SIAM J. SCI. COMPUT.
Vol. 44, No. 5, pp. A3366–A3392

INVARIANT-DOMAIN-PRESERVING HIGH-ORDER TIME STEPPING: I. EXPLICIT RUNGE–KUTTA SCHEMES*

ALEXANDRE ERN† AND JEAN-LUC GUERMOND‡

Abstract. We introduce a technique that makes every explicit Runge–Kutta (ERK) time stepping method invariant domain preserving and mass conservative when applied to high-order discretizations of the Cauchy problem associated with systems of nonlinear conservation equations. The key idea is that at each stage of the ERK scheme one computes a low-order update, a high-order update, both defined from the same intermediate stage, and then one applies the nonlinear, mass conservative limiting operator. The main advantage over the strong stability preserving (SSP) paradigm is more flexibility in the choice of the ERK scheme, thus allowing for less stringent restrictions on the time step. The technique is agnostic to the space discretization. It can be combined with continuous finite elements, discontinuous finite elements, and finite volume discretizations in space. Numerical experiments are presented to illustrate the theory.

Key words. time integration, Runge–Kutta, invariant domain preserving, strong stability preserving, conservation equations, hyperbolic systems, high-order method

MSC codes. 35L65, 65M60, 65M12, 65N30

DOI. 10.1137/21M145793X

1. Introduction.

This paper is the first part of a work devoted to the construction of invariant-domain preserving, high-order time stepping schemes. In this first part, we deal with explicit Runge–Kutta (ERK) schemes. In the forthcoming second part, we extend the proposed techniques to implicit-explicit (IMEX) Runge–Kutta schemes. The goal of this section is to motivate the problem under consideration and to discuss our objectives.

1.1. Position of the problem.

Our main motivation lies in the approximation of the Cauchy problem for nonlinear conservation equations posed over a space domain $D \subset \mathbb{R}^d$ and a time interval $[0,T]$ with $T>0$:

$$\partial_t \boldsymbol{u} = -\nabla\cdot\mathbb{f}(\boldsymbol{u}), \qquad \boldsymbol{u}(0)=\boldsymbol{u}_0. \tag{1.1}$$

The dependent variable, $\boldsymbol{u}$, is assumed to take values in $\mathbb{R}^m$, $m\geq 1$. The $\mathbb{R}^{m\times d}$-valued function $\mathbb{f}$ is called flux. Since there is no general theory for the existence and uniqueness of solutions to the Cauchy problem (1.1), we assume that (1.1) admits a reasonable definition of solutions and there exists a nontrivial invariant domain $\mathcal{A}\subset\mathbb{R}^m$ for these solutions. This means that if the initial datum takes values in $\mathcal{A}$ everywhere in D (and in the absence of perturbations due to the boundary conditions),

*Submitted to the journal's Methods and Algorithms for Scientific Computing section November 8, 2021; accepted for publication (in revised form) August 5, 2022; published electronically October 17, 2022.

https://doi.org/10.1137/21M145793X

Funding: This work was supported by the National Science Foundation grant DMS2110868, the Air Force Office of Scientific Research, USAF, under grant/contract FA9550-18-1-0397, the Army Research Office, under grant W911NF-19-1-0431, and the U.S. Department of Energy by Lawrence Livermore National Laboratory under contracts B640889. The support of INRIA through the International Chair program is acknowledged.

†CERMICS, Ecole des Ponts, 77455 Marne-la-Vallee Cedex 2, France, and INRIA Paris, 75589 Paris, France (alexandre.ern@enpc.fr).

‡Department of Mathematics, Texas A&M University, College Station, TX 77843 USA (guermond@math.tamu.edu).

then the solution also takes values in $\mathcal{A}$ everywhere in D at all times $t \in [0, T]$. Another important property of (1.1) is conservation, meaning that (again up to perturbations due to the boundary conditions) the integral over D of the solution is constant in time.

For scalar conservation equations with Lipchitz flux, $m = 1$, the solutions one is interested in are the entropy solutions. These solutions satisfy the maximum principle, and the invariant domains are intervals $[\alpha, \beta] \subset \mathbb{R}$. More precisely, if the initial datum is bounded from below by α and from above by β in D, then it is also the case of the entropy solution at all times $t \in [0, T]$. In the general case of hyperbolic systems, $m > 1$, the existence of invariant domains can be established by adding a second-order perturbation like $\epsilon \boldsymbol{\Delta u}$ on the right-hand side of (1.1). These domains are independent of $\epsilon > 0$ (see Chueh, Conley, and Smoller [5]). For instance, for the compressible Euler equations equipped with the co-volume equation of state, the set $\mathcal{A}$ is composed of the states with positive density, positive internal energy, density less than the maximal compressibility constant from the co-volume equation of state, and specific entropy larger than the minimum of the specific entropy of the initial state. For the shallow water equations, the set $\mathcal{A}$ is composed of the states with positive water height. In the theory of hyperbolic systems, invariant domains are usually convex.

When approximating the Cauchy problem (1.1), it is important to devise approximation methods that are high-order accurate in space and time, invariant-domain preserving (IDP), and conservative. When the invariant domain $\mathcal{A}$ is a convex set, the usual paradigm in the literature to achieve this goal is to resort to strong stability preserving (SSP) ERK (SSPRK) methods. We refer the reader to Ferracina and Spijker [10], Gottlieb, Shu, and Tadmor [11], Higueras [24], Kraaijevanger [26] for reviews on SSPRK methods. The key idea in the SSPRK paradigm is that the higher-order update in time is obtained as a convex combination of limited forward Euler steps; see Shu and Osher [37, eq. (2.12)]. Since the limited forward Euler step is IDP under a CFL restriction on the time step and since the set $\mathcal{A}$ is convex, the SSP update stays in this set. Notice that the nonlinear conservative limiter has to be applied at each stage of the SSPRK method.

1.2. Objectives of the paper. The objective of this work is to go beyond the SSPRK paradigm and introduce a technique that makes every ERK method IDP and conservative. We call the resulting time stepping techniques "IDP-ERK methods." There are three main reasons that led us to investigate this question.

The first reason is that the class of SSP methods is restricted in accuracy. For instance, SPPRK methods are restricted to fourth-order if one insists on never stepping backward in time. This statement is proved in Ruuth and Spiteri [35, Thm. 4.1]. The methodology we propose in this paper breaks this order barrier.

The second reason is efficiency. The efficiency of an explicit s-stage Runge–Kutta method is defined as follows.

DEFINITION 1.1 (efficiency ratio). *Let τ^* be the maximal time step that makes the forward Euler method IDP. Consider some s-stage ERK method and let $\tilde{\tau}$ be the maximal time step that makes this method IDP as well. We call the* efficiency ratio *of the s-stage ERK method the ratio $c_{\text{eff}} := \frac{\tilde{\tau}}{s\tau^*}$.*

The rationale behind this definition is that the number of operations to reach a fixed time T for an s-stage ERK method using the time step $\tilde{\tau}$ is roughly equal to $\frac{1}{c_{\text{eff}}}$ times the number of operations needed by the forward Euler method using the time step τ^*. As one always has $c_{\text{eff}} \leq 1$, one is interested in devising ERK methods

that have an efficiency ratio equal to 1. Unfortunately, many SSPRK methods have an efficiency ratio that is (significantly) smaller than one. We show in this paper that every s-stage ERK method, in particular those for which $c_{\rm eff} = 1$, can be made IDP.

The third, and foremost, reason is that the SSP setting is difficult to deploy in the context of methods combining implicit and explicit (IMEX) time stepping. This problem is particularly evident when solving the compressible Navier–Stokes equations (see, e.g., Demkowicz, Oden, and Rachowicz [7], Guermond et al. [19]). Indeed, the inviscid compressible Euler equations satisfy a minimum principle on the specific entropy (which is then an invariant-domain property of the explicit part of the problem), whereas the viscous effects of the Navier–Stokes equations (which are treated implicitly) violate this minimum principle. Also, the invariant-domain properties of the compressible Euler equations are expressed in terms of the conserved variables, whereas the invariant-domain properties induced by the viscous part of the problem are expressed in terms of the primitive variables. Furthermore, it is established in Gottlieb, Shu, and Tadmor [11, Prop. 6.2] that implicit SSPRK methods cannot be more than first-order accurate. We show in the forthcoming Part II of this work that the methodology described in this paper naturally extends to IMEX methods, i.e., every IMEX method can be made IDP.

The above difficulties with the SSP paradigm come from the requirement that updates be convex combinations of elementary steps that are IDP. In this paper, we address this difficulty by developing an alternative technique where the main idea is to perform at each stage of the IDP-ERK method the following three operations: (i) one introduces a low-order update based on a forward Euler step (from a previous stage that is already IDP); (ii) one also computes a high-order update that results from an incremental rewriting of the ERK update and which can step out of the invariant domain; (iii) one combines these two updates by applying a nonlinear, conservative limiting operation to evaluate the final IDP update of the stage. During the revision of this work, we became aware of the work by Kuzmin et al. [29, sect. 3.3], which shares the ideas of going beyond the SSP paradigm and applying a limiter after each stage of the ERK method. However, the central idea of rewriting the ERK method in incremental form and maximizing the efficiency of the method appears to be original to the present approach.

The IDP-ERK methods developed herein rely on ERK methods whose radius of absolute monotonicity can be zero. The crucial point, however, is that another concept of stability is embedded into IDP-ERK methods by means of the nonlinear limiting operation. Indeed, this operation, which is anyway needed for high-order space discretizations, ties the high-order approximate solution to the low-order IDP update produced by the forward Euler substeps. Just like for SSP methods, high-order accuracy in time is recovered if the excursions outside the invariant domain of the unlimited high-order update are small and infrequent; see, e.g., Sanders [36, Lem. 3.3], Coquel and LeFloch [6, Thm .4.3], Liu and Osher [32, Thm. 2], or Zhang and Shu [41, Lem. 2.4].

The rest of this paper is organized as follows. We introduce the main ideas behind IDP-ERK methods in section 2. To pinpoint the key ideas while avoiding distracting technicalities, we ignore conservation issues in this section. Then, in section 3, we show how to modify the IDP-ERK methods from section 2 to make them conservative. Numerical tests illustrating the proposed methodology on various IDP-ERK methods of order $p \in \{2, 3, 4, 5\}$ are reported in section 4. Finally, examples of implementation of the methods (including possible choices for the space discretization and the limiters)

are briefly outlined in section 5 for completeness.

2. Main ideas on IDP-ERK time stepping. In this section, we briefly present the discrete setting for the space and time approximation of the Cauchy problem (1.1), and we state structural assumptions that are meant to reflect the state of the art in the literature on how to make the forward Euler scheme IDP. Then we present the main novel idea on how to devise higher-order IDP-ERK schemes. To avoid distracting technicalities, we do not discuss conservation here. Conservation is addressed in the next section.

2.1. Discrete setting. We start the approximation process of (1.1) by applying the method of lines, i.e., we start with the space approximation. Let I be the total number of degrees of freedom involved in the space approximation. This leads us to consider (time-dependent) vectors $\mathbf{U}(t) \in (\mathbb{R}^m)^I$ with components $\mathbf{U}_{p,i}$, where $p \in \{1{:}m\}$ and $i \in \mathcal{V} := \{1{:}I\}$. For all $i \in \mathcal{V}$, we assume that the (sub)vectors $\mathbf{U}_i(t) \in \mathbb{R}^m$ refer to an approximation of the exact solution at some point in D. In this context, the IDP property means that $\mathbf{U}_i(t) \in \mathcal{A}$ for all $i \in \mathcal{V}$ and all $t \in [0,T]$.

We consider two space discretization schemes. The low-order scheme is based on a low-order invertible mass matrix $\mathbb{M}^{\mathrm{L}} \in \mathbb{R}^{I\times I}$ and a low-order flux $\mathbf{F}^{\mathrm{L}} : \mathcal{A}^I \to (\mathbb{R}^m)^I$. The high-order scheme is based on a high-order invertible mass matrix $\mathbb{M}^{\mathrm{H}} \in \mathbb{R}^{I\times I}$ and a high-order flux $\mathbf{F}^{\mathrm{H}} : \mathcal{A}^I \to (\mathbb{R}^m)^I$. For every matrix $\mathbb{M} \in \mathbb{R}^{I\times I}$ and every vector $\mathbf{V} \in (\mathbb{R}^m)^I$ with components $\mathbf{V}_{p,i}$ where $p \in \{1{:}m\}$ and $i \in \mathcal{V}$, the components of the vector $\mathbb{M}\mathbf{V} \in (\mathbb{R}^m)^I$ are defined to be $(\mathbb{M}\mathbf{V})_{p,i} = \sum_{j\in\mathcal{V}} m_{ij}\mathbf{V}_{p,j}$. Further details on the mass matrices and fluxes are given in section 3.1. Examples using continuous finite elements and finite differences are presented in section 5. Thus, one considers the following two nonlinear systems of ordinary differential equations:

$$\mathbb{M}^{\mathrm{L}}\partial_t\mathbf{U}^{\mathrm{L}} = \mathbf{F}^{\mathrm{L}}(\mathbf{U}^{\mathrm{L}}), \qquad \mathbf{U}^{\mathrm{L}}(0) = \mathbf{U}_0, \tag{2.1a}$$

$$\mathbb{M}^{\mathrm{H}}\partial_t\mathbf{U}^{\mathrm{H}} = \mathbf{F}^{\mathrm{H}}(\mathbf{U}^{\mathrm{H}}), \qquad \mathbf{U}^{\mathrm{H}}(0) = \mathbf{U}_0, \tag{2.1b}$$

where $\mathbf{U}^{\mathrm{L}}(t)$ and $\mathbf{U}^{\mathrm{H}}(t)$ take values in $(\mathbb{R}^m)^I$ and $\mathbf{U}_0 \in \mathcal{A}^I$ is some approximation of the initial datum.

