Numerische Mathematik (2020) 146:875–906
https://doi.org/10.1007/s00211-020-01158-4

Numerische
Mathematik

# General relaxation methods for initial-value problems with application to multistep schemes

Hendrik Ranocha[1] · Lajos Lóczi[2,3] · David I. Ketcheson[1]

Received: 6 March 2020 / Revised: 17 October 2020 / Accepted: 20 October 2020 /
Published online: 28 October 2020

**Abstract**
Recently, an approach known as relaxation has been developed for preserving the correct evolution of a functional in the numerical solution of initial-value problems, using Runge–Kutta methods. We generalize this approach to multistep methods, including all general linear methods of order two or higher, and many other classes of schemes. We prove the existence of a valid relaxation parameter and high-order accuracy of the resulting method, in the context of general equations, including but not limited to conservative or dissipative systems. The theory is illustrated with several numerical examples.

**Mathematics Subject Classification** 65L06 · 65L20 · 65M12 · 65M70 · 65P10 · 37M99

## 1 Introduction

Consider an initial-value ordinary differential equation (ODE) in a Banach space:

$$u'(t) = f(u(t)) \quad u(0) = u^0. \tag{1}$$

✉ Hendrik Ranocha
mail@ranocha.de

Lajos Lóczi
LLoczi@inf.elte.hu

David I. Ketcheson
david.ketcheson@kaust.edu.sa

[1] Extreme Computing Research Center (ECRC), Computer Electrical and Mathematical Science and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

[2] Department of Numerical Analysis, Eötvös Loránd University, Budapest, Hungary

[3] Department of Differential Equations, Budapest University of Technology and Economics, Budapest, Hungary

Here and in the following, we use upper indices for $t$ and $u$ to denote the index of the corresponding time step. We say the problem (1) is *dissipative* with respect to a smooth functional $\eta$ if

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\eta(u(t)) \leq 0 \tag{2a}$$

for all solutions $u$ of (1), i.e. if

$$\forall u\colon \quad \eta'(u) f(u) \leq 0. \tag{2b}$$

In the case of equality in (2), we say the problem is *conservative*. In the numerical solution of dissipative or conservative problems, it is desirable to enforce the same property discretely. For a $k$-step method we thus require

$$\eta(u^n) \leq \max\{\eta(u^{n-1}), \eta(u^{n-2}), \dots, \eta(u^{n-k})\} \tag{3}$$

for dissipative problems, or

$$\eta(u^n) = \eta(u^{n-1}) \tag{4}$$

for conservative problems. A numerical method satisfying this requirement is also said to be dissipative (also known as monotone) or conservative, respectively.

For instance, initial-value problems for hyperbolic or parabolic partial differential equations (PDEs) usually have a conserved or dissipated quantity, but in the presence of boundary and/or source terms this quantity may sometimes increase. In that case, energy/entropy estimates are still important and the methods developed in this article are still applicable.

## 1.1 Related work

Conservative or dissipative ODEs arise in a variety of applications and various approaches exist for enforcing these properties discretely; for conservative problems see e.g., [25], and for dissipative problems see e.g., [15,22] and references therein. Besides classical examples such as Hamiltonian systems, many hyperbolic or hyperbolic-parabolic PDEs such as the Euler and Navier–Stokes equations are equipped with an entropy whose evolution in time is important both physically and for mathematical and numerical stability estimates [13, Chapter 5]. While there are many semidiscretely entropy-conservative or -dissipative numerical methods [11,17,19,34,42,43,61,68], transferring such semidiscrete results to fully discrete schemes is not easy in general. Proofs of monotonicity for fully discrete schemes have mainly been limited to semidiscretizations including certain amounts of dissipation [28,30,47,70], linear equations [52,63,65,67], or fully implicit time integration schemes [4,6,7,18,34,39,44]. For explicit methods and general equations, there are negative experimental and theoretical results concerning energy/entropy stability [36,37,45,48].

