lambda_i}$, results in

$$\sum_{i=0}^{\gamma} \lambda_i\left(\tilde{v}_i + \Gamma_i u\right)^2 + \left(\mathsf{N}_{(\gamma+1)(\gamma+1)} - \sum_{i=0}^{\gamma}\lambda_i\Gamma_i^2\right)u^2, \tag{A.81}$$

and by choosing $\mathsf{N}_{(\gamma+1)(\gamma+1)} > \sum_{i=0}^{\gamma}\lambda_i\Gamma_i^2$, (A.81) is > 0 and it has been proven that $\mathsf{N}_{(\gamma+1)(\gamma+1)}$ can be chosen such that $\mathsf{N}(0:(\gamma+1), 0:(\gamma+1))$ is symmetric positive definite. This same procedure is then applied to specify the remaining N_{ii} such that the resultant N is symmetric positive definite.

The above procedure works under the assumption that the diagonal elements N_{ii} are specified before they appear in the $\mathbf{b}$ vectors. From Lemma 6, we can see that the only free parameters in the solution of the compatibility equations are the diagonal elements of N. Moreover, any entries of N that depend on a element along the diagonal of N appears in a $\mathbf{b}$ vector to the right or below the diagonal element. Therefore, the N_{ii} can be sequentially set such that N is symmetry positive definite. We have only discussed the sub-matrix that includes all of the $\mathbf{b}$ vectors, however, the same process can be apple to the rest of the matrix; for example, the $\mathbf{z}$ vectors can be chosen to be vectors of zeros and the remaining diagonal entries of N to be positive numbers. □

Now we know that dense-norm GSBP operators of all degrees exist. However, we would like a procedure to construct H from quadrature rules, in analogy with diagonal-norm operators. This is the topic of the next theorem.

Theorem A.6. *Given a quadrature,* $\mathbf{w} = [w_0, w_1, \ldots, w_{n-1}]^{\mathrm{T}}$*, of degree* τ*, it is always possible to construct a dense-norm GSBP operator with* $\mathsf{H}\mathbf{1} = \mathbf{w}$ *of odd degree* $p \leq \tau$ *and* $p \leq n-1$*. Furthermore, if* τ *is odd, it is possible to construct an even-degree dense-norm GSBP operator, such that* $p = \tau + 1$*, for* $p \leq n-1$*, if the coefficient of the truncation term of the quadrature rule satisfies*

$$\mathsf{V}_{\left(\frac{p}{2}\right)\left(\frac{p}{2}\right)} + \frac{\mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w} - \mathsf{V}_{0p}}{(-1)^{\frac{p}{2}} \prod_{r=0}^{\frac{p-2}{2}} \frac{(p+2+2r)}{(p-2r)}} > \sum_{i=0}^{\frac{p-2}{2}} \lambda_i \Gamma_i^2. \tag{A.82}$$

Proof. The GSBP dense-norm, H, needs to satisfy the compatibility equations for $i, j \in [0, p]$, be symmetric positive definite, and satisfy $\mathsf{H}\mathbf{1} = \mathbf{w}$. Consider H as in Theorem A.5, which satisfies the compatibility equations,

$$\mathsf{X}^{\mathrm{T}}\mathsf{H}\mathsf{X} = \begin{bmatrix} \mathsf{V}\,(0:\gamma, 0:\gamma) & \mathbf{b}_{(\gamma+1)} & \ldots & \mathbf{b}_p & \mathbf{z}_{p+1} & \ldots & \mathbf{z}_{n-1} \\ \mathbf{b}_{(\gamma+1)}^{\mathrm{T}} & \mathsf{z}_{(\gamma+1)(\gamma+1)} & & & & & \\ \vdots & & \ddots & & & & \\ \mathbf{b}_p^{\mathrm{T}} & & & \mathsf{N}_{pp} & & & \\ \mathbf{z}_{p+1}^{\mathrm{T}} & & & & \mathsf{N}_{(p+1)(p+1)} & & \\ \vdots & & & & & \ddots & \\ \mathbf{z}_{n-1}^{\mathrm{T}} & & & & & & \mathsf{N}_{(n-1)(n-1)} \end{bmatrix}, \tag{A.83}$$

where $\gamma = \left\lfloor \frac{p-1}{2} \right\rfloor$. The $\mathbf{b}$ vectors are not arbitrary and it will be necessary to be more specific about some of their entries. First, it is proven that the elements of $\mathbf{b}$ can be chosen such that $\mathsf{H}\mathbf{1} = \mathbf{w}$, for $p \leq \tau$, $p \leq n-1$, for odd p. A symmetric positive definite H can then be constructed as in Theorem A.5. For odd τ, it is proven that an H for a GSBP operator of degree $p = \tau + 1$, $p \leq n-1$, can be constructed if the truncation coefficient of the quadrature rule satisfies inequality (A.82). If the inequality (A.82) is satisfied, then a symmetric positive definite H can be constructed as in Theorem A.5. First, it is necessary to satisfy

$$\mathsf{H}\mathbf{1} = \mathbf{w}. \tag{A.84}$$

Left multiplying (A.84) by X^{T} and using the fact that $\mathsf{X}\mathsf{X}^{-1}\mathbf{1} = \mathbf{1}$ results in

$$\mathsf{X}^{\mathrm{T}}\mathsf{H}\mathsf{X}\mathsf{X}^{-1}\mathbf{1} = \mathsf{X}^{\mathrm{T}}\mathbf{w}. \tag{A.85}$$

Inserting (A.83) and using the fact that $\mathsf{X}^{-1}\mathbf{1} = \mathbf{e}_0 = [1, 0, \ldots, 0]^{\mathrm{T}}$ results in

$$\begin{bmatrix} \mathsf{V}(0:\gamma, 0:\gamma) & \mathbf{b}_{(\gamma+1)} & \cdots & \mathbf{b}_p & \mathbf{N}_{p+1} & \cdots & \mathbf{N}_{n-1} \\ \mathbf{b}_{(\gamma+1)}^{\mathrm{T}} & \mathsf{N}_{(\gamma+1)(\gamma+1)} & & & & & \\ \vdots & & \ddots & & & & \\ \mathbf{b}_p^{\mathrm{T}} & & & \mathsf{N}_{pp} & & & \\ \mathbf{N}_{p+1}^{\mathrm{T}} & & & & \mathsf{N}_{(p+1)(p+1)} & & \\ \vdots & & & & & \ddots & \\ \mathbf{N}_{n-1}^{\mathrm{T}} & & & & & & \mathsf{N}_{(n-1)(n-1)} \end{bmatrix} \mathbf{e}_0 = \mathsf{X}^{\mathrm{T}}\mathbf{w}. \tag{A.86}$$

To prove the first part of the theorem, consider odd p such that $p \leq \tau$. By definition,

$$\mathsf{X}^{\mathrm{T}}\mathbf{w} = \begin{bmatrix} \mathsf{V}_{00} & \mathsf{V}_{01} & \ldots & \mathsf{V}_{0(p)} & \mathsf{V}_{0(p+1)} & \ldots & \mathsf{V}_{0(\tau)} & \mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w} & \ldots & \mathbf{x}_{n-1}^{\mathrm{T}}\mathbf{w} \end{bmatrix}^{\mathrm{T}}. \tag{A.87}$$

Expanding (A.86), and remembering that for odd p, by Corollary 6, H must be at least associated with a quadrature rule of degree $\tau = p$, gives

$$\begin{aligned} &\begin{bmatrix} \mathsf{V}_{00} & \mathsf{V}_{01} & \ldots & \mathsf{V}_{0(p)} & \mathsf{N}_{0(p+1)} & \ldots & \mathsf{N}_{0(\tau)} & \mathsf{N}_{0(\tau+1)} & \ldots & \mathsf{N}_{0(n-1)} \end{bmatrix}^{\mathrm{T}} = \\ &\begin{bmatrix} \mathsf{V}_{00} & \mathsf{V}_{01} & \ldots & \mathsf{V}_{0(p)} & \mathsf{V}_{0(p+1)} & \ldots & \mathsf{V}_{0(\tau)} & \mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w} & \ldots & \mathbf{x}_{n-1}^{\mathrm{T}}\mathbf{w} \end{bmatrix}^{\mathrm{T}}. \end{aligned} \tag{A.88}$$

Thus, for $\mathsf{H}\mathbf{1} = \mathbf{w}$, $\mathsf{N}_{0i} = \mathsf{V}_{0i}$, for $i \in [p+1, \tau]$ and $\mathsf{N}_{0i} = \mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w}$ for $i \in [\tau+1, n-1]$. To construct a symmetric positive definite H, the same steps as in Theorem A.5 are used.

