

Locally conservative and flux consistent iterative methods

Viktor Linders¹, Philipp Birken¹

¹*Centre for mathematical sciences, Lund University, Lund, Sweden.
email: viktor.linders@math.lu.se
philipp.birken@na.lu.se*

Abstract

Conservation and consistency are fundamental properties of discretizations of systems of hyperbolic conservation laws. Here, these concepts are extended to the realm of iterative methods by formally defining *locally conservative* and *flux consistent* iterations. These concepts are of both theoretical and practical importance: Based on recent work by the authors, it is shown that pseudo-time iterations using explicit Runge-Kutta methods are locally conservative but not necessarily flux consistent. An extension of the Lax-Wendroff theorem is presented, revealing convergence towards weak solutions of a temporally retarded system of conservation laws. Each equation is modified in the same way, namely by a particular scalar factor multiplying the spatial flux terms. A technique for enforcing flux consistency, and thereby recovering convergence, is presented. Further, local conservation is established for all Krylov subspace methods, with and without restarts, and for Newton's method under certain assumptions on the discretization. Thus it is shown that Newton-Krylov methods are locally conservative, although not necessarily flux consistent. Numerical experiments with the 2D compressible Euler equations corroborate the theoretical results. Further numerical investigations of the impact of flux consistency on Newton-Krylov methods indicate that its effect is case dependent, and diminishes as the number of iterations grow.

Keywords: Iterative methods, Conservation laws, Lax-Wendroff theorem, Pseudo-time iterations, Newton-Krylov methods

1 Introduction

Conservation laws arise ubiquitously in the modelling of physical phenomena and their discretizations remain the subject of intense research. Fundamen-

tal properties of successful schemes include conservation, consistency and convergence. These concepts are well defined for both space and time discretizations; explicit and implicit.

Implicit discretizations typically result in a large, sparse systems of nonlinear equations to be solved in each time step. The solution is usually approximated through the application of iterative methods; see e.g. [2, Chapters 5 & 6]. Yet, discussions about conservation, consistency and convergence (in the sense of grid refinement) for schemes involving iterative methods are rare. In [9, 1], studies were conducted of particular implicit finite volume schemes applied to the steady Euler equations, solved using a variety of modified Newton-type methods. The results indicate that the choice of iterative method has a significant impact on the convergence of the scheme. Based on these results, a study of similar schemes applied to the Reynolds-Averaged Navier-Stokes (RANS) equations were carried out in [10], where it was found that the less performant methods violate mass conservation.

In [4], the authors considered general finite volume discretizations of 1D scalar conservation laws, discretized in time with the implicit Euler method. Global (i.e. mass) conservation was proven for many methods, including pseudo-time iterations, Krylov subspace methods, Newton's method and certain multigrid techniques. On the other hand, the Jacobi and Gauss-Seidel iterations were shown to violate mass conservation in general, corroborating the observations in [10]. A stronger notion is that of *local conservation*, defined formally below, which loosely means that mass is not only conserved but also not teleported. This notion is important for physical correctness, and allows to prove extensions of the Lax-Wendroff theorem [12], thus giving a much stronger mathematical backing of such nonlinear schemes. In [4], a start was made in this vein for pseudo-time iterations. It was shown that in case of convergence, the resulting scheme converges to a solution of a conservation law, where the flux is multiplied by a scheme dependent factor, unless particular care is taken. We say that the iterative method lacks *flux consistence*, which manifests as a temporal retardation.

Throughout this article, we work with systems of conservation laws. After introducing relevant notation and terminology, we formally define *locally conservative* and *flux consistent* iterative methods in 2. We extend the results on pseudotime iterations from [4] in 3, while considering a large class of implicit Runge-Kutta (RK) methods in place of Euler's method. As it turns out, even for systems flux inconsistency manifests through a scheme dependent scalar factor. Thus, a method to enforce flux consistency, first introduced in [4], applies here too.

The second focus of this work is local conservation of the important class of Newton-Krylov methods. In 4, we first prove that Newton's method is both locally conservative and flux consistent under certain assumptions on the spatial discretization and when solving all linear systems exactly. We

can thus establish that if there are problems with conservation within an implicit solver using Newton’s method, they stem from the iterative solver for the linear systems. Secondly, by relating Krylov subspace methods to pseudo-time iterations, local conservation is shown also for these in 5. This subsequently leads to a proof of local conservation for Newton-Krylov methods.

Numerical examples corroborate the theoretical findings in 6. We further explore the impact of flux consistency on Newton-Krylov methods by applying the aforementioned technique for ensuring flux consistency of pseudo-time iterations. The results indicate that role of flux consistency is case dependent, and that its effect diminishes as the number of iterations grow.

2 Preliminaries

This section introduces relevant notation and the theoretical background upon which the remaining paper rests.

2.1 Notation

Scalar quantities are denoted by letters in normal font. Vectors and matrices are bold, with vectors being lower case and matrices upper case. Vectors and matrices of several different dimensions are treated in the manuscript:

- We consider systems of m conservation laws. Vectors in \mathbb{R}^m are written with an underline e.g. $\underline{\mathbf{u}}$.
- We consider s -stage explicit Runge-Kutta methods. Vectors in \mathbb{R}^s are written with a right-pointing arrow, e.g. $\vec{\mathbf{b}}$, and similarly for matrices in $\mathbb{R}^{s \times s}$.
- We consider \tilde{s} -stage implicit Runge-Kutta methods. Vectors in $\mathbb{R}^{\tilde{s}}$ are written with a left-pointing arrow e.g. $\overleftarrow{\mathbf{b}}$, and similarly for matrices in $\mathbb{R}^{\tilde{s} \times \tilde{s}}$.
- Vectors whose dimension is the product of the dimensions listed above are written with a combination of attributes, e.g. $\overleftrightarrow{\mathbf{x}} \in \mathbb{R}^{\tilde{s}s}$, $\overrightarrow{\mathbf{y}} \in \mathbb{R}^{sm}$, $\overleftarrow{\mathbf{z}} \in \mathbb{R}^{\tilde{s}sm}$.
- Vectors representing quantities on spatial grids are expressed in a bold sans serif font, e.g.

$$\mathbf{u}^\top = (\dots, u_{i-1}, u_i, u_{i+1}, \dots).$$

This notation is combined with the accents above if the evaluated quantity in question is a vector. Thus, a vector of m -element vectors

is represented as

$$\underline{\mathbf{u}}^\top = (\dots, \underline{\mathbf{u}}_{i-1}^\top, \underline{\mathbf{u}}_i^\top, \underline{\mathbf{u}}_{i+1}^\top, \dots).$$

Matrices operating on these vectors are denoted similarly with capital letters and are constructed block-diagonally:

$$\underline{\mathbf{A}} = \text{blkdiag}(\dots, \underline{\mathbf{A}}_{i-1}, \underline{\mathbf{A}}_i, \underline{\mathbf{A}}_{i+1}, \dots).$$

- A flux function that takes $(p+q+1)$ arguments, e.g. $\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}_{i-p}, \dots, \underline{\mathbf{u}}_{i+q})$, is sometimes denoted with the abbreviated argument $\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}})$.

2.2 Conservation laws and the Lax-Wendroff theorem

Consider the system of 1D conservation laws

$$\underline{\mathbf{u}}_t + \underline{\mathbf{f}}_x = \underline{\mathbf{0}}, \quad \underline{\mathbf{u}}(x, 0) = \underline{\mathbf{u}}_0(x), \quad x \in \Omega, \quad t > 0. \quad (1)$$

Here, $\underline{\mathbf{u}}(x, t)$, $\underline{\mathbf{f}}$, $\underline{\mathbf{u}}_0 \in \mathbb{R}^m$. Throughout, it is assumed that (1) is posed either as a Cauchy problem or on a periodic domain. Under these circumstances, the quantity $\int_\Omega \underline{\mathbf{u}} dx$ is conserved. A space-time discretization of (1) that discretely mimics this property is said to be *globally conservative*.

We consider discretizations of (1) that may be expressed in the form

$$\frac{\underline{\mathbf{u}}_i^{n+1} - \underline{\mathbf{u}}_i^n}{\Delta t} + \frac{1}{\Delta x} \left(\underline{\mathbf{f}}_{i+\frac{1}{2}} - \underline{\mathbf{f}}_{i-\frac{1}{2}} \right) = \underline{\mathbf{0}}, \quad i = \dots, -1, 0, 1, \dots \quad (2)$$

Here, $\underline{\mathbf{u}}_i^n \approx \underline{\mathbf{u}}(x_i, t_n) \equiv \underline{\mathbf{u}}(i\Delta x, n\Delta t)$. In this paper we consider implicit discretizations and therefore restrict our attention to numerical fluxes of the type

$$\underline{\mathbf{f}}_{i+\frac{1}{2}} \equiv \underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}^{n+1}) = \underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}_{i-p}^{n+1}, \dots, \underline{\mathbf{u}}_{i+q}^{n+1}),$$

where p and q are nonnegative integers with $p+q > 0$. Throughout, it is assumed that these numerical fluxes are consistent:

Definition 1. *The numerical flux $\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}) = \underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}_{i-p}, \dots, \underline{\mathbf{u}}_{i+q})$ is said to be consistent with $\underline{\mathbf{f}}(\underline{\mathbf{u}})$ if it is Lipschitz continuous in each argument and if $\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}, \dots, \underline{\mathbf{u}}) = \underline{\mathbf{f}}(\underline{\mathbf{u}})$.*

The concept of local conservation will be central in the remainder. It applies to explicit and implicit discretizations alike:

Definition 2. *A discretization of (1) that can be expressed in the form (2) is said to be locally conservative.*

Local conservation is a useful property for both physical and mathematical reasons (1). It enforces that solution components leaving one computational cell necessarily enter the neighbouring one. Conservation of "total mass", $\sum_i \Delta x \underline{\mathbf{u}}_i^n$ is thereby ensured, in analogy with the continuous problem (i.e. global conservation). Further, it is an essential ingredient in the ubiquitous Lax-Wendroff theorem.

The Lax-Wendroff theorem applies to the Cauchy problem for (1) and considers locally conservative discretizations with consistent numerical flux. If the numerical solution of such a scheme converges to a function $\underline{\mathbf{u}}$ in the limit of vanishing Δx and Δt , the theorem provides sufficient conditions for $\underline{\mathbf{u}}$ to be a weak solution of the conservation law (1) [13, Chapter 12]. More precisely, consider a sequence of grids $(\Delta x_\ell, \Delta t_\ell)$ such that $\Delta x_\ell, \Delta t_\ell \rightarrow 0$ as $\ell \rightarrow \infty$. Let $\underline{\mathbf{u}}_\ell(x, t)$ denote the piecewise constant function that takes the solution value $\underline{\mathbf{u}}_i^n$ in $(x_i, x_{i+1}] \times (t_{n-1}, t_n]$ on the ℓ th grid. We make the following assumptions:

Assumption 1.