To cope with the limitations of time integration schemes, several methods for enforcing discrete conservation or dissipation have been proposed. These include orthogonal projections, for one-step methods [23,59] [25, Section IV.4] and multistep methods [16,20,60], as well as more problem-dependent techniques for dissipative ODEs such as artificial dissipation or filtering [21,41,64]. For one-step methods, there are also extensions to projection methods employing more general search directions via embedded Runge–Kutta (RK) methods [9,10,33]. Kojima [32] reviewed some related methods and proposed another kind of projection scheme for conservative systems.

The ideas of relaxation methods can be traced back to [56,58] and [15, pp. 265–266]. A relaxation approach was applied in the first two references to the leapfrog method, and in the third reference to the fourth-order Runge–Kutta method, in each case to conserve or dissipate an inner-product norm; see also [9]. General relaxation Runge–Kutta methods without order reduction have been proposed and analyzed recently in [31,46,50,53]. Herein we further generalize the relaxation approach to multistep methods; we focus on linear multistep methods but the theoretical results apply to virtually any conceivable method for (1), including for instance all general linear methods. In the context of partial differential equations, the relaxation approach is not even limited to a method-of-lines framework.

### 1.2 Outline of the article

Firstly, we introduce the general relaxation approach for time integration methods in Sect. 2. The proofs of accuracy and existence of solutions for the relaxation parameter $\gamma$ are divided into multiple steps and presented in Sects. 2.1–2.3, where the latter section contains the most general result. Section 3 shows how to compute useful estimates for the evolution of $\eta(u)$ for non-conservative problems. In Sect. 4, we study the accuracy of multistep relaxation methods in the case when the method coefficients are not adapted to account for the variable step size. Afterwards, we study stability and accuracy properties of relaxation methods in Sect. 5 and present numerical results supporting our analysis in Section 6. Finally, we summarize our findings and present some directions of future research in Sect. 7. An additional analysis of superconvergence results for relaxation methods is contained in the extended version of this article available on arXiv.org [51].

## 2 A general relaxation approach

To describe the general relaxation approach, we first write a $k$-step, order $p$ (with $p \geq 2$) time integration method for the ODE (1) in the form

$$u^{\text{new}} = \psi(f, \Delta t, u^{n-1}, u^{n-2}, \dots, u^{n-k}). \tag{5}$$

here $u^{\text{new}} \approx u(t^{\text{new}})$ is the numerical solution that ordinarily would be used to continue marching in time, and $t^{\text{new}} = t^{n-1} + \Delta t$ is the corresponding time of approximation. But since $u^{\text{new}}$ might violate a desired dissipativity (3) or conservation (4) property, we perform a line search along the (approximate) secant line connecting $u^{\text{new}}$ and a convex combination $u^{\text{old}}$ of previous solution values, where

$$u^{\text{old}} = \sum_{i=0}^{m-1} \nu_i u^{n-m+i}, \qquad t^{\text{old}} = \sum_{i=0}^{m-1} \nu_i t^{n-m+i}, \tag{6}$$

for a fixed $m \geq 1$ with $\nu_i \geq 0$ and $\sum_i \nu_i = 1$, to find a conservative or dissipative solution:

$$u^n_\gamma = u^{\text{old}} + \gamma (u^{\text{new}} - u^{\text{old}}). \tag{7a}$$

As we will show, under quite general assumptions, there is always a positive value of $\gamma$ that guarantees (3) or (4) and is very close to unity, so that $u^n_\gamma$ approximates $u(t^n_\gamma)$, where

$$t^n_\gamma = t^{\text{old}} + \gamma (t^{\text{new}} - t^{\text{old}}), \tag{7b}$$

to the same order of accuracy as the original approximate solution $u^{\text{new}}$. We will usually suppress the subscript $\gamma$ unless there is a reason to emphasize this dependence. Additionally, the dependence of the relaxation parameter $\gamma$ on the time step is also not written out explicitly.

