

Stability of Overintegration Methods for Nodal Discontinuous Galerkin Spectral Element Methods

David A. Kopriva

Received: date / Accepted: date

Abstract

We perform stability analyses for discontinuous Galerkin spectral element approximations of linear variable coefficient hyperbolic systems in three dimensional domains with curved elements. Although high order, the precision of the quadratures used are typically too low with respect to polynomial order associated with their arguments, which introduces aliasing errors that can destabilize an approximation, especially when the solution is underresolved. We show that using a larger number of points in the volume quadrature, often called “overintegration”, can eliminate the aliasing term associated with the volume, but introduces new aliasing errors at the surfaces that can destabilize the solution. Increased quadrature precision on both the volume and surface terms, on the other hand, leads to a stable approximation. The results support the findings of Mengaldo *et al.* [*Dealiasing techniques for high-order spectral element methods on regular and irregular grids. Journal of Computational Physics, 299:56 – 81, 2015*] who found that fully consistent integration was more robust for the solution of compressible flows than the volume only version.

1 Introduction

Discontinuous Galerkin (DG) spectral element methods (DGSEM) for hyperbolic systems of equations are characterized by an approximation of a weak form of the equations where the solutions and fluxes are approximated by polynomials of degree N (typically large, $N \geq 4$) with nodal values represented at Gauss (or Gauss-Lobatto) points, and inner products are approximated by the associated Gauss quadrature. Unfortunately, the volume and surface flux terms usually are underintegrated with respect to the polynomial order of the advective fluxes. This happens in nonlinear problems, but also for variable coefficient linear problems, or even constant coefficient problems posed on curved element meshes.

Although the method usually works in practice, aliasing errors associated with the underintegration of the fluxes can lead to instability, especially if the solution is severely underresolved [1],[4]. In [7], the authors proposed a method to stabilize nodal DGSEMs for hyperbolic systems, which they called “super-collocation” and which, though perhaps better called “consistent integration”, is commonly called “overintegration”. In [7], Gauss quadratures of sufficiently high order

were used to approximate the volume integrals so as to eliminate the polynomial aliasing associated with the quadrature. Later, in [14], numerical experiments suggested that overintegrating the surface integrals increases the robustness of approximations of the Euler gas dynamics equations. Overintegration is used in large codes such as Nektar++[2] and Flexi[6] for stabilization. However theoretical justification of the procedures has not formally been established to date.

In this paper we study the linear (energy) stability for hyperbolic systems with variable coefficients and for curved elements in three space dimensions of (i) the original DG spectral element approximation, (ii) overintegration of the volume terms only, and (iii) overintegration of the surface and volume terms together. Our results support the conclusion of [14] that overintegration of both the surface and volume terms is more robust than overintegration of the volume terms alone.

2 The Hyperbolic Boundary-Value Problem

In this paper we study the stability of discontinuous Galerkin spectral element methods for boundary-value problems for variable coefficient linear hyperbolic systems of the form

$$\mathbf{u}_t + \nabla \cdot \vec{\mathbf{f}} = 0 \quad (1)$$

defined on a domain Ω . Here, $\mathbf{u}(\vec{x}, t) = [u_1(\vec{x}, t) \ u_2(\vec{x}, t) \ \dots \ u_p(\vec{x}, t)]^T$ is the state vector and

$$\vec{\mathbf{f}}(\mathbf{u}) = \sum_{m=1}^3 \underline{\mathbf{A}}_m(\vec{x}) \mathbf{u} \hat{x}_m = \vec{\underline{\mathbf{A}}}(\vec{x}) \mathbf{u} \quad (2)$$

is the linear flux space vector, where $\underline{\mathbf{A}}_m$ is the coefficient matrix of the derivative in the m^{th} space dimension and \hat{x}_m is a unit vector in that direction. We assume that the system has already been symmetrized and is hyperbolic, that is

$$\underline{\mathbf{A}}_m = (\underline{\mathbf{A}}_m)^T \quad \text{and} \quad \sum_{m=1}^3 \alpha_m \underline{\mathbf{A}}_m = \underline{\mathbf{R}} \underline{\Lambda} \underline{\mathbf{R}}^{-1} \quad (3)$$

for any $\sum_{m=1}^3 \alpha_m^2 \neq 0$ and some real diagonal matrix $\underline{\Lambda}$. We also assume that the matrices are continuous in their argument and that the $\underline{\mathbf{A}}_m$ have bounded derivatives in the sense that

$$\left\| \nabla \cdot \vec{\underline{\mathbf{A}}} \right\|_2 < \infty, \quad (4)$$

where $\|\cdot\|_2$ is the matrix 2-norm. Adding the initial condition $\mathbf{u}(\vec{x}, 0) = \mathbf{u}_0(\vec{x})$ and external state boundary values \mathbf{g} properly imposed on $\partial\Omega$ completes the specification of the boundary-value problem.

With characteristic boundary conditions specified along the $\partial\Omega$, the \mathbb{L}^2 norm of the solution $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ defined through the inner product $\langle \mathbf{u}, \mathbf{v} \rangle = \int_{\Omega} \mathbf{u}^T \mathbf{v} dx dy dz$ is bounded in terms of the initial and boundary data as [12]

$$\frac{d}{dt} \|\mathbf{u}\|^2 = - \int_{\partial\Omega} \{ \mathbf{u}^T \underline{\mathbf{A}}^+ \mathbf{u} - \mathbf{g}^T |\underline{\mathbf{A}}^-| \mathbf{g} \} dS + 2\gamma \|u\|^2, \quad (5)$$

where $\gamma = \frac{1}{2} \max_{\Omega} \left\| \nabla \cdot \underline{\vec{A}} \right\|_2$,

$$\underline{\vec{A}} \cdot \hat{n} = \sum_{m=1}^3 \underline{A}_m \hat{n}_m = \underline{R} \underline{\Lambda} \underline{R}^{-1} = \underline{R} \underline{\Lambda}^+ \underline{R}^{-1} + \underline{R} \underline{\Lambda}^- \underline{R}^{-1} \equiv \underline{A}^+ + \underline{A}^-, \quad (6)$$

$\underline{\Lambda}^{\pm} = (\underline{\Lambda} \pm |\underline{\Lambda}|) / 2$ are diagonal matrices, and \hat{n} is the outward normal to the surface of the domain. Integrating (5) in time gives

$$\begin{aligned} \|\mathbf{u}(T)\|^2 + \int_0^T \int_{\partial\Omega} \mathbf{u}^T \underline{A}^+ \mathbf{u} dS dt &\leq e^{2\gamma T} \|\mathbf{u}(0)\|^2 + \int_0^T \int_{\partial\Omega} e^{2\gamma(T-t)} \mathbf{g}^T |\underline{A}^-| \mathbf{g} dS dt \\ &\leq e^{2\gamma T} \left\{ \|\mathbf{u}_0\|^2 + \int_0^T \int_{\partial\Omega} \mathbf{g}^T |\underline{A}^-| \mathbf{g} dS dt \right\}. \end{aligned} \quad (7)$$

Equations (5) and (7) show that energy is dissipated at the physical boundary along outgoing characteristics (\underline{A}^+ term), added through the incoming characteristics (\underline{A}^- term), and can grow as a result of the variable coefficients in the problem ($e^{2\gamma t}$ factor). If the external state $\mathbf{g} = 0$ and $\nabla \cdot \underline{\vec{A}} = 0$ then $\gamma = 0$ and (7) reduces to

$$\|\mathbf{u}(T)\| \leq \|\mathbf{u}_0\|, \quad (8)$$

so that the energy does not grow.