To prove the second part of the theorem, we consider odd τ and $p = \tau + 1$. This implies that p is even and by Corollary 6 H must be at least associated with a quadrature rule of

degree $p-1$. Therefore, expanding (A.86) gives

$$\begin{aligned}\begin{bmatrix} \mathsf{V}_{00} & \mathsf{V}_{01} & \dots & \mathsf{V}_{0(p-1)} & \mathbf{b}_p(0) & \mathsf{N}_{0(p+1)} & \dots & \mathsf{N}_{0(n-1)} \end{bmatrix}^{\mathrm{T}} = \\ \begin{bmatrix} \mathsf{V}_{00} & \mathsf{V}_{01} & \dots & \mathsf{V}_{0(\tau)} & \mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w} & \mathbf{x}_{\tau+2}^{\mathrm{T}}\mathbf{w} & \dots & \mathbf{x}_{n-1}^{\mathrm{T}}\mathbf{w} \end{bmatrix}^{\mathrm{T}}.\end{aligned} \tag{A.89}$$

Now, for H to satisfy $\mathsf{H}\mathbf{1} = \mathbf{w}$, $\mathbf{b}_p(0) = \mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w}$ and $\mathsf{N}_{0(p+i-1)} = \mathbf{x}_{\tau+i}^{\mathrm{T}}\mathbf{w}$ for $i \in [2, n-1-\tau]$. However, because of the condition $\mathbf{b}_p(0) = \mathsf{V}_{0(\tau)}$, the steps used to construct a PD H used in Theorem A.5 cannot immediately be applied. This results because $\mathbf{b}_p(0)$ contains the diagonal element of N, used to enforce the PD condition. In particular, $\mathbf{b}_p(0)$ corresponds to the $p+1$ odd anti-diagonal and by Lemma 6,

$$\mathbf{b}_p(0) = \mathsf{N}_{0p} = \mathsf{V}_{0p} + (-1)^{\frac{p}{2}} \prod_{r=0}^{\frac{p-2}{2}} \frac{(p+2+2r)}{(p-2r)} \left(\mathsf{N}_{(\frac{p}{2})(\frac{p}{2})} - \mathsf{V}_{(\frac{p}{2})(\frac{p}{2})} \right). \tag{A.90}$$

Therefore,

$$\mathsf{N}_{(\frac{p}{2})(\frac{p}{2})} = \mathsf{V}_{(\frac{p}{2})(\frac{p}{2})} + \frac{\mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w} - \mathsf{V}_{0p}}{(-1)^{\frac{p}{2}} \prod_{r=0}^{\frac{p-2}{2}} \frac{(p+2+2r)}{(p-2r)}}. \tag{A.91}$$

Now it is clear that the problem is that it is no longer possible to choose $\mathsf{N}_{(\frac{p}{2})(\frac{p}{2})}$ to force the sub-matrix of N, $i, j \in \left[0, \frac{p}{2}\right]$ to be PD. This implies a condition on the truncation error, $\mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w}$, which using the notation in Theorem A.5 is given as

$$\mathsf{V}_{(\frac{p}{2})(\frac{p}{2})} + \frac{\mathbf{x}_{\tau+1}^{\mathrm{T}}\mathbf{w} - \mathsf{V}_{0p}}{(-1)^{\frac{p}{2}} \prod_{r=0}^{\frac{p-2}{2}} \frac{(p+2+2r)}{(p-2r)}} > \sum_{i=0}^{\frac{p-2}{2}} \lambda_i \Gamma_i^2. \tag{A.92}$$

□

Appendix B

Periodic Simultaneous Approximation Terms

Here we discuss the imposition of boundary conditions for periodic problems and further discuss interface SATs for such problems, where the focus is on the linear convection equation. Our interest is to show that using the GSBP-SAT approach we can construct approximations to periodic problems that mimic the eigenspectrum of the continuous problem. That is to say, for periodic problems the Fourier transform of the first derivative gives an imaginary eigenvalue. In the same way, using centred FD approximations results in a circulant matrix with imaginary eigenvalues. In practical computations, we do not use the set of penalty parameters used in this appendix. Nevertheless, what we highlight here, besides the aforementioned, is that the penalty parameters can be tuned to achieve additional objectives besides stability and conservation.

To impose periodic boundary conditions, the boundary SAT at the first element is replaced with an interface SAT coupling the left boundary of the first element to the right boundary of the last element. We now prove that with a specific choice for the value of the penalty parameters, the resultant spatial operator has zero or imaginary eigenvalues. We do so by showing that the resultant spatial operator is the product of a symmetric positive-definite matrix and a skew-symmetric matrix. Such a product is guaranteed to have eigenvalues with zero real parts, thus, if we can show that the spatial operator is the product of a symmetric positive-definite matrix and a skew-symmetric matrix, then the eigenvalues have zero real parts. Now we can prove the following:

Theorem B.1. *The discretization of $-a\frac{\partial \mathcal{U}}{\partial x}$ with GSBP operators and SATs using N elements for a periodic problem has purely imaginary eigenvalues if the penalty parameters are chosen as $\tau_{u,i} = \frac{a}{2}$ and $\tau_{v,i} = -\frac{a}{2}$ to enforce the interface conditions for the $i \in [1, N+1]$ interface SATs counting the left boundary as the first interface.*

Proof. We show that the result is true for the single-element case and the two-element case;

the general case follows identically.

For the single-element case, the GSBP-SAT discretization of the linear convection equation with periodic boundary conditions is given as

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{H}^{-1}\mathsf{Q}\mathbf{u}_h + \tau_{\mathbf{u}_h}\mathsf{H}^{-1}\left(\mathsf{E}_{x_\mathrm{R}}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{R}}\mathbf{t}_{x_\mathrm{L}}^\mathrm{T}\mathbf{u}_h\right) + \tau_{\mathbf{v}_h}\mathsf{H}^{-1}\left(\mathsf{E}_{x_\mathrm{L}}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{L}}\mathbf{t}_{x_\mathrm{R}}^\mathrm{T}\mathbf{u}_h\right). \tag{B.1}$$

For conservation and stability we can show that $\tau_{\mathbf{v}_h} = \tau_{\mathbf{u}_h} - a$ and $\tau_{\mathbf{u}_h} \leq \frac{a}{2}$. Now consider setting $\tau_{\mathbf{u}_h} = \frac{a}{2}$; therefore $\tau_{\mathbf{v}_h} = -\frac{a}{2}$ and (B.1) becomes

$$\begin{aligned}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} &= -a\mathsf{H}^{-1}\left[\mathsf{Q}^{(\mathrm{A})} + \tfrac{1}{2}\left(\mathsf{E}_{x_\mathrm{R}} - \mathsf{E}_{x_\mathrm{L}}\right)\right]\mathbf{u}_h + \tfrac{a}{2}\mathsf{H}^{-1}\left(\mathsf{E}_{x_\mathrm{R}}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{R}}\mathbf{t}_{x_\mathrm{L}}^\mathrm{T}\mathbf{u}_h\right) \\ &\quad -\tfrac{a}{2}\mathsf{H}^{-1}\left(\mathsf{E}_{x_\mathrm{L}}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{L}}\mathbf{t}_{x_\mathrm{R}}^\mathrm{T}\mathbf{u}_h\right),\end{aligned} \tag{B.2}$$

which reduces to

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{H}^{-1}\left[\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathbf{t}_{x_\mathrm{R}}\mathbf{t}_{x_\mathrm{L}}^\mathrm{T} - \frac{1}{2}\mathbf{t}_{x_\mathrm{L}}\mathbf{t}_{x_\mathrm{R}}^\mathrm{T}\right]\mathbf{u}_h. \tag{B.3}$$

The matrix $\mathsf{Q}^{(\mathrm{A})} + \frac{1}{2}\mathbf{t}_{x_\mathrm{R}}\mathbf{t}_{x_\mathrm{L}}^\mathrm{T} - \frac{1}{2}\mathbf{t}_{x_\mathrm{L}}\mathbf{t}_{x_\mathrm{R}}^\mathrm{T}$ is skew symmetric; therefore, the spatial operator has eigenvalues with zero real parts.

For the two-element case, the left element has solution $\mathbf{u}_h$ with equations

$$\begin{aligned}\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{Q}_{\mathbf{u}_h}^{(\mathrm{A})} + \tfrac{1}{2}\mathsf{E}_{x_\mathrm{R},\mathbf{u}_h} - \tfrac{1}{2}\mathsf{E}_{x_\mathrm{L},\mathbf{u}_h}\right)\mathbf{u}_h &+ \tau_{v1}\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{E}_{x_\mathrm{L},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{\mathrm{x_L},\mathbf{u}_h}\mathbf{t}_{\mathrm{x_R},\mathbf{v}_h}^\mathrm{T}\mathbf{v}_h\right) \\ &+ \tau_{u2}\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{E}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{u}_h - \mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{t}_{x_{L},\mathbf{v}_h}\mathbf{v}_h\right),\end{aligned} \tag{B.4}$$

and the right element has solution $\mathbf{v}_h$ with equations

$$\begin{aligned}\frac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t} = -a\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{Q}_{\mathbf{v}_h}^{(\mathrm{A})} + \tfrac{1}{2}\mathsf{E}_{x_\mathrm{R},\mathbf{v}_h} - \tfrac{1}{2}\mathsf{E}_{x_\mathrm{L},\mathbf{v}_h}\right)\mathbf{v}_h &+ \tau_{u1}\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{E}_{x_\mathrm{R},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{\mathrm{x_R},\mathbf{v}_h}\mathbf{t}_{\mathrm{x_L},\mathbf{u}_h}^\mathrm{T}\mathbf{u}_h\right) \\ &+ \tau_{v2}\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{E}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{v}_h - \mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}^\mathrm{T}\mathbf{u}_h\right),\end{aligned} \tag{B.5}$$

where the penalty parameters τ_{u1} and τ_{v1}, and τ_{u2} and τ_{v2}, are for the first and second interface counting from left to right. For conservation and stability, it is necessary that $\tau_{u1} \leq \frac{a}{2}$ and $\tau_{v1} = \tau_{u1} - a$, and $\tau_{u2} \leq \frac{a}{2}$ and $\tau_{v2} = \tau_{u2} - a$. As in the single-element case, consider taking $\tau_{u1} = \frac{a}{2}$; therefore, $\tau_{v1} = -\frac{a}{2}$, and $\tau_{u2} = \frac{a}{2}$, thus, $\tau_{v2} = -\frac{a}{2}$. With these choices, (B.4) becomes

$$\frac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} = -a\mathsf{H}_{\mathbf{u}_h}^{-1}\left(\mathsf{Q}_{\mathbf{u}_h}^{(\mathrm{A})}\mathbf{u}_h - \mathbf{t}_{\mathrm{x_L},\mathbf{u}_h}\mathbf{t}_{\mathrm{x_R},\mathbf{v}_h}^\mathrm{T} + \frac{1}{2}\mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}^\mathrm{T}\mathbf{v}_h\right) \tag{B.6}$$

and (B.5) becomes

$$\frac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t} = -a\mathsf{H}_{\mathbf{v}_h}^{-1}\left(\mathsf{Q}_{\mathbf{v}_h}^{(\mathrm{A})}\mathbf{v}_h + \frac{1}{2}\mathbf{t}_{\mathrm{x_R},\mathbf{v}_h}\mathbf{t}_{\mathrm{x_L},\mathbf{u}_h}^{\mathrm{T}}\mathbf{u}_h - \frac{1}{2}\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}\mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}^{\mathrm{T}}\mathbf{u}_h\right). \tag{B.7}$$

