SIAM J. NUMER. ANAL.
Vol. 60, No. 1, pp. 423–449

HIGH ORDER STRONG STABILITY PRESERVING MULTIDERIVATIVE IMPLICIT AND IMEX RUNGE–KUTTA METHODS WITH ASYMPTOTIC PRESERVING PROPERTIES*

SIGAL GOTTLIEB†, ZACHARY J. GRANT‡, JINGWEI HU§, AND RUIWEN SHU¶

Abstract. In this work we present a class of high order unconditionally strong stability preserving (SSP) implicit two-derivative Runge–Kutta schemes and SSP implicit-explicit (IMEX) multiderivative Runge–Kutta schemes where the time-step restriction is independent of the stiff term. The unconditional SSP property for a method of order $p > 2$ is unique among SSP methods and depends on a backward-in-time assumption on the derivative of the operator. We show that this backward derivative condition is satisfied in many relevant cases where SSP IMEX schemes are desired. We devise unconditionally SSP implicit Runge–Kutta schemes of order up to $p = 4$ and IMEX Runge–Kutta schemes of order up to $p = 3$. For the multiderivative IMEX schemes, we also derive and present the order conditions, which have not appeared previously. The unconditional SSP condition ensures that these methods are positivity preserving, and we present sufficient conditions under which such methods are also asymptotic preserving when applied to a range of problems, including a hyperbolic relaxation system, the Broadwell model, and the Bhatnagar–Gross–Krook kinetic equation. We present numerical results to support the theoretical results on a variety of problems.

Key words. asymptotic preserving, strong stability preserving, implicit-explicit scheme, multiderivative, positivity preserving, high order

AMS subject classification. 65

DOI. 10.1137/21M1403175

1. Introduction.

Explicit strong stability preserving (SSP) Runge–Kutta methods were first developed for use with total variation diminishing spatial discretizations for hyperbolic conservation laws with discontinuous solutions [22, 23]. They have proven useful in a wide variety of problems where we need to evolve an ODE, as they preserve any convex functional property satisfied by the forward Euler method while giving higher order solutions. Given a system of ODEs, generally resulting from a spatial discretization of a PDE, of the form

$$u_t = G(u) \tag{1}$$

*Received by the editors March 8, 2021; accepted for publication (in revised form) September 13, 2021; published electronically February 15, 2022.
https://doi.org/10.1137/21M1403175
Funding: The work of the first author was partially supported by the AFOSR grant FA9550-18-1-0383. The work of the second author was supported by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). The work of the third author was partially supported by the National Science Foundation CAREER grant DMS-1654152.
†Department of Mathematics, University of Massachusetts Dartmouth, Dartmouth, MA 02747 USA (sgottlieb@umassd.edu).
‡Multiscale Methods and Dynamics Group, Mathematics in Computation subsection, Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (zackjgrant@gmail.com).
§Department of Applied Mathematics, University of Washington, Seattle, WA 98195 USA (hujw@uw.edu).
¶Department of Mathematics, University of Maryland, College Park, College Park, MD 20742 USA (rshu@cscamm.umd.edu).

that satisfies some *forward Euler condition*:

$$\|u + \Delta t G(u)\| \le \|u\| \quad \text{for all} \quad \Delta t \le \Delta t_{\text{FE}}, \tag{2}$$

where $\|\cdot\|$ is some convex functional (e.g., positivity). In practice, we don't want to use Euler's method. Instead, we desire a higher order method that preserves the forward Euler condition, perhaps under a modified time-step restriction $\Delta t \le \mathcal{C}\Delta t_{\text{FE}}$. Higher order methods that can be written as convex combinations of forward Euler steps with $\mathcal{C} > 0$ will preserve the forward Euler condition and are called SSP. The value $\mathcal{C}$ is called the SSP coefficient, and we generally want to devise methods that have a large $\mathcal{C}$.

When concerned with linear stability properties, we turn to implicit methods, or to implicit-explicit methods, to alleviate the time-step restriction. When considering the more strict SSP property, even implicit methods suffer from a step-size restriction that is quite severe: the SSP coefficient is usually bounded by twice the number of stages for a Runge–Kutta method [20]. This is true for all implicit methods that have been tested: Runge–Kutta, multistep methods, and general linear methods. However, by using a second operator $\tilde{G}$ that approximates G and satisfies a *downwind condition*,

$$\|u - \Delta t \tilde{G}(u)\| \le \|u\| \quad \text{for all} \quad \Delta t \le \ \Delta t_{\text{FE}}, \tag{3}$$

Ketcheson found a family of implicit second order methods that are unconditionally SSP [20].

In [6, 12] the SSP properties of multiderivative Runge–Kutta methods were studied. For such methods, in addition to the forward Euler condition (2), we need some condition on the second derivative $\dot{G} = \frac{dG}{dt} = G'G$. One candidate was a *second derivative condition* [6]:

$$\|u + \Delta t^2 \dot{G}(u)\| \le \|u\| \quad \text{for all} \quad \Delta t^2 \le \tilde{k}\ \Delta t_{\text{FE}}^2, \tag{4}$$

where $\tilde{k} > 0$. The other possibility was a *Taylor series condition* [12]:

$$\|u + \Delta t G(u) + \frac{1}{2}\Delta t^2 \dot{G}(u)\| \le \|u\| \quad \text{for all} \quad \Delta t \le \hat{k}\ \Delta t_{\text{FE}}, \tag{5}$$

where $\hat{k} > 0$. Previously, explicit SSP two-derivative methods were developed that preserved the forward Euler (2) and second derivative (4) conditions [6] or the forward Euler (2) and Taylor series (5) conditions [12]. However, unconditionally implicit methods that preserve the forward Euler condition (2) cannot exist [11]. Furthermore, the proof in [11] can be easily applied to the two-derivative case to show that there are no unconditionally implicit methods that preserve (2) and (4), or (2) and (5) (see Appendix A). This leads us to consider the backward derivative condition as an alternative to (4) and (5).

To obtain unconditional SSP methods, we consider in this work a new condition on the second derivative, the *backward derivative condition*:

$$\|u - \Delta t^2 \dot{G}(u)\| \le \|u\| \quad \text{for all} \quad \Delta t^2 \le \dot{k}\ \Delta t_{\text{FE}}^2, \tag{6}$$

for some $\dot{k} > 0$. Under this condition, we *require* negative coefficients on the derivative and in this way are able to obtain unconditionally SSP two-derivative Runge–Kutta methods. In subsection 2.2, we show the conditions under which such an implicit two-derivative method is unconditionally SSP, in the sense that it preserves the strong

stability condition (6) for any positive time-step Δt. In subsection 2.3 we proceed to present unconditionally SSP methods of this type of order up to $p = 4$.

After establishing that unconditionally SSP implicit two-derivative Runge–Kutta methods of up to fourth order exist in subsection 2.3, we proceed to expand the theory in subsection 2.2 to implicit-explicit (IMEX) multiderivative Runge–Kutta methods. We devise IMEX methods that are SSP under a time-step restriction resulting only from the operator treated explicitly. We consider equations of the type

$$u_t = F(u) + G(u), \tag{7}$$

where F and G satisfy a forward Euler condition and $\dot{G}$ satisfies a backward derivative condition. Here, the condition on F requires a reasonable size time-step, but the condition on G requires an inconveniently small time-step. To alleviate this restriction, we present the multiderivative IMEX approach in section 3 and give sufficient conditions under which we can ensure the method is SSP under a time-step that depends only on F. We then derive the order conditions for multiderivative IMEX methods. In subsection 3.3 we present our new second and third order methods. A rich area of applications is described in subsection 3.4.1, where the backward derivative condition appears throughout.

One property that is desired in the problems presented in subsection 3.4.1 is positivity for time-steps that depend only on F. Being SSP, the methods in subsection 3.3 automatically preserve this positivity property. Furthermore, our methods satisfy an additional condition: that either or both G and $\dot{G}$ appear in each stage. This condition is not needed for SSP (or, equivalently, positivity), but it is valuable for an additional property that is of interest: they are asymptotic preserving, as we prove in subsection 3.4.2.

Taken together, we present unconditionally SSP—and thus positivity preserving—methods: both implicit two-derivative Runge–Kutta methods and IMEX multiderivative Runge–Kutta methods, where the time-step restriction comes from the explicit part. The IMEX methods are also asymptotic preserving, which is valuable for the problems in subsection 3.4.1. These results are significant in that unconditionally SSP methods of order $p > 1$ are rare. We are limited only by the fact that the function and its derivative must satisfy a forward Euler (2) and backward derivative (6) conditions, respectively. While the forward Euler condition (2) seems standard, the backward derivative condition (6) seems, at first glance, to be a bit unusual. However, there is a similarity between it and the downwinding condition (3). Furthermore, it turns out that it is a natural condition and quite useful for a variety of problems, as we show in subsection 3.4.1.

2. SSP implicit two-derivative Runge–Kutta methods. In this section, we consider two-derivative Runge–Kutta methods for the ODE

$$u_t = G(u).$$

As discussed in [6], the two-derivative Runge–Kutta method can be written in the Butcher form

$$u^{(i)} = u^n + \Delta t \sum_{j=1}^{i} a_{ij} G(u^{(j)}) + \Delta t^2 \sum_{j=1}^{i} \dot{a}_{ij} \dot{G}(u^{(j)}), \qquad i = 1, \dots, s, \tag{8a}$$

$$u^{n+1} = u^{(s)}. \tag{8b}$$

In matrix form, this becomes

$$U = \mathbf{e}u^n + \Delta t \mathbf{A} G(U) + \Delta t^2 \dot{\mathbf{A}} \dot{G}(U), \tag{9}$$

where $\mathbf{e}$ is a vector of ones.

We proceed to define the order conditions of such a method in the next subsection.

2.1. Formulating the order conditions. Given the Butcher form (9), the vectors $\mathbf{b}$ and $\dot{\mathbf{b}}$ are given by the last row of $\mathbf{A}$ and $\dot{\mathbf{A}}$, respectively. The vectors $\mathbf{c} = \mathbf{A}\mathbf{e}$ and $\dot{\mathbf{c}} = \dot{\mathbf{A}}\mathbf{e}$ define the time-levels at which the stages are happening; these values are known as the abscissas. The order conditions for methods of this form are given in [6] up to sixth order. We repeat them here up to fourth order:

$p = 1$: $\mathbf{b}^T\mathbf{e} = 1$,

$p = 2$: $\mathbf{b}^T\mathbf{c} + \dot{\mathbf{b}}^T\mathbf{e} = \frac{1}{2}$,

$p = 3$: $\mathbf{b}^T\mathbf{c}^2 + 2\dot{\mathbf{b}}^T\mathbf{c} = \frac{1}{3}$, $\quad \mathbf{b}^T A\mathbf{c} + \mathbf{b}^T\dot{\mathbf{c}} + \dot{\mathbf{b}}^T\mathbf{c} = \frac{1}{6}$,

$p = 4$: $\mathbf{b}^T\mathbf{c}^3 + 3\dot{\mathbf{b}}^T\mathbf{c}^2 = \frac{1}{4}$, $\quad \mathbf{b}^T\mathbf{c}A\mathbf{c} + \mathbf{b}^T\mathbf{c}\dot{\mathbf{c}} + \dot{\mathbf{b}}^T\mathbf{c}^2 + \dot{\mathbf{b}}^T A\mathbf{c} + \dot{\mathbf{b}}^T\dot{\mathbf{c}} = \frac{1}{8}$,

$\mathbf{b}^T A\mathbf{c}^2 + 2\mathbf{b}^T\dot{A}\mathbf{c} + \dot{\mathbf{b}}^T\mathbf{c}^2 = \frac{1}{12}$,

$\mathbf{b}^T A^2\mathbf{c} + \mathbf{b}^T A\dot{\mathbf{c}} + \mathbf{b}^T\dot{A}\mathbf{c} + \dot{\mathbf{b}}^T A\mathbf{c} + \dot{\mathbf{b}}^T\dot{\mathbf{c}} = \frac{1}{24}$.