The rest of the paper consists of constructing IDP time approximations of (2.1b) using the IDP properties of the forward Euler approximation of (2.1a). Let $t^n \in [0,T]$ be the current time for all $n \in \{0{:}N\}$, with the convention that $t^0 = 0$ and $t^N = T$. Let τ be the current time step and let $t^{n+1} := t^n + \tau$. A priori, the time step τ depends on the index n, but we omit this dependency to simplify the notation. Let $(\mathbf{U}^n)_{n\in\{0:N\}}$, with $\mathbf{U}^n \in (\mathbb{R}^m)^I$ for all $n \in \{0{:}N\}$, be the sequence of vectors produced by the time stepping method. Since $\mathcal{A}$ is an invariant domain for the continuous system, it is natural to require that the whole discretization process satisfies the following invariant-domain property:

$$(\mathbf{U}^0 \in \mathcal{A}^I) \Longrightarrow (\mathbf{U}^n \in \mathcal{A}^I \ \forall n \in \{1{:}N\}). \tag{2.2}$$

Moreover, global conservation is expressed by the additional requirement that

$$\sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^n = \sum_{i\in\mathcal{V}} m_i\mathbf{U}_i^0 \qquad \forall n \in \{1{:}N\}, \tag{2.3}$$

where m_i denotes the mass associated with the ith dof.

2.2. Structural assumptions and first-order (Euler) IDP-ERK. We first recall the main steps that are usually invoked in the literature to make the forward Euler scheme IDP (no originality is claimed here); see, e.g., Boris and Book [3], Harten [21], Osher and Chakravarthy [33], Zalesak [39]. Starting from the state vector $\mathbf{U}^n$, we consider the following low-order and high-order updates at t^{n+1}:

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},n+1} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{L}}(\mathbf{U}^n), \tag{2.4}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},n+1} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{H}}(\mathbf{U}^n). \tag{2.5}$$

The first key assumption we make is that the low-order flux is constructed so that when starting from a state $\mathbf{U}^n \in \mathcal{A}^I$ in the invariant domain, the update $\mathbf{U}^{\mathrm{L},n+1}$ stays in $\mathcal{A}^I$ under a CFL restriction on the time step. In general, this assumption cannot be made for the high-order update, i.e., $\mathbf{U}^{\mathrm{H},n+1}$ may step out of the invariant domain $\mathcal{A}^I$ for every τ (this is a loose form of Godunov's theorem). It is, however, possible to devise a nonlinear limiting procedure that combines the starting state $\mathbf{U}^n$, the low-order flux $\mathbf{\Phi}^{\mathrm{L},n} := \mathbf{F}^{\mathrm{L}}(\mathbf{U}^n)$, and the high-order flux $\mathbf{\Phi}^{\mathrm{H},n} := \mathbf{F}^{\mathrm{H}}(\mathbf{U}^n)$ into an update that belongs to the invariant domain. Hence, our second key assumption is that there is a limiting operator, $\boldsymbol{\ell}$, such that

$$\mathbf{U}^{n+1} := \boldsymbol{\ell}(\mathbf{U}^n, \mathbf{\Phi}^{\mathrm{L},n}, \mathbf{\Phi}^{\mathrm{H},n}) \in \mathcal{A}^I. \tag{2.6}$$

The nonlinear limiting operator $\boldsymbol{\ell}$ is always devised so that $\mathbf{U}^{n+1}$ is as close as possible to $\mathbf{U}^{\mathrm{H},n+1}$.

Let us now formalize the above heuristic ideas into the following two structural assumptions, which we are going to invoke later:

(i) There exists a real number $\tau^* > 0$ such that the forward Euler scheme combined with the low-order space discretization is IDP, i.e., for all $\tau \in (0, \tau^*]$, we have

$$(\mathbf{V} \in \mathcal{A}^I) \Longrightarrow (\mathbf{V} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{F}^{\mathrm{L}}(\mathbf{V}) \in \mathcal{A}^I). \tag{2.7}$$

(ii) There exists a nonlinear limiting operator $\boldsymbol{\ell} : \mathcal{A}^I \times (\mathbb{R}^m)^I \times (\mathbb{R}^m)^I \to (\mathbb{R}^m)^I$ such that for all $(\mathbf{V}, \mathbf{\Phi}^{\mathrm{L}}, \mathbf{\Phi}^{\mathrm{H}}) \in \mathcal{A}^I \times (\mathbb{R}^m)^I \times (\mathbb{R}^m)^I$,

$$(\mathbf{V} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{\Phi}^{\mathrm{L}} \in \mathcal{A}^I) \Longrightarrow (\boldsymbol{\ell}(\mathbf{V}, \mathbf{\Phi}^{\mathrm{L}}, \mathbf{\Phi}^{\mathrm{H}}) \in \mathcal{A}^I). \tag{2.8}$$

Other details on the action of the nonlinear limiting operator are given in section 3 and in section 5.3; the above formalism is sufficient at this stage for our purpose.

The structural assumptions (2.7)–(2.8) are all that is needed to make the forward Euler scheme IDP. Indeed, if the time step is chosen so that $\tau \in (0, \tau^*]$, then assumption (2.7) implies that $\mathbf{U}^{\mathrm{L},n+1} \in \mathcal{A}^I$, and thus assumption (2.8) implies that $\mathbf{U}^{n+1} \in \mathcal{A}^I$. We are now going to show how these two structural assumptions allow one to make every ERK scheme IDP.

Example 2.1 (assumptions (2.7)–(2.8)). In the context of discontinuous Galerkin or finite volume approximations of conservation equations, the assumption (2.7) is achieved by using a piecewise constant approximation (often called dG(0)) with the upwind numerical flux (or Godunov numerical flux). This can also be done in the context of continuous finite elements by adding some graph viscosity, as shown in Guermond and Popov [13, eq. (3.13)]. An abstract unifying framework achieving (2.7) for continuous finite elements, discontinuous Galerkin, finite volumes, and finite differences is introduced in Guermond et al. [17, sect. 3.1]. The assumption (2.8) is

realized in the discontinuous Galerkin and finite volume settings by squeezing the high-order approximation toward the piecewise constant approximation over each mesh cell (see Sanders [36, Thm. 2.1], Coquel and LeFloch [6, Thm .4.3], Liu and Osher [32, Thm. 1], and Zhang and Shu [40, Thm. 2.5]). This can also be done for all the above discrete frameworks by using either the flux transport corrected method by Zalesak [39, eq. (4)] (for scalar conservation equations, $m = 1$) or the convex limiting method by Guermond, Popov, and Tomas [18, sect. 7] (for any $m \geq 1$).

2.3. High-order IDP-ERK. Let $s \geq 2$ be a natural number ($s = 1$ corresponds to the forward Euler scheme), and consider an s-stage ERK method described by its Butcher tableau

$$
\begin{array}{c|ccccc}
c_1 & 0 & & & & \\
c_2 & a_{2,1} & 0 & & & \\
c_3 & a_{3,1} & a_{3,2} & 0 & & \\
\vdots & \vdots & & \ddots & \ddots & \\
c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & 0 \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array}
\tag{2.9}
$$

Examples are given in section 4.1. Since we consider explicit methods, we have $a_{j,j} = 0$ for all $j \in \{1{:}s\}$. Notice that consistency requires that $\sum_{j\in\{1:s\}} b_j = 1$. Moreover, we assume that $\sum_{l\in\{1:j\}} a_{j,l} = c_j$ for all $j \in \{1{:}s\}$. This is one of Butcher's simplifying assumptions (sometimes called row-sum condition), and it implies here that $c_1 = 0$. Recall that the coefficient c_j defines the intermediate time steps $t^{n,j} := t^n + c_j\tau$. In what follows, we assume that $c_j \geq 0$ for all $j \in \{2{:}s\}$. Moreover, since it is convenient to rewrite the final stage of the ERK scheme involving the b_j's using the same formalism as in the previous stages, we conventionally set

$$
a_{s+1,k} := b_k \quad \forall k \in \{1{:}s\} \qquad \text{and} \qquad c_{s+1} := 1 \quad (\text{so that } t^{n,s+1} = t^{n+1}).
\tag{2.10}
$$

Let $\mathbf{U}^n$ be the approximation at the time t^n, which we assume to be IDP, i.e., $\mathbf{U}^n \in \mathcal{A}^I$. Our goal is to construct an approximation at the time t^{n+1} in such a way that it is also IDP, i.e., $\mathbf{U}^{n+1} \in \mathcal{A}^I$. The technique we propose is based on two key ideas. The first one is that at each stage $l \in \{2{:}s+1\}$ of the IDP-ERK method, one computes a low-order update $\mathbf{U}^{\mathrm{L},l}$ and a high-order update $\mathbf{U}^{\mathrm{H},l}$. The low-order update is IDP (under a CFL restriction on the time step), whereas the high-order update may not be, i.e., we have $\mathbf{U}^{\mathrm{L},l} \in \mathcal{A}^I$ but $\mathbf{U}^{\mathrm{H},l} \in (\mathbb{R}^m)^I$. These two updates are then combined by using the nonlinear limiting operator to deliver an update $\mathbf{U}^{n,l}$ that is again IDP, i.e., $\mathbf{U}^{n,l} \in \mathcal{A}^I$. However, the limiting process formalized in assumption (2.8) is operative only if the two updates $\mathbf{U}^{\mathrm{L},l}$ and $\mathbf{U}^{\mathrm{H},l}$ are constructed from the same starting state. Thus, the second main and original idea is to rewrite the lth-stage in incremental form. To do this, we use as starting value at the lth-stage the IDP state $\mathbf{U}^{n,l'}$, where the stage index $l' < l$ is defined to be the closest to l in the sense that the difference $c_l - c_{l'}$ is nonnegative ($c_l \geq c_{l'}$) and the smallest. More precisely, we set

$$
l'(l) := \min\{k \in \{1{:}l-1\} \mid c_l - c_k \geq 0\} \qquad \forall l \in \{2{:}s+1\}.
\tag{2.11}
$$

Notice that the set $\{k \in \{1{:}l-1\} \mid c_l - c_k \geq 0\}$ is nonempty because it contains the index 1 (since $c_l \geq 0 = c_1$ for all $l \in \{2{:}s+1\}$). Notice also that the above definition remains meaningful for so-called confluent ERK methods for which several c_l's take

the same value. If the sequence $(c_l)_{l\in\{1:s\}}$ is nondecreasing, then $l'(l) = l-1$ for all $l \in \{2{:}s+1\}$. The reason for looking for the smallest difference $c_l - c_{l'(l)}$ is that we want to minimize the CFL restriction on the time step (see Lemma 2.2). Although l' depends on l, we will simply write l' in what follows to alleviate the notation. For further reference, we define

$$\Delta c^{\max} := \max_{2\le l\le s+1} \big(c_l - c_{l'(l)}\big). \tag{2.12}$$

Notice that $\Delta c^{\max} \ge \frac{1}{s}$. Moreover, we have $\Delta c^{\max} = \frac{1}{s}$ when all the stages are equidistributed, i.e., $c_l = \frac{l-1}{s}$, $l \in \{1{:}s+1\}$.

We now describe the IDP-ERK time stepping scheme. Since $c_1 = 0$, we start by setting $\mathbf{U}^{n,1} := \mathbf{U}^n$. Then, for every stage $l \in \{2{:}s+1\}$, we assume that all the states $\mathbf{U}^{n,1}, \dots, \mathbf{U}^{n,l-1}$ have been computed and are all in $\mathcal{A}^I$ (this property will be established by induction). We first compute the provisional low-order update

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},l} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^{n,l'} + \tau(c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}). \tag{2.13}$$

Notice that this update corresponds to a forward Euler step from $t^{n,l'}$ to $t^{n,l}$. In principle, the high-order update $\mathbf{U}^{\mathrm{H},l}$ could be obtained by using the standard ERK expression which directly follows from the Butcher tableau (and by using our above convention (2.10) when $l = s+1$):

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n} + \tau \sum_{k\in\{1:l-1\}} a_{l,k}\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{2.14}$$

But, to be able to compare $\mathbf{U}^{\mathrm{H},l}$ and $\mathbf{U}^{\mathrm{L},l}$ and to perform the limiting process, we want to define $\mathbf{U}^{\mathrm{H},l}$ by using $\mathbf{U}^{n,l'}$ as the starting value. For this purpose, we proceed in two steps. First, we subtract the equation for the high-order update at the l'th-stage from the equation for the high-order update at the lth-stage. Using that the terms $a_{l',k}$ are zero for all $k \ge l'$, this gives

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l'} + \tau \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{2.15}$$

Then, we replace the invariant-domain violating state $\mathbf{U}^{\mathrm{H},l'}$ by the IDP state $\mathbf{U}^{n,l'}$ in the above equation. Thus, instead of (2.15), the equation we use for the evaluation of the provisional high-order update $\mathbf{U}^{\mathrm{H},l}$ is

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{2.16}$$

Notice that $c_l - c_{l'} = \sum_{k\in\{1:l-1\}}(a_{l,k} - a_{l',k})$ owing to Butcher's simplifying assumption. Hence, both $\mathbf{U}^{\mathrm{L},l}$ and $\mathbf{U}^{\mathrm{H},l}$ are approximations of the solution at $t^{n,l}$. Since both $\mathbf{U}^{\mathrm{L},l}$ and $\mathbf{U}^{\mathrm{H},l}$ use the IDP state $\mathbf{U}^{n,l'}$ as the starting value, it makes sense to employ the limiting operator and to set

$$\mathbf{U}^{n,l} := \boldsymbol{\ell}(\mathbf{U}^{n,l'}, \boldsymbol{\Phi}^{\mathrm{L},l}, \boldsymbol{\Phi}^{\mathrm{H},l}) \tag{2.17}$$

with the low-order and high-order fluxes defined as follows:

$$\boldsymbol{\Phi}^{\mathrm{L},l} := (c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}), \qquad \boldsymbol{\Phi}^{\mathrm{H},l} := \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{2.18}$$

To sum up, the s-stage IDP-ERK method proceeds as described in Algorithm 2.1.

Algorithm 2.1 s-stage IDP-ERK scheme.

