Invariant-domain-preserving high-order time stepping: I. Explicit Runge-Kutta schemes

Alexandre Ern, Jean-Luc Guermond

Invariant-domain-preserving high-order time stepping: I. Explicit Runge–Kutta schemes

Alexandre Ern[†] and Jean-Luc Guermond[‡]

Draft version November 10, 2021

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 to 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.

Keywords. Time integration, Runge–Kutta, invariant domain preserving, strong stability preserving, conservation equations, hyperbolic systems, high-order method.

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

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 $T > 0$. The dependent variable is assumed to take values in $\mathbb{R}^m$, $m \geq 1$. We have $m = 1$ for scalar conservation equations and $m > 1$ for hyperbolic systems. A fundamental property of the model problem is the existence of some invariant domain $\mathcal{A} \subset \mathbb{R}^m$. This means that if the initial data takes values in $\mathcal{A}$ everywhere in D (and in the absence of perturbations due to the boundary conditions), then any admissible

[†]CERMICS, Ecole des Ponts, 77455 Marne-la-Vallee Cedex 2, France and INRIA Paris, 75589 Paris, France
[‡]Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843, USA.

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

A first important example is the class of scalar conservation equations ($m = 1$). In this case, the only admissible solutions are the entropy solutions. For these solutions, the maximum principle holds, and the invariant domains are intervals $[\alpha, \beta] \subset \mathbb{R}$. More precisely, if the initial data 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 $t \in [0,T]$. When $m > 1$, the invariant domain property is a generalized version of the maximum principle. 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 said Cauchy problem, it is important to devise approximation methods that are high-order accurate in space and time, invariant domain preserving (IDP), and conservative. The objective of the paper is to introduce a time-stepping technique that does so.

Applying first some space discretization technique leads to a system of ordinary differential equations (ODEs) in $(\mathbb{R}^m)^I$, where the natural number $I \geq 1$ refers to the number of degrees of freedom (dofs) employed in the space discretization. Applying some time-marching scheme then leads to the sequence of vectors $(\mathbf{U}^n)_{n\in\{0:N\}}$, where $\mathbf{U}^n := (\mathbf{U}_1^n, \dots, \mathbf{U}_I^n)^\mathsf{T} \in (\mathbb{R}^m)^I$ for all $n \in \{0{:}N\} := \{0, \dots, N\}$, and $\mathbf{U}_i^n \in \mathbb{R}^m$ for all $i \in \mathcal{V} := \{1, \dots, I\}$. Since $\mathcal{A}$ is an invariant domain for the continuous system, it is then natural to require that the whole discretization process satisfies the following invariant-domain property:

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

Moreover, global conservation is expressed by the additional requirement that

$$\sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^n = \sum_{i\in\mathcal{V}} m_i \mathbf{U}_i^0, \qquad \forall n \in \{1{:}N\}, \tag{2}$$

where m_i denotes the mass associated with the i-th dof.

Ensuring that (1)-(2) hold true is by no means straightforward. It is possible to satisfy (1)-(2) using the forward Euler scheme under a CFL restriction on the time step if one resorts to a low-order space discretization based on some suitable (graph) viscosity. Achieving (1) with a higher-order space discretization is then made possible by applying some nonlinear, conservative limiting technique. In practice, however, one is interested in developing higher-order time discretization techniques that are also IDP and conservative. One well-known way to achieve the IDP when the invariant domain $\mathcal{A}$ is a convex set (which is the case in practice) is to resort to strong stability preserving (SSP) ERK methods (in short, SSPRK). 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. Since the limited forward Euler step is IDP and $\mathcal{A}^I$ is convex, the SSP update stays in $\mathcal{A}^I$. We refer the reader to Ferracina and Spijker [8], Gottlieb et al. [9], Higueras [20], Kraaijevanger [22] for reviews on SSPRK methods.

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. There are two main reasons that lead us to investigate this question.

The first reason is that the class of SSP methods is restricted in accuracy and efficiency. For instance, these methods are restricted to fourth-order if one insists on never stepping backward in time. Also, denoting by $c_{\rm os}$ the Shu–Osher coefficient [28] and by s the number of stages of the SSPRK method, the efficiency ratio of the method, defined as $s^{-1}c_{\rm os}$, is often significantly less than 1. For instance, the efficiency ratio of Heun's method (which is a standard second-order SSPRK method) is equal to $\frac{1}{2}$. The efficiency ratio for SSPRK(3,3) is equal to $\frac{1}{3}$, it is equal to $\frac{1}{2}$ for SSPRK(3,4), and it is approximately equal to 0.51 for SSPRK(5,4). (Here and in what follows, the acronym SSPRK(s,p) refers to an s-stage, p-th order SSPRK method.) The SSP paradigm also excludes methods like the second-order midpoint rule. This is unfortunate since the efficiency ratio of the midpoint rule is exactly 1, which is two times larger than that of the second-order Heun's method.

The second reason is that the SSP setting is difficult to deploy in the context of methods combining implicit and explicit time stepping. This problem is particularly evident when solving the compressible Navier–Stokes equations (see, e.g., Demkowicz et al. [5], Guermond et al. [17]). It is known that 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. One then understands that the one-size fits all strategy that underpins the SSP implicit and explicit paradigm cannot properly handle this problem (which set of variables should be used? which invariant domain property should be enforced?). Implicit-explicit schemes are addressed in the forthcoming Part II of this work.

The above difficulties with the SSP paradigm come from the requirement that updates are convex combinations of elementary steps that are IDP. In this paper, we address this difficulty by developing an alternative technique that actually makes every ERK method IDP without constructing convex combinations. We call the resulting time stepping techniques "IDP-ERK methods". Given any ERK method, the main idea is to perform at each stage 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.

We observe that the IDP-ERK methods developed herein rely on ERK methods that are not necessarily SSP (or contractive), i.e., their 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.

The rest of this paper is organized as follows. We introduce the main ideas behind IDP-ERK methods in §2. To pinpoint the key ideas while avoiding distracting technicalities, we do not bother about conservation properties in this section. Then, in §3, we show how to modify the IDP-ERK methods from §2 to make them conservative. Examples of implementation of the methods (including possible choices for the space discretization and the limiters) are given in §4. Numerical tests illustrating the proposed methods are reported in §5.

2 Main ideas on IDP-ERK time stepping

In this section, we briefly present the discrete setting in space and in time, 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 forget the conservation issue in this section. Conservation is addressed in the next section.

2.1 Discrete setting

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.

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 $\mathbf{F}^{\mathrm{H}} : \mathcal{A}^I \to (\mathbb{R}^m)^I$. Further details are given in §3.1. Examples using continuous finite elements, discontinuous finite elements, and finite differences are presented in §4.

In what follows, 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 $(\mathbb{M}\mathbf{V})_{p,i} = \sum_{j\in\mathcal{V}} m_{ij}\mathbf{V}_{p,j}$.

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). Starting from the state vector $\mathbf{U}^n$, we consider the following low-order and high-order updates at t^{n+1}:

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},n+1} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{L}}(\mathbf{U}^n), \tag{3}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},n+1} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^n + \tau\mathbf{F}^{\mathrm{H}}(\mathbf{U}^n). \tag{4}$$

The low-order flux is constructed so that when starting from an admissible state $\mathbf{U}^n \in \mathcal{A}^I$, the update $\mathbf{U}^{\mathrm{L},n+1}$ stays in the admissible domain $\mathcal{A}^I$ under a restriction on the time step (which we also refer to as CFL time step restriction). Unfortunately, this is not the case for the high-order update $\mathbf{U}^{\mathrm{H},n+1}$ which can step out of the admissible domain $\mathcal{A}^I$ for every τ. 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 admissible update. We formalize this limiting process as follows:

$$\mathbf{U}^{n+1} := \ell(\mathbf{U}^n, \mathbf{\Phi}^{\mathrm{L},n}, \mathbf{\Phi}^{\mathrm{H},n}) \in \mathcal{A}^I. \tag{5}$$

The nonlinear limiting operator ℓ 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 two structural assumptions which we are going to invoke later. More precisely, we assume the following:

(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

$$(\mathbf{V} \in \mathcal{A}^I) \Longrightarrow (\mathbf{V} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{F}^{\mathrm{L}}(\mathbf{V}) \in \mathcal{A}^I). \tag{6}$$

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

$$(\mathbf{V} + \tau(\mathbb{M}^{\mathrm{L}})^{-1}\mathbf{\Phi}^{\mathrm{L}} \in \mathcal{A}^I) \Longrightarrow (\ell(\mathbf{V}, \mathbf{\Phi}^{\mathrm{L}}, \mathbf{\Phi}^{\mathrm{H}}) \in \mathcal{A}^I). \tag{7}$$

Other details on the action of the nonlinear limiting operator are given in §3 and in §4.5; the above formalism is sufficient at this stage for our purpose.