We now describe how $\gamma$ is chosen at each step. Given an invariant $\eta$, $\gamma$ is chosen such that

$$\eta(u^n_\gamma) = \eta^{\text{old}}, \qquad \eta^{\text{old}} = \sum_{i=0}^{m-1} \nu_i \eta(u^{n-m+i}). \tag{8}$$

Obviously, if $\eta$ is an invariant and the previous step values were computed in a conservative way, then

$$\eta^{\text{old}} = \sum_{i=0}^{m-1} \nu_i \eta(u^{n-m+i}) = \sum_{i=0}^{m-1} \nu_i \eta(u^0) = \eta(u^0). \tag{9}$$

If $\eta$ is not an invariant, a suitable estimate

$$\eta^{\text{new}} \approx \eta(u(t^{\text{new}})) = \eta(u^{\text{new}}) + \mathscr{O}(\Delta t^{p+1}) \tag{10}$$

has to be obtained first. In particular, this estimate $\eta^{\text{new}}$ should be obtained such that the correct sign of the discrete rate of change can be guaranteed. Then, $\gamma$ has to be chosen such that

$$\eta(u^n_\gamma) = \mathscr{H}(t^n_\gamma) := \eta^{\text{old}} + \gamma (\eta^{\text{new}} - \eta^{\text{old}}). \tag{11}$$

Obviously, a suitable choice for invariants is $\eta^{\text{new}} = \eta^{\text{old}}$. Hence, this approach is a strict generalization of relaxation methods for invariants to general functionals $\eta$.

In summary, at each time step the general relaxation algorithm consists of the following substeps:

1. Define the values
$$\begin{pmatrix} t^{\text{old}} \\ u^{\text{old}} \\ \eta^{\text{old}} \end{pmatrix} = \sum_{i=0}^{m-1} \nu_i \begin{pmatrix} t^{n-m+i} \\ u^{n-m+i} \\ \eta(u^{n-m+i}) \end{pmatrix} \tag{12}$$
as base points of the secants in time, phase space, and "entropy" space. These old values are convex combinations, i.e. $\nu_i \geq 0$ and $\sum_i \nu_i = 1$.
2. Compute new values $t^{\text{new}}$ and $u^{\text{new}} \approx u(t^{\text{new}}) + \mathscr{O}(\Delta t^{p+1})$ using a given time integration scheme (5) and a suitable estimate $\eta^{\text{new}} = \eta(u^{\text{new}}) + \mathscr{O}(\Delta t^{p+1})$.
3. Solve the system
$$\begin{pmatrix} t^n_\gamma \\ u^n_\gamma \\ \eta(u^n_\gamma) \end{pmatrix} = \begin{pmatrix} t^{\text{old}} \\ u^{\text{old}} \\ \eta^{\text{old}} \end{pmatrix} + \gamma \begin{pmatrix} t^{\text{new}} - t^{\text{old}} \\ u^{\text{new}} - u^{\text{old}} \\ \eta^{\text{new}} - \eta^{\text{old}} \end{pmatrix} \tag{13}$$
for $\gamma \approx 1$ and continue the integration with $t^n_\gamma$, $u^n_\gamma$ instead of $t^{\text{new}}$, $u^{\text{new}}$.