Input: $\mathbf{U}^n \in \mathcal{A}^I$
Set $\mathbf{U}^{n,1} := \mathbf{U}^n$
for $l = 2, \dots, s+1$ **do**
1. $\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},l} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^{n,l'} + \tau(c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'})$ (Low-order update (2.13))
2. $\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau\sum_{k\in\{1:l-1\}}(a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k})$ (High-order (2.16))
3. $\mathbf{U}^{n,l} := \boldsymbol{\ell}(\mathbf{U}^{n,l'}, \boldsymbol{\Phi}^{\mathrm{L},l}, \boldsymbol{\Phi}^{\mathrm{H},l})$ with fluxes defined in (2.18) (Limiting)

end for
Set $\mathbf{U}^{n+1} := \mathbf{U}^{n,s+1}$

LEMMA 2.2 (IDP). *Let τ^* be the maximal time step from assumption* (2.7). *Assume that $\tau\Delta c^{\max} \le \tau^*$ with $\Delta c^{\max}$ defined in* (2.12). *Assume that $\mathbf{U}^n \in \mathcal{A}^I$ and that $\mathbf{U}^{n+1}$ is computed by Algorithm* 2.1. *Then, $\mathbf{U}^{n+1} \in \mathcal{A}^I$.*

Proof. We argue by induction to establish that $\mathbf{U}^{n,l} \in \mathcal{A}^I$ for all $l \in \{1{:}s+1\}$. The definition $\mathbf{U}^{n,1} := \mathbf{U}^n$ implies that the assumption holds true for $l = 1$. The assumptions $\tau(c_l - c_{l'}) \le \tau^*$, the IDP assumption (2.7), the property $\mathbf{U}^{n,l'} \in \mathcal{A}^I$ already established (since $l' < l$ by construction), and the definition of the low-order update (2.13) imply that $\mathbf{U}^{\mathrm{L},l} \in \mathcal{A}^I$. As a result, the definition (2.17) makes sense, and $\mathbf{U}^{n,l} \in \mathcal{A}^I$ by construction of the limiting operator. Hence, the induction assumption holds true for all $l \in \{1{:}s+1\}$. This implies that $\mathbf{U}^{n+1} := \mathbf{U}^{n,s+1} \in \mathcal{A}^I$. ∎

We now discuss the efficiency of the above method (see Definition 1.1). Assume that one wants to reach some fixed time T using an s-stage IDP-ERK method. Then the number of time steps required to do so is approximately $T/\tilde{\tau}$ (recall we defined $\tilde{\tau}$ to be the maximal time step that makes the method IDP). Since each time step requires estimating s fluxes and performing s limiting operations, the algebraic complexity of the method scales like $sT/\tilde{\tau}$. Similarly, the complexity of the forward Euler method to reach the same time scales like T/τ^*. Hence, the ratio of the complexity of the forward Euler method to that of the s-stage method is $\frac{T}{\tau^*}\frac{\tilde{\tau}}{sT} = \frac{\tilde{\tau}}{s\tau^*} =: c_{\mathrm{eff}}$, thereby Definition 1.1.

LEMMA 2.3 (maximal efficiency). *Consider an s-stage IDP-ERK method.*
(i) *The efficiency ratio of the method is (at least) $c_{\mathrm{eff}} \ge \frac{1}{s\Delta c^{\max}}$.*
(ii) *Maximal efficiency is reached if the stages are equidistributed, and in this case, we have $c_{\mathrm{eff}} = 1$.*

Proof. (i) From Lemma 2.2, we infer that the method is IDP for all $t \in (0, \tilde{\tau}]$ with $\tilde{\tau} = \frac{\tau^*}{\Delta c^{\max}}$. Hence, the efficiency ratio of the method is (at least) $\frac{\tau^*}{\tau^* s\Delta c^{\max}} = \frac{1}{s\Delta c^{\max}}$. (ii) Whenever the time stages are equidistributed, i.e., $c_l = \frac{l-1}{s}$ for all $l \in \{1{:}s+1\}$, we have $\Delta c^{\max} = \frac{1}{s}$ and the efficiency ratio is 1. ∎

2.4. Comparison with SSP. We now summarize the main points of comparison between the present IDP-ERK methods and the more traditional SSPRK paradigm:

1. Every ERK method with efficiency 1 can be made IDP with the proposed algorithm, whereas the SSP paradigm excludes methods that are maximally efficient. For instance, one has $c_{\mathrm{eff}} = \frac{1}{2}$ for Heun's second-order method (which is a popular second-order SSPRK method), $c_{\mathrm{eff}} = \frac{1}{3}$ for SSPRK$(3,3)$, $c_{\mathrm{eff}} = \frac{1}{2}$ for SSPRK$(4,3)$, and $c_{\mathrm{eff}} \approx 0.51$ for SSPRK$(5,4)$. (Here and in what follows, the acronym SSPRK(s,p) refers to an s-stage, pth-order SSPRK method.) Instead,

the efficiency ratio of the midpoint rule, which is not SSP, is exactly 1, which is two times larger than that of Heun's second-order method. The efficiency ratio of Heun's third-order method, which is not SSP, is exactly 1, which is three times larger than that of the popular SSPRK$(3,3)$ method. In section 4.1, we give examples of optimally efficient ERK methods of order four and five as well.

2. The computational effort deployed in each stage of an IDP-ERK(s,p) method is the same as that deployed for an SSPRK(s,p) method, i.e., for each method one needs to compute a low-order update by means of a forward Euler step, compute a high-order update, and apply the limiting operator. The flexibility of IDP-ERK methods compared to SSPRK methods is that they do not invoke a convex combination of limited states. This flexibility is paid with a slight loss in simplicity in the actual implementation of the IDP-ERK method since an incremental form of the Butcher tableau is considered.
3. The only accuracy barriers on the present methods are those on the ERK methods (for instance, one must have $s > p$ for $p \geq 5$; see Hairer, Nørsett, and Wanner [20, sect. II.5]), whereas SSPRK methods are reduced to fourth-order if one insists on never stepping backward in time (see Ruuth and Spiteri [35, Thm. 4.1]).
4. As will be shown in the forthcoming second part of this work, the present framework naturally extends to IMEX methods, which is not the case in the SSP paradigm since it is established in Gottlieb, Shu, and Tadmor [11, Prop. 6.2] that implicit SSP Runge–Kutta methods cannot be more than first-order accurate.

3. Conservative IDP-ERK time stepping. In the previous section, we only focused on the invariant-domain property (2.2). In this section, we show how to achieve the conservation property (2.3) as well. Being conservative is essential for the approximation of conservation equations, since (up to an appropriate boundedness assumption) conservation implies convergence to weak solutions with shocks moving at the right speed. The material presented in this section is inspired by the flux corrected transport literature (see Boris and Book [3], Kuzmin and Turek [27], Kuzmin, Löhner, and Turek [28], Zalesak [39]) and from the convex limiting literature (see, e.g., [17], [18]). Originality is only claimed for the content of sections 3.3 and 3.4.

3.1. Low-order and high-order mass matrices and fluxes. The assumptions we are going to make concerning the space discretization are independent of the time stepping strategy. They are common to every finite volume, finite difference, or finite element approximation technique for conservation equations.

Recall that $\mathcal{V} := \{1{:}I\}$ denotes the collection of the degrees of freedom resulting from the space discretization. The components of the low-order and high-order fluxes are denoted $\mathbf{F}_i^{\mathrm{L}}(\mathbf{V}) \in \mathbb{R}^m$ and $\mathbf{F}_i^{\mathrm{H}}(\mathbf{V}) \in \mathbb{R}^m$ for all $i \in \mathcal{V}$ and all $\mathbf{V} \in \mathcal{A}^I$. We assume that for every $i \in \mathcal{V}$, there exists a subset $\mathcal{I}(i) \subsetneq \mathcal{V}$, which we call stencil at i, so that for every $\mathbf{V} \in \mathcal{A}^I$, we have

$$\mathbf{F}_i^{\mathrm{L}}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{V}), \qquad \mathbf{F}_i^{\mathrm{H}}(\mathbf{V}) = \sum_{j\in\mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{V}), \tag{3.1}$$

where $\mathbf{F}_{ij}^{\mathrm{L}}, \mathbf{F}_{ij}^{\mathrm{H}} : \mathcal{A}^I \to \mathbb{R}^m$ are Lipschitz mappings. We assume that the stencil is symmetric, i.e., $j \in \mathcal{I}(i)$ if and only if $i \in \mathcal{I}(j)$. To express that the fluxes result from the space discretization of a conservation equation, we assume that the following skew-symmetry property holds true for all $\mathbf{V} \in \mathcal{A}^I$, all $i \in \mathcal{V}$, and all $j \in \mathcal{I}(i)$:

$$\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{V}) = -\mathbf{F}_{ji}^{\mathrm{L}}(\mathbf{V}), \qquad \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{V}) = -\mathbf{F}_{ji}^{\mathrm{H}}(\mathbf{V}). \tag{3.2}$$

Examples of fluxes are given in section 5.1 for continuous finite elements (see (5.6)–(5.7)) and in section 5.2 for fourth-order finite differences (see (5.10)–(5.11)).

Concerning the low-order mass matrix, we assume that $\mathbb{M}^{\mathrm{L}}$ is diagonal and positive, i.e.,

$$\mathbb{M}^{\mathrm{L}}_{ij} = m_i\delta_{ij}\ \forall (i,j)\in\mathcal{V}^2 \quad \text{and} \quad m_i>0\ \forall i\in\mathcal{V}. \tag{3.3}$$

This assumption is justified in [16], where is it is established that it is necessary that the mass matrix be diagonal and positive for the maximum principle to hold for scalar conservation equations. Concerning the high-order mass matrix, we assume that $\mathbb{M}^{\mathrm{H}}$ is invertible, symmetric, and sparse with the same sparsity pattern as the fluxes. Denoting by m_{ij} the entries of $\mathbb{M}^{\mathrm{H}}$, we thus assume that

$$(\mathbb{M}^{\mathrm{H}}\mathsf{X})_i = \sum_{j\in\mathcal{I}(i)} m_{ij}\mathsf{X}_j \quad \forall \mathsf{X}\in\mathbb{R}^I \quad \text{and} \quad m_{ij}=m_{ji} \quad \forall i\in\mathcal{V},\forall j\in\mathcal{I}(i). \tag{3.4}$$

We also assume that $\mathbb{M}^{\mathrm{L}}$ and $\mathbb{M}^{\mathrm{H}}$ are related by the following identity:

$$m_i = \sum_{j\in\mathcal{I}(i)} m_{ij} \qquad \forall i\in\mathcal{V}. \tag{3.5}$$

In the finite element terminology, this means that $\mathbb{M}^{\mathrm{L}}$ is the lumped version of $\mathbb{M}^{\mathrm{H}}$. This identity also means the $\mathbb{M}^{\mathrm{L}}$ and $\mathbb{M}^{\mathrm{H}}$ carry the same mass, i.e., we have $\sum_{i\in\mathcal{V}} m_i\mathbf{V}_i = \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)} m_{ij}\mathbf{V}_j$ for all $\mathbf{V}\in(\mathbb{R}^m)^I$.

Remark 3.1 (low-order and high-order stencil). For simplicity, we assumed in (3.1) that the low-order and the high-order fluxes can be decomposed over the same stencil. An example showing that this is indeed possible with finite differences using three-point and five-point stencils is discussed in section 5.2. In the finite volume/difference context, this is usually done by using multivariate numerical flux functions; see, e.g., Harten [21, eq. (1.4b)], Harten, Lax, and van Leer [23, eq. (1.10)], Osher and Chakravarthy [33, eq. 2.3]. We also refer the reader to Abgrall et al. [2] and Pazner [34], where the stencil mismatch is addressed in the finite element context.

3.2. Conservative limiting operator for forward Euler step. We present in this section and the next one a possible realization of the limiting operator introduced in section 2.8 that is conservative. We first explain the method using the forward Euler method. The method is explained in full generality in section 3.3.

We rewrite (2.4)–(2.5) in component form as follows: For all $i\in\mathcal{V}$,

$$\begin{aligned} m_i\mathbf{U}_i^{\mathrm{L},n+1} &= m_i\mathbf{U}_i^n + \tau\sum_{j\in\mathcal{I}(i)}\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^n),\\ m_i\mathbf{U}_i^{\mathrm{H},n+1} &= m_i\mathbf{U}_i^n + \sum_{j\in\mathcal{I}(i)}(m_{ij}-m_i\delta_{ij})(\mathbf{U}_j^n-\mathbf{U}_j^{\mathrm{H},n+1}) + \tau\sum_{j\in\mathcal{I}(i)}\mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^n). \end{aligned}$$

We subtract the first equation from the second one and obtain

$$\mathbf{U}_i^{\mathrm{H},n+1} = \mathbf{U}_i^{\mathrm{L},n+1} + m_i^{-1}\tau\sum_{j\in\mathcal{I}(i)}\mathbf{A}_{ij}^n \tag{3.6}$$

with

$$\mathbf{A}_{ij}^n := \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^n) - \mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^n) + \frac{m_{ij}-m_i\delta_{ij}}{\tau}(\mathbf{U}_j^n - \mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_i^n + \mathbf{U}_i^{\mathrm{H},n+1}), \tag{3.7}$$

where we used $\sum_{j\in\mathcal{I}(i)}(m_{ij}-m_i\delta_{ij})(\mathbf{U}_i^n+\mathbf{U}_i^{\mathrm{H},n+1})=\mathbf{0}$ (this is a consequence of (3.5)). The purpose of this manipulation is to obtain the following skew-symmetry property:

$$\mathbf{A}_{ij}^n=-\mathbf{A}_{ji}^n,\qquad \forall i\in\mathcal{V}\ \forall j\in\mathcal{I}(i). \tag{3.8}$$

The $\mathbb{R}^m$-valued terms $\boldsymbol{A}_{ij}^n$ are sometimes called antidiffusion coefficients in the flux corrected transport literature. We are now ready to formalize the conservative limiting operator.