The structural assumptions (i)-(ii) are all what is needed to make the forward Euler scheme IDP. Indeed, if the time step is chosen so that $\tau \in (0, \tau^*]$, then Assumption (i) implies that $\mathbf{U}^{\mathrm{L},n+1} \in \mathcal{A}^I$, and thus Assumption (ii) 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.

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

$$\begin{array}{c|ccccc} c_1 & 0 & & & & \\ c_2 & a_{2,1} & 0 & & & \\ c_3 & a_{3,1} & a_{3,2} & 0 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \tag{8}$$

Examples are given in §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, 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 $a_{s+1,k} := b_k$ for all $k \in \{1{:}s\}$ and $c_{s+1} := 1$ (so that $t^{n,s+1} = t^{n+1}$).

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 (ii) 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 idea is to rewrite the l-th stage in incremental form. To do this, we use as starting value at the stage l the invariant domain preserving 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. (This is always possible because the set $\{k < l \mid c_l - c_k \geq 0\}$ contains $\{1\}$ since $c_j \geq 0$ for all $j \in \{2{:}s\}$, i.e., the minimising set is nonempty.) If the sequence $(c_l)_{l\in\{1:s\}}$ is nondecreasing (as it often happens), then $l' = l - 1$ for all $l \in \{2{:}s + 1\}$. The reason for looking for the smallest difference $c_l - c_{l'}$ is that we want to minimize the CFL restriction on the time step (see Lemma 2.1). Notice that l' depends on l, so that we should write $l'(l)$, but we simply write l' to alleviate the notation.

Let us now provide more details. Since $c_1 = 0$, we start by setting $\mathbf{U}^{n,1} := \mathbf{U}^n$. Then, for all $l \in \{2{:}s + 1\}$, we assume that all the states $\mathbf{U}^{n,1}, \ldots, \mathbf{U}^{n,l-1}$ have been computed and are all in $\mathcal{A}^I$ (this property will be established by induction). First, we compute the provisional low-order update

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},l} := \mathbb{M}^{\mathrm{L}}\mathbf{U}^{n,l'} + \tau(c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}). \tag{9}$$

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 our above convention if $l = s+1$):

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n} + \tau \sum_{k\in\{1:l-1\}} a_{l,k}\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{10}$$

But, to be able to compare $\mathbf{U}^{\mathrm{H},l}$ and $\mathbf{U}^{\mathrm{L},l}$ and to perform a limiting process, we want to define $\mathbf{U}^{\mathrm{H},l}$ by using $\mathbf{U}^{n,l'}$ as the starting value. To 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 $k \geq l'$, this gives

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l'} + \tau \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{11}$$

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 (11), the equation we use for the evaluation of the provisional high-order update $\mathbf{U}^{\mathrm{H},l}$ is

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{12}$$

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

$$\mathbf{U}^{n,l} := \ell(\mathbf{U}^{n,l'}, \mathbf{\Phi}^{\mathrm{L},l}, \mathbf{\Phi}^{\mathrm{H},l}), \tag{13}$$

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

$$\mathbf{\Phi}^{\mathrm{L},l} := (c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}), \qquad \mathbf{\Phi}^{\mathrm{H},l} := \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{14}$$

To sum up, the s-stage IDP-ERK method proceeds as described in Algorithm 1.

Algorithm 1 s-stage IDP-ERK scheme.

Input: $\mathbf{U}^n \in \mathcal{A}^I$
Set $\mathbf{U}^{n,1} := \mathbf{U}^n$
for $l = 2, \ldots, s+1$ **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. $\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} := \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau\sum_{k\in\{1:l-1\}}(a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k})$ (High-order update)
3. $\mathbf{U}^{n,l} := \ell(\mathbf{U}^{n,l'}, \mathbf{\Phi}^{\mathrm{L},l}, \mathbf{\Phi}^{\mathrm{H},l})$ with fluxes defined in (14) (Limiting)

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

Lemma 2.1 (IDP)**.** *Let τ^* be the maximal time-step from Assumption* (6) *and assume that $\tau \max_{l\in\{2:s+1\}}(c_l - c_{l'}) \leq \tau^*$. Assume that $\mathbf{U}^n \in \mathcal{A}^I$ and that $\mathbf{U}^{n+1}$ is computed by Algorithm 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 (6), 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 (9) imply that $\mathbf{U}^{\mathrm{L},l} \in \mathcal{A}^I$. As a result, the definition (13) 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 $\mathbf{U}^{n+1} := \mathbf{U}^{n,s+1} \in \mathcal{A}^I$. □

Definition 2.2 (Efficiency ratio)**.** *Let* $c_{\mathrm{eff}} := \max_{l\in\{2:s+1\}}(c_l - c_{l'})$ *(using again the convention* $c_{s+1} = 1$*). The quantity* $c_{\mathrm{eff}}^{-1} s^{-1}$ *is called* efficiency ratio *of the* s*-stage IDP-ERK method.*

To motivate the above definition (inspired from Shu and Osher [28]), let us compare the algorithmic complexity of an s-stage IDP-ERK scheme with that of the forward Euler method. The maximum time step that makes the s-stage method IDP is $c_{\mathrm{eff}}^{-1}\tau^*$. Assuming that one wants to reach some fixed time T, the number of time steps required to do so is approximately $c_{\mathrm{eff}}T/\tau^*$. Since each time step requires estimating s fluxes and performing s limiting operations, the algebraic complexity of the method scales like $sc_{\mathrm{eff}}T/\tau^*$. 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 the above-defined efficiency ratio $c_{\mathrm{eff}}^{-1}s^{-1}$. Since the minimal value of c_{eff} is obtained when the coefficients $\{c_l\}_{l\in\{1:s+1\}}$ are ordered and equi-distributed, we infer that $s^{-1} \le c_{\mathrm{eff}}$. Hence $c_{\mathrm{eff}}^{-1}s^{-1} \le 1$. This computation shows that the s-stage IDP-ERK methods that are the most efficient according to this criterion are those for which the stages are equi-distributed, i.e., $c_l = \frac{l-1}{s}$, $l \in \{1{:}s+1\}$, leading to an efficiency ratio of one. Choosing equi-distributed coefficients is not always possible (it is definitively possible for $s \in \{2, 3\}$), but in any case, it indicates the ideal situation.

Remark 2.3 (Comparison with SSP)**.** *The computational effort deployed in each stage of an IDP-ERK method is the same as that deployed for an SSPRK 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 algorithm.*

3 Conservative IDP-ERK time stepping

In the previous section, we only focused on the invariant domain property (1). In this section, we show how to achieve the conservation property (2) as well. Being conservative is essential for the approximation of conservation equations, since (up to appropriate boundedness assumptions) conservation implies convergence to weak solutions with shocks moving at the right speed. The material presented in this section is inspired from the flux corrected transport literature (see Boris and Book [3], Kuzmin and Turek [23], Kuzmin et al. [24, 25], Zalesak [30]) and from the convex limiting literature (see e.g., [15, 16]). Originality is only claimed for the content of §3.3.

