SIAM J. SCI. COMPUT.
Vol. 42, No. 5, pp. B1271–B1301

DIFFUSION SYNTHETIC ACCELERATION PRECONDITIONING FOR DISCONTINUOUS GALERKIN DISCRETIZATIONS OF S_N TRANSPORT ON HIGH-ORDER CURVED MESHES*

TERRY S. HAUT†, BEN S. SOUTHWORTH‡, PETER G. MAGINOT§, AND VLADIMIR Z. TOMOV†

Abstract. This paper derives and analyzes new diffusion synthetic acceleration (DSA) preconditioners for the S_N transport equation when discretized with a high-order (HO) discontinuous Galerkin (DG) discretization. DSA preconditioners address the need to accelerate the S_N transport equation when the mean free path ε of particles is small and the condition number of the S_N transport equation scales like $\mathcal{O}(\varepsilon^{-2})$. By expanding the S_N transport operator in ε and employing a rigorous singular matrix perturbation analysis, we derive a DSA matrix that reduces to the symmetric interior penalty (SIP) DG discretization of the standard continuum diffusion equation when the mesh is first-order and the total opacity is constant. We prove that preconditioning the HO DG S_N transport equation with the SIP DSA matrix results in an $\mathcal{O}(\varepsilon)$ perturbation of the identity, and fixed-point iteration therefore converges rapidly for optically thick problems. However, the SIP DSA matrix is conditioned like $\mathcal{O}(\varepsilon^{-1})$, making it difficult to invert for small ε. We further derive a new two-part, additive DSA preconditioner based on a continuous Galerkin discretization of diffusion-reaction, which has a condition number independent of ε, and prove that this DSA variant has the same theoretical efficiency as the SIP DSA preconditioner in the optically thick limit. The analysis is extended to the case of HO (curved) meshes, where so-called mesh cycles can result from elements both being upwind of each other (for a given discrete photon direction). In particular, we prove that performing two additional transport sweeps, with fixed scalar flux, in between DSA steps yields the same theoretical conditioning of fixed-point iterations as in the cycle-free case. Theoretical results are validated by numerical experiments on a HO, highly curved two- and three-dimensional meshes that are generated from an arbitrary Lagrangian–Eulerian hydrodynamics code, where the additional inner sweeps between DSA steps offer up to a 4× reduction in total number of sweeps required for convergence.

Key words. S_N transport, diffusion synthetic acceleration, high-order DG

AMS subject classifications. 65Fxx, 65Bxx, 65Mxx, 65Zxx

DOI. 10.1137/19M124993X

*Submitted to the journal's Computational Methods in Science and Engineering section March 14, 2019; accepted for publication (in revised form) July 2, 2020; published electronically October 27, 2020.

https://doi.org/10.1137/19M124993X

Funding: This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contracts DE-AC52-07NA27344, B614452, and B627942, Lawrence Livermore National Security, LLC (LLNL-JRNL-759881). This work was performed under the auspices of the U.S. Department of Energy under grant number (NNSA) DE-NA0002376. Disclaimer: This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

†Lawrence Livermore National Laboratory, Livermore, CA 94550 (haut3@llnl.gov, tomov2@llnl.gov).

‡Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO 80309 (ben.s.southworth@gmail.com).

§Los Alamos National Laboratory, Los Alamos, NM 87545 (pmaginot@lanl.gov).

1. Introduction.

1.1. Background. The S_N transport equation forms a key component in modeling the interaction of radiation and a background medium, and its accurate solution is critical in the simulation of astrophysics, interial confinement fusion, and a number of other fields. In this paper, we derive and analyze diffusion-based preconditioners for a high-order (HO) discontinuous Galerkin (DG) discretization of the monoenergetic S_N transport equations in the challenging (but typical) case of scattering-dominated regimes. One motivation of this research is in the context of HO arbitrary Lagrangian–Eulerian (ALE) hydrodynamics on HO (curved) meshes [4], where standard diffusion-based preconditioners are inadequate.

The standard approach for solving the S_N transport equations involves a fixed-point iteration, referred to as source iteration in the transport literature. It is well known that source iteration can converge arbitrarily slowly in the optically thick limit of large scattering and small absorption. To quantify this, it is useful to introduce the diffusion scaling. In particular, let ε be a nondimensional parameter representing the ratio of a typical mean free path of a particle to a dimension of the domain [19]. Then, for characteristic mesh spacing $h_{\mathbf{x}}$ and total cross section σ_t, the optically thick limit corresponds to $\varepsilon/(h_{\mathbf{x}}\sigma_t) \sim \varepsilon^2\sigma_a/\sigma_t \ll 1$. In this case, the matrix corresponding to source iteration has a condition number that scales like $(h_{\mathbf{x}}\sigma_t/\varepsilon)^2$ and, therefore, will converge very slowly without specialized preconditioners. Such preconditioners typically involve a two-level acceleration scheme and fall within two broad classes: (i) using a diffusion equation to solve for a corrected scalar flux, referred to as diffusion synthetic acceleration (DSA), and (ii) solving the S_N transport equations with a reduced number of angular quadrature points, referred to as transport synthetic acceleration (TSA; cf. [20, 31, 39]). This paper focuses on DSA-type algorithms. An excellent discussion on the historical development of DSA can be found in [18].

Some of the earliest work on accelerating transport equations with a diffusion-based preconditioner can be found in, for example, [14, 21, 24]. It was shown in [11, 32] that diffusion-based acceleration for source iteration is effective for fine spatial meshes ($\varepsilon \geq h_{\mathbf{x}}\sigma_t$), but its performance can degrade for coarse meshes (that is, $\varepsilon \ll h_{\mathbf{x}}\sigma_t$). Further seminal work in [3] contained a derivation and theory for a diffusion-equation accelerator whose discretization is consistent with the S_N transport diamond-difference scheme (a finite-volume type scheme for transport) and which yields fast acceleration independent of the spatial mesh size. Since, DSA methods have been significantly refined and expanded to other spatial discretizations [1, 2, 16, 17, 18, 35, 38]. In the context of HO DG discretizations, the authors in [34] develop a modified interior penalty (MIP) DSA scheme for HO DG (on first-order meshes) and numerically demonstrate that source iteration converges rapidly with the MIP DSA preconditioner.

Here we present a rigorous, discrete analysis of DSA in the context of HO DG, on potentially HO (curved) meshes. The paper proceeds as follows. Section 2 introduces the DG discretization of the S_N transport equations as well as the standard fixed-point iteration to solve the discrete S_N system, known as "source iteration." The primary theoretical contributions are formally stated in section 3.1, with the proofs provided in section 4. As a byproduct of this analysis, three new DG DSA preconditioners are developed that are effective for HO discretizations on HO meshes. Two of these preconditioners reduce to interior penalty DG discretizations of diffusion when the mesh is straight-edged (i.e., noncurved); the third preconditioner is new even for straight-edged meshes and avoids the numerical difficulties associated with inverting the interior penalty DSA matrices while provably having the same efficacy

in the thick limit. Section 3.2 relates our analysis to both the MIP preconditioner [34] and the consistent DSA preconditioner [38]. In section 5, the efficacy of our DSA preconditioners is demonstrated for HO DG discretizations on highly curved two- and three-dimensional meshes generated by [4] (an HO ALE hydrodynamics code). With the newly developed HO DSA algorithm and DSA discretization, rapid fixed-point convergence is obtained for all tested values of the mean free path (while iterations diverge on the HO mesh without the proposed algorithmic modification). Numerical results also demonstrate the new additive DSA preconditioner to be robust on optically thick and thin problems in one spatial dimension, with preconditioning in the optically thick limit equally as effective as traditional DSA using our new discretization. Brief conclusions are given in section 6.

1.2. Outline of contributions. In general, the discrete source-iteration propagation operator has singular modes with singular values on the order of $\mathcal{O}(\varepsilon^2)$, where ε is the characteristic mean free path. These modes are referred to as the near nullspace of source iteration and are extremely slow to converge when $\varepsilon \ll 1$. By directly expanding the discrete DG source-iteration operator in ε, we derive a DSA preconditioner that exactly represents the problematic error modes that are slow to decay. For first-order meshes and constant opacities, we also show that the DSA matrix exactly corresponds to the symmetric interior penalty (SIP) DG discretization of the diffusion equation. In Theorem 3, we prove that the corresponding DSA-preconditioned S_N transport equation is an $\mathcal{O}(\varepsilon)$ perturbation of the identity, and the resulting fixed-point iteration therefore converges rapidly for sufficiently small mean free path. In the optically thick limit of $\varepsilon/(h_{\mathbf{x}}\sigma_t) \sim \varepsilon^2\sigma_a/\sigma_t \ll 1$ (and assuming constant total opacity and a first order mesh), this diffusion discretization is identical to the MIP DSA preconditioner that is numerically analyzed in [34], and Theorem 3 provides a rigorous justification for its efficacy. In section 3.2 we discuss stabilization in thin regimes and formulate a nonsymmetric interior penalty (IP) DSA preconditioner as an alternative to the SIP DSA approach.

It turns out the SIP DSA matrix is in the form of a singular matrix perturbation: the dominant term is of order $1/\varepsilon$ relative to the other terms and has a nullspace consisting of continuous functions with zero boundary values. This term acts as a large penalization and constrains the solution to be continuous in the limit of $\varepsilon \to 0$. This term also leads to the SIP DSA matrix having a condition number that scales like $\mathcal{O}(1/\varepsilon)$, one of the primary reasons that DG DSA discretizations such as SIP can be difficult to precondition (although, see [5, 27, 36] for several approaches to preconditioning these systems). Appealing to the singular perturbation, we then derive a two-part additive DSA preconditioner based on projecting onto the spaces of continuous and discontinuous functions. Theorem 4 proves that the resulting preconditioned S_N transport equation fixed-point iteration is also an $\mathcal{O}(\varepsilon)$ perturbation of the identity and therefore has the same theoretical efficiency as the SIP DSA preconditioner in the optically thick limit. Moreover, the condition numbers of linear systems in the additive preconditioner are independent of ε. We note that the leading order term in this two-part additive DSA preconditioner corresponds to the continuous Galerkin (CG) discretization of the diffusion equation obtained in [12].

Next, we modify the analysis to account for HO curved meshes. In this case, neighboring mesh elements can both be upwind of each other, leading to so-called mesh cycles. With mesh cycles, the discrete streaming plus collision operator that is inverted in source iteration is no longer block lower triangular in any element ordering, and so it cannot be easily inverted through a forward solve. We prove in Theorem 5 that performing two additional transport sweeps on the S_N transport equations, with

a fixed scalar flux, yields a preconditioner that has the same asymptotic efficiency as that obtained on cycle-free meshes.

Finally, we perform a series of numerical tests to compare the behavior, in thick and thin regimes, of the three major preconditioning approaches presented in this work, namely, the SIP DSA (Theorem 3), its IP modification (section 3.2.2), and the additive DSA preconditioner (Theorem 4).

2. HO DG discretization of S_N transport and the need for preconditioning in scattering dominated regimes.

2.1. DG discretization. Consider the monoenergetic, steady-state, discrete-ordinates linear Boltzmann equation with isotropic scattering, given by

(1)
$$\mathbf{\Omega}_d \cdot \nabla_{\mathbf{x}} \psi_d(\mathbf{x}) + \frac{\sigma_t(\mathbf{x})}{\varepsilon} \psi_d(\mathbf{x}) = \frac{1}{4\pi} \left(\frac{\sigma_t(\mathbf{x})}{\varepsilon} - \varepsilon \sigma_a(\mathbf{x}) \right) \sum_{d'=1}^{N_\Omega} w_{d'} \psi_{d'}(\mathbf{x}) + \varepsilon q_d(\mathbf{x}), \quad \mathbf{x} \in \mathcal{D},$$
$$\psi_d(\mathbf{x}) = \psi_{d,\mathrm{inc}}(\mathbf{x}), \quad \mathbf{x} \in \partial\mathcal{D}, \quad \text{and} \quad \mathbf{n}(\mathbf{x}) \cdot \mathbf{\Omega}_d < 0.$$

In (1), $\mathcal{D}$ denotes the spatial domain with boundary $\partial\mathcal{D}$, $\psi_d(\mathbf{x})$ denotes the specific intensity associated with the discrete ordinate direction $\mathbf{\Omega}_d$, and $q_d(\mathbf{x})$ denotes a fixed (direction-dependent) source. Here, the total opacity, $\varepsilon\sigma_t^{-1}(\mathbf{x})$, and the absorption opacity, $\varepsilon\sigma_a(\mathbf{x})$, are scaled according to the diffusion limit, where ε is a nondimensional parameter proportional to the characteristic mean free path and which goes to zero in the optically thick limit [19]. The quadrature angle vectors $\mathbf{\Omega}_d \in \mathbb{S}^2$ and weights $w_d > 0$ are constructed to have desirable symmetry properties and integrate spherical harmonics up to a given degree that depends on the number of angles, N_Ω.

We consider a DG discretization of the S_N transport equation. To do so, we set some notation. First, consider a decomposition of the domain $\mathcal{D}$ in to a set $\mathcal{E}$ of HO (curved) elements $\kappa \in \mathcal{E}$, and let $\mathcal{F}$ denote the set of interior and boundary finite element faces $\Gamma \in \mathcal{F}$. We further decompose the set $\mathcal{F} = \mathcal{F}_{\mathrm{int}} \cup \mathcal{F}_{\mathrm{ext}}$ into the set of interior$\mathcal{F}_{\mathrm{int}}$ faces and the set $\mathcal{F}_{\mathrm{ext}}$ of boundary faces. The finite element space $\mathcal{U}$ corresponds to the collection of piecewise polynomial functions of fixed degree r on each reference element $\hat{\kappa}$, $\mathcal{P}_r(\hat{\kappa})$,

$$\mathcal{U} = \left\{ u \in L^2(\mathcal{D}) : \hat{u} \in \mathcal{P}_r(\hat{\kappa}) \right\}.$$

The values of u in physical space are obtained simply by $u(x) = \hat{u}(\hat{x}) = \sum_{i=1}^{N_\kappa} \hat{v}_i(\hat{x}) u_i$, where $\hat{x} \to x$ is the mapping from reference to physical coordinates, $\{\hat{v}_i\}_1^{N_\kappa}$ is the basis of $\mathcal{U}$ on $\hat{\kappa}$, and $\boldsymbol{u} = (u_1 \dots u_{N_\kappa})$ are the finite element coefficients of u. Basis functions in physical space are also obtained by $v(x) = \hat{v}(\hat{x})$. The order r of the solution space $\mathcal{U}$ is generally independent of the order of the mesh. All methods presented in this work are fully algebraic and do not involve geometric operations, thus they are independent of the discrete mesh representation; interested readers can find technical details about our mesh representation approach in [10]. For an interior mesh face $\Gamma \in \mathcal{F}$ shared by two neighboring elements κ and κ', we let $\mathbf{n}$ denote the normal vector that points from κ and κ'. Given this (fixed but arbitrary) choice for the sign of the normal vector $\mathbf{n}$ on each element face, the jump $[\![u]\!]$ and average $\{u\}$ for a function $u \in \mathcal{U}$ are defined by

$$[\![u]\!] = \begin{cases} u_\kappa - u_{\kappa'} & \text{if } \Gamma \text{ is an interior face shared by elements } \kappa \text{ and } \kappa', \\ u_\kappa & \text{if } \Gamma \text{ is a boundary face of element } \kappa, \end{cases}$$

and

$$\{u\} = \begin{cases} (u_\kappa + u_{\kappa'})/2 & \text{if } \Gamma \text{ is an interior face shared by elements } \kappa \text{ and } \kappa', \\ u_\kappa & \text{if } \Gamma \text{ is a boundary face of element } \kappa. \end{cases}$$

Although the definitions of the jump $[\![u]\!]$ and average $\{u\}$ depend on arbitrarily choosing a sign for the normal vector $\mathbf{n}$, it turns out that the bilinear forms below are invariant with respect to this choice.