3 DG Spectral Element Methods for Variable Coefficient Problems on Curved Elements

We consider here the domain Ω subdivided into N_{el} nonoverlapping hexahedral elements, $e^k, k = 1, 2, \dots, N_{el}$. We assume that the subdivision is conforming. Each element is mapped from the reference element E by a transformation $\vec{x} = \vec{X}(\vec{\xi})$. Directions in physical space can be represented in terms of the mapping by the three covariant basis vectors

$$\vec{a}_i = \frac{\partial \vec{X}}{\partial \xi^i} \quad i = 1, 2, 3, \quad (9)$$

and the (volume weighted) contravariant vectors, formally written as

$$\mathcal{J} \vec{a}^i = \vec{a}_j \times \vec{a}_k, \quad (i, j, k) \text{ cyclic}, \quad (10)$$

where

$$\mathcal{J} = \vec{a}_1 \cdot (\vec{a}_2 \times \vec{a}_3) \quad (11)$$

is the Jacobian of the transformation. We will assume that the metric terms are approximated as polynomials of degree N so that

$$\sum_{i=1}^3 \frac{\partial \mathbb{I}^N(\vec{J} \vec{a}^i)}{\partial \xi^i} = 0, \quad (12)$$

as described in [8],[13], where $\mathbb{I}^N : L^2(E) \rightarrow \mathbb{P}^N$ is the polynomial interpolation operator and \mathbb{P}^N is the space of polynomials of degree less than or equal to N on E .

Under the mapping, the divergence of a spatial vector flux can be written compactly in terms of the reference space variables as

$$\nabla \cdot \vec{\mathbf{f}} = \frac{1}{\mathcal{J}} \sum_{i=1}^3 \frac{\partial}{\partial \xi^i} \left(\mathcal{J} \vec{a}^i \cdot \vec{\mathbf{f}} \right) = \frac{1}{\mathcal{J}} \sum_{i=1}^3 \frac{\partial \tilde{\mathbf{f}}^i}{\partial \xi^i} = \frac{1}{\mathcal{J}} \nabla_\xi \cdot \tilde{\mathbf{f}}. \quad (13)$$

The vector $\tilde{\mathbf{f}}$ is the volume weighted contravariant flux whose components are $\tilde{\mathbf{f}}^i = \mathcal{J} \vec{a}^i \cdot \vec{\mathbf{f}}$. The conservation law can therefore be represented on the reference domain by another conservation law

$$\mathcal{J} \mathbf{u}_t + \nabla_\xi \cdot (\tilde{\mathbf{A}} \mathbf{u}) = 0, \quad (14)$$

where we define the (volume weighted) contravariant coefficient matrices

$$\tilde{\mathbf{A}}^i = \mathcal{J} \vec{a}^i \cdot \vec{\mathbf{A}} \quad (15)$$

and

$$\tilde{\mathbf{A}} = \sum_{i=1}^3 \tilde{\mathbf{A}}^i \xi^i. \quad (16)$$

The discontinuous Galerkin approximation is created from a weak formulation formed by multiplying (14) by a test function ϕ , forming the inner product over the reference element, E , and integrating by parts.

$$\langle \mathcal{J} \mathbf{u}, \phi \rangle + \int_{\partial E} \phi^T \tilde{\mathbf{f}} \cdot \hat{n} dS - \langle \tilde{\mathbf{f}}, \nabla \phi \rangle = 0. \quad (17)$$

The solution and fluxes are then approximated by polynomials global within each element. The solution \mathbf{u} is approximated by a polynomial of degree N , $\mathbf{u} \approx \mathbf{U} \in \mathbb{P}^N$ as is the Jacobian, $\mathcal{J} \approx J \in \mathbb{P}^N$. They are polynomial interpolants with nodes at the nodes of the Gauss quadrature used later to approximate the inner products.

With variable coefficients and variable metric terms, the fluxes are not necessarily polynomials of degree N or less even if the solution is a polynomial of degree N . They can be approximated as polynomials of degree N in several ways. One can, for instance, approximate the contravariant flux as a polynomial of degree $2N$

$$\tilde{\mathbf{F}} = \tilde{\mathcal{A}} \mathbf{U} \in \mathbb{P}^{2N}. \quad (18)$$

by approximating the coefficient matrices as a polynomial of degree N

$$\tilde{\mathcal{A}} = \mathbb{I}^N \left(\mathbb{I}^N (J \vec{a}^i) \cdot \vec{\mathbf{A}} \right) = \mathbb{I}^N \left(\mathbb{I}^N (J \vec{a}^i) \cdot \mathbb{I}^N \left(\vec{\mathbf{A}} \right) \right) \in \mathbb{P}^N. \quad (19)$$

Alternatively, one could approximate the contravariant coefficient matrices by

$$\tilde{\mathcal{A}} = \mathbb{I}^N (J \vec{a}^i) \cdot \mathbb{I}^N \left(\vec{\mathbf{A}} \right) \in \mathbb{P}^{2N}, \quad (20)$$

leading to a contravariant flux

$$\tilde{\mathbf{F}} = \tilde{\mathcal{A}}\mathbf{U} \in \mathbb{P}^{3N}. \quad (21)$$

Finally, the usual (and underintegrated) approximation interpolates the entire flux,

$$\tilde{\mathbf{F}} = \mathbb{I}^N \left(\mathbb{I}^N (J\vec{a}^i) \cdot \vec{\underline{\mathbf{A}}}\mathbf{U} \right) \in \mathbb{P}^N. \quad (22)$$

If none of these is done, and $\vec{\underline{\mathbf{A}}}$ is not a polynomial function of the reference space variables, then the flux $\tilde{\mathbf{F}}$ is not a polynomial function approximation.

In the following, we will assume that

$$\mathbf{U} \in \mathbb{P}^N, \quad \tilde{\mathbf{F}} \in \mathbb{P}^{pN}, \quad (23)$$

for some p . As an important aside, the fluxes of the Euler equations of gas dynamics are not polynomials of the state vector of the conservative variables, and hence there is no exact representation of the fluxes as a polynomial when the state vector is a polynomial [5].

The normal surface flux in (17) is replaced by an interpolant, $\tilde{\mathbf{F}}^*$, of a numerical flux function, $\tilde{\mathbf{f}}^*(\mathbf{U}^L, \mathbf{U}^R)$, that is a function of the left and right states at the faces and is continuous across the face. The exact upwind numerical flux for the linear flux function is

$$\tilde{\mathbf{f}}^*(\mathbf{U}^L, \mathbf{U}^R) = \frac{\tilde{\underline{\mathbf{A}}} \cdot \hat{n} \mathbf{U}^L + \tilde{\underline{\mathbf{A}}} \cdot \hat{n} \mathbf{U}^R}{2} - \frac{1}{2} \left| \tilde{\underline{\mathbf{A}}} \cdot \hat{n} \right| (\mathbf{U}^R - \mathbf{U}^L) \equiv \tilde{\underline{\mathbf{A}}} \cdot \hat{n} \llbracket \mathbf{U} \rrbracket - \frac{1}{2} \left| \tilde{\underline{\mathbf{A}}} \cdot \hat{n} \right| \llbracket \mathbf{U} \rrbracket, \quad (24)$$

where \hat{n} is the reference space normal and the usual jump and average operators are $\llbracket \cdot \rrbracket$ and $\{\!\!\{ \cdot \}\!\!\}$, respectively.

The final equality in (24) will be true for the discrete approximation if the interpolant of the coefficient matrices, $\tilde{\underline{\mathbf{A}}}$, is continuous at the interfaces. We can ensure continuity if the interpolation points include the boundary points, as would be the case if the Gauss-Lobatto nodes are used as the interpolation points, but not if the Gauss points are used. For this reason, and the fact that they were used in [14], we consider only approximations that use the Gauss-Lobatto points in this paper. For simplicity, we define $\tilde{\underline{\mathbf{A}}}^\pm = \frac{1}{2} \left(\tilde{\underline{\mathbf{A}}} \cdot \hat{n} \pm \left| \tilde{\underline{\mathbf{A}}} \cdot \hat{n} \right| \right)$ so that $\tilde{\mathbf{f}}^*(\mathbf{U}^L, \mathbf{U}^R) = \tilde{\underline{\mathbf{A}}}^+ \mathbf{U}^L + \tilde{\underline{\mathbf{A}}}^- \mathbf{U}^R$. As with the interior fluxes, (18), (21), (22), the numerical surface flux would be a polynomial of degree $3N$, $2N$ or N depending on how it is approximated.