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 dofs 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

$$\mathbf{F}_i^{\mathrm{L}}(\mathbf{V}) = \sum_{j \in \mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{V}), \qquad \mathbf{F}_i^{\mathrm{H}}(\mathbf{V}) = \sum_{j \in \mathcal{I}(i)} \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{V}), \tag{15}$$

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 $j \in \mathcal{I}(i)$:

$$\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{V}) = -\mathbf{F}_{ji}^{\mathrm{L}}(\mathbf{V}), \qquad \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{V}) = -\mathbf{F}_{ji}^{\mathrm{H}}(\mathbf{V}). \tag{16}$$

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

$$\mathbb{M}_{ij}^{\mathrm{L}} = m_i \delta_{ij}, \ \forall (i,j) \in \mathcal{V}^2 \quad \text{and} \quad m_i > 0, \ \forall i \in \mathcal{V}. \tag{17}$$

This assumption is justified in [14], 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

$$(\mathbb{M}^{\mathrm{H}}\mathsf{X})_i = \sum_{j \in \mathcal{I}(i)} m_{ij} \mathsf{X}_j, \quad \forall \mathsf{X} \in \mathbb{R}^I, \quad \text{and} \quad m_{ij} = m_{ji}, \quad \forall i \in \mathcal{V}, \forall j \in \mathcal{I}(i). \tag{18}$$

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

$$m_i = \sum_{j \in \mathcal{I}(i)} m_{ij}, \qquad \forall i \in \mathcal{V}. \tag{19}$$

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* (15) *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 §4.4. We also refer the reader to Abgrall et al. [2] and Pazner [27] where the stencil mismatch question is addressed.*

Remark 3.2 (Two-point fluxes)**.** *In the literature, one often considers two-point fluxes, meaning that $\mathbf{F}_{ij}$ only depends on the pair $(\mathbf{V}_i, \mathbf{V}_j)$ (see e.g.,* (39)-(40) *and* (45)*). We are not making this assumption since it is too restrictive. In particular, the two-point flux assumption cannot be used when using high-order space discretization schemes (see §4.4 for an example).*

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 §2.2(ii) that is conservative. We first explain the method using the forward Euler method. The method is explained in full generality in §3.3.

We rewrite (3)-(4) in component form as follows: For all $i \in \mathcal{V}$,

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

Following the flux corrected transport strategy (see Boris and Book [3], Kuzmin and Turek [23], Kuzmin et al. [24, 25], Zalesak [30]), we subtract the first equation from the second one and obtain

$$\mathbf{U}_i^{\mathrm{H},n+1} = \mathbf{U}_i^{\mathrm{L},n+1} + m_i^{-1}\tau \sum_{j\in\mathcal{I}(i)} \mathbf{A}_{ij}^n, \tag{20}$$

with

$$\mathbf{A}_{ij}^n := \mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^n) - \mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^n) + \frac{m_{ij} - m_i\delta_{ij}}{\tau}(\mathbf{U}_j^n - \mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_i^n + \mathbf{U}_i^{\mathrm{H},n+1}), \tag{21}$$

where we used (19) to infer that $\sum_{j\in\mathcal{I}(i)}(m_{ij} - m_i\delta_{ij})(\mathbf{U}_i^n + \mathbf{U}_i^{\mathrm{H},n+1}) = \mathbf{0}$. The purpose of this manipulation is to obtain the following skew-symmetry property:

$$\mathbf{A}_{ij}^n = -\mathbf{A}_{ji}^n, \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{22}$$

We are now ready to formalize the conservative limiting operator.

Definition 3.3 (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 values 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 conservative limiter any operator from $\mathcal{A}^I\times\mathfrak{M}$ to $\mathfrak{L}$, say $(\{\mathbf{V}_i\}_{i\in\mathcal{V}}, \{\mathbf{A}_{ij}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}) \mapsto \{\ell_{ij}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$, so that*

$$\mathbf{V}_i + m_i^{-1}\tau \sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathbf{A}_{ij} \in \mathcal{A}, \quad \forall i \in \mathcal{V}. \tag{23}$$

The limited 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 $\ell(\mathbf{V},\mathbf{A})$.

Notice that the existence of conservative limiters is always guaranteed since the trivial limiter $\ell(\mathbf{V},\mathbf{A}) = \mathbf{V}$ (i.e., $\ell_{ij} = 0$ for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$) is always admissible 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 §4.5.

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

$$\mathbf{U}^{n+1} := \ell(\mathbf{U}^{\mathrm{L},n+1}, \mathbf{A}^n), \tag{24}$$

with $\mathbf{A}^n$ defined in (21). The definition (24) can be recast into the following form:

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n + m_i^{-1}\tau\sum_{i\in\mathcal{V}} \ell_{ij}\mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^n) + (1-\ell_{ij})\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^n)$$
$$+ \sum_{i\in\mathcal{V}} \ell_{ij}(m_{ij} - m_i\delta_{ij})(\mathbf{U}_j^n - \mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_i^n + \mathbf{U}_i^{\mathrm{H},n+1}). \tag{25}$$

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

(i) *Let τ^* be the maximal time-step from Assumption* (6)*. 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 (6), we conclude that $\mathbf{U}^{\mathrm{L},n+1} \in \mathcal{A}^I$. As a result, the definition (24) make sense, and the construction of the limiter implies that $\mathbf{U}^{n+1} \in \mathcal{A}^I$.
(ii) The definition (15), the skew-symmetry assumption (16), and the property (22) 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 (24) 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 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:

$$\mathbb{M}^{\mathrm{L}}\mathbf{U}^{\mathrm{L},l} = \mathbb{M}^{\mathrm{L}}\mathbf{U}^{n,l'} + \tau(c_l - c_{l'})\mathbf{F}^{\mathrm{L}}(\mathbf{U}^{n,l'}), \tag{26}$$

$$\mathbb{M}^{\mathrm{H}}\mathbf{U}^{\mathrm{H},l} = \mathbb{M}^{\mathrm{H}}\mathbf{U}^{n,l'} + \tau \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}^{\mathrm{H}}(\mathbf{U}^{n,k}). \tag{27}$$

By proceeding as in (21), we now define the following skew-symmetric fluxes:

$$\mathbf{A}_{ij}^{n,l} := \sum_{k\in\{1:l-1\}} (a_{l,k} - a_{l',k})\mathbf{F}_{ij}^{\mathrm{H}}(\mathbf{U}^{n,k}) - (c_l - c_{l'})\mathbf{F}_{ij}^{\mathrm{L}}(\mathbf{U}^{n,l'}) + \tau^{-1}(m_{ij} - m_i\delta_{ij})(\mathbf{U}_j^{n,l'} - \mathbf{U}_j^{\mathrm{H},l} - \mathbf{U}_i^{n,l'} + \mathbf{U}_i^{\mathrm{H},l}). \tag{28}$$

Given some conservative limiting operator ℓ satisfying the requirements of Definition 3.3, we then set

$$\mathbf{U}^{n,l} := \ell(\mathbf{U}^{\mathrm{L},l}, \mathbf{A}^{n,l}), \qquad \forall l \in \{2{:}s+1\}. \tag{29}$$

Lemma 3.5 (IDP and conservation)**.** *Recall from Definition 2.2 that $c_{\mathrm{eff}} := \max_{l\in\{2:s+1\}}(c_l - c_{l'})$. Then the following holds true.*

(i) *Let τ^* be the maximal time-step from Assumption* (6)*. Assume that $\mathbf{U}^n \in \mathcal{A}^I$ and $\tau \in (0, c_{\mathrm{eff}}^{-1}\tau^*]$. 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 is omitted for brevity since it combines the arguments of the proofs of Lemma 2.1 and Lemma 3.4. □

4 Examples

In this section, we illustrate the theory presented above by giving examples. We list some ERK techniques in §4.1. We also briefly describe in §4.2–§4.5 the use of the method in the context of the approximation of the nonlinear conservation equation $\partial_t \boldsymbol{u} + \nabla\cdot\mathbb{f}(\boldsymbol{u}) = 0$ with $\boldsymbol{u} : D \to \mathbb{R}^m$, $D \subset \mathbb{R}^d$, and $\mathbb{f} \in C^1(\mathbb{R}^m;\mathbb{R}^{m\times d})$. We show how the low- and high-order fluxes can be computed with continuous and discontinuous finite elements and with finite differences.

4.1 Examples of ERK methods

We give here examples of ERK methods that are not SSP. These methods are used in the numerical section (see §5). We use the notation RK$(s, p;\, e)$ where s indicates the number of stages, p the order, and e the efficiency ratio (see Definition 2.2 and recall that a method with optimally equi-distributed sub-steps (i.e., with increment $\frac{1}{s}$) reaches the best possible value $e = 1$).

