Flux-Corrected Transport Techniques Applied to the Radiation Transport Equation Discretized with Continuous Finite Elements

Joshua E. Hansel[a], Jean C. Ragusa[a]

[a]*Department of Nuclear Engineering, Texas A&M University, College Station, TX 77840*

Abstract

The Flux-Corrected Transport (FCT) algorithm is applied to the unsteady and steady-state particle transport equation. The proposed FCT method employs the following: (1) a low-order, positivity-preserving scheme, based on the application of M-matrix properties, (2) a high-order scheme based on the entropy viscosity method introduced by Guermond [1], and (3) local, discrete solution bounds derived from the integral transport equation. The resulting scheme is second-order accurate in space, enforces an entropy inequality, mitigates the formation of spurious oscillations, and guarantees the absence of negativities. Space discretization is achieved using continuous finite elements. Time discretizations for unsteady problems include theta schemes such as explicit and implicit Euler, and strong-stability preserving Runge-Kutta (SSPRK) methods. The developed FCT scheme is shown to be robust with explicit time discretizations but may require damping in the nonlinear iterations for steady-state and implicit time discretizations.

Keywords: entropy viscosity, FCT, particle transport equation

1. Introduction

The radiation transport equation, or linear Boltzmann equation, describes the transport of particles interacting with a background medium [2]. Some of its applications include the modeling of nuclear reactors, radiation therapy, astrophysical applications, radiation shielding, and high energy density physics [2, 3, 4, 5, 6]. This paper focuses on solution techniques applicable to the first-order form of the transport equation in Cartesian geometries, recalled below in Equation (1). The transport equation is a particle balance statement in a six-dimensional phase-space volume where $\mathbf{x}$ denotes the particle's position, $\mathbf{\Omega}$

Email addresses: joshua.hansel89@gmail.com (Joshua E. Hansel), jean.ragusa@tamu.edu (Jean C. Ragusa)

its direction of flight, and E its energy:

$$\frac{1}{v(E)}\frac{\partial \psi}{\partial t} + \mathbf{\Omega}\cdot\nabla\psi(\mathbf{x},\mathbf{\Omega},E,t) + \Sigma_{\mathrm{t}}(\mathbf{x},E,t)\psi(\mathbf{x},\mathbf{\Omega},E,t) = Q_{\mathrm{tot}}(\mathbf{x},\mathbf{\Omega},E,t)\,. \quad (1)$$

$Q_{\mathrm{tot}}(\mathbf{x},\mathbf{\Omega},E,t)$ denotes the total particle source gains in an infinitesimal phase-space volume due to particle scattering, extraneous source of particles (if any), and fission sources (in the case of neutron transport in multiplying media):

$$Q_{\mathrm{tot}}(\mathbf{x},\mathbf{\Omega},E,t) \equiv Q_{\mathrm{sca}}(\mathbf{x},\mathbf{\Omega},E,t) + Q_{\mathrm{ext}}(\mathbf{x},\mathbf{\Omega},E,t) + Q_{\mathrm{fis}}(\mathbf{x},\mathbf{\Omega},E,t)\,. \quad (2)$$

The source terms Q_{sca} and Q_{fis} linearly depend on the solution variable, the angular flux, denoted by ψ. Only simple configurations are amenable to an analytical solution of Equation (1). In most cases of relevance, the transport equation must be solved numerically; ~~this~~ transport calculations fall under two main categories: stochastic calculations and deterministic calculations. The former category is referred to as Monte Carlo and relies on sampling large numbers of particle histories using random number generators [2], and the latter involves discretization of the phase-space and the use of iterative techniques. This work applies to the latter category. One common angular discretization is the discrete-ordinate or S_N method [2, 5, 7]; it is a collocation method in angle whereby the transport equation is solved only along discrete directions $\mathbf{\Omega}_d$ ($1 \leq d \leq n_{\mathbf{\Omega}}$, with $n_{\mathbf{\Omega}}$ the total number of discrete directions). One of the main advantages of the S_N technique is that it enables an iterative approach, called source iteration in the transport literature [2, 5, 7], to resolve both the particle's streaming and interaction processes and the scattering events as follows:

$$\frac{1}{v}\frac{\partial \psi_d^{(\ell)}}{\partial t} + \mathbf{\Omega}_d\cdot\nabla\psi_d^{(\ell)} + \Sigma_{\mathrm{t}}\psi_d^{(\ell)} = Q_{\mathrm{tot},d}^{(\ell-1)} \quad \forall d \in [1, n_{\mathbf{\Omega}}]\,, \quad (3)$$

where ℓ is the iteration index and $\psi_d^{(\ell)}(\mathbf{x},E,t) = \psi^{(\ell)}(\mathbf{x},\mathbf{\Omega}_d,E,t)$. Hence, a system of $n_{\mathbf{\Omega}}$ decoupled equations are to be solved at a given source iteration index ℓ. For curvilinear geometries, an angular derivative term is present at iteration ℓ, and thus the equations are not decoupled; in this case, the scalar FCT methodology discussed in this work requires amendment. This allows solution techniques for scalar conservation laws to be leveraged in solving ~~Equations~~the system given by Equation (3). For brevity, the discrete-ordinate subscript d will be omitted hereafter and our model transport equation will consist in one of the $n_{\mathbf{\Omega}}$ transport equations for a given fixed source (right-hand side).

~~Traditionally, the~~ A common spatial discretization method for the S_N equations has been a Discontinuous Galerkin finite element method (DGFEM) with upwinding [8, 9]. Here, however, a Continuous Galerkin finite element method (CGFEM) is applied. Some recent work by Guermond and Popov [1] on solution techniques for conservation laws with CGFEM addresses some of the main disadvantages of CGFEM versus DGFEM, including the formation of spurious oscillations. The purpose of the present paper is to demonstrate a proof of concept for the application of such solution techniques to the transport equation.

Furthermore, some or all of the methodology explored in this paper can be later extended to DGFEM as well; see, for instance, Zingan et al. [10] where the techniques proposed by Guermond and Popov [1] have been ported to DGFEM schemes for Euler equations.

One of the main objectives of this paper is to present a method that precludes the formation of spurious oscillations and the negativities that result from these oscillations; these issues have been a long-standing issue in the numerical solution of the transport equation [11]. Not only are these negativities physically incorrect (a particle's distribution density must be non-negative), but they can cause simulations to terminate prematurely, for example in radiative transfer where the radiation field is nonlinearly coupled to a material energy equation. Many attempts to remedy the negativities in transport solutions rely on ad-hoc fix-ups, such as the set-to-zero fix-up for the classic diamond difference scheme [5]. Recent work by Hamilton introduced a similar fix-up for the linear discontinuous finite element method (LDFEM) that conserves local balance and preserves third-order accuracy [12]. Walters and Wareing developed characteristic methods [13], but Wareing later notes that these characteristic methods are difficult to implement and offers a nonlinear positive spatial differencing scheme known as the exponential discontinuous scheme [14]. Maginot has recently developed a consistent set-to-zero (CSZ) LDFEM method [15], as well as a non-negative method for bilinear discontinuous FEM [16][17].

In fluid dynamics, ~~tradiational~~ traditional approaches to remedy the issue of spurious oscillations include the flux-corrected transport (FCT) algorithm, introduced in 1973 as the SHASTA algorithm for finite difference discretizations by Boris and Book [18], where it was applied to linear discontinuities and gas dynamic shock waves. To the best of our knowledge, these FCT techniques have not been applied to the particle transport equation. The main idea of the FCT algorithm is to blend a low-order scheme having desirable properties with a high-order scheme which may lack these properties. Zalesak improved methodology of the algorithm and introduced a fully multi-dimensional limiter [19]. Parrott and Christie extended the algorithm to the finite element method on unstructured grids [20], thus beginning the FEM-FCT methodology. Löhner et. al. applied FEM-FCT to the Euler and Navier-Stokes equations and began using FCT with complex geometries [21]. Kuzmin and Möller introduced an algebraic FCT approach for scalar conservation laws [22] and later introduced a general-purpose FCT scheme, which is designed to be applicable to both steady-state and transient problems [23]. In these FEM-FCT works and others [24, 25, 26], the high-order scheme used in the FCT algorithm was the Galerkin finite element method, but this work uses the entropy viscosity method developed by Guermond and others [1].

Recent work by Guermond and Popov addresses the issue of spurious oscillations for general conservation laws by using artificial dissipation based on local entropy production, a method known as entropy viscosity [1]. The idea of entropy viscosity is to enforce an entropy inequality on the weak solution, and thus filter out weak solutions containing spurious oscillations. However, entropy viscosity solutions may still contain spurious oscillations, albeit smaller in

magnitude, and consequently negativities are not precluded. To circumvent this deficiency, Guermond proposed using the entropy viscosity method in conjunction with the FCT algorithm [27]; the high-order scheme component in FCT, traditionally the unmodified Galerkin scheme, is replaced with the entropy viscosity scheme. For the low-order scheme, Guermond also introduced a discrete maximum principle (DMP) preserving (and positivity-preserving) scheme for scalar conservation laws [28].

The algorithm described in this paper takes a similar approach to the algorithm described in the work by Guermond and Popov for scalar conservation laws, but is extended to allow application to the transport equation, which does not fit the precise definition of a conservation law but is instead a balance law since it includes sinks and sources, namely the reaction term $\Sigma_t \psi$ and the source term Q_{tot}. The presence of these terms is also a novelty in the context of the FCT algorithm. In addition, much of the work on FCT to date has been for fully explicit time discretizations. Because speeds in radiation transport (such as the speed of light in the case of photons) are so large, implicit and steady-state time discretization are important considerations, given the CFL time step size restriction for fully explicit methods. Thus this paper also considers implicit and steady-state FCT.

This paper is organized as follows. Section 2 gives some preliminaries such as the problem formulation and discretization. Recall that the FCT algorithm uses a low-order scheme and a high-order scheme. Section 3.1 presents the low-order scheme, Section 3.2 presents the high-order scheme (which is based on entropy viscosity), and Section 3.3 presents the FCT scheme that combines the two. Section 4 presents results for a number of test problems, and Section 5 gives conclusions.