Inserting the polynomial approximations and the numerical surface flux gives us the exactly integrated discontinuous Galerkin approximation

$$\langle \mathbf{U}_t, \phi \rangle + \int_{\partial E} \mathbf{F}^{*,T} \phi dS - \left\langle \vec{\mathbf{F}}, \nabla \phi \right\rangle = 0, \quad (25)$$

where $\mathbf{F}^*, \vec{\mathbf{F}} \in \mathbb{P}^{pN}$.

Because the arguments are high order polynomials, the inner products are not integrated exactly in practice. Instead, they are approximated by quadrature. We write the Gauss-Lobatto quadrature over E as

$$\int_{E,N} g d\xi d\eta d\zeta \equiv \sum_{i,j,k=0}^N g_{ijk} \omega_i \omega_j \omega_k, \quad (26)$$

where ω_i , ω_j and ω_k are the quadrature weights for the i, j , and k coordinate directions, and $g_{ijk} = g(\xi_i, \eta_j, \zeta_k)$ are the values of g evaluated at the quadrature points. Two-dimensional surface integral approximations of a vector function \vec{g} are

$$\begin{aligned} \int_{\partial E, N} \vec{g} \cdot \hat{n} dS &\equiv \sum_{i,j=0}^N \omega_{ij} g^{(1)}(\xi, \eta_i, \zeta_j) \Big|_{\xi=-1}^1 + \sum_{i,j=0}^N \omega_{ij} g^{(2)}(\xi_i, \eta, \zeta_j) \Big|_{\eta=-1}^1 + \sum_{i,j=0}^N \omega_{ij} g^{(3)}(\xi_i, \eta_j, \zeta) \Big|_{\zeta=-1}^1 \\ &\equiv \int_N g^{(1)} d\eta d\zeta \Big|_{\xi=-1}^1 + \int_N g^{(2)} d\xi d\zeta \Big|_{\eta=-1}^1 + \int_N g^{(3)} d\xi d\eta \Big|_{\zeta=-1}^1, \end{aligned} \quad (27)$$

where $\omega_{ij} \equiv \omega_i \omega_j$, etc. Finally, we define the discrete inner product of two functions \mathbf{f} and \mathbf{g} and the discrete norm of \mathbf{f} from the quadrature

$$\langle \mathbf{f}, \mathbf{g} \rangle_{E,N} = \int_{E,N} \mathbf{f}^T \mathbf{g} d\xi d\eta d\zeta \equiv \sum_{i,j,k=0}^N \mathbf{f}_{ijk}^T \mathbf{g}_{ijk} \omega_{ijk}, \quad \|\mathbf{f}\|_{E,N} = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle_{E,N}}. \quad (28)$$

We often use the fact that the definition of the discrete inner product implies that

$$\langle \mathbf{f}, \mathbf{V} \rangle_{E,N} = \langle \mathbb{I}^N(\mathbf{f}), \mathbf{V} \rangle_{E,N} \quad \forall \mathbf{V} \in \mathbb{P}^N. \quad (29)$$

Stability analysis hangs on the fact that the Gauss-Lobatto quadrature rule satisfies a summation by parts property [9] and from that, the discrete Gauss law [12]

$$\left\langle \nabla \cdot \vec{\mathbf{F}}, \mathbf{V} \right\rangle_{E,N} = \int_{\partial E, N} \mathbf{V}^T \vec{\mathbf{F}} \cdot \hat{n} dS - \left\langle \vec{\mathbf{F}}, \nabla \mathbf{V} \right\rangle_{E,N}, \quad \forall \mathbf{V} \in \mathbb{P}^N. \quad (30)$$

With the definitions for the quadrature, we write a general DG spectral element approximation for (25), where the inner products are approximated by possibly different polynomial orders N, M, L by restricting $\phi \in \mathbb{P}^N$ and writing

$$\langle J\mathbf{U}, \phi \rangle_N + \int_{\partial E, L} \phi^T \tilde{\mathbf{F}}^* dS - \left\langle \tilde{\mathbf{F}}, \nabla \phi \right\rangle_M = 0, \quad (31)$$

where we have left off the subscript E to be understood in context. The form (31) is generally known as the “weak” form of the approximation.

4 Stability Analysis

We now study the stability of three approximations. The first is the standard approximation with $N = L = M$. We show that instability can be caused by large aliasing errors in severely under-resolved situations. The second is fully integrated, with $L = M > N$ chosen so that quadrature is exact and the approximation is stable. Finally, we examine the approximation where only the volume term is fully integrated, $L = N, M > N$ to see that for severely underresolved problems, interpolation error along the element surfaces can lead to instability.

4.1 Standard Approximation: $L = M = N$

The standard approximation with $L = M = N$ is equivalent to approximating $\tilde{\mathbf{F}} = \mathbb{I}^N (\tilde{\mathcal{A}}\mathbf{U}) \in \mathbb{P}^N$ using (22) by way of (29). Using quadrature of the same order for all of the inner products cannot be shown to be stable unless $\tilde{\mathbf{F}} = \tilde{\mathcal{A}}\mathbf{U} \in \mathbb{P}^N$ and $\tilde{\mathcal{A}}$ is a constant, which corresponds to constant coefficients and rectangular elements. When we set $\phi = \mathbf{U}$ in (31),

$$\langle J\mathbf{U}, \mathbf{U} \rangle_N + \int_{\partial E, N} \mathbf{U}^T \tilde{\mathbf{F}}^* dS - \langle \tilde{\mathbf{F}}, \nabla \mathbf{U} \rangle_N = 0, \quad (32)$$

or writing the first term as the time derivative of the J -weighted norm,

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J, N}^2 + \int_{\partial E, N} \mathbf{U}^T \tilde{\mathbf{F}}^* dS - \langle \tilde{\mathbf{F}}, \nabla \mathbf{U} \rangle_N = 0. \quad (33)$$

We then use the discrete Gauss law (30) on the volume term

$$\langle \tilde{\mathbf{F}}, \nabla \mathbf{U} \rangle_N + \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_N = \int_{\partial E, N} \mathbf{U}^T \tilde{\mathbf{F}} \cdot \hat{n} dS \quad (34)$$

to re-write (33) as

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J, N}^2 + \int_{\partial E, N} \mathbf{U}^T \left\{ \tilde{\mathbf{F}}^* - \tilde{\mathbf{F}} \cdot \hat{n} \right\} dS + \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_N = 0. \quad (35)$$

To account for the fact that the product rule does not in general hold for the interpolant of a product, we add and subtract

$$\frac{1}{2} \langle \tilde{\mathcal{A}} \cdot \nabla \mathbf{U} + (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U}, \mathbf{U} \rangle_N, \quad (36)$$

where $\tilde{\mathcal{A}} = \mathbb{I}^N \left(\mathbb{I}^N (J\vec{a}^i) \cdot \vec{\mathcal{A}} \right) \in \mathbb{P}^N$ to write

$$\begin{aligned} \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_N &= \frac{1}{2} \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_N + \frac{1}{2} \langle \tilde{\mathcal{A}} \cdot \nabla \mathbf{U}, \mathbf{U} \rangle_N + \frac{1}{2} \langle \nabla \cdot \tilde{\mathcal{A}} \mathbf{U}, \mathbf{U} \rangle_N \\ &\quad - \frac{1}{2} \langle \tilde{\mathcal{A}} \cdot \nabla \mathbf{U} + (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U} - \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_N. \end{aligned} \quad (37)$$

Now the discrete Gauss law (30), the collocation of the solution and flux, and symmetry of $\tilde{\mathcal{A}}$ imply that

$$\begin{aligned} \langle \tilde{\mathcal{A}} \cdot \nabla \mathbf{U}, \mathbf{U} \rangle_N &= \int_{\partial E, N} \mathbf{U}^T \tilde{\mathbf{F}} dS - \langle \mathbf{U}, \nabla \cdot \mathbb{I}^N (\tilde{\mathcal{A}} \mathbf{U}) \rangle_N \\ &= \int_{\partial E, N} \mathbf{U}^T \tilde{\mathbf{F}} dS - \langle \mathbf{U}, \nabla \cdot \tilde{\mathbf{F}} \rangle_N. \end{aligned} \quad (38)$$