1. There is a function $\underline{\mathbf{u}}(x, t)$ such that over every bounded set $\Omega = [a, b] \times [0, T]$ in x - t space,

$$\|\underline{\mathbf{u}}_\ell(x, t) - \underline{\mathbf{u}}(x, t)\|_{1, \Omega} \rightarrow 0 \quad \text{as } \ell \rightarrow \infty.$$

2. For each $T \geq 0$ there is a constant $R > 0$ such that the total variation

$$TV(\underline{\mathbf{u}}_\ell(\cdot, t)) < R \quad \text{for all } 0 \leq t \leq T, \quad \ell = 1, 2, \dots$$

The Lax-Wendroff theorem can then be stated as follows:

Theorem 1. Consider a sequence of grids $(\Delta x_\ell, \Delta t_\ell)$ such that $\Delta x_\ell, \Delta t_\ell \rightarrow 0$ as $\ell \rightarrow \infty$. Suppose that the numerical flux $\underline{\mathbf{f}}_{i \pm \frac{1}{2}}$ in (2) is consistent with $\underline{\mathbf{f}}$ and that 1 is satisfied. Then, $\underline{\mathbf{u}}(x, t)$ is a weak solution of (1).

1 is not strictly speaking necessary for the Lax-Wendroff theorem. Both conditions can be relaxed somewhat. Indeed, the original proof due to Lax and Wendroff instead assumes that $\underline{\mathbf{u}}_\ell$ converges boundedly almost everywhere to $\underline{\mathbf{u}}$ [12].

The discretization (2) is implicit, hence the solution generally must be approximated using iterative methods. Since we are interested in the convergence properties of schemes involving iterative methods, the concepts of local conservation and consistency must be extended to this setting. To this end, we propose the following definition:

Definition 3. Suppose that the solution to (2) is approximated by a sequence of iterates $\underline{\mathbf{u}}_i^{(k)}$, $k = 0, \dots, N$ and set $\underline{\mathbf{u}}^{n+1} = \underline{\mathbf{u}}_i^{(N)}$. If there is a

numerical flux function $\underline{\mathbf{h}}_{i+\frac{1}{2}}^{(N)}$ such that the approximate numerical solution $\underline{\mathbf{u}}^{n+1}$ satisfies

$$\frac{\underline{\mathbf{u}}_i^{n+1} - \underline{\mathbf{u}}_i^n}{\Delta t} + \frac{1}{\Delta x} \left(\underline{\mathbf{h}}_{i+\frac{1}{2}}^{(N)} - \underline{\mathbf{h}}_{i-\frac{1}{2}}^{(N)} \right) = \underline{\mathbf{0}}, \quad i = \dots, -1, 0, 1, \dots \quad (3)$$

then the iterative method is said to be locally conservative. If $\underline{\mathbf{h}}_{i+\frac{1}{2}}^{(N)}$ further satisfies 1, then the iterative method is said to be flux consistent.

3 Conservation and consistency of pseudo-time iterations

In order to approximate the solution of the nonlinear system (2) using pseudo-time iterations, we introduce a pseudo-time derivative,

$$\frac{\partial \underline{\mathbf{u}}_i}{\partial \tau} + \underline{\mathbf{g}}_i(\underline{\mathbf{u}}) = 0, \quad \underline{\mathbf{u}}_i(0) = \underline{\mathbf{u}}_{0_i}, \quad i = \dots, -1, 0, 1, \dots$$

where the nonlinear function $\underline{\mathbf{g}}_i$ is given by

$$\underline{\mathbf{g}}_i(\underline{\mathbf{u}}) = \frac{\underline{\mathbf{u}}_i - \underline{\mathbf{u}}_i^n}{\Delta t} + \frac{1}{\Delta x} \left(\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}) - \underline{\mathbf{f}}_{i-\frac{1}{2}}(\underline{\mathbf{u}}) \right). \quad (4)$$

Several different methods are available for iterating in pseudo-time [19, 3]. Herein, we use an s -stage explicit Runge-Kutta (ERK) method. Let $(\vec{\mathbf{A}}, \vec{\mathbf{b}}, \vec{\mathbf{c}})$ denote the coefficient matrix and vectors of the ERK method. We denote the k th pseudo-time iterate by $\underline{\mathbf{u}}_i^{(k)}$. The subsequent iterate $\underline{\mathbf{u}}_i^{(k+1)}$ is computed from $\underline{\mathbf{u}}_i^{(k)}$ as

$$\underline{\mathbf{u}}_i^{(k+1)} = \underline{\mathbf{u}}_i^{(k)} - \Delta \tau_k \sum_{j=1}^s b_j \underline{\mathbf{g}}_i \left(\underline{\mathbf{u}}_j^{(k)} \right), \quad i = \dots, -1, 0, 1, \dots \quad (5)$$

where the stage vectors $\underline{\mathbf{u}}_j^{(k)}$, $j = 1, \dots, s$ have elements

$$\underline{\mathbf{u}}_{j_\iota}^{(k)} = \underline{\mathbf{u}}_\iota^{(k)} - \Delta \tau_k \sum_{l=1}^{j-1} a_{j,l} \underline{\mathbf{g}}_\iota \left(\underline{\mathbf{u}}_l^{(k)} \right), \quad \iota = i - p, \dots, i + q. \quad (6)$$

As previously, p and q determine the bandwidth of the finite volume stencil.

3.1 Systems of conservation laws

In [4] the scalar version (i.e. $m = 1$) of (1) was considered. There, it was shown that the scheme (5)–(6) preserves the local conservation of the space-time discretization. Further, an extension of the Lax-Wendroff theorem was provided that incorporates a fixed number N of pseudo-time iterations.

Here, we present two theorems that generalize the results in [4] to systems of conservation laws (i.e. $m \geq 1$). We define a step in physical time by setting $\underline{\mathbf{u}}_i^{n+1} = \underline{\mathbf{u}}_i^{(N)}$. Throughout, $\underline{\mathbf{u}}_i^{(0)} = \underline{\mathbf{u}}_i^n$ is chosen as initial guess.

Recall that the *stability function* $\phi(z)$ of an RK method $(\vec{\mathbf{A}}, \vec{\mathbf{b}}, \vec{\mathbf{c}})$ is given by

$$\phi(z) = 1 + z\vec{\mathbf{b}}^\top (\vec{\mathbf{I}} - z\vec{\mathbf{A}})^{-1}\vec{\mathbf{1}}, \quad (7)$$

where $\vec{\mathbf{I}}$ is the $s \times s$ identity matrix and $\vec{\mathbf{1}} \in \mathbb{R}^s$ is the vector of all ones; see e.g. [21, Chapter IV.3]. The *stability region* of the RK method is defined as the subset of the complex plane for which $|\phi(z)| < 1$.

The proofs of the following theorems are very similar to those presented for the scalar case in [4]. The details are therefore omitted. The following is a generalization to systems of conservation laws, which also allows each pseudo-time step to be taken with different ERK methods.

Theorem 2. *Choose the initial guess $\underline{\mathbf{u}}_i^{(0)} = \underline{\mathbf{u}}_i^n$. Apply N pseudo-time iterations to (2), where the k th iteration is performed with an ERK method $(\vec{\mathbf{A}}_k, \vec{\mathbf{b}}_k, \vec{\mathbf{c}}_k)$ with stability function $\phi_k(z)$ and pseudo-time step $\Delta\tau_k$. Let $\mu_k = \Delta\tau_k/\Delta t$ for $k = 0, \dots, N-1$. The pseudo-time iterations are locally conservative with numerical flux*

$$\underline{\mathbf{h}}_{i+\frac{1}{2}}^{(N)} = \sum_{k=0}^{N-1} \left(\mu_k \vec{\mathbf{b}}_k^\top (\vec{\mathbf{I}} + \mu_k \vec{\mathbf{A}}_k)^{-1} \otimes \underline{\mathbf{I}} \right) \left(\prod_{l=k+1}^{N-1} \phi_l(-\mu_l) \right) \vec{\mathbf{f}}_{i+\frac{1}{2}}^{(k)}. \quad (8)$$

Here, $\underline{\mathbf{I}}$ is the $m \times m$ identity matrix and

$$\vec{\mathbf{f}}_{i+\frac{1}{2}}^{(k)} = \left(\underline{\mathbf{f}}_{i+\frac{1}{2}}^\top \left(\underline{\mathbf{u}}_1^{(k)} \right), \dots, \underline{\mathbf{f}}_{i+\frac{1}{2}}^\top \left(\underline{\mathbf{u}}_s^{(k)} \right) \right)^\top \in \mathbb{R}^{sm}.$$

The numerical flux is consistent with $c\underline{\mathbf{f}}(u)$, where

$$c \equiv c(\mu_0, \dots, \mu_{N-1}) = 1 - \prod_{l=0}^{N-1} \phi_l(-\mu_l). \quad (9)$$

Thus, pseudo-time iterations are flux consistent if and only if $c = 1$.

Remark 1. *The product in (8) is empty when $k = N - 1$. To handle this case we use the convention*

$$\prod_{l=N}^{N-1} \phi_l(-\mu_l) = 1.$$

Proof. The proof is step by step the same as those of [4, Lemma 3 & Theorem 2]. The only changes necessary are to replace $\vec{\mathbf{A}}$ by $\vec{\mathbf{A}} \otimes \underline{\mathbf{I}}$, and to swap each multiplication of the form $\vec{\mathbf{1}}\phi$, where ϕ is a scalar, to a corresponding Kronecker product $\vec{\mathbf{1}} \otimes \underline{\phi}$. \square

2 reveals that pseudo-time iterations are locally conservative but not necessarily flux consistent, as characterized by the scalar c in (9). To remove the inconsistency, it suffices to make a single iteration with any explicit RK method for which $\phi_l(-\mu_l) = 0$. In [4] it was suggested to iterate once with the explicit Euler method, choosing $\Delta\tau_0 = \Delta t$ so that $\mu_0 = 1$. The stability function is $\phi(-\mu_0) = 1 - \mu_0$, hence $\phi(-1) = 0$. This rids the iterative method of its flux inconsistency and the remaining iterations can be made with any other ERK method as preferred. Numerical experiments in [4] showed that enforcing flux consistency can have a profound impact on the convergence of the pseudo-time iterations.

The coefficient c causes the numerical fluxes in (8) to be consistent with the modified system of conservation laws

$$\underline{\mathbf{u}}_t + c(\mu_0, \dots, \mu_{N-1})\underline{\mathbf{f}}_x = 0. \quad (10)$$

Note that c affects all components of the system identically. An interpretation of its presence is that the pseudo-time iterations alter the rate of flow of time. Defining the modified time $t_c = ct$ it follows that $\underline{\mathbf{u}}_t = c\underline{\mathbf{u}}_{t_c}$ so that $\underline{\mathbf{u}}_{t_c} + \underline{\mathbf{f}}_x = \underline{\mathbf{0}}$. Hence, t_c rather than t is the governing time variable in (10).

