SIAM J. SCI. COMPUT.
Vol. 41, No. 3, pp. A1500–A1526

A POSITIVE ASYMPTOTIC-PRESERVING SCHEME FOR LINEAR KINETIC TRANSPORT EQUATIONS*

M. PAUL LAIU†, MARTIN FRANK‡, AND CORY D. HAUCK†

Abstract. We present a positive- and asymptotic-preserving numerical scheme for solving linear kinetic transport equations that relax to a diffusive equation in the limit of infinite scattering. The proposed scheme is developed using a standard spectral angular discretization and a classical micro-macro decomposition. The three main ingredients are a semi-implicit temporal discretization, a dedicated finite difference spatial discretization, and realizability limiters in the angular discretization. Under mild assumptions, the scheme becomes a consistent numerical discretization for the limiting diffusion equation when the scattering cross-section tends to infinity. The scheme also preserves positivity of the particle concentration on the space-time mesh and therefore fixes a common defect of spectral angular discretizations. The scheme is tested on the well-known line source benchmark problem with the usual uniform material medium as well as a medium composed of different materials that are arranged in a checkerboard pattern. We also tested the scheme on a Riemann problem with a nonuniform material medium. The observed order of space-time accuracy of the proposed scheme is reported.

Key words. kinetic transport equations, diffusion limit, positive-preserving schemes, asymptotic-preserving schemes, finite difference methods

AMS subject classifications. 35B09, 35L40, 41A60, 65M06, 65M70, 82C70, 82D75

DOI. 10.1137/18M1196297

1. Introduction.

Kinetic transport equations are widely used to model particle systems in many applications, including thermal radiative transfer [63, 67], rarefied gas dynamics [10], plasmas [33], and neutron transport [15, 56]. These equations track the temporal evolution of a particle distribution function in a position-velocity phase space.

For kinetic equations that model propagation through a background medium, the kinetic distribution function is often approximated by the solution of a much simpler diffusion equation. Such an approximation is accurate when the dynamics of the particle system are dominated by scattering interactions with the medium. However, many problems of interest are "multiscale" in the sense that scattering and other material cross-sections may vary in space by several orders of magnitude. In regions of moderate scattering, the diffusion approximation may not be sufficiently accurate, while in

*Submitted to the journal's Methods and Algorithms for Scientific Computing section June 25, 2018; accepted for publication (in revised form) March 7, 2019; published electronically May 9, 2019. This manuscript has been authored, in part, by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for the United States Government purposes.

http://www.siam.org/journals/sisc/41-3/M119629.html

Funding: The first author was supported by the U.S. Department of Energy, under the SCGSR program administered by the Oak Ridge Institute for Science and Education under contract DE-AC05-06OR23100. The third author's research was sponsored by the Office of Advanced Scientific Computing Research and performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725.

†Computational Mathematics Group, Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 (laiump@ornl.gov, hauckc@ornl.gov).

‡Karlsruhe Institute of Technology, Steinbuch Centre for Computing, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany (martin.frank@kit.edu).

strongly scattering regions, classical numerical schemes for kinetic equations must resolve collisional length scales, making them computationally prohibitive. In addition, explicit time integrators for kinetic equations in scattering dominated regimes require very small time steps in order to maintain accuracy and stability, while implicit time integrators can be exceptionally stiff. For these reasons, it is desirable to solve multiscale problems with numerical schemes that are consistent in both the kinetic and diffusive regions and uniformly stable under a reasonable time-step restriction. Such schemes are often referred to as "asymptotic-preserving" (AP) schemes [38].

AP schemes for kinetic equations with diffusion limits were first considered in the context of steady-state neutron transport [50, 51]. Since then, a variety of approaches have been taken, including discontinuous Galerkin methods [1, 25, 32, 50, 61], methods based on even-odd parity [41, 42, 64, 69], micro-macro decompositions [53, 57], numerical fluxes that depend on the scattering cross-section [8, 40], temporal regularization [30, 38], well-balanced schemes [20, 21, 22, 23], and unified gas-kinetic schemes [62, 73]. Many of these approaches are related or overlapping, and despite any differences, they all seek to address two fundamental issues: first, that the numerical dissipation induced by the discretization of the hyperbolic advection operator in the kinetic equation must be controlled in the diffusion limit; second, that for time-dependent problems, stiffness must be overcome either with semi-implicit time integrators [4] or with fully implicit time integrators that leverage acceleration and/or preconditioning techniques [2, 52].

Deterministic numerical simulations of kinetic equations require discretization in space, velocity, and time. Spherical harmonic (P_N) methods [9, 56, 67] discretize the angular component of the velocity using a polynomial approximation with coefficients that are functions of space and time. The kinetic equations are then converted using the standard Galerkin approach into a system of reduced equations for these coefficients. As with other spectral methods, approximate solutions generated in this way converge spectrally to the solution of the kinetic equation when the latter is sufficiently smooth. However, when the solution to the kinetic equation is discontinuous, the P_N method produces oscillatory solutions which may cause the approximation of the particle concentration (defined as the integral of the kinetic distribution over the velocity variable) to become negative.

In [60], filtering techniques [24, 27] for mitigating the Gibbs phenomena in spectral approximations were proposed as a way to reduce oscillations in the solution of the P_N equations. Later in [68], this idea was used to derive a system of modified P_N equations, referred to as the filtered P_N (FP_N) equations. While filtering significantly reduces oscillations in the profile of the particle concentration, negative values are still possible. Thus in [46], several positive-preserving schemes were proposed to augment the filtering strategy. These schemes combine a second-order, explicit, finite-volume discretization in space and time [3, 19] with limiters that force the spectral approximation in angle to be nonnegative on a finite set of quadrature points in the angular domain. The schemes do preserve positivity of the particular concentration. However, the limiters may reduce accuracy or be computationally expensive, and the finite-volume discretization is not AP.

In this paper, we propose a positive-preserving AP scheme for solving the FP_N equations. The proposed scheme follows the approach proposed in [53] and analyzed in [57], where a one-dimensional kinetic equation was solved using a classical micro-macro decomposition [58] of the kinetic distribution. Here we use the micro-macro decomposition to formulate the FP_N equations as a coupled system for the expansion coefficients that correspond to the macro and micro parts of the kinetic distribution.

The numerical scheme we use to solve the micro-macro system involves three main ingredients: a semi-implicit temporal discretization, a dedicated finite difference spatial discretization, and realizability limiters in the angular discretization. The space-time discretization is designed so that the realizability limiters in the angular discretization are physically reasonable but also less strict than the pointwise limiters used in [46]. In designing the spatial discretization, we focus on a simplified geometry that allows for a formulation in two space dimensions. However, we also discuss how to extend the proposed numerical scheme to the full three-dimensional setting.

The remainder of the paper is organized as follows. In section 2, we introduce the linear kinetic equation, the FP_N equations, their diffusion limits, and the derivation of the associated micro-macro systems. In section 3, we present the space-time discretization for the micro-macro FP_N system and show that under mild assumptions on the initial condition and time step, the fully discretized scheme gives a consistent explicit numerical scheme for the diffusion limit when the scattering cross-section tends to infinity. In section 4, we give sufficient conditions for preserving positivity of the particle concentration and we detail the approach, including the realizability limiters and time-step restriction needed to enforce these conditions. We test the proposed scheme on three benchmark problems in both kinetic and diffusive regimes and report the results in section 5. Conclusions and discussion are given in section 6.

2. Linear kinetic equations, FP_N equations, and the diffusion limits.

2.1. Linear kinetic equation and its diffusion limit. We consider the one-speed linear kinetic transport equation [9, 14, 15, 56, 66]

$$\partial_t f + \Omega \cdot \nabla_r f = \sigma_{\mathrm{s}} \overline{f} - \sigma_{\mathrm{t}} f \tag{2.1}$$

for the kinetic distribution function $f = f(r, \Omega, t)$. Here $r = (x, y, z) \in \mathbb{R}^3$ is the position; $\Omega = (\Omega_x, \Omega_y, \Omega_z) \in \mathbb{S}^2$ is the angle; $\sigma_{\mathrm{s}}(r) \geq \sigma_{\mathrm{s}}^{\min} > 0$, $\sigma_{\mathrm{a}}(r) \geq 0$, and $\sigma_{\mathrm{t}} = \sigma_{\mathrm{s}} + \sigma_{\mathrm{a}}$ are, respectively, the scattering, absorption, and total cross-sections; and $\overline{f}(r, t) = (4\pi)^{-1}\langle f \rangle$, where $\langle \cdot \rangle$ denotes integration over $\mathbb{S}^2$ with respect to Ω, is the angular average of f. We denote the particle concentration associated to f by $\rho = \langle f \rangle = 4\pi \overline{f}$. With appropriate initial and boundary conditions, (2.1) is known to have a unique solution [14].

Given $\epsilon > 0$, letting $\sigma_{\mathrm{s}} \to \epsilon^{-1}\sigma_{\mathrm{s}}$, $\sigma_{\mathrm{a}} \to \epsilon\sigma_{\mathrm{a}}$, and $t \to \epsilon^{-1}t$ in (2.1) leads to the scaled equation

$$\epsilon\partial_t f + \Omega \cdot \nabla_r f = \frac{\sigma_{\mathrm{s}}}{\epsilon}\left(\overline{f} - f\right) - \epsilon\sigma_{\mathrm{a}} f \,. \tag{2.2}$$

It is well known [29, 49] that when $\epsilon \ll 1$, the kinetic distribution f in (2.2) is given by $f = \overline{f} + O(\epsilon)$. Meanwhile, the particle concentration $\rho = 4\pi\overline{f}$ is governed approximately by a diffusion equation

$$\partial_t \rho - \nabla_r \cdot (D \nabla_r \rho) + \sigma_{\mathrm{a}} \rho = O(\epsilon) \,, \tag{2.3}$$

where the matrix of diffusion coefficients D is given by

$$D = \frac{1}{4\pi\sigma_{\mathrm{s}}} \operatorname{diag}\left(\langle \Omega_x^2 \rangle, \langle \Omega_y^2 \rangle, \langle \Omega_z^2 \rangle\right) = \frac{1}{3\sigma_{\mathrm{s}}} I_{3\times 3} \,. \tag{2.4}$$

When $\epsilon \to 0$, the right-hand side of (2.3) vanishes, and the resulting equation (2.3) is referred to as the diffusion limit for (2.2).

2.2. P_N and FP_N equations and their diffusion limit. The P_N method [9, 56] approximates the kinetic distribution by a polynomial expansion in Ω with coefficients that are functions of space and time. When the angular space is the unit sphere, spherical harmonics are commonly used as the basis for spectral approximations. Specifically, let $\mathbb{P}_N(\mathbb{S}^2) \subset L^2(\mathbb{S}^2)$ be the vector space of polynomials in Ω with degree at most N, and let $\mathbf{m}\colon \mathbb{S}^2 \to \mathbb{R}^n$, where $n = \dim(\mathbb{P}_N(\mathbb{S}^2))$, be a vector-valued function that takes the form $\mathbf{m} := [m_0^0, m_1^{-1}, m_1^0, m_1^1, \dots]^T$, where m_ℓ^k denotes the real-valued spherical harmonic of degree ℓ and order k, normalized such that $\langle m_\ell^k m_{\ell'}^{k'} \rangle = \delta_{\ell,\ell'} \cdot \delta_{k,k'}$ with $\delta_{\ell,\ell'}$ the Kronecker delta function. For example, the first few components of $\mathbf{m}$ are

$$m_0^0 = \sqrt{\frac{1}{4\pi}}\,, \quad m_1^{-1} = \sqrt{\frac{3}{4\pi}}\Omega_y\,, \quad m_1^0 = \sqrt{\frac{3}{4\pi}}\Omega_z\,, \quad m_1^1 = \sqrt{\frac{3}{4\pi}}\Omega_x\,. \tag{2.5}$$

The components of $\mathbf{m}$ form an orthonormal basis of $\mathbb{P}_N(\mathbb{S}^2)$, and the scaled P_N equations corresponding to (2.2) are given by

$$\epsilon\partial_t \mathbf{u}_{\mathrm{P}_N} + \left\langle \mathbf{m}\mathbf{m}^T\Omega \right\rangle \cdot \nabla_r \mathbf{u}_{\mathrm{P}_N} = -\frac{\sigma_\mathrm{s}}{\epsilon} R\mathbf{u}_{\mathrm{P}_N} - \epsilon\sigma_\mathrm{a}\mathbf{u}_{\mathrm{P}_N}\,, \tag{2.6}$$

where $R = \mathrm{diag}([0, 1, \dots, 1])$, and

$$\left\langle \mathbf{m}\mathbf{m}^T\Omega \right\rangle \cdot \nabla_r := \left\langle \mathbf{m}\mathbf{m}^T\Omega_x \right\rangle \partial_x + \left\langle \mathbf{m}\mathbf{m}^T\Omega_y \right\rangle \partial_y + \left\langle \mathbf{m}\mathbf{m}^T\Omega_z \right\rangle \partial_z\,. \tag{2.7}$$

The solution $\mathbf{u}_{\mathrm{P}_N}\colon \mathbb{R}^3 \times \mathbb{R}^+ \to \mathbb{R}^n$ to (2.6) is an approximation to the spectral expansion coefficients of f, and the initial condition is given by $\mathbf{u}_{\mathrm{P}_N}(r, 0) := \langle \mathbf{m} f(r, \Omega, 0)\rangle$. The P_N equations form a symmetric, linear hyperbolic system of PDEs.

When the solution to (2.2) is not smooth, the P_N method produces oscillatory solutions [6, 19]. To reduce oscillations, a filtering term was introduced into (2.6) in [68], resulting in the following system of modified equations:

$$\epsilon\partial_t \mathbf{u}_{\mathrm{FP}_N} + \left\langle \mathbf{m}\mathbf{m}^T\Omega \right\rangle \cdot \nabla_r \mathbf{u}_{\mathrm{FP}_N} = -\frac{\sigma_\mathrm{s}}{\epsilon} R\mathbf{u}_{\mathrm{FP}_N} - \epsilon\sigma_\mathrm{a}\mathbf{u}_{\mathrm{FP}_N} - \epsilon\sigma_\mathrm{f} F\mathbf{u}_{\mathrm{FP}_N}\,, \tag{2.8}$$

where $\sigma_\mathrm{f} > 0$ is a tunable parameter which determines the strength of the filter (see [17, 60, 68]), the filtering matrix $F \in \mathbb{R}^{n\times n}$ is a diagonal matrix with elements $F_{(\ell,k),(\ell,k)} = -\ln\left(\kappa\left(\frac{\ell}{N+1}\right)\right)$, and $\kappa\colon \mathbb{R}^+ \to [0, 1]$ is a filter function with $\kappa(0) = 1$. (See, for example, [17] for a detailed definition of κ.) The solution $\mathbf{u}_{\mathrm{FP}_N}\colon \mathbb{R}^3 \times \mathbb{R}^+ \to \mathbb{R}^n$ to (2.8) is also an approximation to the spectral expansion coefficients of f, and the initial condition is given by $\mathbf{u}_{\mathrm{FP}_N}(r, 0) = \langle \mathbf{m} f(r, \Omega, 0)\rangle$. The modified equations (2.8), referred to as the filtered P_N (FP_N) equations [17, 68], also form a linear hyperbolic system. Analogous to (2.3), the diffusion limit of (2.8) is given by

$$\partial_t \bar{u}_{\mathrm{FP}_N} - \nabla_r \cdot (D\nabla_r \bar{u}_{\mathrm{FP}_N}) + \sigma_\mathrm{a}\bar{u}_{\mathrm{FP}_N} = 0\,, \tag{2.9}$$

where D is as defined in (2.4) and $\bar{u}_{\mathrm{FP}_N}$ denotes the first component of $\mathbf{u}_{\mathrm{FP}_N}$. Note that the filtering term $\sigma_\mathrm{f} F\mathbf{u}_{\mathrm{FP}_N}$ in (2.8) is scaled such that it vanishes as $\epsilon \to 0$, since the solution is generally not oscillatory in the diffusion limit.