Three examples of ERK methods with optimally equi-distributed sub-steps are as follows:

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

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

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

(30)

The method in the leftmost tableau in (30) is the midpoint rule. It is second-order accurate. The second and third methods in (30) are both third-order accurate. The second method in (30) is often called Heun's third-order method. The third method in (30) satisfies all the necessary and sufficient conditions to be fourth-order on linear problems (and it is the only one with optimally equi-distributed sub-steps to do so), but it is only third-order accurate on nonlinear problems. This proves in passing that there does not exist a four-stage ERK method that is genuinely fourth-order and has optimally equi-distributed sub-steps.

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

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

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

(31)

The first method in (31) is a popular fourth-order accurate ERK method, but its efficiency ratio is rather small (only $\frac{1}{2}$). The second method in (31) is often called $\frac{3}{8}$ rule. It has equi-distributed sub-steps, but the distribution is slightly sub-optimal (the increment is $\frac{1}{3}$ instead of $\frac{1}{4}$).

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

$$\begin{array}{c|cccccc} 0 & 0 & & & & & \\ \frac{1}{5} & \frac{1}{5} & 0 & & & & \\ \frac{2}{5} & 0 & \frac{2}{5} & 0 & & & \\ \frac{3}{5} & \frac{3}{20} & 0 & \frac{9}{20} & 0 & & \\ \frac{4}{5} & \frac{4}{5} & -\frac{8}{5} & \frac{8}{5} & 0 & 0 & \\ 1 & -\frac{71}{4} & 40 & -\frac{75}{4} & -10 & \frac{15}{2} & 0 \\ \hline 1 & \frac{17}{144} & 0 & \frac{25}{36} & -\frac{25}{72} & \frac{25}{48} & \frac{1}{72} \end{array} \qquad \begin{array}{c|cccccc} 0 & 0 & & & & & \\ \frac{1}{4} & \frac{1}{4} & 0 & & & & \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & 0 & & & \\ \frac{1}{2} & 0 & -\frac{1}{2} & 1 & 0 & & \\ \frac{3}{4} & \frac{3}{16} & 0 & 0 & \frac{9}{16} & 0 & \\ 1 & -\frac{3}{7} & \frac{2}{7} & \frac{12}{7} & -\frac{12}{7} & \frac{8}{7} & 0 \\ \hline 1 & \frac{7}{90} & 0 & \frac{32}{90} & \frac{12}{90} & \frac{32}{90} & \frac{7}{90} \end{array} \tag{32}$$

RK(6,5;$\frac{5}{6}$) RK(6,5;$\frac{2}{3}$)

The first method in (32) has equi-distributed sub-steps, but the distribution is slightly sub-optimal (the increment is $\frac{1}{5}$ instead of $\frac{1}{6}$). Moreover, some of the coefficients in the fifth subtep are rather large. The second method in (32) has been proposed in Butcher [4], and its efficiency ratio is a bit smaller than the first method ($\frac{2}{3}$ vs. $\frac{5}{6}$).

We are also going to test standard SSPRK techniques of second-, third-, and fourth-order with optimal efficiency ratio. Using the present notation, we refer to these methods as SSPRK(2,2;$\frac{1}{2}$), SSPRK(3,3;$\frac{1}{3}$), and SSPRK(5,4;0.51). The Butcher tableau of the first two methods is as follows:

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array} \tag{33}$$

SSPRK(2,2;$\frac{1}{2}$) SSPRK(3,3;$\frac{1}{3}$)

The Butcher tableau of the third method can be found in Kraaijevanger [22, p. 522]. The first method in (33) is known as Heun's method, and the second method is sometimes called Fehlberg's method [7].

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 SSPRK(3,3;$\frac{1}{3}$) and SSPRK(5,4;0.51) 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 cG

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. [13], [11], [16, §4.2], Ern and Guermond [6, Chap. 81], Kuzmin and Turek [23]). 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:

$$P(\mathcal{T}_h) := \{v \in \mathcal{C}^0(D;\mathbb{R}) \mid v_{|K} \circ \boldsymbol{T}_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\}, \qquad \boldsymbol{P}(\mathcal{T}_h) := [P(\mathcal{T}_h)]^m. \tag{34}$$

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

$$m_{ij}^{\mathrm{H}} := \int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}x, \qquad m_{ij}^{\mathrm{L}} := \delta_{ij}\int_D \varphi_i(\boldsymbol{x})\,\mathrm{d}x, \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i), \tag{35}$$

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

$$\boldsymbol{c}_{ij} := \int_D \varphi_i(\boldsymbol{x}) \nabla \varphi_j(\boldsymbol{x}) \,\mathrm{d}x, \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{36}$$

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

$$\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}, \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{37}$$

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

$$d_{ij}^{\mathrm{L}} = d_{ji}^{\mathrm{L}} \geq 0 \quad \text{and} \quad d_{ij}^{\mathrm{H}} = d_{ji}^{\mathrm{H}} \geq 0, \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{38}$$

We refer to [11] and [6, Chap. 81] for the construction of $\{d_{ij}^{\mathrm{L}}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$ (essentially d_{ij}^{L} scales as $\|\boldsymbol{c}_{ij}\|_{\ell^2}$ multiplied by a maximum wave speed associated with a suitable local Riemann problem), and to [13, 23], [16, §6], and [6, Chap. 82-83] for examples of constructions of $\{d_{ij}^{\mathrm{H}}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$. With these definitions, we set

$$\boldsymbol{F}_{ij}^{\mathrm{L}}(\mathbf{V}) := -(\mathbb{f}(\mathbf{V}_j) + \mathbb{f}(\mathbf{V}_i))\boldsymbol{c}_{ij} + d_{ij}^{\mathrm{L}}(\mathbf{V}_j - \mathbf{V}_i), \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i), \tag{39}$$

$$\boldsymbol{F}_{ij}^{\mathrm{H}}(\mathbf{V}) := -(\mathbb{f}(\mathbf{V}_j) + \mathbb{f}(\mathbf{V}_i))\boldsymbol{c}_{ij} + d_{ij}^{\mathrm{H}}(\mathbf{V}_j - \mathbf{V}_i), \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{40}$$

(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 (16) is satisfied (up to details regarding the boundary conditions which are beyond the scope of the paper). Finally, setting $\mathcal{I}^*(i) := \mathcal{I}(i) \setminus \{i\}$, it can be shown that the structural assumption (6) holds true for

$$\tau \leq \tau^* := \frac{1}{2} \min_{i\in\mathcal{V}} \frac{m_i}{\sum_{j\in\mathcal{I}^*(i)} d_{ij}^{\mathrm{L}}}. \tag{41}$$

4.3 dG

The construction described above can be repeated verbatim for discontinuous finite elements. The main difference is that the finite element spaces $P(\mathcal{T}_h)$ and $\boldsymbol{P}(\mathcal{T}_h)$ are now defined by

$$P(\mathcal{T}_h) := \{v \in L^\infty(D;\mathbb{R}) \mid v_{|K} \circ \boldsymbol{T}_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\}, \quad \boldsymbol{P}(\mathcal{T}_h) := [P(\mathcal{T}_h)]^m. \tag{42}$$

Denoting by n the dimension of the reference space $\widehat{P}$, the global shape functions can be enumerated as follows $\{\varphi_{lK}\}_{l\in\{1:\mathsf{n}\},K\in\mathcal{T}_h}$. Assuming that the cells in $\mathcal{T}_h$ are enumerated from 1 to M, and denoting by $\mathsf{m} : \mathcal{T}_h \to \{1{:}M\}$ the corresponding enumeration map, we define a global enumeration of the shape functions by saying that $i := \mathsf{n}\times(\mathsf{m}(K)-1)+l$ is the global enumeration index of φ_{lK}; hence, an alternative notation for φ_{lK} is φ_i. Notice that $\mathcal{V} := \{1{:}\mathsf{n}\times M\}$.