2. Preliminaries

For the remainder of this paper, the scalar transport model given by Equation (1) will be generalized to a scalar balance equation having reaction terms and source terms, with the following notation:

$$\frac{\partial u}{\partial t} + v\boldsymbol{\Omega} \cdot \nabla u(\mathbf{x}, t) + \sigma(\mathbf{x}) u(\mathbf{x}, t) = q(\mathbf{x}, t)\,, \tag{4}$$

where u is the balanced quantity, v is the transport speed, $\boldsymbol{\Omega}$ is a constant, uniform unit direction vector, σ is the reaction coefficient, and q is the source function~~. Note that $q \geq 0$ for particle transport: the~~ , possibly including contributions from an external source, ~~the inscatteringsource, and the fissionsource are all particle production terms and hence are positive~~inscattering, and fission. These contributions are all physically non-negative, but it should be noted that in practical deterministic simulations, if the scattering source is not isotropic, the scattering term may be negative due to its approximation as a truncated Legendre polynomial expansion. However, this work makes the assumption that the source is non-negative: $q \geq 0$; the proof of non-negativity of the solution

relies on this assumption. With anisotropic sources, extra precautions may need to be taken, but this topic is not explored in this preliminary work on the subject.

The problem formulation is completed by supplying initial conditions on the problem domain $\mathcal{D}$ (for transient problems):

$$u(\mathbf{x}, 0) = u^0(\mathbf{x}) \quad \mathbf{x} \in \mathcal{D}\,, \tag{5}$$

as well as boundary conditions, which will be assumed to be incoming flux boundary conditions:

$$u(\mathbf{x}, t) = u^{\text{inc}}(\mathbf{x}, t) \quad \mathbf{x} \in \partial\mathcal{D}^-\,, \tag{6}$$

where $u^{\text{inc}}(\mathbf{x}, t)$ is the incoming boundary data function, and $\partial\mathcal{D}^-$ is the incoming portion of the domain boundary:

$$\partial\mathcal{D}^- \equiv \{\mathbf{x} \in \partial\mathcal{D} : \mathbf{n}(\mathbf{x}) \cdot \mathbf{\Omega} \leq 0\}\,, \tag{7}$$

where $\mathbf{n}(\mathbf{x})$ is the outward-pointing normal vector on the domain boundary at point $\mathbf{x}$.

Application of the standard Galerkin method with piecewise linear basis functions gives the following semi-discrete system:

$$\mathbf{M}^C \frac{d\mathbf{U}}{dt} + \mathbf{A}\mathbf{U}(t) = \mathbf{b}(t)\,, \tag{8a}$$

where the consistent (i.e., not lumped) mass matrix is given by

$$M^C_{i,j} \equiv \int\limits_{S_{i,j}} \varphi_i(\mathbf{x})\varphi_j(\mathbf{x}) dV\,, \tag{8b}$$

the (steady-tstate) transport matrix is

$$A_{i,j} \equiv \int\limits_{S_{i,j}} \left(v\mathbf{\Omega} \cdot \nabla\varphi_j(\mathbf{x}) + \sigma(\mathbf{x})\varphi_j(\mathbf{x})\right) \varphi_i(\mathbf{x}) dV\,, \tag{8c}$$

and the right-hand-side is

$$b_i(t) \equiv \int\limits_{S_i} q(\mathbf{x})\varphi_i(\mathbf{x}) dV\,. \tag{8d}$$

The components of the solution vector $\mathbf{U}(t)$ are denoted by $U_j(t)$ and represent the degrees of freedom of the approximate solution u_h:

$$u_h(\mathbf{x}, t) = \sum_j U_j(t)\varphi_j(\mathbf{x})\,, \tag{9}$$

where $\varphi_j(\mathbf{x})$ is a finite element test function. S_i is the support of basis function i and $S_{i,j}$ is the shared support of basis functions i and j.

A number of temporal discretizations are considered in this paper. Fully explicit temporal discretizations considered include forward Euler:

$$\mathbf{M}^C \frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} + \mathbf{A}\mathbf{U}^n = \mathbf{b}^n \,, \tag{10}$$

as well as Strong Stability Preserving Runge Kutta (SSPRK) methods that can be expressed in the following form:

$$\hat{\mathbf{U}}^0 = \mathbf{U}^n \,, \tag{11a}$$

$$\hat{\mathbf{U}}^i = \gamma_i \mathbf{U}^n + \zeta_i \left[\hat{\mathbf{U}}^{i-1} + \Delta t \mathbf{G}(t^n + c_i \Delta t, \hat{\mathbf{U}}^{i-1}) \right] \,, \quad i = 1, \ldots, s \,, \tag{11b}$$

$$\mathbf{U}^{n+1} = \hat{\mathbf{U}}^s \,. \tag{11c}$$

where s is the number of stages, γ_i, ζ_i, and c_i are coefficients that correspond to the particular SSPRK method, and $\mathbf{G}$ represents the right-hand-side function of an ODE

$$\frac{d\mathbf{U}}{dt} = \mathbf{G}(t, \mathbf{U}(t)) \,, \tag{12}$$

which in this case is the following:

$$\mathbf{G}(t, \mathbf{U}(t)) = (\mathbf{M}^C)^{-1} \left(\mathbf{b}(t) - \mathbf{A}\mathbf{U}(t) \right) \,. \tag{13}$$

SSPRK methods are a subclass of Runge Kutta methods that offer high-order accuracy while preserving stability [29, 30]. The form given in Equation (11) makes it clear that these SSPRK methods can be expressed as a linear combination of steps resembling forward Euler steps, with the only difference being that the explicit time dependence of the source is not necessarily on the old time t^n but instead is on a stage time $t^n + c_i \Delta t$. An example is the 3-stage, 3rd-order accurate SSPRK method has the following coefficients:

$$\gamma = \begin{bmatrix} 0 \\ \frac{3}{4} \\ \frac{1}{3} \end{bmatrix} \,, \quad \zeta = \begin{bmatrix} 1 \\ \frac{1}{4} \\ \frac{2}{3} \end{bmatrix} \,, \quad c = \begin{bmatrix} 0 \\ 1 \\ \frac{1}{2} \end{bmatrix} \,. \tag{14}$$

This work also considers the Theta-family of temporal discretizations:

$$\mathbf{M}^C \frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} + \mathbf{A}((1-\theta)\mathbf{U}^n + \theta \mathbf{U}^{n+1}) = (1-\theta)\mathbf{b}^n + \theta \mathbf{b}^{n+1} \,, \tag{15}$$

where $0 \leq \theta \leq 1$ is the implicitness parameter. For example, θ values of 0, $\frac{1}{2}$, and 1 correspond to forward Euler, Crank-Nicohlson, and backward Euler discretizations, respectively.

Finally, in the case of a steady-state solve, we have the following system of equations:

$$\mathbf{A}\mathbf{U} = \mathbf{b} \,. \tag{16}$$

3. FCT Methodology Applied to Particle Transport

Recall that the FCT algorithm is built from a low-order scheme and a high-order scheme. Section 3.1 describes the low-order scheme, and Section 3.2 describes the high-order scheme. Section 3.3 describes the FCT scheme combined from these components.

3.1. Low-Order Scheme

The role of a low-order scheme in the context of the FCT algorithm is to provide a fail-safe solution, which has desirable properties such as positivity-preservation and lack of spurious oscillations. These properties come at the cost of excessive artificial diffusion and thus a lesser degree of accuracy. However, the idea of the FCT algorithm is to undo some of the over-dissipation of the low-order scheme as much as possible without violating some physically-motivated solution bounds.

Here positivity-preservation and monotonicity are achieved by requiring that that the matrix of the low-order system satisfies the M-matrix property. M-matrices are a subset of inverse-positive matrices and have the monotone property. For instance, consider the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$; If $\mathbf{A}$ is an M-matrix, then the following property is verified:

$$\text{If } \mathbf{b} \geq 0, \text{ then } \mathbf{x} \geq 0\,. \tag{17}$$

Hence, ~~to prove positivity-preservation,~~ given that the linear system matrix is an M-matrix, ~~is achieved~~ positivity-preservation is proven by proving positivity of the right-hand-side vector $\mathbf{b}$. This monotonicity property of the linear system matrix is also responsible for the ~~lack of spurious oscillations~~satisfaction of a discrete maximum principle [28].

In this section, a first-order viscosity method introduced by Guermond [28] will be adapted to the transport equation given by Equation (4). This method uses an element-wise artificial viscosity definition in conjunction with a graph-theoretic local viscous bilinear form that makes the method valid for arbitrary element shapes and dimensions. These definitions will be shown to ensure that the system matrix is a non-singular M-matrix.

The graph-theoretic local viscous bilinear form has the following definition.

Definition 1 (Local Viscous Bilinear Form). The local viscous bilinear form for element K is defined as follows:

$$d_K(\varphi_j, \varphi_i) \equiv \begin{cases} -\frac{1}{n_K - 1} V_K & i \neq j\,, \quad i, j \in \mathcal{I}_K\,, \\ V_K & i = j\,, \quad i, j \in \mathcal{I}_K\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{18}$$

where V_K is the volume of cell K, $\mathcal{I}_K$ is the set of degree of freedom indices such that the corresponding test function has support on cell K, and n_K is the number of indices in that set.

This bilinear form bears resemblance to a standard Laplacian bilinear form: the diagonal entries are positive, the off-diagonal entries are negative, and the row sums are zero. These facts will be invoked in the proof of the M-matrix conditions later in this section.

The element-wise low-order viscosity definition from [28] is adapted to account for the reaction term in the transport equation, Equation (4), but otherwise remains unchanged.

Definition 2 (Low-Order Viscosity). The low-order viscosity for cell K is defined as follows:

$$\nu_K^L \equiv \max_{i \neq j \in \mathcal{I}_K} \frac{\max(0, A_{i,j})}{-\sum_{T \in \mathcal{K}(S_{i,j})} d_T(\varphi_j, \varphi_i)}, \tag{19}$$

where $A_{i,j}$ is the i,j entry of matrix $\mathbf{A}$ given by Equation (8c), $\mathcal{I}_K$ is the set of degree of freedom indices corresponding to basis functions that have support on cell K (this is illustrated in Figure 1 – the indicated nodes have degree of freedom indices belonging to $\mathcal{I}_K$), and $\mathcal{K}(S_{i,j})$ is the set of cell indices for which the cell domain and the shared support $S_{i,j}$ overlap.

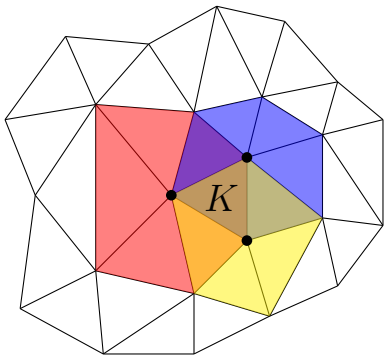


Figure 1: Illustration of Cell Degree of Freedom Indices $\mathcal{I}_K$

This viscosity definition is designed to give the minimum amount of artificial diffusion without violating the M-matrix conditions.