Convergent pseudo-time iterations require that the pseudo-time steps are restricted to the stability domain of the explicit RK method, at least for most of the iterations. Thus, the product in (9) is generally taken over factors bounded by unity, and consequently $c \leq 1$. Thus, t_c represents a time retardation, i.e. time flows slower in (10) than in the original conservation law (1). This is in line with the experimental observations made in [4].

3.2 Higher order implicit Runge-Kutta methods

Let us return to the system of conservation laws (1) and introduce a spatial semi-discretization

$$\underline{\mathbf{u}}_t + \frac{1}{\Delta x} \left(\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}) - \underline{\mathbf{f}}_{i-\frac{1}{2}}(\underline{\mathbf{u}}) \right) = \underline{\mathbf{0}}. \quad (11)$$

If the implicit Euler method is used to discretize (11) in time, then (2) is recovered. However, suppose that we instead wish to discretize in time using an \tilde{s} -stage implicit Runge-Kutta (IRK) method with Butcher matrix and vectors $(\overleftarrow{\mathbf{A}}, \overleftarrow{\mathbf{b}}, \overleftarrow{\mathbf{c}})$. The resulting scheme can be expressed as

$$\begin{aligned}\underline{\mathbf{u}}_i^{(j)} &= \underline{\mathbf{u}}_i^n - \frac{\Delta t}{\Delta x} \sum_{l=1}^{\tilde{s}} \overleftarrow{a}_{jl} \left(\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}^{(j)}) - \underline{\mathbf{f}}_{i-\frac{1}{2}}(\underline{\mathbf{u}}^{(j)}) \right), \quad j = 1, \dots, \tilde{s}, \\ \underline{\mathbf{u}}_i^{n+1} &= \underline{\mathbf{u}}_i^n - \frac{\Delta t}{\Delta x} \sum_{j=1}^{\tilde{s}} \overleftarrow{b}_j \left(\underline{\mathbf{f}}_{i+\frac{1}{2}}(\underline{\mathbf{u}}^{(j)}) - \underline{\mathbf{f}}_{i-\frac{1}{2}}(\underline{\mathbf{u}}^{(j)}) \right).\end{aligned}\tag{12}$$

Define the quantities

$$\begin{aligned}\overleftarrow{\underline{\mathbf{U}}}_i &= \left(\underline{\mathbf{u}}_i^{(1),\top}, \dots, \underline{\mathbf{u}}_i^{(\tilde{s}),\top} \right)^\top \in \mathbb{R}^{\tilde{s}m}, \\ \overleftarrow{\underline{\mathbf{f}}}_{i+\frac{1}{2}}(\overleftarrow{\underline{\mathbf{U}}}) &= \left(\underline{\mathbf{f}}_{i+\frac{1}{2}}^\top(\underline{\mathbf{u}}^{(1)}), \dots, \underline{\mathbf{f}}_{i+\frac{1}{2}}^\top(\underline{\mathbf{u}}^{(\tilde{s})}) \right)^\top \in \mathbb{R}^{\tilde{s}m}.\end{aligned}$$

Then, (12) can be reformulated as

$$\begin{aligned}\overleftarrow{\underline{\mathbf{U}}}_i &= \overleftarrow{\mathbf{1}} \otimes \underline{\mathbf{u}}_i^n - \frac{\Delta t}{\Delta x} \left(\overleftarrow{\mathbf{A}} \otimes \underline{\mathbf{I}} \right) \left(\overleftarrow{\underline{\mathbf{f}}}_{i+\frac{1}{2}}(\overleftarrow{\underline{\mathbf{U}}}) - \overleftarrow{\underline{\mathbf{f}}}_{i-\frac{1}{2}}(\overleftarrow{\underline{\mathbf{U}}}) \right), \\ \underline{\mathbf{u}}_i^{n+1} &= \underline{\mathbf{u}}_i^n - \frac{\Delta t}{\Delta x} \left(\overleftarrow{\mathbf{b}}^\top \otimes \underline{\mathbf{I}} \right) \left(\overleftarrow{\underline{\mathbf{f}}}_{i+\frac{1}{2}}(\overleftarrow{\underline{\mathbf{U}}}) - \overleftarrow{\underline{\mathbf{f}}}_{i-\frac{1}{2}}(\overleftarrow{\underline{\mathbf{U}}}) \right),\end{aligned}\tag{13}$$

where $\overleftarrow{\mathbf{1}}$ is the vector of ones in $\mathbb{R}^{\tilde{s}}$.

We seek an approximation of the solution $\underline{\mathbf{u}}_i^{n+1}$. To this end we must find an approximate solution of the nonlinear equation system in the second line of (13). Rewriting this equation as

$$\frac{\overleftarrow{\underline{\mathbf{U}}}_i - \overleftarrow{\mathbf{1}} \otimes \underline{\mathbf{u}}_i^n}{\Delta t} + \frac{1}{\Delta x} \left(\overleftarrow{\mathbf{A}} \otimes \underline{\mathbf{I}} \right) \left(\overleftarrow{\underline{\mathbf{f}}}_{i+\frac{1}{2}}(\overleftarrow{\underline{\mathbf{U}}}) - \overleftarrow{\underline{\mathbf{f}}}_{i-\frac{1}{2}}(\overleftarrow{\underline{\mathbf{U}}}) \right) = \overleftarrow{\underline{\mathbf{0}}},\tag{14}$$

we see that it is of precisely the same form as the discretization (2), although with $\overleftarrow{\underline{\mathbf{U}}}_i$ taking the place of $\underline{\mathbf{u}}_i^{n+1}$, $\overleftarrow{\mathbf{1}} \otimes \underline{\mathbf{u}}_i^n$ replacing $\underline{\mathbf{u}}_i^n$ and $(\overleftarrow{\mathbf{A}} \otimes \underline{\mathbf{I}}) \overleftarrow{\underline{\mathbf{f}}}_{i+\frac{1}{2}}$ in place of $\underline{\mathbf{f}}_{i+\frac{1}{2}}$. Thus, if pseudo-time iterations are used to approximate a solution, then 2 applies and we can immediately conclude that the resulting scheme can be written in the conservative form

$$\frac{\overleftarrow{\underline{\mathbf{U}}}_i - \overleftarrow{\mathbf{1}} \otimes \underline{\mathbf{u}}_i^n}{\Delta t} + \frac{1}{\Delta x} \left(\overleftarrow{\underline{\mathbf{h}}}_{i+\frac{1}{2}}^{(N)} - \overleftarrow{\underline{\mathbf{h}}}_{i-\frac{1}{2}}^{(N)} \right) = \overleftarrow{\underline{\mathbf{0}}},\tag{15}$$

where the numerical flux is given by

$$\underline{\mathbf{h}}_{i+\frac{1}{2}}^{(N)} = \sum_{k=0}^{N-1} \left(\mu_k \vec{\mathbf{b}}^\top (\vec{\mathbf{I}} + \mu_k \vec{\mathbf{A}}) \otimes \overleftarrow{\mathbf{A}} \otimes \underline{\mathbf{I}} \right) \left(\prod_{l=k+1}^{N-1} \phi_l(\mu_l) \right) \underline{\mathbf{f}}_{i+\frac{1}{2}}^{(k)},$$

and

$$\underline{\mathbf{f}}_{i+\frac{1}{2}}^{(k)} = \left(\underline{\mathbf{f}}^\top \left(\underline{\mathbf{u}}_1^{(k)} \right), \dots, \underline{\mathbf{f}}^\top \left(\underline{\mathbf{u}}_s^{(k)} \right) \right)^\top.$$

At this point we note that if the scheme is such that a vector $\overleftarrow{\mathbf{v}}$ exists, satisfying

$$\overleftarrow{\mathbf{v}}^\top \overleftarrow{\mathbf{A}} = \overleftarrow{\mathbf{b}}^\top, \quad \overleftarrow{\mathbf{v}}^\top \overleftarrow{\mathbf{1}} = 1, \quad (16)$$

then the second line in (13) can be evaluated by left-multiplying the first line by $\overleftarrow{\mathbf{v}}^\top \otimes \underline{\mathbf{I}}$. In other words, it follows that

$$\underline{\mathbf{u}}_i^{n+1} = \left(\overleftarrow{\mathbf{v}}^\top \otimes \underline{\mathbf{I}} \right) \underline{\mathbf{U}}_i. \quad (17)$$

If we adopt this principle and use it to compute $\underline{\mathbf{u}}_i^{n+1}$ based on the approximation of $\underline{\mathbf{U}}_i$ obtained by applying pseudo-time iterations to (14), then the resulting scheme is conservative:

Theorem 3. *Apply N pseudo-time iterations to the stage equations (14) with the same assumptions as in 2. Suppose that a vector $\overleftarrow{\mathbf{v}}$ exists that satisfies conditions (16) and compute $\underline{\mathbf{u}}_i^{n+1}$ using (17). Then the pseudo-time iterations are locally conservative with the numerical flux*

$$\underline{\mathbf{h}}_{i+\frac{1}{2}}^{(N)} = \sum_{k=0}^{N-1} \left(\mu_k \vec{\mathbf{b}}^\top (\vec{\mathbf{I}} + \mu_k \vec{\mathbf{A}}) \otimes \overleftarrow{\mathbf{b}}^\top \otimes \underline{\mathbf{I}} \right) \left(\prod_{l=k+1}^{N-1} \phi_l(\mu_l) \right) \underline{\mathbf{f}}_{i+\frac{1}{2}}^{(k)}. \quad (18)$$

The numerical flux is consistent with $c(\mu_0, \dots, \mu_{N-1}) \underline{\mathbf{f}}$. Thus, the pseudo-time iterations are flux consistent if and only if $c = 1$.