Definition 3.2 (conservative limiting operator). *Let $\mathfrak{L}$ be the collection of the symmetric matrices in $\mathbb{R}^{I\times I}$ with the same sparsity pattern as $\mathbb{M}^{\mathrm{H}}$ and with coefficients in $[0,1]$. Let $\mathfrak{M}$ be the collection of the skew-symmetric matrices in $(\mathbb{R}^m)^{I\times I}$ with the same block-sparsity pattern as $\mathbb{M}^{\mathrm{H}}$. We call a conservative limiter any operator ℓ from $\mathcal{A}^I\times\mathfrak{M}$ to $\mathfrak{L}$, so that for all $\mathbf{V}\in\mathcal{A}^I$ (with coefficients $(\mathbf{V}_i)_{i\in\mathcal{V}}$) and all $\mathbf{A}\in\mathfrak{M}$ (with coefficients $(\mathbf{A}_{ij})_{i\in\mathcal{V},j\in\mathcal{I}(i)}$), the matrix $\ell(\mathbf{V},\mathbf{A})\in\mathfrak{L}$ (with coefficients $(\ell_{ij})_{i\in\mathcal{V},j\in\mathcal{I}(i)}$) is such that*

$$\mathbf{V}_i+m_i^{-1}\tau\sum_{j\in\mathcal{I}(i)}\ell_{ij}\mathbf{A}_{ij}\in\mathcal{A}\quad\forall i\in\mathcal{V}. \tag{3.9}$$

For brevity, the state $(\mathbf{V}_i+m_i^{-1}\tau\sum_{j\in\mathcal{I}(i)}\ell_{ij}\mathbf{A}_{ij})_{i\in\mathcal{V}}\in\mathcal{A}^I$ is denoted $\boldsymbol{\ell}^{\mathrm{cons}}(\mathbf{V},\mathbf{A})$.

Notice that the existence of conservative limiters is always guaranteed since the trivial limiter $\boldsymbol{\ell}^{\mathrm{cons}}(\mathbf{V},\mathbf{A})=\mathbf{V}$ (i.e., $\ell_{ij}=0$ for all $i\in\mathcal{V}$ and all $j\in\mathcal{I}(i)$) is always possible because $\mathbf{V}\in\mathcal{A}^I$. Of course, the trivial limiter is inefficient. The goal of limiters is to construct the limiting coefficients ℓ_{ij} as close to 1 as possible. Examples of conservative limiting techniques based on the above formalism are given in section 5.3.

With the help of the above definition, we can now define the conservative limited update of the forward Euler step as follows:

$$\mathbf{U}^{n+1}:=\boldsymbol{\ell}^{\mathrm{cons}}(\mathbf{U}^{\mathrm{L},n+1},\mathbf{A}^n) \tag{3.10}$$

with $\mathbf{A}^n$ defined in (3.7). The definition (3.10) can be recast into the following form:

$$\begin{aligned}\mathbf{U}_i^{n+1}=\mathbf{U}_i^n&+m_i^{-1}\tau\sum_{i\in\mathcal{V}}\ell_{ij}\mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^n)+(1-\ell_{ij})\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^n)\\&+m_i^{-1}\sum_{i\in\mathcal{V}}\ell_{ij}(m_{ij}-m_i\delta_{ij})(\mathbf{U}_j^n-\mathbf{U}_j^{\mathrm{H},n+1}-\mathbf{U}_i^n+\mathbf{U}_i^{\mathrm{H},n+1}).\end{aligned} \tag{3.11}$$

This expression shows that $\mathbf{U}^{n+1}=\mathbf{U}^{\mathrm{L},n+1}$ if all the limiter coefficients are equal to 0 and that $\mathbf{U}^{n+1}=\mathbf{U}^{\mathrm{H},n+1}$ if all the limiter coefficients are equal to 1. In what follows, we say that two generic state vectors $\mathbf{V},\mathbf{W}\in(\mathbb{R}^m)^I$ carry the same mass if $\sum_{i\in\mathcal{V}}m_i\mathbf{V}_i=\sum_{i\in\mathcal{V}}m_i\mathbf{W}_i$.

Lemma 3.3 (IDP and conservation). *The following assertions hold true:*

(i) *Let τ^* be the maximal time step from assumption (2.7). For all $\tau\in(0,\tau^*]$ and all $\mathbf{U}^n\in\mathcal{A}^I$, we have $\mathbf{U}^{n+1}\in\mathcal{A}^I$.*

(ii) *The states $\mathbf{U}^n$, $\mathbf{U}^{\mathrm{L},n+1}$, $\mathbf{U}^{\mathrm{H},n+1}$, and $\mathbf{U}^{n+1}$ all carry the same mass.*

Proof. (i) Owing to the assumptions $\tau\le\tau^*$ and (2.7), we conclude that $\mathbf{U}^{\mathrm{L},n+1}\in\mathcal{A}^I$. As a result, the definition (3.10) makes sense, and the construction of the limiting operator implies that $\mathbf{U}^{n+1}\in\mathcal{A}^I$.

(ii) The definition (3.1), the skew-symmetry assumption (3.2), and the property (3.8) imply that $\sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{\mathrm{L},n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^n = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{\mathrm{H},n+1}$. Recalling that the definition (3.10) is equivalent to

$$\mathbf{U}_i^{n+1} := \mathbf{U}_i^{\mathrm{L},n+1} + m_i^{-1}\tau \sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathbf{A}_{ij}^n \qquad \forall i \in \mathcal{V},$$

the skew-symmetry property $\ell_{ij}\mathbf{A}_{ij}^n = -\ell_{ji}\mathbf{A}_{ij}^n$ implies that $\mathbf{U}^{n+1}$ and $\mathbf{U}^{\mathrm{L},n+1}$ carry the same mass, i.e., $\sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^{\mathrm{L},n+1}$. ∎

3.3. Mass conservative limiting operator for IDP-ERK. We now rewrite the s-stage IDP-ERK scheme presented in Algorithm 2.1 using the above conservative setting. Recall that we assume $\mathbf{U}^n \in \mathcal{A}^I$ and that we set $\mathbf{U}^{n,1} := \mathbf{U}^n$. Then, at the stage $l \in \{2{:}s+1\}$ of the IDP-ERK scheme, we compute the low-order and the high-order updates by using the following definitions:

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},l} = \mathbb{M}^{\mathrm{L}}\mathbf{U}^{n,l'} + \tau(c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}), \tag{3.12}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{3.13}$$

By proceeding as in (3.7), we now define the following skew-symmetric fluxes:

$$\mathbf{A}_{ij}^{n,l} := \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^{n,k}) - (c_l - c_{l'})\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^{n,l'}) + \frac{m_{ij} - m_i\delta_{ij}}{\tau}(\mathbf{U}_j^{n,l'} - \mathbf{U}_j^{\mathrm{H},l} - \mathbf{U}_i^{n,l'} + \mathbf{U}_i^{\mathrm{H},l}). \tag{3.14}$$

Given some conservative limiting operator $\boldsymbol{\ell}^{\mathrm{cons}}$ satisfying the requirements of Definition 3.2, we then set

$$\mathbf{U}^{n,l} := \boldsymbol{\ell}^{\mathrm{cons}}(\mathbf{U}^{\mathrm{L},l}, \mathbf{A}^{n,l}) \qquad \forall l \in \{2{:}s+1\}. \tag{3.15}$$

LEMMA 3.4 (IDP and conservation). *Let $\Delta c^{\max}$ be defined in* (2.12) *(recall that $\frac{1}{s} \le \Delta c^{\max} \le 1$). Then the following holds true.*

(i) *Let τ^* be the maximal time step from assumption* (2.7). *Assume that $\mathbf{U}^n \in \mathcal{A}^I$ and $\tau \in (0, \tau^*/\Delta c^{\max}]$. Let $\mathbf{U}^{n+1}$ be computed as above. Then, $\mathbf{U}^{n+1} \in \mathcal{A}^I$.*

(ii) *The states $\mathbf{U}^n$, $\{\mathbf{U}^{n,l}\}_{l\in\{1:s\}}$, and $\mathbf{U}^{n+1}$ all carry the same mass.*

Proof. The proof combines the arguments of the proofs of Lemmas 2.2 and 3.3. ∎

3.4. High-order viscosity. In the context of nonlinear conservation equations, the low-order and the high-order fluxes always contain some artificial viscosity. The high-order update proposed in (3.13) may lose the benefit of the artificial viscosity because the sign of $a_{l,k} - a_{l',k}$ is arbitrary. We now propose a variation of the algorithm (3.12)–(3.15) that addresses this issue. We leave (3.12) unchanged, but we are more precise in our definition of the high-order flux and rewrite (3.13) as follows:

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \sum_{k\in\{1:l-1\}} \tau(a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}) + \tau(c_l - c_{l'})\mathbf{D}^{\mathrm{H}}(\mathbf{U}^{n,l'}) \tag{3.16}$$

with the artificial viscosity operator $\mathbf{D}^{\mathrm{H}}$ defined by

$$\mathbf{D}_i^{\mathrm{H}}(\mathbf{U}^{n,l'}) := \sum_{j\in\mathcal{I}(i)} d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^{n,l'} - \mathbf{U}_i^{n,l'}) \qquad \forall i \in \mathcal{V}. \tag{3.17}$$

Here, $d_{ij}^{\mathrm{H},n} \geq 0$ is some high-order viscosity coefficient satisfying $d_{ij}^{\mathrm{H},n} = d_{ji}^{\mathrm{H},n}$. Letting $(\mathcal{V}, \mathcal{E})$ be the graph such that $(i,j) \in \mathcal{E}$ if $j \in \mathcal{I}(i)$ and $i \in \mathcal{I}(j)$, the operator $\mathbf{D}^{\mathrm{H}}$ is a graph Laplacian acting $(\mathcal{V}, \mathcal{E})$.

4. Numerical illustrations. In this section, we illustrate our methodology by giving examples of ERK methods and presenting numerical results showing that all these methods can be successfully applied to approximate the nonlinear conservation equation (1.1) even if they are not SSP. The tests are performed with continuous finite elements and finite differences for the space approximation. We use the notation $\mathrm{RK}(s, p; c_{\mathrm{eff}})$, where s indicates the number of stages, p the order, and c_{eff} the efficiency ratio (see Definition 1.1 and recall that a method with optimally equidistributed substeps (i.e., with increment $\frac{1}{s}$) reaches the best possible value $c_{\mathrm{eff}} = 1$). The SSP methods are identified with the superindex ‡ instead of the prefix "SSP" to save horizontal space in the tables.

4.1. Examples of ERK methods. Three examples of ERK methods with optimally equidistributed substeps are as follows:

(4.1)

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline 1 & 0 & 1 \end{array} \qquad \mathrm{RK}(2,2;1)$$

$$\begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{3} & \frac{1}{3} & 0 & \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline 1 & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \qquad \mathrm{RK}(3,3;1)$$

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{4} & \frac{1}{4} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ \frac{3}{4} & 0 & \frac{1}{4} & \frac{1}{2} & 0 \\ \hline 1 & 0 & \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{array} \qquad \mathrm{RK}(4,3;1)$$

The method in the leftmost tableau in (4.1) is the midpoint rule. It is second-order accurate. The second and third methods in (4.1) are both third-order accurate. The second method in (4.1) is often called Heun's third-order method. The third method in (4.1) satisfies all the necessary and sufficient conditions to be fourth-order accurate on linear problems (and it is the only one with optimally equidistributed substeps to do so), but it is only third-order accurate on nonlinear problems. This proves in passing that there does not exist any four-stage ERK method that is genuinely fourth-order and has optimally equidistributed substeps.

Two examples of fourth-order accurate ERK methods are as follows:

(4.2)

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline 1 & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array} \qquad \mathrm{RK}(4,4;\tfrac{1}{2})$$

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ \frac{1}{3} & \frac{1}{3} & 0 & & & \\ \frac{2}{3} & -\frac{1}{3} & 1 & 0 & & \\ 1 & 1 & -1 & 1 & 0 & \\ \hline 1 & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} & \end{array} \qquad \mathrm{RK}(4,4;\tfrac{3}{4})$$

The first method in (4.2) is a popular fourth-order accurate ERK method, but its efficiency ratio is rather small (only $\frac{1}{2}$). The second method in (4.2) is often called the $\frac{3}{8}$ rule. It has equidistributed substeps, but the distribution is slightly suboptimal (the increment is $\frac{1}{3}$ instead of $\frac{1}{4}$) so that the efficiency ratio is $\frac{3}{4}$. We are also going to consider two other fourth-order accurate methods, called $\mathrm{RK}(5,4;1)$ and $\mathrm{RK}(6,4;1)$, which will be discussed in detail in the forthcoming second part of this work. Their explicit tableaux are constructed to be compatible with an implicit Butcher tableau

to be used in an IMEX method. The method RK$(6,4;1)$ is actually constructed to be fifth-order accurate on linear problems.

One six-stage, fifth-order accurate ERK method is the RK$(6,5;\frac{2}{3})$ method proposed by Lawson [30, Tab. 1] (using ideas from Butcher [4]). Its Butcher tableau is as follows:

(4.3)

$$\begin{array}{c|cccccc}
0 & 0 & & & & & \\
\frac{1}{4} & \frac{1}{4} & 0 & & & & \\
\frac{1}{4} & \frac{1}{8} & \frac{1}{8} & 0 & & & \\
\frac{1}{2} & 0 & -\frac{1}{2} & 1 & 0 & & \\
\frac{3}{4} & \frac{3}{16} & 0 & 0 & \frac{9}{16} & 0 & \\
1 & -\frac{3}{7} & \frac{2}{7} & \frac{12}{7} & -\frac{12}{7} & \frac{8}{7} & 0 \\
\hline
1 & \frac{7}{90} & 0 & \frac{32}{90} & \frac{12}{90} & \frac{32}{90} & \frac{7}{90}
\end{array}$$

The efficiency ratio of this method is $\frac{2}{3}$. This method belongs to a general class described in Hairer, Nørsett, and Wanner [20, p. 176], which requires that $c_6 = 1$. This constraint limits the efficiency ratio to be at most $\frac{5}{6}$.