The equations (B.6) and (B.7) can be recast in matrix form as

$$\begin{bmatrix} \dfrac{\mathrm{d}\mathbf{u}_h}{\mathrm{d}t} \\ \\ \dfrac{\mathrm{d}\mathbf{v}_h}{\mathrm{d}t} \end{bmatrix} = -a \begin{bmatrix} \mathsf{H}_{\mathbf{u}_h}^{-1} & \\ & \mathsf{H}_{\mathbf{v}_h}^{-1} \end{bmatrix} \overbrace{\begin{bmatrix} \mathsf{Q}_{\mathbf{u}_h}^{(\mathrm{A})} & \mathsf{G} \\ \\ -\mathsf{G}^{\mathrm{T}} & \mathsf{Q}_{\mathbf{v}_h}^{(\mathrm{A})} \end{bmatrix}}^{\mathsf{A}} \begin{bmatrix} \mathbf{u}_h \\ \\ \mathbf{v}_h \end{bmatrix}, \tag{B.8}$$

where $\mathsf{G} = -\frac{1}{2}\mathbf{t}_{x_\mathrm{L},\mathbf{u}_h}\mathbf{t}_{\mathrm{x_R},\mathbf{v}_h}^{\mathrm{T}} + \frac{1}{2}\mathbf{t}_{x_\mathrm{R},\mathbf{u}_h}\mathbf{t}_{x_\mathrm{L},\mathbf{v}_h}^{\mathrm{T}}$, and as in the single-element case, the matrix A is skew symmetric and again we conclude that the spatial operator has eigenvalues that have zero real parts. □

Appendix C

Numerical results

C.1 Linear convection equation

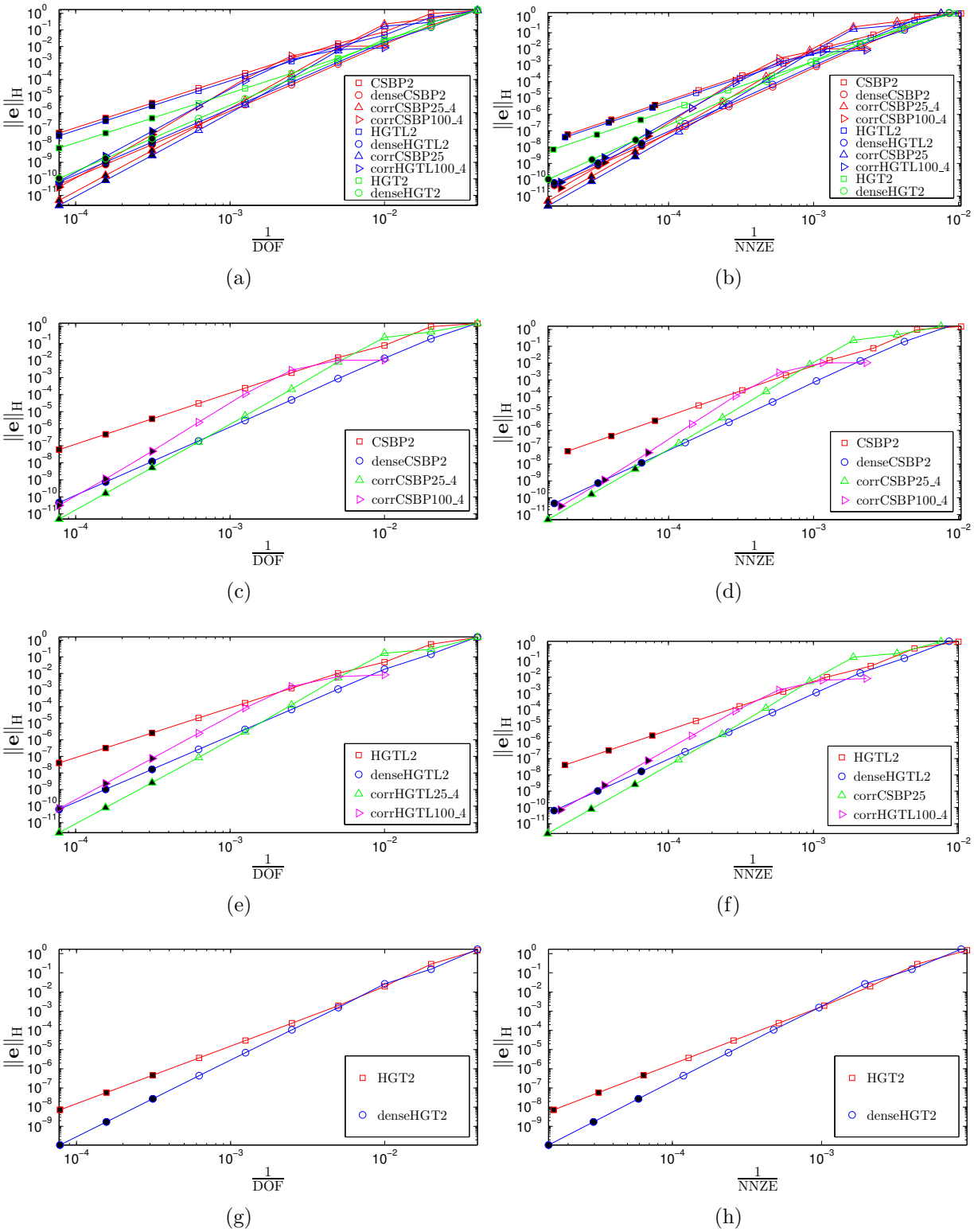


Figure C.1: Operators with a repeating interior operator of order 4 implemented as elements with 25 nodes, with the exception of the corner-corrected operator with 100 nodes. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$, (a), (c), (e), and (g) or versus $\frac{1}{\text{NNZE}}$, (b), (d), (f), and (h). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 2$.

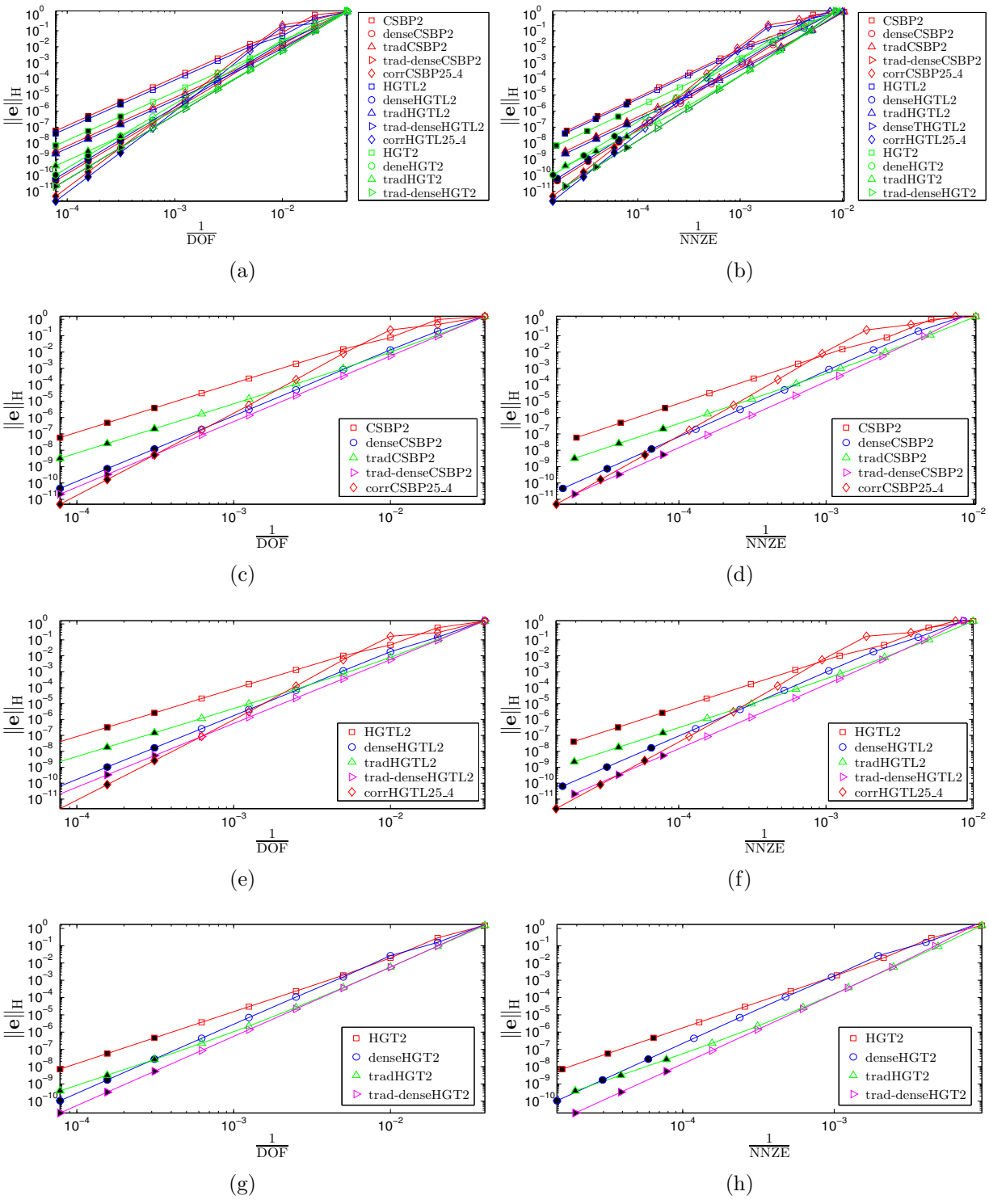


Figure C.2: Operators with a repeating interior operator of order 4 implemented in a traditional FD manner. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$ (a), (c), (e), and (g) or versus $\frac{1}{\text{NNZE}}$, (b), (d), (f), and (h). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 2$.