2.2. SSP properties. To ensure that a method of the form (8) does not result in an SSP time-step restriction, we write the method in a special Shu–Osher form with only implicit computations:

$$u^{(i)} = r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \Delta t d_{ii} G(u^{(i)}) + \Delta t^2 \dot{d}_{ii} \dot{G}(u^{(i)}), \qquad i = 1, \ldots, s, \tag{10a}$$

$$u^{n+1} = u^{(s)}. \tag{10b}$$

This form ensures that only implicit evaluations of G and $\dot{G}$ are present, so that we do not have a time-step restriction due to a forward Euler, second derivative, or Taylor series term. The form (10) ensures that any explicit terms in the method (8) enter only after they were introduced implicitly in a prior stage. This is a necessary (but not sufficient) condition so that an SSP time-step restriction will not occur [10].

In matrix form, this becomes

$$U = \mathbf{R}\mathbf{e}u^n + \mathbf{P}U + \Delta t \mathbf{D} G(U) + \Delta t^2 \dot{\mathbf{D}} \dot{G}(U), \tag{11}$$

where $\mathbf{P}$ and $\mathbf{R} = I - \mathbf{P}$ are $s \times s$ matrices, r_i are the ith row sum of $\mathbf{R}$, and $\mathbf{D}$ and $\dot{\mathbf{D}}$ are $s \times s$ diagonal matrices. The numerical solution u^{n+1} is then given by the final element of the vector U. Note that the relationship between the Butcher form (9) and the Shu–Osher form (11) is given by

$$\mathbf{A} = \mathbf{R}^{-1}\mathbf{D}, \quad \dot{\mathbf{A}} = \mathbf{R}^{-1}\dot{\mathbf{D}}.$$

Note that given a method of the form (9), it is not always possible to select some matrix of coefficients $\mathbf{R}$ and thus obtain matrices $\mathbf{P}$, $\mathbf{D}$, and $\dot{\mathbf{D}}$, where the matrices $\mathbf{D}$

and $\dot{\mathbf{D}}$ are diagonal. (However, if $\mathbf{A}$ has only nonzero elements on the diagonal, then it is possible). On the other hand, is always possible to start from a two-derivative method of the form (11) and write it in the form (9).

A method of the form (10) will be unconditionally SSP under the following conditions.

THEOREM 1. *Let the operators G and $\dot{G}$ satisfy the forward Euler condition*

$$\|u + \Delta t G(u)\| \le \|u\| \quad \textit{for all} \quad \Delta t \le \Delta t_{\mathrm{FE}}$$

and the backward derivative condition

$$\|u - \Delta t^2 \dot{G}(u)\| \le \|u\| \quad \textit{for all} \quad \Delta t^2 \le \dot{k}\, \Delta t_{\mathrm{FE}}^2,$$

for some $\Delta t_{\mathrm{FE}} > 0$ and $\dot{k} > 0$, and for some convex functional $\|\cdot\|$. A method given by (11) *which satisfies the conditions*

$$\mathbf{Re} \ge 0, \quad \mathbf{P} \ge 0, \quad \mathbf{D} \ge 0, \quad \dot{\mathbf{D}} \le 0 \tag{12}$$

(where the inequalities are understood componentwise) will preserve the strong stability property

$$\|u^{n+1}\| \le \|u^n\|$$

for any positive time-step $\Delta t > 0$.

Proof. The first stage of the method is given by

$$u^{(1)} = u^n + \Delta t d_{11} G(u^{(1)}) + \Delta t^2 \dot{d}_{11} \dot{G}(u^{(1)}).$$

Using the forward Euler and backward derivative conditions, we can show that $\|u^{(1)}\| \le \|u^n\|$ whenever $d_{11} \ge 0$ and $\dot{d}_{11} \le 0$. To see this add $(\alpha + \beta)u^{(1)}$ to both sides and rearrange

$$\begin{aligned} u^{(1)} = {} & \frac{u^n}{1+\alpha+\beta} + \frac{\alpha}{1+\alpha+\beta}\left(u^{(1)} + \frac{1}{\alpha}\Delta t d_{11} G(u^{(1)})\right) \\ & + \frac{\beta}{1+\alpha+\beta}\left(u^{(1)} - \frac{1}{\beta}\Delta t^2 |\dot{d}_{11}| \dot{G}(u^{(1)})\right). \end{aligned}$$

Assuming that $\alpha \ge 0$ and $\beta \ge 0$ we have (from the forward Euler condition and backward derivative condition)

$$\|u^{(1)}\| \le \frac{1}{1+\alpha+\beta}\|u^n\| + \frac{\alpha}{1+\alpha+\beta}\left\|u^{(1)}\right\| + \frac{\beta}{1+\alpha+\beta}\left\|u^{(1)}\right\|;$$

hence

$$\|u^{(1)}\| \le \|u^n\|,$$

for any Δt such that $\frac{1}{\alpha}\Delta t d_{11} \le \Delta t_{\mathrm{FE}}$ and $\frac{1}{\beta}|\dot{d}_{11}|\Delta t^2 \le \dot{k}\Delta t_{\mathrm{FE}}^2$. Since we can choose α and β to be arbitrarily large, then this is true for any Δt.

Each subsequent stage of the method is given by

$$u^{(i)} = \left(r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)}\right) + \Delta t d_{ii} G(u^{(i)}) + \Delta t^2 \dot{d}_{ii} \dot{G}(u^{(i)}),$$

where we can now assume that $\|u^{(j)}\| \le \|u^n\|$ for all $j < i$. The explicitly computed terms are

$$\begin{aligned}\|u_e^{(i)}\| &= \left\| r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} \right\| \le \|r_i u^n\| + \left\| \sum_{j=1}^{i-1} p_{ij} u^{(j)} \right\| \\ &\le r_i \|u^n\| + \sum_{j=1}^{i-1} p_{ij} \|u^{(j)}\| \le \left(r_i + \sum_{j=1}^{i-1} p_{ij} \right) \|u^n\| \\ &= \|u^n\|.\end{aligned}$$

due to the nonnegativity of r_i and p_{ij} and the fact that they sum to one. Note that this condition is independent of Δt. Finally we write each stage as

$$u^{(i)} = u_e^{(i)} + \Delta t d_{ii} G(u^{(i)}) + \Delta t^2 \dot{d}_{ii} \dot{G}(u^{(i)})$$

and use the same argument as for the first stage above to show that $\|u^{(i)}\| \le \|u_e^{(i)}\| \le \|u^n\|$ under any time-step Δt, provided only that $d_{ii} \ge 0$ and $\dot{d}_{ii} \le 0$. ∎

2.3. New SSP implicit two-derivative Runge–Kutta methods up to order $p = 4$. We found second, third, and fourth order methods that satisfy the conditions above and are unconditionally SSP.

Second order. The one-stage, second order method is simply the implicit Taylor series method

$$u^{n+1} = u^n + \Delta t G(u^{n+1}) - \frac{1}{2} \Delta t^2 \dot{G}(u^{n+1}).$$

Third order. A two-stage, third order unconditionally SSP implicit two-derivative Runge–Kutta method is given by the Shu–Osher coefficients

$$\mathbf{D} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \dot{\mathbf{D}} = \begin{bmatrix} -\frac{1}{6} & 0 \\ 0 & -\frac{1}{3} \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{Re} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and the Butcher coefficients

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \dot{\mathbf{A}} = \begin{bmatrix} -\frac{1}{6} & 0 \\ -\frac{1}{6} & -\frac{1}{3} \end{bmatrix}.$$

Fourth order. A five-stage, fourth order unconditionally SSP implicit two-derivative Runge–Kutta method is given by the Shu–Osher coefficients

$$diag(\mathbf{D}) = \begin{bmatrix} 0.660949255604937 \\ 0.242201390400848 \\ 1.137542996287740 \\ 0.191388711018110 \\ 0.625266691721946 \end{bmatrix}, \quad diag(\dot{\mathbf{D}}) = \begin{bmatrix} -0.177750705279127 \\ -0.354733903778084 \\ -0.403963513682271 \\ -0.161628266349058 \\ -0.218859021269943 \end{bmatrix},$$

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0.084036809261019 & 0.915963190738981 & 0 & 0 & 0 \\ 0.001511648458457 & 0 & 0.090254853867587 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{Re} = [1, 0, 0, 0.908233497673956, 0]^T.$$

and Butcher coefficients

$$a_{ii} = d_{ii}, \;\; a_{12} = a_{13} = a_{11}, \quad a_{14} = a_{15} = 0.060653001401867,$$

$$a_{23} = 0.221847558352979, \; a_{24} = a_{25} = 0.020022818960029,$$

$$a_{34} = a_{35} = 0.102668776898047, \;\; a_{45} = a_{44},$$

and

$$\dot{a}_{ii} = \dot{d}_{ii}, \;\; \dot{a}_{12} = a_{13} = \dot{a}_{11}, \quad \dot{a}_{14} = \dot{a}_{15} = -0.016311560509453,$$

$$\dot{a}_{23} = -0.324923198367868, \;\; \dot{a}_{24} = \dot{a}_{25} = -0.029325895786881,$$

$$\dot{a}_{34} = \dot{a}_{35} = -0.036459667895230, \;\; \dot{a}_{45} = \dot{a}_{44}.$$

We were unable to find any fifth order methods that satisfy the conditions in Theorem 1.

2.4. Numerical tests. We test all three of our methods on the nonlinear scalar problem

$$u_t = -10u^2$$

with initial condition $u(0) = 10$ with $T_{final} = 2$. Here, $G = -10u^2$ and $\dot{G} = 200u^3$. This problem satisfies the forward Euler condition for positivity:

$$u^n > 0 \;\; \Rightarrow \;\; u^{n+1} = u^n + \Delta t G(u^n) = u^n \left(1 - 10\Delta t u^n\right) > 0 \quad \text{for } \; \Delta t \leq \frac{0.1}{u^n}$$

and the backward derivative conditions for positivity:

$$u^n > 0 \; \Rightarrow \; u^{n+1} = u^n - \Delta t^2 \dot{G}(u^n) = u^n \left(1 - 200\Delta t^2 (u^n)^2\right) > 0 \text{ for } \; \Delta t^2 \leq \frac{0.005}{(u^n)^2}.$$

Note that these restrictions induce a severe time-constraint, especially as u^n is large, on an explicit method. However, as long as these (explicit-type) conditions hold for nonzero Δt, we preserve this positivity property *unconditionally* for the implicit methods we found above.

We compare our second and third order methods in the subsection above to diagonally implicit stiffly stable methods (or diagonally implicit Runge–Kutta (DIRK) methods) in the literature with Butcher tableau [19]:

Second order DIRK

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

Third order DIRK

0	0	0	0	0
$\frac{3}{2}$	$\frac{3}{4}$	$\frac{3}{4}$	0	0
$\frac{7}{5}$	$\frac{447}{675}$	$-\frac{357}{675}$	$\frac{855}{675}$	0
1	$\frac{13}{42}$	$\frac{84}{42}$	$-\frac{125}{42}$	$\frac{70}{42}$
	$\frac{13}{42}$	$\frac{84}{42}$	$-\frac{125}{42}$	$\frac{70}{42}$

As expected, the SSP methods preserve positivity up to a large time-step, while the DIRK methods lose positivity for relatively small time-steps. The second order DIRK method loses positivity for $\Delta t > \frac{1}{50}$ and the third order for $\Delta t > \frac{1}{75}$. This loss of positivity has significant consequences to the convergence of the schemes. We see in Figure 1 that the DIRK methods converge to a solution that is qualitatively poor if the time-step is not small. On the other hand, the unconditionally SSP methods converge to a solution that is qualitatively correct even for much larger time-steps.