Proof. Local conservation with the flux (18) follows from left-multiplying (14) by $(\overleftarrow{\mathbf{v}}^\top \otimes \underline{\mathbf{I}})$ and using (16) to conclude that $(\overleftarrow{\mathbf{v}}^\top \otimes \underline{\mathbf{I}}) (\overleftarrow{\mathbf{1}} \otimes \underline{\mathbf{u}}_i^n) = \underline{\mathbf{u}}_i^n$. Consistency with $c \underline{\mathbf{f}}$ follows from the fact that $\overleftarrow{\mathbf{b}}^\top \overleftarrow{\mathbf{1}} = 1$ for every consistent RK method. \square

It should be noted that we cannot find a vector $\overleftarrow{\mathbf{v}}$ that satisfies conditions (16) for all IRK methods. For example, the implicit midpoint rule is given by $(\overleftarrow{\mathbf{A}}, \overleftarrow{\mathbf{b}}, \overleftarrow{\mathbf{c}}) = (1, 1/2, 1/2)$. Thus, the first condition in (16) gives $\overleftarrow{\mathbf{v}} = 1/2$ whereas the second one gives $\overleftarrow{\mathbf{v}} = 1$, both of which cannot be satisfied. However, we remark that the important class of IRK methods

associated with Summation-By-Parts (SBP) methods all have such a $\widehat{\mathbf{v}}$ by construction [5, 14]. This is a broad class of methods with the ability to preserve L^2 -type estimates of the solution to systems of differential equations [15] and encompass the ubiquitous Radau IA and IIA and Lobatto IIC methods [16] as special cases. For further details about the theoretical and practical aspects of SBP methods for time marching, see [20] and the references therein.

A generalization of the Lax-Wendroff theorem can be obtained if we restrict ourselves to considering a fixed number of iterations on a sequence of ever finer grids:

Theorem 4. *Consider a sequence of grids $(\Delta x_\ell, \Delta t_\ell)$ such that $\Delta x_\ell, \Delta t_\ell \rightarrow 0$ as $\ell \rightarrow \infty$. Fix N independently of ℓ , set $\mathbf{u}_i^{(0)} = \mathbf{u}_i^n$. Apply N pseudo-time iterations to the conservative discretization (2), or to the stage equations (6) followed by the RK step (5). Let $\Delta\tau_{k,\ell}/\Delta t_\ell = \mu_{k,\ell} = \mu_k$ be constants independent of ℓ for each $k = 0, \dots, N-1$. Suppose that the numerical flux $\underline{\mathbf{f}}_{i\pm\frac{1}{2}}$ in (2) is consistent with $\underline{\mathbf{f}}$ and that 1 is satisfied. Then, $\mathbf{u}(x, t)$ is a weak solution of the conservation law (10).*

Proof. The proof is identical to that of [4, Theorem 3], with the same changes as those in the proof of 2. \square

4 Newton's method

We now return to the nonlinear system (2), or equivalently to the implicit Runge-Kutta stage equations in (13) and consider Newton's method, which replaces the nonlinear system with a sequence of linear ones. The solution to each linear system can then be approximated using pseudo-time iterations, or more commonly, using a Krylov subspace method.

In this section we limit our attention to the case when the numerical flux is bivariate, i.e. when $\underline{\mathbf{f}}_{i+\frac{1}{2}} = \underline{\mathbf{f}}(\mathbf{u}_i, \mathbf{u}_{i+1})$ for some function $\hat{\mathbf{f}}(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}})$. It will be convenient to split the flux into a symmetric (or *convective*) and an anti-symmetric (or *dissipative*) component;

$$\hat{\mathbf{f}} = \hat{\mathbf{f}}^{(+)} + \hat{\mathbf{f}}^{(-)}, \quad \hat{\mathbf{f}}^{(+)}(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}}) = \hat{\mathbf{f}}^{(+)}(\underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\theta}}), \quad \hat{\mathbf{f}}^{(-)}(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}}) = -\hat{\mathbf{f}}^{(-)}(\underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\theta}}).$$

Any bivariate function can be expressed in this way by setting

$$\hat{\mathbf{f}}^{(+)}(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}}) = \frac{1}{2}(\hat{\mathbf{f}}(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}}) + \hat{\mathbf{f}}(\underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\theta}})), \quad \hat{\mathbf{f}}^{(-)}(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}}) = \frac{1}{2}(\hat{\mathbf{f}}(\underline{\boldsymbol{\theta}}, \underline{\boldsymbol{\phi}}) - \hat{\mathbf{f}}(\underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\theta}})).$$

By anti-symmetry, the dissipative component satisfies $\hat{\mathbf{f}}^{(-)}(\mathbf{u}, \mathbf{u}) = \mathbf{0}$. Consistency of the numerical flux is therefore equivalent to consistency of the convective component.

In the following subsections, we demonstrate that Newton's method is locally conservative and flux consistent when bivariate fluxes are used. To simplify the presentation, we begin by proving these results for scalar conservation laws. The extension to systems is straightforward but notationally complicated. We therefore postpone this to a separate subsection.

4.1 Scalar conservation laws

Consider the scalar conservation law

$$u_t + f_x = 0,$$

posed on a periodic spatial domain and adjoined with appropriate initial data. Let $\hat{f}^{(\pm)}(\theta, \phi)$ be a consistent bivariate numerical flux. Without loss of generality we may assume that the flux is either symmetric or anti-symmetric. The analysis will proceed in the same way in both cases and we may thereafter form linear combinations of such fluxes as we like. We discretize with a finite volume method,

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + \frac{\hat{f}^{(\pm)}(u_i^{n+1}, u_{i+1}^{n+1}) - \hat{f}^{(\pm)}(u_{i-1}^{n+1}, u_i^{n+1})}{\Delta x} = 0. \quad (19)$$

The discretization (19) may equivalently be expressed in vector form as

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \frac{1}{\Delta x} (\mathbf{Q}^{(\mp)} \circ \mathbf{F}^{(\pm)}(\mathbf{u}^{n+1})) \mathbf{1} = \mathbf{0}. \quad (20)$$

The dimensions of the vectors and matrices match the number of cells in the computational grid. Here, \circ denotes the Hadamard product. The elements of the matrix $\mathbf{F}^{(\pm)}(\mathbf{u})$ are given by $(\mathbf{F}^{(\pm)})_{ij} = \hat{f}^{(\pm)}(u_i, u_j)$ and $\mathbf{Q}^{(\mp)}$ is given by

$$\mathbf{Q}^{(\mp)} = \begin{bmatrix} 0 & 1 & \dots & \mp 1 \\ \mp 1 & 0 & 1 & \\ & \mp 1 & 0 & 1 \\ \vdots & & \ddots & 1 \\ 1 & & 0 & \mp 1 & 0 \end{bmatrix}.$$

We define the function $\mathbf{g}(\mathbf{v})$ as

$$\mathbf{g}(\mathbf{v}) := \frac{\mathbf{v} - \mathbf{u}^n}{\Delta t} + \frac{1}{\Delta x} (\mathbf{Q}^{(\mp)} \circ \mathbf{F}^{(\pm)}(\mathbf{v})) \mathbf{1}. \quad (21)$$

Newton's method applied to the nonlinear system (20) is then given by

$$\mathbf{g}'(\mathbf{v}^{(k)}) \Delta \mathbf{v} + \mathbf{g}(\mathbf{v}^{(k)}) = \mathbf{0}, \quad \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \Delta \mathbf{v}, \quad (22)$$

where \mathbf{g}' is the Jacobian of \mathbf{g} . An explicit expression for \mathbf{g}' is given in [6, Theorems 2.1 & 5.1]. Let $\hat{f}_\phi^{(\pm)} = \partial \hat{f}^{(\pm)} / \partial \phi$ and introduce the matrix $\mathbf{F}_\phi^{(\pm)}(\mathbf{v})$ with elements $(\mathbf{F}_\phi^{(\pm)})_{ij} = \hat{f}_\phi^{(\pm)}(v_i, v_j)$. Defining the matrix

$$\partial \mathbf{F}^{(\pm)}(\mathbf{v}) := \mathbf{Q}^{(\mp)} \circ \mathbf{F}_\phi^{(\pm)}(\mathbf{v}) - \text{diag}(\mathbf{1}^\top (\mathbf{Q}^{(\mp)} \circ \mathbf{F}_\phi^{(\pm)}(\mathbf{v}))),$$

the Jacobian is given by $\mathbf{g}'(\mathbf{v}^{(k)}) = \Delta t^{-1} \mathbf{I} + \Delta x^{-1} \partial \mathbf{F}^{(\pm)}(\mathbf{v}^{(k)})$. This is a tridiagonal matrix, which we may explicitly write as

$$\mathbf{g}'(\mathbf{v}^{(k)}) = \text{tri} \left(\mp \frac{\hat{f}_\phi^{(\pm)}(v_i^{(k)}, v_{i-1}^{(k)})}{\Delta x}, \frac{1}{\Delta t} - \frac{\hat{f}_\phi^{(\pm)}(v_{i-1}^{(k)}, v_i^{(k)}) \mp \hat{f}_\phi^{(\pm)}(v_{i+1}^{(k)}, v_i^{(k)})}{\Delta x}, \frac{\hat{f}_\phi^{(\pm)}(v_i^{(k)}, v_{i+1}^{(k)})}{\Delta x} \right)$$

Inserting this expression for $\mathbf{g}'(\mathbf{v}^{(k)})$ into (22) and collecting terms leads to a system of equations of the form

$$\frac{v_i^{(k+1)} - u_i^n}{\Delta t} + \frac{1}{\Delta x} \left(h_{i+\frac{1}{2}}^{(k+1)} - h_{i-\frac{1}{2}}^{(k+1)} \right) = 0, \quad i = \dots, -1, 0, 1, \dots, \quad (23)$$

where the numerical flux function $h_{i+\frac{1}{2}}^{(k+1)}$ is given by

$$h_{i+\frac{1}{2}}^{(k+1)} = \hat{f}_\phi^{(\pm)}(v_i^{(k)}, v_{i+1}^{(k)}) \Delta v_{i+1} \pm \hat{f}_\phi^{(\pm)}(v_{i+1}^{(k)}, v_i^{(k)}) \Delta v_i + \hat{f}^{(\pm)}(v_{i+1}^{(k)}, v_i^{(k)}). \quad (24)$$

Theorem 5. *Choose $v_i^{(0)} = u_i^n$ and suppose that $\mathbf{g}'(u\mathbf{1})$ is nonsingular for any non-zero scalar u . Then Newton's method, applied to the discretization (20), is locally conservative and flux consistent.*

Proof. Setting $u_i^{n+1} = v_i^{(k+1)}$ in (23) shows that Newton's method applied to (20) is locally conservative. To establish flux consistency, let $\mathbf{v}^{(k)}$ and \mathbf{u}^n both be given as $u\mathbf{1}$ for some nonzero scalar u and some k (e.g. $k = 0$). Observe that, by the consistency of $\hat{f}^{(\pm)}$, we have $(\mathbf{F}^{(+)}(u\mathbf{1}))_{ij} = f(u)$ and $(\mathbf{F}^{(-)}(u\mathbf{1}))_{ij} = 0$. Consequently, $\mathbf{Q}^{(+)} \circ \mathbf{F}^{(-)}(u\mathbf{1}) = \mathbf{0}$ and $\mathbf{Q}^{(-)} \circ \mathbf{F}^{(+)}(u\mathbf{1}) = f(u)\mathbf{Q}^{(-)}$. Since $\mathbf{Q}^{(-)}\mathbf{1} = \mathbf{0}$ it therefore follows from (21) that $\mathbf{g}(u\mathbf{1}) = \mathbf{0}$. Since $\mathbf{g}'(u\mathbf{1})$ is nonsingular by assumption, (22) implies that $\Delta \mathbf{v} = \mathbf{0}$. Thus, from (24) we have

$$h_{i+\frac{1}{2}}^{(k+1)} = \hat{f}_\phi^{(\pm)}(u, u) \cdot 0 \pm \hat{f}_\phi^{(\pm)}(u, u) \cdot 0 + \hat{f}^{(\pm)}(u, u) = \hat{f}^{(\pm)}(u, u),$$

the latter of which is consistent by assumption. Note that $h_{i+\frac{1}{2}}^{(k+1)}$ exclusively depends on the current iterate $v^{(k+1)}$ via Δv and is therefore a linear function. The numerical flux is thus Lipschitz continuous, hence consistent.