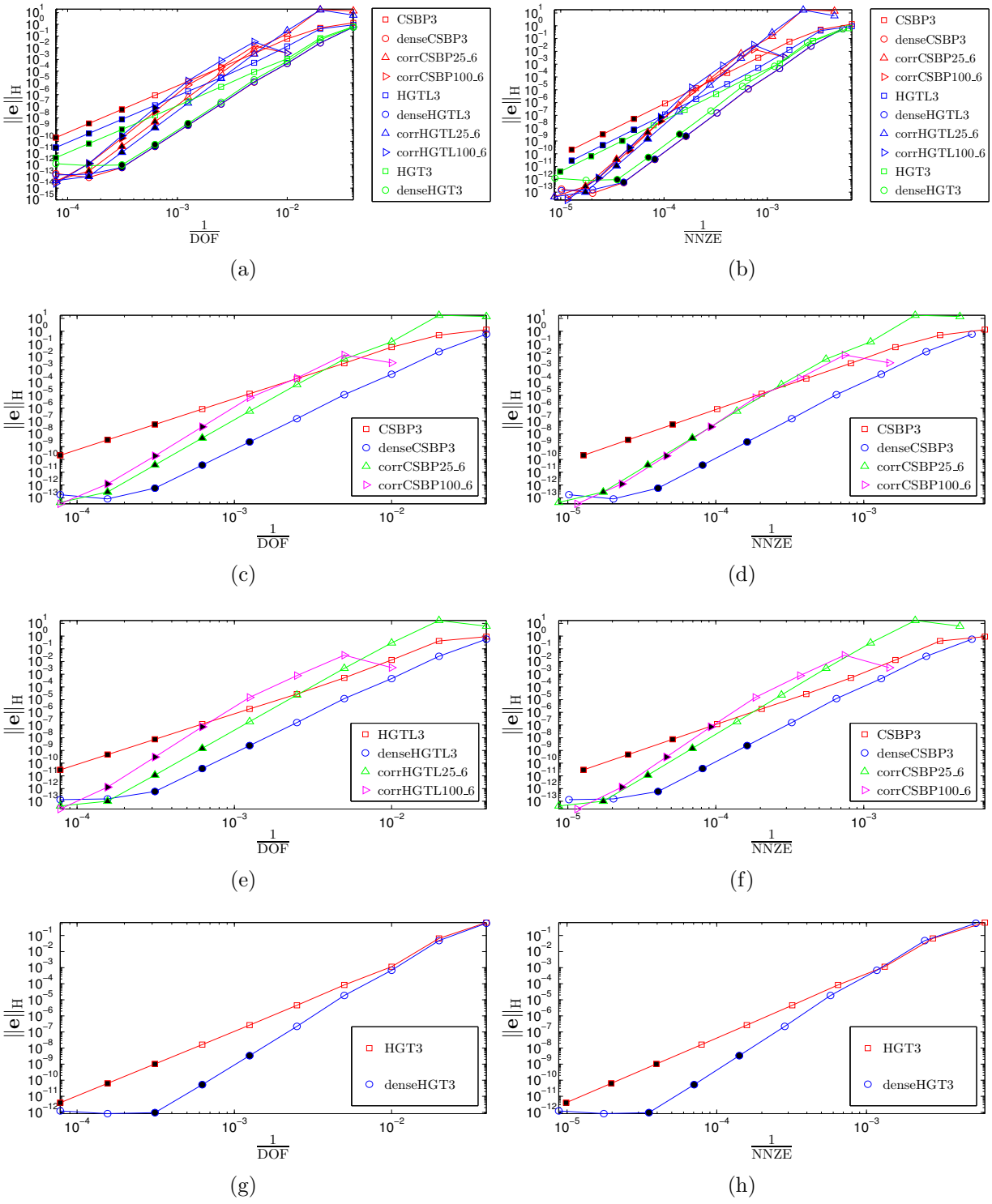


Figure C.3: Operators with a repeating interior operator of order 6 implemented as elements with 25 nodes, with the exception of the corner-corrected operator with 100 nodes. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$, (a), (c), (e), and (g) or versus $\frac{1}{\text{NNZE}}$, (b), (d), (f), and (h). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 3$.

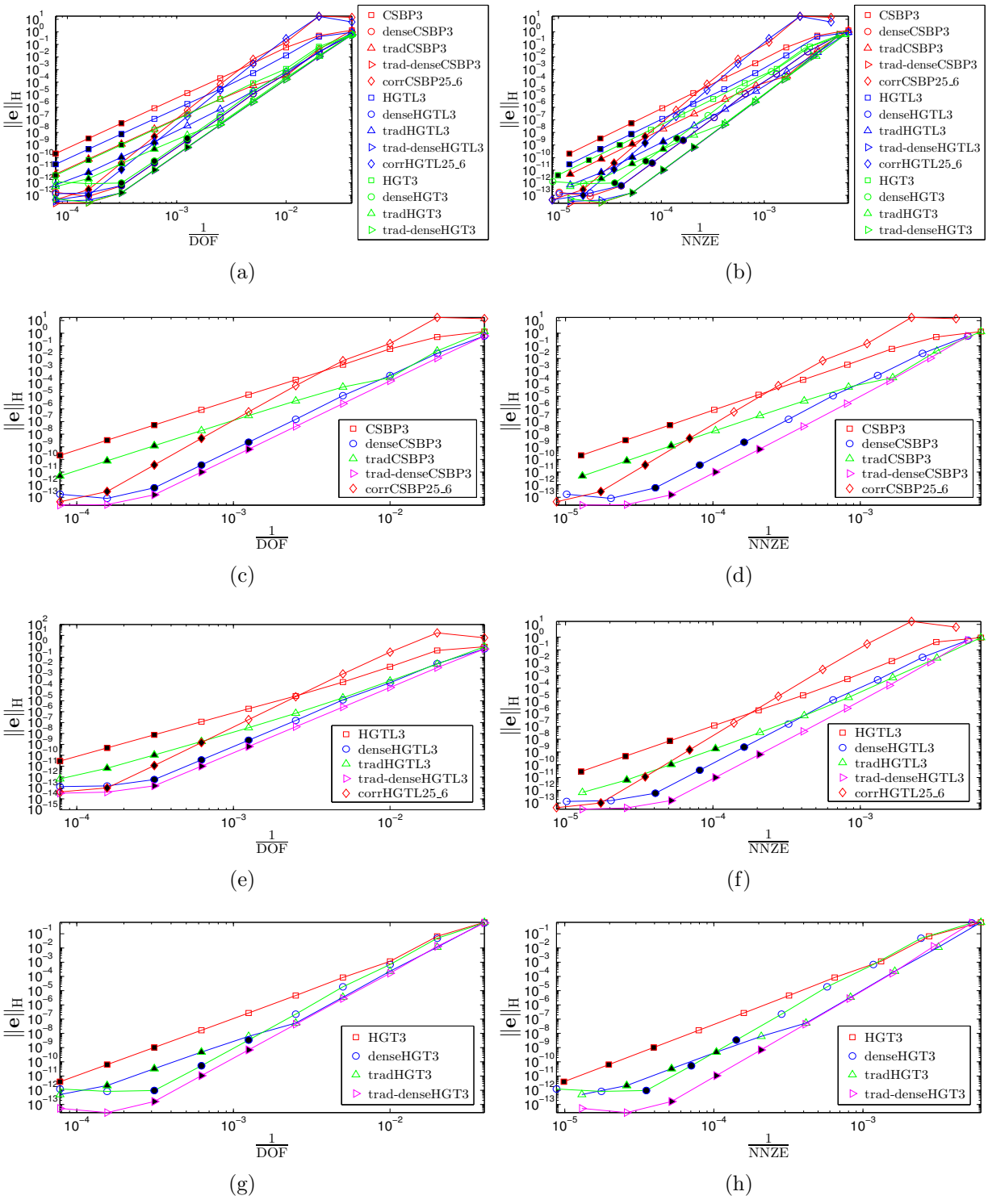


Figure C.4: Operators with a repeating interior operator of order 6 implemented in a traditional FD manner. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$ (a), (c), (e), and (g) or versus $\frac{1}{\text{NNZE}}$, (b), (d), (f), and (h). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 3$.

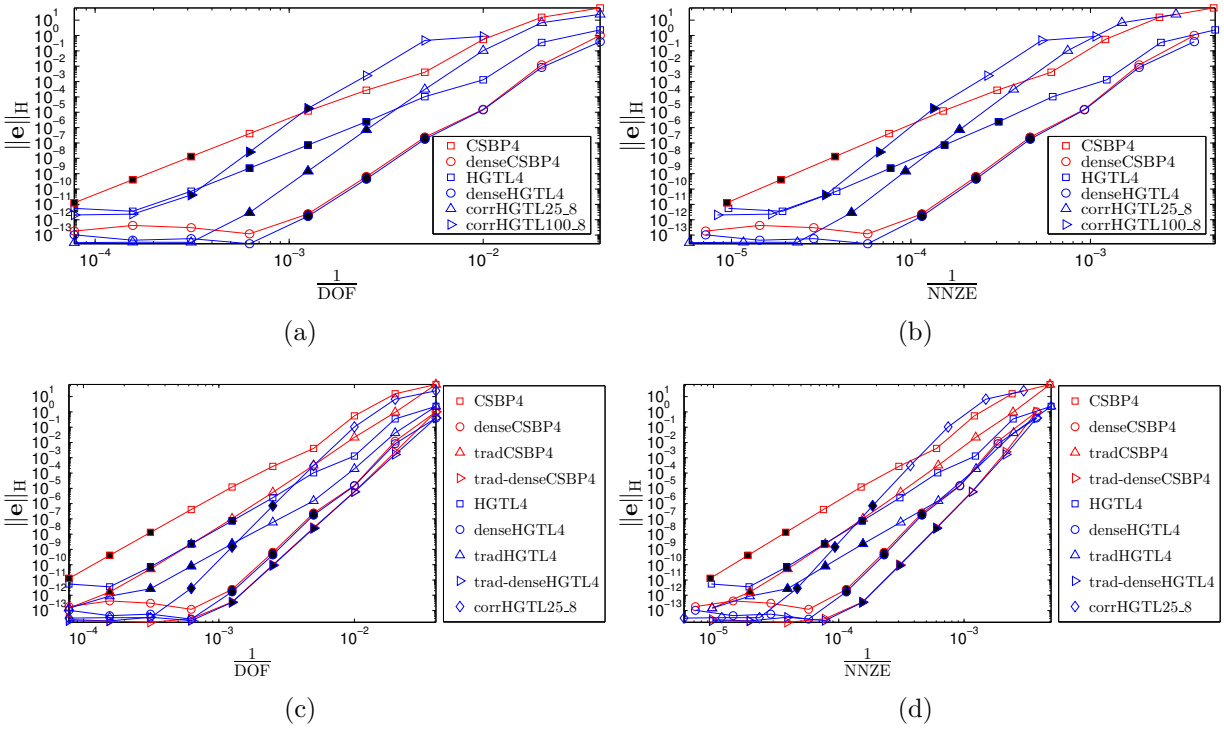


Figure C.5: Operators with a repeating interior operator of order 8 implemented as elements with 25 nodes, with the exception of the corner-corrected operator with 100 nodes, (a) and (b) or in a traditional FD manner (c) and (d). H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$ (a) and (c) or versus $\frac{1}{\text{NNZE}}$, (b) and (d). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 4$ while the corner-corrected operators have $\tilde{a} = \tilde{j} = 3$.