2.3. The micro-macro decomposition. For the scaled kinetic equation (2.2), the kinetic distribution f can be decomposed into $f(r, \Omega, t) = \overline{f}(r, t) + \epsilon\widetilde{f}(r, \Omega, t)$ (see, e.g., [53] for details), where the macro component $\overline{f}$ is constant with respect to Ω and

the micro component $\widetilde{f}$ satisfies $\langle \widetilde{f} \rangle = 0$. The governing equations for $\overline{f}$ and $\widetilde{f}$ are

$$\partial_t \overline{f} + \frac{1}{4\pi} \langle \Omega \cdot \nabla_r \widetilde{f} \rangle + \sigma_{\mathrm{a}} \overline{f} = 0 \,, \tag{2.10a}$$

$$\partial_t \widetilde{f} + \frac{1}{\epsilon} \Omega \cdot \nabla_r \widetilde{f} - \frac{1}{4\pi\epsilon} \langle \Omega \cdot \nabla_r \widetilde{f} \rangle + \sigma_{\mathrm{a}} \widetilde{f} = -\frac{\sigma_{\mathrm{s}}}{\epsilon^2} \widetilde{f} - \frac{1}{\epsilon^2} \Omega \cdot \nabla_r \overline{f} \,. \tag{2.10b}$$

We apply a micro-macro decomposition to the FP_N equations (2.8). Specifically, we decompose $\mathbf{u}_{\mathrm{FP}_N} \in \mathbb{R}^n$ into the macro expansion coefficients $\overline{u} \in \mathbb{R}$ and micro expansion coefficients $\widetilde{\mathbf{u}} \in \mathbb{R}^{\widetilde{n}}$, where $\widetilde{n} := n - 1$ and $\mathbf{u}_{\mathrm{FP}_N} = [\overline{u}, \epsilon \widetilde{\mathbf{u}}^T]^T$. To simplify the notation, we drop subscripts and set $\mathbf{u} = \mathbf{u}_{\mathrm{FP}_N}$ for the remainder of this paper. We also let $\overline{m}$ denote $m_0^0 = (4\pi)^{-\frac{1}{2}}$ and let $\widetilde{\mathbf{m}} \colon \mathbb{S}^2 \to \mathbb{R}^{\widetilde{n}}$ denote the remaining components of $\mathbf{m}$, i.e., $\mathbf{m} =: [\overline{m}, \widetilde{\mathbf{m}}^T]^T$. Then, similar to (2.10a)–(2.10b), $\overline{u}$ and $\widetilde{\mathbf{u}}$ are governed by the *micro-macro system*

$$\partial_t \overline{u} + \langle \overline{m} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \widetilde{\mathbf{u}} + \sigma_{\mathrm{a}} \overline{u} = 0 \,, \tag{2.11a}$$

$$\partial_t \widetilde{\mathbf{u}} + \frac{1}{\epsilon} \langle \widetilde{\mathbf{m}} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \widetilde{\mathbf{u}} + \sigma_{\mathrm{a}} \widetilde{\mathbf{u}} = -\frac{\sigma_{\mathrm{s}}}{\epsilon^2} \widetilde{\mathbf{u}} - \sigma_{\mathrm{f}} \widetilde{F} \widetilde{\mathbf{u}} - \frac{1}{\epsilon^2} \langle \widetilde{\mathbf{m}} \overline{m} \Omega \rangle \cdot \nabla_r \overline{u} \,, \tag{2.11b}$$

where $\widetilde{F} \in \mathbb{R}^{\widetilde{n} \times \widetilde{n}}$ is formed by removing the first column and first row of F.

3. Space-time discretization. We present a semi-implicit time discretization for the micro-macro system (2.11a)–(2.11b) in section 3.1. In section 3.2, we introduce a reduced two-dimensional linear kinetic equation that is considered in the remainder of this paper. The finite difference spatial discretization for the reduced equation is given in section 3.3. We verify the AP property of the fully discretized micro-macro scheme in section 3.4. Extensions of the proposed spatial discretization to the original three-dimensional kinetic equation are discussed in section 4.4.

3.1. Time discretization. To discretize (2.11a) and (2.11b) in time, we assume a uniform time step Δt with time levels $t^n := n\Delta t$ and let $\overline{u}^n \approx \overline{u}(t^n, \cdot)$ and $\widetilde{\mathbf{u}}^n \approx \widetilde{\mathbf{u}}(t^n, \cdot)$ satisfy

$$\frac{\overline{u}^{n+1} - \overline{u}^n}{\Delta t} + \langle \overline{m} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \widetilde{\mathbf{u}}^{n+1} + \sigma_{\mathrm{a}} \overline{u}^{n+1} = 0 \,, \tag{3.1a}$$

$$\frac{\widetilde{\mathbf{u}}^{n+1} - \widetilde{\mathbf{u}}^n}{\Delta t} + \frac{1}{\epsilon} \langle \widetilde{\mathbf{m}} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \widetilde{\mathbf{u}}^n + \left(\sigma_{\mathrm{a}} + \frac{\sigma_{\mathrm{s}}}{\epsilon^2} + \sigma_{\mathrm{f}} \widetilde{F} \right) \widetilde{\mathbf{u}}^{n+1} = -\frac{1}{\epsilon^2} \langle \widetilde{\mathbf{m}} \overline{m} \Omega \rangle \cdot \nabla_r \overline{u}^n \,. \tag{3.1b}$$

We rewrite (3.1b) as

$$\widetilde{\mathbf{u}}^{n+1} = \widetilde{\Gamma} \widetilde{\mathbf{u}}^n - \frac{\Delta t}{\epsilon} \widetilde{\Gamma} \langle \widetilde{\mathbf{m}} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \widetilde{\mathbf{u}}^n - \frac{\Delta t}{\epsilon^2} \widetilde{\Gamma} \langle \widetilde{\mathbf{m}} \overline{m} \Omega \rangle \cdot \nabla_r \overline{u}^n \,, \tag{3.2}$$

where $\widetilde{\Gamma} \in \mathbb{R}^{\widetilde{n} \times \widetilde{n}}$ is a nonsingular, diagonal matrix with elements

$$\widetilde{\Gamma}_{(\ell,k),(\ell,k)} = \epsilon^2 \left(\epsilon^2 (1 + \sigma_{\mathrm{a}} \Delta t + \sigma_{\mathrm{f}} \Delta t \widetilde{F}_{(\ell,k),(\ell,k)}) + \sigma_{\mathrm{s}} \Delta t \right)^{-1} \in (0, 1) \,. \tag{3.3}$$

To obtain an explicit update for (3.1a), we replace the implicit term $\widetilde{\mathbf{u}}^{n+1}$ in (3.1a) with the right-hand side of (3.2). This gives

$$\begin{aligned} (1 + \sigma_{\mathrm{a}} \Delta t) \overline{u}^{n+1} = \overline{u}^n &- \Delta t \langle \overline{m} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r (\widetilde{\Gamma} \widetilde{\mathbf{u}}^n) \\ &+ \frac{\Delta t^2}{\epsilon} \langle \overline{m} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \left(\widetilde{\Gamma} \langle \widetilde{\mathbf{m}} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \widetilde{\mathbf{u}}^n \right) \\ &+ \frac{\Delta t^2}{\epsilon^2} \langle \overline{m} \widetilde{\mathbf{m}}^T \Omega \rangle \cdot \nabla_r \left(\widetilde{\Gamma} \langle \widetilde{\mathbf{m}} \overline{m} \Omega \rangle \cdot \nabla_r \overline{u}^n \right) \,. \end{aligned} \tag{3.4a}$$

Since $\sigma_{\rm a}$ and $\sigma_{\rm s}$ are functions of r, to avoid nonconservative products in (3.2), we perform a change of variables before and after solving (3.2). Specifically, we update $\widetilde{\mathbf{u}}^{n+1}$ by computing

$$\begin{aligned}
\widetilde{\mathbf{v}}^n &= \epsilon^2\widetilde{\Gamma}^{-1}\widetilde{\mathbf{u}}^n\,,\\
\widetilde{\mathbf{v}}^{n+1} &= \widetilde{\Gamma}\widetilde{\mathbf{v}}^n - \frac{\Delta t}{\epsilon}\langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega\rangle\cdot\nabla_r(\widetilde{\Gamma}\widetilde{\mathbf{v}}^n) - \Delta t\langle\widetilde{\mathbf{m}}\overline{m}\Omega\rangle\cdot\nabla_r\overline{u}^n\,, \qquad (3.4b)\\
\widetilde{\mathbf{u}}^{n+1} &= \epsilon^{-2}\widetilde{\Gamma}\widetilde{\mathbf{v}}^{n+1}\,.
\end{aligned}$$

3.2. Reduced linear kinetic equation and the micro-macro system. For the remainder of the paper, we restrict ourselves to a reduced two-dimensional linear kinetic equation that is valid when $\partial_z f = 0$:

$$\epsilon\partial_t f + \Omega_x\partial_x f + \Omega_y\partial_y f = \frac{\sigma_{\rm s}}{\epsilon}\left(\frac{1}{4\pi}\langle f\rangle - f\right) - \epsilon\sigma_{\rm a} f\,. \qquad (3.5)$$

In this setting, the micro-macro system (2.11a)–(2.11b) becomes

$$\begin{aligned}
&\partial_t\overline{u} + (\langle\overline{m}\widetilde{\mathbf{m}}^T\Omega_x\rangle\partial_x + \langle\overline{m}\widetilde{\mathbf{m}}^T\Omega_y\rangle\partial_y)\widetilde{\mathbf{u}} + \sigma_{\rm a}\overline{u} = 0\,, \qquad (3.6a)\\
&\partial_t\widetilde{\mathbf{u}} + \frac{1}{\epsilon}(\langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_x\rangle\partial_x + \langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_y\rangle\partial_y)\widetilde{\mathbf{u}} + \sigma_{\rm a}\widetilde{\mathbf{u}}\\
&\qquad = -\frac{\sigma_{\rm s}}{\epsilon^2}\widetilde{\mathbf{u}} - \sigma_{\rm f}\bar{F}\widetilde{\mathbf{u}} - \frac{1}{\epsilon^2}(\langle\widetilde{\mathbf{m}}\overline{m}\Omega_x\rangle\partial_x + \langle\widetilde{\mathbf{m}}\overline{m}\Omega_y\rangle\partial_y)\overline{u}\,. \qquad (3.6b)
\end{aligned}$$

Applying the time discretization in (3.4) to (3.6) leads to the reduced scheme

$$\begin{aligned}
(1+\sigma_{\rm a}\Delta t)\overline{u}^{n+1} &= \overline{u}^n - \Delta t(\widetilde{\mathbf{a}}_x^T\partial_x + \widetilde{\mathbf{a}}_y^T\partial_y)(\widetilde{\Gamma}\widetilde{\mathbf{u}}^n) + \frac{\Delta t^2}{\epsilon}\nabla_{(x,y)}\cdot\left(\widetilde{\mathbf{Q}}\nabla_{(x,y)}\widetilde{\mathbf{u}}^n\right)\\
&\quad + \frac{\Delta t^2}{\epsilon^2}\nabla_{(x,y)}\cdot\left(\overline{Q}\nabla_{(x,y)}\overline{u}^n\right)\,, \qquad (3.7a)\\
\widetilde{\mathbf{v}}^n &= \epsilon^2\widetilde{\Gamma}^{-1}\widetilde{\mathbf{u}}^n\,,\\
\widetilde{\mathbf{v}}^{n+1} &= \widetilde{\Gamma}\widetilde{\mathbf{v}}^n - \frac{\Delta t}{\epsilon}(\widetilde{A}_x\partial_x + \widetilde{A}_y\partial_y)(\widetilde{\Gamma}\widetilde{\mathbf{v}}^n) - \Delta t(\widetilde{\mathbf{a}}_x\partial_x + \widetilde{\mathbf{a}}_y\partial_y)\overline{u}^n\,, \qquad (3.7b)\\
\widetilde{\mathbf{u}}^{n+1} &= \epsilon^{-2}\widetilde{\Gamma}\widetilde{\mathbf{v}}^{n+1}\,,
\end{aligned}$$

where $\widetilde{\mathbf{a}}_x := \langle\overline{m}\widetilde{\mathbf{m}}\Omega_x\rangle \in \mathbb{R}^{\widetilde{n}}$, $\widetilde{\mathbf{a}}_y := \langle\overline{m}\widetilde{\mathbf{m}}\Omega_y\rangle \in \mathbb{R}^{\widetilde{n}}$, $\widetilde{A}_x := \langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_x\rangle \in \mathbb{R}^{\widetilde{n}\times\widetilde{n}}$, $\widetilde{A}_y := \langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_y\rangle \in \mathbb{R}^{\widetilde{n}\times\widetilde{n}}$, and $\widetilde{\mathbf{Q}}$ and $\overline{Q}$ are 2×2 block matrices:

$$\widetilde{\mathbf{Q}} := \begin{pmatrix}\widetilde{\mathbf{Q}}_{x^2} & \widetilde{\mathbf{Q}}_{xy}\\ \widetilde{\mathbf{Q}}_{yx} & \widetilde{\mathbf{Q}}_{y^2}\end{pmatrix} = \begin{pmatrix}\widetilde{\mathbf{a}}_x^T\widetilde{\Gamma}\widetilde{A}_x & \widetilde{\mathbf{a}}_x^T\widetilde{\Gamma}\widetilde{A}_y\\ \widetilde{\mathbf{a}}_y^T\widetilde{\Gamma}\widetilde{A}_x & \widetilde{\mathbf{a}}_y^T\widetilde{\Gamma}\widetilde{A}_y\end{pmatrix} = \begin{pmatrix}\widetilde{\gamma}\widetilde{\mathbf{a}}_{x^2}^T & \widetilde{\gamma}\widetilde{\mathbf{a}}_{xy}^T\\ \widetilde{\gamma}\widetilde{\mathbf{a}}_{yx}^T & \widetilde{\gamma}\widetilde{\mathbf{a}}_{y^2}^T\end{pmatrix} \qquad (3.8)$$

and

$$\overline{Q} := \begin{pmatrix}\overline{Q}_{x^2} & \overline{Q}_{xy}\\ \overline{Q}_{yx} & \overline{Q}_{y^2}\end{pmatrix} = \begin{pmatrix}\widetilde{\mathbf{a}}_x^T\widetilde{\Gamma}\widetilde{\mathbf{a}}_x & \widetilde{\mathbf{a}}_x^T\widetilde{\Gamma}\widetilde{\mathbf{a}}_y\\ \widetilde{\mathbf{a}}_y^T\widetilde{\Gamma}\widetilde{\mathbf{a}}_x & \widetilde{\mathbf{a}}_y^T\widetilde{\Gamma}\widetilde{\mathbf{a}}_y\end{pmatrix} = \begin{pmatrix}\widetilde{\gamma}\overline{a}_{x^2} & \widetilde{\gamma}\overline{a}_{xy}\\ \widetilde{\gamma}\overline{a}_{yx} & \widetilde{\gamma}\overline{a}_{y^2}\end{pmatrix} = \begin{pmatrix}\frac{1}{3}\widetilde{\gamma} & 0\\ 0 & \frac{1}{3}\widetilde{\gamma}\end{pmatrix}\,, \qquad (3.9)$$

with $\widetilde{\mathbf{a}}_{x^2} := \langle\overline{m}\widetilde{\mathbf{m}}\Omega_x^2\rangle \in \mathbb{R}^{\widetilde{n}}$, $\widetilde{\mathbf{a}}_{y^2} := \langle\overline{m}\widetilde{\mathbf{m}}\Omega_y^2\rangle \in \mathbb{R}^{\widetilde{n}}$, $\widetilde{\mathbf{a}}_{xy} = \widetilde{\mathbf{a}}_{yx} := \langle\overline{m}\widetilde{\mathbf{m}}\Omega_x\Omega_y\rangle \in \mathbb{R}^{\widetilde{n}}$, $\overline{a}_{x^2} := \langle\overline{m}^2\Omega_x^2\rangle = \frac{1}{3}$, $\overline{a}_{y^2} := \langle\overline{m}^2\Omega_y^2\rangle = \frac{1}{3}$, and $\overline{a}_{xy} = \overline{a}_{yx} := \langle\overline{m}^2\Omega_x\Omega_y\rangle = 0$. The reductions in (3.8) and (3.9) follow from direct evaluation using the fact that all

but one entries of $\widetilde{\mathbf{a}}_x$ and $\widetilde{\mathbf{a}}_y$ are zero,[1] and $\widetilde{\gamma} \in \mathbb{R}$ is the diagonal element of $\widetilde{\Gamma}$ corresponding to the index of nonzero entry of $\widetilde{\mathbf{a}}_x$.[2] The detailed calculation is given in Appendix A.

3.3. Spatial discretization. In this subsection, we introduce the finite difference scheme used to discretize the spatial derivatives in the micro-macro scheme (3.4). Here we consider $\mathbb{R}^2$ as the spatial domain and a uniform mesh on $\mathbb{R}^2$ with points (x_i, y_j). The distances between mesh points in the x- and y-directions are denoted by Δx and Δy, respectively. We also assume that the aspect ratio $(\Delta x/\Delta y)$ of the spatial mesh is bounded from above and away from zero. We use w and $\mathbf{w}$ to denote the scalar- and vector-valued functions on $\mathbb{R}^2$. For $w\colon \mathbb{R}^2 \to \mathbb{R}$, we denote $w_{i,j} = w(x_i, y_j)$. For $\mathbf{w}\colon \mathbb{R}^2 \to \mathbb{R}^m$, we change the notation and use $\mathbf{w}_{i,j}$ to denote $\mathbf{w}(x_i, y_j)$ instead of the entries of $\mathbf{w}$. We summarize the spatial discretization for each term in (3.4) as follows.