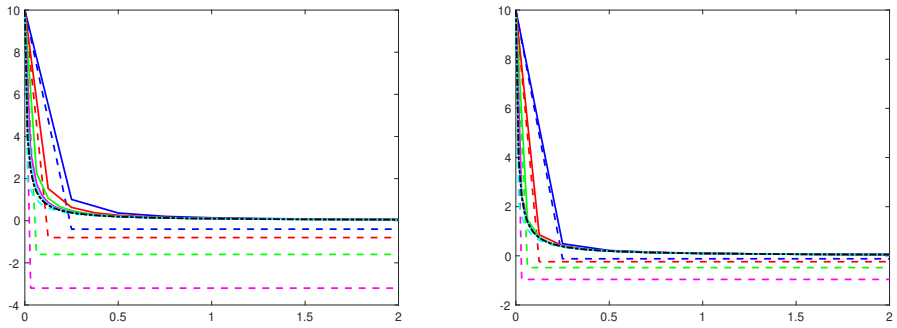

FIG. 1. *The solution of* $u' = -10u^2$ *for the diagonally implicit Runge–Kutta method (DIRK) (dashed lines) and SSP implicit multiderivative Runge–Kutta (SSP-iMDRK) (solid lines) compared to the correct solution (dash-dot line) for* $\Delta t = \frac{1}{n}$ *where* $n = 4, 8, 16, 32, 64$ *in blue, red, green, magenta, and cyan, respectively. We see that if* Δt *is not small enough the qualitative behavior of the numerical solution using the DIRK methods is poor. However, the SSP-iMDRK methods converge to a solution that is qualitatively correct for all values of* Δt *tested. Left: second order methods. Right: third order methods.*

3. Multiderivative IMEX methods. In this section we consider equations of the form (7):

$$u_t = F(u) + G(u),$$

where the time-step restriction coming from F is of a reasonable size (i.e., F is nonstiff), but the time-step restriction coming from G is very small (i.e., G is stiff). We wish to alleviate this time-step restriction. When dealing with linear stability, we typically turn to IMEX methods to alleviate the time-step restriction coming from G. However, when we consider more general norms, seminorms, or convex functionals, the use of IMEX schemes *does not* result in the removal of the time-step restriction caused by the operator G, as shown in [13, 7]. Now that we have showed that unconditional multiderivative SSP methods exist under the backward derivative conditions, we wish to leverage this knowledge to develop SSP IMEX methods that avoid a time-step restriction coming from G. We do this by using an explicit SSP solver for the nonstiff term F, coupled with a purely (or diagonally) implicit solver for the stiff term G.

We assume that the operators F and G preserve some nonlinear stability properties under a convex functional $\|\cdot\|$:

$$\textbf{Condition 1:} \quad \|u + \Delta t F(u)\| \le \|u\| \quad \text{for all} \quad \Delta t \le \Delta t_{\text{FE}},$$

for some $\Delta t_{\text{FE}} > 0$, and

$$\textbf{Condition 2:} \quad \|u + \Delta t G(u)\| \le \|u\| \quad \text{for all} \quad \Delta t \le k \Delta t_{\text{FE}},$$

for some $k > 0$, which may be very small.

The backward derivative condition is natural and relevant in many cases (see subsection 3.4.1); we assume that $\dot{G}(u) = G'(u)G(u)$ satisfies

$$\textbf{Condition 3:} \quad \|u - \Delta t^2 \dot{G}(u)\| \le \|u\| \quad \text{for all} \quad \Delta t^2 \le \dot{k}\, \Delta t_{\text{FE}}^2$$

(where $\dot{k} > 0$ can be of any size). Just as above for the implicit methods, we can devise SSP IMEX methods where there is no time-step restriction coming from G or $\dot{G}$, so that the time-step restriction depends only on F.

For problem (7), we propose an s-stage multiderivative IMEX method, written in the Shu–Osher formulation, as follows:

$$u^{(i)} = r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(u^{(j)} + \frac{\Delta t}{r} F(u^{(j)}) \right) + \Delta t d_{ii} G(u^{(i)}) + \Delta t^2 \dot{d}_{ii} \dot{G}(u^{(i)}), \qquad i = 1, \ldots, s, \tag{13a}$$

$$u^{n+1} = u^{(s)}. \tag{13b}$$

The value of $r > 0$ in the canonical Shu–Osher formulation gives us the SSP coefficient of the explicit method. While at first glance it seems that requiring all the forward Euler steps in the method to have the same time-step $\frac{\Delta t}{r}$ is restrictive, in fact this form does not result in loss of generality, as discussed in [10]. Note that the terms G and $\dot{G}$ appear only implicitly, so that there is no SSP restriction arising from the implicit method.

The intermediate stages can be conveniently written in a matrix form:

$$U = \mathbf{R}\mathbf{e} u^n + \mathbf{P} U + \mathbf{W} \left(U + \frac{\Delta t}{r} F(U) \right) + \Delta t \mathbf{D} G(U) + \Delta t^2 \dot{\mathbf{D}} \dot{G}(U), \tag{14}$$

where $\mathbf{P}$, $\mathbf{W}$, and $\mathbf{R} = I - \mathbf{P} - \mathbf{W}$ are $s \times s$ matrices, r_i are the ith row sum of $\mathbf{R}$, $\mathbf{D}$ and $\dot{\mathbf{D}}$ are $s \times s$ diagonal matrices, and $\mathbf{e}$ is a vector of ones. The numerical solution u^{n+1} is then given by the final element of the vector U.

3.1. SSP properties of multiderivative IMEX Runge–Kutta. The Shu–Osher form allows us to easily observe the ssp properties of the method.

THEOREM 2. *Given operators F and G that satisfy Conditions* 1, 2, *and* 3, *with values $\Delta t_{\mathrm{FE}} > 0$, $k > 0$, $\dot{k} > 0$, for some convex functional $\|\cdot\|$, and if the method given by* (14) *with $r > 0$ satisfies the componentwise conditions*

$$\mathbf{R}\mathbf{e} \geq 0, \quad \mathbf{P} \geq 0, \quad \mathbf{W} \geq 0, \quad \mathbf{D} \geq 0, \quad \dot{\mathbf{D}} \leq 0, \tag{15}$$

then it preserves the strong stability property

$$\|u^{n+1}\| \leq \|u^n\|$$

under the time-step condition

$$\Delta t \leq r \Delta t_{\mathrm{FE}}.$$

Proof. Each stage of the method is

$$u^{(i)} = \left(r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(u^{(j)} + \frac{\Delta t}{r} F(u^{(j)}) \right) \right) + \left(\Delta t d_{ii} G(u^{(i)}) + \Delta t^2 \dot{d}_{ii} \dot{G}(u^{(i)}) \right).$$

In particular, the first stage is

$$u^{(1)} = u^n + \left(\Delta t d_{11} G(u^{(i)}) + \Delta t^2 \dot{d}_{11} \dot{G}(u^{(i)}) \right).$$

Following the argument in the proof of Theorem 1, we easily show that $\|u^{(1)}\| \leq \|u^n\|$.

Now we assume that for the ith stage, we start with the $i-1$ previous stage values, each of which satisfy $\|u^{(j)}\| \le \|u^n\|$.The explicit part of the ith stage is defined by

$$u_e^i = r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(u^{(j)} + \frac{\Delta t}{r} F(u^{(j)}) \right).$$

We note that this value depends only on previous stages and the operator F. Given the nonnegativity of all the coefficients (12) we can show that

$$\begin{aligned}
\|u_e^i\| &= \left\| r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(u^{(j)} + \frac{\Delta t}{r} F(u^{(j)}) \right) \right\| \\
&\le r_i \|u^n\| + \sum_{j=1}^{i-1} p_{ij} \|u^{(j)}\| + \sum_{j=1}^{i-1} w_{ij} \left\| u^{(j)} + \frac{\Delta t}{r} F(u^{(j)}) \right\| \\
&\le r_i \|u^n\| + \sum_{j=1}^{i-1} p_{ij} \|u^{(j)}\| + \sum_{j=1}^{i-1} w_{ij} \left\| u^{(j)} \right\|
\end{aligned}$$

for all $\Delta t \le r \Delta t_{\text{FE}}$. Now, recalling that $\|u^{(j)}\| \le \|u^n\|$ for $j < i$, we obtain $\left\|u_e^i\right\| \le \|u^n\|$, from the condition $R + W + P = I$.

We now have $u^{(i)} = u_e^i + \Delta t d_{ii} G(u^{(i)}) + \Delta t^2 \dot{d}_{ii} \dot{G}(u^{(i)})$, where $\left\|u_e^i\right\| \le \|u^n\|$. Using Conditions 2 and 3 and the argument in the proof of Theorem 1 above, we can show that $\|u^{(i)}\| \le \|u_e^i\|$, whenever $d_{ii} \ge 0$ and $\dot{d}_{ii} \le 0$, and so $\|u^{(i)}\| \le \|u^n\|$ under the time-step $\Delta t \le r \Delta t_{\text{FE}}$. ∎

In subsection 3.3 we will show that it is indeed possible to find second and third order methods that satisfy the requirements in Theorem 2. However, we first present the order conditions a method of this form must satisfy.

3.2. Formulating order conditions. The order conditions for a method (13) are generally easier to formulate if the method is written in its Butcher form:

$$u^{(i)} = u^n + \Delta t \sum_{j=1}^{i-1} \hat{a}_{ij} F(u^{(j)}) + \Delta t \sum_{j=1}^{i} a_{ij} G(u^{(j)}) + \Delta t^2 \sum_{j=1}^{i} \dot{a}_{ij} \dot{G}(u^{(j)}), \quad i = 1, \ldots, s, \tag{16a}$$

$$u^{n+1} = u^n + \Delta t \sum_{j=1}^{i-1} \hat{b}_j F(u^{(j)}) + \Delta t \sum_{j=1}^{i} b_j G(u^{(j)}) + \Delta t^2 \sum_{j=1}^{i} \dot{b}_j \dot{G}(u^{(j)}). \tag{16b}$$

To be consistent with (13), we require that $u^{n+1} = u^{(s)}$, so that $\hat{b}_j = \hat{a}_{sj}$, $b_j = a_{sj}$, $\dot{b}_j = \dot{a}_{sj}$. The intermediate stages of this method can be written in a matrix form:

$$U = \mathbf{e} u^n + \Delta t \widehat{\mathbf{A}} F(U) + \Delta t \mathbf{A} G(U) + \Delta t^2 \dot{\mathbf{A}} \dot{G}(U). \tag{17}$$

The conversion between the two formulations (14) and (17) is given by

$$\widehat{\mathbf{A}} = \frac{1}{r} \mathbf{R}^{-1} \mathbf{W}, \quad \mathbf{A} = \mathbf{R}^{-1} \mathbf{D}, \quad \dot{\mathbf{A}} = \mathbf{R}^{-1} \dot{\mathbf{D}}. \tag{18}$$