Finally, since any bivariate function can be written as a linear combination of symmetric and anti-symmetric components, the argument above extends by linearity to arbitrary bivariate numerical fluxes. \square

4.2 Systems of conservation laws

We now return to the system of conservation laws (1) and the discretization (2), which may correspond either to the implicit Euler method or to the stage equations of a higher order RK method. The analysis proceeds very similarly to the scalar case, although we are now dealing with a flux function $\underline{f} : \mathbb{R}^m \mapsto \mathbb{R}^m$. Suppose that the corresponding numerical flux is of the form

$$\hat{\underline{f}}^{(\pm)}(\underline{\theta}, \underline{\phi}) = (\hat{f}_1^{(\pm)}(\underline{\theta}, \underline{\phi}), \dots, \hat{f}_m^{(\pm)}(\underline{\theta}, \underline{\phi}))^\top,$$

where each $\hat{f}_\ell^{(\pm)}$, $\ell = 1, \dots, m$, is bivariate, consistent and either symmetric or anti-symmetric.

This time, Newton's method is given by

$$\underline{\mathbf{g}}'(\underline{\mathbf{v}}^{(k)})\Delta\underline{\mathbf{v}} + \underline{\mathbf{g}}(\underline{\mathbf{v}}^{(k)}) = \underline{\mathbf{0}}, \quad \underline{\mathbf{v}}^{(k+1)} = \underline{\mathbf{v}}^{(k)} + \Delta\underline{\mathbf{v}}, \quad (25)$$

where we define $\underline{\mathbf{g}}(\underline{\mathbf{v}})$ as

$$\begin{aligned} \underline{\mathbf{g}}(\underline{\mathbf{v}}) &:= \frac{\underline{\mathbf{v}} - \underline{\mathbf{u}}^n}{\Delta t} + \frac{1}{\Delta x} ((\underline{\mathbf{I}} \otimes \mathbf{Q}^{(\mp)}) \circ \underline{\mathbf{F}}^{(\pm)}(\underline{\mathbf{v}})) \underline{\mathbf{1}} \\ &= \frac{\underline{\mathbf{v}} - \underline{\mathbf{u}}^n}{\Delta t} + \frac{1}{\Delta x} \begin{bmatrix} (\mathbf{Q}^{(\mp)} \circ \mathbf{F}_1) \underline{\mathbf{1}} \\ \vdots \\ (\mathbf{Q}^{(\mp)} \circ \mathbf{F}_m) \underline{\mathbf{1}} \end{bmatrix}. \end{aligned} \quad (26)$$

Here we have introduced the matrix $\underline{\mathbf{F}}^{(\pm)} = \mathbf{F}_1^{(\pm)} \oplus \dots \oplus \mathbf{F}_m^{(\pm)}$ with $(\mathbf{F}_\ell^{(\pm)})_{ij} = \hat{f}_\ell^{(\pm)}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j)$. We denote the partial derivatives of the numerical fluxes as $\hat{f}_{\ell, \phi_\kappa}^{(\pm)} = \partial \hat{f}_\ell^{(\pm)} / \partial \phi_\kappa$ and define the matrices $\mathbf{F}_{\ell, \phi_\kappa}^{(\pm)}(\underline{\mathbf{v}})$ with elements $(\mathbf{F}_{\ell, \phi_\kappa}^{(\pm)})_{ij} = \hat{f}_{\ell, \phi_\kappa}^{(\pm)}(\underline{\mathbf{v}}_i, \underline{\mathbf{v}}_j)$. Finally we obtain

$$\partial \mathbf{F}_{\ell, \phi_\kappa}^{(\pm)}(\underline{\mathbf{v}}) := \mathbf{Q}^{(\mp)} \circ \mathbf{F}_{\ell, \phi_\kappa}^{(\pm)}(\underline{\mathbf{v}}) - \text{diag}(\mathbf{1}^\top (\mathbf{Q}^{(\mp)} \circ \mathbf{F}_{\ell, \phi_\kappa}^{(\pm)}(\underline{\mathbf{v}}))).$$

In [6], the Jacobian $\underline{\mathbf{g}}'(\underline{\mathbf{v}}^{(k)})$ is shown to be given by

$$\underline{\mathbf{g}}'(\underline{\mathbf{v}}^{(k)}) = \frac{\underline{\mathbf{1}}}{\Delta t} + \frac{1}{\Delta x} \begin{bmatrix} \partial \mathbf{F}_{1, \phi_1}^{(\pm)}(\underline{\mathbf{v}}^{(k)}) & \dots & \partial \mathbf{F}_{1, \phi_m}^{(\pm)}(\underline{\mathbf{v}}^{(k)}) \\ \vdots & \ddots & \vdots \\ \partial \mathbf{F}_{m, \phi_1}^{(\pm)}(\underline{\mathbf{v}}^{(k)}) & \dots & \partial \mathbf{F}_{m, \phi_m}^{(\pm)}(\underline{\mathbf{v}}^{(k)}) \end{bmatrix}.$$

This is a block matrix formed by m^2 submatrices, each tridiagonal as in the scalar case (although the off-diagonal blocks do not depend on Δt). If we write $\Delta \mathbf{v} = (\Delta \mathbf{v}_1^\top, \dots, \Delta \mathbf{v}_m^\top)^\top$ and collect terms, then the Newton iteration (25) takes the form

$$\begin{aligned} \frac{\mathbf{v}^{(k+1)} - \mathbf{u}^n}{\Delta t} + \frac{1}{\Delta x} \begin{bmatrix} \partial \mathbf{F}_{1,\phi_1}^{(\pm)}(\mathbf{v}^{(k)}) & \dots & \partial \mathbf{F}_{1,\phi_m}^{(\pm)}(\mathbf{v}^{(k)}) \\ \vdots & \ddots & \vdots \\ \partial \mathbf{F}_{m,\phi_1}^{(\pm)}(\mathbf{v}^{(k)}) & \dots & \partial \mathbf{F}_{m,\phi_m}^{(\pm)}(\mathbf{v}^{(k)}) \end{bmatrix} \begin{bmatrix} \Delta \mathbf{v}_1 \\ \vdots \\ \Delta \mathbf{v}_m \end{bmatrix} \\ + \frac{1}{\Delta x} \begin{bmatrix} (\mathbf{Q}^{(\mp)} \circ \mathbf{F}_1^{(\pm)}(\mathbf{v}^{(k)})) \mathbf{1} \\ \vdots \\ (\mathbf{Q}^{(\mp)} \circ \mathbf{F}_m^{(\pm)}(\mathbf{v}^{(k)})) \mathbf{1} \end{bmatrix} = \mathbf{0}. \end{aligned}$$

Looking at the i th row of the ℓ th block in this linear system we find that the scheme once again may be expressed on the locally conservative form

$$\frac{v_{\ell_i}^{(k+1)} - u_{\ell_i}^n}{\Delta t} + \frac{1}{\Delta x} \left(h_{\ell_i + \frac{1}{2}}^{(k+1)} - h_{\ell_i - \frac{1}{2}}^{(k+1)} \right) = 0, \quad i = \dots, -1, 0, 1, \dots, \quad (27)$$

where $\ell = 1, \dots, m$. The numerical flux is given by

$$h_{\ell_i + \frac{1}{2}}^{(k+1)} = \sum_{\kappa=1}^m \left(\hat{f}_{\ell, \phi_\kappa}^{(\pm)}(\mathbf{v}_{\kappa_i}^{(k)}, \mathbf{v}_{\kappa_{i+1}}^{(k)}) \Delta v_{\kappa_{i+1}} \pm \hat{f}_{\ell, \phi_\kappa}^{(\pm)}(\mathbf{v}_{\kappa_{i+1}}^{(k)}, \mathbf{v}_{\kappa_i}^{(k)}) \Delta v_{\kappa_i} \right) + \hat{f}_\ell^{(\pm)}(\mathbf{v}_{\ell_{i+1}}^{(k)}, \mathbf{v}_{\ell_i}^{(k)}).$$

Further, if $\underline{\mathbf{g}}'(\mathbf{u} \otimes \mathbf{1})$ is nonsingular for any vector \mathbf{u} with non-zero elements, then by the same argument as in the scalar case, the numerical flux is consistent. We may thus conclude:

Theorem 6. *Choose $v_{\ell_i}^{(0)} = u_{\ell_i}^n$ for each i and $\ell = 1, \dots, m$ and suppose that $\underline{\mathbf{g}}'(\mathbf{u} \otimes \mathbf{1})$ is nonsingular for any vector \mathbf{u} with non-zero elements. Then Newton's method, applied to the discretization (26), is locally conservative and flux consistent.*

Motivated by the conservation and flux consistency of Newton's method we conjecture that under conditions similar to 1, the numerical approximation converges to a weak solution of the original conservation law (1). In 6 we present numerical results supporting this conjecture.

The solution $\Delta \mathbf{v}$ to the linear systems arising within Newton's method will in practice be approximated using a second iterative method. If we for this purpose consider pseudo-time iterations, then by the conservation and flux consistency of Newton's method, 2 applies. The only change necessary is to choose the initial guess $\Delta \mathbf{u}^{(0)}$ as the null-vector. We summarize these observations in the following:

Theorem 7. *Apply K iterations of Newton's method to the implicit scheme (2) with initial guess $\mathbf{v}^{(0)} = \mathbf{u}^n$. Approximate the solution to the linear system in the j th Newton iteration using N_j pseudo-time iterations with ERK methods and initial guess $\Delta \mathbf{u}^{(0)} = \mathbf{0}$. The Newton-pseudo-time iterations are locally conservative with numerical fluxes consistent with $c\mathbf{f}$, where*

$$c = 1 - \prod_{l=0}^{\sum_j N_j - 1} \phi_l(-\mu_l). \quad (28)$$

Proof. Since Newton's method is locally conservative and flux consistent, the result is a direct consequence of 2. \square

5 Krylov subspace methods

Krylov subspace methods are a more common choice than pseudo-time iterations for approximating the solutions of the linear systems arising within Newton's method. The goal of this section is to establish local conservation of Krylov subspace methods, and thereby also of Newton-Krylov methods, by relating them to pseudo-time iterations.