For the advection term in the macro equation (3.7a), we use the standard central difference scheme with additional artificial dissipation terms. For the diffusion terms in (3.7a), we adopt the symmetric scheme proposed in [26] and modify the scheme by introducing some averaging coefficients into the diffusion stencil. For the micro equation (3.7b), we discretize the micro and macro advection terms with a second-order kinetic upwind scheme and a central difference scheme, respectively. The artificial dissipation terms and the modified symmetric scheme in the discretization of (3.7a) are needed for proving the *positive-preserving property*. Specifically, they guarantee that $\overline{u}^{n+1} \geq 0$ on the spatial mesh, provided $\overline{u}^n \geq 0$ on the spatial mesh.

3.3.1. Macro equation—advection term. For the advection term in the macro equation (3.7a), we use the central difference scheme with additional artificial dissipation. Specifically, the advection term in (3.7a) is approximated by

(3.10)
$$\left((\widetilde{\mathbf{a}}_x^T \partial_x + \widetilde{\mathbf{a}}_y^T \partial_y)(\widetilde{\Gamma}\widetilde{\mathbf{u}}^n)\right)_{i,j} \approx (\widetilde{\mathbf{a}}_x^T D_x^{\mathrm{c}} + \widetilde{\mathbf{a}}_y^T D_y^{\mathrm{c}})(\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{u}}_{i,j}^n) - C_{\mathrm{AD}}(\Delta x^3 \delta_x^4 + \Delta y^3 \delta_y^4)\overline{u}_{i,j}^n ,$$

where C_{AD} is the artificial dissipation parameter and D_x^{c}, D_y^{c} are central difference operators. For functions $\mathbf{w}$ on the spatial domain,

$$D_x^{\mathrm{c}}(\mathbf{w}_{i,j}) := \frac{\mathbf{w}_{i+1,j} - \mathbf{w}_{i-1,j}}{2\Delta x} \quad \text{and} \quad D_y^{\mathrm{c}}(\mathbf{w}_{i,j}) := \frac{\mathbf{w}_{i,j+1} - \mathbf{w}_{i,j-1}}{2\Delta y} . \tag{3.11}$$

For functions $w = w(x, y)$, we define the artificial dissipation operator in the x-direction by

$$\delta_x^4(w_{i,j}) := \frac{1}{\Delta x^4}\left((w^+_{i+\frac{1}{2},j} - w^-_{i+\frac{1}{2},j}) - (w^+_{i-\frac{1}{2},j} - w^-_{i-\frac{1}{2},j})\right) , \tag{3.12}$$

where

$$w^+_{i+\frac{1}{2},j} := w_{i+1,j} - \frac{\Delta x}{2} s^x_{i+1,j} , \quad w^-_{i+\frac{1}{2},j} := w_{i,j} + \frac{\Delta x}{2} s^x_{i,j} , \tag{3.13}$$

and

$$s^x_{i,j} = \operatorname{minmod}\left\{\theta \frac{w_{i+1,j} - w_{i,j}}{\Delta x}, \frac{w_{i+1,j} - w_{i-1,j}}{2\Delta x}, \theta \frac{w_{i,j} - w_{i-1,j}}{\Delta x}\right\} . \tag{3.14}$$

[1] This is due to the facts that $\overline{m}$ is a constant, that Ω_x and Ω_y are scalar multiples of some entries in $\widetilde{\mathbf{m}}$, and that entries of $\mathbf{m}$ are orthogonal.

[2] An equivalent definition of $\widetilde{\gamma}$ is the diagonal element of $\widetilde{\Gamma}$ corresponding to the index of nonzero entry of $\widetilde{\mathbf{a}}_y$. See Appendix A for details.

Here $\theta \in [1,2]$ is a parameter, and the minmod limiter returns the real number with the smallest absolute value in the convex hull of the three arguments (see [54, section 16.3]). It can be verified that when $\theta = 1$, the minmod limiter leads to an inconsistent solution in the diffusion limit (see [61] for relevant discussion in the discontinuous Galerkin setting). On the other hand, it will be shown in section 4.1 that to prove the positive-preserving property, it is required that $\theta < 2$. Thus, we consider $\theta \in (1,2)$ in the remainder of this paper. The operator δ_y^4 in the y-direction is defined analogously. In this paper, we choose

$$C_{\mathrm{AD}} = \Theta \frac{\widetilde{\gamma}_{\max}}{\epsilon}, \quad \text{with} \quad \widetilde{\gamma}_{\max} := \frac{\epsilon^2}{\epsilon^2 + \sigma_{\mathrm{s}}^{\min}\Delta t}, \tag{3.15}$$

where $\Theta := \frac{1}{2-\theta}$ and $\sigma_{\mathrm{s}}^{\min} > 0$ is the lower bound of σ_{s} on the space. This choice of C_{AD} ensures that the artificial dissipation term vanishes at the diffusion limit while guaranteeing the positive-preserving property of the scheme. (See discussions in sections 3.4 and 4.1 for details.)

3.3.2. Macro equation—diffusion term. For the two diffusion terms in (3.7a), we apply a modified version of the symmetric scheme proposed in [26]. As the original scheme, the modified scheme is also formally second-order accurate and conservative. Specifically, we introduce some averaging coefficients into the discretization of the second derivatives, while keeping the mixed derivative discretizations identical to the ones used in [26]. At each point (x_i, y_j), we approximate $(\nabla_{(x,y)} \cdot (\overline{Q}\nabla_{(x,y)}\overline{u}^n))_{i,j}$ and $(\nabla_{(x,y)} \cdot (\widetilde{\mathbf{Q}}\nabla_{(x,y)}\widetilde{\mathbf{u}}^n))_{i,j}$ by

$$\begin{aligned} D_{\overline{Q}}^2(\overline{u}_{i,j}^n) = {} & \frac{\overline{a}_{x^2}}{\Delta x^2} \sum_{\ell=0,\pm1} c_\ell \left(\widetilde{\gamma}_{i+\frac{1}{2},j+\frac{\ell}{2}} (\overline{u}_{i+1,j+\ell}^n - \overline{u}_{i,j}^n) - \widetilde{\gamma}_{i-\frac{1}{2},j+\frac{\ell}{2}} (\overline{u}_{i,j}^n - \overline{u}_{i-1,j+\ell}^n) \right) \\ & + \frac{\overline{a}_{y^2}}{\Delta y^2} \sum_{k=0,\pm1} c_k \left(\widetilde{\gamma}_{i+\frac{k}{2},j+\frac{1}{2}} (\overline{u}_{i+k,j+1}^n - \overline{u}_{i,j}^n) - \widetilde{\gamma}_{i+\frac{k}{2},j-\frac{1}{2}} (\overline{u}_{i,j}^n - \overline{u}_{i+k,j-1}^n) \right) \end{aligned} \tag{3.16}$$

and

$$\begin{aligned} D_{\widetilde{\mathbf{Q}}}^2(\widetilde{\mathbf{u}}_{i,j}^n) = {} & \frac{\widetilde{\mathbf{a}}_{x^2}^T}{\Delta x^2} \sum_{\ell=0,\pm1} c_\ell \left(\widetilde{\gamma}_{i+\frac{1}{2},j+\frac{\ell}{2}} (\widetilde{\mathbf{u}}_{i+1,j+\ell}^n - \widetilde{\mathbf{u}}_{i,j}^n) - \widetilde{\gamma}_{i-\frac{1}{2},j+\frac{\ell}{2}} (\widetilde{\mathbf{u}}_{i,j}^n - \widetilde{\mathbf{u}}_{i-1,j+\ell}^n) \right) \\ & + \frac{\widetilde{\mathbf{a}}_{y^2}^T}{\Delta y^2} \sum_{k=0,\pm1} c_k \left(\widetilde{\gamma}_{i+\frac{k}{2},j+\frac{1}{2}} (\widetilde{\mathbf{u}}_{i+k,j+1}^n - \widetilde{\mathbf{u}}_{i,j}^n) - \widetilde{\gamma}_{i+\frac{k}{2},j-\frac{1}{2}} (\widetilde{\mathbf{u}}_{i,j}^n - \widetilde{\mathbf{u}}_{i+k,j-1}^n) \right) \\ & + \frac{\widetilde{\mathbf{a}}_{xy}^T}{2\Delta x \Delta y} \sum_{k=\pm1} \left(\widetilde{\gamma}_{i+\frac{k}{2},j+\frac{k}{2}} (\widetilde{\mathbf{u}}_{i+k,j+k}^n - \widetilde{\mathbf{u}}_{i,j}^n) - \widetilde{\gamma}_{i+\frac{k}{2},j-\frac{k}{2}} (\widetilde{\mathbf{u}}_{i+k,j-k}^n - \widetilde{\mathbf{u}}_{i,j}^n) \right), \end{aligned} \tag{3.17}$$

respectively, with averaging coefficients $c_0 = \frac{1}{2}$ and $c_{\pm1} = \frac{1}{4}$. Mixed derivatives do not appear in (3.16) since $\overline{Q}_{xy} = \overline{Q}_{yx} = 0$ (see (3.9)). Here we compute $\widetilde{\gamma}_{i\pm\frac{1}{2},j}$, $\widetilde{\gamma}_{i,j\pm\frac{1}{2}}$, and $\widetilde{\gamma}_{i\pm\frac{1}{2},j\pm\frac{1}{2}}$ by taking the harmonic averages of the adjacent values as proposed in [70]. Specifically,

$$\widetilde{\gamma}_{i+\frac{1}{2},j} := 2\left((\widetilde{\gamma}_{i+1,j})^{-1} + (\widetilde{\gamma}_{i,j})^{-1}\right)^{-1}, \quad \widetilde{\gamma}_{i,j+\frac{1}{2}} := 2\left((\widetilde{\gamma}_{i,j+1})^{-1} + (\widetilde{\gamma}_{i,j})^{-1}\right)^{-1}, \tag{3.18}$$

and

$$\widetilde{\gamma}_{i+\frac{1}{2},j+\frac{1}{2}} := 4\Big(\textstyle\sum_{k=0,1}\sum_{\ell=0,1}(\widetilde{\gamma}_{i+k,j+\ell})^{-1}\Big)^{-1}. \tag{3.19}$$

In the original symmetric scheme [26], there is no averaging, i.e., $c_0 = 1$ and $c_{\pm 1} = 0$. We introduce the averaging coefficients to include $(x_{i\pm1}, y_{j\pm1})$ in the second-derivative stencils. These coefficients are needed for proving the positive-preserving property (see (4.7)–(4.9) for details).

3.3.3. Micro equation. For the micro equation (3.7b), we adopt the spatial discretization used in [53] in the one-dimensional setting, but with second-order discretizations for more accurate solutions. Specifically, we discretize the micro advection term by a second-order upwind kinetic scheme (see, for example, [3, 16, 19, 65]):

$$\left((\widetilde{A}_x\partial_x + \widetilde{A}_y\partial_y)\widetilde{\Gamma}\widetilde{\mathbf{v}}^n\right)_{i,j} \approx (\widetilde{A}_x^+ D_x^- + \widetilde{A}_x^- D_x^+ + \widetilde{A}_y^+ D_y^- + \widetilde{A}_y^- D_y^+)(\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{v}}_{i,j}^n)\,, \tag{3.20}$$

where $\widetilde{A}_x^\pm := \langle \widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_x^\pm\rangle$, $\widetilde{A}_y^\pm := \langle \widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_y^\pm\rangle$, with $\Omega_x^\pm := \max\{\pm\Omega_x, 0\}$ and $\Omega_y^\pm := \max\{\pm\Omega_y, 0\}$. In the x-direction, D_x^+ and D_x^- are defined as

$$\begin{aligned} D_x^+(\mathbf{w}_{i,j}) &:= \frac{1}{\Delta x}\left(\left(\mathbf{w}_{i+1,j} - \frac{\mathbf{w}_{i+2,j} - \mathbf{w}_{i,j}}{4}\right) - \left(\mathbf{w}_{i,j} + \frac{\mathbf{w}_{i+1,j} - \mathbf{w}_{i-1,j}}{4}\right)\right)\,, \\ D_x^-(\mathbf{w}_{i,j}) &:= \frac{1}{\Delta x}\left(\left(\mathbf{w}_{i,j} - \frac{\mathbf{w}_{i+1,j} - \mathbf{w}_{i-1,j}}{4}\right) - \left(\mathbf{w}_{i-1,j} + \frac{\mathbf{w}_{i,j} - \mathbf{w}_{i-2,j}}{4}\right)\right)\,. \end{aligned} \tag{3.21}$$

In the y-direction, D_y^+ and D_y^- are defined similarly. The macro advection term is discretized using the central difference operators given in (3.11), i.e.,

$$((\widetilde{\mathbf{a}}_x\partial_x + \widetilde{\mathbf{a}}_y\partial_y)\overline{u}^n)_{i,j} \approx (\widetilde{\mathbf{a}}_x D_x^{\mathrm{c}} + \widetilde{\mathbf{a}}_y D_y^{\mathrm{c}})\overline{u}_{i,j}^n\,. \tag{3.22}$$

3.4. Fully discretized micro-macro scheme and the AP property. We now show in Theorem 3.2 that under some reasonable assumptions, the fully discretized scheme for the micro-macro system (3.6) recovers a standard explicit discretization of the diffusion equation

$$\partial_t\overline{u} - \partial_x\left(\frac{1}{3\sigma_{\mathrm{s}}}\partial_x\overline{u}\right) - \partial_y\left(\frac{1}{3\sigma_{\mathrm{s}}}\partial_y\overline{u}\right) + \sigma_{\mathrm{a}}\overline{u} = 0 \tag{3.23}$$

when $\epsilon \to 0$. For reference, the scheme is

$$\begin{aligned} (1 + (\sigma_{\mathrm{a}})_{i,j}\Delta t)\overline{u}_{i,j}^{n+1} &= \overline{u}_{i,j}^n + \frac{\Delta t^2}{\epsilon}D_{\widetilde{\mathbf{Q}}}^2(\widetilde{\mathbf{u}}_{i,j}^n) + \frac{\Delta t^2}{\epsilon^2}D_{\widetilde{Q}}^2(\overline{u}_{i,j}^n) \\ &\quad -\Delta t\left((\widetilde{\mathbf{a}}_x^T D_x^{\mathrm{c}} + \widetilde{\mathbf{a}}_y^T D_y^{\mathrm{c}})(\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{u}}_{i,j}^n) - \Theta\frac{\widetilde{\gamma}_{\max}}{\epsilon}(\Delta x^3\delta_x^4 + \Delta y^3\delta_y^4)\overline{u}_{i,j}^n\right)\,, \end{aligned} \tag{3.24a}$$

$$\widetilde{\mathbf{v}}_{i,j}^n = \epsilon^2\widetilde{\Gamma}_{i,j}^{-1}\widetilde{\mathbf{u}}_{i,j}^n\,, \tag{3.24b}$$

$$\begin{aligned} \widetilde{\mathbf{v}}_{i,j}^{n+1} &= \widetilde{\Gamma}_{i,j}\widetilde{\mathbf{v}}_{i,j}^n - \frac{\Delta t}{\epsilon}(\widetilde{A}_x^+ D_x^- + \widetilde{A}_x^- D_x^+ + \widetilde{A}_y^+ D_y^- + \widetilde{A}_y^- D_y^+)(\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{v}}_{i,j}^n) \\ &\quad - \Delta t(\widetilde{\mathbf{a}}_x D_x^{\mathrm{c}} + \widetilde{\mathbf{a}}_y D_y^{\mathrm{c}})\overline{u}_{i,j}^n\,, \end{aligned} \tag{3.24c}$$

$$\widetilde{\mathbf{u}}_{i,j}^{n+1} = \epsilon^{-2}\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{v}}_{i,j}^{n+1}\,. \tag{3.24d}$$

To prove the AP property of this fully discretized scheme, we first show that the artificial dissipation term in (3.24a) vanishes at the diffusion limit.