Now that the low-order artificial diffusion operator (bilinear form + viscosity definitions) has been provided, we describe the low-order system. Consider a modification of the Galerkin scheme given in Equation (8) which lumps the mass matrix ($\mathbf{M}^C \rightarrow \mathbf{M}^L$) and adds an artificial diffusion operator $\mathbf{D}^L$, hereafter called the low-order diffusion matrix:

$$\mathbf{M}^L \frac{d\mathbf{U}^L}{dt} + (\mathbf{A} + \mathbf{D}^L)\mathbf{U}^L(t) = \mathbf{b}(t), \tag{20}$$

where $\mathbf{U}^L(t)$ denotes the discrete low-order solution values. Defining the low-order steady-state system matrix $\mathbf{A}^L \equiv \mathbf{A} + \mathbf{D}^L$, the low-order system for the

steady-state system, explicit Euler system, and Theta system, respectively, are
Steady-state:

$$\mathbf{A}^L \mathbf{U}^L_{\sim} = \mathbf{b}\,, \tag{21a}$$

Explicit Euler:

$$\mathbf{M}^L \mathbf{U}^{n+1\,L,n+1}_{\sim\sim} = \mathbf{M}^L \mathbf{U}^n + \Delta t \left(\mathbf{b}^n - \mathbf{A}^L \mathbf{U}^n\right)\,, \tag{21b}$$

Theta scheme:

$$\left(\mathbf{M}^L + \theta \Delta t \mathbf{A}^L\right) \underline{U^{n+1}}\,\underline{\mathbf{U}^{L,n+1}}_{\sim\sim} = \mathbf{M}^L \mathbf{U}^n + \Delta t(\mathbf{b}^\theta - \mathbf{A}^L(1-\theta)\mathbf{U}^n)\,, \tag{21c}$$

where $\mathbf{b}^\theta \equiv (1-\theta)\mathbf{b}^n + \theta \mathbf{b}^{n+1}$. The low-order diffusion matrix is assembled element-wise using the local viscous bilinear form and low-order viscosity definitions:

$$D^L_{i,j} \equiv \sum_{K \in \mathcal{K}(S_{i,j})} \nu^L_K d_K(\varphi_j, \varphi_i)\,. \tag{22}$$

Now the low-order scheme has been fully described, some statements will be made on its properties. Firstly the M-matrix property will be shown for the low-order matrix $\mathbf{A}^L$.

Theorem 1 (M-matrix property). *The low-order steady-state system matrix $\mathbf{A}^L$ is a non-singular M-matrix.*

Proof. There are many definitions that can be used to identify a non-singular M-matrix; one definition gives that an M-matrix can be identified by verifying both of the following properties [31]:

1. strict positivity of diagonal entries: $A_{i,i} > 0$, $\forall i$ and
2. non-positivity of off-diagonal entries: $A_{i,j} \leq 0$, $\forall i$, $\forall j \neq i$.

First, we show that the off-diagonal elements of the matrix $\mathbf{A}^L$ are non-positive. The diffusion matrix entry $D^L_{i,j}$ is bounded as follows:

$$\begin{aligned} D^L_{i,j} &= \sum_{K \in \mathcal{K}(S_{i,j})} \nu^L_K d_K(\varphi_j, \varphi_i) \\ &= \sum_{K \in \mathcal{K}(S_{i,j})} \max_{k \neq \ell \in \mathcal{I}_K} \left(\frac{\max(0, A_{k,\ell})}{-\sum_{T \in \mathcal{K}(S_{k,\ell})} d_T(\varphi_\ell, \varphi_k)} \right) d_K(\varphi_j, \varphi_i)\,. \end{aligned}$$

For an arbitrary quantity $c_{k,\ell} \geq 0\,, \forall k \neq \ell \in \mathcal{I}$, the following is true for $i \neq j \in \mathcal{I}$: $\max\limits_{k \neq \ell \in \mathcal{I}} c_{k,\ell} \geq c_{i,j}$, and thus for $a \leq 0$, $a \max\limits_{k \neq \ell \in \mathcal{I}} c_{k,\ell} \leq a c_{i,j}$. Recall that

$d_K(\varphi_j, \varphi_i) < 0$ for $j \neq i$. Thus, we have:

$$\begin{aligned} D^L_{i,j} &\leq \sum_{K \in \mathcal{K}(S_{i,j})} \frac{\max(0, A_{i,j})}{-\sum_{T \in \mathcal{K}(S_{i,j})} d_T(\varphi_j, \varphi_i)} d_K(\varphi_j, \varphi_i)\,, \qquad j \neq i, \\ &= -\max(0, A_{i,j}) \frac{\sum_{K \in \mathcal{K}(S_{i,j})} d_K(\varphi_j, \varphi_i)}{\sum_{T \in \mathcal{K}(S_{i,j})} d_T(\varphi_j, \varphi_i)}\,, \qquad j \neq i, \\ &= -\max(0, A_{i,j})\,, \qquad j \neq i, \\ &\leq -A_{i,j}\,, \qquad j \neq i\,. \end{aligned}$$

Then applying this relation to the definition of the low-order steady state matrix gives

$$A^L_{i,j} = A_{i,j} + D^L_{i,j} \leq 0\,.$$

Next it will be shown that the row sums are non-negative. Using the fact that $\sum_j \varphi_j(\mathbf{x}) = 1$ and $\sum_j d_K(\varphi_j, \varphi_i) = 0$,

$$\begin{aligned} \sum_j A^L_{i,j} &= \sum_j \int_{S_{i,j}} \left(\mathbf{f}'(u_h) \cdot \nabla \varphi_j + \sigma \varphi_j\right) \varphi_i dV + \sum_j \sum_{K \in \mathcal{K}(S_{i,j})} \nu^L d_K(\varphi_j, \varphi_i)\,, \\ &= \int_{S_i} \left(\mathbf{f}'(u_h) \cdot \nabla \sum_j \varphi_j(\mathbf{x}) + \sigma(\mathbf{x}) \sum_j \varphi_j(\mathbf{x}) \right) \varphi_i(\mathbf{x}) dV\,, \\ &= \int_{S_i} \sigma(\mathbf{x}) \varphi_i(\mathbf{x}) dV\,, \\ &\geq 0\,. \end{aligned}$$

Remark 1. If incoming flux boundary conditions are weakly imposed, then the steady-state system matrix is modified: $\mathbf{A} \to \tilde{\mathbf{A}}$, and the low-order viscosity then uses the *modified* steady-state matrix $\tilde{\mathbf{A}}$. The non-positivity property of the off-diagonal elements still holds. The non-negativity property of the row sums also holds, owing to the relation $\tilde{A}_{i,j} \geq A_{i,j}$.

If the support S_i is not entirely vacuum ($\sigma(\mathbf{x}) \geq 0$ with $\sigma(\mathbf{x}) > 0$ for some $\mathbf{x}$), then the row sum is *strictly* positive. Proof of strict positivity of the diagonal elements directly follows from proof of non-positivity of the off-diagonal elements and strict positivity of the row sums. Thus both conditions for the non-singular M-matrix property have been met.

Thus far, we have been proven that the system matrix for the low-order steady-state system is an M-matrix, and it remains to demonstrate the same for each

of the transient systems. For the explicit Euler/SSPRK systems, the system matrix is just the lumped mass matrix $\mathbf{M}^L$, which is easily shown to be an M-matrix since it is a positive, diagonal matrix. For the θ temporal discretization, the system matrix is a linear combination of the lumped mass matrix and the low-order steady-state system matrix; this linear combination is also an M-matrix since it is a combination of two M-matrices with non-negative combination coefficients.

To complete the proof of positivity preservation for the low-order scheme, we need to show that the system right-hand-side vectors for each temporal discretization are non-negative.

This is immediate for the steady-state case due to the assumption that the source q is non-negative. The following theorem gives that the system right-hand-side vector for the theta system is non-negative. This theorem extends to explicit Euler discretization since explicit Euler is a special case of the Theta discretization.

Theorem 2. *Non-Negativity of the Theta Low-Order System Right-Hand-Side: If the old solution $\mathbf{U}^n$ is non-negative and the time step size Δt satisfies*

$$\Delta t \leq \frac{M^L_{i,i}}{(1-\theta)A^L_{i,i}}\,, \quad \forall i\,, \tag{23}$$

then the new solution $\mathbf{U}^{L,n+1}$ of the Theta low-order system given by Equation (21c) is non-negative, i.e., $U_i^{L,n+1} \geq 0$, $\forall i$.

Proof. The right-hand-side vector $\mathbf{y}$ of this system has the entries

$$y_i = \Delta t b_i^\theta + \left(M^L_{i,i} - (1-\theta)\Delta t A^L_{i,i}\right) U_i^n - (1-\theta)\Delta t \sum_{j\neq i} A^L_{i,j} U_j^n\,.$$

As stated previously, the source function q is non-negative and thus $b_i^\theta \geq 0$. Due to the time step size assumption given by Equation (23),

$$M^L_{i,i} - (1-\theta)\Delta t A^L_{i,i} \geq 0\,,$$

and because the off-diagonal terms of $\mathbf{A}^L$ are ~~non-negative~~non-positive, the off-diagonal sum term is ~~also~~ non-negative. Thus y_i is a sum of non-negative terms, and the theorem is proven.

It can also be shown that the described low-order scheme satisfies a local discrete maximum principle, which is easily shown given the M-matrix property. One may decide to use these bounds as the imposed bounds in the FCT algorithm; however, this approach has been found to yield less accurate solutions than the approach to be outlined in Section 3.3 and is thus not discussed here for brevity.

3.2. High-Order Scheme

This section describes the entropy viscosity method applied to the scalar conservation law given by Equation (4). Recall that the entropy viscosity method is to be used as the high-order scheme in the FCT algorithm, instead of the standard Galerkin method as has been used previously in FCT-FEM schemes; for Galerkin FCT-FEM examples, see, for instance, [22, 24, 21, 25, 26]. Usage of the entropy viscosity method in the FCT algorithm ensures convergence to the entropy solution [27].

The entropy viscosity method has been applied to a number of PDEs such as general scalar conservation laws of the form

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = 0\,, \tag{24}$$

the inviscid Euler equations [1, 32], and the two-phase seven-equation fluid model [33]. The scalar model studied in this paper does not fit into the general form given by Equation (24), due to the addition of the reaction term σu and source term q. Application of entropy viscosity method to the transport equation model is novel and it is described below.