Therefore,

$$\begin{aligned} \langle \tilde{\mathbf{F}}, \nabla \mathbf{U} \rangle_N &= \frac{1}{2} \int_{\partial E, N} \mathbf{U}^T \tilde{\mathbf{F}} \cdot \hat{n} dS - \frac{1}{2} \langle (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U}, \mathbf{U} \rangle_N \\ &\quad + \frac{1}{2} \langle \tilde{\mathcal{A}} \cdot \nabla \mathbf{U} + (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U} - \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_N. \end{aligned} \quad (39)$$

Then (33) becomes

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J,N}^2 &= - \int_{\partial E,N} \mathbf{U}^T \left\{ \tilde{\mathbf{F}}^* - \frac{1}{2} \tilde{\mathbf{F}} \cdot \hat{n} \right\} dS \\ &\quad - \frac{1}{2} \left\langle \left(\nabla \cdot \tilde{\mathcal{A}} \right) \mathbf{U}, \mathbf{U} \right\rangle_N \\ &\quad + \frac{1}{2} \left\langle \tilde{\mathcal{A}} \cdot \nabla \mathbf{U} + \left(\nabla \cdot \tilde{\mathcal{A}} \right) \mathbf{U} - \nabla \cdot \mathbb{I}^N \left(\tilde{\mathbf{F}} \right), \mathbf{U} \right\rangle_N. \end{aligned} \quad (40)$$

The last inner product in (40) is the projection of the amount by which the product rule is not satisfied by polynomial functions. Note that if $\tilde{\mathcal{A}}$ is constant, which happens when the original PDE is constant coefficient and the elements are rectangular, then $\tilde{\mathbf{F}} = \mathbb{I}^N \left(\tilde{\mathcal{A}} \mathbf{U} \right) = \tilde{\mathcal{A}} \mathbf{U}$ and the last term vanishes. Otherwise the last term can be positive or negative.

The product rule error term can be bounded as

$$\left\langle \tilde{\mathcal{A}} \cdot \nabla \mathbf{U} + \left(\nabla \cdot \tilde{\mathcal{A}} \right) \mathbf{U} - \nabla \cdot \mathbb{I}^N \left(\tilde{\mathbf{F}} \right), \mathbf{U} \right\rangle_N \leq 2\varepsilon_A \|\mathbf{U}\|_{J,N}^2, \quad (41)$$

where

$$\varepsilon_A = \frac{1}{2} \max_{E, \|\mathbf{U}\|_{J,N} \neq 0} \frac{\left\| \tilde{\mathcal{A}} \cdot \nabla \mathbf{U} + \left(\nabla \cdot \tilde{\mathcal{A}} \right) \mathbf{U} - \nabla \cdot \mathbb{I}^N \left(\tilde{\mathbf{F}} \right) \right\|_2^2}{J \|\mathbf{U}\|_{J,N}^2}. \quad (42)$$

Note also that the sign of $\nabla \cdot \tilde{\mathcal{A}}$ can be positive or negative and that

$$- \left\langle \left(\nabla \cdot \tilde{\mathcal{A}} \right) \mathbf{U}, \mathbf{U} \right\rangle_N \leq 2\hat{\gamma} \|\mathbf{U}\|_{J,N} \quad (43)$$

where

$$\hat{\gamma} = \max_E \frac{\left\| \nabla \cdot \tilde{\mathcal{A}} \right\|_2}{J}. \quad (44)$$

Therefore, we can bound the elemental energy in (40) as

$$\frac{d}{dt} \|\mathbf{U}\|_{J,N}^2 \leq -2 \int_{\partial E,N} \mathbf{U}^T \left\{ \tilde{\mathbf{F}}^* - \frac{1}{2} \tilde{\mathbf{F}} \cdot \hat{n} \right\} dS + 2(\hat{\gamma} + \varepsilon_A) \|\mathbf{U}\|_{J,N}^2. \quad (45)$$

Equation (45) shows how the energy changes on a single element. Summing over all elements e^k gives the total energy change

$$\begin{aligned} \frac{d}{dt} \sum_{k=1}^{N_{el}} \|\mathbf{U}^k\|_{J,N}^2 &\leq 2 \sum_{\substack{\text{Interior} \\ \text{Faces}}} \int_{\partial E,N} \left\{ \mathbf{F}^{*,T} \llbracket \mathbf{U} \rrbracket - \frac{1}{2} \left[\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right] \right\} dS - \text{PBT} \\ &\quad + 2(\hat{\gamma} + \varepsilon_A) \sum_{n=1}^{N_{el}} \|\mathbf{U}^k\|_{J,N}^2, \end{aligned} \quad (46)$$

where $\hat{\gamma} = \max_k \hat{\gamma}^k$, $\varepsilon_A = \max_k \hat{\varepsilon}_A^k$, and PBT represents the physical boundary terms [11],

$$\begin{aligned} \text{PBT} &= 2 \sum_{\substack{\text{Boundary} \\ \text{Faces}}} \int_{\partial E, N} \left(\tilde{\mathbf{F}}^* - \frac{1}{2} \mathbf{F} \cdot \hat{n} \right)^T \mathbf{U} dS \\ &= \sum_{\substack{\text{Boundary} \\ \text{Faces}}} \int_{\partial E, N} \left\{ \mathbf{U}^T \tilde{\mathcal{A}}^+ \mathbf{U} + (\mathbf{U} - \mathbf{g})^T \left| \tilde{\mathcal{A}}^- \right| (\mathbf{U} - \mathbf{g}) - \mathbf{g}^T \left| \tilde{\mathcal{A}}^- \right| \mathbf{g} \right\} dS. \end{aligned} \quad (47)$$

The middle term in the PBT represents additional damping due to the weak imposition of the boundary conditions through the numerical flux. The interior face contributions satisfy [10],

$$\tilde{\mathbf{F}}^{*T} \llbracket \mathbf{U} \rrbracket - \frac{1}{2} \llbracket \tilde{\mathbf{F}}^T \mathbf{U} \rrbracket = -\frac{1}{2} \llbracket \mathbf{U} \rrbracket^T \left| \tilde{\mathcal{A}} \cdot \hat{n} \right| \llbracket \mathbf{U} \rrbracket \leq 0, \quad (48)$$

pointwise.

If we define the discrete broken norm

$$\|\mathbf{U}\|_N^2 = \sum_{k=1}^K \|\mathbf{U}^k\|_{J,N}^2, \quad (49)$$

then the time derivative of the energy over the whole domain satisfies

$$\frac{d}{dt} \|\mathbf{U}\|_N^2 \leq -\text{PBT} - \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E, N} \llbracket \mathbf{U} \rrbracket^T \left| \tilde{\mathcal{A}} \cdot \hat{n} \right| \llbracket \mathbf{U} \rrbracket dS + 2(\hat{\gamma} + \varepsilon_a) \|\mathbf{U}\|_N^2. \quad (50)$$

Equation (50) matches that of the continuous analysis, (5), except for the product rule error term. If no energy is added through the boundaries, i.e. if $\mathbf{g} = 0$, and if $\hat{\gamma} = 0$ when the same is true of the continuous problem, then

$$\begin{aligned} \frac{d}{dt} \|\mathbf{U}\|_N^2 &\leq - \sum_{\substack{\text{Boundary} \\ \text{Faces}}} \int_{\partial E, N} \left\{ \mathbf{U}^T \tilde{\mathcal{A}}^+ \mathbf{U} + \mathbf{U}^T \left| \tilde{\mathcal{A}}^- \right| \mathbf{U} \right\} dS \\ &\quad - \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E, N} \llbracket \mathbf{U} \rrbracket^T \left| \tilde{\mathcal{A}} \right| \llbracket \mathbf{U} \rrbracket dS \\ &\quad + 2\varepsilon_A \|\mathbf{U}\|_N^2. \end{aligned} \quad (51)$$

We see in (51) that if the dissipation due to the internal face jumps and the weak imposition of the physical boundary conditions are not sufficiently large, then the approximation can grow in time with a growth rate, ε_A , that depends on the size of the aliasing error.