To achieve optimality, we instead devise a seven-step, fifth-order method with the optimal choice $c_l = \frac{l-1}{7}$ for all $l \in \{1{:}7\}$, which we call RK$(7,5;1)$. There are 28 unknowns (the 21 entries of the strictly lower triangular matrix (a_{ij}) of order 7 and the 7 b_j's). We enforce Butcher's simplifying conditions on the row sums of the matrix (a_{ij}) (6 conditions) and the (linear and nonlinear) order conditions up to 5 (17 conditions). We also promote stability along the imaginary axis by requiring that the amplification function $R(z)$ satisfies $|R(i\epsilon)| = 1 + \rho_6\epsilon^6 + \mathcal{O}(\epsilon^8)$ with $\rho_6 = -0.009$ (1 condition). The resulting underdetermined system of nonlinear equations is solved using `julia` with 10^{-15} tolerance. The Butcher tableau is as follows:

(4.4)

$$\begin{array}{c|ccccc}
0 & 0 & & & & \\
\frac{1}{7} & 0.1428571428571428 & 0 & & & \\
\frac{2}{7} & 0.0107112392440216 & 0.2750030464702641 & 0 & & \\
\frac{3}{7} & 0.4812641640977338 & -0.9634955610240432 & 0.9108028254977381 & 0 & \\
\frac{4}{7} & 0.3718168921589701 & -0.5615016072648120 & 0.5590150320681445 & 0.2020982544662687 & \\
\frac{5}{7} & 0.2210152091353413 & 0.3526985345185138 & -0.8940286416537777 & 0.8097519357352928 & \\
\frac{6}{7} & 0.2038005573304709 & -0.4759394836772968 & 1.0938423462712870 & -0.2853403360392873 & \\
\hline
1 & 0.0979996468518433 & -0.0044680013474903 & 0.3592897484042552 & 0.0225280828210172 &
\end{array}$$

$$\vdots$$

$$\begin{array}{c|ccc}
\frac{5}{7} & 0.2248486765503442 & 0 & \\
\frac{6}{7} & -0.1249739792585496 & 0.4457537525162331 & 0 \\
\hline
1 & 0.2680292384753375 & -0.1064595934043553 & 0.3630808781993925
\end{array}$$

A contour plot of the amplification function and a line plot of this function along the imaginary axis is reported in Figure 4.1. We have $|\Re(is)| \leq 1$ for $s \in [-1.562, 1.562]$.

We are also going to test standard SSPRK techniques of second-, third-, and fourth-order. Using the present notation, we refer to these methods as RK$^\ddagger(2,2;\frac{1}{2})$, RK$^\ddagger(3,3;\frac{1}{3})$, and RK$^\ddagger(5,4;\frac{1}{2})$ (the efficiency ratio of the fourth-order method is actually 0.51, but we write $\frac{1}{2}$ to save some horizontal space in the tables). The Butcher tableau of the second-order and third-order methods is as follows:

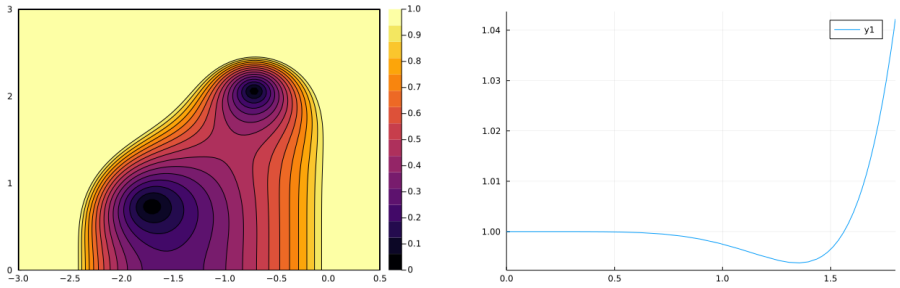

FIG. 4.1. *Amplification function* $\mathrm{RK}(7,5;1)$. *Left: contour plot (truncated at the value* 1*) in the domain* $\{z \in \mathbb{C} \mid \Re(z) \in [-3, 0.5], \Im(z) \in [0, 3]\}$. *Right: line plot along the imaginary axis.*

(4.5)

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

$\mathrm{RK}^{\ddagger}(2,2;\frac{1}{2})$

$$\begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}$$

$\mathrm{RK}^{\ddagger}(3,3;\frac{1}{3})$

$\mathrm{RK}^{\ddagger}(2,2;\frac{1}{2})$ is known as Heun's second-order method, and $\mathrm{RK}^{\ddagger}(3,3;\frac{1}{3})$ is sometimes called Fehlberg's method [9]. The Butcher tableau of $\mathrm{RK}^{\ddagger}(5,4;\frac{1}{2})$ can be found in Kraaijevanger [26, p. 522].

Finally, we observe that for all the above ERK methods, we have $l'(l) = l - 1$ for all $l \in \{2{:}s+1\}$, except for $\mathrm{RK}^{\ddagger}(3,3;\frac{1}{3})$ and $\mathrm{RK}^{\ddagger}(5,4;\frac{1}{2})$ for which the values for $l'(l)$ are $(1,1,2)$ for $l \in \{2{:}4\}$ and $(1,2,2,3,5)$ for $l \in \{2{:}6\}$, respectively.

4.2. Methodology for the numerical tests. For all the tests reported here, the limiting is performed with two iterations of the convex limiting technique introduced in [17], [18]. Since we illustrate the method only for scalar conservation equations (for brevity and simplicity), the relaxed bounds are not allowed to exceed the minimum and the maximum values of the initial data. Hence, the global minimum principle and the global maximum principle are always strictly enforced up to machine accuracy. Relaxed local bounds are also enforced; see Remark 5.2.

In all the tests, the time step is computed by using the expression

$$\tau := \mathrm{CFL} \times s \times \tau^{*}, \tag{4.6}$$

where $\mathrm{CFL} > 0$ is a fixed parameter, s the number of stages of the ERK method, and τ^{*} the maximum Euler time step. Notice that τ^{*} a priori depends on n, but we omit this dependence to simplify the notation; τ^{*} is recomputed at the beginning of each time step t^{n}. Given some final time T and some mesh $\mathcal{T}_h$ determining the value of τ^{*}, the choice (4.6) for τ guarantees that all the ERK methods described in section 4.1 perform exactly the same number of flux evaluations to reach the final time T, i.e., the complexity of all the algorithms is the same. Since

$$\Delta c^{\max} \tau = \mathrm{CFL} \times \Delta c^{\max} \times s \times \tau^{*},$$

and all the methods are IDP if $\Delta c^{\max} \tau \le \tau^{*}$, we conclude that all the methods are IDP provided $\mathrm{CFL} \in (0, \frac{1}{\Delta c^{\max} s}]$ (recall from Lemma 2.3 that $\frac{1}{\Delta c^{\max} s} \le 1$).

TABLE 4.1
Linear transport, 1D finite differences, second-order methods, error in the L^∞-norm.

	CFL = 0.2				CFL = 0.25			
I	RK$(2,2;1)$	Rate	RK$^\ddagger(2,2;\frac{1}{2})$	Rate	RK$(2,2;1)$	Rate	RK$^\ddagger(2,2;\frac{1}{2})$	Rate
50	4.72E-02	–	1.23E-01	–	4.91E-02	–	1.30E-01	–
100	2.81E-03	4.07	1.50E-02	3.03	4.51E-03	3.44	4.32E-02	1.60
200	1.16E-03	1.28	1.24E-03	3.60	2.01E-03	1.17	2.14E-03	4.34
400	3.38E-04	1.78	3.47E-04	1.84	5.41E-04	1.89	5.67E-04	1.91
800	8.79E-05	1.94	9.28E-05	1.90	1.38E-04	1.97	1.48E-04	1.94
1600	2.22E-05	1.98	2.33E-05	1.99	3.47E-05	1.99	3.78E-05	1.97
3200	5.58E-06	1.99	5.92E-06	1.98	8.73E-06	1.99	5.36E-05	-.50

TABLE 4.2
Linear transport, 1D finite differences, third-order methods, error in the L^∞-norm.

	CFL = 0.05						CFL = 0.25					
I	RK(3,3;1)	Rate	RK$^\ddagger(3,3;\frac{1}{3})$	Rate	RK(4,3;1)	Rate	RK(3,3;1)	Rate	RK$^\ddagger(3,3;\frac{1}{3})$	Rate	RK(4,3;1)	Rate
50	5.15E-02	–	4.76E-02	–	5.15E-02	–	5.48E-02	–	1.55E-01	–	6.08E-02	–
100	5.41E-03	3.25	5.41E-03	3.14	5.41E-03	3.25	5.15E-03	3.41	6.12E-02	1.35	6.15E-03	3.31
200	3.79E-04	3.83	3.79E-04	3.83	3.79E-04	3.83	3.92E-04	3.72	1.07E-03	5.84	3.83E-04	4.01
400	2.27E-05	4.06	2.27E-05	4.06	2.27E-05	4.06	2.89E-05	3.76	2.18E-04	2.29	2.30E-05	4.06
800	1.58E-06	3.85	1.58E-06	3.85	1.58E-06	3.85	3.20E-06	3.18	6.41E-05	1.77	1.59E-06	3.85
1600	9.12E-08	4.12	1.22E-07	3.69	8.13E-08	4.28	8.23E-07	1.96	1.83E-05	1.81	8.25E-08	4.27
3200	1.52E-08	2.58	6.84E-08	0.84	5.31E-09	3.94	2.40E-07	1.78	5.39E-06	1.76	5.39E-09	3.94

4.3. Convergence tests with smooth solutions. We illustrate the method by solving the following linear transport equation in one and two space dimensions:

$$\partial_t u + \nabla\cdot(\boldsymbol{\beta} u) = 0. \tag{4.7}$$

All the errors reported in this section are evaluated at the final time and are relative; for instance, for the L^∞-norm, we report $\|u(T) - u_h(T)\|_{L^\infty(D)}/\|u(T)\|_{L^\infty(D)}$.

4.3.1. Fourth-order finite differences in 1D. We consider the problem (4.7) in the one-dimensional (1D) domain $D := (0,1)$ with $\boldsymbol{\beta} := 1$ and the initial data $u_0(x) := (4\frac{(x-x_0)(x_1-x)}{(x_1-x_0)^2})^6$ if $x \in (x_0, x_1)$, and $u_0(x) := 0$ otherwise, with $x_0 := 0.1$ and $x_1 := 0.4$. We enforce periodic boundary conditions. The tests are realized up to $T := 1$ with fourth-order finite differences on uniform meshes (see section 5.2). The space approximation is formally fourth-order accurate. In this section, we only report L^∞-errors, which are more challenging than L^1-errors; recall in particular that the maximum principle is strictly enforced up to machine precision in all cases. We have verified that all the methods achieve the expected convergence order in the L^1-norm with CFL of the order of 0.5 (results not shown for brevity).

We show in Table 4.1 the relative error in the L^∞-norm for the second-order methods RK$(2,2;1)$ and RK$^\ddagger(2,2;\frac{1}{2})$. Both methods perform as expected at CFL $=$ 0.2. Instead, there is a slight difference of behavior at CFL $= 0.25$: RK$^\ddagger(2,2;\frac{1}{2})$ is no longer in the asymptotic convergence regime on the finest mesh (orange column), whereas RK$(2,2;1)$ still behaves properly (green column).

We show in Table 4.2 the relative error in the L^∞-norm for the third-order methods RK$(3,3;1)$, RK$^\ddagger(3,3;\frac{1}{3})$, and RK$(4,3;1)$. All the methods deliver (at least) third-order accuracy for CFL $= 0.05$. The performance of RK$^\ddagger(3,3;\frac{1}{3})$ somewhat degrades

TABLE 4.3
Linear transport, 1D finite differences, fourth-order methods, error in the L^∞-norm.

	CFL = 0.05						CFL = 0.2					
I	RK(4,4;$\frac{1}{2}$)	Rate	RK‡(5,4;$\frac{1}{2}$)	Rate	RK(5,4;1)	Rate	RK(4,4;$\frac{1}{2}$)	Rate	RK‡(5,4;$\frac{1}{2}$)	Rate	RK(5,4;1)	Rate
50	4.32E-02	–	5.37E-02	–	5.95E-02	–	1.26E-01	–	5.63E-02	–	5.55E-02	–
100	5.41E-03	3.00	5.09E-03	3.40	5.09E-03	3.54	1.65E-02	2.93	7.82E-03	2.85	5.72E-03	3.28
200	3.79E-04	3.84	3.04E-04	4.07	3.04E-04	4.07	4.10E-04	5.33	3.80E-04	4.36	3.82E-04	3.90
400	2.27E-05	4.06	1.91E-05	3.99	1.91E-05	3.99	5.02E-05	3.03	2.27E-05	4.06	2.29E-05	4.06
800	1.58E-06	3.85	1.19E-06	4.00	1.19E-06	4.00	1.10E-05	2.19	1.79E-06	3.67	1.60E-06	3.84
1600	8.13E-08	4.28	7.45E-08	4.00	7.45E-08	4.00	2.70E-06	2.03	3.66E-07	2.29	8.26E-08	4.28
3200	5.36E-09	3.92	4.65E-09	4.00	4.65E-09	4.00	7.69E-07	1.81	9.29E-08	1.98	5.38E-09	3.94

TABLE 4.4
Linear transport, 1D finite differences, fifth-order methods, error in the L^∞-norm.

	CFL = 0.02				CFL = 0.025			
I	RK(6,5;$\frac{1}{3}$)	Rate	RK(7,5;1)	Rate	RK(6,5;$\frac{2}{3}$)	Rate	RK(7,5;1)	Rate
50	5.19E-02	–	5.19E-02	–	5.19E-02	–	5.19E-02	–
100	5.41E-03	3.26	5.41E-03	3.26	5.41E-03	3.26	5.41E-03	3.26
200	3.79E-04	3.83	3.79E-04	3.83	3.79E-04	3.84	3.79E-04	3.83
400	2.27E-05	4.06	2.27E-05	4.06	2.27E-05	4.06	2.27E-05	4.06
800	1.58E-06	3.85	1.58E-06	3.85	1.58E-06	3.85	1.58E-06	3.85
1600	8.48E-08	4.22	8.13E-08	4.28	8.71E-08	4.18	8.13E-08	4.28
3200	7.10E-09	3.58	5.92E-09	3.78	1.16E-08	2.91	5.56E-09	3.87

as the mesh size is refined though. Moreover, the performance significantly degrades at CFL = 0.25 (orange column). The performance of RK$(3,3;1)$ also degrades, but far less. The optimal RK$(4,3;1)$ method behaves extremely well at CFL = 0.25 (green column). It delivers fourth-order accuracy. This is coherent since, although the method is only third-order accurate on nonlinear problems, it satisfies all the necessary and sufficient conditions to be fourth-order accurate on linear problems.