The definitions of the consistent and of the lumped mass matrices are unchanged. Assuming that $i = \mathsf{n}\times(\mathsf{m}(K)-1)+l$ and $j = \mathsf{n}\times(\mathsf{m}(T)-1)+m$, the $\mathbb{R}^d$-valued coefficients $\boldsymbol{c}_{ij}$ are defined by using the centered numerical flux as follows:

$$\boldsymbol{c}_{ij} := \begin{cases} -\displaystyle\int_K \varphi_{mT}\nabla\varphi_{lK}\,\mathrm{d}x + \frac{1}{2}\int_{\partial K} \varphi_{mT}\varphi_{lK}\boldsymbol{n}_K\,\mathrm{d}s & \text{if } K = T. \\ \displaystyle\frac{1}{2}\int_{\partial K} \varphi_{mT}\varphi_{lK}\boldsymbol{n}_K\,\mathrm{d}s & \text{if } K \neq T,\ |\partial K \cap \partial T| > 0. \end{cases} \tag{43}$$

Here, $\boldsymbol{n}_K$ is the unit outward normal on ∂K. Notice that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$. The definitions of the low-order and high-order graph viscosity matrices $\{d_{ij}^{\mathrm{L}}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$, $\{d_{ij}^{\mathrm{H}}\}_{i\in\mathcal{V},j\in\mathcal{I}(i)}$ are similar to what can be done for continuous finite elements. Then the low-order and the high-order fluxes can be defined as in (39)-(40), and one can again establish that the structural assumption (6) holds true under the condition (41).

4.4 Finite differences

We finish with a short example involving finite differences in one space dimension with periodic boundary conditions. This setting is used in §5 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 $\mathcal{V} := \{1{:}N-1\}$.

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 set $\mathcal{I}(i) := \{i-1, i, i+1\}$ with the convention that 0 and N are replaced by $N-1$ and 1, respectively. For every $i \in \mathcal{V}$, we also define the coefficients

$$\boldsymbol{c}_{i,i-1} := -\frac{1}{2}, \qquad \boldsymbol{c}_{i,i} := 0, \qquad \boldsymbol{c}_{i,i+1} := \frac{1}{2}. \tag{44}$$

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

$$\mathbf{F}^{\mathrm{L}}_{ij}(\mathbf{V}) := -(\mathbb{f}(\mathbf{V}_j) + \mathbb{f}(\mathbf{V}_i))\boldsymbol{c}_{ij} + d^{\mathrm{L}}_{ij}(\mathbf{V}_j - \mathbf{V}_i), \qquad \forall i \in \mathcal{V},\ \forall j \in \mathcal{I}(i). \tag{45}$$

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 (6) holds true under the condition (41).

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). As announced in Remark 3.1 and Remark 3.2, we 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

$$\begin{aligned}\mathbf{F}^{\mathrm{H}}_{i,i-1}(\mathbf{V}) := &-\frac{1}{12}(\mathbb{f}(\mathbf{V}_{i-2}) - \mathbb{f}(\mathbf{V}_{i-1}) - \mathbb{f}(\mathbf{V}_i) + \mathbb{f}(\mathbf{V}_{i+1})) \\ &+ \frac{6}{12}(\mathbb{f}(\mathbf{V}_{i-1}) + \mathbb{f}(\mathbf{V}_i)) + d^{\mathrm{H}}_{i,i-1}(\mathbf{V}_{i-1} - \mathbf{V}_i),\end{aligned} \tag{46}$$

$$\begin{aligned}\mathbf{F}^{\mathrm{H}}_{i,i+1}(\mathbf{V}) := &\frac{1}{12}(\mathbb{f}(\mathbf{V}_{i-1}) - \mathbb{f}(\mathbf{V}_i) - \mathbb{f}(\mathbf{V}_{i+1}) + \mathbb{f}(\mathbf{V}_{i+2})) \\ &- \frac{6}{12}(\mathbb{f}(\mathbf{V}_i) + \mathbb{f}(\mathbf{V}_{i+1})) + d^{\mathrm{H}}_{i,i+1}(\mathbf{V}_{i+1} - \mathbf{V}_i).\end{aligned} \tag{47}$$

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

$$\mathbf{F}^{\mathrm{H}}_i(\mathbf{V}) := \sum_{j\in\mathcal{I}(i)} \mathbf{F}^{\mathrm{H}}_{i,j}(\mathbf{V}) = -\frac{1}{12}(\mathbb{f}(\mathbf{V}_{i-2}) - 8\mathbb{f}(\mathbf{V}_{i-1}) + 8\mathbb{f}(\mathbf{V}_{i+1}) - \mathbb{f}(\mathbf{V}_{i+2})), \tag{48}$$

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

$$\mathbf{F}^{\mathrm{H}}_{ij}(\mathbf{V}) = -\mathbf{F}^{\mathrm{H}}_{ji}(\mathbf{V}), \quad \forall i \in \mathcal{I},\ \forall j \in \mathcal{I}(i), \tag{49}$$

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

4.5 Limiting

There are many ways to perform the limiting operation mentioned in (5), (13), Definition 3.3, and (29). The so-called flux transport correction technique of Boris and Book [3] and Zalesak [30] is

probably the most well-known limiting technique for scalar conservation equations. The reader is also referred to Kuzmin and Turek [23] and the books Kuzmin et al. [24, 25] for other extensions on this method. But, as observed in [15, 16], 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 set $\mathcal{A}$ introduced in the structural assumption (6) is of the following form:

$$\mathcal{A} = \bigcap_{l\in\mathcal{L}} \{\mathbf{V} \in \mathbb{R}^m \mid \Psi_l(\mathbf{V}) \geq 0\}, \tag{50}$$

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 [15, 16]. We refer the reader to these two references for the details on how convex limiting can be used in (29).

Remark 4.1 (Local bounds)**.** *The generic invariant domain property assumed in* (6) *can be localized. More precisely, given $i \in \mathcal{V}$ and some initial states $\{\mathbf{V}_j\}_{j\in\mathcal{I}(i)}$, with $\mathbf{V}_j \in \mathcal{A}$ for all $j \in \mathcal{I}(i)$, one can often construct a subset $\mathcal{A}_i \subset \mathcal{A}$ that is such that $\mathbf{V}_i + \tau m_i^{-1}\mathbf{F}_i^{\mathrm{L}}(\mathbf{V}) \in \mathcal{A}_i$ for all $\tau \leq \tau^*$. Let us illustrate this point for scalar conservation equations. Given some initial data $\mathbf{V} \in \mathbb{R}^I$, one first defines the global invariant set $\mathcal{A} := [\mathsf{V}^{\min}, \mathsf{V}^{\max}]$ 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* (6) *simply formalizes that the low-order method satisfies the global maximum/minimum principle, i.e., $\mathbf{V}_i + \tau m_i^{-1}\mathbf{F}_i^{\mathrm{L}}(\mathbf{V}) \in [\mathsf{V}^{\min}, \mathsf{V}^{\max}]$ for all $i \in \mathcal{V}$. But, very often one can show that for every $i \in \mathcal{V}$, 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 $\mathsf{V}_i^{\max} := \max_{j\in\mathcal{I}(i)} \mathbf{V}_j$, 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 invoked in* (29) *to be implemented with local bounds, which gives a tighter control on the approximate solution.*

Remark 4.2 (Bounds relaxation)**.** *As mentioned by many authors working on conservation equations, strictly enforcing the maximum/minimum principle degrades the converge rate to first-order close to extrema (see e.g., Khobalatte and Perthame [21, §3.3], Zhang and Shu [32, p. 2753]). A typical way to address this issue in the finite volume literature consists of relaxing the slope reconstructions; see e.g., Harten [18, Eq. (5.7)], Harten and Osher [19]. Similar techniques can be used with discontinuous Galerkin approximations as in Zhang and Shu [31, 32]. In the numerical results reported in §5, the local bounds mentioned in Remark 4.1 are all relaxed using the process explained in Guermond et al. [15, §4.7] and Guermond et al. [16, §7.6].*

5 Numerical illustrations

We numerically illustrate the method described above on conservation equations. The tests are performed with continuous finite elements and finite differences for the space approximation.

5.1 Numerical details

For all the tests reported here, the limiting is performed with two iterations of the convex limiting technique introduced in [15, 16]. 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 enforced up to machine accuracy.