Following the standard DG discretization procedure and using upwinding to define the numerical flux, (1) can be discretized as

(2)
$$\mathbf{\Omega}_d \cdot \mathbf{G}\boldsymbol{\psi}^{(d)} + F^{(d)}\boldsymbol{\psi}^{(d)} + \frac{1}{\varepsilon}M_t\boldsymbol{\psi}^{(d)} - \frac{1}{4\pi}\left(\frac{1}{\varepsilon}M_t - \varepsilon M_a\right)\boldsymbol{\varphi} = \frac{1}{4\pi}\left(\boldsymbol{q}_{\text{inc}}^{(d)} + \varepsilon \boldsymbol{q}^{(d)}\right).$$

Here the vector $\boldsymbol{\varphi}$ of coefficients for the scalar flux φ is given by

$$\boldsymbol{\varphi} = \sum_d w_d \boldsymbol{\psi}^{(d)}, \tag{3}$$

the vectors $\boldsymbol{q}_{\text{inc}}^{(d)}$ and $\boldsymbol{q}^{(d)}$ on the right-hand side of (2) correspond to the linear forms

$$\left[\boldsymbol{q}_{\text{inc}}^{(d)}\right]_m = -\sum_{\Gamma\in\mathcal{F}_{\text{ext}}}\int_\Gamma \mathbf{\Omega}_d\cdot\mathbf{n} v_m \psi_{\text{inc}}^{(d)}\, dS + \frac{1}{2}\sum_{\Gamma\in\mathcal{F}_{\text{ext}}}\int_\Gamma |\mathbf{\Omega}_d\cdot\mathbf{n}|\, v_m\psi_{\text{inc}}^{(d)}\, dS, \tag{4}$$

$$\left[\boldsymbol{q}^{(d)}\right]_m = \sum_{\kappa\in\mathcal{E}}\int_\kappa v_m q^{(d)} d\mathbf{x}, \tag{5}$$

where $\{v_m\}_1^N$ is the finite element basis of $\mathcal{U}$, and N is the total number of degrees of freedom in $\mathcal{U}$. We will also denote by $\mathbf{u}$ and $\mathbf{v}$ the vectors of coefficients corresponding to some discrete functions u and v in the finite element space $\mathcal{U}$. The matrices $\mathbf{\Omega}_d\cdot\mathbf{G}$, $F^{(d)}$, M_t, and M_a in (2) correspond, respectively, to the bilinear forms,

$$\mathbf{v}^T\left(\mathbf{\Omega}_d\cdot\mathbf{G}\right)\mathbf{u} = \sum_{\kappa\in\mathcal{E}}\int_\kappa \left(\mathbf{\Omega}_d\cdot\nabla_{\mathbf{x}}u\right) v d\mathbf{x}, \tag{6}$$

$$\mathbf{v}^T F^{(d)}\mathbf{u} = -\sum_{\Gamma\in\mathcal{F}}\int_\Gamma \mathbf{\Omega}_d\cdot\mathbf{n}\,[\![u]\!]\,\{v\}\,dS + \frac{1}{2}\sum_{\Gamma\in\mathcal{F}}\int_\Gamma |\mathbf{\Omega}_d\cdot\mathbf{n}|\,[\![u]\!]\,[\![v]\!]\,dS, \tag{7}$$

$$\mathbf{v}^T M_t\mathbf{u} = \sum_{\kappa\in\mathcal{E}}\int_\kappa \sigma_t uv d\mathbf{x}, \tag{8}$$

$$\mathbf{v}^T M_a\mathbf{u} = \sum_{\kappa\in\mathcal{E}}\int_\kappa \sigma_a uv d\mathbf{x}. \tag{9}$$

Note that in our convention bold symbols indicate vectors and capital (from the Latin alphabet) symbols indicate matrices. In addition, the notation $\mathbf{G}$ is shorthand for a vector with three matrix components, $\mathbf{G} = (G_1, G_2, G_3)$, so that

$$\mathbf{\Omega}_d\cdot\mathbf{G} = \sum_{j=1}^3 \left(\mathbf{\Omega}_d\right)_j G_j.$$

Also recall that each direction $\mathbf{\Omega}_d$ has a corresponding reversed direction $\mathbf{\Omega}_{d'} = -\mathbf{\Omega}_d$ with identical weight $w_{d'} = w_d$, and note the useful identities, $\sum_d w_d = 4\pi$, $\sum_d w_d\mathbf{\Omega}_d\mathbf{\Omega}_d^T = \frac{4\pi}{3}I$, $\sum_d w_d\mathbf{\Omega}_d = \mathbf{0}$, and $\sum_d w_d\mathbf{\Omega}_d|\mathbf{\Omega}_d\cdot\mathbf{n}| = \mathbf{0}$.

To reformulate (2), define the column vector $\boldsymbol{\psi} = (\boldsymbol{\psi}^{(1)}; \ldots; \boldsymbol{\psi}^{(N_\Omega)})$ and projection

$$(P_0\boldsymbol{\psi})^{(d)} = \frac{1}{4\pi}\sum_{d'} w_{d'}\boldsymbol{\psi}_{d'} = \frac{1}{4\pi}\boldsymbol{\varphi}, \quad d = 1, \ldots, N_\Omega. \tag{10}$$

P_0 is a weighted average over direction d that projects the average on to all vector blocks. In the matrix sense, P_0 is an $NN_\Omega \times NN_\Omega$ operator, where each block row takes the form $\frac{1}{4\pi}[w_0 I_N, w_1 I_N, \ldots, w_{N_\Omega} I_N]$. P_0 being a projection relies on the fact that $\sum_d w_d = 4\pi$. Defining

$$W = \operatorname{diag}\left[w_0 I_N, w_1 I_N, \ldots, w_{N_\Omega} I_N\right], \quad \langle \mathbf{x}, \mathbf{y}\rangle_W = \langle W\mathbf{x}, \mathbf{y}\rangle, \tag{11}$$

P_0 is an orthogonal projection in the W-inner product. Letting $Q_0 := I - P_0$ denote the orthogonal complement to P_0, recall that for any vector $\boldsymbol{\psi}$, $\|\boldsymbol{\psi}\|_W = \|P_0\boldsymbol{\psi}\|_W + \|Q_0\boldsymbol{\psi}\|_W$. Now, rewrite (2) as

$$\left[I + \varepsilon M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\right]\boldsymbol{\psi}^{(d)} - \frac{1}{4\pi}\left(I - \varepsilon^2 M_t^{-1} M_a\right)\boldsymbol{\varphi} = \frac{1}{4\pi}\varepsilon M_t^{-1}\left(\boldsymbol{q}_{\text{inc}}^{(d)} + \varepsilon \boldsymbol{q}^{(d)}\right). \tag{12}$$

In matrix form, over all angles, the first term in (12) operating on $\boldsymbol{\psi}^{(d)}$ is block diagonal in d, with each block corresponding to a fixed direction $\boldsymbol{\Omega}_d$, and the second term a global angular coupling through projection P_0. A standard technique in transport is to invert the first, block-diagonal term. This approach corresponds to solving the linear transport equation independently, for all directions d, and is known as a transport sweep. Define T_ε as the block-diagonal operator over direction d, multiplied by P_0, when a transport sweep is applied:

$$T_\varepsilon = \frac{1}{4\pi}\operatorname{diag}_d\left[\left(I + \varepsilon M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\right)^{-1}\left(I - \varepsilon^2 M_t^{-1} M_a\right)\right] P_0.$$

Then, (2) can be rewritten as the preconditioned linear system

$$(I - T_\varepsilon)\boldsymbol{\psi} = \tilde{\boldsymbol{q}}, \tag{13}$$

where

$$\tilde{\boldsymbol{q}}^{(d)} = \left(I + \varepsilon M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\right)^{-1}\frac{1}{4\pi}\varepsilon M_t^{-1}\left(\boldsymbol{q}_{\text{inc}}^{(d)} + \varepsilon \boldsymbol{q}^{(d)}\right).$$

Multiplying (13) by the quadrature weight, w_d, and summing over direction index, d, yields a linear system for the scalar flux,

$$\left(I - S_\varepsilon\right)\boldsymbol{\varphi} = \mathbf{s}, \tag{14}$$

where

$$S_\varepsilon = \sum_d w_d\left(I + \varepsilon M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\right)^{-1}\frac{1}{4\pi}\left(I - \varepsilon^2 M_t^{-1} M_a\right), \tag{15}$$

$$\mathbf{s} = \sum_d w_d\left(I + \varepsilon M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \; + F^{(d)}\right)\right)^{-1}\varepsilon M_t^{-1}\varepsilon\frac{1}{4\pi}\left(\boldsymbol{q}_{\text{inc}}^{(d)} + \boldsymbol{q}^{(d)}\right). \tag{16}$$

We note that, in applying the operator T_ε, we need to invert $[I + \varepsilon M_t^{-1}(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)})]$. As it turns out, this is not always computationally tractable, particularly in the case of HO curved meshes. Theorem 5 and section 4.3 analyze a more general case where this term is not inverted exactly.

Remark 1. Our analysis of (13) is valid under the assumption that

$$\varepsilon \left\| M_t^{-1} \left(\mathbf{\Omega}_d \cdot \mathbf{G} + F^{(d)} \right) \right\| < 1, \quad \varepsilon^2 \| M_t^{-1} M_a \| < 1.$$

Since $\| M_t^{-1} (\mathbf{\Omega}_d \cdot \mathbf{G} + F^{(d)}) \|$ scales like $1/(\sigma_t h_{\mathbf{x}})$, where $h_{\mathbf{x}}$ denotes the characteristic mesh spacing, the error bounds in Theorems 3–5 below are small as long as

$$\eta = \min \left\{ \varepsilon / (h_{\mathbf{x}} \sigma_t), \varepsilon \sqrt{\sigma_a / \sigma_t} \right\} \ll 1.$$

The regime $\eta \ll 1$ corresponds to the standard optically thick limit.

2.2. Useful identities. Next we present two identities that will be used regularly in further derivations. First, applying integration by parts to the term

$$\mathbf{v}^T \mathbf{\Omega}_d \cdot \mathbf{G} \mathbf{u} = \sum_{\kappa \in \mathcal{E}} \int_\kappa (\mathbf{\Omega}_d \cdot \nabla_{\mathbf{x}} u)\, v d\mathbf{x}$$

in $\mathbf{v}^T (\mathbf{\Omega}_d \cdot \mathbf{G} + F^{(d)}) \mathbf{u}$ yields the identity

$$\mathbf{\Omega}_d \cdot \mathbf{G} + F^{(d)} = -\mathbf{\Omega}_d \cdot \mathbf{G}^T + \tilde{F}^{(d)}, \tag{17}$$

where the matrix $\tilde{F}^{(d)}$ corresponding to the bilinear form

$$\mathbf{v}^T \tilde{F}^{(d)} \mathbf{u} = -\sum_{\Gamma \in \mathcal{F}} \int_\Gamma \mathbf{\Omega}_d \cdot \mathbf{n} \{u\} [\![v]\!]\, dS + \sum_{\Gamma \in \mathcal{F}} \int_\Gamma \frac{1}{2} |\mathbf{\Omega}_d \cdot \mathbf{n}| [\![u]\!] [\![v]\!]\, dS. \tag{18}$$

A second property follows immediately from (7) and (18). Let P denote a projection onto the space of continuous functions with zero boundary values. Then, $P\mathbf{v}$ corresponds to a continuous function with zero boundary value and, therefore, $[\![v]\!] = 0$ on each interior mesh face Γ and $v = 0$ on each boundary face. From (7), we see that $(P\mathbf{v})^T \tilde{F}^{(d)} \mathbf{u} = \mathbf{v}^T P^T \tilde{F}^{(d)} \mathbf{u} = \mathbf{0}$, for any $\mathbf{u}$ and $\mathbf{v}$. Since $\mathbf{u}$ and $\mathbf{v}$ are arbitrary, $P^T \tilde{F}^{(d)} = \mathbf{0}$. A similar identity follows from (18), yielding the two identites

$$F^{(d)} P = \mathbf{0}, \qquad P^T \tilde{F}^{(d)} = \mathbf{0}. \tag{19}$$

2.3. Need for preconditioning in the optically thick limit. To motivate DSA and further analysis in this paper, we state the following proposition which shows that preconditioning the linear system in (13) is important in the optically thick limit of small ε. The proof of Proposition 2 is given in the appendix.

PROPOSITION 2. *Assume that the matrix $I - T_\varepsilon$ in the linear system* (13) *is invertible. Then the condition number of the matrix $I - T_\varepsilon$ from* (13) *satisfies*

$$\mathrm{cond}(I - T_\varepsilon) = \|I - T_\varepsilon\|_W \|(I - T_\varepsilon)^{-1}\|_W \geq \mathcal{O}\left(\varepsilon^{-2}\right),$$

where the norm $\|\cdot\|_W$ is defined by (11). *In addition, suppose that E_ε inverts $P_0(I - T_\varepsilon)P_0$ on the range of P_0 to within $\mathcal{O}(\varepsilon)$, $E_\varepsilon P_0 (I - T_\varepsilon) P_0 = P_0 + \mathcal{O}(\varepsilon)$. Then the preconditioned matrix $((I - P_0) + E_\varepsilon P_0)(I - T_\varepsilon)$ is an $\mathcal{O}(\varepsilon)$ perturbation of the identity*

$$((I - P_0) + E_\varepsilon P_0)(I - T_\varepsilon) = I + \mathcal{O}(\varepsilon). \tag{20}$$

The relationship

$$(P_0 (I - T_\varepsilon) P_0 \boldsymbol{\psi})^{(d)} = (I - S_\varepsilon) \frac{\varphi}{4\pi}, \quad d = 1, \ldots, N_\Omega, \tag{21}$$

connects the Theorems in section 3.1 with Proposition 2.

3. DSA preconditioners for HO DG discretizations on curved meshes.

3.1. Overview of the DSA preconditioners and statement of the theorems. This section presents the main theoretical contributions of this paper, the proofs of which are contained in the following subsections.

First we present results on an SIP DSA preconditioner. To do so, define the SIP DSA matrix,

$$D_\varepsilon = \frac{1}{\varepsilon} F_0 + D_0, \tag{22}$$

where

$$D_0 = \frac{1}{3}\mathbf{G}^T \cdot M_t^{-1}\mathbf{G} - \tilde{\mathbf{F}}_1 \cdot M_t^{-1}\mathbf{G} + \mathbf{G}^T \cdot M_t^{-1}\mathbf{F}_1 + M_a, \tag{23}$$

and

$$F_0 = \frac{1}{4\pi}\sum_d w_d F^{(d)}, \quad \mathbf{F}_1 = \frac{1}{4\pi}\sum_d w_d \mathbf{\Omega}_d F^{(d)}, \quad \tilde{\mathbf{F}}_1 = \frac{1}{4\pi}\sum_d w_d \mathbf{\Omega}_d \tilde{F}^{(d)}. \tag{24}$$

In the previous equations, $\mathbf{F}_1$ and $\tilde{\mathbf{F}}_1$ correspond to vectors of matrices; for example, in three spatial dimensions

$$\left(\mathbf{F}_1\right)_j = \frac{1}{4\pi}\sum_d w_d \left(\mathbf{\Omega}_d\right)_j F^{(d)}, \quad j = 1, 2, 3.$$

Assuming that the mesh is first order and that the opacities, σ_t and σ_a, are constants, it turns out (see section 4.2) that D_ε corresponds to the bilinear form,

$$\mathbf{v}^T D_\varepsilon \mathbf{u} = \mathcal{B}_{\text{SIP}}\left(\cdot,\cdot\right), \tag{25}$$

where

$$\begin{aligned}\mathcal{B}_{\text{SIP}}\left(u,v\right) := {} & \frac{1}{\varepsilon}\sum_{\Gamma\in\mathcal{F}}\int_\Gamma \alpha \left[\!\left[u\right]\!\right]\left[\!\left[v\right]\!\right] dS + \sum_{\kappa\in\mathcal{E}}\int_\kappa \frac{1}{3\sigma_t}\nabla_{\mathbf{x}} u\cdot\nabla_{\mathbf{x}} v d\mathbf{x} + \sum_{\kappa\in\mathcal{E}}\int_\kappa \sigma_a uv d\mathbf{x} \\ & - \sum_{\Gamma\in\mathcal{F}}\int_\Gamma \left[\!\left[u\right]\!\right]\left\{\mathbf{n}\cdot\frac{1}{3\sigma_t}\nabla_{\mathbf{x}} v\right\} dS - \sum_{\Gamma\in\mathcal{F}}\int_\Gamma \left[\!\left[v\right]\!\right]\left\{\mathbf{n}\cdot\frac{1}{3\sigma_t}\nabla_{\mathbf{x}} u\right\} dS.\end{aligned} \tag{26}$$

Here, the function $\alpha(\cdot)$ in the first integral is defined as

$$\alpha\left(\mathbf{x}\right) = \frac{1}{4\pi}\sum_d w_d \left|\mathbf{\Omega}_d\cdot\mathbf{n}\left(\mathbf{x}\right)\right|, \quad \mathbf{x}\in\Gamma\in\mathcal{F}, \tag{27}$$

and converges to 1/4 in the limit of a large number of angles, $\mathbf{\Omega}_d$. The bilinear form in (26) corresponds to a variant of the SIP discretization of the reaction-diffusion operator,

$$\nabla_{\mathbf{x}}\cdot\left(\frac{1}{3\sigma_t}\nabla_{\mathbf{x}}\right) - \sigma_a.$$

Theorem 3 shows that preconditioning the fixed-point iteration based on $(I - S_\varepsilon)$ (14) with the DSA matrix D_ε results in fast convergence in the optically thick limit.