LEMMA 3.1. *Suppose that* (i) *the minmod limiter in* (3.14) *returns the center argument, and* (ii) *the time step* Δt *satisfies* $\Delta t \geq C(\Delta x + \Delta y)^3$ *for some constant* $C > 0$. *The artificial dissipation term* $\Theta \frac{\widetilde{\gamma}_{\max}}{\epsilon}(\Delta x^3 \delta_x^4 + \Delta y^3 \delta_y^4)\overline{u}_{i,j}$ *goes to zero as* $\epsilon \to 0$ *at each* (x_i, y_j) *on the spatial mesh.*

Proof. Under assumption (i), it is straightforward to verify that $\delta_x^4 \overline{u}_{i,j}$ and $\delta_y^4 \overline{u}_{i,j}$ are second-order approximations to $-\frac{1}{4}\partial_x^4 \overline{u}$ and $-\frac{1}{4}\partial_y^4 \overline{u}$ at (x_i, y_j), respectively. Therefore, the artificial dissipation term vanishes at the diffusion limit if $\Theta \frac{\widetilde{\gamma}_{\max}}{\epsilon}\Delta x^3$ and $\Theta \frac{\widetilde{\gamma}_{\max}}{\epsilon}\Delta y^3$ go to zero as $\epsilon \to 0$. By Young's inequality,

$$\Theta \frac{\widetilde{\gamma}_{\max}}{\epsilon}\Delta x^3 = \Theta \frac{\epsilon \Delta x^3}{\epsilon^2 + \sigma_{\mathrm{s}}^{\min}\Delta t} \leq \Theta \epsilon^{1/3} \Delta x \frac{\epsilon^2 + 2\Delta x^3}{3(\epsilon^2 + \sigma_{\mathrm{s}}^{\min}\Delta t)}, \tag{3.25}$$

and a similar upper bound can be obtained for $\Theta \frac{\widetilde{\gamma}_{\max}}{\epsilon}\Delta y^3$. Assumption (ii) then implies that such upper bounds are $O(\epsilon^{1/3})$ terms, which completes the proof. ∎

Note that (3.25) also implies that the artificial dissipation goes to zero as $\Delta x \to 0$, which shows the consistency of the proposed scheme as discussed later in Remark 3. We make the following remark on the two assumptions in Lemma 3.1.

Remark 1. Since $\theta > 1$, assumption (i) is always satisfied when away from the extrema, Δx is sufficiently small, and $\overline{u}$ is sufficiently smooth. Assumption (ii) imposes a mild lower bound on the time step. Such lower bound will be justified later in Remark 3, section 4.3.

We next show in Theorem 3.2 that the proposed scheme is *weakly* asymptotic-preserving (see [39] and discussions therein).

THEOREM 3.2. *Assume that* (i) *the initial conditions* $\overline{u}_{i,j}^0$ *and* $\widetilde{\mathbf{u}}_{i,j}^0$ *are both* $O(1)$ *quantities,* (ii) $\Delta t \geq C(\Delta x + \Delta y)^3$, *and* (iii) *the minmod limiter in* (3.14) *returns the center argument. When* $\epsilon \to 0$, *the macro update scheme* (3.24a) *becomes a consistent* 9*-point discretization of the diffusion equation* (3.23)*:*

$$\begin{aligned}(1 + (\sigma_{\mathrm{a}})_{i,j}\Delta t)\overline{u}_{i,j}^{n+1} = \overline{u}_{i,j}^n &+ \frac{\Delta t}{\Delta x^2} \sum_{\ell=0,\pm 1} c_\ell \left(\frac{\overline{u}_{i+1,j+\ell}^n - \overline{u}_{i,j}^n}{3(\sigma_{\mathrm{s}})_{i+\frac{1}{2},j+\frac{\ell}{2}}} - \frac{\overline{u}_{i,j}^n - \overline{u}_{i-1,j+\ell}^n}{3(\sigma_{\mathrm{s}})_{i-\frac{1}{2},j+\frac{\ell}{2}}} \right) \\ &+ \frac{\Delta t}{\Delta y^2} \sum_{k=0,\pm 1} c_k \left(\frac{\overline{u}_{i+k,j+1}^n - \overline{u}_{i,j}^n}{3(\sigma_{\mathrm{s}})_{i+\frac{k}{2},j+\frac{1}{2}}} - \frac{\overline{u}_{i,j}^n - \overline{u}_{i+k,j-1}^n}{3(\sigma_{\mathrm{s}})_{i+\frac{k}{2},j-\frac{1}{2}}} \right),\end{aligned} \tag{3.26}$$

where $c_0 = \frac{1}{2}$ *and* $c_{\pm 1} = \frac{1}{4}$.

Proof. We first prove by induction that when $\epsilon \to 0$, $\overline{u}_{i,j}^n$ and $\widetilde{\mathbf{u}}_{i,j}^n$ are $O(1)$ quantities for all $n \in \mathbb{N}$. The initial case is given by assumption (i). Suppose then that $\overline{u}_{i,j}^n$ and $\widetilde{\mathbf{u}}_{i,j}^n$ are $O(1)$ quantities. From (3.3), each element of $\widetilde{\Gamma}_{i,j}$ is in $(0, 1)$,

$$\epsilon^2 \widetilde{\Gamma}_{i,j}^{-1} = (\sigma_{\mathrm{s}})_{i,j}\Delta t + O(\epsilon^2), \quad \text{and} \quad \frac{\Delta t}{\epsilon^p}\widetilde{\Gamma}_{i,j} \leq \frac{\epsilon^{2-p}}{(\sigma_{\mathrm{s}})_{i,j}} = O(\epsilon^{2-p}), \quad p = 0,\, 1,\, 2. \tag{3.27}$$

Together (3.24b) and (3.27) imply that $\widetilde{\mathbf{v}}_{i,j}^n$ is $O(1)$. It then follows from (3.27) that the micro advection term in (3.24c) is an $O(\epsilon)$ term and thus vanishes as $\epsilon \to 0$. Combining (3.24b)–(3.24d) and taking $\epsilon \to 0$ then leads to

$$\widetilde{\mathbf{u}}_{i,j}^{n+1} = \lim_{\epsilon \to 0} \left(\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{u}}_{i,j}^n - \frac{\Delta t}{\epsilon^2}\widetilde{\Gamma}_{i,j}(\widetilde{\mathbf{a}}_x D_x^{\mathrm{c}} + \widetilde{\mathbf{a}}_y D_y^{\mathrm{c}})\overline{u}_{i,j}^n \right), \tag{3.28}$$

which implies that $\widetilde{\mathbf{u}}_{i,j}^{n+1}$ is an $O(1)$ term since, by (3.27), both terms on the right-hand side are $O(1)$.

On the other hand, it follows from (3.27) that the advection term in (3.24a) is $O(\epsilon^2)$ and that the diffusion term associated to $\widetilde{\mathbf{Q}}$ is $O(\epsilon)$ (see (3.8)). Thus, these two terms vanish as $\epsilon \to 0$. In addition, Lemma 3.1 implies that the artificial dissipation in (3.24a) also vanishes as $\epsilon \to 0$ under assumptions (ii) and (iii). Substituting (3.16) into (3.24a) and taking $\epsilon \to 0$ then leads to the 9-point discretization (3.26) that guarantees $\overline{u}_{i,j}^{n+1}$ is $O(1)$. Hence, it is proved by induction that for all $n \in \mathbb{N}$, $\overline{u}_{i,j}^{n}$ and $\widetilde{\mathbf{u}}_{i,j}^{n}$ are both $O(1)$ when $\epsilon \to 0$. Thus (3.24a) becomes the 9-point discretization (3.26) at the diffusion limit. ∎

In the following remark, we consider three types of diffusion limits and discuss possible relaxations of some assumptions in Theorem 3.2 at different limits.

Remark 2. As discussed in [51], the diffusion limits can be classified into three types based on the scaling of the spatial mesh. Specifically, the three types of diffusion limits are the *thick* ($(\Delta x, \Delta y) = O(1)$), *intermediate* ($(\Delta x, \Delta y) = O(\epsilon)$), and *thin* ($(\Delta x, \Delta y) = O(\epsilon^\ell)$, $\ell \geq 2$) limits. We note that the AP and positivity-preserving properties proved in Theorems 3.2 and 4.1 hold for all three types of diffusion limits. Further, it follows from (3.25) that the artificial dissipation vanishes in the intermediate and thin diffusion limits regardless of Δt. Thus assumption (ii) is necessary only when the thick diffusion limit is considered. In addition, with the time-step restriction given in (4.13), one can verify that assumption (iii) is needed only when the intermediate diffusion limit is considered.

4. The positive-preserving property.

In section 4.1, we state and prove Theorem 4.1, which gives sufficient conditions to yield the positive-preserving property of the fully discretized scheme (3.24). The approach we use to enforce these conditions in the proposed scheme is presented in sections 4.2 and 4.3. In section 4.4, we discuss the difficulties in extending the proposed scheme and the associated positivity conditions to the three-dimensional case, and we propose a possible approach.

4.1. Sufficient conditions for preserving positivity.

Here we state Theorem 4.1 on the positive-preserving property of the fully discretized scheme (3.24). For convenience, in addition to $\overline{u} \in \mathbb{R}$ and $\widetilde{\mathbf{u}} \in \mathbb{R}^{\widetilde{n}}$, we also use the notation $\mathbf{u} := [\overline{u}, \epsilon\widetilde{\mathbf{u}}^T]^T \in \mathbb{R}^n$ in Theorem 4.1. The corresponding vectors are defined as $\mathbf{a}_{x^2} := [\overline{a}_{x^2}, \widetilde{\mathbf{a}}_{x^2}^T]^T = [\frac{1}{3}, \widetilde{\mathbf{a}}_{x^2}^T]^T \in \mathbb{R}^n$, $\mathbf{a}_{xy} := [\overline{a}_{xy}, \widetilde{\mathbf{a}}_{xy}^T]^T = [0, \widetilde{\mathbf{a}}_{xy}^T]^T \in \mathbb{R}^n$, and $\mathbf{a}_{y^2} := [\overline{a}_{y^2}, \widetilde{\mathbf{a}}_{y^2}^T]^T = [\frac{1}{3}, \widetilde{\mathbf{a}}_{y^2}^T]^T \in \mathbb{R}^n$, respectively. Also, we recall the definition of $\widetilde{\gamma}_{\max}$ in (3.15).

THEOREM 4.1. *At time t^n, suppose that $\overline{u}_{i,j}^n \geq 0$ and that $\mathbf{u}_{i,j}^n := [\overline{u}_{i,j}^n, \epsilon(\widetilde{\mathbf{u}}_{i,j}^n)^T]^T$ satisfies the positivity conditions*

(C1) $\overline{u}_{i,j}^n \pm \epsilon\widetilde{\mathbf{a}}_x^T\widetilde{\mathbf{u}}_{i,j}^n \geq 0\,,$ (C2) $\overline{u}_{i,j}^n \pm \epsilon\widetilde{\mathbf{a}}_y^T\widetilde{\mathbf{u}}_{i,j}^n \geq 0\,,$

(C3) $\overline{u}_{i,j}^n \geq \mathbf{a}_{x^2}^T\mathbf{u}_{i,j}^n \geq 0,$ (C4) $\overline{u}_{i,j}^n \geq \mathbf{a}_{y^2}^T\mathbf{u}_{i,j}^n \geq 0\,,$

(C5) $\overline{u}_{i,j}^n \pm 2\mathbf{a}_{xy}^T\mathbf{u}_{i,j}^n \geq 0,$ (C6) $\left(\dfrac{\mathbf{a}_{x^2}^T}{\Delta x^2} \pm 2\dfrac{\mathbf{a}_{xy}^T}{\Delta x\Delta y} + \dfrac{\mathbf{a}_{y^2}^T}{\Delta y^2}\right)\mathbf{u}_{i,j}^n \geq 0$

for each (x_i, y_j) on the spatial mesh. Further, assume that Δt satisfies

$$1-\widetilde{\gamma}_{\max}\left(\frac{2\Theta\Delta t}{\epsilon\Delta x}+\frac{2\Theta\Delta t}{\epsilon\Delta y}+\frac{2\Delta t^2}{\epsilon^2\Delta x^2}+\frac{\Delta t^2}{2\epsilon^2\Delta x\Delta y}+\frac{2\Delta t^2}{\epsilon^2\Delta y^2}\right)\geq 0 \tag{C7}$$

where $\Theta=\frac{1}{2-\theta}$ depends on the minmod parameter $\theta\in(1,2)$ in (3.14). *Then the macro scheme* (3.24a) *guarantees that $\overline{u}_{i,j}^{n+1}\geq 0$ for each (x_i, y_j) on the spatial mesh.*

Proof. For simplicity, we write (3.24a) as

$$(1+(\sigma_{\mathrm{a}})_{i,j}\Delta t)\overline{u}_{i,j}^{n+1}=\overline{u}_{i,j}^{n}+(\mathbf{T}_x)+(\mathbf{T}_y)+(\mathbf{D})\,, \tag{4.1}$$

where $(\mathbf{T}_x)$ and $(\mathbf{T}_y)$ denote the discretized advection terms in the x- and y-directions, respectively, and $(\mathbf{D})$ denotes the sum of the two diffusion terms. Specifically,

$$(\mathbf{T}_x):=-\Delta t\left(\widetilde{\mathbf{a}}_x^T D_x^{\mathrm{c}}(\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{u}}_{i,j}^n)-\Theta\frac{\widetilde{\gamma}_{\max}}{\epsilon}\Delta x^3\delta_x^4(\overline{u}_{i,j}^n)\right)\,, \tag{4.2a}$$

$$(\mathbf{T}_y):=-\Delta t\left(\widetilde{\mathbf{a}}_y^T D_y^{\mathrm{c}}(\widetilde{\Gamma}_{i,j}\widetilde{\mathbf{u}}_{i,j}^n)-\Theta\frac{\widetilde{\gamma}_{\max}}{\epsilon}\Delta y^3\delta_y^4(\overline{u}_{i,j}^n)\right)\,, \tag{4.2b}$$

$$(\mathbf{D}):=\frac{\Delta t^2}{\epsilon}D_{\widetilde{\mathbf{Q}}}^2(\widetilde{\mathbf{u}}_{i,j}^n)+\frac{\Delta t^2}{\epsilon^2}D_{\overline{Q}}^2(\overline{u}_{i,j}^n)\,. \tag{4.2c}$$

Since σ_{a} is assumed to be nonnegative, we know from (4.1) that $\overline{u}_{i,j}^{n+1}\geq 0$ if

$$\overline{u}_{i,j}^{n}+(\mathbf{T}_x)+(\mathbf{T}_y)+(\mathbf{D})\geq 0\,, \tag{4.3}$$

which we now show.