Since the weak form of the problem does not have a unique solution, one must supply additional conditions called *admissibility* conditions or *entropy* conditions to filter out spurious weak solutions, leaving only the physical weak solution, often called the entropy solution. A number of entropy conditions are valid, but usually the most convenient entropy condition for use in numerical methods takes the form of an *entropy inequality*, such as the following, which is valid for the general scalar conservation law given by Equation (24):

$$\frac{\partial \eta(u)}{\partial t} + \nabla \cdot \mathbf{\Psi}(u) \leq 0\,, \tag{25}$$

which holds for any convex entropy function $\eta(u)$ and associated entropy flux $\mathbf{\Psi}(u) \equiv \int \eta'(u)\mathbf{f}'(u)du$. If one can show that this inequality holds for an arbitrary convex entropy function, then one proves it holds for all convex entropy functions [34, 1]. For the scalar PDE considered in this paper, the entropy inequality becomes the following:

$$\frac{\partial \eta(u)}{\partial t} + \nabla \cdot \mathbf{\Psi}(u) + \eta'(u)\sigma u - \eta'(u)q \leq 0\,. \tag{26}$$

One can verify this inequality by multiplying the governing PDE by $\eta'(u)$ and applying reverse chain rule.

The entropy viscosity method enforces the entropy inequality by measuring local entropy production and dissipating accordingly. In practice, one defines the entropy residual:

$$\mathcal{R}(u) \equiv \frac{\partial \eta(u)}{\partial t} + \nabla \cdot \mathbf{\Psi}(u) + \eta'(u)\sigma u - \eta'(u)q\,, \tag{27}$$

which can be viewed as the amount of violation of the entropy inequality. The entropy viscosity for an element K is then defined to be proportional to this violation, for example:

$$\nu_K^\eta = \frac{c_\mathcal{R} \|\mathcal{R}(u_h)\|_{L^\infty(K)}}{\hat{\eta}_K}\,, \tag{28}$$

where $\hat{\eta}_K$ is a normalization constant with the units of entropy, $c_\mathcal{R}$ is a proportionality constant that can be modulated for each problem, and $\|\mathcal{R}(u_h)\|_{L^\infty(K)}$ is the maximum of the entropy residual on element K, which can be approximated as the maximum over the quadrature points of element K. Designing a universally appropriate normalization constant $\hat{\eta}_K$ remains a challenge for the entropy viscosity method (see [32] for an alternate normalization for low-Mach flows). The objective of this normalization coefficient is to prevent the user from needing to make significant adjustments to the tuning parameter $c_\mathcal{R}$ for different problems. A definition that produces reasonable results for a large number of problems is the following:

$$\hat{\eta}_K \equiv \|\eta - \bar{\eta}\|_{L^\infty(\mathcal{D})}\,, \tag{29}$$

where $\bar{\eta}$ is the average entropy over the entire computational domain.

In addition to the entropy residual, it can also be beneficial to measure the jump in the gradient of the entropy flux across cell interfaces. Note that given the definition of the entropy flux, the gradient of the entropy flux is $\nabla\boldsymbol{\Psi}(u) = \nabla\eta(u)\mathbf{f}'(u)$. Then let $\mathcal{J}_F$ denote the jump of the normal component of the entropy flux gradient across face F:

$$\mathcal{J}_F \equiv |\mathbf{f}'(u)\cdot\mathbf{n}_F| [\![\nabla\eta(u)\cdot\mathbf{n}_F]\!]\,, \tag{30}$$

where the double square brackets denote a jump quantity. Then we define the maximum jump on a cell:

$$\mathcal{J}_K \equiv \max_{F\in\mathcal{F}(K)} |\mathcal{J}_F|\,. \tag{31}$$

Finally, putting everything together, one can define the entropy viscosity for a cell K to be

$$\nu_K^\eta = \frac{c_\mathcal{R}\|\mathcal{R}(u_h)\|_{L^\infty(K)} + c_\mathcal{J}\mathcal{J}_K}{\hat{\eta}_K}\,. \tag{32}$$

However, it is known that the low-order viscosity for an element, as computed in Section 3.1, gives enough local artificial diffusion for regularization; any amount of viscosity larger than this low-order viscosity would be excessive. Thus, the low-order viscosity for an element is imposed as the upper bound for the high-order viscosity:

$$\nu_K^H \equiv \min(\nu_K^L, \nu_K^\eta)\,. \tag{33}$$

One can note that, in smooth regions, this high-order viscosity will be small, and, in regions of strong gradients or discontinuities, the entropy viscosity can be relatively large.

Finally, the high-order system of equations for the various time discretizations are as follows: Steady-state:

$$\mathbf{A}^H \mathbf{U}^H = \mathbf{b}\,, \tag{34a}$$

Explicit Euler:

$$\mathbf{M}^C \frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} \frac{\mathbf{U}^{H,n+1} - \mathbf{U}^n}{\Delta t} + \mathbf{A}^{H,n}\mathbf{U}^n = \mathbf{b}^n\,, \tag{34b}$$

Theta scheme:

$$\mathbf{M}^C \frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} \frac{\mathbf{U}^{H,n+1} - \mathbf{U}^n}{\Delta t} + \theta \mathbf{A}^{H,n+1}\mathbf{U}^{n+1\,H,n+1} + (1-\theta)\mathbf{A}^{H,n}\mathbf{U}^n = \mathbf{b}^\theta\,, \tag{34c}$$

where the high-order steady-state system matrix is defined as $\mathbf{A}^H \equiv \mathbf{A} + \mathbf{D}^H$, and the high-order diffusion matrix $\mathbf{D}^H$ is defined similarly to the low-order case but using the high-order viscosity instead of the low-order viscosity:

$$D^H_{i,j} \equiv \sum_{K \in \mathcal{K}(S_{i,j})} \nu^H_K d_K(\varphi_j, \varphi_i)\,. \tag{35}$$

Note that unlike the low-order scheme, the high-order scheme does not lump the mass matrix.

Remark 2. Note that due to the nonlinearity of the entropy viscosity, the entropy viscosity scheme is nonlinear for implicit and steady-state temporal discretizations, and thus some nonlinear solution technique must be utilized. For the results in this paper, a simple fixed-point iteration scheme is used. An alternative such as Newton's method is likely to be advantageous in terms of the number of nonlinear iterations; however, fixed-point is used here for comparison with the nonlinear FCT scheme to be described in Section 3.3.

3.3. FCT Scheme

3.3.1. The FCT System

The entropy viscosity method described in Section 3.2 enforces the entropy condition and thus produces numerical approximations that converge to the entropy solution. However, numerical solutions may still contain spurious oscillations and negativities, although these effects are smaller in magnitude than for the corresponding Galerkin solution. In this paper, the flux-corrected transport (FCT) algorithm is used to further mitigate the formation of spurious oscillations and to guarantee the absence of negativities.

The first ingredient of the FCT algorithm is the definition of the antidiffusive fluxes. To arrive at this definition, the low-order systems, given by Equations (21a), (21b), and (21c) for each temporal discretization, are augmented with

the addition of the *antidiffusion source* $\mathbf{p}$, which now, instead of producing the low-order solution $\mathbf{U}^L$, produces the high-order solution $\mathbf{U}^H$:

$$\mathbf{A}^L\mathbf{U}^H = \mathbf{b} + \mathbf{p}\,, \tag{36a}$$

$$\mathbf{M}^L\frac{\mathbf{U}^H - \mathbf{U}^n}{\Delta t} + \mathbf{A}^L\mathbf{U}^n = \mathbf{b}^n + \mathbf{p}^n\,, \tag{36b}$$

$$\mathbf{M}^L\frac{\mathbf{U}^H - \mathbf{U}^n}{\Delta t} + \mathbf{A}^L\left(\theta\mathbf{U}^H + (1-\theta)\mathbf{U}^n\right) = \mathbf{b}^\theta + \mathbf{p}^\theta\,. \tag{36c}$$

Then the corresponding high-order systems, given by Equations (34a), (34b), (34c) are subtracted from these equations to give the following definitions for $\mathbf{p}$:

$$\mathbf{p} \equiv \left(\mathbf{D}^L - \mathbf{D}^H\right)\mathbf{U}^H\,, \tag{37a}$$

$$\mathbf{p}^n \equiv -\left(\mathbf{M}^C - \mathbf{M}^L\right)\frac{\mathbf{U}^H - \mathbf{U}^n}{\Delta t} + \left(\mathbf{D}^L - \mathbf{D}^H\right)\mathbf{U}^n\,, \tag{37b}$$

$$\mathbf{p}^\theta \equiv -\left(\mathbf{M}^C - \mathbf{M}^L\right)\frac{\mathbf{U}^H - \mathbf{U}^n}{\Delta t} + (1-\theta)\left(\mathbf{D}^L - \mathbf{D}^{H,n}\right)\mathbf{U}^n + \theta\left(\mathbf{D}^L - \mathbf{D}^{H,n+1}\right)\mathbf{U}^{n+1\,H}\,, \tag{37c}$$

The next step is to decompose each antidiffusive source p_i into a sum of antidiffusive fluxes: $p_i = \sum_j P_{i,j}$. Because the matrices $\mathbf{M}^C - \mathbf{M}^L$ and $\mathbf{D}^L - \mathbf{D}^H$ are symmetric and feature row sums of zero, the following are valid antidiffusive flux decompositions:

$$P_{i,j} = \left(D^L_{i,j} - D^H_{i,j}\right)\left(U^H_j - U^H_i\right)\,, \tag{38a}$$

$$P^n_{i,j} = -M^C_{i,j}\left(\frac{U^H_j - U^n_j}{\Delta t} - \frac{U^H_i - U^n_i}{\Delta t}\right) + \left(D^L_{i,j} - D^{H,n}_{i,j}\right)\left(U^n_j - U^n_i\right)\,, \tag{38b}$$

$$P^\theta_{i,j} = -M^C_{i,j}\left(\frac{U^H_j - U^n_j}{\Delta t} - \frac{U^H_i - U^n_i}{\Delta t}\right) + (1-\theta)\left(D^L_{i,j} - D^{H,n}_{i,j}\right)\left(U^n_j - U^n_i\right) + \theta\left(D^L_{i,j} - D^{H,n+1}_{i,j}\right)\left(U^H_j - U^H_i\right)\,. \tag{38c}$$

Note that this decomposition yields equal and opposite antidiffusive flux pairs since the antidiffusion matrix $\mathbf{P}$ is skew symmetric: $P_{j,i} = -P_{i,j}$ (and likewise for $P^n_{j,i}$ and $P^\theta_{j,i}$). Up until this point, no changes have been made to the high-order scheme: solving Equations (36a) through (36c) still produces the high-order solution. FCT is applied to these equations by applying limiting coefficients $L_{i,j}$ to each antidiffusive flux $P_{i,j}$. Thus the FCT systems are the following:

$$\mathbf{A}^L\mathbf{U}^H = \mathbf{b} + \hat{\mathbf{p}}\,, \tag{39a}$$

$$\mathbf{M}^L \frac{\mathbf{U}^H - \mathbf{U}^n}{\Delta t} + \mathbf{A}^L \mathbf{U}^n = \mathbf{b}^n + \hat{\mathbf{p}}^n \,, \tag{39b}$$

$$\mathbf{M}^L \frac{\mathbf{U}^H - \mathbf{U}^n}{\Delta t} + \mathbf{A}^L \left(\theta \mathbf{U}^H + (1-\theta)\mathbf{U}^n\right) = \mathbf{b}^\theta + \hat{\mathbf{p}}^\theta \,, \tag{39c}$$

where the hat denotes limitation: $\hat{p}_i \equiv \sum_j L_{i,j} P_{i,j}$. The limiting coefficients range between zero and one, representing full limitation and no limitation, respectively. For example, setting all limiting coefficients to zero would result in the low-order solution, and setting all to one would result in the high-order solution. The actual values of the limiting coefficients are determined by the limiter, which operates on the following goal: maximize the limiting coefficients such that the imposed solution bounds are not violated.