THEOREM 3 (SIP DSA preconditioner). *Assume that the function $\alpha(\cdot)$ defined in (27) is uniformly bounded away from zero on each interior and boundary mesh faces. Then*

$$\left(\varepsilon^2 D_\varepsilon\right)^{-1} M_t \left(I - S_\varepsilon\right) = I + \mathcal{O}\left(\varepsilon\right).$$

Theorem 3 states that the preconditioned iteration matrix looks like the identity plus an $\mathcal{O}(\varepsilon)$ perturbation. For small ε, this ensures a well-conditioned iteration matrix and fast convergence. Under the assumptions of Theorem 3, it follows from the identity (see section 4.2)

$$\mathbf{v}^T F_0 \mathbf{u} = \sum_{\Gamma \in \mathcal{F}} \int_\Gamma \alpha \left[\!\left[u\right]\!\right] \left[\!\left[v\right]\!\right] dS$$

that F_0 has a nullspace consisting of continuous functions with zero boundary values. For example, if $\mathbf{u}$ is in the nullspace of F_0, then

$$\mathbf{u}^T F_0 \mathbf{u} = \frac{1}{\varepsilon} \sum_{\Gamma \in \mathcal{F}} \int_\Gamma \alpha \left[\!\left[u\right]\!\right]^2 dS = 0,$$

and so the jump $[\![u]\!]$ must vanish on each interior mesh face and u must vanish on each boundary face. It follows that the condition number of D_ε scales like $\mathcal{O}(\varepsilon^{-1})$, and a good preconditioner is required to efficiently invert the IP DSA matrix. Unfortunately, HO DG discretizations can prove difficult for fast linear solvers and preconditioners, such as algebraic multigrid (AMG), even when considering elliptic problems [8, 26, 30, 33]. This difficulty is compounded on highly unstructured grids, which are some of the motivating problems here.

Fortunately, the proof of Theorem 3 also yields a better-conditioned DSA preconditioner for optically thick problems. In fact, let P denote an (arbitrary) projection of functions in the DG space onto the continuous functions, and $Q = I - P$ be its complement. Then Theorem 4 develops a two-part, additive DSA matrix; a single DSA step involves three applications of $P(P^T D_0 P)^{-1} P^T$ (that is, solving a CG diffusion discretization), and one application of $Q(Q^T F_0 Q)^{-1} Q^T$ (solving in the complement). In the optically thick limit, this DSA matrix is proven to have the same theoretical iteration efficiency as the SIP DSA matrix discussed in Theorem 3, and its application requires inverting matrices with condition number independent of ε.

THEOREM 4. *Let P denote a projection on to the subspace of $\mathcal{U}$ containing continuous polynomials with zero boundary values, and let $Q = I - P$. Define the operators*

$$E_P = P\left(P^T D_0 P\right)^{-1} P^T, \qquad E_Q = Q\left(Q^T F_0 Q\right)^{-1} Q^T,$$

and

$$E_\varepsilon = \frac{1}{\varepsilon} E_P + \left(I - E_P D_0\right) E_Q \left(I - D_0 E_P\right),$$

with D_0 as in (23). Then

$$\frac{1}{\varepsilon} E_\varepsilon M_t \left(I - S_\varepsilon\right) = I + \mathcal{O}\left(\varepsilon\right).$$

As in Theorem 3, Theorem 4 proves that the preconditioned operator is an $\mathcal{O}(\varepsilon)$ perturbation of the identity and is thus well-conditioned for small ε, and the corresponding fixed-point iteration will converge rapidly. Note that, using (26) for the

bilinear form corresponding to $P(P^TDP)^{-1}P^T$, it is straightforward to see that the matrix $P(P^TD_0P)^{-1}P^T$ corresponds to solving a CG discretization of the diffusion equation (3.1) (for constant opacities σ_a and σ_t). In practice, the projection matrix P is formed as a sparse matrix that (i) interpolates the DG solution at the internal (per element) Gauss–Lobatto nodes of the CG space, (ii) averages overlapping CG degrees of freedom on element faces, and (iii) zeroes out CG degrees-of-freedom on each mesh boundary face. Note that a number of works have considered preconditioning elliptic DG discretizations with a projection onto continuous functions. To our knowledge, this was first considered in the widely unrecognized paper by Warsa et al. [37] and has been considered in a number of other papers more recently [6, 9, 28, 29]. Such approaches are similar in principle to Theorem 4, but here we directly precondition the larger transport iteration by projecting onto continuous functions rather than trying to solve the DG DSA matrix with a continuous preconditioner. However, examples of projections P and Q can be found in [6, 9, 28, 29, 37].

The final result of this paper regards applying DSA to HO (curved) meshes. In particular, consider the general linear system in (12), expressed as a single operator on $\boldsymbol{\psi}$:

$$\left[(I+\varepsilon H) - \left(I - \varepsilon^2 M_t^{-1} M_a\right) P_0\right] \boldsymbol{\psi} = \mathbf{s}. \tag{28}$$

Often it is possible to order the mesh elements so that $H = \mathrm{diag}_d[M_t^{-1}(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)})]$ is block lower triangular, with blocks corresponding to mesh elements. In such cases, $I + \varepsilon H$ can be inverted directly to give the equivalent (but better conditioned) system

$$(I - T_\varepsilon)\,\boldsymbol{\psi} = (I+\varepsilon H)^{-1}\,\mathbf{s}, \tag{29}$$

where

$$T_\varepsilon = (I+\varepsilon H)^{-1}\left(I - \varepsilon^2 M_t^{-1} M_a\right) P_0.$$

However, for HO meshes, it is typically the case that H is no longer block lower triangular and cannot be easily inverted through a forward solve. Recent work developed a nonsymmetric AMG algorithm that has proved effective to invert HO DG transport discretizations on HO meshes [22, 23], albeit with a larger overhead cost compared with a forward solve. Alternatively, a graph-based algorithm was developed in [13] to replace the inversion with a Gauss–Seidel type iteration in a pseudo-optimal ordering when mesh cycles are present. To consider an approximate inversion through an ordered Gauss–Seidel, suppose that we choose a mesh element ordering that leads to a decomposition,

$$H = H_{\leq} + H_{>},$$

where

$$H_{\leq} = \mathrm{diag}_d\left[M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F_{\leq}^{(d)}\right)\right], \quad H_{>} = \mathrm{diag}_d\left[M_t^{-1} F_{>}^{(d)}\right].$$

Here, we invert $H_{\leq}$ exactly and move $H_{>}$ to the right-hand side. For example, $H_{\leq}$ corresponds to the lower-triangular part of the matrix ordering in [13], which is inverted in an ordered Gauss–Seidel iteration.

The following theorem shows that three transport sweeps with lagging—that is, three applications of $(I+\varepsilon H_{\leq})^{-1}$— yields an efficient preconditioner using the DSA matrix from Theorem 3 or 4.

THEOREM 5. *Let $I - T_\varepsilon$ be the preconditioned linear system in* (29) *that corresponds to applying $(I+\varepsilon H)^{-1}$ as a preconditioner. Define $I-\widetilde{T}_\varepsilon$ as the preconditioned linear system associated with applying three iterations of $(I + \varepsilon H_{\leq})^{-1}$ to* (28)*, while keeping the term $(I - \varepsilon^2 M_t^{-1} M_a) P_0 \boldsymbol{\psi}$ fixed. Then*

$$\widetilde{T}_\varepsilon = T_\varepsilon + \mathcal{O}\left(\varepsilon^3\right),$$

and, letting E_ε correspond to the DSA preconditioner in Theorem 3 *or* 4*,*

$$E_\varepsilon P_0 \left(I - T_\varepsilon\right) P_0 = E_\varepsilon P_0 \big(I - \widetilde{T}_\varepsilon\big) P_0 + \mathcal{O}\left(\varepsilon\right).$$

Remark 6. Note that moving the term $(I - \varepsilon^2 M_t^{-1} M_a) P_0 \boldsymbol{\psi}$ in the linear system to the right-hand side and fixing it—that is, not updating $(I - \varepsilon^2 M_t^{-1} M_a) P_0 \boldsymbol{\psi}$ based on an updated $\boldsymbol{\psi}$—is not typical in a fixed-point iterative method. One can also work out the error-propagation matrix for multiple iterations that include updating this term each iteration. For this variant, the asymptotics in ε do not clearly indicate a well-conditioned system for $\varepsilon \ll 0$, as obtained in Theorem 5. However, numerically, updating $(I - \varepsilon^2 M_t^{-1} M_a) P_0 \boldsymbol{\psi}$ each iteration proves to be more robust for larger ε, which is discussed in section 5.

Remark 7. When $\varepsilon \gtrapprox h_x \sigma_t \ll 1$, the preconditioned matrix $(\varepsilon^2 D_\varepsilon)^{-1} M_t (I - S_\varepsilon)$ from Theorem 3 becomes ill-conditioned (since the spectrum of D_ε^{-1} goes to zero for high-frequency eigenvectors of D_ε). In practice, we therefore use the preconditioner

$$I + \left(\varepsilon^2 D_\varepsilon\right)^{-1} M_t. \tag{30}$$

In the thick limit, Theorem 3 yields $(\varepsilon^2 D_\varepsilon)^{-1} M_t (I - S_\varepsilon) = I + \mathcal{O}(\varepsilon)$. In addition, $\|I - S_\varepsilon\| \sim \mathcal{O}(1)$ and is well-conditioned in the thin limit, and thus the resulting preconditioned matrix using (30) has the same asymptotic efficiency for optically thick problems and is well-conditioned for optically thin problems.

Remark 8. As noted previously, the symmetric interior penalty DSA matrix in Theorem 3 has a penalty parameter that scales like ε^{-1}, and this can present challenges for standard AMG preconditioners. However, the recent preconditioner developed in [5] utilizes a decomposition of the DG space into a continuous component and a correction. The resulting preconditioner involves solving a CG diffusion matrix, and a correction step involving a Jacobi iteration. In this way, the method has a close connection to the preconditioner in [27]. The preconditioner in [5] results in a number of iterations on the preconditioned IP DSA matrix that is provably independent of the local DG polynomial order, the mesh spacing, and the penalty parameter. Therefore, the interior penalty DSA matrix from Theorem 3, in conjunction with the preconditioner in [5], can serve as an effective alternative to the DSA matrix in Theorem 4.

Proofs of Theorems 3–5 are given in section 4.

3.2. Connection to previous work.

3.2.1. The MIP DSA preconditioner. We first connect our derivation and analysis of the SIP DSA preconditioner to the MIP DSA preconditioner in [34] and then relate the SIP DSA preconditioner to the consistent DSA preconditioner derived in [38] for linear DG discretizations.

In [34] the authors numerically demonstrate that using the MIP DSA matrix yields uniformly good convergence in both optically thick and thin regimes. The corresponding bilinear form is similar to (26), but the penalty coefficient γ in the penalty term,

$$\sum_{\Gamma\in\mathcal{F}}\int_{\Gamma}\gamma[\![u]\!][\![v]\!]dS,$$

is modified outside of the optically thick limit. In particular, letting $h_{\mathbf{x}}$ denote the characteristic mesh spacing, the MIP coefficient γ in [34] scales like $\max(1/(4\varepsilon), C_p/(\sigma_t h_{\mathbf{x}}))$, where C_p is a constant that depends on the finite element local polynomial order. Notice that, when $\varepsilon \lessapprox \sigma_t h_{\mathbf{x}}$, the MIP coefficient reduces to $1/(4\varepsilon) \approx \alpha/\varepsilon$ (this inequality becomes an equality in the limit of an infinite number of quadrature angles), which is identical to the SIP DSA coefficient in (26). Therefore, Theorem 3 justifies the numerically observed behavior in [34] when $\varepsilon \lessapprox \sigma_t h_{\mathbf{x}}$.

When $\varepsilon \gtrapprox \sigma_t h_{\mathbf{x}}$, the analysis in Theorem 3 breaks down. Nevertheless, at this point the mesh spacing $h_{\mathbf{x}}$ is small enough to numerically resolve the continuum transport equation (1). It is then expected that the analysis of DSA for the continuum S_N transport equation using the continuum diffusion equation can describe this situation (for example, see [18]). In particular, as long as the discrete diffusion equation remains a valid discretization of the continuum diffusion equation when $\varepsilon \gtrapprox \sigma_t h_{\mathbf{x}}$, we expect rapid acceleration for both optically thick and thin regimes. However, it is well-known that the penalty parameter must be at least as large as $\mathcal{O}(\frac{1}{\sigma_t h_{\mathbf{x}}})$ in order for the SIP discretization to remain a stable discretization of the continuum diffusion equation (for example, see [7]). This motivates choosing κ to scale like $\max\{1/(4\varepsilon), C_p/(\sigma_t h_{\mathbf{x}})\}$ to ensure that the MIP DSA matrix both approximates the near-nullspace in the optically thick (ill-conditioned) limit $\varepsilon \lessapprox \sigma_t h_{\mathbf{x}}$ and also remains a good approximation to the continuum diffusion equation as $\varepsilon \gtrapprox \sigma_t h_{\mathbf{x}}$ and $h_{\mathbf{x}}$ begins to resolve the mean free path.

3.2.2. The nonsymmetric IP DSA preconditioner. In section 5, the SIP DSA preconditioner is shown to be robust for $\varepsilon \ll 1$ but does not converge for moderate ε (relative to the characteristic mesh spacing). Consider the nonsymmetric IP version of the DSA matrix

$$\frac{1}{\varepsilon}F_0 + \frac{1}{3}\boldsymbol{G}^T M_t^{-1}\boldsymbol{G} - \tilde{\mathbf{F}}_1 \cdot M_t^{-1}\mathbf{G} + M_a, \tag{31}$$

where we have neglected the term $\mathbf{G}^T \cdot M_t^{-1}\mathbf{F}_1$ from the SIP DSA preconditioner defining D_ε (see (23)). Dropping this term results in a nonsymmetric IP discretization of the diffusion equation when the opacities are constant, and we observe empirically that uniformly good convergence is obtained for all tested values of ε using this DSA matrix (31). In fact, for linear DG discretizations and straight-edged meshes, the nonsymmetric IP DSA matrix (31) reduces to the Warsa–Wareing–Morel consistent diffusion discretization [38].