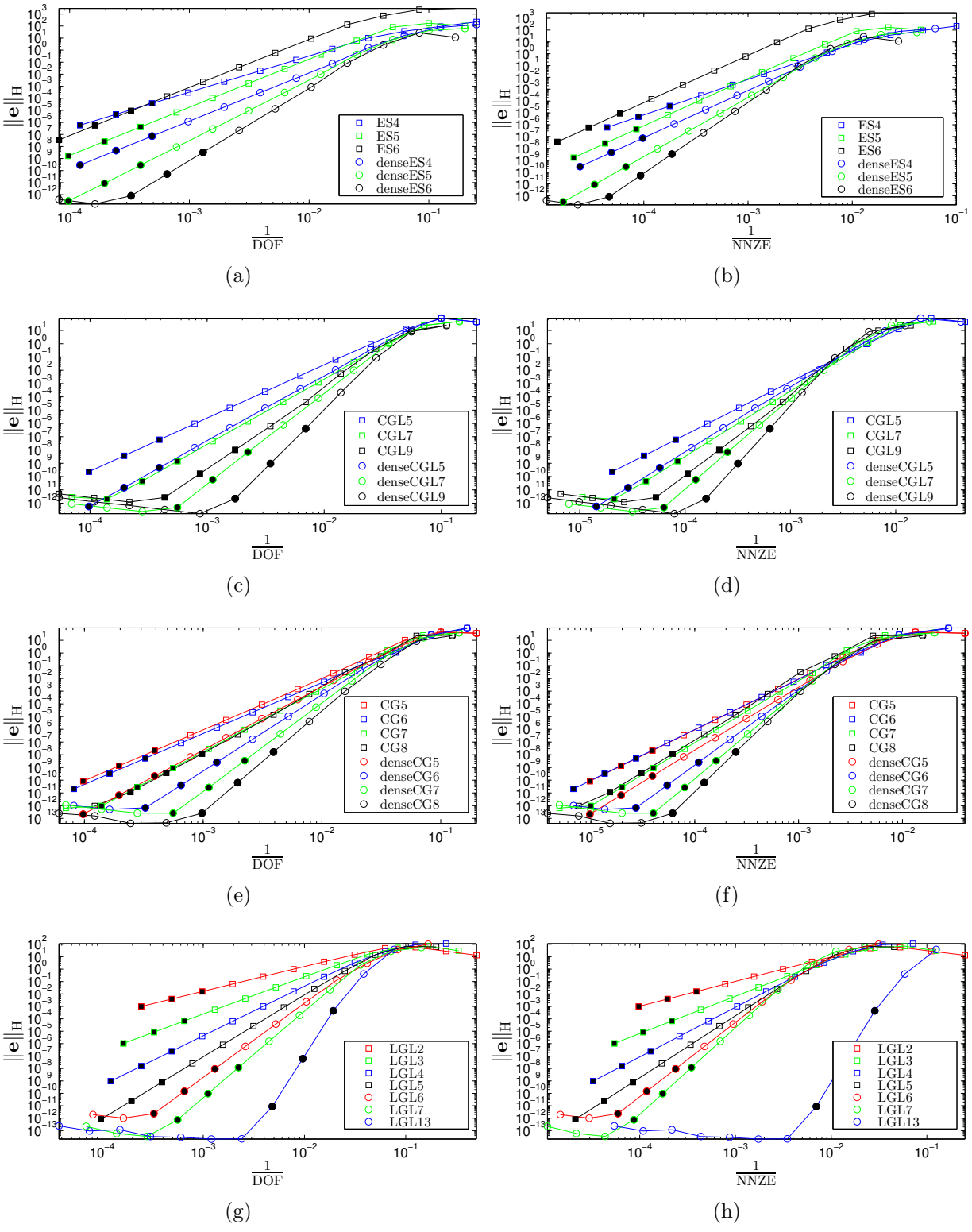


Figure C.6: Element-type GSBP operators. H norm of the error in the solution to problem (8.1) versus $\frac{1}{\text{DOF}}$, (a), (c), (e), and (g) or versus $\frac{1}{\text{NNZE}}$ (b), (d), (f), and (h).

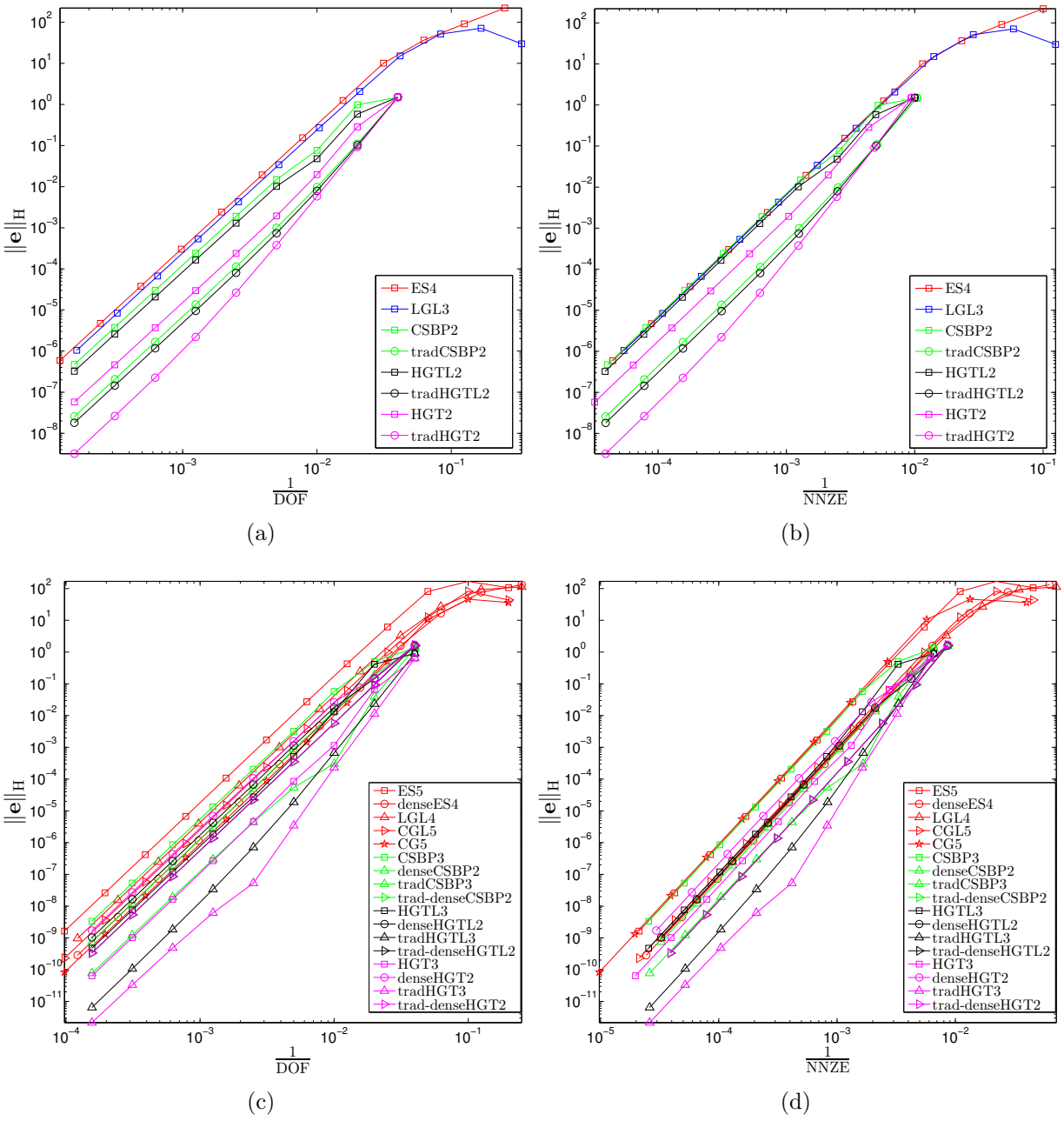


Figure C.7: H norm of the error in the solution to problem (8.1), for operators with solution error of order $3-4$, versus $\frac{1}{\mathrm{DOF}}$, (a) and (c) or versus $\frac{1}{\mathrm{NNZE}}$, (c) and (d).