As will be discussed in Section 3.3.2, the solution bounds for implicit FCT and steady-state FCT are implicit, and thus the systems given by Equations (39c) and (39a) are nonlinear, since the limiting coefficients contained in $\hat{\mathbf{p}}$ are nonlinear. In this paper, a fixed-point iteration scheme is used to resolve the nonlinearities. For any nonlinear iteration scheme, the imposed solution bounds must be computed using the previous solution iterate:

$$U_i^{-,(\ell)} \leq U_i^{(\ell+1)} \leq U_i^{+,(\ell)} \,. \tag{40}$$

Though the solution bounds are lagged, the antidiffusion bounds $\hat{p}_i^{\pm}$ still contains terms at iteration $\ell + 1$; these terms must be lagged as well. As a consequence, the solution bounds for implicit/steady-state FCT schemes are only satisfied upon nonlinear convergence, not at each iteration.

3.3.2. Solution Bounds

The integral form of the transport equation can be derived using the method of characteristics. Consider a frame of reference moving with the radiation field so that position is a function of time, resulting in a family of characteristic curves (since the transport equation is linear, these curves are straight lines) $\mathbf{x}(t)$ that solve the following ODE:

$$\frac{d\mathbf{x}}{dt} = v\boldsymbol{\Omega} \,, \quad \mathbf{x}(0) = \mathbf{x}_0 \,. \tag{41}$$

Then taking the time derivative of $u(\mathbf{x}(t), t)$ gives

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \nabla \cdot (u(\mathbf{x}(t), t)) \frac{d\mathbf{x}}{dt} \tag{42}$$

$$= \frac{\partial u}{\partial t} + \nabla \cdot (u(\mathbf{x}(t), t)) \frac{d\mathbf{x}}{dt} \tag{43}$$

Finally, combining this with Equation (4) and solving the resulting ODE gives the integral transport equation [2]:

$$u(\mathbf{x},t) = u_0(\mathbf{x}-vt\boldsymbol{\Omega})e^{-\int_0^t \sigma(\mathbf{x}-v(t-t')\boldsymbol{\Omega})vdt'} + \int_0^t q(\mathbf{x}-v(t-t')\boldsymbol{\Omega},t')e^{-\int_{t'}^t \sigma(\mathbf{x}-v(t-t'')\boldsymbol{\Omega})vdt''} vdt' \,. \quad (44)$$

If the time step size Δt satisfies the condition

$$v\Delta t \leq h_{\min}\,, \quad h_{\min} \equiv \min_K h_K\,, \quad (45)$$

where h_K is the diameter of cell K, then the following discrete solution bounds apply:

$$U_i^- \leq U_i^{n+1} \leq U_i^+\,, \quad (46a)$$

where

$$U_i^- \equiv \begin{cases} U_{\min,i}^n e^{-v\Delta t\sigma_{\max,i}} + \frac{q_{\min,i}}{\sigma_{\max,i}}(1-e^{-v\Delta t\sigma_{\max,i}})\,, & \sigma_{\max,i}\neq 0 \\ U_{\min,i}^n + v\Delta t q_{\min,i}\,, & \sigma_{\max,i}=0 \end{cases}\,, \quad (46b)$$

and

$$U_i^+ \equiv \begin{cases} U_{\max,i}^n e^{-v\Delta t\sigma_{\min,i}} + \frac{q_{\max,i}}{\sigma_{\min,i}}(1-e^{-v\Delta t\sigma_{\min,i}})\,, & \sigma_{\min,i}\neq 0 \\ U_{\max,i}^n + v\Delta t q_{\max,i}\,, & \sigma_{\min,i}=0 \end{cases}\,. \quad (46c)$$

The other quantities used in the above expressions are:

$$U_{\max,i}^n \equiv \max_{j\in\mathcal{I}(S_i)} U_j^n\,, \quad U_{\min,i}^n \equiv \min_{j\in\mathcal{I}(S_i)} U_j^n\,, \quad (46d)$$

$$\sigma_{\max,i} \equiv \max_{\mathbf{x}\in S_i}\sigma(\mathbf{x})\,, \quad \sigma_{\min,i} \equiv \min_{\mathbf{x}\in S_i}\sigma(\mathbf{x})\,, \quad (46e)$$

$$q_{\max,i} \equiv \max_{\mathbf{x}\in S_i} q(\mathbf{x})\,, \quad q_{\min,i} \equiv \min_{\mathbf{x}\in S_i} q(\mathbf{x})\,. \quad (46f)$$

Note the time step size condition given by Equation (45) implies that when using CFL numbers greater than 1 with implicit time discretizations, these bounds no longer apply. Similar bounds can be derived for $v\Delta t > h_{min}$; however, these bounds for a node i will no longer only depend on the solution values of the immediate neighbors of i; instead, a larger neighborhood must be used in the bounds, making the local solution bounds wider and thus less restrictive and arguably less useful in the FCT algorithm. This represents a significant disadvantage for implicit FCT, not only because the converged FCT solution could contain more undesirable features but also because the wider bounds typically result in a greater number of nonlinear iterations because of the increased freedom in the limiting coefficients.

Steady-state FCT solution bounds can be inferred from Equation (46) by making the substitution $v\Delta t \to s$, where $0 \leq s \leq h_{min}$. This restriction of s similarly ensures that only the nearest neighbors of i are needed for the solution bounds of i. Steady-state FCT unfortunately suffers many of the same drawbacks as implicit FCT because like implicit FCT, its solution bounds are implicit and thus change with each iteration.

3.3.3. Antidiffusion Bounds

Bounds imposed on a solution value i, such as the bounds described in Section 3.3.2, directly translate into bounds on the limited antidiffusion source $\hat{p}_i$. These antidiffusion bounds $\hat{p}_i^{\pm}$ for steady-state, explicit Euler, and Theta discretization are respectively derived by solving Equations (39a), (39b), and (39c) for $\hat{p}_i$ and manipulating the inequality $U_i^- \leq U_i \leq U_i^+$. This yields:

$$\hat{p}_i^{\pm} \equiv A_{i,i}^L U_i^{\pm} + \sum_{j \neq i} A_{i,j}^L U_j - b_i \,, \tag{47a}$$

$$\hat{p}_i^{\pm} \equiv M_{i,i}^L \frac{U_i^{\pm} - U_i^n}{\Delta t} + \sum_j A_{i,j}^L U_j^n - b_i^n \,, \tag{47b}$$

$$\hat{p}_i^{\pm} \equiv \left(\frac{M_{i,i}^L}{\Delta t} + \theta A_{i,i}^L \right) U_i^{\pm} + \left((1-\theta) A_{i,i}^L - \frac{M_{i,i}^L}{\Delta t} \right) U_i^n + \sum_{j \neq i} A_{i,j}^L U_j^{\theta} - b_i^{\theta} \,. \tag{47c}$$

We note that, if the limiting coefficients $L_{i,j}$ are selected such that $\hat{p}_i^- \leq \hat{p}_i \leq \hat{p}_i^+$, then the solution bounds are satisfied: $U_i^- \leq U_i \leq U_i^+$.

Limiters such as the Zalesak ~~'s~~ limiter described in Section 3.3.4 are algebraic ~~operator~~operators, taking as input the antidiffusion bounds $\hat{p}_i^{\pm}$ and the antidiffusive fluxes $P_{i,j}$ and returning as output the limiting coefficients $L_{i,j}$. It is important to note that most limiters, including the limiter described in this paper, assume the following: $\hat{p}_i^- \leq 0$, $\hat{p}_i^+ \geq 0$; the reasoning for this assumption is as follows. Recall that FCT starts from the low-order scheme, which is equivalent to the solution with $\hat{p}_i = 0$. The limiter should start from this point so that there is a fail-safe solution for the FCT algoritm: the low-order solution. Otherwise, there is no guarantee that any combination of values of limiting coefficients will achieve the desired condition $\hat{p}_i^- \leq \hat{p}_i \leq \hat{p}_i^+$. If $\hat{p}_i^- > 0$ or $\hat{p}_i^+ < 0$, then the starting state, the low-order solution, with $\hat{p}_i = 0$ is an invalid solution of the FCT algorithm. Some solution bounds automatically satisfy $\hat{p}_i^- \leq 0$ and $\hat{p}_i^+ \geq 0$, but in general these conditions must be enforced. In this paper, the solution bounds are possibly widened by directly enforcing these assumptions:

$$\hat{p}_i^- \leftarrow \min(0, \hat{p}_i^-) \,, \tag{48}$$

$$\hat{p}_i^+ \leftarrow \max(0, \hat{p}_i^+) \,. \tag{49}$$

We have noted that omitting this step can lead to poor results. Without this step, the assumptions of the limiter are violated, and thus limiting coefficients that do not satisfy the imposed solution bounds may be generated.