Also, a straightforward (but tedious) calculation shows that one can obtain the SIP DSA preconditioner by taking the first two (discrete) angular moments of the discrete equation (2) and employing the following discrete version of Fick's law

$$\boldsymbol{\psi}^{(d)} = \frac{1}{4\pi}\boldsymbol{\varphi} - \varepsilon\frac{1}{4\pi}M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\boldsymbol{\varphi} + \mathcal{O}\left(\varepsilon^2\right). \tag{32}$$

Equation (32) results from (2),

$$\begin{aligned}\boldsymbol{\psi}^{(d)} &= \left(I + \varepsilon M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\right)^{-1}\left(\frac{1}{4\pi}\boldsymbol{\varphi} + \varepsilon\frac{1}{4\pi}\boldsymbol{q}_{\text{inc}}^{(d)}\right)\\ &= \frac{1}{4\pi}\boldsymbol{\varphi} - \varepsilon M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\frac{1}{4\pi}\boldsymbol{\varphi} + \varepsilon\frac{1}{4\pi}\boldsymbol{q}_{\text{inc}}^{(d)} + \mathcal{O}\left(\varepsilon^2\right),\end{aligned}$$

where the constant vector $\varepsilon(4\pi)^{-1}\boldsymbol{q}_{\text{inc}}^{(d)}$ is neglected for simplicity since it only contributes to the right-hand side. Similarly, by instead employing the following modified version of Fick's law in the discrete moment equations,

$$\boldsymbol{\psi}^{(d)} \approx \frac{1}{4\pi}\boldsymbol{\varphi} + \varepsilon\frac{1}{4\pi}M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{J}\right), \tag{33}$$

an analogous calculation shows that $\mathbf{J} = -\mathbf{G}\boldsymbol{\varphi}$ and leads to the nonsymmetric IP DSA matrix (31). In particular, the modified version of Fick's law (33) results from neglecting the term $\varepsilon(4\pi)^{-1}M_t^{-1}F^{(d)}\boldsymbol{\varphi}$ in (32).

Remark 9. The consistent P1 formulation is usually written in terms of both the current $\mathbf{J}$ and the scalar flux $\boldsymbol{\varphi}$. However, upon using the first moment equation to express the discrete current in terms of the scalar flux and plugging the resulting expression in to the 0th moment equation, one obtains an equation for the scalar flux only, and this equation exactly corresponds to the nonsymmetric IP DSA preconditioner.

4. Proofs of main results.

4.1. Proofs of the theorems.

We first establish the following lemma.

LEMMA 10. *Consider the matrix*

$$\widehat{D} = F_0 + \varepsilon D,$$

where F_0 is a symmetric, singular matrix. Define P as a projection on to the nullspace of F_0, let $Q = I - P$ denote its complement, and define

$$E_P = P\left(P^T D P\right)^{-1} P^T, \quad E_Q = Q\left(Q^T F_0 Q\right)^{-1} Q^T.$$

Then,

$$\widehat{D}^{-1} = \frac{1}{\varepsilon}E_P + \left(I - E_P D\right)E_Q\left(I - DE_P\right) + \varepsilon\left(I - E_P D\right)R_\varepsilon\left(I - DE_P\right), \tag{34}$$

where

$$R_\varepsilon = \varepsilon\Big(I + \varepsilon E_Q\left(D - DE_P D\right)Q\Big)^{-1} E_Q\left(I - DE_P\right)DE_Q.$$

In addition, suppose that $D = D_0 + D_1$, where $P^T D_1 = D_1 P = \mathbf{0}$. Then,

$$\left(F_0 + \varepsilon D\right)^{-1} = \left(F_0 + \varepsilon D_0\right)^{-1} + \mathcal{O}\left(\varepsilon\right). \tag{35}$$

Proof. Consider the equation $\widehat{D}\mathbf{x} = \mathbf{y}$, and let P be a projection onto the null space of F_0 and $Q = I - P$ its complement. Similar to the proof of Proposition 2, $\widehat{D}\mathbf{x} = \mathbf{y}$ can be expanded based on P and Q as a 2×2 system. First, note that $\widehat{D}\mathbf{x} = \mathbf{y}$ can be written as

$$\widehat{D}\begin{pmatrix}P & Q\end{pmatrix}\begin{pmatrix}P\\ Q\end{pmatrix}\mathbf{x} = (P+Q)\mathbf{y}.$$

Now, we can multiply on the left by the full column-rank operator $\left(\begin{smallmatrix} P^T \\ Q^T \end{smallmatrix}\right)$ to yield

$$\begin{pmatrix} P^T \\ Q^T \end{pmatrix} \widehat{D} \begin{pmatrix} P & Q \end{pmatrix} \begin{pmatrix} P \\ Q \end{pmatrix} \mathbf{x} = \begin{pmatrix} P^T \\ Q^T \end{pmatrix} (P+Q)\mathbf{y},$$
$$\begin{pmatrix} P^T\widehat{D}P & P^T\widehat{D}Q \\ Q^T\widehat{D}P & Q^T\widehat{D}Q \end{pmatrix} \begin{pmatrix} P\mathbf{x} \\ Q\mathbf{x} \end{pmatrix} = \begin{pmatrix} P^T\mathbf{y} \\ Q^T\mathbf{y} \end{pmatrix}.$$

Denote $\mathbf{x}_P = P\mathbf{x}$ and $\mathbf{x}_Q := Q\mathbf{x}$. Using the equations $P^T F_0 = F_0 P = \mathbf{0}$, we can rewrite the linear system as

$$\begin{pmatrix} \varepsilon P^T D P & \varepsilon P^T D Q \\ \varepsilon Q^T D P & Q^T \widehat{D} Q \end{pmatrix} \begin{pmatrix} \mathbf{x}_P \\ \mathbf{x}_Q \end{pmatrix} = \begin{pmatrix} P^T\mathbf{y} \\ Q^T\mathbf{y} \end{pmatrix}. \tag{36}$$

Then,

$$\begin{aligned} \mathbf{x}_P &= \frac{1}{\varepsilon} \left(P^T D P\right)^{-1} P^T \mathbf{y} - \left(P^T D P\right)^{-1} \left(P^T D Q\right) \mathbf{x}_Q \\ &= \frac{1}{\varepsilon} E_P \mathbf{y} - E_P D Q \mathbf{x}_Q, \end{aligned}$$

where the second equality follows from noting that $\mathbf{x}_P = P\mathbf{x}$ and multiplying both sides by P. Equation (36) also yields

$$\begin{aligned} Q^T \widehat{D} Q \mathbf{x}_Q + \varepsilon Q^T D P \mathbf{x}_P &= Q^T \mathbf{y}, \\ \left(Q^T \widehat{D} Q - \varepsilon \left(Q^T D P\right)\left(E_P D Q\right)\right) \mathbf{x}_Q &= Q^T \mathbf{y} - \left(Q^T D E_P\right) \mathbf{y}, \\ \left(Q^T F_0 Q + \varepsilon Q^T \left(D - D E_P D\right) Q\right) \mathbf{x}_Q &= Q^T \mathbf{y} - \left(Q^T D E_P\right) \mathbf{y}. \end{aligned}$$

Now, since the matrix $Q^T F_0 Q$ above is invertible on the range of Q^{T}, we can apply $\left(Q^T F_0 Q\right)^{-1} Q^T$ to both sides to get

$$\left(I + \varepsilon Q \left(Q^T F_0 Q\right)^{-1} Q^T \left(D - D E_P D\right) Q\right) \mathbf{x}_Q = \left(Q^T F_0 Q\right)^{-1} Q^T \left(I - D E_P\right) \mathbf{y}. \tag{37}$$

Substituting $E_Q = Q \left(Q^T F_0 Q\right)^{-1} Q^T$ in the left-hand side and applying the matrix identity $(I + A)^{-1} = I - (I + A)^{-1} A$ to $(I + \varepsilon E_Q (D - D E_P D) Q)^{-1}$ yields

$$\left(I + \varepsilon E_Q \left(D - D E_P D\right) Q\right)^{-1} \left(Q^T F_0 Q\right)^{-1} Q^T = \left(Q^T F_0 Q\right)^{-1} Q^T - \tilde{R}_\varepsilon, \tag{38}$$

where

$$\begin{aligned} \tilde{R}_\varepsilon &= \varepsilon \Big(I + \varepsilon E_Q \left(D - D E_P D\right) Q\Big)^{-1} E_Q \left(D - D E_P D\right) Q \left(\left(Q^T F_0 Q\right)^{-1} Q^T\right) \\ &= \varepsilon \Big(I + \varepsilon E_Q \left(D - D E_P D\right) Q\Big)^{-1} E_Q \left(I - D E_P\right) D E_Q. \end{aligned}$$

Therefore, using (38) in (37), and noting that $Q\mathbf{x}_Q = \mathbf{x}_Q$,

$$\begin{aligned} \mathbf{x}_Q &= \left(Q^T F_0 Q\right)^{-1} Q^T \left(I - D E_P\right) \mathbf{y} - \tilde{R}_\varepsilon \left(I - D E_P\right) \mathbf{y}, \\ &= E_Q \left(I - D E_P\right) \mathbf{y} + \tilde{R}_\varepsilon \left(I - D E_P\right) \mathbf{y}. \end{aligned}$$

Solving for $\widehat{D}^{-1}\mathbf{y}$ yields the final result

$$\begin{aligned}\widehat{D}^{-1}\mathbf{y} &= \mathbf{x}_P + \mathbf{x}_Q \\ &= \frac{1}{\varepsilon}E_P\mathbf{y} + (I - E_PDQ)\mathbf{x}_Q \\ &= \frac{1}{\varepsilon}E_P\mathbf{y} + (I - E_PDQ)E_Q\left(I - DE_P\right)\mathbf{y} + (I - E_PD)\tilde{R}_\varepsilon\left(I - DE_P\right).\end{aligned}$$

Equation (35) follows by noting that

$$(I - E_PDQ)E_Q\left(I - DE_P\right)\mathbf{y} = (I - E_PD_0Q)E_Q\left(I - D_0E_P\right)\mathbf{y},$$

and that terms involving D_1 only come up in the $\mathcal{O}(\varepsilon)$ remainder term, $\tilde{R}_\varepsilon$. ∎

We can now use Lemma 10 to prove Theorems 3 and 4

Proof of Theorem 3. Recall the definition from Lemma 10, $H^{(d)} = M_t^{-1}(\mathbf{\Omega}_d\cdot\mathbf{G} + F^{(d)})$, and the identities $\sum_d w_d = 4\pi$, $\sum_d w_d\mathbf{\Omega}_d = \mathbf{0}$ and $F_0 = \frac{1}{4\pi}\sum_d w_dF^{(d)}$. Using the identity for $(I + \varepsilon H^{(d)})^{-1}$ in (55) of the appendix, $I - S_\varepsilon$ can be expanded as

$$\begin{aligned}I - S_\varepsilon &= I - \frac{1}{4\pi}\sum_d w_d\left(I + \varepsilon H^{(d)}\right)^{-1}\left(I - \varepsilon^2M_t^{-1}M_a\right) \\ &= I - \frac{1}{4\pi}\sum_d w_d\left(I - \varepsilon H^{(d)} + \varepsilon^2\left(\left(H^{(d)}\right)^2 - M_t^{-1}M_a\right)\right) + \mathcal{O}\left(\varepsilon^3\right) \\ &= \varepsilon\left[\frac{1}{4\pi}\sum_d w_dH^{(d)} - \frac{1}{4\pi}\varepsilon\sum_d w_d\left(\left(H^{(d)}\right)^2 - M_t^{-1}M_a\right)\right] + \mathcal{O}\left(\varepsilon^3\right) \\ &= \varepsilon M_t^{-1}\left(F_0 + \varepsilon\tilde{D}_\varepsilon\right) + \mathcal{O}\left(\varepsilon^3\right),\end{aligned} \tag{39}$$

where $\tilde{D}_\varepsilon$ corresponds to the latter term in (39) and is given by

$$\tilde{D}_\varepsilon = -\frac{1}{4\pi}M_t\sum_d w_d\left(\left(M_t^{-1}\left(\mathbf{\Omega}_d\cdot\mathbf{G} + F^{(d)}\right)\right)^2 - M_t^{-1}M_a\right). \tag{40}$$

Recall the identity from (17), $\mathbf{\Omega}_d\cdot\mathbf{G} + F^{(d)} = -\mathbf{\Omega}_d\cdot\mathbf{G}^T + \tilde{F}^{(d)}$, the definitions of $\mathbf{F}_1 = \frac{1}{4\pi}\sum_d w_d\mathbf{\Omega}_dF^{(d)}$ and $\tilde{\mathbf{F}}_1 = \frac{1}{4\pi}\sum_d w_d\mathbf{\Omega}_d\tilde{F}^{(d)}$ from (24), and also that because $\mathbf{\Omega}_d$ is a scalar vector, it commutes in a certain sense; for example,

$$\begin{aligned}\mathbf{\Omega}_d\cdot\boldsymbol{G}^TM_t^{-1}F^{(d)} &= (\Omega_{d_1},\Omega_{d_2},\Omega_{d_3})\cdot\left(\mathbf{G}_1^T,\mathbf{G}_2^T,\mathbf{G}_3^T\right)M_t^{-1}F^{(d)} \\ &= \left[\left(\mathbf{G}_1^T,\mathbf{G}_2^T,\mathbf{G}_3^T\right)M_t^{-1}F^{(d)}\right]\cdot(\Omega_{d_1},\Omega_{d_2},\Omega_{d_3}) \\ &= \boldsymbol{G}^TM_t^{-1}\cdot\left(F^{(d)}\mathbf{\Omega}_d\right).\end{aligned}$$

Also recall the outer product summation $\sum_d w_d\mathbf{\Omega}_d\mathbf{\Omega}_d^T = \frac{4\pi}{3}I$. Expanding the quadratic term in $\tilde{D}_\varepsilon$ and plugging in these identities yields

$$\begin{aligned}\tilde{D}_\varepsilon &= -\frac{1}{4\pi}\sum_d w_d\Big(-\boldsymbol{\Omega}_d\cdot\boldsymbol{G}^T M_t^{-1}\boldsymbol{\Omega}_d\cdot\boldsymbol{G} + \tilde{F}^{(d)}M_t^{-1}\boldsymbol{\Omega}_d\cdot\boldsymbol{G} \\ &\qquad - \boldsymbol{\Omega}_d\cdot\boldsymbol{G}^T M_t^{-1}F^{(d)} + \tilde{F}^{(d)}M_t^{-1}F^{(d)} - M_a\Big) \\ &= \frac{1}{3}\boldsymbol{G}^T M_t^{-1}\boldsymbol{G} - \tilde{\mathbf{F}}_1\cdot M_t^{-1}\mathbf{G} + \mathbf{G}^T\cdot M_t^{-1}\mathbf{F}_1 + M_a \\ &\quad - \frac{1}{4\pi}\sum_d w_d\tilde{F}^{(d)}M_t^{-1}F^{(d)}.\end{aligned}$$

Decompose $\tilde{D}_\varepsilon = D_0 + D_1$, where

$$\begin{aligned}D_0 &= \left(\frac{1}{3}\mathbf{G}^T\cdot M_t^{-1}\mathbf{G} - \tilde{\mathbf{F}}_1\cdot M_t^{-1}\mathbf{G} + \mathbf{G}^T\cdot M_t^{-1}\mathbf{F}_1 + M_a\right), \\ D_1 &= -\sum_d w_d\tilde{F}^{(d)}M_t^{-1}F^{(d)}.\end{aligned} \tag{41}$$

Then,

$$I - S_\varepsilon = \varepsilon M_t^{-1}\left(F_0 + \varepsilon\left(D_0 + D_1\right)\right) + \mathcal{O}\left(\varepsilon^3\right). \tag{42}$$

In the right-hand side of (42), the lower-order terms in ε exactly take the form of the operator in Lemma 10, where $PD_1 = D_1P = \mathbf{0}$. To that end, from (35) in Lemma 10,

$$\left(F_0 + \varepsilon\left(D_0 + D_1\right)\right)^{-1} = \frac{1}{\varepsilon}\left(\frac{1}{\varepsilon}F_0 + D_0\right)^{-1} + \mathcal{O}(\varepsilon). \tag{43}$$