In all the tests, the time step is computed once at the beginning of each simulation by using the expression

$$\tau := \mathrm{CFL}\times s\times\tau^*, \tag{51}$$

where CFL > 0 is a fixed parameter, s the number of stages of the ERK method, and τ^* the maximum time step defined in (41). Given some final time T and given some mesh $\mathcal{T}_h$ determining the value of τ^*, the choice (51) of τ guarantees that all the ERK methods described in §4 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

$$c_{\mathrm{eff}}\tau = \mathrm{CFL}\times c_{\mathrm{eff}}\times s\times\tau^*,$$

and all the methods are invariant-domain preserving if $c_{\mathrm{eff}}\tau \le \tau^*$, we conclude that all the methods are invariant-domain preserving provided CFL $\in (0, \frac{1}{c_{\mathrm{eff}}s}]$ (recall that c_{eff} is the efficiency ratio introduced in Definition 2.2 and that $\frac{1}{c_{\mathrm{eff}}s} \le 1$).

5.2 Convergence tests with smooth solutions

We illustrate the method by solving the following linear transport equation in one and two space dimensions:

$$\partial_t u + \nabla\cdot(\boldsymbol{\beta} u) = 0. \tag{52}$$

5.2.1 Fourth-order finite differences in 1D

We consider the problem (52) in the one-dimensional domain $D := (0,1)$ with $\boldsymbol{\beta} := 1$ and the initial data $u_0(x) := (4\frac{(x-x_0)(x_1-x)}{x_1-x_0})^6$ if $x \in (x_0, x_1)$, and $u_0(x) := 0$ otherwise, with $x_0 := 0.1$ and $x_0 := 0.4$. We enforce periodic boundary conditions. The tests are realized as explained in §4.4 with fourth-order finite differences using uniform meshes. The space approximation is formally fourth-order accurate. We test the ten ERK methods described in §4. All the errors are reported at $T := 1$.

We show in Table 5.2.1 the relative error in the L^∞-norm for the second-order methods RK(2,2;1) and RK(2,2;$\frac{1}{2}$). Both methods perform as expected at CFL $= 0.2$. There is a slight difference of behavior at CFL $= 0.25$. Heun's method is no longer in the asymptotic convergence regime for CFL $= 0.25$, whereas the midpoint method still behaves properly.

Table 1: Linear transport, 1D finite differences, second-order methods.

	CFL = 0.2				CFL = 0.25			
I	RK(2,2;1)	rate	RK(2,2;$\frac{1}{2}$)	rate	RK(2,2;1)	rate	RK(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

We show in Table 5.2.1 the relative error in the L^∞-norm for the third-order methods RK(3,3;1), RK(3,3;$\frac{1}{3}$), and RK(4,3;1). The method RK(3,3;1) delivers third-order accuracy for CFL $= 0.05$. The performance of the method RK(3,3;$\frac{1}{3}$) degrades as the mesh size is refined. The performance of both methods degrades as the CFLnumber grows. The simulations at CFL $= 0.25$ show that the performance of the SSP method RK(3,3;$\frac{1}{3}$) degrades faster than its optimal counterpart RK(3,3;1). Notice also that the method RK(4,3;1) behaves extremely well since it delivers fourth-order accuracy for both CFLnumbers. 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.

Table 2: Linear transport, 1D finite differences, third-order methods.

	CFL = 0.05						CFL = 0.25					
I	RK(3,3;1)	rate	RK(3,3;$\frac{1}{3}$)	rate	RK(4,3;1)	rate	RK(3,3;1)	rate	RK(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

We show in Table 5.2.1 the relative error in the L^∞-norm for the fourth-order methods RK(4,4;$\frac{1}{2}$), RK(4,4;$\frac{3}{4}$), and RK(5,4;0.51) (we use the notation RK(5,4;$\frac{1}{2}$) instead of RK(5,4;0.51) in the table to save some horizontal space). We observe that the method RK(4,4;$\frac{1}{2}$) delivers fourth-order accuracy in the L^∞-norm up to CFL = 0.1. The SSP method RK(5,4;0.51) delivers fourth-order accuracy for the lower value CFL = 0.05. The method RK(4,4;$\frac{3}{4}$) is not yet in its asymptotic range for CFL = 0.05. This result shows that considerations on the efficiency coefficient are not enough to classify the methods. This test nevertheless shows that the popular fourth-order method RK(4,4;$\frac{1}{2}$) vastly outperforms the SSP method RK(5,4;0.51). We recall that the definition of the time step (51) is such that for any given CFL number, all the methods perform exactly the same number of flux evaluations, so that the above comparison is, in our opinion, fair.

Table 3: Linear transport, 1D finite differences, fourth-order methods.

	CFL = 0.05						CFL = 0.1					
I	RK(4,4;$\frac{1}{2}$)	rate	RK(4,4;$\frac{3}{4}$)	rate	RK(5,4;$\frac{1}{2}$)	rate	RK(4,4;$\frac{1}{2}$)	rate	RK(4,4;$\frac{3}{4}$)	rate	RK(5,4;$\frac{1}{2}$)	rate
50	4.32E-02	–	4.72E-02	–	4.32E-02	–	6.35E-02	–	5.18E-02	–	6.28E-02	–
100	5.41E-03	3.00	5.40E-03	3.13	5.41E-03	3.00	5.36E-03	3.57	5.20E-03	3.31	5.66E-03	3.47
200	3.79E-04	3.84	3.79E-04	3.83	3.79E-04	3.83	3.79E-04	3.82	3.79E-04	3.78	3.79E-04	3.90
400	2.27E-05	4.06	2.27E-05	4.06	2.27E-05	4.06	2.27E-05	4.06	2.59E-05	3.87	2.27E-05	4.06
800	1.58E-06	3.85	1.58E-06	3.85	1.58E-06	3.84	1.58E-06	3.84	4.05E-06	2.68	1.58E-06	3.85
1600	8.13E-08	4.28	2.88E-07	2.46	8.58E-08	4.20	8.13E-08	4.28	9.94E-07	2.03	1.13E-07	3.80
3200	5.36E-09	3.92	6.98E-08	2.04	8.95E-09	3.26	4.97E-09	4.03	2.45E-07	2.02	2.72E-08	2.06

We show in Table 5.2.1 the relative error in the L^∞-norm for the fifth-order methods RK(6,5;$\frac{5}{6}$) and RK(6,5;$\frac{2}{3}$). We observe that the two methods reach their asymptotic convergence range for CFL = 0.025 (recall that here the space discretization errors should dominate on fine meshes). We notice that, as anticipated, the method RK(6,5;$\frac{5}{6}$) performs slightly better than RK(6,5;$\frac{2}{3}$) although the differences are rather small.

Table 4: Linear transport, 1D finite differences, fifth-order methods.

	CFL = 0.02				CFL = 0.025			
I	RK(6,5;$\frac{5}{6}$)	rate	RK(6,5;$\frac{2}{3}$)	rate	RK(6,5;$\frac{5}{6}$)	rate	RK(6,5;$\frac{2}{3}$)	rate
50	5.19E-02	–	5.19E-02	–	5.20E-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	3.84
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	3.84	1.58E-06	3.85	1.58E-06	3.85	1.58E-06	3.85
1600	8.13E-08	4.28	8.48E-08	4.22	8.24E-08	4.26	8.71E-08	4.18
3200	6.24E-09	3.70	7.10E-09	3.58	6.32E-09	3.70	1.16E-08	2.91

5.2.2 Second-order finite elements

We now solve the problem (52) in the two-dimensional domain $D := (0,1)^2$ with $\boldsymbol{\beta} := (0.9,1)^\intercal$ and the initial data $u_0(\boldsymbol{x}) := (4\frac{(x-x_0)(x_1-x)}{x_1-x_0})^4 \times (4\frac{(y-y_0)(y_1-y)}{y_1-y_0})^4$ 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 $T := 0.5$ with CFL $= 0.2$. 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 be tangential to any mesh edge to avoid any extraneous super-convergence effects.) The Lagrange shape functions are invariant by central symmetry in the support of every nodal basis function; this guarantees that the method is super-convergent up to third-order at the mesh nodes (see [10, Prop. 2.1] and Thompson [29, Prop. 4.4]).

Table 5: Linear transport, 2D $\mathbb{P}_1$ finite elements on uniform meshes. Relative error in the L^1- and L^∞-norms for the second- and third-order ERK methods at CFL $= 0.2$.