The vectors $\widehat{\mathbf{b}}$, $\mathbf{b}$, and $\dot{\mathbf{b}}$ are given by the last row of $\widehat{\mathbf{A}}$, $\mathbf{A}$, and $\dot{\mathbf{A}}$, respectively. The vectors $\mathbf{c} = \mathbf{A}\mathbf{e}$, $\dot{\mathbf{c}} = \dot{\mathbf{A}}\mathbf{e}$, and $\widehat{\mathbf{c}} = \widehat{\mathbf{A}}\mathbf{e}$ define the time-levels at which the stages are happening; these values are known as the abscissas. The order conditions for methods of this form are as follows:

For $p \geq 1$	$\mathbf{b}^t\mathbf{e} = 1$
	$\widehat{\mathbf{b}}^t\mathbf{e} = 1$
For p ≥ 2	$\mathbf{b}^t\mathbf{c} + \dot{\mathbf{b}}^t\mathbf{e} = \frac{1}{2}$
	$\mathbf{b}^t\widehat{\mathbf{c}} = \frac{1}{2}$
	$\widehat{\mathbf{b}}^t\mathbf{c} = \frac{1}{2}$
	$\widehat{\mathbf{b}}^t\widehat{\mathbf{c}} = \frac{1}{2}$
For $p \geq 3$	$\mathbf{b}^t\mathbf{A}\mathbf{c} + \dot{\mathbf{b}}^t\mathbf{c} + \mathbf{b}^t\dot{\mathbf{c}} = \frac{1}{6}$
	$\mathbf{b}^t\mathbf{A}\widehat{\mathbf{c}} + \dot{\mathbf{b}}^t\widehat{\mathbf{c}} = \frac{1}{6}$
	$\mathbf{b}^t\widehat{\mathbf{A}}\mathbf{c} = \frac{1}{6}$
	$\mathbf{b}^t\widehat{\mathbf{A}}\widehat{\mathbf{c}} = \frac{1}{6}$
For $p \geq 3$ (continued)	$\widehat{\mathbf{b}}^t\mathbf{A}\mathbf{c} + \widehat{\mathbf{b}}^t\dot{\mathbf{c}} = \frac{1}{6}$
	$\widehat{\mathbf{b}}^t\mathbf{A}\widehat{\mathbf{c}} = \frac{1}{6}$
	$\widehat{\mathbf{b}}^t\widehat{\mathbf{A}}\mathbf{c} = \frac{1}{6}$
	$\widehat{\mathbf{b}}^t\widehat{\mathbf{A}}\widehat{\mathbf{c}} = \frac{1}{6}$
	$\mathbf{b}^t(\mathbf{c}\cdot\mathbf{c}) + 2\dot{\mathbf{b}}^t\mathbf{c} = \frac{1}{3}$
	$\mathbf{b}^t(\mathbf{c}\cdot\widehat{\mathbf{c}}) + \dot{\mathbf{b}}^t\widehat{\mathbf{c}} = \frac{1}{3}$
	$\mathbf{b}^t(\widehat{\mathbf{c}}\cdot\widehat{\mathbf{c}}) = \frac{1}{3}$
	$\widehat{\mathbf{b}}^t(\mathbf{c}\cdot\mathbf{c}) = \frac{1}{3}$
	$\widehat{\mathbf{b}}^t(\mathbf{c}\cdot\widehat{\mathbf{c}}) = \frac{1}{3}$
	$\widehat{\mathbf{b}}^t(\widehat{\mathbf{c}}\cdot\widehat{\mathbf{c}}) = \frac{1}{3}$

3.3. New SSP IMEX multiderivative Runge–Kutta methods. Given functions F and G that satisfy Conditions 1–3, these IMEX methods have an explicit part that is SSP for a time-step that depends only on F, and an implicit part that is unconditionally SSP. We will later show that these methods are positivity preserving and also asymptotic preserving for the problems described in subsection 3.4.1.

3.3.1. Second order method. We begin with a method that has Shu–Osher coefficients

$$\mathbf{W} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/2 & 0 & 0 \end{bmatrix}, \quad \mathbf{Re} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

and

$$diag(\mathbf{D}) = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}, \quad diag(\dot{\mathbf{D}}) = -\begin{bmatrix} 0 \\ \frac{1}{2} \\ 0 \end{bmatrix},$$

with $r = 1$.

In Butcher form, these become

$$\widehat{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}, \quad \dot{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & -1/4 & 0 \end{bmatrix}.$$

The benefit of this method over the one in [16] is that the positivity preserving coefficient $r = 1$ for this method is larger than the positivity preserving coefficient $r = 0.8125$ in the method given in subsection 2.6.2 of [16]. The two methods each require the implicit solution of three stages.

3.3.2. Third order method. We found a third order method of this form, as well. This method has $r = 0.904402174130635$ with coeffiicients:

$$diag(\mathbf{D}) = \begin{pmatrix} 0 \\ 2 \\ 0.388820513661584 \\ 0.083529464436389 \\ 1.793313488277995 \\ 0 \end{pmatrix}, \quad diag(\dot{\mathbf{D}}) = -\begin{pmatrix} 0.871358934880525 \\ 0.856842702601821 \\ 0 \\ 0 \\ 2 \\ 0.205134529930013 \end{pmatrix}.$$

Note that $d_{ii} + \left|\dot{d}_{ii}\right| > 0$ for each stage i:

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.058453072749259 & 0 & 0 & 0 & 0 & 0 \\ 0.764266518291495 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.292520982667463 & 0 & 0 & 0 \\ 0.173788618990251 & 0 & 0 & 0.281050180194829 & 0 & 0 \\ 0.016811671845949 & 0 & 0 & 0.448630511341543 & 0 & 0 \end{pmatrix},$$

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.253395246357353 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.235733481708505 & 0 & 0 & 0 & 0 \\ 0 & 0.123961833526104 & 0 & 0 & 0 & 0 \\ 0.409037644509411 & 0.136123556305509 & 0 & 0 & 0 & 0 \\ 0.203353399602184 & 0 & 0 & 0 & 0.331204417210324 & 0 \end{pmatrix},$$

$$\mathbf{Re} = \begin{pmatrix} 1 \\ 0.688151680893388 \\ 0 \\ 0.583517183806433 \\ 0 \\ 0 \end{pmatrix}.$$

We have the Butcher form coefficient matrices:

$$\hat{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.064631725156397 & 0 & 0 & 0 & 0 & 0 \\ 0.860287477078593 & 0 & 0 & 0 & 0 & 0 \\ 0.259664005325885 & 0 & 0.323441264334256 & 0 & 0 & 0 \\ 0.273935075266107 & 0 & 0.090903225623586 & 0.310757966128278 & 0 & 0 \\ 0.225810414773773 & 0 & 0.175213169672431 & 0.598976415553796 & 0 & 0 \end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0.471466963417009 & 0.388820513661584 & 0 & 0 & 0 \\ 0 & 0.385837646486197 & 0.113738158737554 & 0.083529464436389 & 0 & 0 \\ 0 & 0.380686852681912 & 0.031966130008218 & 0.023475971031425 & 1.793313488277995 & 0 \\ 0 & 0.299183707820065 & 0.061613731773316 & 0.045249211646092 & 0.593953348760527 & 0 \end{pmatrix},$$

and

$$\dot{\mathbf{A}} = -\begin{pmatrix} 0.871358934880525 & 0 & 0 & 0 & 0 & 0 \\ 0.271731819181020 & 0.856842702601821 & 0 & 0 & 0 & 0 \\ 0.730006747169852 & 0.201986513560852 & 0 & 0 & 0 & 0 \\ 0.247226665569066 & 0.165301085890380 & 0 & 0 & 0 & 0 \\ 0.614323072678900 & 0.163094375848475 & 0 & 0 & 2 & 0 \\ 0.506222742811925 & 0.128176688391489 & 0 & 0 & 0.662408834420649 & 0.205134529930013 \end{pmatrix}.$$

To achieve a third order method we required six stages. However, this allowed us to design a third order method that is SSP with a time-step restriction that does not depend on G.

3.4. Applications. The new SSP multiderivative IMEX methods developed in subsection 3.3 are of particular use for a number of models we describe in subsection 3.4.1. These are all problems that lead to ODE systems of the form

$$\frac{\mathrm{d}u}{\mathrm{d}t} = T(u) + \frac{1}{\varepsilon}Q(u), \tag{19}$$

where the solution $u(t) \in \mathbb{R}^N$ and the operators T, Q: $\mathbb{R}^N \to \mathbb{R}^N$, $N \geq 2$. The parameter $0 < \varepsilon \leq O(1)$ indicates the regime of the problem: $\varepsilon = O(1)$ corresponds to the nonstiff regime, $\varepsilon \ll 1$ to the stiff regime. For such systems, we require a high order time discretization that preserves the physical properties at the discrete level, in particular positivity and the asymptotic limit.

Positivity. Problems of the form (19) that are of interest to us have positive solutions. It is preferable that the numerical solution will preserve this positivity property, for a time-step not dependent on ε. It should be pointed out that positivity is an important property when solving kinetic equations. For example, the Bhatnagar–Gross–Krook (BGK) model (see (41) below) requires the macroscopic quantities to be positive, and even small negative values of the solution f may cause some macroscopic quantities, especially the temperature, to fail to be well-defined. In such cases, the requirement that the numerical solution remains positive for time-steps independent of ε is critical to the success of the simulation. Ssp methods are also positivity preserving, so the multiderivative IMEX methods given in subsection 3.3 will preserve these properties, with a time-step independent of ε.

Asymptotic limit. Very often the operator Q satisfies the following properties: Q is "conservative" in the sense that there exists a linear operator $\mathcal{R}$: $\mathbb{R}^N \to \mathbb{R}^n$, $n < N$, such that $\mathcal{R}Q(u) = 0$ for all u; Q is dissipative and has a unique local equilibrium of the form $E(\mathcal{R}u)$, where E: $\mathbb{R}^n \to \mathbb{R}^N$ is some operator. Using these properties, applying $\mathcal{R}$ to (19) yields

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \mathcal{R}T(u), \quad \omega := \mathcal{R}u, \tag{20}$$

which is not a closed system. However, if $\varepsilon \to 0$, (19) implies $Q(u) \to 0$; hence $u \to E(\omega)$. Substituting this u into (20) gives a closed (reduced) system:

$$\frac{\mathrm{d}\omega}{\mathrm{d}t} = \mathcal{R}T(E(\omega)). \tag{21}$$

The above simple analysis reveals that when $\varepsilon \to 0$, (19) not only is stiff but also possesses a *nontrivial asymptotic limit*. (Recall that $n < N$ and note that the original variable u is in $\mathbb{R}^N$ while the reduced variable ω is in $\mathbb{R}^n$.)

Systems of the form (19) arise (after the method of lines discretization of a PDE) from many physical problems in multiscale modeling. A prominent example is the Boltzmann equation in kinetic theory [4]:

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\varepsilon}Q(f), \quad x, v \in \mathbb{R}^d, \tag{22}$$

where $f = f(t, x, v) \geq 0$ is the probability density function of time t, position x, and velocity v. The term $v \cdot \nabla_x f$ describes the particle transport, and $Q(f)$ describes the collisions between particles, which is a complicated nonlinear integral operator. The dimensionless parameter ε, called the Knudsen number, is defined as the ratio of the mean free path and characteristic length scale. When $\varepsilon = O(1)$, the transport and collision balance so the system is in the fully kinetic regime. When $\varepsilon \ll 1$, the collision effect dominates, i.e., collisions happen so frequently that the overall system is close to the local equilibrium or fluid regime. In this case, one can derive the limiting fluid equations (the compressible Euler equations) as $\varepsilon \to 0$ from (22). The process is similar to the abstract model reduction procedure described above for (19).