Let us consider a system of linear conservation laws and a corresponding linearly implicit discretization in the form (2). Due to the linearity of the numerical flux, there is a matrix \mathbf{D} such that the scheme can be expressed as

$$\underbrace{\left(\frac{\mathbf{I}}{\Delta t} + \frac{\mathbf{D}}{\Delta x} \right)}_{\mathbf{M}} \mathbf{u}^{n+1} = \underbrace{\frac{\mathbf{u}^n}{\Delta t}}_{\mathbf{d}}.$$

We assume that \mathbf{M} is an invertible matrix.

By definition, the $(k+1)$ st iteration of a Krylov subspace method finds an approximate solution $\mathbf{w}^{(k+1)}$ to $\mathbf{M}\mathbf{u} = \mathbf{d}$ such that $\mathbf{w}^{(k+1)} - \mathbf{w}^{(0)} \in \mathcal{K}_{k+1}(\mathbf{M}, \mathbf{r}^{(0)})$. The Krylov subspace $\mathcal{K}_{k+1}(\mathbf{M}, \mathbf{r}^{(0)})$ is defined as

$$\mathcal{K}_{k+1}(\mathbf{M}, \mathbf{r}^{(0)}) := \text{span}(\{\mathbf{r}^{(0)}, \mathbf{M}\mathbf{r}^{(0)}, \dots, \mathbf{M}^k \mathbf{r}^{(0)}\}), \quad (29)$$

and $\mathbf{r}^{(0)} = \mathbf{d} - \mathbf{M}\mathbf{w}^{(0)}$ is the initial residual. In other words, there are coefficients $\alpha_0, \dots, \alpha_k$ such that

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(0)} + \left(\sum_{\ell=0}^k \alpha_\ell \mathbf{M}^\ell \right) \mathbf{r}^{(0)}. \quad (30)$$

We will now show that all Krylov subspace methods are locally conservative. To this end, we first establish a connection between Krylov subspace methods and pseudo-time iterations with explicit RK methods:

Lemma 8. Let $\underline{\mathbf{w}}^{(k+1)}$ denote the $(k+1)$ st iterate of a Krylov subspace method. There exists an ERK method with $s = k+1$ stages such that the first iterate $\underline{\mathbf{u}}^{(1)}$ of a pseudo-time iteration using this method satisfies $\underline{\mathbf{u}}^{(1)} = \underline{\mathbf{w}}^{(k+1)}$.

Proof. We prove the lemma in two steps: The first step is to show that

$$\underline{\mathbf{u}}^{(1)} - \underline{\mathbf{u}}^{(0)} \in \mathcal{K}_{k+1}(\underline{\mathbf{M}}, \underline{\mathbf{r}}^{(0)})$$

when $\underline{\mathbf{u}}^{(0)} = \underline{\mathbf{w}}^{(0)}$. In other words, a single pseudo-time iteration with a $(k+1)$ -stage explicit RK methods yields an approximation from the appropriate Krylov subspace. The second step is to choose $(\vec{\mathbf{A}}, \vec{\mathbf{b}}, \vec{\mathbf{c}})$ for the Runge-Kutta method such that the coefficients $\alpha_0, \dots, \alpha_k$ in (30) are recovered.

For the first step, recall that pseudo-time iterations approximate solutions to the steady state problem

$$\underline{\mathbf{u}}_\tau = \underline{\mathbf{d}} - \underline{\mathbf{M}}\underline{\mathbf{u}} =: \underline{\mathbf{r}}(\underline{\mathbf{u}}), \quad \underline{\mathbf{u}}(0) = \underline{\mathbf{u}}^{(0)}.$$

A single iteration with an explicit RK method is obtained as

$$\begin{aligned} \vec{\underline{\mathbf{U}}} &= \vec{\mathbf{I}} \otimes \underline{\mathbf{u}}^{(0)} + \Delta\tau(\vec{\mathbf{A}} \otimes \mathbf{I})\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}), \\ \underline{\mathbf{u}}^{(1)} &= \underline{\mathbf{u}}^{(0)} + \Delta\tau(\vec{\mathbf{b}}^\top \otimes \mathbf{I})\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}), \end{aligned} \tag{31}$$

where

$$\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}) = ((\underline{\mathbf{d}} - \underline{\mathbf{M}}\underline{\mathbf{u}}_1)^\top, \dots, (\underline{\mathbf{d}} - \underline{\mathbf{M}}\underline{\mathbf{u}}_s)^\top)^\top = (\vec{\mathbf{I}} \otimes \underline{\mathbf{d}}) - (\vec{\mathbf{I}} \otimes \underline{\mathbf{M}})\vec{\underline{\mathbf{U}}}.$$

Inserting $\vec{\underline{\mathbf{U}}}$ from (31) into the latter expression leads to

$$\begin{aligned} \vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}) &= (\vec{\mathbf{I}} \otimes \underline{\mathbf{d}}) - (\vec{\mathbf{I}} \otimes \underline{\mathbf{M}}) \left[\vec{\mathbf{I}} \otimes \underline{\mathbf{u}}^{(0)} + \Delta\tau(\vec{\mathbf{A}} \otimes \mathbf{I})\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}) \right] \\ &= (\vec{\mathbf{I}} \otimes \underline{\mathbf{d}}) - (\vec{\mathbf{I}} \otimes \underline{\mathbf{M}}\underline{\mathbf{u}}^{(0)}) - \Delta\tau(\vec{\mathbf{A}} \otimes \underline{\mathbf{M}})\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}) \\ &= (\vec{\mathbf{I}} \otimes \underline{\mathbf{r}}^{(0)}) - \Delta\tau(\vec{\mathbf{A}} \otimes \underline{\mathbf{M}})\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}). \end{aligned}$$

Solving for $\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}})$ results in

$$\vec{\underline{\mathbf{R}}}(\vec{\underline{\mathbf{U}}}) = \left[\vec{\mathbf{I}} \otimes \mathbf{I} + \Delta\tau(\vec{\mathbf{A}} \otimes \underline{\mathbf{M}}) \right]^{-1} (\vec{\mathbf{I}} \otimes \underline{\mathbf{r}}^{(0)}).$$

Inserting this expression into $\underline{\mathbf{u}}^{(1)}$ from (31) yields

$$\begin{aligned} \underline{\mathbf{u}}^{(1)} &= \underline{\mathbf{u}}^{(0)} + \Delta\tau(\vec{\mathbf{b}} \otimes \mathbf{I}) \left[\vec{\mathbf{I}} \otimes \mathbf{I} + \Delta\tau(\vec{\mathbf{A}} \otimes \underline{\mathbf{M}}) \right]^{-1} (\vec{\mathbf{I}} \otimes \underline{\mathbf{r}}^{(0)}) \\ &= \underline{\mathbf{u}}^{(0)} + \underline{\mathbf{M}}^{-1} \left(\Delta\tau(\vec{\mathbf{b}} \otimes \underline{\mathbf{M}}) \left[\vec{\mathbf{I}} \otimes \mathbf{I} + \Delta\tau(\vec{\mathbf{A}} \otimes \underline{\mathbf{M}}) \right]^{-1} (\vec{\mathbf{I}} \otimes \mathbf{I}) \right) \underline{\mathbf{r}}^{(0)}. \end{aligned}$$

We identify the paranthesised portion of this expression as $\mathbf{I} - \phi(-\Delta\tau\mathbf{M})$, where $\phi(z)$ is the stability function of the RK method. Thus,

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} + \mathbf{M}^{-1}(\mathbf{I} - \phi(-\Delta\tau\mathbf{M}))\mathbf{r}^{(0)}. \quad (32)$$

Now, $\phi(z)$ is a polynomial of degree $s = k + 1$ and from (7) we see that its constant term is 1. Hence, $\mathbf{M}^{-1}(\mathbf{I} - \phi(-\mathbf{M}))$ is a polynomial in \mathbf{M} of degree k . Consequently, $\mathbf{u}^{(1)} - \mathbf{u}^{(0)} \in \mathcal{K}_{k+1}(\mathbf{M}, \mathbf{r}^{(0)})$, which completes the first step of the proof.

For the second step we must find an explicit RK method whose stability polynomial satisfies $z^{-1}(1 - \phi(-z)) = \sum_{\ell=0}^k \alpha_\ell z^\ell$. We do not require this RK method to be accurate or even consistent; it suffices that it exists.

Consider the $(k + 1)$ -stage explicit RK method defined by the Butcher tableau

$$\begin{array}{c|cccc} 0 & 0 & & & \\ c_1 & -a & 0 & & \\ c_2 & & -a & 0 & \\ \vdots & & & \ddots & \ddots \\ c_k & & & & -a & 0 \\ \hline & b_0 & b_1 & \dots & b_{k-1} & b_k \end{array}$$

By direct computation we can verify that $(\vec{\mathbf{I}} - z\vec{\mathbf{A}})^{-1}$ is a lower triangular Toeplitz matrix with elements $(-az)^\ell$ on the ℓ th subdiagonal ($\ell = 0$ being the main diagonal and $\ell = k$ being the bottom left element). From (7), the stability function of this method is therefore

$$\phi(z) = 1 + z \sum_{i=0}^k \sum_{j=i}^k b_j (-az)^i.$$

Consequently we are looking for coefficients b_j that satisfy the relation

$$\frac{1 - \phi(-z)}{z} = \sum_{i=0}^k \sum_{j=i}^k b_j (az)^i = \sum_{\ell=0}^k \alpha_\ell z^\ell.$$

Matching exponents of z , this amounts to solving the triangular system

$$\begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 0 & 1 & \dots & 1 & 1 \\ & & \ddots & & \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} a^{-0}\alpha_0 \\ a^{-1}\alpha_1 \\ \vdots \\ a^{-k}\alpha_k \end{pmatrix}.$$

This system is uniquely solvable for any coefficients α_ℓ , $\ell = 0, \dots, k$ and $a \neq 0$. Hence, an explicit RK method can always be found such that a single pseudo-time iteration reproduces the $(k + 1)$ th Krylov vector when the initial guess is given by $\underline{\mathbf{u}}^{(0)} = \underline{\mathbf{w}}^{(0)}$. \square

With 8 established, the following is a direct consequence of 2:

Theorem 9. *With the initial guess $\underline{\mathbf{w}}^{(0)} = \underline{\mathbf{u}}^n$, Krylov subspace methods applied to the locally conservative linear discretization (2) are locally conservative.*

Proof. For any iteration of any Krylov subspace method, 8 shows that there is an equivalent explicit RK method that yields the same numerical solution in one pseudo-time iteration. By 2, all explicit RK methods are locally conservative, hence so are Krylov subspace methods. \square

We do not know what the coefficients $\alpha_0, \dots, \alpha_k$ are and consequently we cannot identify a corresponding RK method. Further, each α_j will in general change with every iteration. We can therefore not know a priori if a given Krylov subspace method is flux consistent. We can also not apply the extension of the Lax-Wendroff theorem in 4 since α_j in general will not remain constant upon grid refinement.