L^1-norm										
I	RK(2,2;1)	rate	RK(2,2;$\frac{1}{2}$)	rate	RK(3,3;1)	rate	RK(3,3;$\frac{1}{3}$)	rate	RK(4,3;1)	rate
51^2	2.34E-02	–	2.38E-02	–	3.09E-02	–	3.16E-02	–	3.13E-02	–
101^2	2.37E-03	3.31	2.37E-03	3.33	1.91E-03	4.01	1.97E-03	4.01	1.93E-03	4.02
201^2	7.85E-04	1.59	7.85E-04	1.59	1.36E-04	3.81	1.36E-04	3.85	1.35E-04	3.84
401^2	2.12E-04	1.89	2.12E-04	1.89	9.94E-06	3.77	1.03E-05	3.72	9.49E-06	3.83
801^2	5.40E-05	1.97	5.40E-05	1.97	7.70E-07	3.69	8.48E-07	3.60	6.62E-07	3.84
L^∞-norm										
I	RK(2,2;1)	rate	RK(2,2;$\frac{1}{2}$)	rate	RK(3,3;1)	rate	RK(3,3;$\frac{1}{3}$)	rate	RK(4,3;1)	rate
51^2	2.58E-02	–	2.61E-02	–	3.27E-02	–	3.33E-02	–	3.29E-02	–
101^2	1.32E-03	4.29	1.32E-03	4.30	7.82E-04	5.39	1.00E-03	5.05	8.02E-04	5.36
201^2	4.73E-04	1.48	4.73E-04	1.49	8.28E-05	3.24	1.09E-04	3.21	8.03E-05	3.32
401^2	1.26E-04	1.90	1.26E-04	1.90	9.44E-06	3.13	2.41E-05	2.17	9.33E-06	3.11
801^2	3.22E-05	1.97	3.22E-05	1.97	1.03E-06	3.19	6.46E-06	1.90	1.06E-06	3.13

We show in Table 5.2.2 the relative errors measured in the L^1-norm and in the L^∞-norm for the second-order methods RK(2,2;1), RK(2,2;$\frac{1}{2}$) and for the third-order methods RK(3,3;1), RK(3,3;$\frac{1}{3}$), RK(4,3;1); in all cases, we set CFL $= 0.2$. We observe that all the methods achieve the expected convergence order in the L^1-norm for this CFL number (recall that all the methods perform exactly the same number of flux evaluations). The observed rate is even larger than 3 for the third-order methods. The methods RK(2,2;1), RK(2,2;$\frac{1}{2}$), RK(3,3;1), and RK(4,3;1) also achieve the expected rate in the L^∞-norm over the entire range of meshes. The rates are suboptimal for the SSP method RK(3,3;$\frac{1}{3}$). This is due to its efficiency coefficient $\frac{1}{3}$ being significantly less than 1. Of course, optimality is recovered by lowering the CFL number (tests not shown). This test illustrates the importance of using RK methods with large efficiency coefficients.

Table 6: Linear transport, 2D $\mathbb{P}_1$ finite elements on uniform meshes. Relative error in the L^1- and L^∞-norms for the fourth- and fifth-order RK methods at CFL $= 0.2$.

L^1-norm										
I	RK(4,4;$\frac{1}{2}$)	rate	RK(4,4;$\frac{3}{4}$)	rate	RK(5,4;0.51)	rate	RK(6,5;$\frac{5}{6}$)	rate	RK(6,5;$\frac{2}{3}$)	rate
50^2	3.14E-02	–	3.22E-02	–	3.15E-02	–	3.17E-02	–	3.12E-02	–
100^2	1.94E-03	4.02	2.03E-03	3.99	1.94E-03	4.02	1.91E-03	4.05	1.92E-03	4.02
200^2	1.34E-04	3.85	1.38E-04	3.88	1.35E-04	3.84	1.50E-04	3.67	1.34E-04	3.84
400^2	9.53E-06	3.82	9.71E-06	3.83	9.48E-06	3.83	1.30E-05	3.53	9.45E-06	3.83
800^2	6.82E-07	3.80	7.13E-07	3.77	6.62E-07	3.84	1.24E-06	3.39	6.63E-07	3.83
L^∞-norm										
I	RK(4,4;$\frac{1}{2}$)	rate	RK(4,4;$\frac{3}{4}$)	rate	RK(5,4;0.51)	rate	RK(6,5;$\frac{5}{6}$)	rate	RK(6,5;$\frac{2}{3}$)	rate
50^2	3.31E-02	–	3.28E-02	–	3.32E-02	–	4.03E-02	–	3.30E-02	–
100^2	8.75E-04	5.24	9.47E-04	5.11	8.69E-04	5.26	2.13E-03	4.24	8.04E-04	5.36
200^2	8.00E-05	3.45	8.95E-05	3.40	8.03E-05	3.44	3.87E-04	2.46	8.03E-05	3.32
400^2	9.32E-06	3.10	1.55E-05	2.53	9.34E-06	3.10	8.97E-05	2.11	9.33E-06	3.11
800^2	2.15E-06	2.12	4.07E-06	1.93	1.03E-06	3.18	2.71E-05	1.73	1.04E-06	3.17

We show in Table 5.2.2 the relative errors measured in the L^1-norm and in the L^∞-norm for the fourth-order methods RK(4,4;$\frac{1}{2}$), RK(4,4;$\frac{3}{4}$), RK(5,4;0.51) and for the fifth-order methods

RK(6,5;$\frac{5}{6}$), RK(6,5;$\frac{2}{3}$); as above, we set CFL = 0.2. We observe that all the methods achieve the same convergence order (about 3.8) in the L^1-norm, which is above the expected convergence order 3. The convergence rate in the L^∞-norm of the methods RK(5,4;0.51) and RK(6,5;$\frac{2}{3}$) is 3, as expected for this CFL number. The convergence rate is suboptimal for the other methods. Here again, optimality is recovered by lowering the CFL number (tests not shown). The slight suboptimality of the method RK(6,5;$\frac{5}{6}$) is probably due to the presence of very large coefficients in the 6th row of its Butcher tableau, which likely induce a large truncation error. Notice that the method RK(4,3;1) performs significantly better than the other two fourth-order methods RK(4,4;$\frac{1}{2}$), RK(4,4;$\frac{3}{4}$) in the present case.

5.3 Linear transport with non-smooth solutions

Here, we solve a standard linear transport problem with non-smooth initial data: $\partial_t u + \nabla\cdot(\boldsymbol{\beta} u) = 0$ in $D := \{\boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x}\|_{\ell^2} < 1\}$ with $\boldsymbol{\beta}(\boldsymbol{x}) := 2\pi(-x_2, x_1)^\intercal$. The initial data is

$$u_0(\boldsymbol{x}) := \begin{cases} 1 & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_d\|_{\ell^2} \le r_0 \text{ and } (|x_1| \ge 0.05 \text{ or } x_2 \ge 0.7), \\ 1 - \frac{\|\boldsymbol{x}-\boldsymbol{x}_c\|_{\ell^2}}{r_0} & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_c\|_{\ell^2} \le r_0, \\ g(\|\boldsymbol{x}-\boldsymbol{x}_h\|_{\ell^2}) & \text{if } \|\boldsymbol{x}-\boldsymbol{x}_h\|_{\ell^2} \le r_0, \\ 0 & \text{otherwise,} \end{cases} \tag{53}$$

where $r_0 := 0.3$, $g(r) := \frac{1}{4}\big[1 + \cos\big(\pi\frac{r}{r_0}\big)\big]$, $\boldsymbol{x}_d := (0, 0.5)$, $\boldsymbol{x}_c := (0, -0.5)$, $\boldsymbol{x}_h := (-0.5, 0)$. This test has been proposed in Leveque [26], Zalesak [30]. 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 non-nested 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 shun in the literature for "its lack of stability", performs actually very well when used with the IDP technique proposed in this paper. We show in Figure 1 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. We observe that the method behaves as expected and that the results are visually of the same quality as what is usually reported in the literature.