We require a time-stepping method that preserves the asymptotic limit of the equation. That is, for a fixed Δt, when $\varepsilon \to 0$, the scheme for (19) automatically reduces to a high order time discretization for the limiting system (21). A numerical scheme with this property is called asymptotic preserving (AP) as initially coined in [18]. To insure the AP property, the time-step Δt should not be limited by the small parameter ε. This necessitates some implicit treatment of the stiff collision term $\frac{1}{\varepsilon}Q(u)$. The need for the AP property further motivates the use of IMEX methods. There is an extensive literature on development of IMEX schemes that possess the AP property; see, for instance, [21, 8, 2, 9] for the application to hyperbolic and kinetic equations.

The need for a high order numerical integrator that is both AP and positivity preserving motivated the work in this paper. We will show that the second and third order methods we presented above are AP high order time discretization methods that preserve the positivity of the solution for arbitrary ε. Previously, designing a time-stepping scheme with both positivity and AP property has proven difficult. First order IMEX schemes with these properties exist, but methods above first order may violate positivity unless the time-step is restricted by ε [13, 14].

Second order IMEX schemes that preserve the AP property and positivity for arbitrary ε have been previously found by incorporating a derivative correction term at the final stage of each time-step. Such an approach was successfully considered in [17, 16] (note that the method in [17] only works for a special relaxation system and can preserve the positivity of one component of the solution vector, while [16] works for a general class of equations and the scope is similar to what we consider in this work); however, this strategy failed to find methods of order three. By formulating IMEX multiderivative Runge–Kutta methods that allow the use of $\dot{Q}$ at every stage, we are able to obtain a third order IMEX method that is AP and positivity preserving independent of ε. Furthermore, the second order method improves upon the previously presented method in [16], in the sense that we obtain a 23% larger allowable time-step.

We present a a summary of the model equations and their properties in subsection 3.4.1. In subsection 3.4.2 we prove the positivity preserving and AP properties of the multiderivative IMEX Runge–Kutta methods. Finally, in subsection 3.5 we demonstrate the numerical performance of these methods on sample problems.

3.4.1. A summary of the models and properties. We assume that the operators T and Q in (19) satisfy the following properties.

PROPERTY 1. *The operator T is conditionally positivity preserving under a forward Euler step:*

$$u > 0 \Longrightarrow u + \Delta t T(u) > 0 \text{ for all } \ 0 \leq \Delta t \leq \Delta t_{\mathrm{FE}}, \tag{23}$$

for some time-step $\Delta t_{FE} > 0$.

PROPERTY 2. *The operator Q is unconditionally positivity preserving under a backward Euler step:*

$$u > 0,\ v = u + \Delta t Q(v) \Longrightarrow v > 0 \text{ for all } \Delta t \geq 0. \tag{24}$$

We observe that the first two properties essentially concern the positivity preserving property of the operators T and Q in (19).

Remark 1. Property 2 plays a similar role to that of Condition 2 in section 3. Condition 2 is a forward Euler condition (2), which we then use to show that the backward Euler method unconditionally preserves this strong stability property. Property 2 states that backward Euler preserves positivity unconditionally. This is necessary because positivity may be preserved under the forward Euler condition but be violated (for certain Δt) for the backward Euler method.

PROPERTY 3. *Conservation of Q: there exists a linear operator $\mathcal{R} : \mathbb{R}^N \to \mathbb{R}^n$, $n < N$, such that*

$$\mathcal{R}Q(u) = 0 \text{ for all } u. \tag{25}$$

PROPERTY 4. *Equilibrium of Q: there exists an (possibly nonlinear) operator $E : \mathbb{R}^n \to \mathbb{R}^N$ such that*

$$Q(u) = 0 \Longrightarrow u = E(\mathcal{R}u). \tag{26}$$

Moreover, E satisfies $\mathcal{R}E(\mathcal{R}u) = \mathcal{R}u$ for all u.

Note that Properties 3 and 4 together imply that (19) has a limiting system (21). Properties 1–4 are satisfied by a large class of kinetic equations of the form (22), where the collision operator Q can be the full Boltzmann collision operator (an integral-type operator), the kinetic Fokker–Planck operator (a diffusion-type operator), the BGK operator (a relaxation-type operator), or its generalized version such as the Ellipsoidal–Statistical BGK operator. For more details about these operators, we refer the readers to [15].

PROPERTY 5. *The Fréchet derivative of Q satisfies*

$$\dot{Q}(u) := Q'(u)Q(u) = -C_{\mathcal{R}u}Q(u), \tag{27}$$

where $C_{\mathcal{R}u}$ is some positive function depending only on $\mathcal{R}u$. The Fréchet derivative of Q at u is defined by

$$Q'(u)v = \lim_{\delta \to 0} \frac{Q(u + \delta v) - Q(u)}{\delta}. \tag{28}$$

Property 5 means the operator Q is dissipative in some sense. This property is not generic, but it is satisfied by quite a few kinetic models including the BGK operator and the Broadwell model. Some stiff ODE systems and hyperbolic relaxation systems also satisfy this property, though for these problems positivity is usually not a big concern compared to the kinetic equations. Since our proposed multiderivative methods highly depend on Property 5, we list below a few examples.

- **An ODE model:**

$$\begin{cases} u_1' = u_2, \\ u_2' = \dfrac{1}{\varepsilon} f(u_1)\left(g(u_1) - u_2\right), \end{cases} \tag{29}$$

where f and g are some functions of u_1, and $f(u_1) > 0$. Define

$$u = (u_1, u_2)^T, \quad T(u) = (u_2, 0)^T, \quad Q(u) = (0, f(u_1)\,(g(u_1) - u_2))^T; \tag{30}$$

then (29) falls into the general form (19). It is easy to see that (29) has a limit as $\varepsilon \to 0$:

$$u_1' = g(u_1). \tag{31}$$

Indeed, one can just take $\mathcal{R}u = u_1$ and $E(\mathcal{R}u) = E(u_1) := (u_1, g(u_1))^T$. It can also be verified by direct calculation that

$$\dot{Q}(u) = -f(u_1)Q(u). \tag{32}$$

- **A PDE model, the hyperbolic relaxation system [5]:**

$$\begin{cases} \partial_t u_1 + \partial_x u_2 = 0, \\ \partial_t u_2 + \partial_x u_1 = \dfrac{1}{\varepsilon}\left(F(u_1) - u_2\right), \end{cases} \tag{33}$$

where F is some function of u_1. Equation (33) again has the form of (19) if we define $u = (u_1, u_2)^T$, $T(u) = -(\partial_x u_2, \partial_x u_1)^T$, $Q(u) = (0, F(u_1) - u_2)^T$. Note that we abused the notation a bit: u, T, and Q should be defined for the system after spatial discretization. It is easy to see that (33) has a limit as $\varepsilon \to 0$:

$$\partial_t u_1 + \partial_x F(u_1) = 0. \tag{34}$$

Indeed, one can just take $\mathcal{R}u = u_1$ and $E(\mathcal{R}u) = E(u_1) := (u_1, F(u_1))^T$. Similarly to the previous model, it can be verified that

$$\dot{Q}(u) = -Q(u). \tag{35}$$

- **The Broadwell model [3]:** The Broadwell model is a simple discrete velocity kinetic model:

$$\begin{cases} \partial_t f_+ + \partial_x f_+ = \dfrac{1}{\varepsilon}(f_0^2 - f_+ f_-), \\ \partial_t f_0 = -\dfrac{1}{\varepsilon}(f_0^2 - f_+ f_-), \\ \partial_t f_- - \partial_x f_- = \dfrac{1}{\varepsilon}(f_0^2 - f_+ f_-), \end{cases} \tag{36}$$

where $f_+ = f_+(t,x)$, $f_0 = f_0(t,x)$, and $f_- = f_-(t,x)$ denote the densities of particles with speed 1, 0, and -1, respectively. Define $f = (f_+, f_0, f_-)^T$, $T(f) = (-\partial_x f_+, 0, \partial_x f_-)^T$, and $Q(f) = (f_0^2 - f_+ f_-, -(f_0^2 - f_+ f_-), f_0^2 - f_+ f_-)^T$ (again these should be defined for the system after spatial discretization). Then (36) falls into the general form (19). To see its limit as $\varepsilon \to 0$, we rewrite (36) using moment variables:

$$\begin{cases} \partial_t \rho + \partial_x m = 0, \\ \partial_t m + \partial_x z = 0, \\ \partial_t z + \partial_x m = \dfrac{1}{2\varepsilon}(\rho^2 + m^2 - 2\rho z), \end{cases} \tag{37}$$

where $\rho := f_+ + 2f_0 + f_-$, $m := f_+ - f_-$, and $z := f_+ + f_-$. From (37), it is clear that when $\varepsilon \to 0$, $z \to \frac{\rho^2+m^2}{2\rho}$. This, when substituted into the first two equations, yields a closed hyperbolic system:

$$\begin{cases} \partial_t \rho + \partial_x m = 0, \\ \partial_t m + \partial_x \left(\dfrac{\rho^2 + m^2}{2\rho} \right) = 0. \end{cases} \tag{38}$$

Indeed, the operators $\mathcal{R}$ and E in Properties 3–4 can be taken as

$$\mathcal{R}f = (\rho, m)^T, \tag{39}$$

$$E(\mathcal{R}f) = E((\rho, m)^T) := \left(\frac{(\rho+m)^2}{4\rho}, \frac{\rho^2 - m^2}{4\rho}, \frac{(\rho-m)^2}{4\rho} \right)^T.$$

Furthermore, it can be verified that

$$\dot{Q}(f) = -\rho Q(f). \tag{40}$$

- **The BGK model [1]:** The BGK model is a widely used kinetic model introduced to mimic the full Boltzmann equation:

$$\partial_t f + v \cdot \nabla_x f = \frac{1}{\varepsilon}(M - f), \quad x, v \in \mathbb{R}^d, \tag{41}$$

where $f = f(t, x, v)$ is the probability density function and M is the so-called Maxwellian given by

$$M(t, x, v) = \frac{\rho(t, x)}{(2\pi T(t, x))^{d/2}} \exp\left(-\frac{|v - u(t, x)|^2}{2T(t, x)} \right), \tag{42}$$

where the density ρ, bulk velocity u, and temperature T are given by the moments of f:

$$\rho = \int_{\mathbb{R}^d} f \, \mathrm{d}v, \quad \rho u = \int_{\mathbb{R}^d} f v \, \mathrm{d}v, \quad \frac{1}{2}\rho d T = \frac{1}{2} \int_{\mathbb{R}^d} f |v - u|^2 \, \mathrm{d}v. \tag{43}$$

To see its asymptotic limit, we multiply (41) by $(1, v, |v|^2/2)^T$ and integrate w.r.t. v to obtain

$$\begin{cases} \partial_t \rho + \nabla_x \cdot \displaystyle\int_{\mathbb{R}^d} v f \, \mathrm{d}v = 0, \\ \partial_t (\rho u) + \nabla_x \cdot \displaystyle\int_{\mathbb{R}^d} v \otimes v f \, \mathrm{d}v = 0, \\ \partial_t \mathcal{E} + \nabla_x \cdot \displaystyle\int_{\mathbb{R}^d} \frac{1}{2} v |v|^2 f \, \mathrm{d}v = 0, \end{cases} \tag{44}$$

where $\mathcal{E} = \frac{1}{2}\rho u^2 + \frac{1}{2}\rho d T$ is the total energy. This system is not closed. However, if $\varepsilon \to 0$, (41) implies $f \to M$. Substituting this f into (44), we can get a closed system:

$$\begin{cases} \partial_t \rho + \nabla_x \cdot (\rho u) = 0, \\ \partial_t (\rho u) + \nabla_x \cdot (\rho u \otimes u + pI) = 0, \\ \partial_t \mathcal{E} + \nabla_x \cdot ((\mathcal{E} + p) u) = 0, \end{cases} \tag{45}$$

where I is the identity matrix and $p = \rho T$ is the pressure. Equation (45) is nothing but the compressible Euler equations. To write the BGK model into the form (19), we define $T(f) = -v \cdot \nabla_x f$ and $Q(f) = M - f$ (these should be defined for (41) after spatial and velocity discretization). Moreover, the operators $\mathcal{R}$ and E are given by

$$\mathcal{R}f = \int_{\mathbb{R}^d} f(1, v, |v|^2/2)^T \, \mathrm{d}v = (\rho, \rho u, \mathcal{E})^T, \tag{46}$$

$$E(\mathcal{R}f) = E((\rho, \rho u, \mathcal{E})^T) = M. \tag{47}$$

Furthermore, it can be verified that

$$\dot{Q}(f) = -Q(f). \tag{48}$$

To summarize, we have introduced four different models (including both ODE and PDEs) which all satisfy Properties 3–5. For the Broadwell model and BGK model, one can check that they also satisfy the positivity preserving Properties 1–2 provided a positivity preserving spatial discretization is used for the transport/convection term; see [16] for more details.