We first consider the term $(\mathbf{T}_x)$. As shown in Appendix B, the operator δ_x^4 defined in (3.12) satisfies

$$\delta_x^4(\overline{u}_{i,j}^n)\geq\frac{1}{2\Delta x^4}\left(\frac{1}{\Theta}\overline{u}_{i+1,j}^n-4\overline{u}_{i,j}^n+\frac{1}{\Theta}\overline{u}_{i-1,j}^n\right)\,. \tag{4.4}$$

Applying (4.4) and the definition of D_x^{c} in (3.11) to (4.2a) leads to

(4.5)

$$\begin{aligned}(\mathbf{T}_x)&\geq\frac{\Delta t}{2\epsilon\Delta x}\left(\widetilde{\gamma}_{\max}\left(\overline{u}_{i+1,j}^n-4\Theta\overline{u}_{i,j}^n+\overline{u}_{i-1,j}^n\right)-\epsilon\widetilde{\mathbf{a}}_x^T(\widetilde{\Gamma}_{i+1,j}\widetilde{\mathbf{u}}_{i+1,j}^n-\widetilde{\Gamma}_{i-1,j}\widetilde{\mathbf{u}}_{i-1,j}^n)\right)\\&\geq\frac{\Delta t}{2\epsilon\Delta x}\Big(\widetilde{\gamma}_{\max}\left(\overline{u}_{i+1,j}^n-4\Theta\overline{u}_{i,j}^n+\overline{u}_{i-1,j}^n\right)-\epsilon\widetilde{\gamma}_{\max}\left(|\widetilde{\mathbf{a}}_x^T\widetilde{\mathbf{u}}_{i+1,j}^n|+|\widetilde{\mathbf{a}}_x^T\widetilde{\mathbf{u}}_{i-1,j}^n|\right)\Big)\\&=\frac{\widetilde{\gamma}_{\max}\Delta t}{2\epsilon\Delta x}\sum_{k=\pm1}\left(\overline{u}_{i+k,j}^n-\epsilon|\widetilde{\mathbf{a}}_x^T\widetilde{\mathbf{u}}_{i+k,j}^n|\right)-\widetilde{\gamma}_{\max}\frac{2\Theta\Delta t}{\epsilon\Delta x}\overline{u}_{i,j}^n\,.\end{aligned}$$

Here the second inequality follows from two facts: (i) all diagonal entries of $\widetilde{\Gamma}$ are bounded from above by $\widetilde{\gamma}_{\max}$, and (ii) all but one entries of $\widetilde{\mathbf{a}}_x$ are zero, as discussed in Appendix A. A similar lower bound for $(\mathbf{T}_y)$ can be obtained analogously. It follows from (C1) that the first term in the lower bound of $(\mathbf{T}_x)$ is nonnegative. Similarly, the corresponding term in the lower bound of $(\mathbf{T}_y)$ is also nonnegative from (C2). Thus, by plugging the lower bounds of $(\mathbf{T}_x)$ and $(\mathbf{T}_y)$ into (4.3), it suffices to show that

$$\left(1-\widetilde{\gamma}_{\max}\left(\frac{2\Theta\Delta t}{\epsilon\Delta x}+\frac{2\Theta\Delta t}{\epsilon\Delta y}\right)\right)\overline{u}_{i,j}^n+(\mathbf{D})\geq 0\,. \tag{4.6}$$

We next consider the term (**D**). Substituting (3.16) and (3.17) into (4.2c) gives

(4.7)
$$
\begin{aligned}
(\mathbf{D}) = &\frac{\Delta t^2}{\epsilon^2 \Delta x^2} \mathbf{a}_{x^2}^T \sum_{\ell=0,\pm 1} c_\ell \left(\widetilde{\gamma}_{i+\frac{1}{2},j+\frac{\ell}{2}} (\mathbf{u}_{i+1,j+\ell}^n - \mathbf{u}_{i,j}^n) - \widetilde{\gamma}_{i-\frac{1}{2},j+\frac{\ell}{2}} (\mathbf{u}_{i,j}^n - \mathbf{u}_{i-1,j+\ell}^n) \right) \\
&+ \frac{\Delta t^2}{\epsilon^2 \Delta y^2} \mathbf{a}_{y^2}^T \sum_{k=0,\pm 1} c_k \left(\widetilde{\gamma}_{i+\frac{k}{2},j+\frac{1}{2}} (\mathbf{u}_{i+k,j+1}^n - \mathbf{u}_{i,j}^n) - \widetilde{\gamma}_{i+\frac{k}{2},j-\frac{1}{2}} (\mathbf{u}_{i,j}^n - \mathbf{u}_{i+k,j-1}^n) \right) \\
&+ \frac{\Delta t^2}{2\epsilon^2 \Delta x \Delta y} \mathbf{a}_{xy}^T \sum_{k=\pm 1} \left(\widetilde{\gamma}_{i+\frac{k}{2},j+\frac{k}{2}} (\mathbf{u}_{i+k,j+k}^n - \mathbf{u}_{i,j}^n) - \widetilde{\gamma}_{i+\frac{k}{2},j-\frac{k}{2}} (\mathbf{u}_{i+k,j-k}^n - \mathbf{u}_{i,j}^n) \right) ,
\end{aligned}
$$

where $\mathbf{u} = [\overline{u}, \epsilon \widetilde{\mathbf{u}}^T]^T$ and the vectors $\mathbf{a}_{x^2}$, $\mathbf{a}_{xy}$, and $\mathbf{a}_{y^2}$ are defined in the beginning of this section. Collecting the terms in (4.7) based on the spatial indices leads to

(4.8)
$$
\begin{aligned}
(\mathbf{D}) \geq &- \frac{\widetilde{\gamma}_{\max}}{\epsilon^2} \left(\frac{2\Delta t^2}{\Delta x^2} \left| \mathbf{a}_{x^2}^T \mathbf{u}_{i,j}^n \right| + \frac{\Delta t^2}{\Delta x \Delta y} \left| \mathbf{a}_{xy}^T \mathbf{u}_{i,j}^n \right| + \frac{2\Delta t^2}{\Delta y^2} \left| \mathbf{a}_{y^2}^T \mathbf{u}_{i,j}^n \right| \right) \\
&+ \frac{\Delta t^2}{2\epsilon^2 \Delta x^2} \sum_{k=\pm 1} \widetilde{\gamma}_{i+\frac{k}{2},j} \mathbf{a}_{x^2}^T \mathbf{u}_{i+k,j}^n + \frac{\Delta t^2}{2\epsilon^2 \Delta y^2} \sum_{\ell=\pm 1} \widetilde{\gamma}_{i,j+\frac{\ell}{2}} \mathbf{a}_{y^2}^T \mathbf{u}_{i,j+\ell}^n \\
&+ \frac{\Delta t^2}{4\epsilon^2} \left(\frac{\mathbf{a}_{x^2}^T}{\Delta x^2} + 2\frac{\mathbf{a}_{xy}^T}{\Delta x \Delta y} + \frac{\mathbf{a}_{y^2}^T}{\Delta y^2} \right) \sum_{k=\pm 1} \widetilde{\gamma}_{i+\frac{k}{2},j+\frac{k}{2}} \mathbf{u}_{i+k,j+k}^n \\
&+ \frac{\Delta t^2}{4\epsilon^2} \left(\frac{\mathbf{a}_{x^2}^T}{\Delta x^2} - 2\frac{\mathbf{a}_{xy}^T}{\Delta x \Delta y} + \frac{\mathbf{a}_{y^2}^T}{\Delta y^2} \right) \sum_{k=\pm 1} \widetilde{\gamma}_{i+\frac{k}{2},j-\frac{k}{2}} \mathbf{u}_{i+k,j-k}^n ,
\end{aligned}
$$

where the inequality follows from dropping nonnegative terms, taking absolute values, and bounding all $\widetilde{\gamma}$'s with $\widetilde{\gamma}_{\max}$ in the first term. From (C3), (C4), and (C6), all but the first term on the right-hand side of (4.8) are nonnegative and thus can be dropped from the inequality. We then apply (C3), (C4), and (C5) to the remaining term, which yields

$$
(\mathbf{D}) \geq -\widetilde{\gamma}_{\max} \left(\frac{2\Delta t^2}{\epsilon^2 \Delta x^2} + \frac{\Delta t^2}{2\epsilon^2 \Delta x \Delta y} + \frac{2\Delta t^2}{\epsilon^2 \Delta y^2} \right) \overline{u}_{i,j}^n . \tag{4.9}
$$

After plugging this lower bound into (4.6), it now suffices to show that

$$
\left(1 - \widetilde{\gamma}_{\max} \left(\frac{2\Theta \Delta t}{\epsilon \Delta x} + \frac{2\Theta \Delta t}{\epsilon \Delta y} + \frac{2\Delta t^2}{\epsilon^2 \Delta x^2} + \frac{\Delta t^2}{2\epsilon^2 \Delta x \Delta y} + \frac{2\Delta t^2}{\epsilon^2 \Delta y^2} \right) \right) \overline{u}_{i,j}^n \geq 0 . \tag{4.10}
$$

From (C7), (4.10) holds and the proof is complete. ∎

Theorem 4.1 provides sufficient conditions (C1)–(C7) to preserve positivity of the particle concentrations with the proposed scheme. In general, these conditions are not readily satisfied. In section 4.2, we review two limiters proposed in [46] and adopt them to enforce conditions (C1)–(C6). For condition (C7), we show in section 4.3 that this condition is satisfied if a CFL-type time-step restriction is imposed.

4.2. Realizability limiters. It is known [13, 43, 44, 55, 71] that for a nonnegative spectral expansion $h := \mathbf{m}^T \mathbf{u}$ on $\mathbb{S}^2$, the moments of h satisfy some conditions referred to as the *realizability* conditions. Conditions (C1)–(C6), which we impose on $\mathbf{u}$ to preserve positivity, correspond to a subset of these realizability conditions. To confirm that (C1)–(C6) are satisfied for all $\mathbf{u}$ such that $h = \mathbf{m}^T \mathbf{u} \geq 0$ on $\mathbb{S}^2$, we

observe that since $|\Omega_x| \leq 1$, $\langle (1 \pm \Omega_x) h \rangle \geq 0$. Therefore, plugging $h = \mathbf{m}^T \mathbf{u}$ into this inequality verifies that $\mathbf{u}$ satisfies (C1). (C2)–(C6) can be verified analogously. We refer to conditions with this property as *physical* conditions, and conclude that enforcing these physical conditions does not affect nonnegative spectral expansions.

For a given angular spectral expansion with nonnegative mean, the limiters considered in [46] give approximations that are nonnegative pointwisely on a specific quadrature set while preserving the mean. We refer to these limiters as pointwise limiters in this paper. It is straightforward to verify that the pointwise nonnegativity condition required by these limiters is stronger than conditions (C1)–(C6). Thus we borrow the idea of these pointwise limiters, but relax them to enforce (C1)–(C6). These new realizability limiters preserve the values of the macro coefficients $\overline{u}$ and modify the micro coefficients $\widetilde{\mathbf{u}}$ to satisfy (C1)–(C6). We expect these realizability limiters to be more efficient than the pointwise limiters in terms of accuracy and computation cost, since the realizability limiters enforce weaker conditions and are less likely to be active.

We first consider the linear scaling (`ls`) limiter [59, 74, 75], which damps the micro coefficients uniformly until some desirable condition (C) on $\mathbf{u}$ is satisfied. Specifically, given $\mathbf{u} = [\overline{u}, \epsilon \widetilde{\mathbf{u}}^T]^T$ with $\overline{u} \geq 0$, the `ls` limiter produces an approximation $\mathbf{u}_{\texttt{ls}} := [\overline{u}_{\texttt{ls}}, \epsilon \widetilde{\mathbf{u}}_{\texttt{ls}}^T]^T$ such that $\overline{u}_{\texttt{ls}} = \overline{u}$ and

$$\widetilde{\mathbf{u}}_{\texttt{ls}} = \alpha_{\texttt{ls}} \widetilde{\mathbf{u}} \text{ with } \alpha_{\texttt{ls}} := \operatorname*{argmax}_{\alpha \in [0,1]} \left\{ \alpha \colon [\overline{u}, \epsilon \alpha \widetilde{\mathbf{u}}^T]^T \text{ satisfies (C)} \right\} . \tag{4.11}$$

The second limiter considered in this paper is the optimization-based (`opt`) limiter [47], which finds the best approximation to the micro coefficients, in the ℓ^2 sense, that still satisfies (C). Specifically, given $\mathbf{u} := [\overline{u}, \epsilon \widetilde{\mathbf{u}}^T]^T$ with $\overline{u} \geq 0$, the `opt` limiter gives an approximation $\mathbf{u}_{\texttt{opt}} := [\overline{u}_{\texttt{opt}}, \epsilon \widetilde{\mathbf{u}}_{\texttt{opt}}^T]^T$ such that $\overline{u}_{\texttt{opt}} = \overline{u}$ and

$$\widetilde{\mathbf{u}}_{\texttt{opt}} = \operatorname*{argmin}_{\widetilde{\mathbf{v}} \in \mathbb{R}^{\tilde{n}}} \left\{ \frac{1}{2} \|\widetilde{\mathbf{v}} - \widetilde{\mathbf{u}}\|_2^2 \colon [\overline{u}, \epsilon \widetilde{\mathbf{v}}^T]^T \text{ satisfies (C)} \right\} . \tag{4.12}$$

Here we apply these limiters to $\mathbf{u}_{i,j}^n$ at each $((x_i, y_j), t^n)$ on the space-time mesh in order to enforce conditions (C1)–(C6). We embed the limiters into the proposed scheme such that when $\mathbf{u}_{i,j}^n$ violates any of (C1)–(C6), we compute the limited coefficients $\mathbf{u}_{i,j,\texttt{ls}}^n$ or $\mathbf{u}_{i,j,\texttt{opt}}^n$ and then proceed with $\mathbf{u}_{i,j}^n$ replaced by $\mathbf{u}_{i,j,\texttt{ls}}^n$ or $\mathbf{u}_{i,j,\texttt{opt}}^n$. We refer to these realizability limiters as `ls-r` and `opt-r`, respectively.

In the numerical experiments reported in section 5, we compare the `ls-r` and `opt-r` limiters as well as their pointwise versions, denoted, respectively, as `ls-pw` and `opt-pw`, considered in [46]. The `ls-pw` and `opt-pw` limiters are formulated by replacing (C) in (4.11) and (4.12), respectively, by the analogous pointwise condition. From the numerical results in section 5, we confirm that the `ls-r` and `opt-r` limiters are more efficient than the `ls-pw` and `opt-pw` limiters.

4.3. Positivity time-step restriction. In this section, we show that condition (C7) in Theorem 4.1 is satisfied under a CFL-type time-step restriction stated in the following lemma.

LEMMA 4.2. *Condition* (C7) *holds if* Δt *satisfies*

$$\Delta t \leq \max \{ \Delta t_{\text{hyp}}, \Delta t_{\text{par}} \} , \tag{4.13}$$

where

$$\Delta t_{\text{hyp}} := 2\left(\sqrt{\frac{9}{8}+\Theta^2}-\Theta\right)\epsilon\left(\frac{\Delta x\Delta y(\Delta x+\Delta y)}{4\Delta x^2+\Delta x\Delta y+4\Delta y^2}\right), \tag{4.14}$$

$$\Delta t_{\text{par}} := \sigma_{\text{s}}^{\min}\left(\frac{\Delta x^2\Delta y^2}{(2+\Theta^2)\Delta x^2+(\frac{1}{2}+2\Theta^2)\Delta x\Delta y+(2+\Theta^2)\Delta y^2}\right), \tag{4.15}$$

and the constant $\Theta = \frac{1}{2-\theta}$ *depends on the minmod parameter* $\theta \in (1,2)$ *in* (3.14).

Proof. From the definition of $\widetilde{\gamma}_{\max}$ in (3.15), (C7) is equivalent to

$$h_{\Delta t}(\epsilon) := \epsilon^2 + \sigma_{\text{s}}^{\min}\Delta t - \Delta t\left(\frac{2\Theta\epsilon}{\Delta x}+\frac{2\Theta\epsilon}{\Delta y}+\frac{2\Delta t}{\Delta x^2}+\frac{\Delta t}{2\Delta x\Delta y}+\frac{2\Delta t}{\Delta y^2}\right)\geq 0\,. \tag{4.16}$$

The minimizer of the parabola $h_{\Delta t}$ is given by $\epsilon^* = \frac{\Theta\Delta t(\Delta x+\Delta y)}{\Delta x\Delta y}$. Therefore, to prove (C7), it suffices to show that

$$h_{\Delta t}(\epsilon^*) = \sigma_{\text{s}}^{\min}\Delta t - \frac{\Delta t^2}{\Delta x^2\Delta y^2}\left(\Theta^2(\Delta x+\Delta y)^2+\left(2\Delta x^2+\frac{1}{2}\Delta x\Delta y+2\Delta y^2\right)\right)\geq 0\,, \tag{4.17}$$

which leads to

$$\Delta t \leq \sigma_{\text{s}}^{\min}\left(\frac{\Delta x^2\Delta y^2}{(2+\Theta^2)\Delta x^2+(\frac{1}{2}+2\Theta^2)\Delta x\Delta y+(2+\Theta^2)\Delta y^2}\right) = \Delta t_{\text{par}}\,. \tag{4.18}$$

On the other hand, since $\sigma_{\text{s}}^{\min} > 0$, it follows from (4.16) that

$$\epsilon^2 - 2\Delta t\left(\frac{\Theta\epsilon}{\Delta x}+\frac{\Theta\epsilon}{\Delta y}+\frac{\Delta t}{\Delta x^2}+\frac{\Delta t}{4\Delta x\Delta y}+\frac{\Delta t}{\Delta y^2}\right)\geq 0 \tag{4.19}$$

is also a sufficient condition for (C7). Solving this quadratic inequality in Δt gives

$$\Delta t \leq 2\epsilon\Delta x\Delta y\left(\frac{\left(\Theta^2(\Delta x+\Delta y)^2+2\Delta x^2+\frac{1}{2}\Delta x\Delta y+2\Delta y^2\right)^{\frac{1}{2}}-\Theta(\Delta x+\Delta y)}{4\Delta x^2+\Delta x\Delta y+4\Delta y^2}\right). \tag{4.20}$$

Applying the inequality $2\Delta x^2+\frac{1}{2}\Delta x\Delta y+2\Delta y^2 \geq \frac{9}{8}(\Delta x+\Delta y)^2$ then yields

$$\Delta t \leq 2\left(\sqrt{\frac{9}{8}+\Theta^2}-\Theta\right)\epsilon\left(\frac{\Delta x\Delta y(\Delta x+\Delta y)}{4\Delta x^2+\Delta x\Delta y+4\Delta y^2}\right) = \Delta t_{\text{hyp}}\,. \tag{4.21}$$

Since (4.18) and (4.21) are sufficient conditions for (C7), the claim is proved. ∎

Since the aspect ratio $(\Delta x/\Delta y)$ is assumed to be bounded from above and away from zero, the time-step restriction (4.13) switches from a hyperbolic CFL condition to a parabolic CFL condition as $\epsilon \to 0$. Specifically, when $\epsilon \gg (\Delta x+\Delta y)$, (4.13) takes the form of a hyperbolic CFL condition, i.e., $\Delta t \leq C\epsilon(\Delta x+\Delta y)$. On the other hand, when $\epsilon \ll (\Delta x+\Delta y)$, (4.13) switches to a parabolic CFL condition, i.e., $\Delta t \leq C\Delta x\Delta y$. The switch between time-step restrictions is desirable for AP schemes, since the hyperbolic CFL condition becomes prohibitive as $\epsilon \to 0$.