3.3.4. Limiting Coefficients

The results in this paper use the classic multi-dimensional limiter introduced by Zalesak [19]:

$$p_i^+ \equiv \sum_j \max(0, P_{i,j}) \,, \qquad p_i^- \equiv \sum_j \min(0, P_{i,j}) \,, \tag{50a}$$

$$L_i^{\pm} \equiv \begin{cases} 1 & p_i^{\pm} = 0 \\ \min\left(1, \frac{\bar{p}_i^{\pm}}{p_i^{\pm}}\right) & p_i^{\pm} \neq 0 \end{cases}, \tag{50b}$$

$$L_{i,j} \equiv \begin{cases} \min(L_i^+, L_j^-) & P_{i,j} \geq 0 \\ \min(L_i^-, L_j^+) & P_{i,j} < 0 \end{cases}. \tag{50c}$$

The objective of a limiter is to maximize the amount of antidiffusion that can be accepted without violating the imposed solution constraints. Zalesak's limiter is one commonly used attempt at this objective due to its relatively simple form.

Remark 3. Note that it is possible to devise limiters accepting more antidiffusion than Zalesak's limiter, but one must sacrifice the simple, closed form of Zalesak's limiter and adapt a sequential algorithm for computing the antidiffusion coefficients. This approach has not been found in literature, most likely because the sequential aspect of the algorithm makes it node-order-dependent, which reduces reproducibility between implementations.

Finally, one could pass antidiffusive fluxes through a given limiter multiple times, using the remainder antidiffusive flux as the input in each pass, to increase the total antidiffusion accepted [22, 35]; however, the results presented in this paper were all produced using the traditional single-pass approach through the Zalesak limiter.

4. Results

This section presents results for a number of test problems, which compare the solutions obtained using:

- the standard Galerkin FEM, labelled as "Galerkin" in the plots,
- the low-order method, labelled in plots as "Low",
- the entropy viscosity method, labelled in plots as "EV",
- the standard Galerkin FEM with FCT, labelled in plots as "Galerkin-FCT", and
- the entropy viscosity method with FCT, labelled in plots as "EV-FCT".

All problems assume a speed of $v = 1$ (the speed effectively just changes the units of Δt) and an entropy function of $\eta(u) = \frac{1}{2}u^2$. Unless otherwise specified, the transport direction is in the positive x direction: $\mathbf{\Omega} = \mathbf{e}_x$, and the entropy viscosity tuning parameters of $c_{\mathcal{R}}$ and $c_{\mathcal{J}}$ are set to 0.1. A third-order Gauss quadrature is used for spatial integration in all test cases.

4.1. Spatial Convergence Tests

This 1-D, steady-state test problem uses the Method of Manufactured Solutions (MMS) with a solution of $u(x) = \sin(\pi x)$ on the domain $x \in (0, 1)$. Zero Dirichlet boundary conditions are imposed on both boundaries. With $\sigma(x) = 1$, the MMS source becomes $q(x) = \pi \cos(\pi x) + \sin(\pi x)$. The number of cells in the study starts at 8 for the coarsest mesh, and cells are refined by a factor of 2 in each cycle, ending with 256 cells.

Figure 2 shows the L^2 norm errors for this convergence study and indicates first-order spatial convergence for the low-order method and second-order spatial convergence for the entropy viscosity (EV) method and EV-FCT method, as expected.

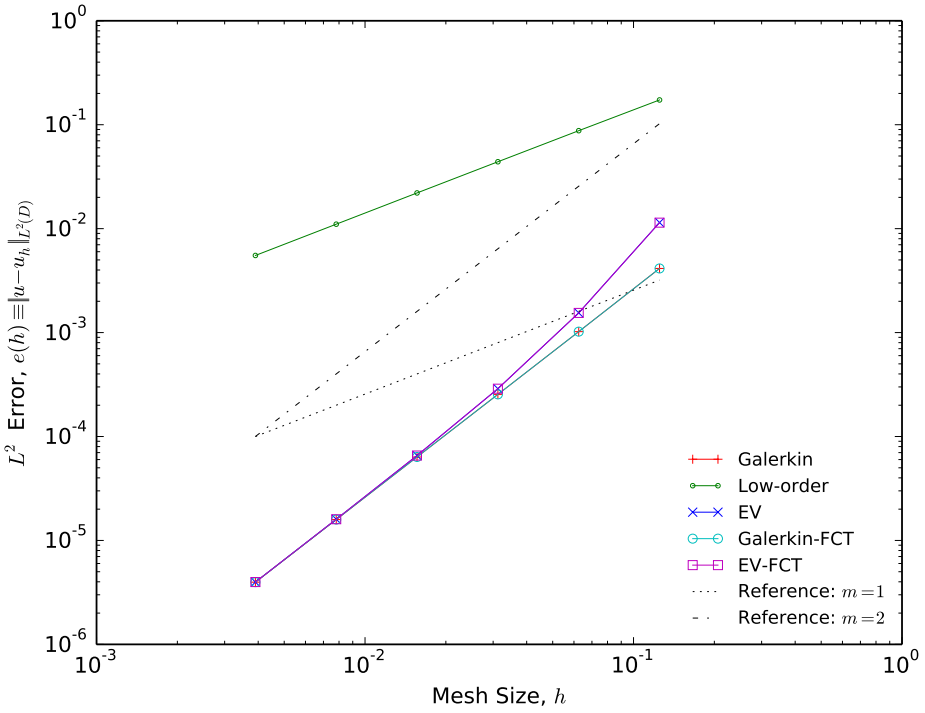
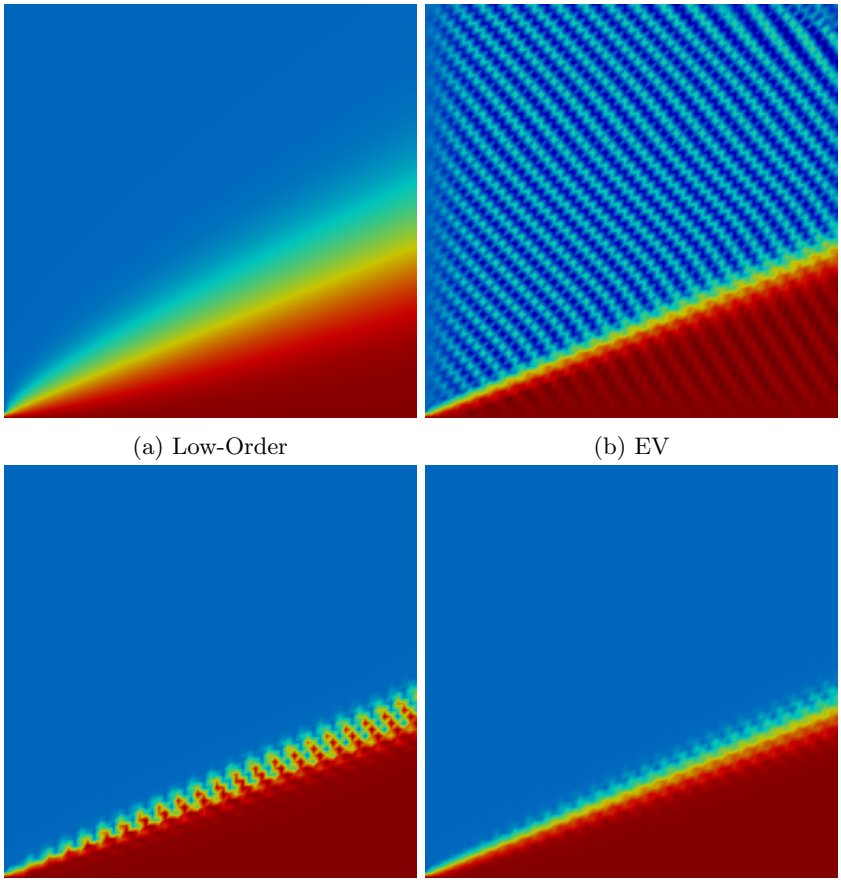


Figure 2: Spatial Convergence for MMS Problem

4.2. Glancing Beam in a Void

This 2-D test problem is on the unit square: $\mathbf{x} \in (0,1)^2$ and simulates a beam incident on the bottom boundary of a void region ($\sigma(\mathbf{x}) = 0$, $q(\mathbf{x}) = 0$) at a shallow angle of 21.94° with respect to the x-axis ($\Omega_x = \cos(21.94°)$, $\Omega_y = \sin(21.94°)$). The exact solution of this problem contains a discontinuity along the line $y = \frac{\Omega_y}{\Omega_x} x$, which presents opportunity for the formation of spurious oscillations. This is run as a pseudo-transient problem with zero initial conditions until steady-state is reached. A Dirichlet boundary condition is imposed on the incoming sides of the problem, with a value of 1 on the bottom boundary and a value of 0 on the left boundary.

This problem is run with Explicit Euler time discretization and a CFL number of 0.5 on a 64×64 mesh. Figure 3 compares the numerical solutions for this problem obtained with the low-order, EV, Galerkin-FCT, and EV-FCT schemes. The Galerkin scheme (without FCT) produced spurious oscillations without bound, so those results are omitted here. The same color scale is used for each image: the dark red shown in the low-order and the two FCT sub-plots corresponds to the incoming value of 1, and the blue in these same sub-plots corresponds to zero. The darker blues and reds shown in the EV sub-plot indicate undershoots and overshoots, respectively. Both FCT solutions keep the solution within the imposed physical bounds, but one can see from the Galerkin-FCT results that some "terracing" effects are present; this behavior is a well-known artifact of traditional FCT schemes [22]. In this case, and in all observed cases of the terracing phenomenon, the EV-FCT scheme shows a reduction of this effect: the addition of the entropy-based artificial viscosity decreases the magnitude of the spurious oscillations in the high-order scheme and thus lessens the burden on the limiter.



(a) Low-Order (b) EV

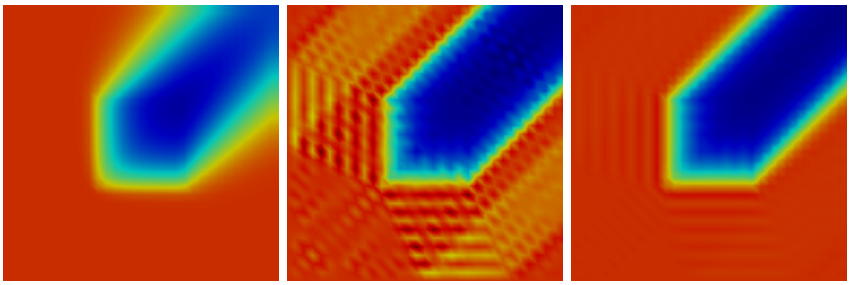
(c) Galerkin-FCT (d) EV-FCT

Figure 3: Comparison of Solutions for the Glance-in-Void Test Problem Using Explicit Euler Time Discretization

4.3. Obstruction Test

This is a 2-D, two-region problem on the unit square $(0,1)^2$ with a beam incident on the left and bottom boundaries at an angle of 45° with the x-axis. The center region $(\frac{1}{3},\frac{2}{3})^2$ is an absorber region with $\sigma(\mathbf{x}) = 10$ and $q(\mathbf{x}) = 0$, and the surrounding region is a void ($\sigma(\mathbf{x}) = 0$, $q(\mathbf{x}) = 0$).