Defining $D_\varepsilon = \frac{1}{\varepsilon}F_0 + D_0$, observe from (42) and (43) that

$$\begin{aligned}(\varepsilon^2 D_\varepsilon)^{-1}M_t(I - S_\varepsilon) &= \frac{1}{\varepsilon}\left(\frac{1}{\varepsilon}F_0 + D_0\right)^{-1}\left(F_0 + \varepsilon\left(D_0 + D_1\right)\right) + \mathcal{O}\left(\varepsilon\right) \\ &= I + \mathcal{O}\left(\varepsilon\right),\end{aligned}$$

which completes the proof. ∎

Proof of Theorem 4. The proof of Theorem 4 follows naturally from that of Theorem 3 and Lemma 10. From Lemma 10,

$$\left(F_0 + \varepsilon D_0\right)^{-1} = \frac{1}{\varepsilon}E_P + \left(I - E_PD_0\right)E_Q\left(I - D_0E_P\right),$$

where $E_P = P(P^TD_0P)^{-1}P^T$ and $E_Q = Q(Q^TF_0Q)^{-1}Q^T$. Defining $E_\varepsilon = (F_0 + \varepsilon D_0)^{-1}$ and appealing to (42) and (43), observe that

$$\begin{aligned}\frac{1}{\varepsilon}E_\varepsilon M_t(I - S_\varepsilon) &= \left(F_0 + \varepsilon D_0\right)^{-1}\left[\left(F_0 + \varepsilon\left(D_0 + D_1\right)\right) + \mathcal{O}(\varepsilon^2)\right] \\ &= \left(F_0 + \varepsilon D_0\right)^{-1}\left(F_0 + \varepsilon\left(D_0 + D_1\right)\right) + \mathcal{O}\left(\varepsilon\right) \\ &= I + \mathcal{O}(\varepsilon).\end{aligned}$$

∎

4.2. Bilinear form for DG near-nullspace. This section proves the identity (25) relating the DSA matrix (22) to the SIP bilinear form (26). In matrix form, (22) corresponds to the bilinear form

$$\mathbf{v}^T\left(\frac{1}{\varepsilon}F_0 + \frac{1}{3}\boldsymbol{G}^TM_t^{-1}\boldsymbol{G} - \tilde{\mathbf{F}}_1\cdot M_t^{-1}\mathbf{G} + \mathbf{G}^T\cdot M_t^{-1}\mathbf{F}_1 + M_a\right)\mathbf{u}. \tag{44}$$

Several of the relations are straightforward. The term $\mathbf{v}^T M_a \mathbf{u} = \sum_{\kappa\in\mathcal{E}} \int_\kappa \sigma_a uv d\mathbf{x}$ follows immediately from (9). Recalling the definitions of $\alpha(\mathbf{x})$ (27) and $\mathbf{v}^T F^{(d)}\mathbf{u}$ (6), along with the identity $\sum_d w_d \mathbf{\Omega}_d = \mathbf{0}$,

$$\mathbf{v}^T \frac{1}{\varepsilon} F_0 \mathbf{u} = \frac{1}{\varepsilon} \sum_d w_d \mathbf{v}^T F^{(d)} \mathbf{u} = \sum_{\Gamma\in\mathcal{F}} \int_\Gamma \alpha \left[\!\left[u\right]\!\right] \left[\!\left[v\right]\!\right] dS.$$

The remaining terms are slightly more technical, and section 4.2.1 proves that

$$\mathbf{v}^T \left(\mathbf{G}^T \cdot M_t^{-1} \mathbf{F}_1\right) \mathbf{u} = -\sum_{\Gamma\in\mathcal{F}} \int_\Gamma \frac{1}{3\sigma_t} \left\{\mathbf{n}\cdot\nabla_{\mathbf{x}} v\right\} \left[\!\left[u\right]\!\right] dS, \tag{45}$$

$$\mathbf{v}^T \left(\tilde{\mathbf{F}}_1 \cdot M_t^{-1} \mathbf{G}\right) \mathbf{u} = \sum_{\Gamma\in\mathcal{F}} \int_\Gamma \frac{1}{3\sigma_t} \left\{\mathbf{n}\cdot\nabla_{\mathbf{x}} u\right\} \left[\!\left[v\right]\!\right] dS. \tag{46}$$

Together the above results combine to yield the identity in (25).

4.2.1. Face matrix terms in bilinear form. This section starts with a lemma expressing the action of $M_t^{-1}\mathbf{G}$ in the context of bilinear forms.

LEMMA 11. *In the case of straight-edged meshes and constant opacities, $1/\sigma_t$, for each mesh element, κ_e, $(M_t^{-1}\mathbf{G})\mathbf{u}$ is related to $\frac{1}{\sigma_t}\nabla_{\mathbf{x}} u$ via*

$$\left(M_t^{-1}\right)_{e,e} \mathbf{G}_{e,e} \left[\boldsymbol{u}_e\right]_m = \frac{1}{\sigma_t} \left[\left(\nabla_{\mathbf{x}} u_e\right)\left(\mathbf{x}_{e,m}\right)\right]_m. \tag{47}$$

Moreover, for any bilinear form $\mathcal{B}(u,v)$ with associated matrix B, $\mathcal{B}(u,v) = \mathbf{v}^T B \mathbf{u}$, it holds that

$$\mathcal{B}\left(\frac{1}{\sigma_t}\partial_{x_j} u, v\right) = \mathbf{v}^T \left(B\left(M_t^{-1} G^{(j)}\right)\right) \mathbf{u}.$$

Proof. Without loss of generality, expand $u(\mathbf{x})$ in a piecewise polynomial basis consisting of interpolating polynomials $\{u_{e,m}(\mathbf{x})\}_{e,m}$,

$$u(\mathbf{x}) = \sum_m u\left(\mathbf{x}_{e,m}\right) u_{e,m}(\mathbf{x}), \quad \mathbf{x}\in\kappa_e.$$

Since the mesh transformation from the reference element $\hat{\kappa}$ to the physical element κ_e is linear, $\partial_{x_j} u_e(\mathbf{x})$ is a polynomial of degree less than or equal to the degree of $u_e(\mathbf{x})$, and so

$$\partial_{x_j} u_e(\mathbf{x}) = \sum_n u_e\left(\mathbf{x}_{e,n}\right) \partial_{x_j} u_{e,n}(\mathbf{x}) = \sum_n \left(\partial_{x_j} u_e\right)\left(\mathbf{x}_{e,n}\right) u_{e,n}(\mathbf{x}).$$

Therefore,

$$\int_{\kappa_e} \left(\partial_{x_j} u_e\right) u_{e,m} d\mathbf{x} = \sum_n \partial_{x_j} u_e\left(\mathbf{x}_{e,n}\right) \int_{\kappa_e} u_{e,n} u_{e,m} d\mathbf{x} = \sum_n u_e\left(\mathbf{x}_{e,n}\right) \int_{\kappa_e} \left(\partial_{x_j} u_{e,n}\right) u_{e,m} d\mathbf{x}.$$

Recalling that $G_e^{(j)} = [\int_{\kappa_e} u_{e,m}(\partial_{x_j} u_{e,n}) d\mathbf{x}]_{mn}$ and $M_{t,e} = \sigma_t[\int_{\kappa_e} u_{e,n} u_{e,m} d\mathbf{x}]_{mn}$, we can write the above identity as

$$\sigma_t^{-1} \left(M_t\right)_{e,e} \left[\partial_{x_j}\varphi_e\left(\mathbf{x}_m\right)\right]_m = G_e^{(j)} \left[\varphi_e\left(\mathbf{x}_m\right)\right]_m.$$

Applying $(M_t^{-1})_{e,e}$ to both sides above yields (47).

Now consider the bilinear form $\mathcal{B}\left(\sigma_t^{-1}\partial_{x_j}u, v\right)$ and suppose that the matrix B is such that

$$\mathbf{v}^T B\mathbf{u} = \mathcal{B}\left(u, v\right),$$

for any $u(\cdot)$ and $v(\cdot)$ in the DG space. Then we use the following: if σ_t is constant, then the bilinear form $\mathcal{B}(u, \partial_{x_j}v)$ corresponds to the matrix $B(M_t^{-1}G^{(j)})$. Indeed, letting $B_{e',e}$ denote the submatrix of B corresponding to elements $\kappa_{e'}$ and κ_e,

$$\begin{aligned}
\mathcal{B}\left(\frac{1}{\sigma_t}\partial_{x_j}u, v\right) &= \sum_{e,e'}\sum_{m,n}\left[v_{e'}\left(\mathbf{x}_{e',m}\right)\right]_m \left(B_{e',e}\right)_{m,n}\frac{1}{\sigma_t}\left[\left(\partial_{x_j}u_e\right)\left(\mathbf{x}_{e,n}\right)\right]_n \\
&= \sum_{e,e'}\mathbf{v}_{e'}^T B_{e',e}\left(\left(M_{t,e}^{-1}G_e^{(j)}\right)\mathbf{u}_e\right) \\
&= \sum_{e,e'}\mathbf{v}_{e'}^T\left(B_{e',e}\left(M_{t,e}^{-1}G_e^{(j)}\right)\right)\mathbf{u}_e \\
&= \mathbf{v}^T\left(B\left(M_t^{-1}G^{(j)}\right)\right)\mathbf{u}.
\end{aligned}$$

Using (24) and (18), and the identities $\sum_d w_d\mathbf{\Omega}_d\mathbf{\Omega}_d^T = \frac{4\pi}{3}I$ and $\sum_d w_d\mathbf{\Omega}_d|\mathbf{\Omega}_d\cdot\mathbf{n}| = \mathbf{0}$,

$$\begin{aligned}
\mathbf{v}^T\left(\mathbf{F}_1\right)_j\mathbf{u} &= -\sum_{\Gamma\in\mathcal{F}}\int_\Gamma\left(\frac{1}{4\pi}\sum_d w_d\left(\mathbf{\Omega}_d\right)_j\mathbf{\Omega}_d^T\right)\mathbf{n}\left[\!\left[u\right]\!\right]\left\{v\right\}dS \\
&\quad + \frac{1}{2}\sum_{\Gamma\in\mathcal{F}^{\mathrm{i}}}\int_\Gamma\left(\frac{1}{4\pi}\sum_d w_d\left(\mathbf{\Omega}_d\right)_j\left|\mathbf{\Omega}_d\cdot\mathbf{n}\right|\right)\left[\!\left[u\right]\!\right]\left[\!\left[v\right]\!\right]dS \\
&= -\frac{1}{3}\sum_{\Gamma\in\mathcal{F}}\int_\Gamma\mathbf{n}_j\left[\!\left[u\right]\!\right]\left\{v\right\}dS,
\end{aligned} \tag{48}$$

where $\mathbf{n}_j$ denotes the jth component of the normal vector $\mathbf{n}$. Similarly,

$$\begin{aligned}
\mathbf{v}^T\left(\tilde{\mathbf{F}}_1\right)_j\mathbf{u} &= \sum_{\Gamma\in\mathcal{F}}\int_\Gamma\left(\frac{1}{4\pi}\sum_d w_d\left(\mathbf{\Omega}_d\right)_j\mathbf{\Omega}_d^T\right)\mathbf{n}\left\{u\right\}\left[\!\left[v\right]\!\right]dS \\
&\quad + \frac{1}{2}\sum_{\Gamma\in\mathcal{F}}\int_\Gamma\left(\frac{1}{4\pi}\sum_d w_d\left(\mathbf{\Omega}_d\right)_j\left|\mathbf{\Omega}_d\cdot\mathbf{n}\right|\right)\left[\!\left[u\right]\!\right]\left[\!\left[v\right]\!\right]dS \\
&= \frac{1}{3}\sum_{\Gamma\in\mathcal{F}}\int_\Gamma\mathbf{n}_j\left\{u\right\}\left[\!\left[v\right]\!\right]dS.
\end{aligned} \tag{49}$$

Applying Lemma 11 yields (45) and (46). ∎

4.3. Fixed-point iteration on HO meshes. Theorem 5 follows from the following lemma.

LEMMA 12. *Consider a linear system of the form in* (28)*, with condensed notation*

$$(I + \varepsilon H - B)\boldsymbol{\psi}^{(d)} = \mathbf{s}^{(d)}. \tag{50}$$

Denote $\mathcal{H} := I + \varepsilon H - B$*, and consider a matrix splitting* $H = H_{\leq} + H_{>}$*. Define* $\widetilde{M}^{-1} = (I + \varepsilon H)^{-1}$ *as the preconditioner associated with inverting* $I + \varepsilon H$*. Now, fix* $B\boldsymbol{\psi}^{(d)}$ *and move it to the right-hand side, for the modified linear system*

$$(I + \varepsilon H)\boldsymbol{\psi}^{(d)} = \mathbf{s}^{(d)} + B\boldsymbol{\psi}_0^{(d)}. \tag{51}$$

Define $\widehat{M}_k^{-1}$ *as the preconditioner associated with performing* k *fixed-point iterations on* (51)*, with approximate inverse* $\widehat{M}_1^{-1} = (I + \varepsilon H_{\leq})^{-1}$*. Then, applying* $\widetilde{M}^{-1}$ *and* $\widehat{M}_k^{-1}$ *as preconditioners for* (50) *is related via*

$$\widehat{M}_k^{-1}\mathcal{H} = \left(I - \left(-\varepsilon(I + \varepsilon H_{\leq})^{-1}H_{>}\right)^k\right)\widetilde{M}^{-1}\mathcal{H}.$$

Proof of Theorem 5. Consider a problem of the form

$$(I + \varepsilon H_{\leq} + \varepsilon H_{>} - B)\boldsymbol{\psi}^{(d)} = \mathbf{s}^{(d)},$$

where $\mathcal{H} := (I + \varepsilon H_{\leq} + \varepsilon H_{>} - B)$. Note the following identities, which will be used regularly:

$$\begin{aligned}
(I + \varepsilon H_{\leq} + \varepsilon H_{>} - B)^{-1} &= \left[I - (I + \varepsilon H_{\leq} + \varepsilon H_{>})^{-1}B\right]^{-1}(I + \varepsilon H_{\leq} + \varepsilon H_{>})^{-1},\\
(I + \varepsilon H_{\leq} + \varepsilon H_{>} - B)^{-1} &= \left[I - (I + \varepsilon H_{\leq})^{-1}(-\varepsilon H_{>} + B)\right]^{-1}(I + \varepsilon H_{\leq})^{-1},\\
(I + \varepsilon H_{\leq} + \varepsilon H_{>})^{-1} &= \left(I + \varepsilon(I + \varepsilon H_{\leq})^{-1}H_{>}\right)^{-1}(I + \varepsilon H_{\leq})^{-1}.
\end{aligned}$$

First, consider a single fixed-point iteration, where we invert $I + \varepsilon H_{\leq} + \varepsilon H_{>}$. Define $\widetilde{M}^{-1} = (I + \varepsilon H_{\leq} + \varepsilon H_{>})^{-1}$. Then, the preconditioned linear system is given by $\widetilde{M}^{-1}(\mathcal{H}\boldsymbol{\psi}^{(d)} - \mathbf{s}^{(d)}) = \mathbf{0}$, where

$$\begin{aligned}
\widetilde{M}^{-1}\mathcal{H} &= (I + \varepsilon H_{\leq} + \varepsilon H_{>})^{-1}(I + \varepsilon H_{\leq} + \varepsilon H_{>} - B)\\
&= I - (I + \varepsilon H_{\leq} + \varepsilon H_{>})^{-1}B.
\end{aligned}$$