3.4.2. Properties of the numerical scheme. A multiderivative IMEX method that is SSP as shown in subsection 3.1 will also be positivity preserving. This is because the SSP property holds for any convex functional, and positivity is preserved under a convex functional. In Proposition 3.1 we show this explicitly, and we also prove that under a mild additional condition satisfied by the methods in subsection 3.3, the AP property is satisfied as well.

PROPOSITION 3.1. *Assume that the problem* (19) *satisfies the Properties* 1–5 *listed in subsection* 3.4.1. *Then the scheme* (13) *that satisfies the inequalities (elementwise)*

$$\mathbf{Re} \geq 0, \quad \mathbf{P} \geq 0, \quad \mathbf{W} \geq 0, \quad \mathbf{D} \geq 0, \quad \dot{\mathbf{D}} \leq 0 \tag{49}$$

will preserve the positivity of the solution for all $\Delta t \leq r\Delta t_{\mathrm{FE}}$. *Furthermore, if we require that at least one of* Q *or* $\dot{Q}$ *appears at every stage, i.e., the strict inequality*

$$d_{ii} + \left|\dot{d}_{ii}\right| > 0 \quad \textit{for all} \quad i = 1, \ldots, s \tag{50}$$

is also satisfied, then the scheme is AP, i.e., when Δt *is fixed, as* $\varepsilon \to 0$, (13) *automatically reduces to an explicit Runge–Kutta scheme, with the same order as the original scheme, applied to the limiting system* (21).

Proof. We consider each stage of (13),

$$\begin{aligned} u^{(i)} &= r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(u^{(j)} + \frac{\Delta t}{r} T(u^{(j)}) \right) \\ &\quad + \frac{\Delta t}{\varepsilon} d_{ii} Q(u^{(i)}) + \frac{\Delta t^2}{\varepsilon^2} \dot{d}_{ii} \dot{Q}(u^{(i)}) \\ &= r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(u^{(j)} + \frac{\Delta t}{r} T(u^{(j)}) \right) \\ &\quad + \left(\frac{\Delta t}{\varepsilon} d_{ii} - \frac{\Delta t^2}{\varepsilon^2} \dot{d}_{ii} C_{\mathcal{R}u^{(i)}} \right) Q(u^{(i)}), \end{aligned} \tag{51}$$

where we applied Property 5 to the last term $\dot{Q}(u^{(i)})$.

At the first stage, we have

$$u^{(1)} = u^n + \left(\frac{\Delta t}{\varepsilon} d_{11} - \frac{\Delta t^2}{\varepsilon^2} \dot{d}_{11} C_{\mathcal{R}u^{(1)}}\right) Q(u^{(1)}).$$

Given a positive u^n, and since $d_{11} \geq 0$, $\dot{d}_{11} \leq 0$, and $C_{\mathcal{R}u^{(1)}} > 0$, using Property 2 we obtain $u^{(1)} > 0$.

Now, given a positive u^n and positive stages $u^{(j)}$ for $j < i$, Property 1 gives us the positivity of the explicit terms:

$$\left(u^{(j)} + \frac{\Delta t}{r} T(u^{(j)})\right) > 0 \quad \text{for all } \frac{\Delta t}{r} \leq \Delta t_{\mathrm{FE}}.$$

Consequently, the nonnegativity of r_i, p_{ij}, and w_{ij}, together with the fact that $r_i + \sum_{j=1}^{i-1}(p_{ij} + w_{ij}) = 1$, ensures the positivity of the explicit terms in $u^{(i)}$:

$$r_i u^n + \sum_{j=1}^{i-1} p_{ij} u^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(u^{(j)} + \frac{\Delta t}{r} T(u^{(j)})\right) > 0.$$

Finally, since $d_{ii} \geq 0$, $\dot{d}_{ii} \leq 0$, and $C_{\mathcal{R}u^{(i)}} > 0$, Property 2 ensures that $u^{(i)} > 0$.

To see the AP property, we apply $\mathcal{R}$ to (51) to obtain (define $\omega^n = \mathcal{R}u^n$, $\omega^{(i)} = \mathcal{R}u^{(i)}$)

$$\omega^{(i)} = r_i \omega^n + \sum_{j=1}^{i-1} p_{ij} \omega^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(\omega^{(j)} + \frac{\Delta t}{r} \mathcal{R}T(u^{(j)})\right), \tag{52}$$

where the collision terms are gone due to Property 3. On the other hand, when Δt is fixed and $\varepsilon \to 0$, since $d_{ii} + |\dot{d}_{ii}| > 0$ and $C_{\mathcal{R}u^{(i)}} > 0$, we have from (51) that $Q(u^{(i)}) \to 0$; hence $u^{(i)} \to E(\omega^{(i)})$ by Property 4. Note that this holds for every $i = 1, \ldots, s$. Replacing $u^{(j)}$ by $E(\omega^{(j)})$ in (52) yields

$$\omega^{(i)} = r_i \omega^n + \sum_{j=1}^{i-1} p_{ij} \omega^{(j)} + \sum_{j=1}^{i-1} w_{ij} \left(\omega^{(j)} + \frac{\Delta t}{r} \mathcal{R}T(E(\omega^{(j)}))\right), \quad i = 1, \ldots, s;$$

together with $\omega^{n+1} = \omega^{(s)}$, this is a high order explicit Runge–Kutta scheme applied to the limiting system (21). In fact, it is the explicit part of (13) applied to (21). ∎

Remark 2. Following the classification of various IMEX Runge–Kutta schemes in [2], the multiderivative IMEX schemes introduced in this paper are both type A and globally stiffly accurate. In other words, since $d_{ii} + |\dot{d}_{ii}| > 0$ for all i, we are solving an implicit collision step at every stage of the scheme; hence any initial condition is allowed to guarantee the AP property.

Remark 3. In the case of the Broadwell model and BGK equation, Theorem 2 can be used to prove the discrete entropy decay property of the numerical method. Taking the following 1D BGK equation as an example,

$$\partial_t f + v \partial_x f = \frac{1}{\varepsilon}(M - f). \tag{53}$$

We set G to be the BGK operator and F to be the transport operator discretized by the first order upwind method (k is the spatial index):

$$(v\partial_x f)_k = \frac{v+|v|}{2}\frac{f_k - f_{k-1}}{\Delta x} + \frac{v-|v|}{2}\frac{f_{k+1}-f_k}{\Delta x}, \tag{54}$$

together with the periodic or compactly supported boundary condition. The convex functional $\|\cdot\|$ is taken as the discrete entropy

$$S[f] = \Delta x \sum_k \int f_k \log f_k \, \mathrm{d}v. \tag{55}$$

Then it can be verified that F and G satisfy the Conditions 1–3 (for more details see [16]). Therefore, the numerical solution obtained by method (14) satisfies

$$S[f^{n+1}] \le S[f^n] \tag{56}$$

under the conditions listed in Theorem 2.

3.5. Numerical results. In this subsection, we verify the accuracy of the proposed second and third order methods in subsection 3.3 on the ODE model, the Broadwell model, and the BGK equation. We will see that the methods exhibit the design accuracy in the kinetic regime $\varepsilon = O(1)$ as well as the fluid regime $\varepsilon \ll 1$. This latter behavior is exactly due to the AP property of the methods. For completeness, we also report the results of the methods in the intermediate regime (i.e., ε lies between 0 and 1), where the methods may exhibit some order reduction as expected. A careful study of this behavior is beyond the scope of the current work and left for future work.

Remark 4. Note that the order conditions in subsection 3.2 do not guarantee that we will not observe order reduction. When $\varepsilon = O(1)$ we expect to see the design accuracy predicted by the order conditions. When $\varepsilon \ll 1$ design accuracy may not be evident due to the order reduction phenomenon. However, the AP property allows us to recover full accuracy in the asymptotic limit $\varepsilon \to 0$.

3.5.1. An ODE model. We consider the ODE model (29) with

$$f(u_1) = 1 + u_1^2, \quad g(u_1) = \sin u_1. \tag{57}$$

We take the initial data as $u(0) = (2, 0)^T$ (which is inconsistent initial data, i.e., we do not start from equilbrium) and solve (29) by the second and third order methods in subsection 3.3, up to final time $T = 1$, with various ε and Δt. To calculate the error of a numerical solution $U = [U_1, U_2]^T$, we compare with a reference solution U^{ref} obtained by the MATLAB solver `ode15s` with relative tolerance `RelTol` $= 1e-13$ and absolute tolerance `AbsTol` $= 1e-15$, and we compute the error by

$$\text{error} = |U_1(T) - U_1^{\mathrm{ref}}(T)| + |U_2(T) - U_2^{\mathrm{ref}}(T)|. \tag{58}$$

The results are shown in Figures 2 and 3. For both methods, one can see the design order accuracy in the kinetic regime ($\varepsilon = O(1)$ and Δt is relatively small) and the fluid regime ($\varepsilon \ll 1$ and Δt is not very small), while in the intermediate regime (when ε and Δt are comparable) one can see some order reduction. In Figure 3 with $\varepsilon = 1$ (and similar for $\varepsilon = 0.01, 1e-10$), one can see that the error increases as Δt decreases

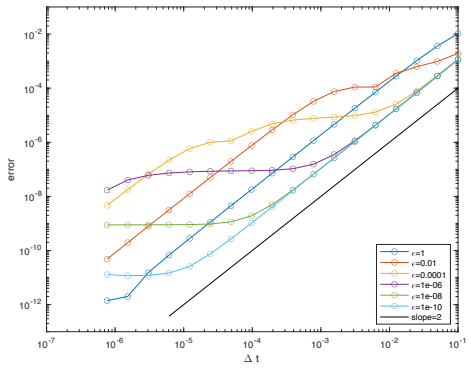

FIG. 2. *Accuracy test of the new second order IMEX scheme for an ODE model.*

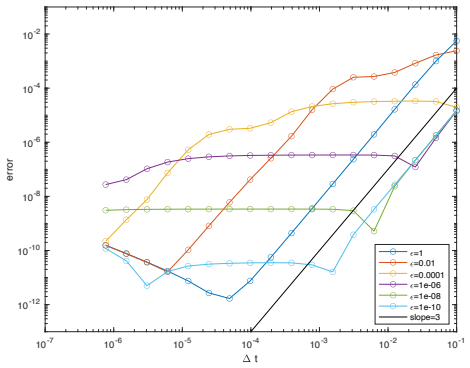

FIG. 3. *Accuracy test of the new third order IMEX scheme for an ODE model.*

when Δt is less than $5e-5$, and this is a consequence of the accumulation of round-off errors.