This problem was run with Implicit Euler with a CFL of 1 to steady-state on a 32×32 mesh. The results are shown in Figure 4. The low-order solution is especially diffusive and shows a fanning of the solution after it passes the corners of the obstruction, which is not present in any of the high-order schemes. The EV solution contains oscillations, although they are much less significant than in the Galerkin solution. Both FCT schemes show a lack of these oscillations, but the Galerkin-FCT solutions shows a terracing effect. The EV-FCT solution also has this effect but to a much smaller degree.



(a) Low-Order (b) Galerkin (c) EV

(d) Galerkin-FCT (e) EV-FCT

Figure 4: Comparison of Solutions for the Obstruction Test Problem Using Implicit Euler Time Discretization

Table 1 shows the results for a parametric study on the number of nonlinear iterations required for EV and EV-FCT for various CFL numbers on a 16×16 mesh, to an end time of $t = 1.5$. Recall that in the FCT algorithm, one first computes the high-order solution (here, EV), which is necessary for computation of the antidiffusive fluxes. The number of nonlinear iterations for this solve is given in the column labeled as "EV". The FCT limiting procedure is also nonlinear; the number of nonlinear iterations for this solve is given in the "FCT" column. ~~These results indicate a trend for both EV and FCT that the number~~

~~of nonlinear iterations increase with CFL number; however, this trend is much more drastic with FCT than EV. For a CFL number of 20, the FCT algorithm fails to converge entirely. Note that in this case, no numbersare shown in the "EV" column, even though entropy viscosity iteration did not fail, because the total iterations could not be determined due to the premature termination of the simulation~~While these results indicate that increasing the CFL number decreases the total computational work in reaching the end of the transient, it should be noted that the quality of the FCT solution deteriorates significantly for large CFL numbers; see the "L^2 Err." column. As discussed previously, the solution bounds are implicit and thus change with each iteration. This challenge is compounded in the case of large time step sizes; as time step size increases, the solution bounds widen, and successive FCT solution iterates differ more than for smaller time step sizes. This trend is ~~shown in the increasing~~ suggested in the rate of increase in the number of FCT iterations per time step in Table 1, which is significantly larger than the rate of increase of EV iterations per time step. However, this issue can be ~~easily~~ mitigated using a relaxation factor on the iterative solution updates; for example, for the failing case of CFL number equal to 20, convergence can be achieved with a relaxation factor of 0.9, giving average iterations per time step of 14.33 and 221.67 for EV and FCT, respectively.

Table 1: Nonlinear Iterations vs. CFL Number for the Obstruction Test Problems

CFL	*Relax*	*EV*		*FCT*		*L^2 Err.*
		Total	*Avg.*	*Total*	*Avg.*	
0.1	–	~~3999~~ 6204	~~8.14~~ 8.43	~~3585~~ 5223	~~7.30~~7.23	5.084×10^{-2}
0.5	–	~~896~~ 1386	~~9.05~~ 9.36	~~1499~~ 2239	~~15.14~~15.13	5.079×10^{-2}
1.0	–	~~501~~ 791	~~10.02~~ 10.69	~~970~~ 1588	~~19.40~~21.46	5.111×10^{-2}
5.0	–	~~157~~ 265	~~15.70~~ 17.67	~~1130~~ 1780	~~113.00~~118.67	5.980×10^{-2}
10.0	–	~~79~~ 150	~~15.80~~ 18.75	~~753~~ 1298	~~150.60~~162.25	9.854×10^{-2}
20.0	–	–	–	(failure)		–
20.0	0.9	~~43~~ 66	~~14.33~~ 16.50	~~665~~ 935	~~221.67~~ 233.50	1.295×10^{-1}

4.4. Two-Region Interface

This 1-D test problem simulates the interface between two regions with varying cross section and source values on the domain $(0, 1)$. The left half of the domain has values $\sigma(\mathbf{x}) = 10$ and $q(\mathbf{x}) = 10$, for which the transport solution will reach a saturation value of $\frac{q}{\sigma} = 1$, while the right half has values of $\sigma(\mathbf{x}) = 40$ and $q(\mathbf{x}) = 20$, giving it a saturation value of $\frac{q}{\sigma} = 0.5$. The transport direction is $\mathbf{\Omega} = \mathbf{e}_x$, with zero incident flux on the left boundary.

This problem was run using SSPRK33 time discretization with a CFL of 1 to steady-state with 32 cells. Figure 5 shows the results for this test problem. The low-order solution suffers from the significant artificial diffusion, while the Galerkin solution suffers from significant spurious oscillations. The EV scheme eliminates some of the first oscillations, but a number of oscillations of a similar magnitude to those produced by the Galerkin scheme still exists to the left of the interface. The sets "U^-" and "U^+" correspond to the minimum and maximum solution bounds, respectively, of each FCT scheme (the differences between the bounds of the two FCT schemes are insignificant here). Both FCT schemes effectively eliminate spurious oscillations without approaching the level of unnecessary artificial diffusion achieved by the low-order scheme. For this test problem, the Galerkin-FCT solution is slightly superior to the EV-FCT solution and the "terracing" phenomenon of FCT is not present.

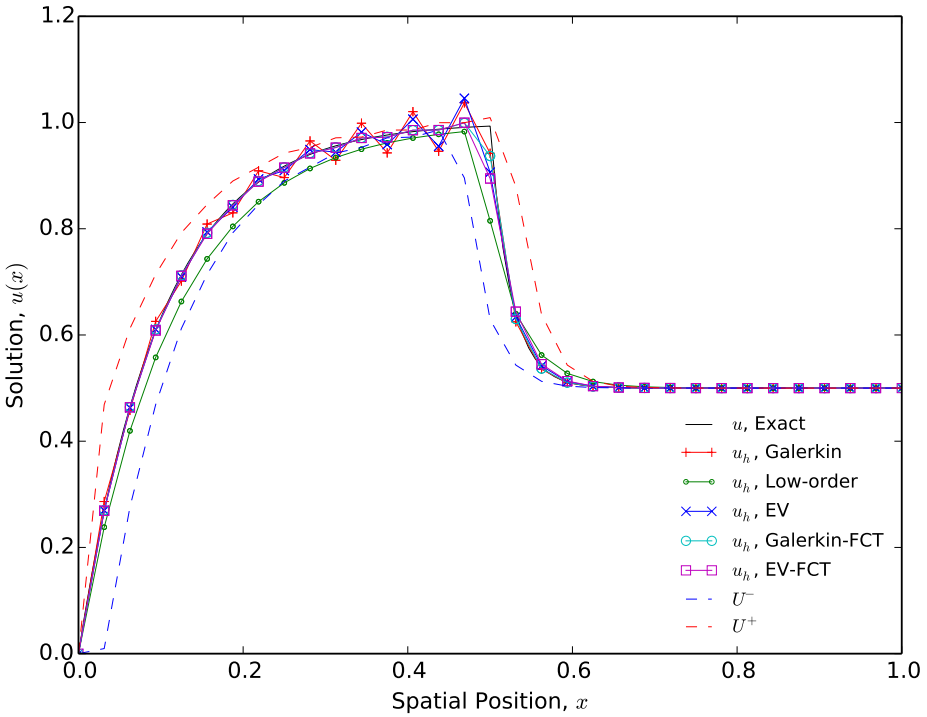


Figure 5: Comparison of Solutions for the Two-Region Interface Test Problem Using SSPRK33 Time Discretization

4.5. Source in a Void Test

This 1-D test problem has two regions, the left being a void but with a source: $\sigma(\mathbf{x}) = 0$ and $q(\mathbf{x}) = 1$, and the right being an absorber without a source: $\sigma(\mathbf{x}) = 10$ and $q(\mathbf{x}) = 0$. This does not represent necessarily a non-physical problem: due to the energy dependence of the particle distribution, it is possible to have a strong inscattering source term in an energy bandwidth where the interaction coefficient is small; here we emphasize the weak interaction probability by zeroing out the reaction term. A zero Dirichlet boundary condition is imposed on the incoming (left) boundary, and this problem is run as a steady-state problem on the domain $(0, 1)$ with 32 cells.

For this problem, the entropy residual and jump coefficients $c_{\mathcal{R}}$ and $c_{\mathcal{J}}$ are set to 0.5; the default value of 0.1 was found to be too small for this problem. Figure 6 shows the results for this problem. The Galerkin solution has "kinks" along the void (left) region, but shows point-wise matching of the exact solution in the absorber (right) region. The EV solution, however, eliminates the terracing effect in the void region but suffers from an inflated peak at the interface, due to entropy production at the interface. The EV-FCT solution bounds generated for this problem are denoted as "U^-, EV-FCT" and "U^+, EV-FCT". For this test problem, the Galerkin-FCT scheme gives superior results to that of the EV-FCT scheme. However, both FCT schemes suffer from the familiar FCT phenomenon known as "peak-clipping", whereby the "mass" that should be present in the peak has been redistributed by the FCT limiter, flattening the peak.