We show in Table 4.3 the relative error in the L^∞-norm for the fourth-order methods RK$(4,4;\frac{1}{2})$, RK$^\ddagger(5,4;\frac{1}{2})$, and RK$(5,4;1)$. The three methods behave optimally at CFL = 0.05, whereas this is the case only for RK$(5,4;1)$ at CFL = 0.2 at least on the finer meshes. Additional tests (not shown) indicate that RK$(4,4;\frac{1}{2})$ still behaves optimally at CFL = 0.125, which is no longer the case for RK$^\ddagger(5,4;\frac{1}{2})$. This test shows that the new method RK$(5,4,1)$ outperforms the SSP method RK$^\ddagger(5,4;\frac{1}{2})$.

We show in Table 4.4 the relative error in the L^∞-norm for the fifth-order methods RK$(6,5;\frac{2}{3})$ and RK$(7,5;1)$. We observe that the two methods reach their asymptotic convergence range for CFL = 0.02 (recall that the space discretization is only fourth-order accurate). At the slightly larger value CFL = 0.025, RK$(7,5;1)$ performs slightly better than RK$(6,5;\frac{2}{3})$ on the finest mesh.

4.3.2. Second-order finite elements. We now solve the problem (4.7) in the two-dimensional (2D) domain $D := (0,1)^2$ with $\boldsymbol{\beta} := (0.9,1)^\mathsf{T}$ and the initial data $u_0(\boldsymbol{x}) := (4\frac{(x-x_0)(x_1-x)}{(x_1-x_0)^2})^4 \times (4\frac{(y-y_0)(y_1-y)}{(y_1-y_0)^2})^4$ if $x \in (x_0,x_1)$ and $y \in (y_0,y_1)$, and $u_0(x) := 0$ otherwise, with $x_0 := y_0 := 0.1$ and $x_1 := y_1 := 0.4$. The simulations are performed up to $T := 0.5$. We use continuous $\mathbb{P}_1$ finite elements on uniform meshes composed of triangles. (Notice that the advection field is on purpose chosen not to

TABLE 4.5

Linear transport. 2D $\mathbb{P}_1$ finite elements on uniform meshes. Relative error in the L^1-norm at $T = 0.5$, using CFL $= 0.4$ *for the second-order methods,* CFL $= 0.7$ *for the third-order methods,* CFL $= 0.5$ *for the fourth-order methods, and* CFL $= 0.4$ *for the fifth-order methods.*

I	RK(2,1;1)	Rate	RK‡(2,2;$\frac{1}{2}$)	Rate
50	2.96E-02	–	3.91E-02	–
100	7.36E-03	2.01	7.47E-03	2.39
200	1.94E-03	1.93	1.94E-03	1.95
400	4.89E-04	1.99	4.89E-04	1.99
800	1.23E-04	1.99	1.26E-04	1.95

I	RK(3,3;1)	Rate	RK‡(3,3;$\frac{1}{3}$)	Rate	RK(4,3;1)	Rate
50	2.80E-02	–	6.48E-02	–	2.40E-02	–
100	3.31E-03	3.08	6.81E-03	3.25	1.48E-03	4.02
200	4.11E-04	3.01	4.23E-04	4.01	8.48E-05	4.13
400	5.15E-05	3.00	5.33E-05	2.99	5.37E-06	3.98
800	6.42E-06	3.00	6.63E-06	3.01	3.57E-07	3.91

I	RK(4,4;$\frac{1}{2}$)	Rate	RK(4,4;$\frac{3}{4}$)	Rate	RK‡(5,4;$\frac{1}{2}$)	Rate	RK(5,4;1)	Rate	RK(6,4;1)	Rate
50	3.79E-02	–	6.25E-02	–	2.32E-02	–	2.20E-02	–	3.32E-02	–
100	1.68E-03	4.49	5.85E-03	3.42	1.30E-03	4.16	1.27E-03	4.12	1.57E-03	4.40
200	6.45E-05	4.71	8.28E-05	6.14	6.43E-05	4.33	7.49E-05	4.08	5.05E-05	4.95
400	3.93E-06	4.04	7.21E-06	3.52	4.56E-06	3.82	4.92E-06	3.93	3.33E-06	3.92
800	2.82E-07	3.80	6.73E-07	3.42	3.59E-07	3.67	3.53E-07	3.80	2.33E-07	3.84

I	RK(6,5;$\frac{2}{3}$)	Rate	RK(7,5;1)	Rate
50	1.87E-02	–	1.66E-02	–
100	1.01E-03	4.21	9.26E-04	4.17
200	5.07E-05	4.31	4.95E-05	4.23
400	3.27E-06	3.95	3.01E-06	4.04
800	2.37E-07	3.79	1.92E-07	3.97

be tangential to any mesh edge to avoid any extraneous superconvergence effects.) The Lagrange shape functions are invariant by central symmetry in the support of every nodal basis function; this guarantees that the method is superconvergent up to third-order at the mesh nodes (see Guermond and Pasquetti [12, Prop. 2.1] and Thompson [38, Prop. 4.4]).

We show in Table 4.5 the relative errors in the L^1-norm for all the RK methods considered herein. All the methods deliver optimal convergence rates for CFL numbers in the range $[0.4, 0.7]$.

We now consider the more challenging error measure based on the L^∞-norm. The relative errors are shown in Table 4.6 with CFL $= 0.2$ for all the methods. Almost all the methods deliver their respective theoretical rate of convergence at this CFL number. We observe again (as expected) that the third-order method RK$(4, 3; 1)$ delivers fourth-order. We also observe that RK$^\ddagger(3, 3; \frac{1}{3})$ behaves poorly at this CFL number compared to the two other third-order IDP-ERK methods. The fourth-order (respectively, fifth-order) method that behaves the best is RK$(5, 4; 1)$ (respectively, RK$(7, 5; 1)$). We observe that the performance of all the fourth- and fifth-order methods slightly deteriorates on the finer meshes at this CFL number. Overall, RK$(4, 3; 1)$ is the method that performs the best in the L^∞-norm at this CFL number.

TABLE 4.6
Linear transport. 2D $\mathbb{P}_1$ finite elements on uniform meshes. $T = 0.5$ at CFL $= 0.2$. Relative error in the L^∞-norm for all the methods.

I	RK(2,1;1)	Rate	RK‡(2,2;$\frac{1}{2}$)	Rate
50	1.81E-02	–	2.20E-02	–
100	1.76E-03	3.37	1.84E-03	3.58
200	3.20E-04	2.46	3.20E-04	2.52
400	7.90E-05	2.02	7.90E-05	2.02
800	1.99E-05	1.99	1.99E-05	1.99

I	RK(3,3;1)	Rate	RK‡(3,3;$\frac{1}{3}$)	Rate	RK(4,3;1)	Rate
50	2.28E-02	–	3.87E-02	–	2.30E-02	–
100	1.13E-03	4.33	2.64E-03	3.87	1.14E-03	4.34
200	4.54E-05	4.64	6.85E-05	5.27	4.81E-05	4.56
400	2.49E-06	4.19	2.01E-05	1.77	2.10E-06	4.52
800	4.09E-07	2.60	5.29E-06	1.93	1.09E-07	4.27

I	RK(4,4;$\frac{1}{2}$)	Rate	RK(4,4;$\frac{3}{4}$)	Rate	RK‡(5,4;$\frac{1}{2}$)	Rate	RK(5,4;1)	Rate	RK(6,4;1)	Rate
50	2.68E-02	–	2.39E-02	–	2.43E-02	–	2.30E-02	–	2.32E-02	–
100	1.55E-03	4.11	1.17E-03	4.35	1.33E-03	4.19	1.14E-03	4.33	1.13E-03	4.36
200	4.98E-05	4.96	5.51E-05	4.41	5.18E-05	4.68	4.80E-05	4.57	4.70E-05	4.59
400	2.61E-06	4.25	1.37E-05	2.01	2.45E-06	4.40	2.30E-06	4.39	2.74E-06	4.10
800	4.37E-07	2.58	3.56E-06	1.94	2.78E-07	3.14	1.77E-07	3.70	7.04E-07	1.96

I	RK(6,5;$\frac{2}{3}$)	Rate	RK(7,5;1)	Rate
50	2.41E-02	–	2.29E-02	–
100	1.14E-03	4.40	1.14E-03	4.34
200	4.65E-05	4.62	4.73E-05	4.59
400	3.37E-06	3.78	2.51E-06	4.24
800	9.76E-07	1.79	5.31E-07	2.24

4.4. Numerical tests with nonsmooth solutions. We now consider test cases with nonsmooth solutions.

4.4.1. Linear transport. Here, we solve a standard linear transport problem with nonsmooth initial data (see Leveque [31], Zalesak [39]): $\partial_t u + \nabla\cdot(\boldsymbol{\beta} u) = 0$ in $D := \{\boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x}\|_{\ell^2} < 1\}$ with $\boldsymbol{\beta}(\boldsymbol{x}) := 2\pi(-x_2, x_1)^\mathsf{T}$. The initial data is

$$u_0(\boldsymbol{x}) := \begin{cases} 1 & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_d\|_{\ell^2} \le r_0 \text{ and } (|x_1| \ge 0.05 \text{ or } x_2 \ge 0.7), \\ 1 - \frac{\|\boldsymbol{x}-\boldsymbol{x}_c\|_{\ell^2}}{r_0} & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_c\|_{\ell^2} \le r_0, \\ g(\|\boldsymbol{x}-\boldsymbol{x}_h\|_{\ell^2}) & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_h\|_{\ell^2} \le r_0, \\ 0 & \text{otherwise,} \end{cases} \tag{4.8}$$

where $r_0 := 0.3$, $g(r) := \frac{1}{4}\big[1 + \cos\big(\pi\frac{r}{r_0}\big)\big]$, $\boldsymbol{x}_d := (0, 0.5)$, $\boldsymbol{x}_c := (0, -0.5)$, $\boldsymbol{x}_h := (-0.5, 0)$. The graph of u_0 consists of three solids: a slotted cylinder of height 1, a smooth hump of height $\frac{1}{2}$, and a cone of height 1.

The simulations are performed up to $T := 1$ using continuous $\mathbb{P}_1$ finite elements on unstructured nonnested Delaunay triangulations. For brevity, we only show the performance of the RK$(2, 2; 1)$ method (i.e., the midpoint rule) to demonstrate that this method, which is often shunned in the literature for "its lack of stability," performs actually very well when used with the IDP technique proposed in this paper. We show

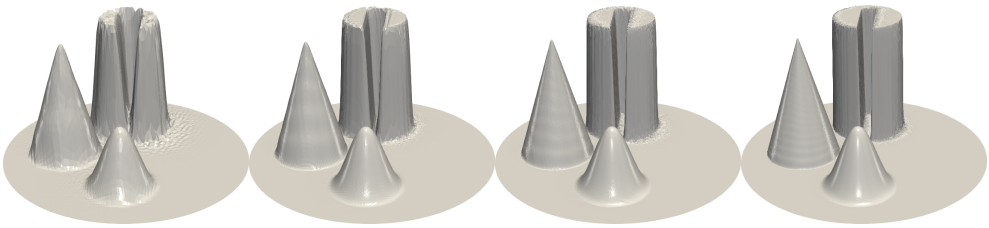

FIG. 4.2. *Three-solids problem at* $T = 1$, *using* RK$(2, 2; 1)$ *at* CFL $= 0.25$. $2D$ $\mathbb{P}_1$ *finite elements on unstructured meshes. From left to right:* $I = 6561$; $I = 24917$; $I = 98648$; $I = 389860$.

TABLE 4.7
Three-solids problem at $T = 1$ *and* CFL $= 0.25$. $2D$ $\mathbb{P}_1$ *finite elements on unstructured meshes. Relative error in the* L^1*-norm for* RK$(2, 2; 1)$ *and* RK$(4, 3; 1)$.

I	RK(2,2;1)	Rate	RK(4,3;1)	rate
1605	2.45E-01	–	2.49E-01	–
6561	1.28E-01	0.93	1.31E-01	0.92
24917	7.34E-02	0.81	7.49E-02	0.84
98648	4.26E-02	0.78	4.44E-02	0.76
389860	2.44E-02	0.81	2.56E-02	0.80

in Figure 4.2 the graph of the solutions computed on four different grids composed of $I = 6561$, $I = 24917$, $I = 98648$, and $I = 389860$ $\mathbb{P}_1$ Lagrange nodes. The computations are done at CFL $= 0.25$. The results produced by RK$(2, 2; 1)$ are visually of the same quality as what is usually reported in the literature.

We show in Table 4.7 the relative error at $T = 1$ measured in the L^1-norm using the two methods RK$(2, 2; 1)$ and RK$(4, 3; 1)$. The convergence rates are similar to those reported in [14, sect. 6.3].

4.4.2. Burgers' equation. In this section, we consider Burgers' equation in the 2D domain $D := (-.25, 1.75)^2$:

$$\partial_t u + \nabla\cdot(\boldsymbol{f}(u)) = 0, \qquad \boldsymbol{f}(u) := \tfrac{1}{2}(u^2, u^2)^\mathsf{T}, \qquad u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) \text{ a.e. } \boldsymbol{x} \in D \tag{4.9}$$

with the initial data

$$u_0(\boldsymbol{x}) := \begin{cases} 1 & \text{if } |x_1 - \tfrac{1}{2}| \le 1 \text{ and } |x_2 - \tfrac{1}{2}| \le 1, \\ -a & \text{otherwise.} \end{cases} \tag{4.10}$$

This problem is considered in [14, sect. 6.1]. The exact solution is given in equations (52)–(53) therein; the invariant domain is $\mathcal{A} := [-a, 1]$. This problem is interesting since it exhibits many sonic points, which makes methods with either too little low-order viscosity or too little high-order viscosity fail (see, e.g., Guermond and Popov [14, Lem. 2.2] for a low-order viscosity counterexample and [14, Lem. 4.6] for a high-order viscosity counterexample). We approximate the solution to this problem with continuous $\mathbb{P}_1$ finite elements on uniform triangular meshes. The computations are done up to $T := 0.65$ with CFL $= 0.25$. The solution obtained on the 801^2 mesh using RK$(4, 3; 1)$ is shown in Figure 4.3.