Remark 3. The time-step restriction (4.13) justifies the time-step assumption $\Delta t \geq C(\Delta x+\Delta y)^3$, which is invoked in the AP property analysis in Lemma 3.1

and Theorem 3.2. In addition, if the time step Δt is chosen to be the largest step allowed by (4.13), then (3.25) can be rewritten as

$$\Theta \frac{\widetilde{\gamma}_{\max}}{\epsilon} \Delta x^3 = \Theta \frac{\epsilon \Delta x^3}{\epsilon^2 + \sigma_{\mathrm{s}}^{\min} \Delta t} \leq \Theta \Delta x^2 \frac{\epsilon^2 + \Delta x^2}{2(\epsilon^2 + \sigma_{\mathrm{s}}^{\min} \Delta t)} \leq C_x \Delta x^2 \,, \tag{4.22}$$

with constant C_x independent of ϵ and Δx. Here the first inequality follows from Young's inequality, and the second inequality uses the fact that Δt is the largest step allowed by (4.13). Inequality (4.22) implies that the artificial dissipation vanishes as $\Delta x \to 0$. Thus, we conclude that the proposed scheme is consistent and AP under (4.13).

4.4. Extension to three dimensions. The spatial discretization and positivity conditions proposed in sections 3.3 and 4.1 are for the micro-macro system corresponding to the reduced linear kinetic equation (3.5) in two dimensions. For the original three-dimensional equation (2.2), a straightforward extension of the proposed spatial discretization leads to a positive-preserving AP scheme, under an extended version of positivity conditions. However, such extension of the diffusion stencils (3.16) and (3.17) gives a discretization that is defined on alternating spatial grids. In the three-dimensional setting, let $C_{i,j,k}$ denote a cuboid of size $\Delta x \times \Delta y \times \Delta z$ centered at (x_i, y_j, z_k). Without averaging coefficients, the extended diffusion stencils calculate the second derivatives at (x_i, y_j, z_k) using the function values at the "face" points $(x_{i\pm1}, y_j, z_k)$, $(x_i, y_{j\pm1}, z_k)$, $(x_i, y_j, z_{k\pm1})$ of $C_{i,j,k}$. On the other hand, the mixed derivatives are computed using the values at the "edge" points $(x_{i\pm1}, y_{j\pm1}, z_k)$, $(x_{i\pm1}, y_j, z_{k\pm1})$, $(x_i, y_{j\pm1}, z_{k\pm1})$ of $C_{i,j,k}$. To form physical positivity conditions (see section 4.2), we need to modify the diffusion stencil by introducing averaging coefficients as in the two-dimensional case. Nevertheless, in the three-dimensional setting, all weights on the face points have to be distributed to the adjacent edge points in the second-derivative stencil to obtain physical positivity conditions. Thus, the face points are no longer in the modified diffusion stencil, and the resulting 13-point stencil includes only the center and edge points. It is then straightforward to verify that this stencil is defined on alternating spatial grids; e.g., the stencil centered at (x_i, y_j, z_k) is completely disjoint with the one centered at (x_{i+1}, y_j, z_k). Such discretization may lead to oscillatory solutions; see, for example, [28] for relevant discussions.

A possible remedy is to modify the extended diffusion stencil so that the mixed derivatives are computed using values at "corner" points $(x_{i\pm1}, y_{j\pm1}, z_{k\pm1})$ of $C_{i,j,k}$. In this case, the second-derivative stencil needs to be distributed from the face points to the adjacent corner points, and physical positivity conditions can be obtained when the weights on the face points are still positive. In the case of constant scattering cross-sections, this procedure leads to a 15-point stencil that includes the center, face, and corner points. This stencil does not suffer from the problem of alternating grids and is expected to preserve positivity under physical conditions. The extension to problems with general scattering cross-sections is left to future work.

5. Numerical results. In this section, we test the performance of the scheme in solving the reduced kinetic equation (3.5) for three benchmark problems. We also run a space-time accuracy test.

5.1. Line source problem. The line source problem and its semi-analytic solution were first considered in [18]. The problem has served as a performance benchmark in studying various numerical schemes for solving linear kinetic equations [6, 19, 31, 60, 68]. It involves an isotropic initial condition supported at the

origin of the spatial domain. In our numerical simulations, the initial condition is approximated by a steep Gaussian distribution centered at the origin with variance $\varsigma^2 = 9 \times 10^{-4}$, i.e.,

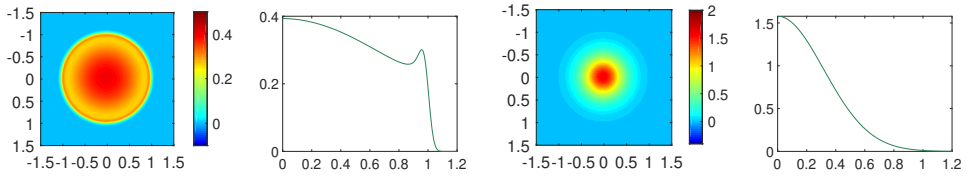
$$f^{\text{in}}(r, \Omega) \approx \frac{1}{4\pi} \left(\frac{1}{2\pi\varsigma^2} e^{\frac{-(x^2+y^2)}{2\varsigma^2}} \right), \tag{5.1}$$

and the cross-sections are chosen to be $\sigma_{\text{t}} = \sigma_{\text{s}} = 1$. Tests are run in the kinetic regime ($\epsilon = 1$) and the diffusive regime ($\epsilon = 10^{-3}$).

The simulation is performed on a truncated spatial domain: a 3×3 square centered at the origin with zero boundary condition. The final time is $t_{\text{final}} = 1$ when $\epsilon = 1$ and $t_{\text{final}} = 0.1$ when $\epsilon = 10^{-3}$. We choose the angular spectral approximation order to be $N = 11$ for the kinetic tests and $N = 3$ for the diffusive tests. We perform the computation on a 150×150 uniform square spatial mesh with the time step chosen as 0.9 times the maximum value allowed by condition (4.13). The filter function κ is given by $\kappa(\lambda) = \frac{1}{1+\lambda^4}$ with the filtering parameter $\sigma_{\text{f}} = 56.2$. The minmod parameter in (3.14) is chosen to be $\theta = 1.5$. In each regime, we solve the problem using the proposed AP scheme with the `ls-r` and `opt-r` realizability limiters and, for comparison, the `ls-pw` and `opt-pw` pointwise limiters considered in [46]. See section 4.2 for the details of these limiters. We implement the `ls-r` and `ls-pw` limiters by solving the maximization problems via direct evaluation. There is no optimization required for these two limiters since conditions (C1)–(C6) are linear inequalities. On the other hand, the minimization problems for the `opt-r` and `opt-pw` limiters are solved, respectively, using the alternating direction method of multipliers (ADMM) [5] and the constraint-reduced Mehrotra-predictor-corrector method (CR-MPC) [48], with tolerance 10^{-6}. The optimization algorithms are chosen such that the computational cost is minimized.

In the kinetic regime, we use the semi-analytic solution given in [18] as the reference solution. In the diffusive regime, the reference solution is generated by solving the diffusion equation (3.23) with the explicit 9-point finite difference scheme (3.26). In Figure 1, we plot the two-dimensional heat maps and one-dimensional line-outs (along the x-axis) of the particle concentration $\rho = \langle f \rangle$ in the reference solutions.

In Figure 1, we illustrate the particle concentration $\rho = \langle f \rangle$ in the reference solutions in two forms—two-dimensional heat maps that present the values of ρ on the x-y spatial domain as colors, and one-dimensional line-outs that plot the values of ρ along the nonnegative x-axis ($x \geq 0$, $y = 0$).



(a) $\epsilon = 1$, $t = 1$. (b) $\epsilon = 10^{-3}$, $t = 0.1$.

FIG. 1. *Reference solutions for the line source problem. Heat maps and line-outs show the particle concentration ρ in the kinetic ($\epsilon = 1$) and diffusive ($\epsilon = 10^{-3}$) regimes.*

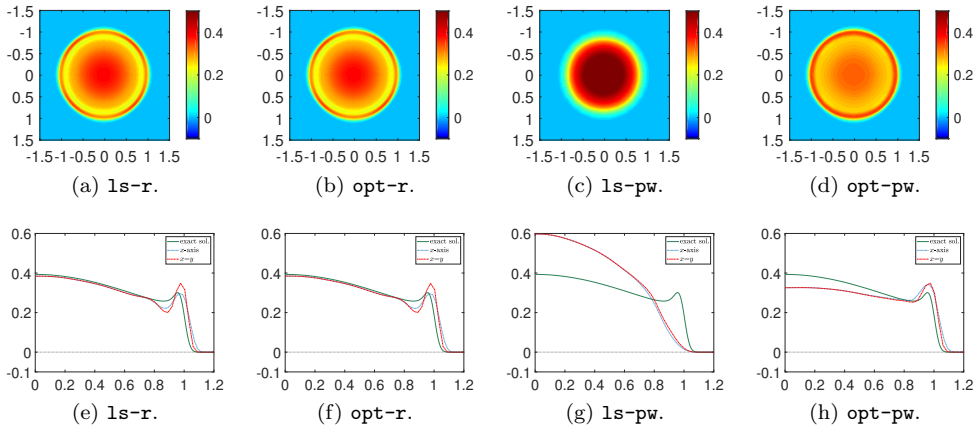
Similar heat maps and line-outs of the numerical solution in the kinetic regime ($\epsilon = 1$) are shown in Figure 2. Each of the one-dimensional line-outs is plotted along the x-axis and along the direction of 45 degrees, which shows the most inaccurate

part of the solution. For comparison, the reference kinetic solution is included in all line-out figures. Plots for the diffusive tests are omitted because the numerical and reference solutions are visually identical.

The run time and relative L^2 spatial errors of the particle concentration in both the kinetic and diffusive regimes are reported in Table 1. The relative L^2 spatial error is defined as

$$E := \frac{\|\rho_c - \rho_{ref}\|_{L_h^2(\mathbb{R}^2)}}{\|\rho_{ref}\|_{L_h^2(\mathbb{R}^2)}}, \quad \text{with} \quad \|\rho\|_{L_h^2(\mathbb{R}^2)} := \left(\sum_{i,j} \rho_{i,j}^2 h^2\right)^{1/2}, \tag{5.2}$$

where ρ_c is the computed solution, ρ_{ref} is the reference solution, the summation in (5.2) is taken over all (i, j) such that (x_i, y_j) belongs to the uniform spatial mesh, and $h = \Delta x = \Delta y$. Figure 2 shows the particle concentrations at $t = t_{final}$ of the computed solutions with various limiters, respectively.



(a) `ls-r`. (b) `opt-r`. (c) `ls-pw`. (d) `opt-pw`.

(e) `ls-r`. (f) `opt-r`. (g) `ls-pw`. (h) `opt-pw`.

FIG. 2. *Numerical solutions for the line source problem. Heat maps and line-outs show the particle concentration ρ generated with the* `ls-r`*,* `opt-r`*,* `ls-pw`*, and* `opt-pw` *limiters in the kinetic ($\epsilon = 1$) regime. The approximation order of the* FP_N *equations is $N = 11$.*

TABLE 1

Run time (sec) and relative L^2 spatial error E for the line source problem without a limiter and with the four limiters. In the diffusive regime, the positivity conditions are never violated due to the smooth solution. Thus, all limiters give identical solutions.

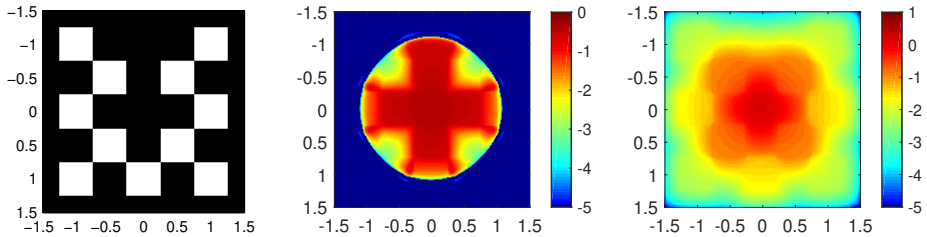
Limiter		`none`	`ls-r`	`opt-r`	`ls-pw`	`opt-pw`
Kinetic ($N = 11$)	run time	280	342	348	413	20847
	E	0.107	0.106	0.106	0.494	0.147
Diffusive ($N = 3$)	run time	149	1040	1055	1082	1044
	E	0.005	0.005	0.005	0.005	0.005

In the kinetic regime ($\epsilon = 1$), we observe in Figure 2 that with the `ls-r`, `opt-r`, and `opt-pw` limiters, the computed solutions are reasonably accurate but slightly more diffusive compared to the reference solution, which we suspect is due to the artificial dissipation terms in (3.10). Meanwhile, the solution with the `ls-pw` limiter is inaccurate. Figure 2 also shows that the solutions with the `ls-r` and `opt-r` limiters have some noticeable artifacts that affect the rotational invariance of the solution. We believe these artifacts come from the axis-dependency of conditions (C1)–(C6), as they

seem to align with the spatial axes. The artifacts are less noticeable in the solution from the `opt-pw` limiter, which enforces a stronger pointwise positivity condition and thus applies more damping on the micro coefficients than the `ls-r` and `opt-r` limiters do.

The results in the kinetic regime ($\epsilon = 1$) reported in Table 1 indicate that the `ls-r` and `opt-r` limiters give solutions that are as accurate as the unlimited solution. The `opt-pw` limiter gives a slightly less accurate solution, while it is about 50x more computationally expensive than the other three limiters. In the diffusive regime ($\epsilon = 10^{-3}$), the reference solution is sufficiently smooth such that conditions (C1)–(C6) and the pointwise positivity condition are never violated. Hence all computed solutions are identical and close to the reference diffusion solution. We also notice that the difference between the computational time of the limited and unlimited cases is more obvious in the diffusive regime. This is due to the lower approximation order $N = 3$ used in the diffusive tests, which reduces the computational cost in each time step and makes the additional cost of the limiters significant.

5.2. Problem with nonuniform scattering/absorption. In this section, we test the proposed AP scheme on problems with nonuniform scattering and absorption cross-sections. The problem is a modification of the lattice problem formulated in [6] and [7], which is motivated by the geometry of an assembly in a nuclear reactor core. As in the original benchmark, a purely scattering medium with strongly absorbing mediums embedded as a checkerboard on a square spatial domain $[-1.5, 1.5] \times [-1.5, 1.5]$ is considered, as shown in Figure 3a. Here the strong absorption regions are colored in white with $\sigma_a = 9.9$ and $\sigma_s = 0.1$; the purely scattering regions are colored in black with $\sigma_a = 0$ and $\sigma_s = 1$. Unlike the original lattice benchmark, there is no source. Rather the initial condition, boundary condition, and all other specifics are identical to those used in the line source experiments in section 5.1. The computation is also run on a 150×150 uniform square mesh with the maximum time step allowed by (4.13) to final time $t_{\text{final}} = 1$ and $t_{\text{final}} = 0.1$ in the kinetic ($\epsilon = 1$) and diffusive ($\epsilon = 10^{-3}$) regimes, respectively. For comparison, we compute a reference kinetic solution using the second-order kinetic scheme proposed in [19] with a high approximation order $N = 37$ on a finely discretized mesh. A reference diffusion solution is computed by solving the diffusion equation (3.23) with the 9-point finite difference scheme (3.26). The reference kinetic and diffusion solutions are shown in Figures 3b and 3c, respectively.



(a) Problem layout. (b) Reference solution $\epsilon = 1$. (c) Reference solution $\epsilon = 10^{-3}$.

FIG. 3. *Problem with nonuniform scattering/absorption cross-sections.*

The run time and relative L^2 spatial errors (as defined in (5.2)) of the proposed scheme with various limiters are reported in Table 2. In the kinetic regime ($\epsilon = 1$),

the solution from the `ls-pw` limiter is still much less accurate than the other solutions. The `opt-r` limiter gives a more accurate solution than the other three limiters, including the expensive `opt-pw` limiter. In the diffusive regime ($\epsilon = 10^{-3}$), all limiters give identical solutions since conditions (C1)–(C6) and the pointwise positivity condition are always satisfied. Similar to the line source results, the difference in the computational time of the limited and unlimited solutions is more significant in the diffusive regime due to the lower approximation order $N = 3$. Here the computed solutions have relatively large errors in the diffusive regime when compared to the computed diffusion solutions in the line source case. It follows from (3.15) that smaller $\sigma_{\mathrm{s}}^{\min}$ leads to stronger artificial dissipation. Thus, we suspect that the stronger artificial dissipation introduced in this nonuniform problem ($\sigma_{\mathrm{s}}^{\min} = 0.1$) leads to diffusion solutions that are less accurate than those in the line source case ($\sigma_{\mathrm{s}}^{\min} = 1$).