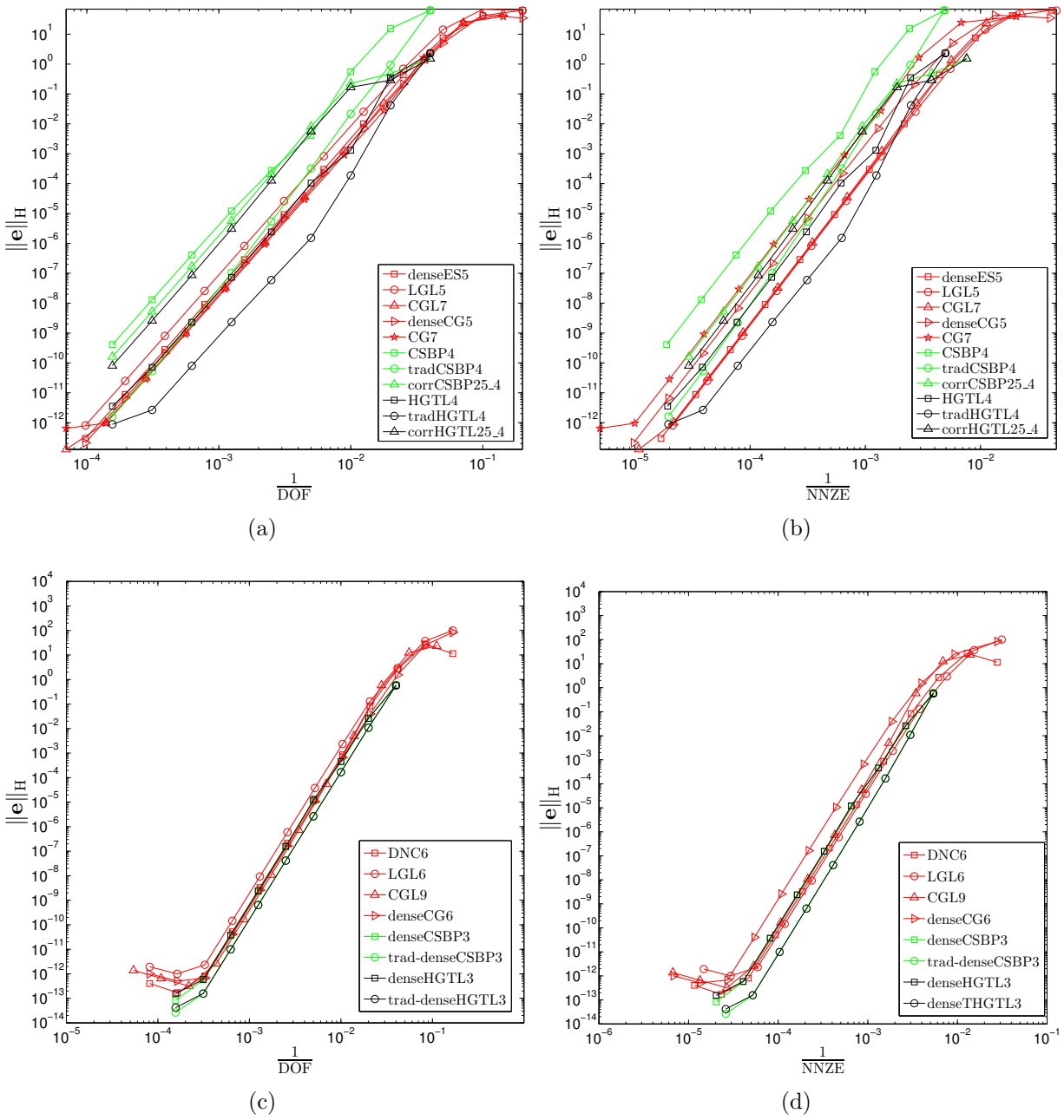


Figure C.8: H norm of the error in the solution to problem (8.1), for operators with solution error of order $5-6$, versus $\frac{1}{\text{DOF}}$, (a) and (c) or versus $\frac{1}{\text{NNZE}}$, (b) and (d).

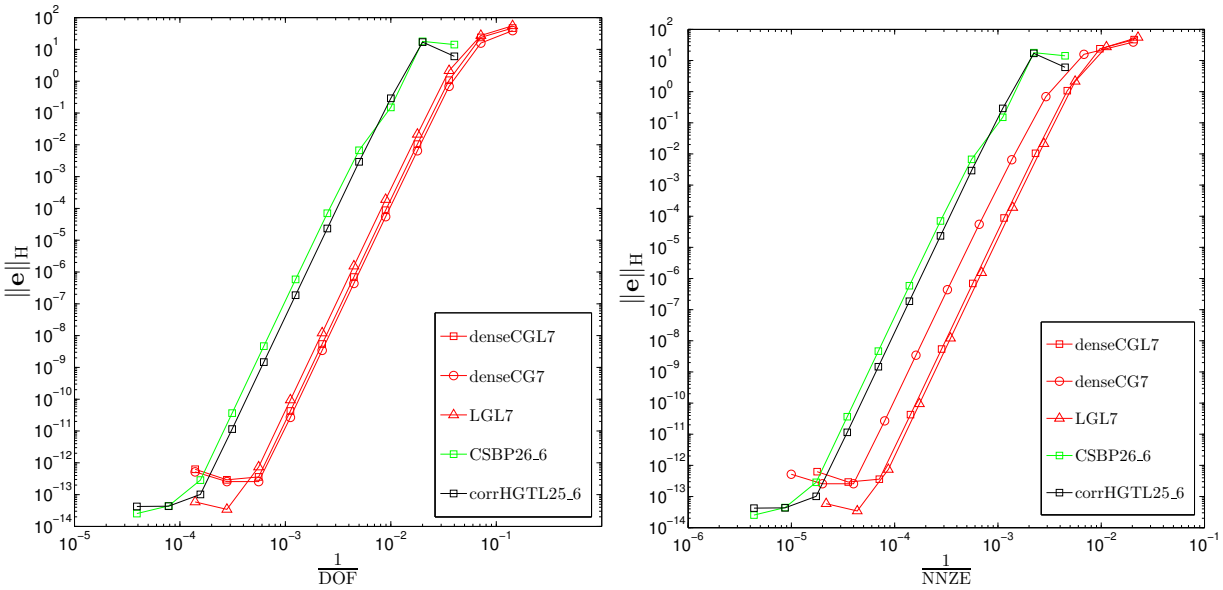


Figure C.9: H norm of the error in the solution to problem (8.1), for operators with solution error of order 7, versus $\frac{1}{\mathrm{DOF}}$, (a) or versus $\frac{1}{\mathrm{NNZE}}$, (b).

C.2 Linear convection-diffusion equation

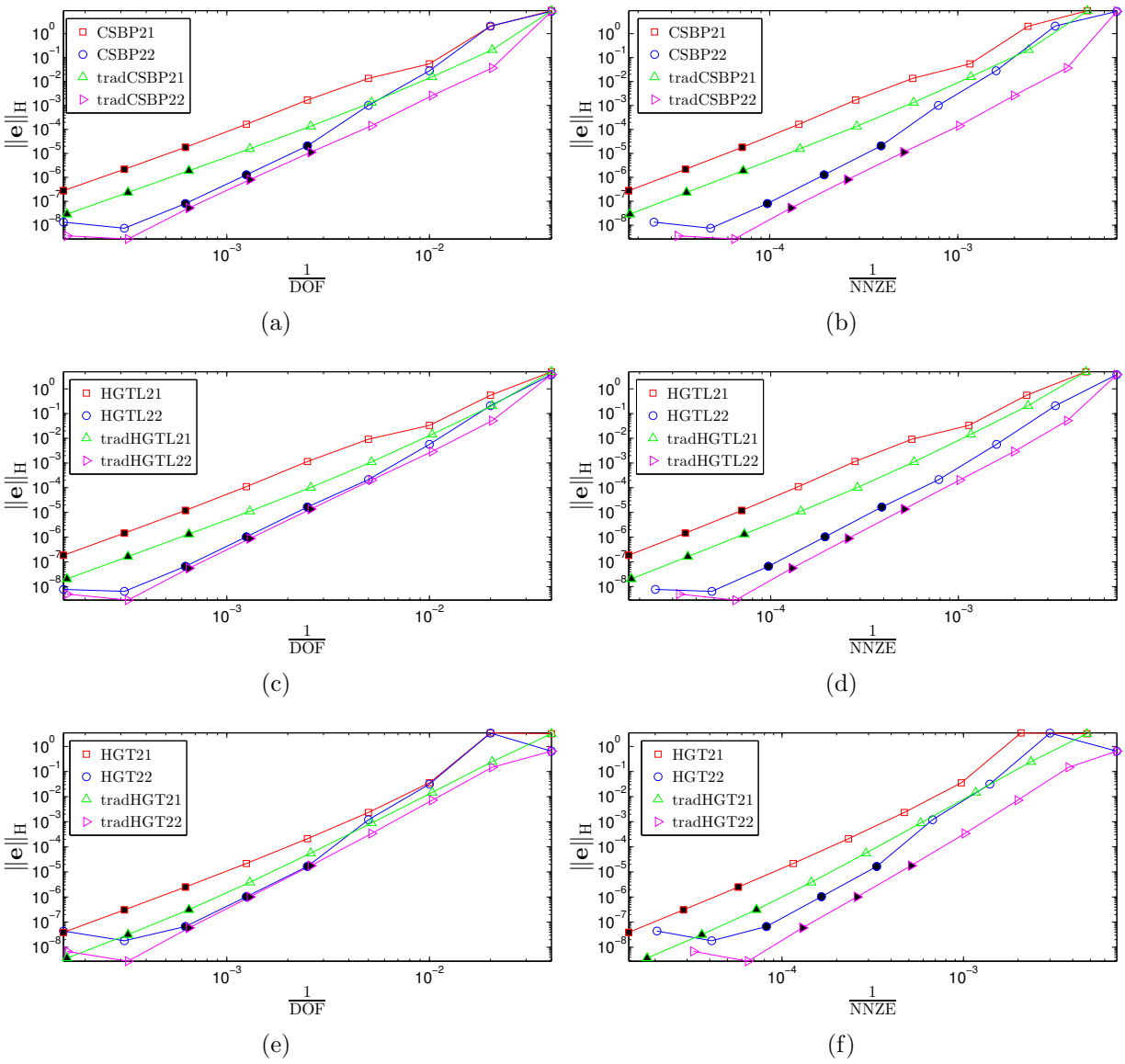


Figure C.10: Operators with a repeating interior operator of order 4 implemented as elements with 25 nodes or in a traditional FD manner. H norm of the error in the solution to problem (8.6) versus $\frac{1}{\text{DOF}}$, (a), (c), and (e) or versus $\frac{1}{\text{NNZE}}$, (b), (d), and (f). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 2$.