We test the IDP-ERK methods considered above. We compute the relative error in the L^1-norm on five consecutively refined meshes. The results are reported in Table 4.8 using CFL $= 0.25$. We observe that all the convergence rates are close to 0.9. The rates are similar to those reported in [14], where the time stepping is done

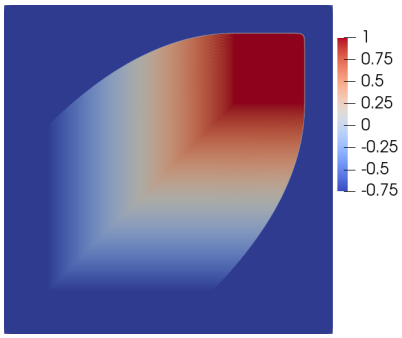

FIG. 4.3. *Solution of* (4.9), $T = 0.65$ *at* CFL $= 0.25$, RK$(4, 3; 1)$, 801^2 *grid points.*

TABLE 4.8
Burgers' equation. $2D\ \mathbb{P}_1$ *finite elements on uniform meshes.* $T = 0.65$ *at* CFL $= 0.25$. *Relative error in the* L^1*-norm for all the methods.*

I	RK(2,1;1)	Rate	RK‡(2,2;$\frac{1}{2}$)	Rate
$(51)^2$	6.61E-02	–	6.70E-02	–
$(101)^2$	3.31E-02	1.00	3.34E-02	1.00
$(201)^2$	2.12E-02	0.65	2.12E-02	0.66
$(401)^2$	1.20E-02	0.82	1.16E-02	0.87
$(801)^2$	6.04E-03	0.99	5.73E-03	1.02

I	RK(3,3;1)	Rate	RK‡(3,3;$\frac{1}{3}$)	Rate	RK(4,3;1)	Rate
$(51)^2$	6.61E-02	–	6.74E-02	–	6.62E-02	–
$(101)^2$	3.31E-02	1.00	3.35E-02	1.01	3.31E-02	1.00
$(201)^2$	2.12E-02	0.65	2.13E-02	0.66	2.12E-02	0.65
$(401)^2$	1.20E-02	0.82	1.15E-02	0.89	1.20E-02	0.82
$(801)^2$	6.04E-03	0.99	5.72E-03	1.01	6.04E-03	0.99

I	RK(4,4;$\frac{1}{2}$)	Rate	RK(4,4;$\frac{3}{4}$)	Rate	RK‡(5,4;$\frac{1}{2}$)	Rate	RK(5,4;1)	Rate	RK(6,4;1)	Rate
$(51)^2$	6.74E-02	–	6.63E-02	–	6.72E-02	–	6.63E-02	–	6.60E-02	–
$(101)^2$	3.35E-02	1.01	3.31E-02	1.00	3.43E-02	0.97	3.32E-02	1.00	3.30E-02	1.00
$(201)^2$	2.13E-02	0.66	2.11E-02	0.65	2.26E-02	0.60	2.12E-02	0.64	2.11E-02	0.64
$(401)^2$	1.17E-02	0.87	1.18E-02	0.84	1.28E-02	0.82	1.20E-02	0.82	1.20E-02	0.82
$(801)^2$	5.75E-03	1.02	5.84E-03	1.02	6.20E-03	1.05	6.06E-03	0.99	6.03E-03	0.99

I	RK(6,5;$\frac{2}{3}$)	Rate	RK(7,5;1)	Rate
$(51)^2$	6.65E-02	–	6.62E-02	–
$(101)^2$	3.32E-02	1.00	3.31E-02	1.00
$(201)^2$	2.11E-02	0.65	2.12E-02	0.65
$(401)^2$	1.18E-02	0.84	1.20E-02	0.82
$(801)^2$	5.79E-03	1.02	6.06E-03	0.99

with RK$^\ddagger(3, 3; \frac{1}{3})$ (the CFL used therein is three times smaller). All the IDP-ERK methods perform as well as the traditional SSPRK methods. This test shows that, as claimed in this paper, the proposed methodology makes every ERK method IDP.

We now test the behavior of the various IDP-ERK methods as the CFL number grows. We use a uniform mesh composed of 401^2 grid points and vary the CFL number in the range $[0.05, 1]$. We show in Figure 4.4 the relative error in the L^1-norm as a

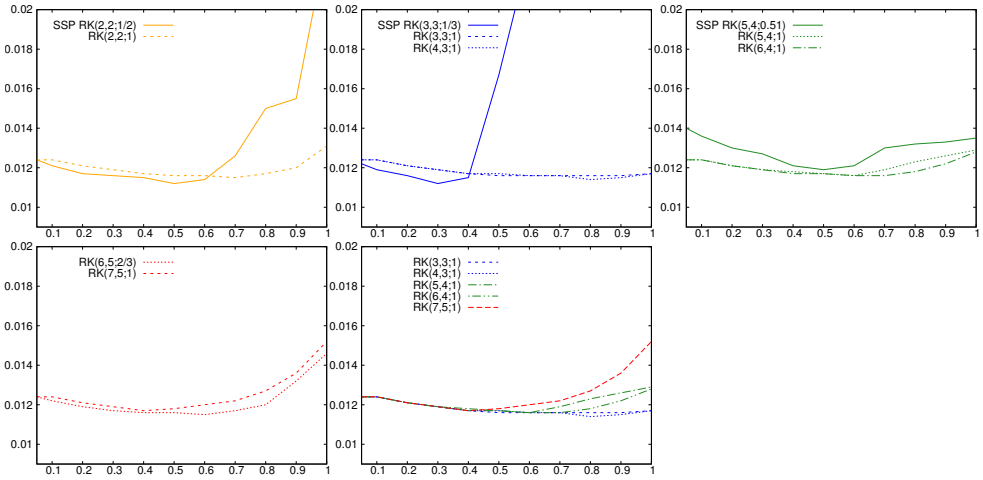

FIG. 4.4. *Burgers' equation.* $2D$ $\mathbb{P}_1$ *finite elements on uniform* 401^2 *grid. Error in the* L^1*-norm versus CFL. Top row: second-, third-, and fourth-order methods. Bottom row: fifth-order methods and all the optimally efficient methods.*

function of the CFL number. On the top row, we show the results for the second-, third-, and fourth-order methods (left, center, and right panels, respectively). On the bottom row, we show the results for the fifth-order methods in the left panel, and we collect all the results for the optimally efficient methods in the right panel. We observe that $\mathrm{RK}^{\ddagger}(2,2;\frac{1}{2})$ (top row, left panel, solid line) starts to lose accuracy when $\mathrm{CFL} \gtrsim \frac{1}{2}$ and eventually loses stability when $\mathrm{CFL} \gtrsim 0.9$. Notice that $\mathrm{RK}(2,2;1)$ behaves properly over the entire CFL range. Similarly, $\mathrm{RK}^{\ddagger}(3,3;\frac{1}{3})$ (top row, central panel, solid line) starts losing accuracy when $\mathrm{CFL} \gtrsim \frac{1}{3}$ and loses stability when $\mathrm{CFL} \gtrsim 0.6$, whereas $\mathrm{RK}(3,3;1)$ and $\mathrm{RK}(4,3;1)$ behave properly over the entire CFL range. All the fourth-order method considered ($\mathrm{RK}^{\ddagger}(5,4;\frac{1}{2})$, $\mathrm{RK}(5,4,1)$, $\mathrm{RK}(6,4;1)$) behave properly. Notice though that the two methods with optimal efficiency are slightly more accurate than $\mathrm{RK}^{\ddagger}(5,4;\frac{1}{2})$ over the whole range of CFL numbers. This test shows that robustness with respect to the CFL number can be achieved for IDP-ERK methods (among other things) by maximizing the efficiency. This test also demonstrates that $\mathrm{RK}^{\ddagger}(3,3;\frac{1}{3})$ is particularly inefficient.

4.5. Conclusions from the numerical tests. The main conclusions we draw from the numerical experiments are as follows. All the IDP-ERK methods proposed herein perform at least as well as, and often better than, the SSPRK methods of the same order. For instance, the midpoint rule outperforms the popular $\mathrm{RK}^{\ddagger}(2,2;\frac{1}{2})$ method, and $\mathrm{RK}(4,3;1)$ (amply) outperforms the popular $\mathrm{RK}^{\ddagger}(3,3;\frac{1}{3})$ method. All the fourth-order methods provide comparable results. Finally, the present methodology allows one to use fifth-order IDP-ERK methods, which is not possible within the SSP paradigm.

5. Examples of space approximation methods. This section briefly outlines how the low-order and the high-order fluxes can be constructed for the two space approximation methods considered in the previous section, i.e., continuous finite elements and finite difference methods. We also briefly discuss limiting. This section is intended for completeness of the presentation, and the reader is referred to the various pointers to the literature given herein for further insight.

5.1. Continuous finite elements. The use of continuous finite elements to construct invariant-domain approximations of nonlinear conservation equations is well documented in the literature (see, e.g., Abgrall [1], Guermond et al. [15], [13], [18, sect. 4.2], Ern and Guermond [8, Chap. 81], Kuzmin and Turek [27]). Given a shape-regular sequence of unstructured matching meshes $(\mathcal{T}_h)_{h\in\mathcal{H}}$, where each mesh covers D exactly, and a reference finite element $(\widehat{K},\widehat{P},\widehat{\Sigma})$, we define the following scalar-valued and vector-valued continuous finite element spaces:

$$P(\mathcal{T}_h) := \{v\in\mathcal{C}^0(D;\mathbb{R}) \mid v_{|K}\circ\boldsymbol{T}_K\in\widehat{P}\ \forall K\in\mathcal{T}_h\},\qquad \boldsymbol{P}(\mathcal{T}_h) := [P(\mathcal{T}_h)]^m. \tag{5.1}$$

Here, for all $K\in\mathcal{T}_h$, $\boldsymbol{T}_K:\widehat{K}\to K$ is the geometric mapping. Let $\{\varphi_i\}_{i\in\mathcal{V}}$ be the global shape functions of $P(\mathcal{T}_h)$. For every $i\in\mathcal{V}$, the stencil $\mathcal{I}(i)$ at i is the collection of the indices $j\in\mathcal{V}$ such that $|\mathrm{supp}(\varphi_i)\cap\mathrm{supp}(\varphi_j)|>0$. The coefficients of the consistent and of the lumped mass matrices are, respectively, defined to be

$$m_{ij}^{\mathrm{H}} := \int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}x,\qquad m_{ij}^{\mathrm{L}} := \delta_{ij}\int_D\varphi_i(\boldsymbol{x})\,\mathrm{d}x\qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i), \tag{5.2}$$

and we set $m_i := m_{ii}^{\mathrm{L}}$. To construct the fluxes, we introduce the vectors

$$\boldsymbol{c}_{ij} := \int_D \varphi_i(\boldsymbol{x})\nabla\varphi_j(\boldsymbol{x})\,\mathrm{d}x\qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i). \tag{5.3}$$

If at least one of the shape functions φ_i and φ_j vanishes at the boundary ∂D, then

$$\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}\qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i). \tag{5.4}$$

We also introduce low-order and high-order graph viscosity matrices $\{d_{ij}^{\mathrm{L},n}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$, $\{d_{ij}^{\mathrm{H},n}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$ with the key assumption that

$$d_{ij}^{\mathrm{L},n} = d_{ji}^{\mathrm{L},n}\ge 0\quad\text{and}\quad d_{ij}^{\mathrm{H},n} = d_{ji}^{\mathrm{H},n}\ge 0\qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i). \tag{5.5}$$

We refer to [13] and [8, Chap. 81] for the construction of $\{d_{ij}^{\mathrm{L},n}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$ (essentially $d_{ij}^{\mathrm{L},n}$ scales as $\|\boldsymbol{c}_{ij}\|_{\ell^2}$ multiplied by a maximum wave speed associated with a suitable local Riemann problem), and to [27], [15], [18, sect. 6], and [8, Chaps. 82–83] for examples of constructions of $\{d_{ij}^{\mathrm{H},n}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$. With these definitions, we set

$$\boldsymbol{F}_{ij}^{\mathrm{L}}(\mathbf{V}) := -(\mathbb{f}(\mathbf{V}_j)+\mathbb{f}(\mathbf{V}_i))\boldsymbol{c}_{ij} + d_{ij}^{\mathrm{L},n}(\mathbf{V}_j-\mathbf{V}_i)\qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i), \tag{5.6}$$

$$\boldsymbol{F}_{ij}^{\mathrm{H}}(\mathbf{V}) := -(\mathbb{f}(\mathbf{V}_j)+\mathbb{f}(\mathbf{V}_i))\boldsymbol{c}_{ij} + d_{ij}^{\mathrm{H},n}(\mathbf{V}_j-\mathbf{V}_i)\qquad \forall i\in\mathcal{V},\ \forall j\in\mathcal{I}(i). \tag{5.7}$$

(Since $\mathbb{f}(\mathbf{V})$ is $\mathbb{R}^{m\times d}$-valued and $\boldsymbol{c}_{ij}$ is $\mathbb{R}^d$-valued, the matrix-vector product $\mathbb{f}(\mathbf{V})\boldsymbol{c}_{ij}$ is $\mathbb{R}^m$-valued.) Notice that the key skew-symmetry property (3.2) is satisfied (up to details regarding the boundary conditions, which are beyond the scope of the paper). Finally, setting $\mathcal{I}^*(i) := \mathcal{I}(i)\backslash\{i\}$, it can be shown that the structural assumption (2.7) holds true for

$$\tau\le\tau^* := \frac{1}{2}\min_{i\in\mathcal{V}}\frac{m_i}{\sum_{j\in\mathcal{I}^*(i)}d_{ij}^{\mathrm{L},n}}. \tag{5.8}$$

Remark 5.1 (discontinuous Galerkin). The above formalism can be readily extended to discontinuous Galerkin methods upon defining the coefficients $\boldsymbol{c}_{ij}$ using centered numerical fluxes. We refer the reader to Guermond et al. [17], Pazner [34].

5.2. Finite differences. We finish with a short example involving finite differences in one space dimension with periodic boundary conditions. This setting is used in section 4 to illustrate the method. Consider the domain $D := (0, L)$. For every $N \geq 2$, we construct the uniform mesh composed of the nodes $x_i := (i-1)h$, $i \in \{1{:}N\}$, with $h := L/(N-1)$.