To confirm that the computed solutions actually converge to the reference diffusion solution as $\epsilon \to 0$, we sample ϵ from 10^{-3} to 10^{-7} and report the L^2 spatial error E and its convergence order ν at each sample of ϵ. Let ϵ_i denote the samples of ϵ; the order ν is computed by $\nu := \log\left(\frac{E_{\epsilon_i}}{E_{\epsilon_{i+1}}}\right) \log\left(\frac{\epsilon_i}{\epsilon_{i+1}}\right)^{-1}$, with E_{ϵ_i} the L^2 spatial error when $\epsilon = \epsilon_i$. The convergence result is reported in Table 3, which shows first-order convergence of the spatial error. Since all limiters are effectively inactive when ϵ is small, we report only one set of the L^2 spatial errors in Table 3.

TABLE 2

Run time (sec) and relative L^2 spatial errors for the computed solutions without limiter and with the four limiters on the problem with nonuniform scattering/absorption cross-sections in the kinetic regime ($\epsilon = 1$, $t_{\mathrm{final}} = 1$) and diffusive regime ($\epsilon = 10^{-3}$, $t_{\mathrm{final}} = 0.1$). In the diffusive regime, the positivity conditions are never violated due to the smoothness of the solution. Thus, all limiters give solutions that are identical to the one without limiter.

Limiter		`none`	`ls-r`	`opt-r`	`ls-pw`	`opt-pw`
Kinetic ($N = 11$)	run time	363	391	406	483	24263
	E	0.092	0.090	0.083	0.501	0.132
Diffusive ($N = 3$)	run time	1481	11718	11922	11754	11662
	E	0.218	0.218	0.218	0.218	0.218

TABLE 3

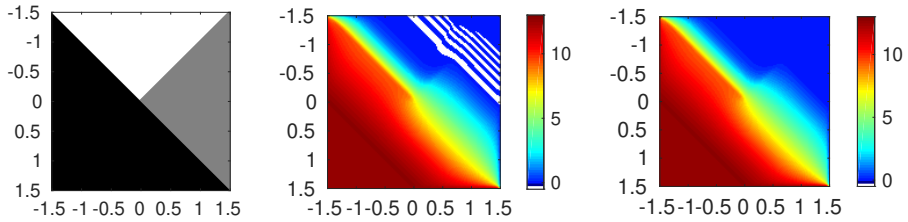
Convergence of the L^2 spatial error as $\epsilon \to 0$ in the diffusive regime at $t_{\mathrm{final}} = 0.1$. The errors E and the convergence orders ν are reported for ϵ sampled between 10^{-3} and 10^{-7}.

ϵ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
E	2.2e-1	1.9e-1	3.3e-2	3.9e-3	4.1e-4
ν	—	0.06	0.72	0.93	0.98

5.3. Riemann problem with nonuniform scattering. In this section, we test the proposed AP scheme on a Riemann problem with nonuniform scattering cross-section. The layout of the Riemann problem on a square spatial domain $[-1.5, 1.5] \times [-1.5, 1.5]$ is shown in Figure 4a, where the black and gray regions represent medium with $\sigma_{\mathrm{s}} = 1$, and the white region represents medium with $\sigma_{\mathrm{s}} = 10$. There is no absorption in the problem; i.e., $\sigma_{\mathrm{a}} = 0$ in all regions. The initial and boundary conditions are given by

(5.3)
$$f^{\mathrm{in}}(r,\Omega) = \begin{cases} 1, & r \text{ in black region,} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f(r,\Omega,t) = \begin{cases} 1, & x = -1.5 \text{ or } y = 1.5, \\ 0, & x = 1.5 \quad \text{or } y = -1.5. \end{cases}$$

All other specifics are still identical to those used in the line source experiments in section 5.1. The computation is also run on a 150×150 uniform square mesh with the maximum time step allowed by (4.13) to final time $t_{\text{final}} = 1$ and $t_{\text{final}} = 0.1$ in the kinetic ($\epsilon = 1$) and diffusive ($\epsilon = 10^{-3}$) regimes, respectively. For comparison, we compute a reference kinetic solution using the second-order kinetic scheme proposed in [19] with a high approximation order $N = 21$ on a finely discretized mesh. A reference diffusion solution is computed by solving the diffusion equation (3.23) with the 9-point finite difference scheme (3.26).



(a) Problem layout. (b) No limiter, $\epsilon = 1$. (c) `ls-r`, $\epsilon = 1$.

FIG. 4. *Problem layout and numerical solutions for the Riemann problem with nonuniform scattering cross-section. In* (b) *and* (c)*, heat maps show the particle concentration ρ generated without limiter and with the* `ls-r` *limiter in the kinetic ($\epsilon = 1$) regime at time $t_{\text{final}} = 1$. Negative particle concentrations are colored in white. The approximation order of the FP_N equations is $N = 11$.*

TABLE 4

Run time (sec) and relative L^2 spatial errors for the computed solutions without limiter and with the four limiters on the Riemann problem with nonuniform scattering cross-section in the kinetic regime ($\epsilon = 1$, $t_{\text{final}} = 1$) and diffusive regime ($\epsilon = 10^{-3}$, $t_{\text{final}} = 0.1$). In the diffusive regime, the positivity conditions are never violated due to the smoothness of the solution. Thus, all limiters give solutions that are identical to the one without limiter.

Limiter		`none`	`ls-r`	`opt-r`	`ls-pw`	`opt-pw`
Kinetic	run time	239	324	333	500	7655
($N = 11$)	E	0.043	0.042	0.042	0.043	0.043
Diffusive	run time	226	1580	1721	1685	1698
($N = 3$)	E	0.044	0.044	0.044	0.044	0.044

To illustrate the positivity-preserving property of the proposed scheme, we show the numerical solutions in the kinetic regime without limiter and with the `ls-r` limiter in Figures 4b and 4c, respectively. The solutions with other limiters are similar to the one shown in Figure 4c. We observe that the negative particle concentrations (colored in white) are removed when the limiters are used. Also, it is visible from the solutions that the particles move faster in the lower scattering regions (black and gray) than they do in the higher scattering region (white).

The run time and relative L^2 spatial errors (defined in (5.2)) of the proposed scheme with various limiters are reported in Table 4. In the kinetic regime ($\epsilon = 1$), the realizability limiters give slightly more accurate solutions than those from the pointwise limiters. The `opt-pw` limiter is still the most expensive one. In the diffusive regime ($\epsilon = 10^{-3}$), all limiters give reasonably accurate solutions. These solutions are all identical since the smooth solution never activated the limiters in the diffusive regime. Similar to the previous results, the lower approximation order in the diffusive regime makes the limited solutions relatively expensive.

5.4. Space-time accuracy. As a final test, we investigate the order of accuracy of the proposed AP scheme in the kinetic ($\epsilon = 1$), transition ($\epsilon = 0.05$), and diffusive ($\epsilon = 10^{-5}$)[3] regimes. As in previous tests, we truncate the spatial domain to a $[-1.5, 1.5] \times [-1.5, 1.5]$ square centered at the origin and impose an artificial zero boundary condition. The computation is run on uniform square meshes of size 20×20 to 1280×1280 with the maximum time step allowed by (4.13). The final time is $t_{\text{final}} = 1$, $t_{\text{final}} = 0.05$, and $t_{\text{final}} = 0.01$ in the kinetic, transition, and diffusive regimes, respectively. Since the steep Gaussian initial condition (5.1) may limit the observed convergence order, we test the scheme with a "gradual" Gaussian initial condition, which takes the same form as (5.1) with variance $\varsigma^2 = 5 \times 10^{-3}$.

All parameter values used in the space-time convergence tests are identical to those listed in section 5.1, except that we choose the approximation order $N = 5$ (instead of 11) for the kinetic and transition tests and $N = 3$ for the diffusive tests. Similar to (5.2), we define the relative L^2 space-time error E_h by

$$E_{h_i} := \|\rho_{h_i} - \rho_{h_{i+1}}\|_{L^2_{h_{i+1}}(\mathbb{R}^2)} / \|\rho_{h_{i+1}}\|_{L^2_{h_{i+1}}(\mathbb{R}^2)}, \quad i = 1, \dots, 6, \tag{5.4}$$

where ρ_h is the particle concentration of the solution computed by the proposed scheme with spatial grid size $h = \Delta x = \Delta y$, and $h_i = \frac{3}{2^i \times 10}$, $i = 1, \dots, 7$, denote the grid sizes associated to the mesh sizes given in Table 5.

Table 5 reports the L^2 space-time errors E_h and observed convergence orders ν for the proposed scheme with `ls-r` and `opt-r` limiters in the three regimes. The order ν is computed by

$$\nu := \log\left(\frac{E_{h_i}}{E_{h_{i+1}}}\right) \log\left(\frac{h_i}{h_{i+1}}\right)^{-1}, \quad i = 1, \dots, 5. \tag{5.5}$$

The results in Table 5 show first-order space-time accuracy in the kinetic and transition regimes, which is to be expected since a first-order time discretization is used in the proposed scheme. In the diffusive regime, we observe second-order accuracy due to the refined time step. In the kinetic and transition regimes, we also noticed that the convergence orders are higher than one when the mesh sizes are between 80×80 and 320×320. We observe in some preliminary numerical results that with these mesh sizes, the first-order temporal error is relatively small when compared to the higher order spatial errors, in particular the artificial dissipation. We conjecture that this leads to the observed faster convergence rates. Finally, there is no noticeable difference between the results from the two limiters, since with the gradual Gaussian initial condition the limiters are rarely active.

6. Conclusions and discussion. We have proposed a new positive AP scheme for solving the FP_N equations, an approximation to the linear kinetic transport equations, in two space dimensions. The scheme applies a micro-macro decomposition to the FP_N equations and solves the resulting system with a suitable semi-implicit temporal discretization and a special finite difference spatial discretization. We give sufficient conditions under which the proposed scheme preserves positivity of particle concentrations, and we show that these sufficient conditions are satisfied under a reasonable time-step restriction with the imposition of realizability limiters. We test the proposed scheme on the notoriously difficult line source benchmark problem, a multiscale lattice problem, and a Riemann problem. Numerical results confirm that in both

[3]Here we choose $\epsilon = 10^{-5}$ for the diffusive regime to ensure that the problem stays in the diffusive regime even when the space-time mesh is refined.

TABLE 5

Convergence of space-time errors: The space-time errors E_h and observed convergence orders ν are reported. The spatial mesh sizes are listed in the first column. We observe at least first order in all three regimes, which confirms the theoretical estimate. There is no noticeable difference in the results from the two limiters.

	Kinetic, $\epsilon = 1$				Transition, $\epsilon = 0.05$				Diffusive, $\epsilon = 10^{-5}$			
	`ls-r`		`opt-r`		`ls-r`		`opt-r`		`ls-r`		`opt-r`	
Mesh	E_h	ν	E_h	ν	E_h	ν	E_h	ν	E_h	ν	E_h	ν
20^2	1.2e-1	—	1.2e-1	—	4.9e-1	—	4.9e-1	—	3.4e-1	—	3.4e-1	—
40^2	1.0e-1	0.2	1.0e-1	0.2	4.0e-1	0.3	4.0e-1	0.3	9.8e-2	1.8	9.8e-2	1.8
80^2	5.2e-2	1.0	5.2e-2	1.0	1.3e-1	1.6	1.3e-1	1.6	2.4e-2	2.0	2.4e-2	2.0
160^2	1.3e-2	2.0	1.3e-2	2.0	2.1e-2	2.7	2.1e-2	2.7	5.7e-3	2.1	5.7e-3	2.1
320^2	3.8e-3	1.8	3.8e-3	1.8	5.4e-3	2.0	5.4e-3	2.0	1.3e-3	2.1	1.3e-3	2.1
640^2	1.9e-3	1.0	1.9e-3	1.0	2.4e-3	1.2	2.4e-3	1.2	2.1e-4	2.7	2.1e-4	2.7

the kinetic (large mean-free-path) and diffusive (small mean-free-path) regimes, the scheme gives accurate solutions and preserves the nonnegativity of particle concentrations. The space-time convergence result shows that the accuracy of the proposed scheme is first order in the kinetic and transition regimes and second order in the diffusive regime.

The uniform stability and accuracy analysis of the proposed scheme is presented in [45] for the one-dimensional case. The analysis indicates that the accuracy of this scheme is limited by the first-order semi-implicit temporal discretization. To achieve higher order of accuracy, it is possible to implement the proposed finite difference method and the realizability limiters together with a second-order implicit-explicit (IMEX) temporal discretization, which has been considered for stiff ordinary differential equations [11] and the stiff Bhatnagar–Gross–Krook (BGK) equation [35]. However, it is known that IMEX schemes require a restrictive time step either to preserve positivity of the solution [34, 35] or to maintain the strong-stability-preserving property for the implicit update [12]. It is also not clear if IMEX schemes resolve the diffusion limit correctly, since only the Euler limit is considered in [35]. On the other hand, the discontinuous Galerkin (DG) spatial discretization has been used together with the micro-macro decomposition to develop high-order AP schemes for the linear kinetic transport equations [36, 37] and the BGK equation [72]. To develop a higher-order positive-preserving AP scheme, it is also possible to apply some modified realizability limiters to these schemes to enforce positivity. However, the derivation of physical positivity conditions under the DG spatial discretization is not straightforward.

Other potential future work includes the following: a rigorous stability and accuracy analysis in the multidimensional case for the proposed scheme; an extension of the proposed scheme to three dimensions, where we believe the naive extension suffers from the issue of alternating grids, and thus a modified extension, such as the one described in section 4.4, is needed; the application of the proposed scheme to other equations, such as the Vlasov–Poisson equation and the linear Boltzmann equation, where the multiscale behavior and the positivity of the solution are of interest.

Appendix A. Calculation of diffusion matrices. In this appendix, we provide detailed calculations in the derivation of diffusion matrices $\widetilde{\mathbf{Q}}$ and $\overline{Q}$ in (3.8) and (3.9). We first write the submatrices of $\widetilde{\mathbf{Q}}$ in (3.8) as $\langle \overline{m}\widetilde{\mathbf{m}}^T\Omega_\alpha\rangle\widetilde{\Gamma}\langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_\beta\rangle$ for $\alpha = x, y$ and $\beta = x, y$. For $\alpha = x, y$, let k_α denote the index such that $c_\alpha\widetilde{\mathbf{m}}_{k_\alpha} = \Omega_\alpha$

with some nonzero constant $c_\alpha \in \mathbb{R}$. We then have

(A.1)
$$\begin{aligned}\langle \overline{m}\widetilde{\mathbf{m}}^T\Omega_\alpha\rangle\widetilde{\Gamma}\langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_\beta\rangle &= \overline{m}\langle\widetilde{\mathbf{m}}^T c_\alpha\widetilde{\mathbf{m}}_{k_\alpha}\rangle\widetilde{\Gamma}\langle\widetilde{\mathbf{m}}\widetilde{\mathbf{m}}^T\Omega_\beta\rangle = \overline{m}c_\alpha\widetilde{\Gamma}_{k_\alpha,k_\alpha}\langle\widetilde{\mathbf{m}}_{k_\alpha}\widetilde{\mathbf{m}}^T\Omega_\beta\rangle \\ &= \widetilde{\Gamma}_{k_\alpha,k_\alpha}\langle\overline{m}\widetilde{\mathbf{m}}^T c_\alpha\widetilde{\mathbf{m}}_{k_\alpha}\Omega_\beta\rangle = \widetilde{\Gamma}_{k_\alpha,k_\alpha}\langle\overline{m}\widetilde{\mathbf{m}}^T\Omega_\alpha\Omega_\beta\rangle\,,\end{aligned}$$

where the second equality follows from the fact that since entries of $\widetilde{\mathbf{m}}$ are orthonormal, $\langle\widetilde{\mathbf{m}}^T\widetilde{\mathbf{m}}_{k_\alpha}\rangle$ is a vector of all zeros except its k_αth entry, which takes value one. Further, we observe that $\widetilde{\Gamma}_{k_x,k_x} = \widetilde{\Gamma}_{k_y,k_y}$, which follows directly from the definition of $\widetilde{\Gamma}$ in (3.3) and the definition of the filtering matrix F introduced in (2.8). Thus, we denote $\widetilde{\gamma} := \widetilde{\Gamma}_{k_x,k_x} = \widetilde{\Gamma}_{k_y,k_y}$ in (3.8) and (3.9). The equalities in (3.8) are now verified, and the equalities in (3.9) can be shown similarly.