4.2 Fully Overintegrated: $L = M, M > N$

The idea behind overintegration is to approximate the integrals by quadrature of order $M > N$ sufficiently large so as to get a stable approximation. The formal statement of the weak form fully

overintegrated approximation is

$$\langle J\mathbf{U}_t, \phi \rangle_N + \int_{\partial E, M} \tilde{\mathbf{F}}^{*,T} \phi dS - \langle \tilde{\mathbf{F}}, \nabla \phi \rangle_M = 0. \quad (52)$$

Taking $\phi = \mathbf{U}$,

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J,N}^2 + \int_{\partial E, M} \tilde{\mathbf{F}}^{*,T} \mathbf{U} dS - \langle \tilde{\mathbf{F}}, \nabla \mathbf{U} \rangle_M = 0. \quad (53)$$

Using the discrete extended Gauss law gives the strong form

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J,N}^2 + \int_{\partial E, M} \mathbf{U}^T \left\{ \tilde{\mathbf{F}}^* - \tilde{\mathbf{F}} \cdot \hat{n} \right\} dS + \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_M = 0. \quad (54)$$

In [7] it was suggested to find M sufficiently large so that the quadrature error is eliminated. We note that if $\tilde{\mathcal{A}} \in \mathbb{P}^N$ (see (19)), then $(\nabla \cdot \tilde{\mathbf{F}})^T \mathbf{U} = (\nabla \cdot (\tilde{\mathcal{A}}\mathbf{U}))^T \mathbf{U} \in \mathbb{P}^{3N}$ so that

$$\langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_M = \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle \quad (55)$$

if $2M - 1 = 3N$, i.e., if $M > 3N/2$. Alternatively, with the flux approximation (21), $2M - 1 = 4N$ and we must take $M > 2N$ for the quadrature to be exact.

The amount by which the quadratures need to be overintegrated for *stability* is greater than that needed to eliminate just the quadrature error. As long as $M \geq N$, the inner product on the right in (54) satisfies the summation by parts property [9] and so

$$\langle \mathbb{I}^M (\tilde{\mathcal{A}}\mathbf{U}), \nabla \mathbf{U} \rangle_M = \int_{\partial E, M} \mathbf{U}^T \mathbb{I}^M (\tilde{\mathcal{A}} \cdot \hat{n} \mathbf{U}) dS - \langle \nabla \cdot \mathbb{I}^M (\tilde{\mathcal{A}}\mathbf{U}), \mathbf{U} \rangle_M. \quad (56)$$

However, the crucial step needed to prove stability is to commute the differentiation and interpolation in $\nabla \cdot \mathbb{I}^M (\tilde{\mathcal{A}}\mathbf{U})$ and use the product rule [12]. The product rule holds if the interpolation is exact, i.e. $\mathbb{I}^M (\tilde{\mathcal{A}}\mathbf{U}) = \tilde{\mathcal{A}}\mathbf{U}$, which occurs if $M \geq 2N$ if $\tilde{\mathcal{A}} \in \mathbb{P}^N$ and $M \geq 3N$ if $\tilde{\mathcal{A}} \in \mathbb{P}^{2N}$.

With M sufficiently large so that $\tilde{\mathbf{F}} = \mathbb{I}^M (\tilde{\mathcal{A}}\mathbf{U}) = \tilde{\mathcal{A}}\mathbf{U}$,

$$\langle \nabla \cdot \mathbb{I}^M (\tilde{\mathcal{A}}\mathbf{U}), \mathbf{U} \rangle_M = \langle \nabla \cdot (\tilde{\mathcal{A}}\mathbf{U}), \mathbf{U} \rangle_M = \langle (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U} + \tilde{\mathcal{A}} \cdot \nabla \mathbf{U}, \mathbf{U} \rangle_M. \quad (57)$$

Therefore, (c.f. (38))

$$\begin{aligned} \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_M &= \frac{1}{2} \langle \nabla \cdot \tilde{\mathbf{F}}, \mathbf{U} \rangle_M + \frac{1}{2} \langle (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U} + \tilde{\mathcal{A}} \cdot \nabla \mathbf{U}, \mathbf{U} \rangle_M \\ &= \frac{1}{2} \int_{\partial E, M} \mathbf{U}^T \tilde{\mathbf{F}} dS + \frac{1}{2} \langle (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U}, \mathbf{U} \rangle_M \end{aligned} \quad (58)$$

and the time derivative of the local energy is

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J,N}^2 + \int_{\partial E, M} \left\{ \tilde{\mathbf{F}}^* - \frac{1}{2} \tilde{\mathbf{F}} \cdot \hat{n} \right\}^T \mathbf{U} dS = -\frac{1}{2} \langle \mathbf{U}, (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U} \rangle_M. \quad (59)$$

The term on the right of (59) can be bounded using the equivalence of the discrete and continuous norms [3] by

$$\begin{aligned} -\langle \mathbf{U}, (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U} \rangle_M &= -\langle \mathbf{U}, (\nabla \cdot \tilde{\mathcal{A}}) \mathbf{U} \rangle \leq C \max_E \|\nabla \cdot \tilde{\mathcal{A}}\|_2 \langle \mathbf{U}, \mathbf{U} \rangle_N \\ &\leq C \frac{\max_E \|\nabla \cdot \tilde{\mathcal{A}}\|_2}{\min_E J} (J\mathbf{U}, \mathbf{U})_N \equiv 2\hat{\gamma}_{OI} \|\mathbf{U}\|_{J,N}^2. \end{aligned} \quad (60)$$

Then for element e^k ,

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}^k\|_{J,N}^2 \leq - \int_{\partial E, M} \left\{ \tilde{\mathbf{F}}^* - \frac{1}{2} \tilde{\mathbf{F}}^k \cdot \hat{n} \right\}^T \mathbf{U}^k dS + \frac{1}{2} 2\hat{\gamma}_{OI}^k \|\mathbf{U}^k\|_{J,N}^2. \quad (61)$$

The surface fluxes are of the same form as for the Standard Approximation, see (45). Therefore, when we sum over all of the elements

$$\frac{d}{dt} \|\mathbf{U}\|_N^2 \leq -PBT - BI + 2\hat{\gamma}_{OI} \|\mathbf{U}\|_N^2, \quad (62)$$

where, now

$$PBT = \sum_{\substack{\text{Boundary} \\ \text{Faces}}} \int_{\partial E, M} \left\{ \mathbf{U}^T \tilde{\mathcal{A}}^+ \mathbf{U} + (\mathbf{U} - \mathbf{g})^T |\tilde{\mathcal{A}}^-| (\mathbf{U} - \mathbf{g}) - \mathbf{g}^T |\tilde{\mathcal{A}}^-| \mathbf{g} \right\} dS, \quad (63)$$

$\tilde{\mathcal{A}}^\pm = \tilde{\mathcal{A}} \cdot \hat{n} \pm |\tilde{\mathcal{A}} \cdot \hat{n}|$ and

$$BI = \int_{\partial E, M} \llbracket \mathbf{U} \rrbracket^T |\tilde{\mathcal{A}}| \llbracket \mathbf{U} \rrbracket dS \geq 0, \quad (64)$$

again, provided that the interpolant of the coefficient matrix is continuous across element interfaces.

In the special case where the energy should not grow, i.e. where $\mathbf{g} = 0$ and $\nabla \cdot \tilde{\mathcal{A}} = 0$, the evolution of the energy of the solution is governed by

$$\frac{d}{dt} \|\mathbf{U}\|_N^2 \leq - \sum_{\substack{\text{Boundary} \\ \text{Faces}}} \int_{\partial E, M} \left\{ \mathbf{U}^T \tilde{\mathcal{A}}^+ \mathbf{U} + \mathbf{U}^T |\tilde{\mathcal{A}}^-| \mathbf{U} \right\} dS - \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E, M} \llbracket \mathbf{U} \rrbracket^T |\tilde{\mathcal{A}}| \llbracket \mathbf{U} \rrbracket dS \leq 0 \quad (65)$$

and does not grow in time, implying stability.