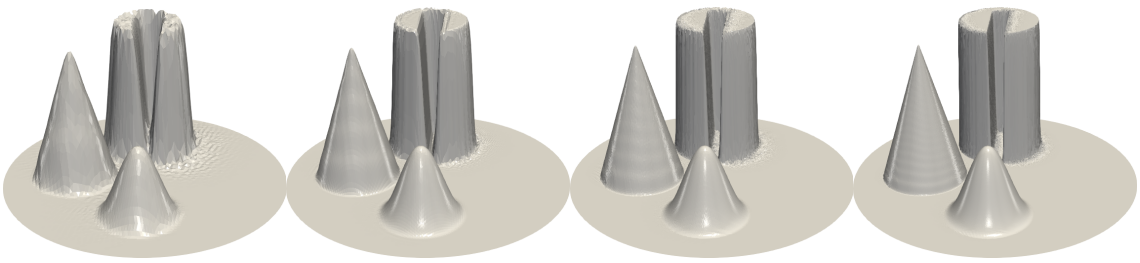

Figure 1: Three-solids problem at $T = 1$, using RK(2,2;1) at CFL = 0.25. 2D $\mathbb{P}_1$ finite elements on unstructured meshes. From left to right: $I = 6561$; $I = 24917$; $I = 98648$; $I = 389860$.

We show in Table 5.3 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 [12, §6.3].

Table 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 methods 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

5.4 Burgers' equation

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

$$\partial_t u + \nabla\cdot(\boldsymbol{f}(u)) = 0, \qquad \boldsymbol{f}(u) := \tfrac{1}{2}(u^2, u^2)^\mathsf{T}, \qquad u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) \text{ a.e. } \boldsymbol{x} \in D, \tag{54}$$

with the initial data

$$u_0(\boldsymbol{x}) := \begin{cases} 1 & \text{if } |x_1 - \frac{1}{2}| \le 1 \text{ and } |x_2 - \frac{1}{2}| \le 1 \\ -a & \text{otherwise.} \end{cases} \tag{55}$$

This problem is considered in [12, §6.1]. The exact solution is given in Eq. (52)-(53) therein. This problem is interesting since it exhibits many sonic points, which makes methods that are too greedy to fail.

We approximate the solution to this problem with continuous $\mathbb{P}_1$ finite elements on uniform triangular meshes. We test the ten RK methods considered above. The computations are done up to $T := 0.65$ with CFL $= 0.25$. We compute the relative error in the L^1-norm on five consecutively refined meshes. The results are reported in Table 5.4. We observe that all the convergence rates are close to 0.9. The rates are similar to those is reported in [12] where the time stepping is done with SSPRK(3,3;$\frac{1}{3}$) (the CFL used therein is CFL $= \frac{0.25}{3}$).

Table 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(2,2;1)	rate	RK(2,2;$\frac{1}{2}$)	rate	RK(3,3;1)	rate	RK(3,3;$\frac{1}{3}$)	rate	RK(4,3;1)	rate
51^2	7.71E-02	–	7.79E-02	–	7.71E-02	–	8.03E-02	–	7.71E-02	–
101^2	3.69E-02	1.06	3.73E-02	1.06	3.69E-02	1.06	3.85E-02	1.06	3.69E-02	1.06
201^2	2.30E-02	0.68	2.32E-02	0.68	2.30E-02	0.68	2.38E-02	0.70	2.30E-02	0.68
401^2	1.24E-02	0.90	1.24E-02	0.90	1.24E-02	0.90	1.27E-02	0.90	1.24E-02	0.90
801^2	6.47E-03	0.93	6.52E-03	0.93	6.48E-03	0.93	6.65E-03	0.93	6.47E-03	0.93

I	RK(4,4;$\frac{1}{2}$)	rate	RK(4,4;$\frac{3}{4}$)	rate	RK(5,4;0.51)	rate	RK(6,5;$\frac{5}{6}$)	rate	RK(6,5;$\frac{2}{3}$)	rate
51^2	7.94E-02	–	8.15E-02	–	7.79E-02	–	1.81E-01	–	9.29E-02	–
101^2	3.80E-02	1.06	3.89E-02	1.07	3.89E-02	1.00	8.56E-02	1.08	4.39E-02	1.08
201^2	2.36E-02	0.69	2.40E-02	0.70	2.47E-02	0.66	4.78E-02	0.84	2.72E-02	0.69
401^2	1.26E-02	0.90	1.28E-02	0.90	1.36E-02	0.86	2.38E-02	1.00	1.41E-02	0.95
801^2	6.61E-03	0.93	6.72E-03	0.94	7.11E-03	0.93	1.22E-02	0.97	7.24E-03	0.96

Acknowledgments

This material is based upon work supported in part by the National Science Foundation via grants DMS2110868; the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-18-1-0397; and by the Army Research Office under grant/contract number W911NF-15-1-0517. The support of INIRIA through the International Chair program is acknowledged.

References

[1] R. Abgrall. Residual distribution schemes: current status and future trends. *Comput. & Fluids*, 35(7):641–669, 2006.

[2] R. Abgrall, Q. Viville, H. Beaugendre, and C. Dobrzynski. Construction of a p-adaptive continuous residual distribution scheme. *J. Sci. Comput.*, 72(3):1232–1268, 2017.

[3] J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. **11** (1973), no. 1, 38–69]. *J. Comput. Phys.*, 135(2):170–186, 1997.

[4] J. C. Butcher. On Runge-Kutta processes of high order. *J. Austral. Math. Soc.*, 4:179–194, 1964. ISSN 0263-6115.

[5] 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(3):275–326, 1990.

[6] A. Ern and J.-L. Guermond. *Finite elements. III. First-order and time-dependent PDEs*, volume 74 of *Texts in Applied Mathematics*. Springer, Cham, 2021.

[7] E. Fehlberg. Klassische Runge-Kutta-Formeln vierter und niedrigerer Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme. *Computing (Arch. Elektron. Rechnen)*, 6:61–71, 1970.

[8] L. Ferracina and M. N. Spijker. An extension and analysis of the Shu-Osher representation of Runge-Kutta methods. *Math. Comp.*, 74(249):201–219, 2005.

[9] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001.

[10] J.-L. Guermond and R. Pasquetti. A correction technique for the dispersive effects of mass lumping for transport problems. *Comput. Methods Appl. Mech. Engrg.*, 253:186–198, 2013.

[11] J.-L. Guermond and B. Popov. Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM J. Numer. Analysis*, 54(4):2466–2489, 2016.

[12] J.-L. Guermond and B. Popov. Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. Numer. Anal.*, 55(6):3120–3146, 2017.

[13] 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(4):2163–2182, 2014.

[14] 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(3):1358–1366, 2017.

[15] 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(5):A3211–A3239, 2018. ISSN 1064-8275.

[16] 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.*, 347:143–175, 2019.

[17] 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:Paper No. 113608, 17, 2021.

[18] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49 (3):357–393, 1983.

[19] A. Harten and S. Osher. Uniformly high-order accurate nonoscillatory schemes. I. *SIAM J. Numer. Anal.*, 24(2):279–309, 1987.

[20] I. Higueras. Representations of Runge-Kutta methods and strong stability preserving methods. *SIAM J. Numer. Anal.*, 43(3):924–948, 2005.

[21] B. Khobalatte and B. Perthame. Maximum principle on the entropy and second-order kinetic schemes. *Math. Comp.*, 62(205):119–131, 1994.

[22] J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta methods. *BIT*, 31(3):482–528, 1991.

[23] D. Kuzmin and S. Turek. Flux correction tools for finite elements. *Journal of Computational Physics*, 175(2):525–558, 2002.

[24] D. Kuzmin, R. Löhner, and S. Turek. *Flux–Corrected Transport*. Scientific Computation. Springer, 2005. 3-540-23730-5.

[25] D. Kuzmin, R. Löhner, and S. Turek. *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Scientific Computation. Springer, 2012. ISBN 9789400740372.

[26] R. J. Leveque. High-resolution conservative algorithms for advection in incompressible flow. *SIAM J. Numer. Anal.*, 33(2):627–665, 1996.

[27] W. Pazner. Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting. *Comput. Methods Appl. Mech. Engrg.*, 382:Paper No. 113876, 28, 2021.

[28] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439 – 471, 1988.

[29] 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.*, 303:229–248, 2016.

[30] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, 31(3):335–362, 1979.

[31] X. Zhang and C.-W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.*, 229(9):3091–3120, 2010.

[32] X. Zhang and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 467(2134):2752–2776, 2011.