Appendix B. Lower bound of the artificial dissipation operator. In this appendix, we prove a lower bound needed in (4.4) in the proof of Theorem 4.1. Specifically we show that for any nonnegative function w defined on the spatial mesh,

$$\delta_x^4(w_{i,j}) \geq \frac{1}{2\Delta x^4}\left(\frac{1}{\Theta}w_{i+1,j} - 4w_{i,j} + \frac{1}{\Theta}w_{i-1,j}\right)\,, \tag{B.1}$$

where δ_x^4 is the artificial dissipation operator defined in (3.12) and $\Theta = \frac{1}{2-\theta}$. From (3.12) and (3.13),

$$\Delta x^4\delta_x^4(w_{i,j}) = w_{i+1,j} - \frac{\Delta x}{2}s^x_{i+1,j} - 2w_{i,j} + w_{i-1,j} + \frac{\Delta x}{2}s^x_{i-1,j}\,, \tag{B.2}$$

where the slope $s^x_{i,j}$ is defined in (3.14). To verify (B.1), we first consider the case that $w_{i+1,j} \geq w_{i,j} \geq w_{i-1,j}$, which, together with (3.14), leads to

$$s^x_{i+1,j} \leq \theta\frac{w_{i+1,j} - w_{i,j}}{\Delta x} \quad \text{and} \quad s^x_{i-1,j} \geq 0\,. \tag{B.3}$$

Thus, (B.2) gives that when $w_{i+1,j} \geq w_{i,j} \geq w_{i-1,j}$,

$$\Delta x^4\delta_x^4(w_{i,j}) \geq \left(1 - \frac{\theta}{2}\right)w_{i+1,j} - \left(2 - \frac{\theta}{2}\right)w_{i,j} + w_{i-1,j}\,. \tag{B.4}$$

By applying similar arguments on other cases, we show that $\Delta x^4\delta_x^4(w_{i,j})$ is bounded from below by

(B.5)
$$\begin{cases}\left(1-\frac{\theta}{2}\right)w_{i+1,j} - \left(2-\frac{\theta}{2}\right)w_{i,j} + w_{i-1,j} & \text{if } w_{i+1,j} \geq w_{i,j} \geq w_{i-1,j}\,, \\ \left(1-\frac{\theta}{2}\right)w_{i+1,j} - (2-\theta)\,w_{i,j} + \left(1-\frac{\theta}{2}\right)w_{i-1,j} & \text{if } w_{i+1,j} \geq w_{i,j}\,, w_{i,j} < w_{i-1,j}\,, \\ w_{i+1,j} - 2w_{i,j} + w_{i-1,j} & \text{if } w_{i+1,j} < w_{i,j}\,, w_{i,j} \geq w_{i-1,j}\,, \\ w_{i+1,j} - \left(2-\frac{\theta}{2}\right)w_{i,j} + \left(1-\frac{\theta}{2}\right)w_{i-1,j} & \text{if } w_{i+1,j} < w_{i,j} < w_{i-1,j}\,.\end{cases}$$

Since $w_{i+1,j}$, $w_{i,j}$, and $w_{i-1,j}$ are all nonnegative and $\theta \in (1,2)$, (B.1) holds.

REFERENCES

[1] M. L. Adams, *Discontinuous finite element transport solutions in thick diffusive problems*, Nucl. Sci. Eng., 137 (2001), pp. 298–333.

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

[3] G. W. Alldredge, C. D. Hauck, and A. L. Tits, *High-order entropy-based closures for linear transport in slab geometry* II: *A computational study of the optimization problem*, SIAM J. Sci. Comput., 34 (2012), pp. B361–B391, https://doi.org/10.1137/11084772X.

[4] S. Boscarino, L. Pareschi, and G. Russo, *Implicit-explicit Runge–Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit*, SIAM J. Sci. Comput., 35 (2013), pp. A22–A51, https://doi.org/10.1137/110842855.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[6] T. A. Brunner, *Forms of Approximate Radiation Transport*, Technical Report SAND2002-1778, Sandia National Laboratories, Albuquerque, NM, 2002.

[7] T. A. Brunner and J. P. Holloway, *Two-dimensional time-dependent Riemann solvers for neutron transport*, J. Comput. Phys., 210 (2005), pp. 386–399.

[8] C. Buet, B. Després, and E. Franck, *Design of asymptotic preserving finite volume schemes for the hyperbolic heat equation on unstructured meshes*, Numer. Math., 122 (2012), pp. 227–278.

[9] K. Case and P. Zweifel, *Linear Transport Theory*, Addison-Wesley, Reading, MA, 1967.

[10] C. Cercignani, *The Boltzmann equation*, in The Boltzmann Equation and Its Applications, Springer, 1988, pp. 40–103.

[11] A. Chertock, S. Cui, A. Kurganov, and T. Wu, *Steady state and sign preserving semi-implicit Runge–Kutta methods for ODEs with stiff damping term*, SIAM J. Numer. Anal., 53 (2015), pp. 2008–2029, https://doi.org/10.1137/151005798.

[12] S. Conde, S. Gottlieb, Z. J. Grant, and J. N. Shadid, *Implicit and implicit–explicit strong stability preserving Runge–Kutta methods with high linear order*, J. Sci. Comput., 73 (2017), pp. 667–690.

[13] R. E. Curto, *Recursiveness, positivity and truncated moment problems*, Houston J. Math., 17 (1991), pp. 603–635.

[14] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Volume* 6: *Evolution Problems.* II, Springer-Verlag, Berlin, 2000.

[15] B. Davison, *Neutron Transport Theory*, Oxford University Press, London, 1973.

[16] S. M. Deshpande, *Kinetic theory based new upwind methods for inviscid compressible flows*, in 24th Aerospace Sciences Meeting, American Institute of Aeronautics and Astronautics, New York, 1986, paper 86-0275.

[17] M. Frank, C. Hauck, and K. Küpper, *Convergence of filtered spherical harmonic equations for radiation transport*, Commun. Math. Sci., 14 (2016), pp. 1443–1465.

[18] B. D. Ganapol, *Homogeneous Infinite Media Time-dependent Analytic Benchmarks for X-TM Transport Methods Development*, Technical report, Los Alamos National Laboratory, Los Alamos, NM, 1999.

[19] C. K. Garrett and C. D. Hauck, *A comparison of moment closures for linear kinetic transport equations: The line source benchmark*, Transport Theory Statist. Phys., 42 (2013), pp. 203–235.

[20] L. Gosse, *Transient radiative transfer in the grey case: Well-balanced and asymptotic-preserving schemes built on Case's elementary solutions*, J. Quant. Spectrosc. Radiat. Transfer, 112 (2011), pp. 1995–2012.

[21] L. Gosse, *Well-balanced schemes using elementary solutions for linear models of the Boltzmann equation in one space dimension*, Kinet. Relat. Models, 5 (2012), pp. 283–323.

[22] L. Gosse and G. Toscani, *An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 337–342.

[23] L. Gosse and G. Toscani, *Asymptotic-preserving & well-balanced schemes for radiative transfer and the Rosseland approximation*, Numer. Math., 98 (2004), pp. 223–250.

[24] D. Gottlieb, S. Gottlieb, and J. Hesthaven, *Spectral Methods for Time-Dependent Problems*, Cambridge University Press, New York, 2007.

[25] J.-L. 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, https://doi.org/10.1137/090746938.

[26] S. Günter, Q. Yu, J. Krüger, and K. Lackner, *Modelling of heat transport in magnetised plasmas using non-aligned coordinates*, J. Comput. Phys., 209 (2005), pp. 354–370.

[27] B. Guo, *Spectral Methods and Their Applications*, World Scientific, Singapore, 1998.

[28] J. R. Haack and C. D. Hauck, *Oscillatory behavior of asymptotic-preserving splitting methods for a linear model of diffusive relaxation*, Kinet. Relat. Models, 1 (2008), pp. 573–590.

[29] G. J. Habetler and B. J. Matkowsky, *Uniform asymptotic expansions in transport theory with small mean free paths, and the diffusion approximation*, J. Math. Phys., 16 (1975), pp. 846–854.

[30] C. D. Hauck and R. B. Lowrie, *Temporal regularization of the P_N equations*, Multiscale Model. Simul., 7 (2009), pp. 1497–1524, https://doi.org/10.1137/07071024X.
[31] C. D. Hauck and R. G. McClarren, *Positive P_N closures*, SIAM J. Sci. Comput., 32 (2010), pp. 2603–2626, https://doi.org/10.1137/090764918.
[32] C. D. Hauck, R. B. Lowrie, and R. G. McClarren, *Methods for diffusive relaxation in the P_N equations*, in Numerical Methods for Balance Laws, Quad. Mat. 24, G. Puppo and G. Russo, eds., Dipartimento di Matematica, Seconda Università di Napoli, Caserta, Italy, 2010, pp. 197–243.
[33] R. D. Hazeltine and F. L. Waelbroeck, *The Framework of Plasma Physics*, CRC Press, Boca Raton, FL, 2018.
[34] I. Higueras and T. Roldán, *Positivity-preserving and entropy-decaying IMEX methods*, in Ninth International Conference Zaragoza-Pau on Applied Mathematics and Statistics, Monogr. Semin. Mat. García Galdeano 33, Prensas Universitarias de Zaragoza, Zaragoza, Spain, 2006, pp. 129–136.
[35] J. Hu, R. Shu, and X. Zhang, *Asymptotic-preserving and positivity-preserving implicit-explicit schemes for the stiff BGK equation*, SIAM J. Numer. Anal., 56 (2018), pp. 942–973, https://doi.org/10.1137/17M1144362.
[36] J. Jang, F. Li, J.-M. Qiu, and T. Xiong, *Analysis of asymptotic preserving DG-IMEX schemes for linear kinetic transport equations in a diffusive scaling*, SIAM J. Numer. Anal., 52 (2014), pp. 2048–2072, https://doi.org/10.1137/130938955.
[37] J. Jang, F. Li, J.-M. Qiu, and T. Xiong, *High order asymptotic preserving DG-IMEX schemes for discrete-velocity kinetic equations in a diffusive scaling*, J. Comput. Phys., 281 (2015), pp. 199–224.
[38] S. Jin, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM J. Sci. Comput., 21 (1999), pp. 441–454, https://doi.org/10.1137/S1064827598334599.
[39] S. Jin, *Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: A review*, Riv. Mat. Univ. Parma. (N.S.), 3 (2012), pp. 177–216.
[40] S. Jin and C. D. Levermore, *Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation*, J. Comput. Phys., 126 (1996), pp. 449–467.
[41] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439, https://doi.org/10.1137/S0036142997315962.
[42] S. Jin, L. Pareschi, and G. Toscani, *Uniformly accurate diffusive relaxation schemes for multiscale transport equations*, SIAM J. Numer. Anal., 38 (2000), pp. 913–936, https://doi.org/10.1137/S0036142998347978.
[43] M. Junk, *Maximum entropy for reduced moment problems*, Math. Models Methods Appl. Sci., 10 (2000), pp. 1001–1025.
[44] D. S. Kershaw, *Flux Limiting Nature's Own Way—A New Method for Numerical Solution of the Transport Equation*, Technical report UCRL-78378, Lawrence Livermore National Laboratory, Livermore, CA, 1976.
[45] M. P. Laiu and C. D. Hauck, *Analysis of a positive asymptotic preserving scheme for linear kinetic transport equations*, in preparation.
[46] M. P. Laiu and C. D. Hauck, *Positivity limiters for filtered spectral approximations of linear kinetic transport equations*, J. Sci. Comput., 78 (2019), pp. 918–950, https://doi.org/10.1007/s10915-018-0790-y.
[47] M. P. Laiu, C. D. Hauck, R. G. McClarren, D. P. O'Leary, and A. L. Tits, *Positive filtered P_N moment closures for linear kinetic equations*, SIAM J. Numer. Anal., 54 (2016), pp. 3214–3238, https://doi.org/10.1137/15M1052871.
[48] M. P. Laiu and A. L. Tits, *A constraint-reduced MPC algorithm for convex quadratic programming, with a modified active set identification scheme*, Comput. Optim. Appl., 72 (2019), pp. 727–768, https://doi.org/10.1007/s10589-019-00058-0.
[49] 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.
[50] E. W. Larsen and J. Morel, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes* II, J. Comput. Phys., 83 (1989), pp. 212–236.
[51] E. W. Larsen, J. Morel, and W. F. Miller, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes*, J. Comput. Phys., 69 (1987), pp. 283–324.
[52] E. W. Larsen and J. E. Morel, *Advances in discrete-ordinates methodology*, in Nuclear Computational Science, Springer, 2010, pp. 1–84.
[53] M. Lemou and L. Mieussens, *A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit*, SIAM J. Sci. Comput., 31 (2008), pp. 334–368, https://doi.org/10.1137/07069479X.

[54] R. LeVeque, *Numerical Methods for Conservation Laws*, Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 1992.
[55] C. D. Levermore, W. J. Morokoff, and B. T. Nadiga, *Moment realizability and the validity of the Navier-Stokes equations for rarefied gas dynamics*, Phys. Fluids, 10 (1998), pp. 3214–3226.
[56] E. E. Lewis and W. F. Miller, Jr., *Computational Methods of Neutron Transport*, John Wiley and Sons, New York, 1984.
[57] J.-G. Liu and L. Mieussens, *Analysis of an asymptotic preserving scheme for linear kinetic equations in the diffusion limit*, SIAM J. Numer. Anal., 48 (2010), pp. 1474–1491, https://doi.org/10.1137/090772770.
[58] T.-P. Liu and S.-H. Yu, *Boltzmann equation: Micro-macro decompositions and positivity of shock profiles*, Comm. Math. Phys., 246 (2004), pp. 133–179.
[59] X.-D. Liu and S. Osher, *Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes* I, SIAM J. Numer. Anal., 33 (1996), pp. 760–779, https://doi.org/10.1137/0733038.
[60] R. G. McClarren and C. D. Hauck, *Robust and accurate filtered spherical harmonics expansions for radiative transfer*, J. Comput. Phys., 229 (2010), pp. 5597–5614.
[61] R. G. McClarren and R. B. Lowrie, *The effects of slope limiting on asymptotic-preserving numerical methods for hyperbolic conservation laws*, J. Comput. Phys., 227 (2008), pp. 9711–9726.
[62] L. Mieussens, *On the asymptotic preserving property of the unified gas kinetic scheme for the diffusion limit of linear kinetic models*, J. Comput. Phys., 253 (2013), pp. 138–156.
[63] D. Mihalis and B. Weibel-Mihalis, *Foundations of Radiation Hydrodynamics*, Dover, Mineola, NY, 1999.
[64] W. F. Miller, *An analysis of the finite differenced, even-parity, discrete ordinates equations in slab geometry*, Nucl. Sci. Eng., 108 (1991), pp. 247–266.
[65] B. Perthame, *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM J. Numer. Anal., 29 (1992), pp. 1–19, https://doi.org/10.1137/0729001.
[66] G. C. Pomraning, *Variational boundary conditions for the spherical harmonics approximation to the neutron transport equation*, Ann. Phys., 27 (1964), pp. 193–215.
[67] G. C. Pomraning, *Radiation Hydrodynamics*, Pergamon Press, New York, 1973.
[68] D. Radice, E. Abdikamalov, L. Rezzolla, and C. D. Ott, *A new spherical harmonics scheme for multi-dimensional radiation transport* I: *Static matter configurations*, J. Comput. Phys., 242 (2013), pp. 648–669.
[69] B. Seibold and M. Frank, *StaRMAP—A second order staggered grid method for spherical harmonics moment equations of radiative transfer*, ACM Trans. Math. Software, 41 (2014), 4.
[70] P. Sharma and G. W. Hammett, *Preserving monotonicity in anisotropic diffusion*, J. Comput. Phys., 227 (2007), pp. 123–142.
[71] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, American Mathematical Society, New York, 1943.
[72] T. Xiong, J. Jang, F. Li, and J.-M. Qiu, *High order asymptotic preserving nodal discontinuous Galerkin IMEX schemes for the BGK equation*, J. Comput. Phys., 284 (2015), pp. 70–94.
[73] K. Xu and J.-C. Huang, *A unified gas-kinetic scheme for continuum and rarefied flows*, J. Comput. Phys., 229 (2010), pp. 7747–7764.
[74] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), pp. 3091–3120.
[75] X. Zhang and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: Survey and new developments*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 467 (2011), pp. 2752–2776.