On the other hand, 9 suggests that we can apply one pseudo-time iteration with the explicit Euler method prior to applying the Krylov subspace method in order to enforce consistency, without violating local conservation. Further yet, since 2 is indifferent to the RK method used in previous iterations, we can conclude that 9 also applies to *restarted* Krylov methods, with each restart corresponding to a pseudo-time iteration with its own RK method. In fact, local conservation will be retained even if we swap Krylov method mid-solve, or if we mix Krylov subspace methods and pseudo-time iterations.

Finally we note that 9 together with 6 implies that Newton-Krylov methods are locally conservative:

Theorem 10. *Apply Newton's method to the implicit scheme (2) with initial guess $\underline{\mathbf{v}}^{(0)} = \underline{\mathbf{u}}^n$. Approximate the solution to the each linear system within the Newton iterations using any Krylov subspace method with initial guess $\Delta \underline{\mathbf{w}}^{(0)} = \underline{\mathbf{0}}$. Then the Newton-Krylov iterations are locally conservative.*

Proof. Since Newton's method is locally conservative and flux consistent by 6, the result is a direct consequence of 9. \square

6 Numerical experiments

As target problem for the experiments in this section we use the 2D compressible Euler equations,

$$\begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (\rho E + p)u \end{bmatrix}_x + \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (\rho E + p)v \end{bmatrix}_y = 0, \quad (33)$$

posed on the domain $(x, y) \in (-5, 15] \times (-5, 5]$. Here, ρ, u, v, E and p respectively denote density, horizontal and vertical velocity components, total energy per unit mass and pressure. The pressure is related to the other variables through the equation of state

$$p = (\gamma - 1)\rho e,$$

where $\gamma = 1.4$ and $e = E - (u^2 + v^2)/2$ is the internal energy density. The domain is taken to be periodic in both spatial coordinates. The setting is the isentropic vortex problem [17] with initial conditions

$$\begin{aligned} \rho_0 &= \left(\frac{1 - \epsilon^2(\gamma - 1)M_\infty^2}{8\pi^2} \exp(r) \right)^{\frac{1}{\gamma-1}}, \\ u_0 &= 1 - \frac{\epsilon y}{2\pi} \exp(r/2), \\ v_0 &= \frac{\epsilon x}{2\pi} \exp(r/2), \\ p_0 &= \frac{\rho_0^\gamma}{\gamma M_\infty^2}, \end{aligned}$$

where $r = 1 - x^2 - y^2$. Here, $\epsilon = 5$ is the circulation and $M_\infty = 0.5$ is the Mach number. As the solution evolves in time, the initial vortex propagates in the horizontal direction with unit speed.

This problem does not adhere to the assumptions made throughout the paper since it is posed in 2D. Nevertheless, we will see shortly that the theoretical results presented so far appear to hold also in this setting.

6.1 Convergence tests

Pseudo-time iterations: SSPRK3

We begin by corroborating the extension of the Lax-Wendroff theorem presented in 4. With implicit Euler used in time, experimental verification of this theorem was provided in [4]. However, 3 implies that the result holds

also when other implicit RK methods are used in time, so long as they fulfill condition (16).

In the following experiments we discretize the Euler equations in space using finite volumes with central fluxes and the three-stage Lobatto IIIC method in time. This fully implicit RK method satisfies (16) with the vector $\tilde{\mathbf{v}} = (0, 0, 1)^\top$. The simulations are run to time $t = 0.1$ with space and time grids satisfying by $\Delta y = \Delta x/2$ and $\Delta t = \Delta x/4$. In each time step, the discrete solution is approximated using three pseudo-time iterations using the explicit strong stability preserving method SSPRK3 from [18]. The pseudo-time steps are chosen to be $\Delta \tau_i = \Delta t/\sqrt{i}$, $i = 1, 2, 3$. The stability function of SSPRK3 is given by

$$\phi(z) = 1 + z + z^2/2 + z^3/6.$$

4 therefore predicts that the solution of the scheme, if convergent, approaches the solution to a modified conservation law with modification constant

$$c(\mu_0, \mu_1, \mu_2) = 1 - \phi(\mu_0)\phi(\mu_1)\phi(\mu_2) \approx 0.9101.$$

For this particular problem, the presence of c causes a corresponding reduction of the vortex propagation speed. 1a shows the L^2 -error in the density component of the solution, as measured with respect to the exact solution of the original system of conservation laws (33) and that of the modified version. Clearly the numerical solution approaches the solution of the modified equations, not the original ones.

Shown also is the error, measured with respect to the original conservation law, of the same scheme but where flux consistency has been enforced using a single pseudo-time iteration with explicit Euler and $\mu = 1$. As expected, this ensures convergence towards the correct solution.

Newton-SSPRK3

We now repeat the experiment but add two Newton iterations per time step. The solutions to the linear systems are then approximated using the same three pseudo-time iterations as before. In addition to the discretization described previously, we also perform the experiment with a different one: In space, the entropy conservative and kinetic energy preserving finite volume scheme of Chandrashekar [7] is used and in time, the two-stage Radau IIA method, which satisfies (16) with the vector $\tilde{\mathbf{v}} = (0, 1)^\top$.

7 suggests that the resulting schemes are consistent with a modified system of conservation laws with modification constant

$$c = 1 - (\phi(\mu_0)\phi(\mu_1)\phi(\mu_2))^2 \approx 0.9919.$$

Rather than computing the Jacobian explicitly, it is approximated using finite differences such that for any two vectors \mathbf{u} and \mathbf{v} ,

$$\log'(\mathbf{u})\mathbf{v} \approx \frac{\log(\mathbf{u} + \epsilon\mathbf{v}) - \log(\mathbf{u})}{\epsilon}, \quad \epsilon = \frac{10^{-7}}{\|\mathbf{v}\|}. \quad (34)$$

Although an extension of the Lax-Wendroff theorem incorporating Newton's method is not yet available, 1b indicates convergence towards the solution of the modified equations, not of the original ones. Once again, this is remedied by enforcing flux consistency with a single explicit Euler iteration. Here, the errors corresponding to the two different space-time discretizations are similar enough that the lines are on top of each other.

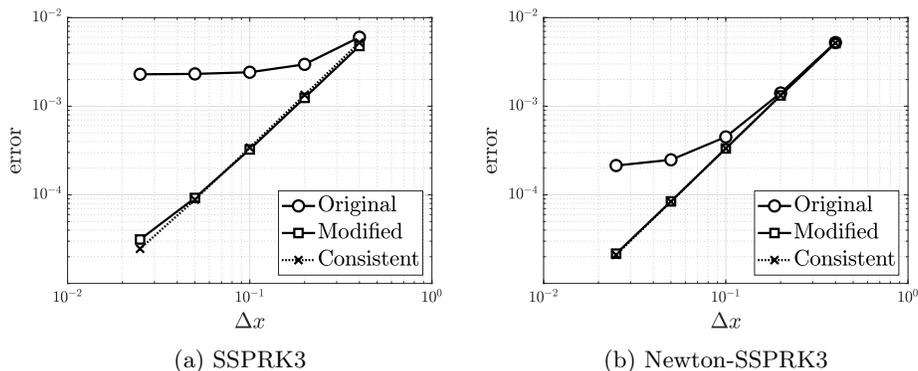


Figure 1: Density error upon grid refinement for the compressible Euler equations. Convergence is seen towards the modified conservation laws (Modified), not the original ones (Original), unless flux consistency is enforced (Consistent). (a) Pseudo-time iterations using SSPRK3. (b) Newton's method with SSPRK3 as subsolver for the linear systems.

Newton-Krylov: Fixed iterations

Newton-Krylov methods are a more common choice of solver than Newton with pseudo-time iterations. 10 establishes that they are all locally conservative. Conceivably, a Lax-Wendroff type result might be available for these methods, although we have not presented one here. It is of interest to explore experimentally if convergence is observed, and if so, towards what solution.

We repeat the previous experiment, this time with the three-stage Lobatto IIIC method in time and Chandrashekar's finite volume scheme in

space, but replace SSPRK3 with GMRES. Here we consider four cases: Using one Newton and one Krylov iteration per time step (N1K1); one Newton and two Krylov iterations (N1K2); one Newton iteration with GMRES run to a tolerance of 10^{-14} (N1), i.e. effectively with 'exact' linear solves; Newton-GMRES run until the Newton residual is smaller than 10^{-15} (Exact), i.e. effectively with 'exact' nonlinear solves. 2a shows the error with respect to the original conservation law of the four schemes. The errors for N1K2, N1 and the exact solver are on top of each other, suggesting that the discretization dominates the error. The numerical solutions appear to converge to the correct solution. However, N1K1 displays a different behaviour, suggesting that the iteration error dominates. The error curve appears to flatten as the grid is refined, although it cannot be deduced whether it will continue towards zero or if it reaches a plateau. Thus, it remains unclear if GMRES is flux consistent.

Newton-Krylov: Tolerance

In practice, the number of Newton iterations will not be preset but rather governed by a relative and/or an absolute tolerance. Here, we set both of these to a value tol and once again explore the convergence behaviour. To set the tolerances of the GMRES iterations we follow the procedure described by Eisenstat and Walker [8] with parameters $\gamma = \eta_{max} = 0.9$; see [11, Chapter 6] for details. The density errors using $tol \in \{10^{-3}, 10^{-4}, 10^{-5}\}$ are shown in 2b.

Two distinct phenomena can be observed: Firstly, when $tol = 10^{-5}$, the convergence behaviour changes from one similar to the exact solver in 2a to one resembling N1K1. Thus, the error is seen to change from being discretization dominated to being iteration dominated as the grid is refined. With $tol = 10^{-4}$ the iteration error appears to dominate throughout.

Secondly, when $tol = 10^{-3}$ the error eventually stops converging. This behavior is explained by the observation that with a fixed tolerance, the initial guess will be a sufficiently accurate approximation of the solution if Δt is small enough. In that case, the Newton and GMRES iterations are terminated without updating the solution.