Now suppose we only invert $I + \varepsilon H_{\leq}$; that is, our preconditioner is given by $\widehat{M}_1^{-1} = (I + \varepsilon H_{\leq})^{-1}$. This arises, for example, in the case of cycles in the mesh, where we can only directly invert the block lower triangular part. In the interest of asymptotics, additionally consider moving $B\boldsymbol{\psi}$ to the right-hand side and applying multiple iterations of $\widehat{M}_1^{-1}$ to the modified linear system, $\widehat{\mathcal{H}}\boldsymbol{\psi}^{(d)} = \widehat{\mathbf{s}}^{(d)}$, given by

$$(I + \varepsilon H_{\leq} + \varepsilon H_{>})\boldsymbol{\psi}^{(d)} = \mathbf{s}^{(d)} + B\boldsymbol{\psi}_0^{(d)},$$

where $\boldsymbol{\psi}_0^{(d)}$ is fixed for all iterations. In a fixed-point sense, this is equivalent to

$$\boldsymbol{\psi}_{k+1}^{(d)} = \boldsymbol{\psi}_k^{(d)} + (I + \varepsilon H_{\leq})^{-1}\left(\mathbf{s}^{(d)} + B\boldsymbol{\psi}_0^{(d)} - \varepsilon(\mathcal{H}_{\leq} + \mathcal{H}_{>})\boldsymbol{\psi}_k^{(d)}\right),$$

with error propagation given by

$$\begin{aligned}
I - \widehat{M}_1^{-1}\widehat{\mathcal{H}} &= I - (I + \varepsilon H_{\leq})^{-1}\widehat{\mathcal{H}}\\
&= -\varepsilon(I + \varepsilon H_{\leq})^{-1}H_{>}.
\end{aligned}$$

Then, we are interested in the preconditioner $\widehat{M}_k$ that results from taking powers of $I - \widehat{M}_k^{-1}\widehat{\mathcal{H}} = (I - \widehat{M}_1^{-1}\widehat{\mathcal{H}})^k$. Solving for $\widehat{M}_k^{-1}$,

$$
\begin{aligned}
\widehat{M}_k^{-1}\widehat{\mathcal{H}} &= I - \left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^k, \\
\widehat{M}_k^{-1} &= \widehat{\mathcal{H}}^{-1} - \left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^k \widehat{\mathcal{H}}^{-1} \\
&= \left(I - -\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right]\right) -1(I+\varepsilon H_{\leq})^{-1} \\
&\quad - \left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^k \left(I - (I+\varepsilon H_{\leq})^{-1}(-\varepsilon H_{>}+B)\right)^{-1}(I+\varepsilon H_{\leq})^{-1} \\
&= \left[\sum_{\ell=0}^{\infty}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell} - \sum_{\ell=k}^{\infty}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell}\right](I+\varepsilon H_{\leq})^{-1} \\
&= \left[\sum_{\ell=0}^{k-1}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell}\right](I+\varepsilon H_{\leq})^{-1}.
\end{aligned}
$$

Now, suppose we apply $\widehat{M}_k^{-1}$ as a preconditioner for the original linear system, $\mathcal{H}\boldsymbol{\psi}^{(d)} = \mathbf{s}^{(d)}$, and consider the difference between $\widehat{M}_k^{-1}$ and $\widetilde{M}^{-1}$:

$$
\begin{aligned}
&\widehat{M}_k^{-1} - \widetilde{M}^{-1} \\
&\quad = \left[\sum_{\ell=0}^{k-1}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell}\right](I+\varepsilon H_{\leq})^{-1} - (I+\varepsilon H_{\leq}+\varepsilon H_{>})^{-1} \\
&\quad = \left[\sum_{\ell=0}^{k-1}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell}\right](I+\varepsilon H_{\leq})^{-1} - \left(I+\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{-1}(I+\varepsilon H_{\leq})^{-1} \\
&\quad = \left[\sum_{\ell=0}^{k-1}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell} - \sum_{\ell=0}^{\infty}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell}\right](I+\varepsilon H_{\leq})^{-1} \\
&\quad = -\left[\sum_{\ell=k}^{\infty}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell}\right](I+\varepsilon H_{\leq})^{-1} \\
&\quad = -\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^k \left[\sum_{\ell=0}^{\infty}\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^{\ell}\right](I+\varepsilon H_{\leq})^{-1} \\
&\quad = -\left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^k \widetilde{M}^{-1}.
\end{aligned}
$$

Then,

$$
\widehat{M}_k^{-1}\mathcal{H} = \left[I - \left(-\varepsilon(I+\varepsilon H_{\leq})^{-1}H_{>}\right)^k\right]\widetilde{M}^{-1}\mathcal{H}. \qquad \square
$$

5. Numerical experiments. In this section we report numerical results from the three major approaches presented in the previous sections, namely, the SIP DSA (Theorem 3), its IP modification (section 3.2.2), and the additive DSA preconditioner (Theorem 4). We also show that performing two additional transport sweeps in between DSA steps (Theorem 5) can greatly accelerate (or prevent divergence of) source iteration on HO meshes with mesh cycles.

We present calculations and comparisons on highly curved two- and three-dimensional meshes that are obtained from moving mesh hydrodynamic simulations [10]. The methods in this paper are implemented by utilizing the finite element infrastructure provided by the MFEM finite element library [25].

5.1. DSA preconditioning on a HO Lagrangian mesh. This section uses DSA to solve the discrete transport equations (2) on a HO hydrodynamics mesh

generated from a purely Lagrangian simulation of the "triple point" problem [15], which is displayed in Figure 1 (the spatial domain for this problem is $[0,7] \times [0,3]$). The mesh is third-order mesh; that is, cubic polynomials are used to map the reference element to physical elements, and our DG discretization uses third-order local basis functions. We also use an S2 quadrature discretization.
For this problem, we use constant opacities

$$\sigma_t(x_1, x_2) = \frac{1}{\varepsilon}, \quad \sigma_a = \varepsilon,$$

and a smooth (but arbitrary) source term

$$q(x_1, x_2) = \varepsilon \cos^2(2x_1 + x_2),$$

where ε is the characteristic mean free path. In our numerical experiments, we vary ε from relatively optically thin regimes $\varepsilon = .75$, to increasingly optically thick regimes $\varepsilon = 10^{-j}$, for $j = 1, 2, 3, 4$. Last, constant inflow boundary conditions are applied,

$$\psi_d(\mathbf{x}) = 1 \quad \text{when} \quad \mathbf{\Omega}_d \cdot \mathbf{n}(\mathbf{x}) < 0 \quad \text{and} \quad \mathbf{x} \in \partial\mathcal{D}.$$

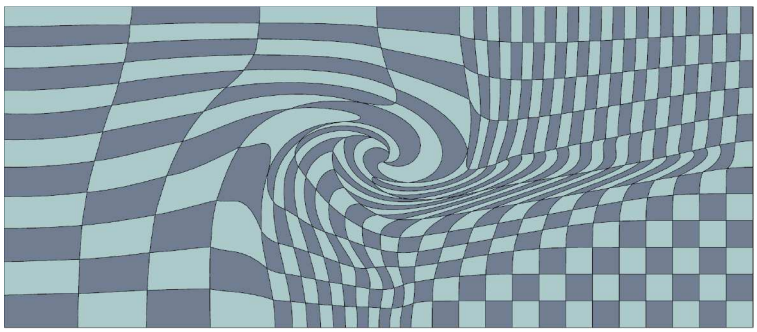
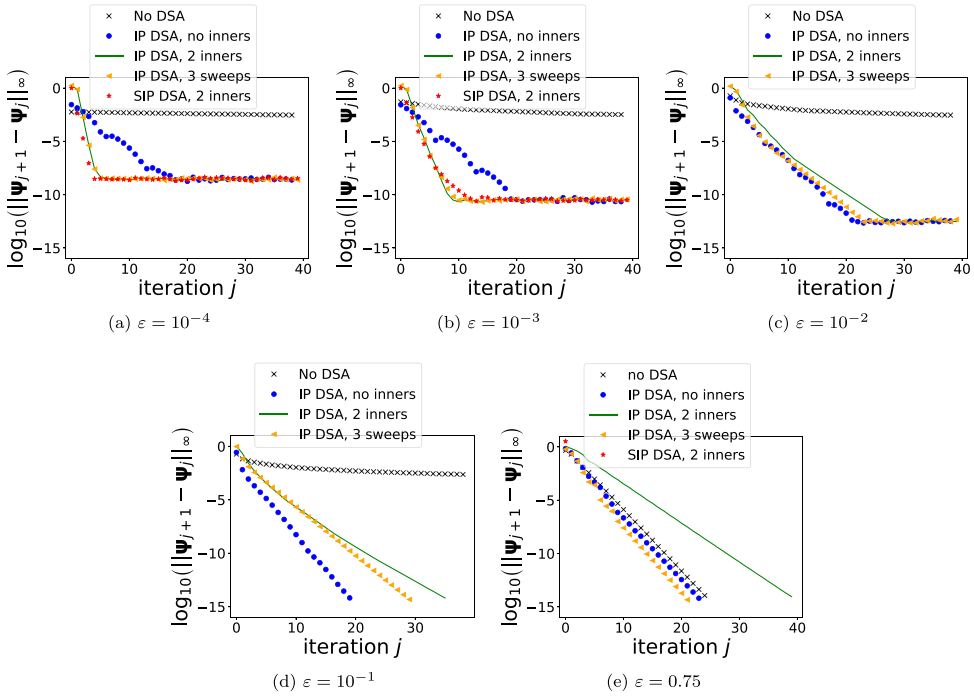


FIG. 1. *A "triple point" third-order Lagrangian mesh.*

Figure 2 shows the iteration error estimate $\|\boldsymbol{\psi}_{j+1} - \boldsymbol{\psi}_{j+1}\|_\infty$ as a function of iteration index j, with and without DSA preconditioning. Recall, due to cycles in the mesh, the transport equation for a fixed angle cannot be easily inverted, so we invert the block lower triangular part of the matrix, and refer to this as a "transport sweep." When DSA preconditioning is included, we consider using a single transport sweep with lagging between DSA steps, as well as using two "inner sweeps," where the scalar flux is not updated, followed by one normal sweep with lagging between DSA steps (see Theorem 5). Finally, we also consider performing, between every DSA step, three transport sweeps, where the scalar flux is updated after each sweep. To ensure a fair comparison, the iteration index j in all five cases displayed in Figure 2 accounts for *the same number of transport sweeps*; however, because of this, each case has a different interpretation:

1. In the "no DSA" case, the iteration index j corresponds to three applications of the fixed-point iteration without any DSA, i.e., (sweep, update flux)^3; for example, $j = 10$ corresponds to 30 fixed-point iterations.
2. In the "IP DSA, no inners" case, the iteration index j represents three applications of a transport sweep and nonsymmetric IP DSA step, i.e., (sweep, update flux, IP DSA)^3.



(a) $\varepsilon = 10^{-4}$ (b) $\varepsilon = 10^{-3}$ (c) $\varepsilon = 10^{-2}$

(d) $\varepsilon = 10^{-1}$ (e) $\varepsilon = 0.75$

FIG. 2. *Iteration error estimate* $\|\boldsymbol{\psi}_{j+1} - \boldsymbol{\psi}_{j+1}\|_\infty$ *as a function of iteration index* j *on the triple point problem, with and without DSA preconditioning using the DSA matrix* (31). *In all cases considered, we always perform* 40 *"iterations." Although the iteration index* j *has a different interpretation for the five displayed cases "no DSA," "IP DSA, no inners," "IP DSA,* 2 *inners," "IP DSA,* 3 *sweeps," and "SIP DSA,* 2 *inners," each iteration involves the same number of transport sweeps. For example, "IP DSA,* 2 *inners" refers to the nonsymmetric IP version of DSA with three total transport sweeps between DSA steps (but with a fixed scalar flux), and "IP DSA,* 3 *sweeps" refers to the nonsymmetric IP version of DSA with a DSA step three transport sweeps (where the scalar flux is updated after each sweep). Similarly, each iteration index in the "no DSA" option corresponds to performing* 3 *transport sweeps per iteration index. In the plots where "SIP DSA,* 2 *inners" is not displayed, the fixed-point iteration diverged for this case.*

3. In the "IP DSA, 2 inners" case, j corresponds to three applications of a transport sweep followed by a scalar flux update and a single IP DSA step, that is, ((sweep)^3, update flux, IP DSA).
4. In the "IP DSA, 3 sweeps" case, the index j represents three applications of both a transport sweep and scalar flux update, followed by a nonsymmetric IP DSA step, i.e., ((sweep, update flux)^3, IP DSA).
5. Finally, the "SIP DSA, 2 inners" case is the same as the "IP DSA, 2 inners" case, but with the SIP DSA matrix used instead.

Because the sweep is typically computationally much more expensive than the DSA step, each iteration index in Figure 2 approximately represents the same computational work for each case. In particular, for small ε, Figure 2 provides numerical confirmation of the asymptotic result (Theorem 5). Using three sweeps before each IP DSA step leads to a $4\times$ speedup for $\varepsilon = 10^{-4}$ when using the nonsymmetric IP DSA matrix (at a slightly lesser cost as well, due to two less diffusion solves), although the cost increases in the optically thin regime relative to using no additional transport sweeps. In addition, although we didn't show this in Figure 2, the SIP DSA

TABLE 1
The final residual (52) *after* 40 *iterations with and without DSA preconditioning.*

ε	IP DSA, 3 sweeps	IP DSA, 2 inners	SIP DSA, 2 inners	IP DSA, no inners	no DSA
0.75	4.84e-15	2.37e-15	2.31e-15	3.02e-15	3.23e-15
1e-1	1.50e-14	3.34e-15	diverged	4.24e-15	6.92e-03
1e-2	1.95e-13	3.04e-13	diverged	2.05e-13	6.16e-02
1e-3	1.67e-11	2.04e-11	1.93e-11	2.23e-11	1.41e-01
1e-4	1.97e-09	1.69e-09	1.98e-09	2.77e-09	8.56e-01

variant actually diverges for $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-4}$ when the inner iterations are not performed.

Table 1 displays the L^∞ residuals of the final iterates,

(52)
$$\max_d \left\| \left(\mathbf{\Omega}_d \cdot \mathbf{G} + F^{(d)} + \frac{1}{\varepsilon} M_t \right) \boldsymbol{\psi}^{(d)} - \frac{1}{4\pi} \left(\frac{1}{\varepsilon} M_t - \varepsilon M_a \right) \boldsymbol{\varphi} - \frac{1}{4\pi} \left(\boldsymbol{q}_{\text{inc}}^{(d)} + \varepsilon \boldsymbol{q}^{(d)} \right) \right\|_\infty ,$$

as well as the iterations counts. Together, Figure 2 and Table 1 confirm that DSA preconditioning on the HO mesh is effective across a wide range of characteristic mean free paths. Interestingly, although using three transport sweeps between DSA steps is more effective for small ε, for larger values of ε it is best to apply a DSA step after each sweep.

Note that as ε gets smaller than 10^{-4}, the DSA preconditioner begins to degrade in effectiveness and ultimately leads to a divergent fixed-point iteration. This degradation in efficiency is likely due to the fact that the condition number of the system (13) scales like $\mathcal{O}(\varepsilon^{-2})$, and for smaller values of ε the delicate cancellations in the derivation of the DSA preconditioner can no longer be adequately captured in floating point arithmetic.

5.2. DSA preconditioning on a three-dimensional curved mesh. This section uses DSA to solve the discrete transport equations (2) on a HO hydrodynamics mesh generated from a purely Lagrangian simulation of a three-dimensional Rayleigh–Taylor instability (see Figure 3). Again we utilize a third-order mesh and third-order basis functions. We also use an S4 quadrature discretization.
For this problem we use spatially dependent opacities,

$$\sigma_t(x_1, x_2, x_3) = \frac{x_1^2 + x_1 x_2 + 1}{\varepsilon}, \qquad \sigma_a(x_1, x_2, x_3) = \frac{x_1^2 + x_1 x_2 + 1}{\varepsilon} - \varepsilon(x_1^2 + x_1 x_2 + 0.5),$$

and a smooth source term,

$$q(x_1, x_2, x_3) = \sin^2(4x_1 + 2x_2 + 2x_2 x_3) + 1,$$

where ε is the characteristic mean free path. As in the previous section, we vary ε from relatively optically thin regimes, $\varepsilon = 0.75$, to increasingly optically thick regimes, $\varepsilon = 10^{-j}$, for $j = 1, 2, 3, 4$. Lastly, constant inflow boundary conditions are applied:

$$\psi_d(\mathbf{x}) = 1 \quad \text{when} \quad \mathbf{\Omega}_d \cdot \mathbf{n}(\mathbf{x}) < 0 \quad \text{and} \quad \mathbf{x} \in \partial\mathcal{D}.$$

We repeat the numerical experiments from section 5.1 to this three-dimensional problem, using the above configuration. The notation in Figure 4 and Table 2 follows the notation in section 5.1. We observe that all DSA preconditioning options for this three-dimensional problem lead to similar convergence trends as in the two-dimensional problem. Again, best convergence is achieved by the IP DSA options.