In summary, when the flux is a polynomial of its argument, and the volume and the surface integrals are approximated to a sufficiently high order, $L = M > 2N$ for $\tilde{\mathcal{A}} \in \mathbb{P}^N$ and $L = M > 3N$ for $\tilde{\mathcal{A}} \in \mathbb{P}^{2N}$, then the product rule error seen in standard, underintegrated approximation, does not appear, and the overintegrated approximation is stable.

4.3 Volume Only OverIntegrated: $L = N, M > N$

As was originally suggested, overintegration could be performed to eliminate the product rule error term ε_A in the volume, but not along the faces [14]. In the partially overintegrated approximation,

we can construct two more “strong” forms, one of which differs from the “weak” form. *In the following, we will assume that $\mathbf{g} = 0$ and $\nabla \cdot \tilde{\mathcal{A}} = 0$ so that any growth in the energy represents instability.*

The strong and weak forms of the approximations are no longer necessarily algebraically equivalent, so we must analyze the two separately. The weak form approximation with the surface terms underintegrated is

$$[W] \quad \langle J\mathbf{U}, \phi \rangle_N + \int_{\partial E, N} \phi^T \tilde{\mathbf{F}}^* dS - \langle \tilde{\mathbf{F}}, \nabla \phi \rangle_M = 0. \quad (66)$$

We can also construct two “strong” form approximations. We can integrate by parts first and then apply quadrature. Starting from (25), applying the Gauss law and then replacing integrals by quadrature gives the first strong form

$$[S1] \quad \langle J\mathbf{U}_t, \phi \rangle + \int_{\partial E, N} \mathbb{I}^N \left(\left(\tilde{\mathbf{F}}^* - \tilde{\mathbf{F}} \cdot \hat{n} \right)^T \phi \right) dS + \langle \nabla \cdot \tilde{\mathbf{F}}, \phi \rangle_M = 0. \quad (67)$$

As an aside, form [S1] is not particularly interesting in practice as it is not conservative if $N \neq M$. The second form we get by applying the discrete Gauss Law to the weak form, (66), giving

$$[S2] \quad \langle J\mathbf{U}_t, \phi \rangle_N + \int_{\partial E, N} \mathbb{I}^N \left(\tilde{\mathbf{F}}^{*,T} \phi \right) dS - \int_{\partial E, M} \mathbf{F}^{*,T} \phi dS + \langle \nabla \cdot \tilde{\mathbf{F}}, \phi \rangle_M = 0 \quad (68)$$

The approximations (67) and (68) are not the same unless $N = M$. Note that (68) is algebraically equivalent to the weak form (66), $[W] \Leftrightarrow [S2]$, (c.f. [9]). Therefore we need to analyze the stability of [S1] and one of [W] or [S2].

4.3.1 Stability of the Approximation [S1]

To find the stability property of [S1], we replace ϕ by \mathbf{U} , as usual. The volume term can then be replaced as in Sec. 4.2 so that the elemental energy is governed by

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J, N}^2 + \int_{\partial E, N} \mathbb{I}^N \left(\left(\mathbf{F}^* - \tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right) dS + \frac{1}{2} \int_{\partial E, M} \left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} dS = 0. \quad (69)$$

We can add and subtract terms to see the contribution from the inconsistent integration. Since M is chosen so that the quadrature is exact, (69) is equivalent to

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J, N}^2 + \int_{\partial E, N} \mathbb{I}^N \left(\left(\mathbf{F}^* - \frac{1}{2} \tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right) dS \\ + \frac{1}{2} \int_{\partial E} \left\{ \left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} - \mathbb{I}^N \left(\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right) \right\} dS = 0. \end{aligned} \quad (70)$$

Summing over all of the elements gives

$$\frac{d}{dt} \sum_{k=1}^K \|\mathbf{U}^k\|_{J, N}^2 = -BI - PBT, \quad (71)$$

where the interior interface contributions are now

$$\begin{aligned}
BI = & -2 \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E, N} \left\{ \mathbf{F}^{*,T} \llbracket \mathbf{U} \rrbracket - \frac{1}{2} \left[\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right] \right\} dS \\
& - \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E} \left\{ \left[\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right] - \left[\mathbb{I}^N \left(\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right) \right] \right\} dS
\end{aligned} \tag{72}$$

and the physical boundary terms are now

$$\begin{aligned}
PBT = & 2 \sum_{\substack{\text{Boundary} \\ \text{faces}}} \int_{\partial E, N} \left(\mathbf{F}^* - \frac{1}{2} \tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} dS \\
& + \sum_{\substack{\text{Boundary} \\ \text{faces}}} \int_{\partial E} \left\{ \left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} - \mathbb{I}^N \left(\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right) \right\} dS.
\end{aligned} \tag{73}$$

We examine the interior boundary contributions first. From the algebraic identity,

$$\left[\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \mathbf{U} \right] = \left\{ \left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \right\} \llbracket \mathbf{U} \rrbracket + \left[\left(\tilde{\mathbf{F}} \cdot \hat{n} \right)^T \right] \{\!\!\{ \mathbf{U} \}\!\!\}, \tag{74}$$

and the particular linear structure of the flux,

$$\left[\left(\tilde{\mathbf{F}} \cdot \hat{n} \right) \mathbf{U} \right]^T = 2 \{\!\!\{ \mathbf{U} \}\!\!\}^T \tilde{\mathcal{A}} \llbracket \mathbf{U} \rrbracket \tag{75}$$

at each surface point. Then using (48), and the fact that the interpolant of the jumps is equal to the jump of the interpolants, we see that the interior interfaces have a dissipative component and an interpolation error component whose sign is indeterminate

$$\begin{aligned}
BI = & \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E, N} \llbracket \mathbf{U} \rrbracket^T \left| \tilde{\mathcal{A}} \cdot \hat{n} \right| \llbracket \mathbf{U} \rrbracket dS \\
& - \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E} \left\{ \{\!\!\{ \mathbf{U} \}\!\!\}^T \tilde{\mathcal{A}} \cdot \hat{n} \llbracket \mathbf{U} \rrbracket - \mathbb{I}^N \left(\{\!\!\{ \mathbf{U} \}\!\!\}^T \tilde{\mathcal{A}} \cdot \hat{n} \llbracket \mathbf{U} \rrbracket \right) \right\} dS.
\end{aligned} \tag{76}$$

A similar error is introduced at the physical boundaries. Following (47),

$$\begin{aligned}
PBT = & \sum_{\substack{\text{Boundary} \\ \text{faces}}} \int_{\partial E, N} \left(\mathbf{U}^T \tilde{\mathcal{A}}^+ \mathbf{U} + \mathbf{U}^T \left| \tilde{\mathcal{A}}^- \right| \mathbf{U} \right) dS \\
& + \sum_{\substack{\text{Boundary} \\ \text{faces}}} \int_{\partial E} \left\{ \mathbf{U}^T \tilde{\mathcal{A}} \cdot \hat{n} \mathbf{U} - \mathbb{I}^N \left(\mathbf{U}^T \tilde{\mathcal{A}} \cdot \hat{n} \mathbf{U} \right) \right\} dS.
\end{aligned} \tag{77}$$

When we insert (76) and (77) into (71),

$$\begin{aligned}
\frac{d}{dt} \|\mathbf{U}\|_N^2 = & - \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E, N} \llbracket \mathbf{U} \rrbracket^T \left| \tilde{\mathcal{A}} \cdot \hat{n} \right| \llbracket \mathbf{U} \rrbracket dS \\
& + \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E} \left\{ \llbracket \mathbf{U} \rrbracket^T \tilde{\mathcal{A}} \cdot \hat{n} \llbracket \mathbf{U} \rrbracket - \mathbb{I}^N \left(\llbracket \mathbf{U} \rrbracket^T \tilde{\mathcal{A}} \cdot \hat{n} \llbracket \mathbf{U} \rrbracket \right) \right\} dS \\
& - \sum_{\substack{\text{Boundary} \\ \text{faces}}} \int_{\partial E, N} \left(\mathbf{U}^T \tilde{\mathcal{A}}^+ \mathbf{U} + \mathbf{U}^T \left| \tilde{\mathcal{A}}^- \right| \mathbf{U} \right) dS \\
& - \sum_{\substack{\text{Boundary} \\ \text{faces}}} \int_{\partial E} \left\{ \mathbf{U}^T \tilde{\mathcal{A}} \cdot \hat{n} \mathbf{U} - \mathbb{I}^N \left(\mathbf{U}^T \tilde{\mathcal{A}} \cdot \hat{n} \mathbf{U} \right) \right\} dS
\end{aligned} \tag{78}$$

We see that the approximation has element face dissipation plus interpolation error terms of indeterminate sign that come from the fact that the inconsistent quadrature does not represent the surface fluxes exactly. If that interpolation error is large, i.e. in severely under resolved approximations, then it is possible that the right hand side of (78) is positive and the energy of the solution grows in time.