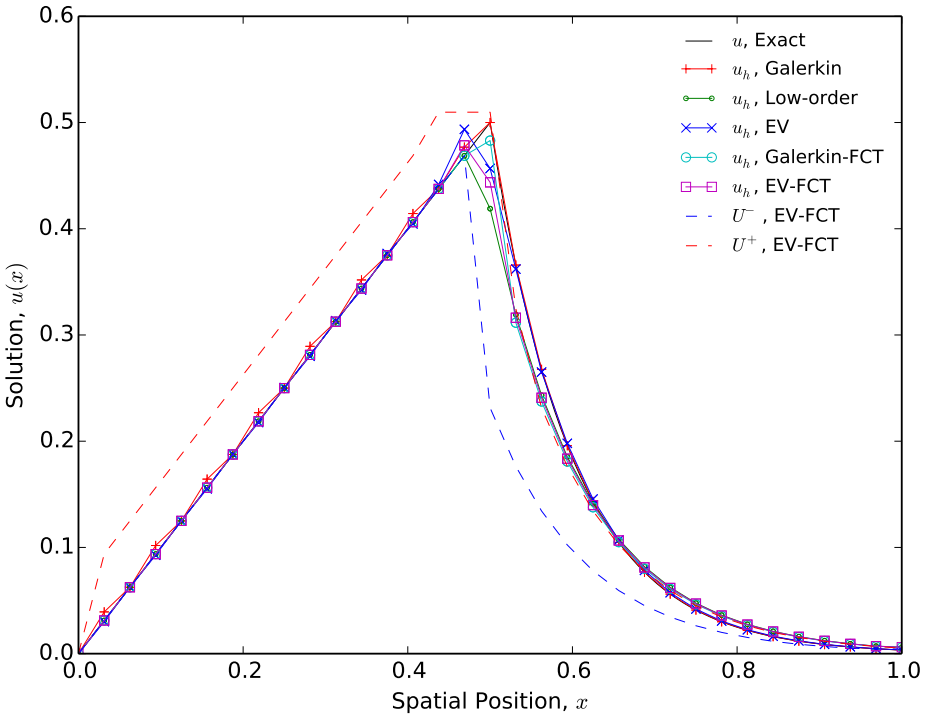


Figure 6: Comparison of Solutions for the Source-in-Void Test Problem Using Steady-State Time Discretization

5. Conclusions

An FCT scheme has been proposed for the particle transport equation. It is has been applied to the following model equation problem, an advection problem with reaction and source terms,

$$\frac{\partial u}{\partial t} + v\boldsymbol{\Omega} \cdot \nabla u(\mathbf{x}, t) + \sigma(\mathbf{x}) u(\mathbf{x}, t) = q(\mathbf{x}, t)\,, \tag{51}$$

which corresponds to the classic Source Iteration equation for particle transport, where extraneous, inscatter, and fission sources are collected in the right-hand-side term q. The FCT methodology relies on a lower-order solution and a high-order solution. For the low-order solution, we have used a first-order viscosity approach, based on the graph-theoric method of Guermond ~~at~~ et al. [28] for scalar conservation laws and ~~extents~~ extends these previous works to situations with reaction and source terms present. The low-order scheme is shown to be positivity-preserving through the use of the M-matrix properties. The high-order solution ~~is~~ employs an entropy-based artificial stabilization. The entropy residual approach is derived for the transport equation shown in Equation (51). Temporal discretizations include explicit and implicit schemes in time, as well as steady state, the latter two cases making the FCT algorithm implicit as well. The standard Zalesak limiter is utilized to limit between the low- and ~~the~~ high-order solutions.

The FCT scheme described in this paper is second-order accurate in space, converges to the entropy solution, and preserves non-negativity. Spurious oscillations are mitigated but are not guaranteed to be eliminated, as smaller magnitude oscillations may exist within the imposed solution bounds.

The traditional FCT phenomenon known as "stair-stepping", "terracing", or "plateauing" is still an open issue, particularly for fully explicit temporal discretizations; however, these effects have been shown to diminish or disappear when using SSPRK33 as opposed to explicit Euler. In addition, these effects are less pronounced for EV-FCT than in the classic FEM-FCT scheme, which uses the standard Galerkin method as the high-order method in FCT.

The explicit temporal discretizations of the described FCT scheme yield a robust algorithm; however, implicit and steady-state discretizations are less robust, suffering from nonlinear convergence difficulties in some problems. The main complication with implicit and steady-state FCT schemes is that the imposed solution bounds are implicit with the solution, and thus the imposed solution bounds change with each iteration of the nonlinear solver.

Future work on this subject should mainly focus on the implicit/steady-state iteration techniques because of the convergence difficulties encountered for some problems. This work used fixed-point iteration, but one can attempt using an alternative such as Newton's method. The main challenge is the evolving solution bounds, which will present difficulty to any nonlinear solution algorithm. Other work could be performed for the FCT algorithm in general: for example, the terracing phenomenon still deteriorates FCT solutions.

6. Acknowledgments

This research was carried out under the auspices of the US Department of Energy (Grant number DE-NE-0000112) and the Idaho National Laboratory, a contractor of the U.S. Government under contract No. DEAC07-05ID14517.

References

[1] J.-L. Guermond, R. Pasquetti, B. Popov, Entropy viscosity method for nonlinear conservation laws, Journal of Computational Physics 230 (2011) 4248–4267.

[2] G. I. Bell, S. Glasstone, Nuclear Reactor Theory, Litton Educational Publishing, Inc., 1970.

[3] M. Schäfer, M. Frank, M. Herty, Optimal treatment planning in radiotherapy based on Boltzmann transport calculations, Mathematical Models and Methods in Applied Sciences 18 (4) (2008) 573–592. doi:10.1142/S0218202508002784.

[4] P. Bodenheimer, et al., Numerical Methods in Astrophysics: An Introduction, CRC Press, 2006.

[5] E. E. Lewis, W. F. Miller, Computational Methods of Neutron Transport, American Nuclear Society, La Grange Park, IL, 1993.

[6] K. Eidmann, Radiation transport and atomic physics modeling in high-energy-density laser-produced plasmas, Laser and Particle Beams 12 (2) (1994) 223–244.

[7] J. J. Duderstadt, W. R. Martin, Transport Theory, John Wiley & Sons, 1979.

[8] P. Lesaint, P. A. Raviart, On a finite element method for solving the neutron transport equation, Publications mathmatiques et informatique de Rennes S4 (1974) 1–40.
URL http://eudml.org/doc/273730

[9] W. Reed, T. Hill, Triangular mesh methods for the neutron transport equation, Tech. Rep. LA-UR-73-479, Los Alamos Scientific Laboratory (1973).

[10] V. Zingan, J.-L. Guermond, J. Morel, B. Popov, Implementation of the entropy viscosity method with the discontinuous Galerkin method, Computer Methods in Applied Mechanics and Engineering 253 (2013) 479–490. doi:10.1016/j.cma.2012.08.018.

[11] K. D. Lathrop, Spatial differencing of the transport equation: Positivity vs. accuracy, Journal of Computational Physics 4 (1969) 475–498.

[12] S. Hamilton, M. Benzi, Negative flux fixups in discontinuous finite element sn transport, in: International Conference on Mathematics, Computational Methods & Reactor Physics, 2009.

[13] W. F. Walters, T. A. Wareing, An accurate, strictly-positive, nonlinear characteristic scheme for the discrete ordinates equations, Transport Theory and Statistical Physics 25 (2) (1996) 197–215.

[14] T. A. Wareing, An exponential discontinuous scheme for discrete-ordinate calculations in cartesian geometries, in: Joint International Conference on Mathematical Methods and Supercomputing in Nuclear Applications, Saratoga Springs, NY, 1997.

[15] P. Maginot, A nonlinear positive extension of the linear discontinuous spatial discretization of the transport equation, Master's thesis, Texas A&M University (December 2010).

[16] P. Maginot, A non-negative, non-linear Petrov-Galerkin method for bilinear discontinuous differencing of the Sn equations, in: Joint International Conference on Mathematics and Computation, Supercomputing in Nuclear Applications, and the Monte Carlo Method (M&C 2015), Nashville, TN, 2015.

[17] P. G. Maginot, J. C. Ragusa, J. E. Morel, Nonnegative methods for bilinear discontinuous differencing of the sn equations on quadrilaterals, Nuclear Science and Engineering 185 (1) (2017) 53–69.

[18] J. P. Boris, D. L. Book, Flux-corrected transport i. SHASTA, a fluid transport algorithm that works, Journal of Computational Physics 11 (1973) 38–69.

[19] S. T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, Journal of Computational Physics 31 (1979) 335–362.

[20] A. K. Parrott, M. A. Christie, Fct applied to the 2-d finite element solution of tracer transport by single phase flow in a porous medium, in: Proceedings on the ICFD Conference on Numerical Methods in Fluid Dynamics, Oxford University Press, 1986, p. 609.

[21] R. Löhner, K. Morgan, J. Peraire, M. Vahdati, Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations, International Journal for Numerical Methods in Fluids 7 (1987) 1093–1109.

[22] D. Kuzmin, R. Löhner, S. Turek, Flux-Corrected Transport, 1st Edition, Springer-Verlag Berlin Heidelberg, Germany, 2005.

[23] D. Kuzmin, On the design of general-purpose flux limiters for finite element schemes. I. scalar convection, Journal of Computational Physics 219 (2006) 513–531.

[24] M. Möller, D. Kuzmin, D. Kourounis, Implicit FEM-FCT algorithms and discrete Newton methods for transient convection problems, International Journal for Numerical Methods in Fluids 57 (2008) 761–792. doi:10.1002/fld.1654.

[25] D. Kuzmin, M. Möller, J. N. Shadid, M. Shashkov, Failsafe flux limiting and constrained data projections for equations of gas dynamics, Journal of Computational Physicsdoi:10.1016/j.jcp.2010.08.009.

[26] D. Kuzmin, Y. Gorb, A flux-corrected transport algorithm for handling the close-packing limit in dense suspensions, Journal of Computational and Applied Mathematics 236 (2012) 4944–4951. doi:10.1016/j.cam.2011.10.019.

[27] J.-L. Guermond, M. Nazarov, B. Popov, Y. Yang, A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations, SIAM Journal on Numerical Analysis 52 (2014) 2163–2182.

[28] J.-L. Guermond, M. Nazarov, A maximum-principle preserving C^0 finite element method for scalar conservation equations, Computational Methods in Applied Mechanics and Engineering 272 (2014) 198–213.

[29] S. Gottlieb, On high order strong stability preserving Runge-Kutta and multi step time discretizations, Journal of Scientific Computing 25 (1).

[30] C. B. Macdonald, Constructing high-order Runge-Kutta methods with embedded strong-stability-preserving pairs, Master's thesis, Acadia University (August 2003).

[31] R. J. Plemmins, M-matrix characterizations I. – nonsingular m-matrices, Linear Algebra and its Applications 18 (2) (1977) 175–188. doi:10.1016/0024-3795(77)90073-8.

[32] M.-O. Delchini, Entropy-based viscous regularization for the multi-d Euler equations in low-mach and transonic flows, Computers and Fluids 118 (225–244).

[33] M.-O. Delchini, J. C. Ragusa, R. A. Berry, Viscous regularization for the non-equilibrium seven-equation two-phase flow model, Journal of Scientific Computingdoi:10.1007/s10915-016-0217-6.
URL http://dx.doi.org/10.1007/s10915-016-0217-6

[34] R. J. LeVeque, Finite Volume Methods for Hyperbolic Problems, Cambridge University Press, 2002.

[35] C. Schär, P. K. Smolarkiewicz, A synchronous and iterative flux-correction formalism for coupled transport equations, Journal of Computational Physics 128 (1) (1996) 101–120. doi:10.1006/jcph.1996.0198.