To account for periodic boundary conditions, we set $\mathcal{V} := \{1{:}N-1\}$. For every $i \in \mathcal{V}$, we set $\mathcal{I}(i) := \{i-1, i, i+1\}$ with the convention that $\mathcal{I}(1) := \{N-1, 1, 2\}$ and $\mathcal{I}(N-1) := \{N-2, N-1, 1\}$ to enforce periodicity. We set the coefficients of the lumped mass matrix, $\mathbb{M}^{\mathrm{L}}$, to be $m_i := h$, and we set $\mathbb{M}^{\mathrm{H}} := \mathbb{M}^{\mathrm{L}}$. For every $i \in \mathcal{V}$, we also define the coefficients

$$\boldsymbol{c}_{i,i-1} := -\frac{1}{2}, \qquad \boldsymbol{c}_{i,i} := 0, \qquad \boldsymbol{c}_{i,i+1} := \frac{1}{2}. \tag{5.9}$$

Here, we abuse the notation by identifying the $\mathbb{R}^1$-valued vectors $\boldsymbol{c}_{ij}$ with scalars. We define the low-order flux such that

$$\mathbf{F}^{\mathrm{L}}_{ij}(\mathbf{V}) := -(\mathbb{f}(\mathbf{V}_j) + \mathbb{f}(\mathbf{V}_i))\boldsymbol{c}_{ij} + d^{\mathrm{L}}_{ij}(\mathbf{V}_j - \mathbf{V}_i) \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{5.10}$$

The definition of the low-order graph viscosity matrices $\{d^{\mathrm{L}}_{ij}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$ is similar to what can be done for continuous finite elements. Here again, one can establish that the structural assumption (2.7) holds true under the condition (5.8).

For the high-order flux, we use the five-point finite difference formula $h\partial_x f(x_i) \approx \frac{1}{12}(f(x_{i-2}) - 8f(x_{i-1}) + 8f(x_{i+1}) - f(x_{i+2}))$ (with the additional convention that -1 and $N+1$ are replaced by $N-2$ and 2, respectively, to enforce periodicity). As announced in Remark 3.1, we introduce two-point fluxes (as in Harten [21, eq. (1.4b)], Harten, Lax, and van Leer [23, eq. (1.10)], Osher and Chakravarthy [33, eq. 2.3]) and transform this formula into an expression that only involves the three-point stencil by setting $\mathbf{F}^{\mathrm{H}}_{i,i}(\mathbf{V}) := \mathbf{0}$ and

$$\begin{aligned}\mathbf{F}^{\mathrm{H}}_{i,i-1}(\mathbf{V}) :=& -\frac{1}{12}(\mathbb{f}(\mathbf{V}_{i-2}) - \mathbb{f}(\mathbf{V}_{i-1}) - \mathbb{f}(\mathbf{V}_i) + \mathbb{f}(\mathbf{V}_{i+1}))\\ &+ \frac{6}{12}(\mathbb{f}(\mathbf{V}_{i-1}) + \mathbb{f}(\mathbf{V}_i)) + d^{\mathrm{H}}_{i,i-1}(\mathbf{V}_{i-1} - \mathbf{V}_i),\end{aligned} \tag{5.11a}$$

$$\begin{aligned}\mathbf{F}^{\mathrm{H}}_{i,i+1}(\mathbf{V}) :=& \frac{1}{12}(\mathbb{f}(\mathbf{V}_{i-1}) - \mathbb{f}(\mathbf{V}_i) - \mathbb{f}(\mathbf{V}_{i+1}) + \mathbb{f}(\mathbf{V}_{i+2}))\\ &- \frac{6}{12}(\mathbb{f}(\mathbf{V}_i) + \mathbb{f}(\mathbf{V}_{i+1})) + d^{\mathrm{H}}_{i,i+1}(\mathbf{V}_{i+1} - \mathbf{V}_i).\end{aligned} \tag{5.11b}$$

Notice that when $d^{\mathrm{H}}_{i,i-1} = 0$ and $d^{\mathrm{H}}_{i,i+1} = 0$, we obtain

$$\mathbf{F}^{\mathrm{H}}_i(\mathbf{V}) := \sum_{j\in\mathcal{I}(i)} \mathbf{F}^{\mathrm{H}}_{i,j}(\mathbf{V}) = -\frac{1}{12}(\mathbb{f}(\mathbf{V}_{i-2}) - 8\mathbb{f}(\mathbf{V}_{i-1}) + 8\mathbb{f}(\mathbf{V}_{i+1}) - \mathbb{f}(\mathbf{V}_{i+2})), \tag{5.12}$$

which is the three-point stencil representation of the five-point finite difference formula approximating $\partial_x \mathbb{f}(\mathbf{V})$ at x_i mentioned above. Notice also that

$$\mathbf{F}^{\mathrm{H}}_{ij}(\mathbf{V}) = -\mathbf{F}^{\mathrm{H}}_{ji}(\mathbf{V}) \quad \forall i \in \mathcal{I},\ \forall j \in \mathcal{I}(i), \tag{5.13}$$

which is the key skew-symmetry property that guarantees conservation.

5.3. Limiting. There are many ways to perform the limiting operation mentioned in (2.6), (2.17), Definition 3.2, and (3.15). The so-called flux transport correction technique of Boris and Book [3] and Zalesak [39] is probably the most well-known limiting technique for scalar conservation equations. The reader is also referred to Kuzmin and Turek [27] and the book by Kuzmin, Löhner, and Turek [28] for other extensions on this method. But, as observed in [17], [18], the flux transport correction technique is not appropriate when the limiting constraints on the states are not affine, which is almost systematically the case for systems of nonlinear conservation equations. To be more precise, in all the applications we have in mind, the invariant domain $\mathcal{A}$ introduced in the structural assumption (2.7) is of the following form:

$$\mathcal{A} = \bigcap_{l\in\mathcal{L}} \{\mathbf{V}\in\mathbb{R}^m \mid \Psi_l(\mathbf{V})\geq 0\}, \tag{5.14}$$

where $\mathcal{L}\subset\mathbb{N}$ is a finite index set and the functions $\{\Psi_l\}_{l\in\mathcal{L}}$ are quasiconcave. The flux transport correction technique can be applied only for those functions Ψ_l that are affine. In the general case, one must resort to other techniques like the convex limiting method introduced in [17], [18]. We refer the reader to these two references for the details on how convex limiting can be used in (3.15).

Remark 5.2 (local bounds and relaxation). The generic property (2.7) can often be localized. More precisely, given $\mathbf{V}\in\mathcal{A}^I$, one can often construct, for all $i\in\mathcal{V}$, a subset $\mathcal{A}_i\subset\mathcal{A}$ such that $\mathbf{V}_i+\tau m_i^{-1}\mathbf{F}_i^{\mathrm{L}}(\mathbf{V})\in\mathcal{A}_i$ for all $\tau\leq\tau^*$. For instance, for scalar conservation equations, the global invariant domain is $\mathcal{A}:=[\mathsf{V}^{\min},\mathsf{V}^{\max}]$ with $\mathsf{V}^{\min}:=\min_{i\in\mathcal{V}}\mathbf{V}_i$ and $\mathsf{V}^{\max}:=\max_{i\in\mathcal{V}}\mathbf{V}_i$. Then (2.7) simply formalizes that the low-order method satisfies the global maximum/minimum principle. But, very often one can show that, setting $\mathcal{A}_i:=[\mathsf{V}_i^{\min},\mathsf{V}_i^{\max}]\subset\mathcal{A}$ with $\mathsf{V}_i^{\min}:=\min_{j\in\mathcal{I}(i)}\mathbf{V}_j$ and $\mathsf{V}_i^{\max}:=\max_{j\in\mathcal{I}(i)}\mathbf{V}_j$ for all $i\in\mathcal{V}$, one also has $\mathbf{V}_i+\tau m_i^{-1}\mathbf{F}_i^{\mathrm{L}}(\mathbf{V})\in\mathcal{A}_i$ for all $\tau\leq\tau^*$. This additional property allows the limiting to be implemented with local bounds, which gives a tighter control on the approximate solution. It is, however, well-known that strictly enforcing local bounds degrades the converge rate to first-order close to extrema (see, e.g., Khobalatte and Perthame [25, sect. 3.3], Zhang and Shu [41, p. 2753]). A typical way to address this issue in the finite volume literature consists of relaxing the slope reconstructions; see, e.g., Harten [21, eq. (5.7)], Harten and Osher [22]. Similar techniques can be used with discontinuous Galerkin approximations as in Zhang and Shu [40], [41]. In the results reported in section 4, the local bounds are all relaxed as explained in Guermond et al. [17, sect. 4.7] and Guermond, Popov, and Tomas [18, sect. 7.6].

REFERENCES

[1] R. Abgrall, *Residual distribution schemes: Current status and future trends*, Comput. & Fluids, 35 (2006), pp. 641–669.

[2] R. Abgrall, Q. Viville, H. Beaugendre, and C. Dobrzynski, *Construction of a p-adaptive continuous residual distribution scheme*, J. Sci. Comput., 72 (2017), pp. 1232–1268.

[3] J. P. Boris and D. L. Book, *Flux-corrected transport.* [I. SHASTA, *a fluid transport algorithm that works*, J. Comput. Phys., 11 (1973), pp. 170–172], J. Comput. Phys., 135 (1997), pp. 172–186.

[4] J. C. Butcher, *On Runge-Kutta processes of high order*, J. Austral. Math. Soc., 4 (1964), pp. 179–194.

[5] K. N. Chueh, C. C. Conley, and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[6] F. Coquel and P. LeFloch, *Convergence of finite difference schemes for conservation laws in several space dimensions: The corrected antidiffusive flux approach*, Math. Comp., 5 (1991), pp. 169–210.

[7] L. Demkowicz, J. T. Oden, and W. Rachowicz, *A new finite element method for solving compressible Navier-Stokes equations based on an operator splitting method and h-p adaptivity*, Comput. Methods Appl. Mech. Engrg., 84 (1990), pp. 275–326.
[8] A. Ern and J.-L. Guermond, *Finite Elements.* III. *First-Order and Time-Dependent PDEs*, Texts Appl. Math. 74, Springer, Cham, 2021.
[9] E. Fehlberg, *Klassische Runge-Kutta-Formeln vierter und niedrigerer Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme*, Computing (Arch. Elektron. Rechnen), 6 (1970), pp. 61–71.
[10] L. Ferracina and M. N. Spijker, *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*, Math. Comp., 74 (2005), pp. 201–219.
[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] J.-L. Guermond and R. Pasquetti, *A correction technique for the dispersive effects of mass lumping for transport problems*, Comput. Methods Appl. Mech. Engrg., 253 (2013), pp. 186–198.
[13] J.-L. Guermond and B. Popov, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489.
[14] J.-L. Guermond and B. Popov, *Invariant domains and second-order continuous finite element approximation for scalar conservation equations*, SIAM J. Numer. Anal., 55 (2017), pp. 3120–3146.
[15] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang, *A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations*, SIAM J. Numer. Anal., 52 (2014), pp. 2163–2182.
[16] J.-L. Guermond, B. Popov, and Y. Yang, *The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations*, J. Sci. Comput., 70 (2017), pp. 1358–1366.
[17] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas, *Second-order invariant domain preserving approximation of the Euler equations using convex limiting*, SIAM J. Sci. Comput., 40 (2018), pp. A3211–A3239.
[18] J.-L. Guermond, B. Popov, and I. Tomas, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, Comput. Methods Appl. Mech. Engrg., 347 (2019), pp. 143–175.
[19] J.-L. Guermond, M. Maier, B. Popov, and I. Tomas, *Second-order invariant domain preserving approximation of the compressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 375 (2021), 113608.
[20] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations.* I. *Nonstiff Problems*, 2nd ed., Springer Ser. Comput. Math. 8, Springer, Berlin, 1993.
[21] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.
[22] A. Harten and S. Osher, *Uniformly high-order accurate nonoscillatory schemes.* I, SIAM J. Numer. Anal., 24 (1987), pp. 279–309.
[23] A. Harten, P. D. Lax, and B. van Leer, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), pp. 35–61.
[24] I. Higueras, *Representations of Runge-Kutta methods and strong stability preserving methods*, SIAM J. Numer. Anal., 43 (2005), pp. 924–948.
[25] B. Khobalatte and B. Perthame, *Maximum principle on the entropy and second-order kinetic schemes*, Math. Comp., 62 (1994), pp. 119–131.
[26] J. F. B. M. Kraaijevanger, *Contractivity of Runge-Kutta methods*, BIT, 31 (1991), pp. 482–528.
[27] D. Kuzmin and S. Turek, *Flux correction tools for finite elements*, J. Comput. Phy., 175 (2002), pp. 525–558.
[28] D. Kuzmin, R. Löhner, and S. Turek, *Flux-Corrected Transport: Principles, Algorithms, and Applications*, Sci. Comput., Springer, 2012.
[29] D. Kuzmin, M. Quezada de Luna, D. I. Ketcheson, and J. Grüll, *Bound-preserving flux limiting for high-order explicit Runge-Kutta time discretizations of hyperbolic conservation laws*, J. Sci. Comput., 91 (2022), 21.
[30] J. D. Lawson, *An order five Runge-Kutta process with extended region of stability*, SIAM J. Numer. Anal., 3 (1966), pp. 593–597.
[31] R. J. Leveque, *High-resolution conservative algorithms for advection in incompressible flow*, SIAM J. Numer. Anal., 33 (1996), pp. 627–665.
[32] X.-D. Liu and S. Osher, *Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes.* I. SIAM J. Numer. Anal., 33 (1996), pp. 760–779.
[33] S. Osher and S. Chakravarthy, *High resolution schemes and the entropy condition*, SIAM J. Numer. Anal., 21 (1984), pp. 955–984.

[34] W. Pazner, *Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting*, Comput. Methods Appl. Mech. Engrg., 382 (2021) 113876.
[35] S. J. Ruuth and R. J. Spiteri, *Two barriers on strong-stability-preserving time discretization methods*, J. Sci. Comput., 17 (2002), pp. 211–220.
[36] R. Sanders, *A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation laws*, Math. Comp., 51 (1988), pp. 535–558.
[37] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[38] T. Thompson, *A discrete commutator theory for the consistency and phase error analysis of semi-discrete C^0 finite element approximations to the linear transport equation*, J. Comput. Appl. Math., 303 (2016), pp. 229–248.
[39] S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.
[40] 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.
[41] X. Zhang and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: Survey and new developments*, Proc. A, 467 (2011), pp. 2752–2776.