4.3.2 Stability of the Approximation $[S2] \Leftrightarrow [W]$

Again, we replace ϕ with \mathbf{U} , this time in (66). Then

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J, N}^2 + \int_{\partial E, N} \mathbf{U}^T \tilde{\mathbf{F}}^* dS - \frac{1}{2} \int_{\partial E, M} \mathbf{U}^T \tilde{\mathbf{F}} \cdot \hat{n} dS = 0. \tag{79}$$

We add and subtract terms and use exactness of the consistently integrated quadrature to rewrite (79) as

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{U}\|_{J, N}^2 + \int_{\partial E, N} \mathbf{U}^T \left(\tilde{\mathbf{F}}^* - \frac{1}{2} \tilde{\mathbf{F}} \cdot \hat{n} \right) dS - \frac{1}{2} \int_{\partial E} \left\{ \mathbf{U}^T \tilde{\mathbf{F}} \cdot \hat{n} - \mathbb{I}^N \left(\mathbf{U}^T \tilde{\mathbf{F}} \cdot \hat{n} \right) \right\} dS = 0. \tag{80}$$

The only difference between (70) and (80) is in the sign of the interpolation error term. Therefore the energy analysis of the previous section applies to the second strong form approximation in the same way.

4.3.3 Analysis of the Element Boundary Contributions

We first show that the errors due to inconsistent integration of the surface terms are due to aliasing of the modes $2N, \dots, pN$, where $p = 3$ or $p = 4$, depending on how the coefficient matrices are approximated. The interpolation errors in (78) are of the form

$$\int_{\partial E} \{ V - \mathbb{I}^N(V) \} dS, \tag{81}$$

where $V \in \mathbb{P}^{pN}$. To simplify the analysis, we will consider two dimensional geometries where the element faces are the element boundary curves. Then along each side, the surface integral reduces to

$$\epsilon \equiv \int_{-1}^1 \{V(\xi) - \mathbb{I}^N(V(\xi))\} d\xi = \langle V - \mathbb{I}^N(V), 1 \rangle = \langle V - \mathbb{I}^N(V), L_0 \rangle, \quad (82)$$

where $L_0 = 1$ is the Legendre polynomial of degree zero. In modal form,

$$V = \sum_{k=0}^{pN} \hat{V}_k L_k(\xi) \quad (83)$$

where $\hat{V}_k = \langle V, L_k \rangle / \|L_k\|^2$, $k = 0, 1, \dots, pN$ are the modal coefficients of the polynomial and L_k is the Legendre polynomial of degree k . The modal representation of the interpolant of V is

$$\mathbb{I}^N(V) = \sum_{k=0}^N \bar{V}_k L_k(\xi), \quad (84)$$

where [3]

$$\bar{V}_k = \hat{V}_k + a_k, \quad (85)$$

and the a_k are the aliases of the true coefficients

$$a_k = \frac{1}{\|L_k\|_N^2} \sum_{n=N+1}^{pN} \langle L_n, L_k \rangle_N \hat{V}_n. \quad (86)$$

Then

$$V - \mathbb{I}^N(V) = \sum_{k=N+1}^{pN} \hat{V}_k L_k - \sum_{k=0}^N a_k L_k. \quad (87)$$

Orthogonality of the Legendre polynomials causes most of the terms in (82) to vanish, leaving

$$\epsilon = -2a_0 = - \sum_{n=N+1}^{pN} \langle L_n, L_0 \rangle_N \hat{V}_n. \quad (88)$$

Finally, $\langle L_n, L_0 \rangle_N = \langle L_n, L_0 \rangle = 0$ for $n = N+1, N+2, \dots, 2N-1$ so

$$\int_{-1}^1 \{V(\xi) - \mathbb{I}^N(V(\xi))\} d\xi = - \sum_{n=2N}^{pN} \langle L_n, L_0 \rangle_N \hat{V}_n. \quad (89)$$

There is no error if, in fact, $V \in \mathbb{P}^{2N-1}$. Otherwise, the error is due to aliasing of the modes from $2N$ to pN .

It remains, then, to see if the aliasing error is smaller or larger than the dissipation due to the numerical flux. First we note that the internal interface dissipation in (78) is $O(\|U\|^2)$, whereas the quadrature aliasing error is $O(\|U\|)$. In lieu of finding the coefficients of a triple product, we illustrate the dissipation and the aliasing errors along an interior edge

$$- \int_{-1,N}^1 \llbracket \mathbf{U} \rrbracket^T \left| \tilde{\mathcal{A}} \cdot \hat{n} \right| \llbracket \mathbf{U} \rrbracket d\xi + \int_{-1}^1 \left\{ \llbracket \mathbf{U} \rrbracket^T \tilde{\mathcal{A}} \cdot \hat{n} \llbracket \mathbf{U} \rrbracket - \mathbb{I}^N \left(\llbracket \mathbf{U} \rrbracket^T \tilde{\mathcal{A}} \cdot \hat{n} \llbracket \mathbf{U} \rrbracket \right) \right\} d\xi \quad (90)$$

for a scalar problem with

$$\begin{aligned}\llbracket U \rrbracket &= \alpha(1 + \xi)^{q/3} \\ \llbracket U \rrbracket &= \beta(1 + \xi)^{q/3} \\ \tilde{\mathcal{A}} \cdot \hat{n} &= \gamma(1 + \xi)^{q/3}.\end{aligned}\tag{91}$$

With this choice, an interior edge contribution to the energy is

$$-\alpha^2 |\gamma| \int_{-1, N}^1 (1 + \xi)^q d\xi + \alpha\beta\gamma \int_{-1}^1 \{(1 + \xi)^q - \mathbb{I}^N((1 + \xi)^q)\} d\xi.\tag{92}$$

As noted above, if $q \leq 2N - 1$ the aliasing contribution should vanish and the edge contribution is dissipative.

If q is large, then the edge contributions can be destabilizing when the aliasing errors are large enough, i.e. in underresolved problems. As a concrete example, we choose $\alpha = 10^{-3}$, $\beta = 1$ and $\gamma = -1$. Table 1 shows the edge dissipation, aliasing contribution and total dissipation for $q = 18$ and approximation orders N between three and 16. We see that for $N \leq 5$, the aliasing error dominates the dissipation due to upwinding and the combined contribution is positive. As the resolution increases, the aliasing error decays exponentially fast (as seen in a semi-log plot of that term vs N) so that the overall contribution is dissipative for $6 \leq N \leq 17$. Finally, for $2N - 1 \geq q$ the aliasing error vanishes as expected and the only edge contribution to the energy comes from the upwinding. Table 1 is therefore an illustration that severe underresolution can lead to destabilizing aliasing errors that might not be significant as the resolution increases.

N	$2N-1$	$-\alpha^2 \gamma \int_{-1, N}^1 (1 + \xi)^q d\xi$	$\alpha\beta\gamma \int_{-1}^1 \{(1 + \xi)^q - \mathbb{I}^N((1 + \xi)^q)\} d\xi$	Sum
3	5	-4.434E-02	1.674E+01	1.670E+01
4	7	-3.092E-02	3.327E+00	3.296E+00
5	9	-2.799E-02	3.967E-01	3.687E-01
6	11	-2.762E-02	2.560E-02	-2.018E-03
7	13	-2.759E-02	7.737E-04	-2.682E-02
8	15	-2.759E-02	8.237E-06	-2.759E-02
9	17	-2.759E-02	1.297E-08	-2.759E-02
10	19	-2.759E-02	-2.310E-15	-2.759E-02
11	21	-2.759E-02	1.102E-14	-2.759E-02
12	23	-2.759E-02	1.826E-14	-2.759E-02
13	25	-2.759E-02	-2.218E-14	-2.759E-02

Table 1: Interior interface dissipation along a single edge as a function of polynomial order, N , for $q = 18$. Bold entries predict instability. The horizontal line marks where the aliasing error associated with the underintegration of the boundary terms vanishes.