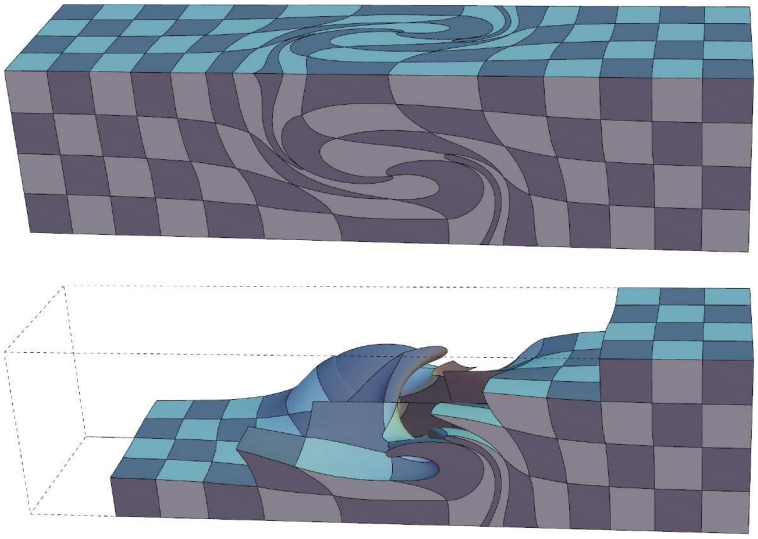
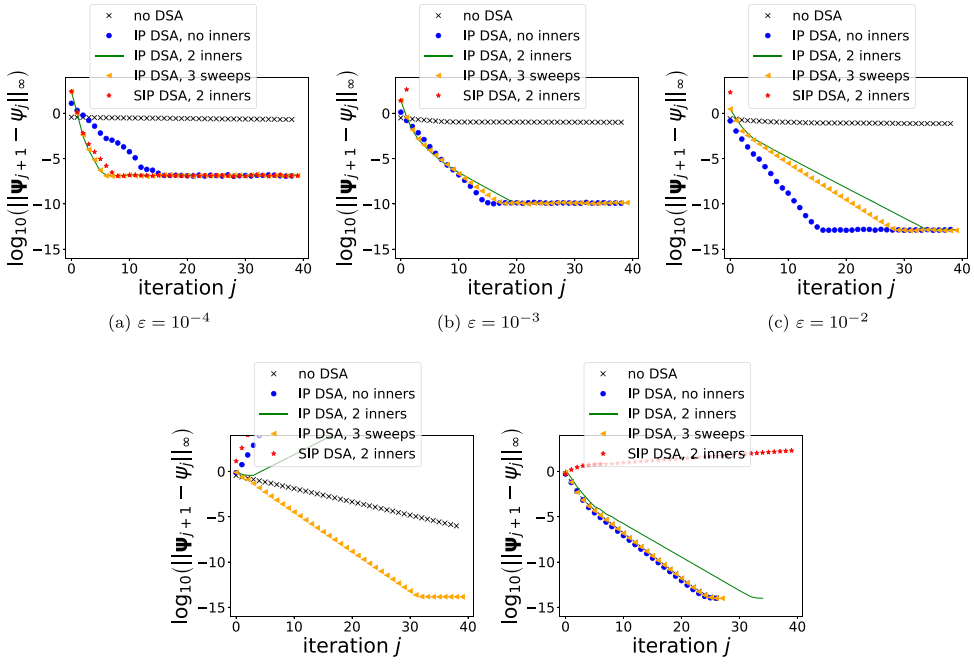


FIG. 3. *Cubic three-dimensional mesh (top) resulting from a Lagrangian simulation of the Rayleigh–Taylor instability. A subset of the mesh elements is shown in the bottom.*



(a) $\varepsilon = 10^{-4}$ (b) $\varepsilon = 10^{-3}$ (c) $\varepsilon = 10^{-2}$

(d) $\varepsilon = 10^{-1}$ (e) $\varepsilon = 0.75$

FIG. 4. *Iteration error estimate* $\|\boldsymbol{\psi}_{j+1} - \boldsymbol{\psi}_{j+1}\|_\infty$ *as a function of iteration index* j *on the three-dimensional Rayleigh–Taylor mesh, with and without DSA preconditioning using the DSA matrix* (31).

TABLE 2
The final residual (52) *after* 40 *iterations with and without DSA preconditioning for the three-dimensional Rayleigh–Taylor mesh.*

ε	IP DSA, 3 sweeps	IP DSA, 2 inners	SIP DSA, 2 inners	IP DSA, no inners	no DSA
0.75	9.97e-15	9.96e-15	diverged	9.90e-15	9.99e-15
1e-1	1.43e-14	diverged	diverged	diverged	6.94e-07
1e-2	1.21e-13	1.45e-13	diverged	1.43e-13	7.31e-02
1e-3	1.36e-10	1.36e-10	diverged	1.14e-10	9.97e-02
1e-4	1.32e-07	1.57e-7	1.25e-07	1.43e-07	2.04e-01

5.3. Additive DSA preconditioning (Theorem 4). In the thick regime, the DSA matrix derived in Theorem 3 is effective when it can be readily inverted; however, its condition number scales like $1/\varepsilon$ (in addition to the standard $1/h^2$ scaling). In addition to the fact that even standard DG discretizations of elliptic problems can be difficult for fast linear solvers such as AMG, inverting the discrete diffusion operator derived here is not trivial. Theorem 4 developed a two-part additive DSA preconditioner that requires inverting a continuous diffusion discretization three times and a mass-matrix like operator once, both of which are more tractable to quickly (approximately) invert in parallel. This section demonstrates on a one-dimensional domain that the derived two-part DSA preconditioner is indeed effective with respect to convergence of the larger iteration. Like the IP DSA variant, our results indicate that the new DSA variant based on Theorem 4 is robust across all values of ε (thin, intermediate, and thick). In contrast, the SIP DSA preconditioner actually results in a divergent iteration for $\varepsilon = 1e-3$ in the problem detailed below (however, as detailed in [34], the SIP DSA preconditioner can be stabilized outside of the thick limit by modifying the penalty coefficient appropriately. We note that the variant of DSA based on Theorem 4 is still significantly less effective as the IP DSA variant when ε is in the intermediate regimes.

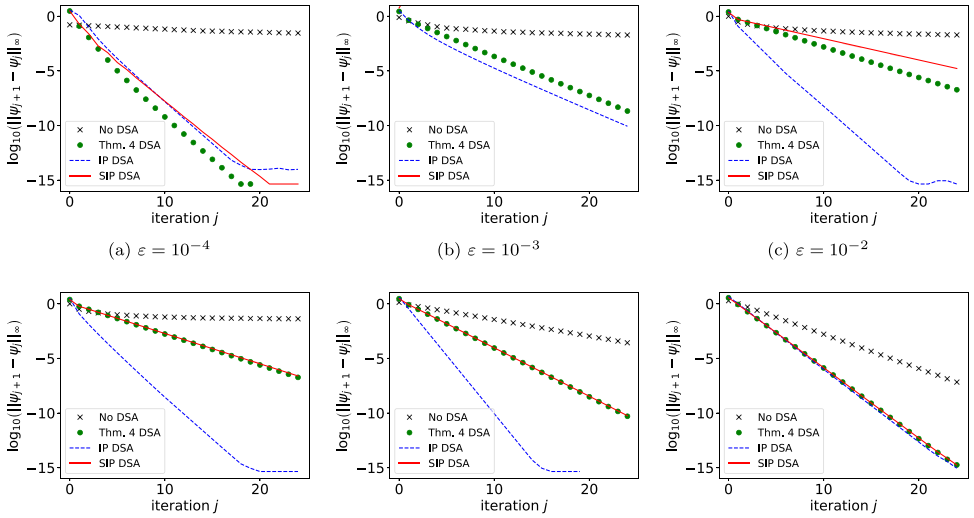
Let σ_a and σ_t be defined as before, with zero inflow boundary conditions, and source term

$$q(x, mu) = \varepsilon \left(2\sin(3x^2)^2 + \cos(x/3)^2\right).$$

Figure 5 plots the global residual as a function of iteration number for no DSA, SIP DSA, NIP DSA, and DSA based on Theorem 4. We use sixth order local basis functions, 100 mesh elements, and an S4 quadrature discretization. Note that, in Figure 5, the caption "IP DSA" denotes the same option as the "IP DSA, 2 inners" option from the two- and three-dimensional results, since in the one-dimensional case there is no need to handle mesh cycles in the transport sweep.

6. Conclusions. This paper derives a discrete analysis of DSA applied to HO DG discretizations of the S_N transport equations. The basis for DSA is taking a simple fixed-point "source iteration," which is slow to converge, and recognizing that the slowly decaying error modes can be represented by a certain diffusion operator. DSA then preconditions source iteration with an appropriate diffusion solve, as a correction for these slowly decaying error modes. When the mean free path of particles is very small, $\varepsilon \ll 1$, conditioning of source iteration is $\mathcal{O}(1/\varepsilon^2)$, and DSA is critical for convergence.

Here, we derive a discrete representation of the slowly decaying error modes for small ε. This leads to the development of a DSA preconditioner that resembles



(a) $\varepsilon = 10^{-4}$ (b) $\varepsilon = 10^{-3}$ (c) $\varepsilon = 10^{-2}$

(d) $\varepsilon = 0.1$ (e) $\varepsilon = 0.5$ (f) $\varepsilon = 0.75$

FIG. 5. *ℓ^∞-error, $\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_j\|_\infty$, as a function of iteration number j, where $\hat{\boldsymbol{\psi}}$ is the exact solution and $\boldsymbol{\psi}_j$ the jth iterate. Error is shown for no DSA, SIP DSA, IP DSA, and a two-part DSA preconditioner based on Theorem 4.*

a SIP DG discretization of diffusion-reaction, where the resulting (preconditioned) fixed-point iteration is conditioned like $1 + \mathcal{O}(\varepsilon)$ (Theorem 3). However, applying this preconditioner requires inverting a DG matrix that is ill-conditioned, $\kappa \sim \mathcal{O}(1/\varepsilon)$, and, furthermore, elliptic DG discretizations are often difficult for fast preconditioners such as multigrid. This motivates further analysis, where a two-part additive DSA preconditioner is developed based on solving a CG discretization of diffusion-reaction, in addition to a second term that involves two CG solves, and one solve of a mass-matrix-like term. These solves are now all conditioned independent of ε and more amenable to fast solvers such as multigrid. Furthermore, the preconditioner leads to a larger fixed-point iteration that is well conditioned, $\kappa \sim 1 + \mathcal{O}(\varepsilon)$, and will converge rapidly for small ε (Theorem 4).

Finally, there is larger interest in discretizing HO DG on HO (curved) meshes. Source iteration relies on the discretization of advection being block triangular in some ordering and, therefore, easily invertible. However, HO meshes lead to cycles in the mesh, and the resulting discretization of advection in the transport equations is no longer block triangular. When cycles are present, a method to approximate the inversion of advection in source iteration through a pseudo-optimal Gauss–Seidel has been developed in [13]. Theorem 5 extends the handling of cycles to cases where DSA is necessary, proving that cycles can be accounted for by performing an additional two source iterations for each larger DSA iteration.

7. Appendix. First we introduce a technical lemma regarding the linear system $(I - T_\varepsilon)\boldsymbol{\psi}^{(d)} = \tilde{\boldsymbol{q}}^{(d)}$; see (13). This then leads to Proposition 2, which proves that the conditioning of $(I - T_\varepsilon)$ is $\mathcal{O}(\varepsilon^{-2})$, making effective preconditioning critical for small ε.

LEMMA 13. *Define*

$$H^{(d)} = M_t^{-1}\left(\mathbf{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right), \tag{53}$$
$$c_0 = \max\left\{\max_d \|H^{(d)}\|, \|M_t^{-1}M_a\|\right\},$$

and assume that $\varepsilon\|H^{(d)}\| < 1$. *Then, the operator in* (13) *satisfies*

$$\begin{aligned}((I - T_\varepsilon)\boldsymbol{\psi})^{(d)} &= \left(\boldsymbol{\psi}^{(d)} - \frac{1}{4\pi}\boldsymbol{\varphi}\right) + \varepsilon\frac{1}{4\pi}H^{(d)}\boldsymbol{\varphi} \\ &\quad - \frac{1}{4\pi}\varepsilon^2\left(\left(H^{(d)}\right)^2 - M_t^{-1}M_a\right)\boldsymbol{\varphi} + R_\varepsilon^{(d)},\end{aligned} \tag{54}$$

where the norm of the remainder $R_\varepsilon^{(d)}$ *is bounded by*

$$\left\|R_\varepsilon^{(d)}\right\| \le \varepsilon^3\frac{1}{4\pi}\left(\frac{c_0^3}{1-\varepsilon c_0}\left(1+\varepsilon^2 c_0\right) + \left(c_0^2 + \varepsilon c_0^3\right)\right)\|\boldsymbol{\varphi}\|.$$

Proof. Note the matrix identity,

$$\left(I + \varepsilon H^{(d)}\right)^{-1} = I - \varepsilon H^{(d)} + \varepsilon^2\left(H^{(d)}\right)^2 - \varepsilon^3\left(H^{(d)}\right)^3\left(I + \varepsilon H^{(d)}\right)^{-1}. \tag{55}$$

Plugging into the definition of T_ε and expanding yields

$$\begin{aligned}((I - T_\varepsilon)\boldsymbol{\psi})^{(d)} &= \left[I - \frac{1}{4\pi}\left(I + \varepsilon H^{(d)}\right)^{-1}\left(I - \varepsilon^2 M_t^{-1}M_a\right)P_0\right]\boldsymbol{\psi}^{(d)} \\ &= \boldsymbol{\psi}^{(d)} - \frac{1}{4\pi}\left[I - \varepsilon H^{(d)} + \varepsilon^2\left(\left(H^{(d)}\right)^2 - M_t^{-1}M_a\right)\right. \\ &\quad - \varepsilon^3\left(\left(H^{(d)}\right)^3\left(I + \varepsilon H^{(d)}\right)^{-1} - H^{(d)}M_t^{-1}M_a\right) \\ &\quad \left.-\varepsilon^4\left(H^{(d)}\right)^2 M_t^{-1}M_a + \varepsilon^5\left(H^{(d)}\right)^3\left(I + \varepsilon H^{(d)}\right)^{-1}M_t^{-1}M_a\right]\boldsymbol{\varphi}.\end{aligned}$$

Equation (54) consists of terms up to $\mathcal{O}(\varepsilon^2)$. Collecting HO terms yields the remainder term, $R_\varepsilon^{(d)}$, given by

$$\begin{aligned}R_\varepsilon^{(d)} &= \frac{1}{4\pi}\varepsilon^3\left(H^{(d)}\right)^3\left(I + \varepsilon H^{(d)}\right)^{-1}\left(I - \varepsilon^2 M_t^{-1}M_a\right)\boldsymbol{\varphi} \\ &\quad + \varepsilon^3 H^{(d)}\left(I - \varepsilon H^{(d)}\right)M_t^{-1}M_a\boldsymbol{\varphi}.\end{aligned}$$

∎

The bound on $\|R_\varepsilon^{(d)}\|$ follows from the identity $\|(I + \varepsilon H^{(d)})^{-1}\| \le \frac{1}{1-\varepsilon\|H^{(d)}\|}$.

Before stating Proposition 2, we set up preliminary notation. First, the linear system (13) can be written in the form

$$I - T_\varepsilon = I - H_\varepsilon P_0, \tag{56}$$

where $T_\varepsilon = H_\varepsilon P_0$ and H_ε is defined via

$$(H_\varepsilon\boldsymbol{\psi})^{(d)} = \left(I + \varepsilon M_t^{-1}\left(\mathbf{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right)\right)^{-1}\frac{1}{4\pi}\left(I - \varepsilon^2 M_t^{-1}M_a\right)\boldsymbol{\psi}^{(d)},$$

for $d = 1, \ldots, N_\Omega$.