We note that the intermediate plateaus that are seen in Figures 2 and 3 are not an indication of the order reduction phenomenon that is usually observed in the AP literature, as we observe the errors are not converging at a rate of $O(\Delta t)$ but leveling off at the order of ε. This result is not caused by numerical round off errors, as the schemes are still converging to a "solution" at the designed order of accuracy. Indeed, if we compare the solution at time-step Δt to the $\Delta t/2$ solution, we observe design-order of convergence. The explanation for these $O(\varepsilon)$ plateaus can likely be found by looking at the higher order asymptotic expansion. In practice, these errors are of $O(\varepsilon)$ which are typically much smaller than other sources of errors in simulations and thus not typically exhibited in practice.

3.5.2. The Broadwell model. We consider the Broadwell model (36) on the domain $x \in [0, 2]$ with periodic boundary condition, with inconsistent initial data

$$f_+(0,\cdot) = 1 + 0.2\exp(0.3\sin(\pi x)), \quad f_-(0,\cdot) = \exp(0.2\cos(2\pi x)),$$
$$f_0(0,\cdot) = \frac{1}{1+0.3\sin(\pi x)}.$$

We discretize in space by the fifth order finite volume weighted essentially nonoscillatory (WENO) scheme, and the collision operator Q is evaluated pointwise on the Gauss quadrature points in each cell, as described in subsection 3.3.2 of [16]. We fix the the CFL number as $\Delta t = \frac{1}{2}\Delta x$ and solve (36) by the second and third order methods in subsection 3.3 up to final time $T = 0.1$. The error is computed by the L^2 norm of the difference between the numerical solution and one with a refined mesh. Note that in order for the fully discrete numerical scheme to be positivity preserving, one has to use the positivity preserving spatial discretization, for example, the positivity preserving finite volume WENO scheme [24], which requires a smaller CFL condition and a positivity preserving limiter. Here since our main focus is to verify the order in time discretization and the AP property, we choose a larger time-step and neglect the limiter.

The results are shown in Figures 4 and 5, and one can see similar behavior as in the previous subsection.

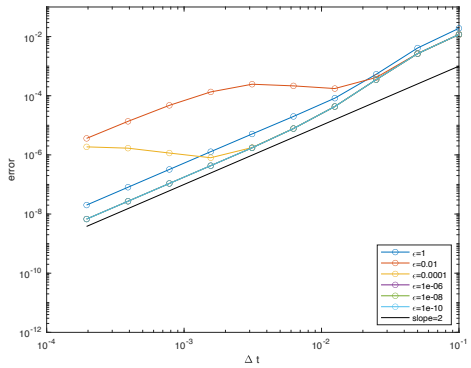

FIG. 4. *Accuracy test of the new second order IMEX scheme for the Broadwell model.*

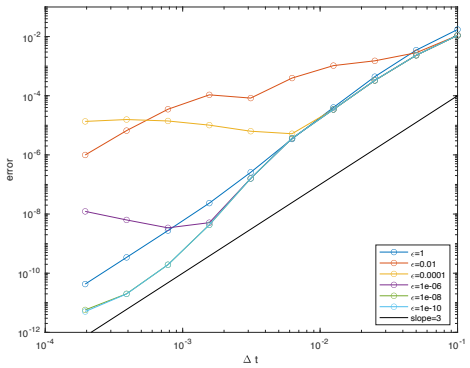

FIG. 5. *Accuracy test of the new third order IMEX scheme for the Broadwell model.*

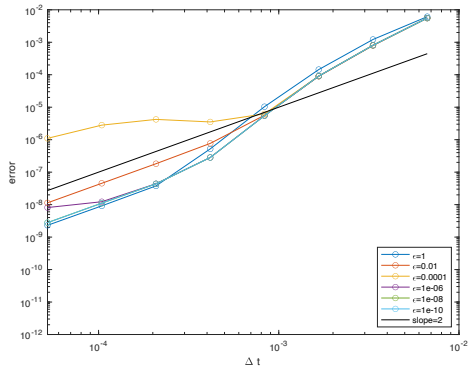

FIG. 6. *Accuracy test of the new second order IMEX scheme for the BGK model.*

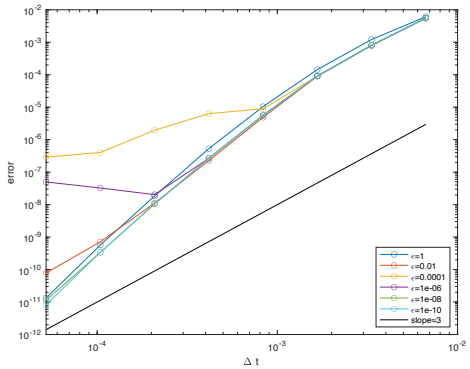

FIG. 7. *Accuracy test of the new third order IMEX scheme for the BGK model.*

3.5.3. The BGK model. We consider the 1D BGK model (41) on the physical domain $x \in [0, 2]$ with periodic boundary condition and inconsistent initial data given by

$$f(0, x, v) = 0.7M[\tilde{\rho}(x), \tilde{u}(x), \tilde{T}(x)](v) + 0.3M[\tilde{\rho}(x), -0.5\tilde{u}(x), \tilde{T}(x)](v) \tag{59}$$

with

$$\tilde{\rho}(x) = 1 + 0.2\sin(2\pi x), \quad \tilde{u}(x) = 1, \quad \tilde{T}(x) = \frac{1}{1 + 0.2\sin(\pi x)}. \tag{60}$$

The velocity domain is truncated into $[-v_{max}, v_{max}]$ with $v_{max} = 15$ and discretized with $N_v = 150$ grid points, and the physical space is discretized in the same way as in the previous subsection. We fix the the CFL number as $\Delta t = \frac{1}{2}\frac{\Delta x}{v_{max}}$ and solve (41) by the second and third order methods in subsection 3.3 up to final time $T = 0.1$. The error is computed by the L^2 norm (in the (x, v) space) of the difference between the numerical solution and one with a refined mesh. Note that the velocity space discretization may introduce some additional error such that Properties 3 and 4 in subsection 3.4.1 may not hold exactly. Here we chose a large velocity domain truncation and many grid points to make sure that the error from the velocity space discretization is negligible.

The results are shown in Figures 6 and 7. For the second order scheme, one can see clearly the second order accuracy when Δt is small enough (so that the temporal

error dominates) for both $\varepsilon = O(1)$ and $\varepsilon \ll 1$, and order reduction is observed in the intermediate regime. For the third order scheme, when $\varepsilon = O(1)$ or $\varepsilon \ll 1$, the error converges at a higher than expected rate even for the smallest Δt in the simulation, which suggests that the spatial error is still dominating. By comparing with the results of the second order scheme we see that the third order scheme indeed gives a much smaller error under the same time-step size.

Finally, to check the AP as well as the positivity preserving properties, we use the second and third order multiderivative IMEX methods in subsection 3.3 to solve a mixed regime problem, i.e., (41) with a variable Knudsen number $\varepsilon = \varepsilon(x)$ specified as below. This numerical example is comparable to the numerical result in section 5.3 of [16].

We take the physical domain as $x \in [0, 2]$ with periodic boundary condition, and the variable Knudsen number

$$\varepsilon(x) = \varepsilon_0 + (\tanh(1 - 11(x-1)) + \tanh(1 + 11(x-1))), \quad \varepsilon_0 = 10^{-5}, \tag{61}$$

so that the problem is in the kinetic regime ($\varepsilon(x) = O(1)$) near $x = 1$, and in the fluid regime ($\varepsilon(x) \approx 10^{-5}$) for x away from 1. The initial data is taken the same as equations (5.1)–(5.2) in [16]. The final time is taken as $T = 0.5$. For the new multiderivative IMEX methods, we discretize the physical space by the fifth order finite volume WENO scheme with positivity preserving limiters in [24], and the velocity space is discretized in the same way as before. The variable Knudsen number is treated by a Gauss–Legendre quadrature in each spatial cell in the same way as section 3.3.3 of [16]. We take $N_x = 40$ and $\Delta t = \frac{1}{24}\frac{\Delta x}{v_{max}}$ to satisfy the positivity preserving CFL condition.

In the simulation we tracked the numerical values (cell averages in the physical space) of f, and no negative cell is observed. The numerical solutions are compared with a reference solution obtained by the explicit second order SSP Runge–Kutta scheme with $N_x = 80$ and $\Delta t = \frac{1}{240}\frac{\Delta x}{v_{max}} \approx 7 \times 10^{-6}$, for which the smallest value of the Knudsen number (around 10^{-5}) is resolved. The result is shown in Figure 8 in terms of the macroscopic quantities. One can see good agreement between the solution by the new schemes and the reference solution. This verifies the AP and positivity preserving properties of the new multiderivative IMEX methods.

4. Conclusions. In this work, we presented a class of unconditionally SSP implicit multiderivative Runge–Kutta schemes. The unconditional SSP methods of order $p > 2$ are novel and are enabled by the backward derivative condition. This condition is an alternative to the second derivative conditions given in [6, 12] and is highly relevant to a range of problems, as shown in section 3.4.1.

The new backward derivative condition, which enabled the unconditionally SSP schemes, was inspired by the work in [16] which derived positivity preserving and AP IMEX Runge–Kutta methods with a derivative correction term. We formulate multiderivative IMEX Runge–Kutta methods that allow us to obtain order $p > 2$ and to ensure that the method is positivity preserving and AP when applied to problems that satisfy the five properties in subsection 3.4.1. In particular, we focus on an application area that includes a hyperbolic relaxation model, the Broadwell model, and the BGK kinetic equation. Such methods require treatment with an IMEX time-stepping approach, and it is desired that the method be AP and positivity preserving.