5 Conclusions

To summarize the results, let us gather the boundary and interface dissipation terms as

$$Dissip \equiv \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E, N} [\mathbf{U}]^T \left| \tilde{\mathcal{A}} \cdot \hat{n} \right| [\mathbf{U}] dS + \sum_{\substack{\text{Boundary} \\ \text{Faces}}} \int_{\partial E, N} \left\{ \mathbf{U}^T \tilde{\mathcal{A}}^+ \mathbf{U} + \mathbf{U}^T \left| \tilde{\mathcal{A}}^- \right| \mathbf{U} \right\} dS \geq 0.$$

Then for linear problems where there should be no energy growth, the energy of the standard, the partially (volume only) overintegrated, and the fully overintegrated approximations satisfy

$$[\text{Standard}] \quad \frac{d}{dt} \|\mathbf{U}\|_N^2 \leq -Dissip + 2\varepsilon_a \|\mathbf{U}\|_N^2, \quad (93)$$

$$\begin{aligned} [\text{Volume Overintegrated}] \quad \frac{d}{dt} \|\mathbf{U}\|_N^2 = & -Dissip \\ & \pm \sum_{\substack{\text{interior} \\ \text{faces}}} \int_{\partial E} \left\{ \{\mathbf{U}\}^T \tilde{\mathcal{A}} \cdot \hat{n} [\mathbf{U}] - \mathbb{I}^N \left(\{\mathbf{U}\}^T \tilde{\mathcal{A}} \cdot \hat{n} [\mathbf{U}] \right) \right\} dS \\ & \pm \sum_{\substack{\text{Boundary} \\ \text{faces}}} \int_{\partial E} \left\{ \mathbf{U}^T \tilde{\mathcal{A}} \cdot \hat{n} \mathbf{U} - \mathbb{I}^N \left(\mathbf{U}^T \tilde{\mathcal{A}} \cdot \hat{n} \mathbf{U} \right) \right\} dS, \end{aligned} \quad (94)$$

where the \pm depends on which strong form is used, and

$$[\text{Fully Overintegrated}] \quad \frac{d}{dt} \|\mathbf{U}\|_N^2 \leq -Dissip \leq 0 \quad (95)$$

provided that M is sufficiently large so that the product rule holds.

The results support the finding of [14] that overintegrating the surface and volume integrals leads to a more robust approximation. The standard approximation has a growth term whose growth rate factor, ε_a , depends on aliasing errors associated with the amount by which the product rule fails to hold for the interpolants of products of polynomials. That term can lead to exponential growth if the dissipation terms associated with the boundary and interface conditions are not sufficiently large, which, in a sense, is a definition of “severely underresolved”. The volume only overintegrated approximation eliminates that aliasing term associated with the volume, so there is no ε_a term, but it introduces aliasing terms along the element faces. In severely underresolved problems, these aliasing terms could also be large enough to destabilize the approximation. Using consistent integration for both the surfaces and the volume leads to a stable approximation where the energy of the solution is nonincreasing.

As a final comment, we note that none of these approximations are dealiased, yet some are stable. There is aliasing in the discrete norm of the solution stemming from the fact that the argument of $\|\mathcal{J}\mathbf{U}\|_N$ will be a polynomial of degree $3N$, which cannot be approximated exactly with the Gauss quadratures of order N . Nevertheless, the discrete norm is equivalent to the continuous norm so discrete stability implies stability in the continuous norm. Also, unless the coefficient matrices $\underline{\mathbf{A}}_m$ are polynomial functions of their argument, there will always be aliasing errors when representing the flux as a polynomial of any finite degree. Nevertheless, the fully overintegrated approximation

is stable (nonincreasing energy when expected) provided that the interpolant of the coefficient matrices is constructed to be divergence free. Aliasing errors are still found, nonetheless, in the dissipation boundary terms in the approximations of the coefficient matrices, but those only affect the rate of dissipation, not stability per se. Therefore, it should be emphasized that the presence of aliasing in an approximation does not imply instability, nor does stability imply no aliasing, and that the “fully overintegrated” approximation has aliasing errors that do not contribute to instability.

Acknowledgement. The author would like to thank Gregor Gassner for helpful discussions. This work was supported by a grant from the Simons Foundation (#426393, David Kopriva).

References

- [1] Andrea Beck, Gregor Gassner, and Claus-Dieter Munz. High order and underresolution. In Rainer Ansorge, Hester Bijl, Andreas Meister, and Thomas Sonar, editors, *Recent Developments in the Numerics of Nonlinear Hyperbolic Conservation Laws*, volume 120 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 41–55. Springer Berlin Heidelberg, 2013.
- [2] C. D. Cantwell, D. Moxey, A. Comerford, A. Bolis, G. Rocco, G. Mengaldo, D. De Grazia, S. Yakovlev, J-E. Lombard, D. Ekelschot, B. Jordi, H. Xu, Y. Mohamied, C. Eskilsson, B. Nelson, P. Vos, C. Biotto, R. M. Kirby, and S. J. Sherwin. Nektar plus plus : An open-source spectral/hp element framework. *COMPUTER PHYSICS COMMUNICATIONS*, 192:205–219, Jul 2015.
- [3] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer, 2006.
- [4] Gregor J. Gassner and Andrea D. Beck. On the accuracy of high-order discretizations for underresolved turbulence simulations. *Theoretical and Computational Fluid Dynamics*, 27(3–4):221–237, 2013.
- [5] Gregor J Gassner, Andrew R Winters, and David A Kopriva. Split form nodal discontinuous Galerkin schemes with summation-by-parts property for the compressible Euler equations. *Journal Of Computational Physics*, 327:39–66, 2016.
- [6] Florian Hindenlang, Gregor J. Gassner, Christoph Altmann, Andrea Beck, Marc Staudenmaier, and Claus-Dieter Munz. Explicit discontinuous Galerkin methods for unsteady problems. *Computers and Fluids*, 61(0):86 – 93, 2012.
- [7] R.M. Kirby and G.E. Karniadakis. De-aliasing on non-uniform grids: algorithms and applications. *Journal of Computational Physics*, 191:249–264, 2003.
- [8] D. A. Kopriva. Metric identities and the discontinuous spectral element method on curvilinear meshes. *Journal of Scientific Computing*, 26(3):301–327, 2006.
- [9] D. A. Kopriva and G. Gassner. On the quadrature and weak form choices in collocation type discontinuous Galerkin spectral element methods. *Journal of Scientific Computing*, 44(2):136–155, 2010.

- [10] D. A. Kopriva and G. J. Gassner. An energy stable discontinuous Galerkin spectral element discretization for variable coefficient advection problems. *SIAM J. Sci. Comp.*, 36(4):A2076–A2099, 2014.
- [11] D.A. Kopriva, A.R. Winters, M Bohm, and G.J. Gassner. A provably stable discontinuous Galerkin spectral element approximation for moving hexahedral meshes. *Computers & Fluids*, doi:10.1016/j.compfluid.2016.05.023, 2016.
- [12] David A. Kopriva. A polynomial spectral calculus for analysis of DG spectral element methods. math 1704.00709, arXiv, 2017.
- [13] David A. Kopriva and Gregor J. Gassner. Geometry effects in nodal discontinuous Galerkin methods on curved elements that are provably stable. *Applied Mathematics and Computation*, 272, Part 2:274 – 290, 2016.
- [14] G. Mengaldo, D. De Grazia, D. Moxey, P.E. Vincent, and S.J. Sherwin. Dealiasing techniques for high-order spectral element methods on regular and irregular grids. *Journal of Computational Physics*, 299:56 – 81, 2015.