Notice that

$$(P_0\,(I-T_\varepsilon)\,P_0\boldsymbol{\psi})^{(d)} = (I-S_\varepsilon)\,(P_0\boldsymbol{\psi})^{(d)}\,, \tag{57}$$

where S_ε is defined in (15). Also, from Lemma 13,

$$I - T_\varepsilon = Q_0 + \varepsilon H_0 P_0 + \varepsilon^2 H_1 P_0 + \mathcal{O}\left(\varepsilon^3\right), \tag{58}$$

where H_i denotes block-diagonal in d matrices, for $i = 1, \ldots, N_\Omega$ as in (54); in particular,

$$(H_0)_{d,d} = \frac{1}{4\pi} H^{(d)} = \frac{1}{4\pi} M_t^{-1}\left(\boldsymbol{\Omega}_d \cdot \mathbf{G} + F^{(d)}\right), \quad (H_1)_{d,d} = \frac{1}{4\pi}\left(\left(H^{(d)}\right)^2 - M_t^{-1} M_a\right).$$

Finally, define $F_0 = \frac{1}{4\pi}\sum_d w_d F^{(d)}$. Then using $\sum_d w_d \boldsymbol{\Omega}_d = \mathbf{0}$, it follows that

$$(P_0 H_0 P_0 \boldsymbol{\psi})^{(d)} = M_t^{-1} F_0 \boldsymbol{\varphi}, \quad d = 1, \ldots, N_\Omega. \tag{59}$$

We now prove Proposition 2.

Proof. First, choose some unit norm vector $\boldsymbol{\psi}$ for which $Q_0\boldsymbol{\psi} = \boldsymbol{\psi}$. Then, using (56),

$$\|I - T_\varepsilon\|_W \geq \|(I - T_\varepsilon)\, Q_0\boldsymbol{\psi}\|_W = \|Q_0\boldsymbol{\psi}\|_W = 1.$$

For the inverse, $(I - T_\varepsilon)\mathbf{x} = \mathbf{y}$ can be decomposed based on P_0 and Q_0 via $(I - T_\varepsilon)(P_0\mathbf{x} + Q_0\mathbf{x}) = (P_0\mathbf{y} + Q_0\mathbf{y})$. Multiplying on the left by the full-column-rank operator $(P_0; Q_0)$ yields the equivalent linear system

$$\begin{pmatrix} P_0 \\ Q_0 \end{pmatrix} (I - T_\varepsilon) \begin{pmatrix} P_0 & Q_0 \end{pmatrix} \begin{pmatrix} P_0\mathbf{x} \\ Q_0\mathbf{x} \end{pmatrix} = \begin{pmatrix} P_0 \\ Q_0 \end{pmatrix} (P_0\mathbf{y} + Q_0\mathbf{y}). \tag{60}$$

Denote $\mathbf{x}_P = P_0\mathbf{x}$ and $\mathbf{x}_Q = Q_0\mathbf{x}$, and likewise for $\mathbf{y}$. Then (60) yields a 2×2 set of equations, which, noting the expansion from Lemma 13 and (58) and the orthogonality of P_0 and Q_0, reduces to

$$\begin{pmatrix} P_0(I - T_\varepsilon)P_0 & \mathbf{0} \\ Q_0(I - T_\varepsilon)P_0 & Q_0 \end{pmatrix} \begin{pmatrix} \mathbf{x}_p \\ \mathbf{x}_Q \end{pmatrix} = \begin{pmatrix} \mathbf{y}_P \\ \mathbf{y}_Q \end{pmatrix}. \tag{61}$$

Here, $\mathbf{x}_P$ is fully determined by inverting $P_0(I - T_\varepsilon)P_0$ on the range of P_0. This is equivalent to inverting $I - S_\varepsilon$ (57), which is assumed to be full rank. Now, choose some vector $\hat{\mathbf{x}}$ for which $P_0\hat{\mathbf{x}} = \hat{\mathbf{x}}$, where each direction block $\hat{\mathbf{x}}_d$ corresponds to a continuous function. From (19), we have that $F_0 P_0 \hat{\mathbf{x}} = \mathbf{0}$ and from (59) $P_0 H^{(d)} P_0 \hat{\mathbf{x}} = \mathbf{0}$. From (58), this yields

$$(P_0\,(I - T_\varepsilon)\,P_0\hat{\mathbf{x}}) = \mathcal{O}\left(\varepsilon^2\right) P_0\hat{\mathbf{y}}. \tag{62}$$

Recall by orthogonality, $\|\mathbf{y}\|_W = \|\mathbf{y}_P\|_W + \|\mathbf{y}_Q\|_W$. Now define a vector $\tilde{\mathbf{y}}$ such that $\tilde{\mathbf{y}}_P = P_0\hat{\mathbf{y}}$ from (62) and $\tilde{\mathbf{y}}_Q = \mathbf{0}$, and let $\tilde{\mathbf{x}} = (I - T_\varepsilon)^{-1}\tilde{\mathbf{y}}$. Then, in the notation of (61),

$$\begin{aligned} \|(I - T_\varepsilon)^{-1}\|_W &= \sup_{\|\mathbf{y}\|_\mathbf{w}=1} \|(I - T_\varepsilon)^{-1}\mathbf{y}\|_W = \sup_{\|\mathbf{y}\|_\mathbf{w}=1} \|\mathbf{x}_P\|_W + \|\mathbf{x}_Q\|_W \\ &\geq \|\tilde{\mathbf{x}}_P\|_W + \|\tilde{\mathbf{x}}_Q\|_W = \mathcal{O}(\varepsilon^{-2})\|\hat{\mathbf{x}}_P\|_W + \|\tilde{\mathbf{x}}_Q\|_W \\ &\geq \mathcal{O}(\varepsilon^{-2})\|\hat{\mathbf{x}}_P\|_W = \mathcal{O}(\varepsilon^{-2}). \end{aligned}$$

To prove (20), note that

$$\begin{aligned}\tilde{A}_\varepsilon &= (E_\varepsilon P_0 + Q_0)(I - T_\varepsilon) \\ &= E_\varepsilon P_0 (I - T_\varepsilon) P_0 + Q_0 (I - T_\varepsilon) \\ &= P_0 + \mathcal{O}(\varepsilon) + Q_0 - Q_0 T_\varepsilon \\ &= I + \mathcal{O}(\varepsilon) - Q_0 T_\varepsilon \\ &= I + \mathcal{O}(\varepsilon).\end{aligned}$$

In the second equality, we used the assumption that $E_\varepsilon P_0(I - T_\varepsilon)P_0 = P_0 + \mathcal{O}(\varepsilon)$ and the identity $P_0(I - T_\varepsilon) = P_0(I - T_\varepsilon)P_0$. In the last equality, we used that $Q_0 T_\varepsilon = \mathcal{O}(\varepsilon)$, which follows from (58). ∎

Remark 14. Letting $h_{\mathbf{x}}$ denote the characteristic mesh spacing, the assumption $\varepsilon \|H^{(d)}\| < 1$ in Lemma 13 holds if $\varepsilon \lesssim \sigma_t h_{\mathbf{x}}$, which corresponds to the optically thick limit.

Acknowledgements. We would like to thank Jim Warsa for pointing out the equivalence between the nonsymmetric interior penalty method explored in this paper and his consistent P1 diffusion discretization, as well as helping us understand many of the nuances of S_N transport preconditioning. We would also like to thank Jim Morel for his many insightful comments and for suggesting the need to perform additional transport sweeps when there are mesh cycles.

REFERENCES

[1] M. L. Adams and E. W. Larsen, *Fast iterative methods for discrete-ordinates particle transport calculations*, Prog. Nuclear Energy, 40 (2002), pp. 3–159.

[2] M. L. Adams and W. R. Martin, *Diffusion synthetic acceleration of discontinuous finite element transport iterations*, Nucl. Sci. Engrg., 111 (1992), pp. 145–167.

[3] R. E. Alcouffe, *Diffusion synthetic acceleration methods for the diamond-differenced discrete-ordinates equations*, Nucl. Sci. Engrg., 64 (1977), pp. 344–355.

[4] R. W. Anderson, V. A. Dobrev, T. V. Kolev, R. N. Rieben, and V. Z. Tomov, *High-order multi-material ALE hydrodynamics*, SIAM J. Sci. Comput., 40 (2018), pp. B32–B58.

[5] P. F. Antonietti, M. Sarti, M. Verani, and L. T. Zikatanov, *A uniform additive schwarz preconditioner for high-order discontinuous galerkin approximations of elliptic problems*, SIAM J. Sci. Comput., 70 (2017), pp. 608–630.

[6] P. F. Antonietti, M. Sarti, M. Verani, and L. T. Zikatanov, *A uniform additive Schwarz preconditioner for high-order discontinuous Galerkin approximations of elliptic problems*, J. Sci. Comput., 70 (2017), pp. 608–630.

[7] D. Arnold, F. Brezzi, B. Cockburn, and L. Marini, *Unified analysis of discontinuous galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[8] P. Bastian, M. Blatt, and R. Scheichl, *Algebraic multigrid for discontinuous Galerkin discretizations of heterogeneous elliptic problems*, Numer. Linear Algebra Appl., 19 (2012), pp. 367–388.

[9] P. Bastian, M. Blatt, and R. Scheichl, *Algebraic multigrid for discontinuous Galerkin discretizations of heterogeneous elliptic problems*, Numer. Linear Algebra Appl., 19 (2012), pp. 367–388.

[10] V. Dobrev, Tz. Kolev, and R. Rieben, *High-order curvilinear finite element methods for Lagrangian hydrodynamics*, SIAM J. Sci. Comput., 34 (2012), pp. 606–641.

[11] E. M. Gelbard and L. A. Hageman, *The synthetic method as applied to the sn equations*, Nucl. Sci. Engrg., 37 (1969), pp. 288–298.

[12] J. Guermond and G. Kanschat, *Asymptotic analysis of upwind discontinuous Galerkin approximation of the radiative transport equation in the diffusive limit*, SIAM J. Numer. Anal., 48 (2010), pp. 53–78.

[13] T. S. Haut, P. G. Maginot, V. Z. Tomov, B. S. Southworth, T. A. Brunner, and T. S. Bailey, *An efficient sweep-based solver for the SN equations on high-order meshes*, Nucl. Sci. Engrg., 193 (2019), pp. 746–759.

[14] H. J. Kopp, *Synthetic method solution of the transport equation*, Nucl. Sci. Engrg., 17 (1963), pp. 65–74.
[15] M. Kucharik, R. V. Garimella, S. P. Schofield, and M. J. Shashkov, *A comparative studdy of interface reconstruction methods for multi-material ALE simulations*, J. Comput. Phys., 229 (2010), pp. 2432–2452.
[16] E. Larsen and D. R. McCoy, *Unconditionally stable diffusion-synthetic acceleration methods for the slab geometry discrete ordinates equations. Part* II*: Numerical results*, 82 (1982), pp. 64–70.
[17] E. W. Larsen, *Unconditionally stable diffusion-synthetic acceleration methods for the slab geometry discrete ordinates equations. Part* I*: Theory*, Nucl. Sci. Engrg., 82 (1982), pp. 47–63.
[18] E. W. Larsen, *Diffusion-synthetic acceleration methods for discrete-ordinates problems*, Trans. Theory Stat. Phys., 13 (1984), pp. 107–126.
[19] E. W. Larsen and J. B. Keller, *Asymptotic solution of neutron transport problems for small mean free paths*, J. Math. Phys., 15 (1974), pp. 75–81.
[20] E. W. Larsen, P. Nowak, and H. L. Hanshaw, *Lawrence Livermore National Laboratory*, United States. Department of Energy. Office of Scientific, and Technical Information. Stretched and Filtered Transport Synthetic Acceleration of Sn Problems: Part 1: Homogeneous Media. United States. Department of Energy, 2003.
[21] V. I. Lebedev, *Convergence of the kp-method for some neutron transfer problems*, USSR Comput. Math. Math. Phys., 9 (1969), pp. 309–323.
[22] T. A. Manteuffel, S. Münzenmaier, J. W. Ruge, and B. S. Southworth, *Nonsymmetric reduction-based algebraic multigrid*, SIAM J. Sci. Comput., in process.
[23] T. A. Manteuffel, J. W. Ruge, and B. S. Southworth, *Nonsymmetric algebraic multigrid based on local approximate ideal restriction (ℓAIR)*, SIAM J. Sci. Comput., 40 (2018), pp. A4105–A4130.
[24] G. Marchuk and V. I. Lebedev, *Numerical Methods in the Theory of Neutron Transport*, United States: Harwood Academic Pub, first edition, 1986.
[25] MFEM: Modular parallel finite element methods library, 2019. http://mfem.org.
[26] L. N. Olson and J. B. Schroder, *Smoothed aggregation multigrid solvers for high-order discontinuous Galerkin methods for elliptic problems*, J. Comput. Phys., 230 (2011), pp. 6959–6976.
[27] B. O'Malley, J. Kópházi, R. P. Smedley-Stevenson, and M. D. Eaton, *Hybrid multilevel solvers for discontinuous galerkin finite element discrete ordinate diffusion synthetic acceleration of radiation transport algorithms*, Ann. Nuclear Energy, 102 (2017), pp. 134–147.
[28] B. O'Malley, J. Kópházi, R. P. Smedley-Stevenson, and M. D. Eaton, *Hybrid Multilevel solvers for discontinuous Galerkin finite element discrete ordinate diffusion synthetic acceleration of radiation transport algorithms*. Ann. Nuclear Energy, 102 (2017), pp. 134–147.
[29] W. Pazner, *Efficient low-order refined preconditioners for high-order matrix-free continuous and discontinuous galerkin methods*, preprint, arXiv:1908.07071, 2019.
[30] F. Prill, M. L. Medvidova, and R. Hartmann, *Smoothed aggregation multigrid for the discontinuous Galerkin method*, SIAM J. Sci. Comput., 31 (2009), pp. 3503–3528.
[31] G. L. Ramone, M. L. Adams, and P. F. Nowak, *A transport synthetic acceleration method for transport iterations*, Nucl. Sci. Engrg., 125 (1997), pp. 257–283.
[32] W. H. Reed, *The effectiveness of acceleration techniques for iterative methods in transport theory*, Nucl. Sci. Engrg., 45 (1971), pp. 245–254.
[33] C. M. Siefert, R. S. Tuminaro, A. Gerstenberger, G. Scovazzi, and S. S. Collis, *Algebraic multigrid techniques for discontinuous Galerkin methods with varying polynomial order*, Comput. Geosci., 18 (2014), pp. 597–612.
[34] Y. Wang and J. C. Ragusa, *Diffusion synthetic acceleration for high-order discontinuous finite element sn transport schemes and application to locally refined unstructured meshes*, Nucl. Sci. Engrg., 166 (2010), pp. 145–166.
[35] T. A. Wareing, J. M. McGhee, J. E. Morel, and S. D. Pautz, *Discontinuous finite element sn methods on three-dimensional unstructured grids*, Nucl. Sci. Engrg., 138 (2001), pp. 256–268.
[36] J. Warsa, T. Wareing, and J. Morel, *Solution of the discontinuous p*1 *equations in two-dimensional cartesian geometry with two-level preconditioning*, SIAM J. Sci. Comput., 24 (2003), pp. 2093–2124.
[37] J. S. Warsa, M. Benzi, T. A. Wareing, and J. E. Morel, *Numerical Mathematics and Advanced Applications*, pp. 967–977, 2003.

[38] J. S. Warsa, T. A. Wareing, and J. E. Morel, *Fully consistent diffusion synthetic acceleration of linear discontinuous sn transport discretizations on unstructured tetrahedral meshes*, Nucl. Sci. Engrg., 141 (2002), pp. 236–251.

[39] M. R. Zika and M. L. Adams, *Transport synthetic acceleration for long-characteristics assembly-level transport problems*, Nucl. Sci. Engrg., 134 (2000), pp. 135–158.