We derived and presented order conditions for SSP IMEX multiderivative Runge–Kutta methods, and we devised implicit methods that achieve fourth order and IMEX methods that are third order and are SSP under a time-step restriction independent

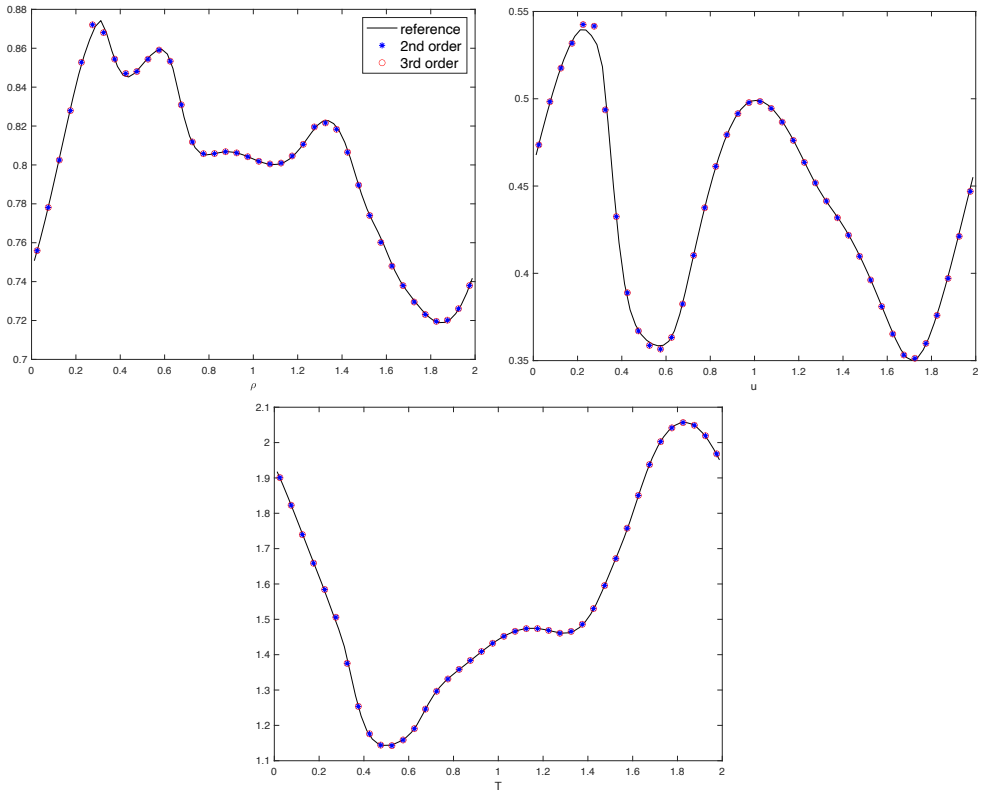

FIG. 8. *The mixed regime problem for the BGK model. Top left: density ρ; top right: bulk velocity u; bottom: temperature T. Asterisks/circles: numerical solutions by the new second/third order schemes. Solid line: the reference solution. The numerical solutions of the two new schemes are very close to each other because the spatial error is dominating.*

of the stiff term. The SSP condition ensures that the multiderivative IMEX schemes are positivity preserving, and we present sufficient conditions under which such methods are also AP when applied to the problems of interest. While we focused in the numerical examples on the IMEX schemes applied to a hyperbolic relaxation system, the Broadwell model, and the BGK equation, we stress that the results in this paper are of broad use. Any problems with operators that satisfy the forward Euler and—if handled implicitly—the backward derivative condition can benefit from these methods which are SSP with a time-step that does not depend on the function handled implicitly.

Appendix A. Unconditionally SSP implicit methods. Previously, explicit SSP two-derivative methods were developed that preserved the forward Euler (2) and second derivative (4) conditions [6] or the forward Euler (2) and Taylor series (5) conditions [12]. Methods that preserve the strong stability properties of these conditions require nonnegative coefficients on the prior stages, the function, and its derivative [6, 12]. In other words, we require that (elementwise)

$$\mathbf{Re} \geq 0, \quad \mathbf{P} \geq 0, \quad \mathbf{D} \geq 0, \quad \dot{\mathbf{D}} \geq 0. \tag{62}$$

We show here that a method of the form (10) that satisfies the conditions (62) cannot be second order. This is simply a restatement of the proof in [11] in the current notation.

The first and second order conditions are

$$\mathbf{b}^T\mathbf{e} = 1, \quad \mathbf{b}^T\mathbf{c} + \dot{\mathbf{b}}^T\mathbf{e} = \frac{1}{2}.$$

Recall that $\mathbf{b}^T$ is the final row of $\mathbf{A}$ and that $\mathbf{c}$ is the row sum of $\mathbf{A}$. Note that the matrix $\mathbf{A} = \mathbf{R}^{-1}\mathbf{D} = (I - \mathbf{P})^{-1}\mathbf{D}$ can be written as

$$\mathbf{A} = (I + \mathbf{P} + \mathbf{P}^2 + \cdots + \mathbf{P}^{s-1})\mathbf{D},$$

a consequence of the fact that $\mathbf{P}$ is strictly lower triangular and so $\mathbf{P}^s$ becomes zero. Let's look at each row of $\mathbf{Ae}$ and $\mathbf{Ac}$ using the recursive nature of the matrix multiplication: The first row is simply $(\mathbf{Ae})_1 = \mathbf{D}_{11}$ and $(\mathbf{Ac})_1 = \mathbf{D}_{11}^2$; the other rows are

$$(\mathbf{Ae})_i = \mathbf{D}_{ii} + \sum_{j=1}^{i-1} p_{ij} (\mathbf{Ae})_j \qquad (\mathbf{Ac})_i = \mathbf{D}_{ii}(\mathbf{Ae})_i + \sum_{j=1}^{i-1} p_{ij} (\mathbf{Ac})_j .$$

For any real number a the first row satisfies

$$(1-a)(\mathbf{Ae})_1 - (\mathbf{Ac})_1 = (1-a)\mathbf{D}_{11} - \mathbf{D}_{11}^2 \le k_1(1-a)^2,$$

where $k_1 = \frac{1}{4}$. We define

$$k_i = \frac{1}{4\,(1 - k_{i-1})}$$

and observe that $\frac{1}{4} = k_1 < k_2 < \cdots k_s < \frac{1}{2}$. Now we can show recursively that if

$$(1-a)(\mathbf{Ae})_j - (\mathbf{Ac})_j \le k_j(1-a)^2 \quad \text{for all } j < i,$$

then

$$\begin{aligned}
&(1-a)(\mathbf{Ae})_i - (\mathbf{Ac})_i \\
&\quad = (1-a)\left(\mathbf{D}_{ii} + \sum_{j=1}^{i-1} p_{ij} (\mathbf{Ae})_j\right) - \left(\mathbf{D}_{ii}(\mathbf{Ae})_i + \sum_{j=1}^{i-1} p_{ij} (\mathbf{Ac})_j\right) \\
&\quad = (1-a)\mathbf{D}_{ii} - \mathbf{D}_{ii}(\mathbf{Ae})_i + \sum_{j=1}^{i-1} p_{ij}\left((1-a)(\mathbf{Ae})_j - (\mathbf{Ac})_j\right) \\
&\quad = (1-a)\mathbf{D}_{ii} - \mathbf{D}_{ii}^2 - \mathbf{D}_{ii}\sum_{j=1}^{i-1} p_{ij} (\mathbf{Ae})_j + \sum_{j=1}^{i-1} p_{ij}\left((1-a)(\mathbf{Ae})_j - (\mathbf{Ac})_j\right) \\
&\quad = (1-a)\mathbf{D}_{ii} - \mathbf{D}_{ii}^2 + \sum_{j=1}^{i-1} p_{ij}\left((1-a-\mathbf{D}_{ii})(\mathbf{Ae})_j - (\mathbf{Ac})_j\right) \\
&\quad < (1-a-\mathbf{D}_{ii})\,\mathbf{D}_{ii} + k_{i-1}\,(1-a-\mathbf{D}_{ii})^2 .
\end{aligned}$$

We look at this final term and observe that it obtains a minimum at

$$\mathbf{D}_{ii} = \frac{1}{2}\frac{(1-a)(2k_{i-1}-1)}{k_{i-1}-1},$$

so that

$$(1-a)(\mathbf{Ae})_i - (\mathbf{Ac})_i \le \frac{1}{4(1-k_{i-1})}(1-a)^2 = k_i(1-a)^2.$$

Using the value $a = 0$ and looking at the final row $i = s$ we obtain

$$\mathbf{b}^T\mathbf{e} - \mathbf{b}^T\mathbf{c} = (\mathbf{A}\mathbf{e})_s - (\mathbf{A}\mathbf{c})_s \leq k_s < \frac{1}{2}.$$

If the method is at least first order, we must then have

$$\mathbf{b}^T\mathbf{c} > \mathbf{b}^T\mathbf{e} - \frac{1}{2} = \frac{1}{2}.$$

We can then conclude that if

$$\mathbf{b}^T\mathbf{c} + \dot{\mathbf{b}}^T\mathbf{e} = \frac{1}{2}$$

and all the coefficients of $\mathbf{A}$ are nonnegative, then $\dot{\mathbf{b}}$ must have negative coefficients or the method cannot be second order.

This argument above shows that the conditions on the method lead to negative coefficients, and as both the forward Euler condition and either the second derivative or Taylor series condition requires positive coefficients on both the function and its derivative, the resulting method is not SSP. Thus, implicit multiderivative Runge–Kutta methods cannot be unconditionally SSP in the sense of preserving the forward Euler and one of the derivative conditions above. This leads us to consider the backward derivative condition.

REFERENCES

[1] P. Bhatnagar, E. Gross, and M. Krook, *A model for collision processes in gases.* I. *Small amplitude processes in charged and neutral one-component systems*, Phys. Rev., 94 (1954), pp. 511–525.

[2] 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.

[3] J. Broadwell, *Shock structure in a simple discrete velocity gas*, Phys. Fluids, 7 (1964), pp. 1013–1037.

[4] C. Cercignani, *The Boltzmann Equation and Its Applications*, Springer-Verlag, New York, 1988.

[5] G.-Q. Chen, C. D. Levermore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Commun. Pure Appl. Math., XLVII (1994), pp. 787–830.

[6] A. Christlieb, S. Gottlieb, Z. Grant, and D. C. Seal, *Explicit strong stability preserving multistage two-derivative time-stepping schemes*, J. Sci. Comput., 68 (2016), pp. 914–942.

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

[8] G. Dimarco and L. Pareschi, *Asymptotic preserving implicit-explicit Runge–Kutta methods for nonlinear kinetic equations*, SIAM J. Numer. Anal., 51 (2013), pp. 1064–1087.

[9] G. Dimarco and L. Pareschi, *Implicit-explicit linear multistep methods for stiff kinetic equations*, SIAM J. Numer. Anal., 55 (2017), pp. 664–690.

[10] S. Gottlieb, D. Ketcheson, and C.-W. Shu, *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*, World Scientific, River Edge, NJ, 2011.

[11] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[12] Z. Grant, S. Gottlieb, and D. Seal, *A strong stability preserving analysis for explicit multistage two-derivative time-stepping schemes based on Taylor series conditions*, Commun. Appl. Math. Comput., 1 (2019), pp. 21–59.

[13] I. Higueras, *Strong stability for additive Runge–Kutta methods*, SIAM J. Numer. Anal., 44 (2006), pp. 1735–1758.

[14] I. Higueras and T. Roldan, *Positivity-preserving and entropy-decaying IMEX methods*, Monogr. Mat. García Galdeano, 33 (2006), pp. 129–136.

[15] J. Hu and R. Shu, *A second-order asymptotic-preserving and positivity-preserving exponential Runge–Kutta method for a class of stiff kinetic equations*, Multiscale Model. Simul., 17 (2019), pp. 1123–1146.

[16] 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.

[17] J. Huang and C.-W. Shu, *A second-order asymptotic-preserving and positivity-preserving discontinuous Galerkin scheme for the Kerr-Debye model*, Math. Models Methods Appl. Sci., 27 (2017), pp. 549–579.

[18] S. Jin, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM J. Sci. Comput., 21 (1999), pp. 441–454.

[19] C. A. Kennedy and M. H. Carpenter, *Diagonally Implicit Runge-Kutta Methods for Ordinary Differential Equations. A Review*, NASA Technical Report, NASA/TM–2016–219173, 2016.

[20] D. I. Ketcheson, *Step sizes for strong stability preserving with downwind-biased operators*, SIAM J. Numer. Anal., 49 (2011), pp. 1649–1660.

[21] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta methods and applications to hyperbolic systems with relaxation*, J. Sci. Comput., 25 (2005), pp. 129–155.

[22] C.-W. Shu, *Total-variation diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), p. 1073–1084.

[23] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), p. 439–471.

[24] 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.