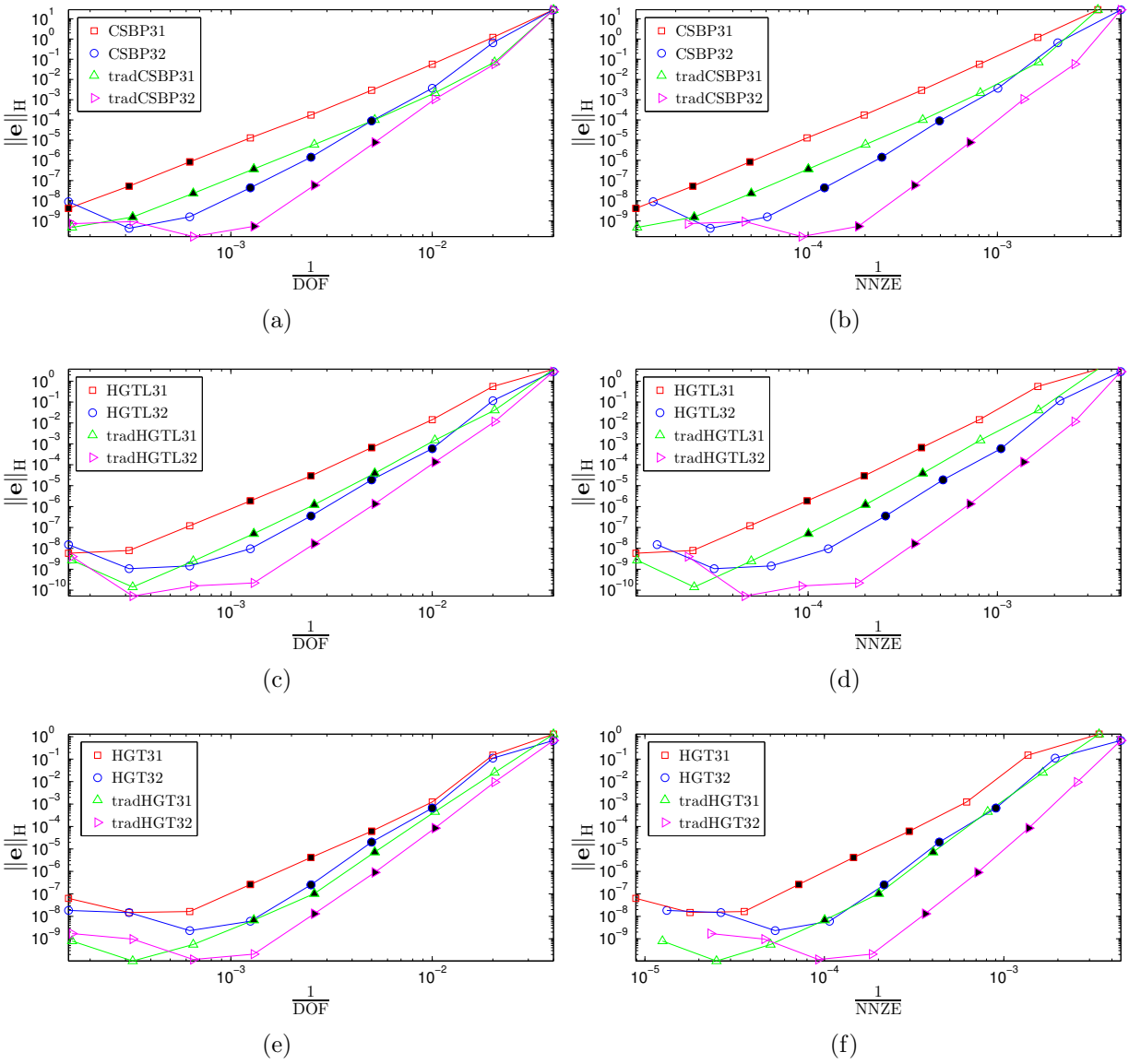


Figure C.11: Operators with a repeating interior operator of order 6 implemented as elements with 25 nodes or in a traditional FD manner. H norm of the error in the solution to problem (8.6) versus $\frac{1}{\text{DOF}}$, (a), (c), and (e) or versus $\frac{1}{\text{NNZE}}$, (b), (d), and (f). The HGTL and HGT nodal distributions were constructed with $\tilde{a} = \tilde{j} = 3$.

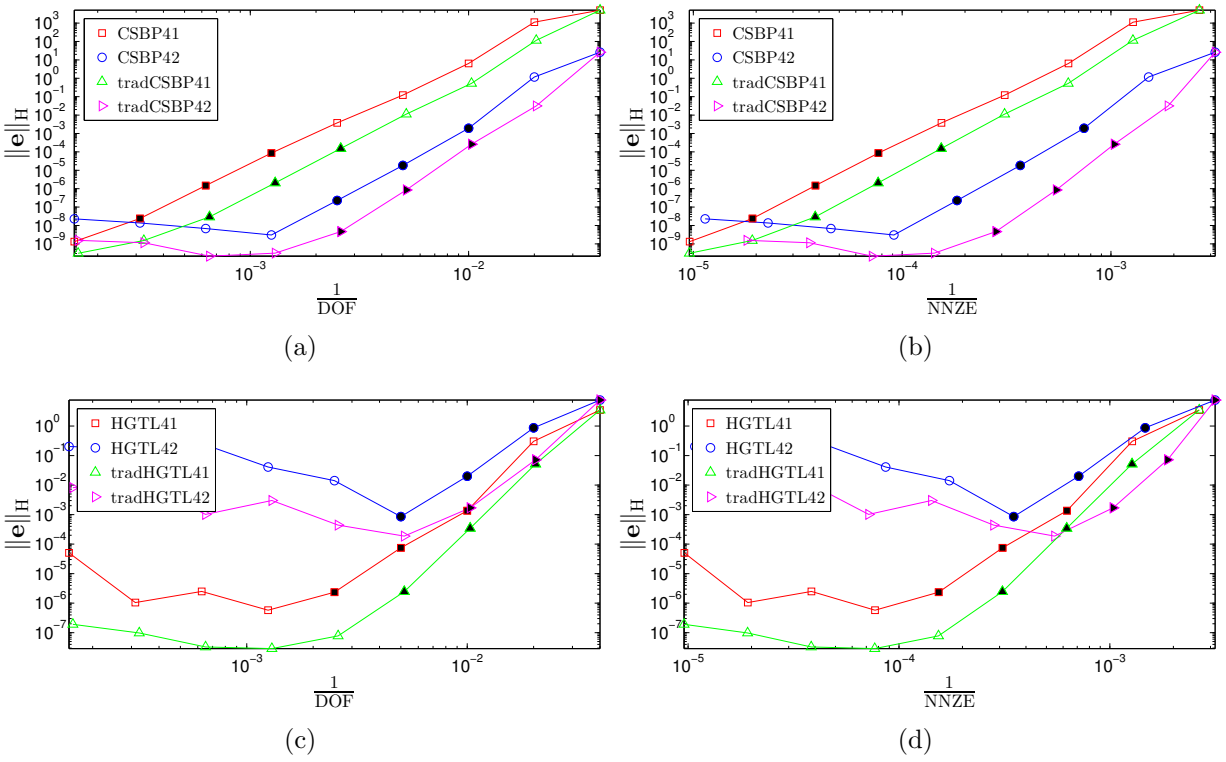


Figure C.12: Operators with a repeating interior operator of order 8 implemented as elements with 25 nodes or in a traditional FD manner. H norm of the error in the solution to problem (8.6) versus $\frac{1}{\text{DOF}}$, (a) and (c) or versus $\frac{1}{\text{NNZE}}$, (b) and (d). The HGTL nodal distributions were constructed with $\tilde{a} = \tilde{j} = 4$.