6.2 Acceleration experiments

In [4], numerical experiments were performed indicating that considerable efficiency gains can be made with pseudo-time iterations by enforcing flux consistency. It is not clear whether Newton-Krylov methods are flux consistent. It is therefore worthwhile exploring if enforced flux consistency leads to efficiency gains also in this case. We use the tolerance governed Newton-GMRES solver with the Eisenstat-Walker procedure and compare

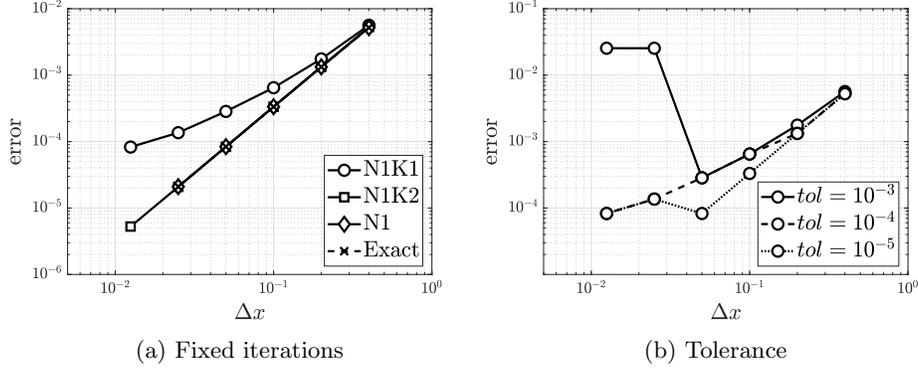


Figure 2: Density error upon grid refinement for the compressible Euler equations. (a) The number of iterations is fixed: One Newton and one GMRES iteration per time step (N1K1); One Newton and two GMRES iterations (N1K2); one Newton with a nearly exact linear solver (N1); nearly exact Newton-GMRES (Exact). (b) Tolerance governed Newton-GMRES with the Eisenstat-Walker procedure.

the standard solver with one where flux consistency is enforced with explicit Euler before each call to GMRES.

In this experiment we use Chandrashekar’s finite volume scheme with the implicit Euler method in time to compute a single time step with $\Delta t = 0.1$. The running cost of the two solvers is measured in terms of the number of evaluations of the full space discretization. Each Newton iteration requires a single such evaluation; see (22). Each GMRES iteration also needs a single function evaluation during the computation of the approximate Jacobian matrix-vector product (34).

The k th pseudo-time step with explicit Euler applied within the j th Newton iteration takes the form

$$\frac{\Delta \mathbf{u}^{(k+1)} - \Delta \mathbf{u}^{(k)}}{\Delta \tau_k} + \mathbf{g}'(\mathbf{v}^{(j)}) \Delta \mathbf{u}^{(k)} + \mathbf{g}(\mathbf{v}^{(j)}) = \mathbf{0}.$$

Consider the case $k = 0$ and recall from 7 that local conservation follows if the initial guess is $\Delta \mathbf{u}^{(0)} = \mathbf{0}$. Thus, enforcing flux consistency by one pseudo-time iteration with explicit Euler simply amounts to setting $\Delta \mathbf{u}^{(1)} = -\Delta t \mathbf{g}(\mathbf{v}^{(j)})$. This single function evaluation is already computed within Newton’s method, hence flux consistency comes at no additional cost. In fact, the only change necessary to the solver is to alter the initial guess for GMRES from $\Delta \mathbf{w}^{(0)} = \mathbf{0}$ to $\Delta \mathbf{w}^{(0)} = -\Delta t \mathbf{g}(\mathbf{v}^{(j)})$.

Three cases with different CFL numbers are considered. 3a shows the

number of function evaluations required to reach a particular residual $\|\underline{\mathbf{g}}(\mathbf{v}^{(j)})\|$, where the L^2 -norm is used. 3b shows the function evaluations distributed across the Newton iterations. In all cases, the flux consistent initial guess (dotted lines) reduces the number of necessary iterations for residuals greater than roughly 10^{-3} , compared to the regular solver (solid lines). However, for smaller residuals the situation varies with the CFL number. At the smallest CFL, flux consistency remains beneficial even for finer tolerances. However, for the largest CFL the opposite trend is seen.

A possible explanation for these observations is that the explicit Euler method introduces significant errors to the numerical solution when the CFL number is large. Its stability region is small, hence problems with large CFL numbers are unsurprising. It is possible to find other explicit Runge-Kutta methods that enforce flux consistency, in principle with much larger stability regions. However, such methods will necessarily have more stages and thus impart additional costs on the solver.

Finally we note that at the smallest tolerances, no discernable difference is seen between the two solvers. Presumably this happens because flux consistency is achieved, either exactly or very nearly, by the standard Newton-GMRES solver when the tolerance is small and the number of iterations is large. In conclusion, whether enforced flux consistency is beneficial for Newton-GMRES is case dependent.

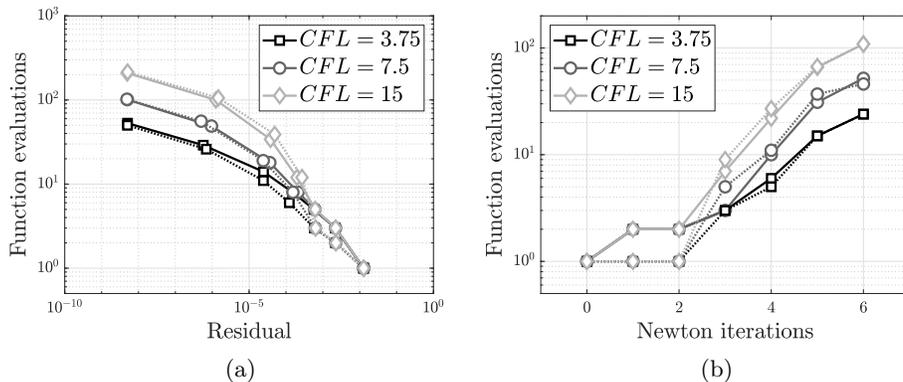


Figure 3: Efficiency study for Newton-GMRES with standard (solid lines) and flux consistent (dotted lines) initial guesses. (a) Function evaluations needed to reach a given residual. (b) Function evaluations per Newton iteration.

7 Conclusions

In this paper, the concepts of locally conservative and flux consistent iterative methods have been introduced and shown to be of both theoretical and practical interest. Based on earlier work in [4], it was shown that pseudo-time iterations using explicit Runge-Kutta methods are locally conservative but not necessarily flux consistent, when applied to conservative discretizations of finite volume-type with a broad class of implicit Runge-Kutta methods. For 1D problems, an extension of the Lax-Wendroff theorem reveals convergence towards weak solutions of a temporally retarded system of conservation laws. Each equation is modified in the same way, namely by a particular scalar factor multiplying the spatial flux terms. Flux consistency, and thereby convergence, is recovered through a technique based on using the explicit Euler method.

Local conservation has further been established for all Krylov subspace methods, with and without restarts, as well as for Newton's method under the assumption of bivariate fluxes. Thus it follows that Newton-Krylov methods are locally conservative, although not necessarily flux consistent. Numerical experiments with the 2D compressible Euler equations suggest that the role enforced flux consistency is case dependent. Its effect diminishes as the number of GMRES iterations grow, presumably because flux consistency is achieved automatically.

References

- [1] T. BARTH, *Analysis of implicit local linearization techniques for upwind and TVD algorithms*, in 25th AIAA Aerospace Sciences Meeting, 1987, p. 595.
- [2] P. BIRKEN, *Numerical methods for unsteady compressible flow problems*, Chapman and Hall/CRC, 2021.
- [3] P. BIRKEN, J. BULL, AND A. JAMESON, *Preconditioned smoothers for the full approximation scheme for the RANS equations*, Journal of Scientific Computing, 78 (2019), pp. 995–1022.
- [4] P. BIRKEN AND V. LINDERS, *Conservative iterative methods for implicit discretizations of conservation laws*, arXiv preprint arXiv:2106.10088, (2021).
- [5] P. D. BOOM AND D. W. ZINGG, *High-order implicit time-marching methods based on generalized Summation-By-Parts operators*, SIAM Journal on Scientific Computing, 37 (2015), pp. A2682–A2709.

- [6] J. CHAN AND C. G. TAYLOR, *Efficient computation of Jacobian matrices for entropy stable summation-by-parts schemes*, J. Comput. Phys., 448 (2022), p. 110701.
- [7] P. CHANDRASHEKAR, *Kinetic energy preserving and entropy stable finite volume schemes for compressible Euler and Navier-Stokes equations*, Communications in Computational Physics, 14 (2013), pp. 1252–1286.
- [8] S. C. EISENSTAT AND H. F. WALKER, *Choosing the forcing terms in an inexact Newton method*, SIAM Journal on Scientific Computing, 17 (1996), pp. 16–32.
- [9] D. JESPERSEN AND T. PULLIAM, *Flux vector splitting and approximate Newton methods*, in 6th Computational Fluid Dynamics Conference Danvers, 1983, p. 1899.
- [10] C. JUNQUEIRA-JUNIOR, L. C. SCALABRIN, E. BASSO, AND J. L. F. AZEVEDO, *Study of conservation on implicit techniques for unstructured finite volume Navier–Stokes solvers*, Journal of Aerospace Technology and Management, 6 (2014), pp. 267–280.
- [11] C. T. KELLEY, *Iterative methods for linear and nonlinear equations*, SIAM, 1995.
- [12] P. LAX AND B. WENDROFF, *Systems of conservation laws*, tech. rep., LOS ALAMOS NATIONAL LAB NM, 1959.
- [13] R. J. LEVEQUE, *Numerical methods for conservation laws*, vol. 3, Springer, 1992.
- [14] V. LINDERS, J. NORDSTRÖM, AND S. H. FRANKEL, *Properties of Runge-Kutta-Summation-By-Parts methods*, Journal of Computational Physics, 419 (2020), p. 109684.
- [15] J. NORDSTRÖM AND T. LUNDQUIST, *Summation-By-Parts in time*, Journal of Computational Physics, 251 (2013), pp. 487–499.
- [16] H. RANOCHA, *Some notes on Summation By Parts time integration methods*, Results in Applied Mathematics, 1 (2019), p. 100004.
- [17] C.-W. SHU, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced numerical approximation of nonlinear hyperbolic equations, Springer, 1998, pp. 325–432.

- [18] C.-W. SHU AND S. OSHER, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of computational physics, 77 (1988), pp. 439–471.
- [19] R. C. SWANSON, E. TURKEL, AND C.-C. ROSSOW, *Convergence acceleration of Runge–Kutta schemes for solving the Navier–Stokes equations*, Journal of Computational Physics, 224 (2007), pp. 365–388.
- [20] L. M. VERSBACH, V. LINDERS, R. KLÖFKORN, AND P. BIRKEN, *Theoretical and practical aspects of space-time DG-SEM implementations*, arXiv preprint arXiv:2201.05800, (2022).
- [21] G. WANNER AND E. HAIRER, *Solving ordinary differential equations II*, Springer Berlin Heidelberg, 1996.