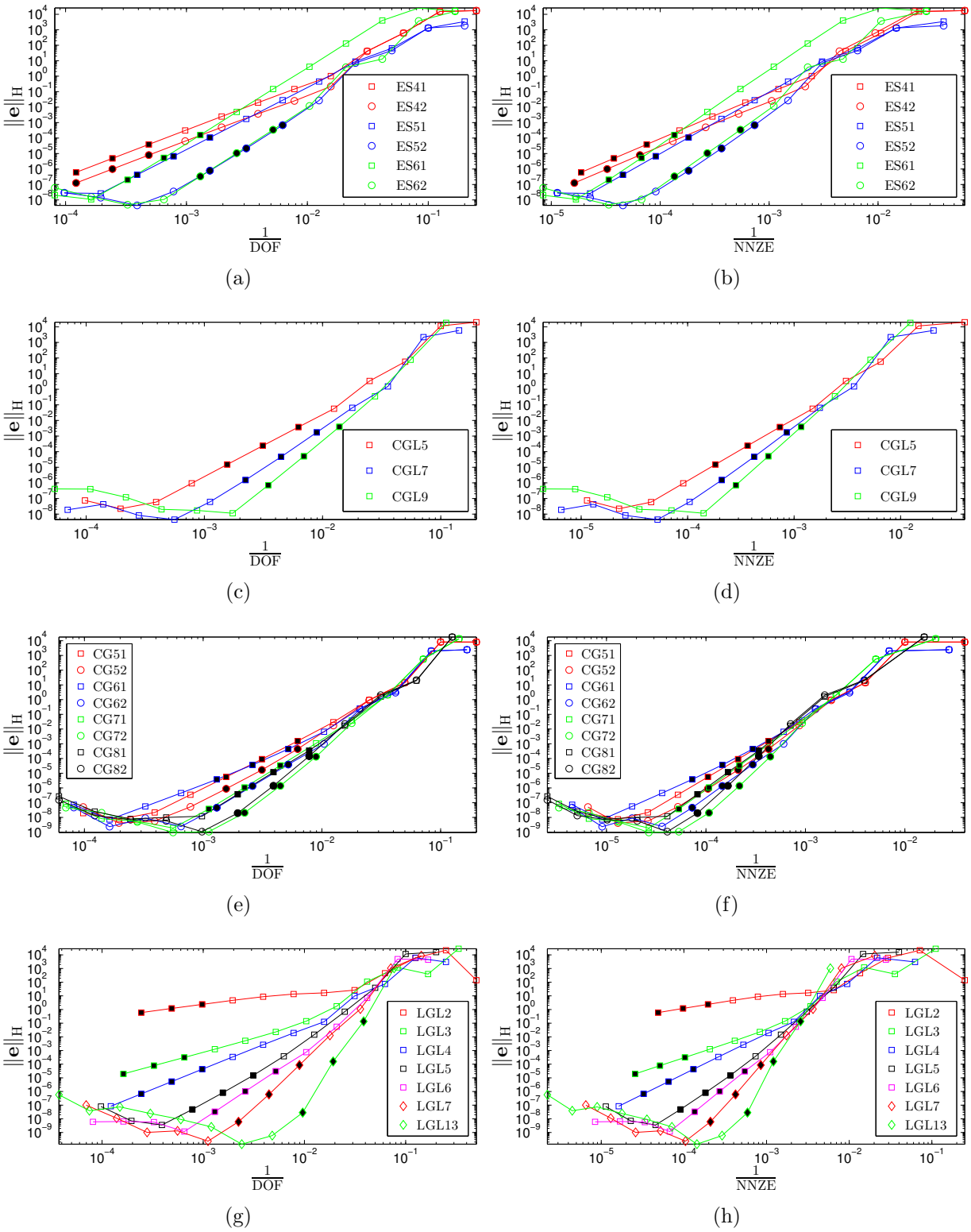


Figure C.13: Element-type GSBP operators. H norm of the error in the solution to problem (8.6) versus $\frac{1}{\mathrm{DOF}}$, (a), (c), and (g) or versus $\frac{1}{\mathrm{NNZE}}$, (b), (d), (f), and (h).

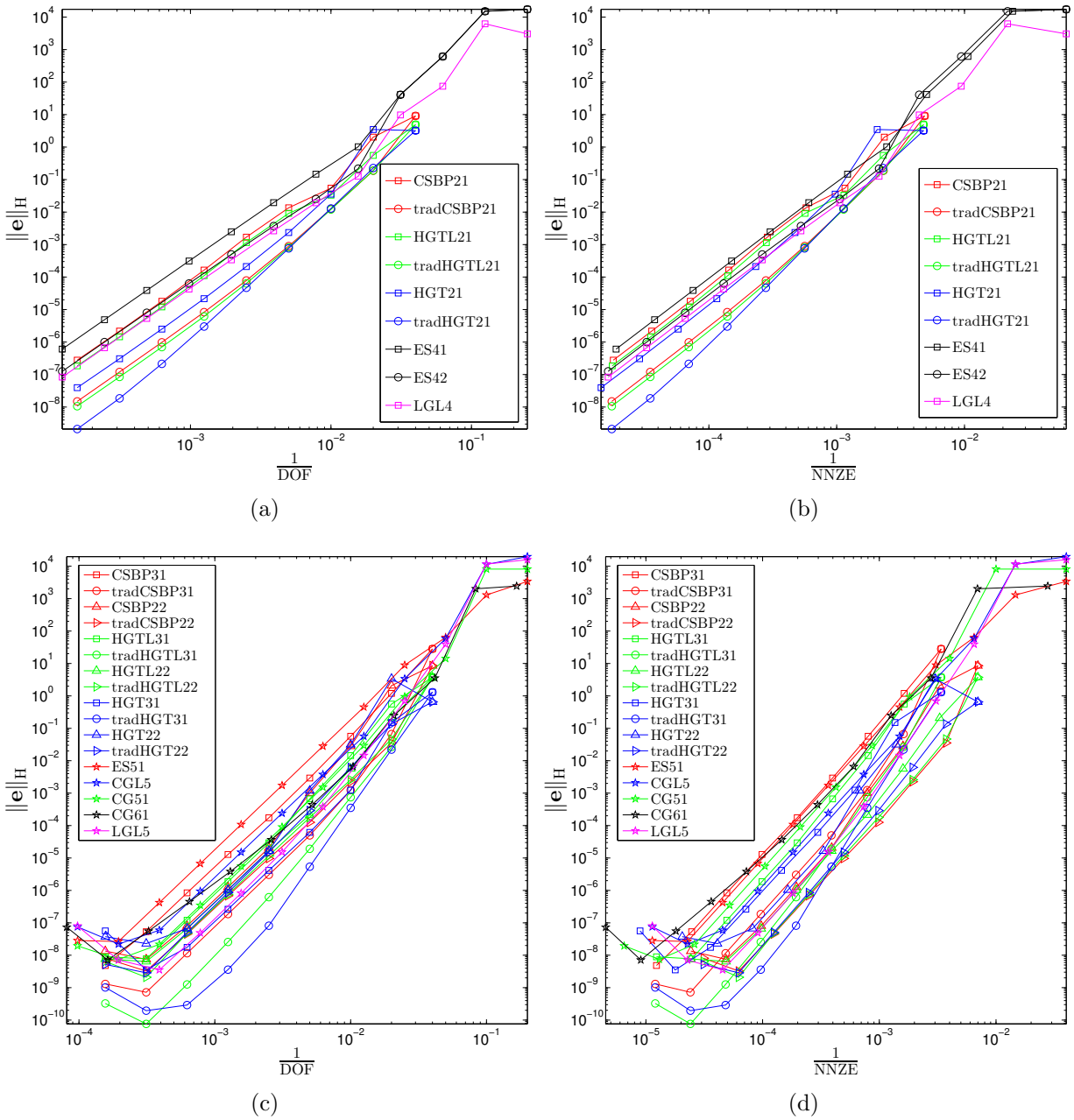


Figure C.14: H norm of the error in the solution to problem (8.6), for operators with solution error of order $3-4$, versus $\frac{1}{\mathrm{DOF}}$, (a), (c), (e), and (g) or versus $\frac{1}{\mathrm{NNZE}}$, (b), (d), (f), and (h).

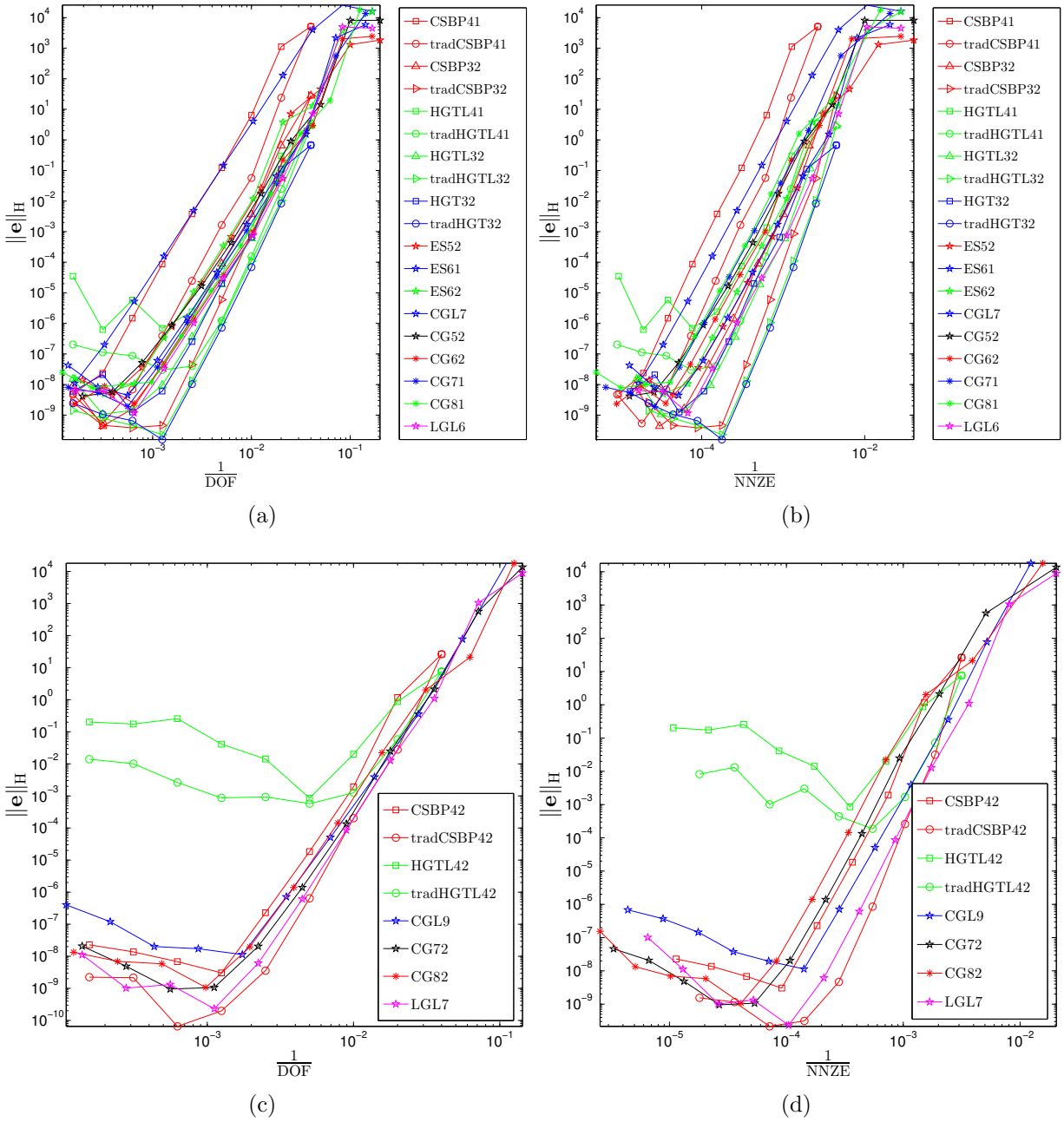


Figure C.15: H norm of the error in the solution to problem (8.6), for operators with solution error of order $5-6$, versus $\frac{1}{\text{DOF}}$, (a), (c), (e), and (g) or versus $\frac{1}{\text{NNZE}}$, (b), (d), (f), and (h).