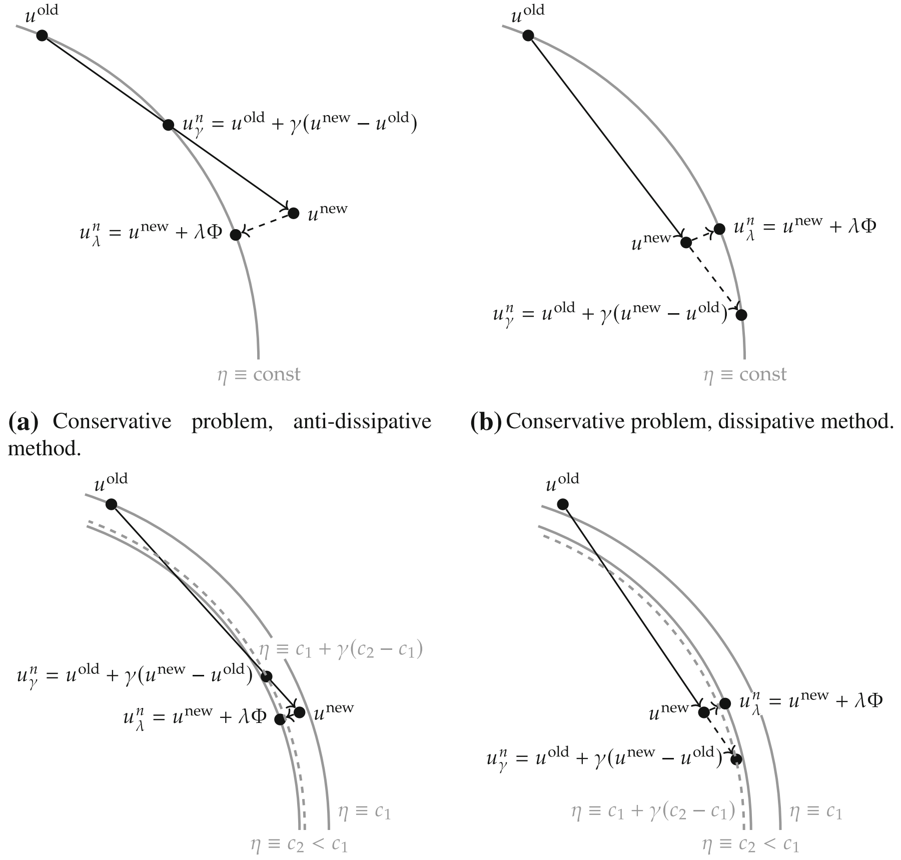
***Remark 2.1*** Solving (13) means inserting the second equation into the third and solving the resulting scalar Eq. (11) for $\gamma$. Then, $t^n_\gamma$ and $u^n_\gamma$ are determined by the first and second equation, respectively.

***Remark 2.2*** Throughout this article, the notation $\mathscr{O}(\cdot)$ refers to the limit $\Delta t \to 0$. As mentioned above, superscripts of $t$ and $u$ denote the time step. Inside $\mathscr{O}(\cdot)$, superscripts of $\Delta t$ and $(\gamma - 1)$ denote exponents.

***Remark 2.3*** The standard choice of the old values (12) is given by $m = 1$ and $\nu_0 = 1$, especially for one-step methods. For dissipative problems, if $m = 1$ and the starting values satisfy $\eta(u^{k-1}) \leq \eta(u^{k-2}) \leq \cdots \leq \eta(u^0)$, the relaxation approach described in the following will guarantee the slightly stronger inequality
$$\eta(u^n) \leq \eta(u^{n-1}) \leq \eta(u^{n-2}) \leq \cdots \leq \eta(u^{n-k}) \leq \cdots \leq \eta(u^0) \tag{14}$$
instead of (3) if the estimate $\eta^{\text{new}}$ is obtained such that the correct sign of the discrete rate of change can be guaranteed.

***Remark 2.4*** One could also consider the case $t^{\text{old}} = t^{\text{new}}$, i.e. $u^{\text{old}} \approx u(t^{\text{new}})$. In that case, for Runge–Kutta methods the relaxation approach reduces to the projection method using an embedded pair studied in [9], where less accuracy of $u^{\text{old}}$ is needed and the new time $t^{\text{new}}$ does not need to be adapted. Here, we focus on the case $t^{\text{old}} < t^{\text{new}}$.

General projection methods replace the numerical solution $u^{\text{new}}$ with
$$u^n_\lambda = u^{\text{new}} + \lambda \Phi \approx u(t^n), \tag{15}$$
where $\Phi$ is a specified search direction and $\lambda$ is chosen such that $\eta(u^n_\lambda) = \eta^{\text{new}}$. Projection methods do not modify the new time $t^{\text{new}} = t^n$. The projection method used most often in applications is orthogonal projection, where $\Phi$ is chosen to minimize the


$u^{\text{old}}$
$u_\gamma^n = u^{\text{old}} + \gamma(u^{\text{new}} - u^{\text{old}})$
$u^{\text{new}}$
$u_\lambda^n = u^{\text{new}} + \lambda\Phi$
$\eta \equiv \text{const}$
$u^{\text{old}}$
$u^{\text{new}}$
$u_\lambda^n = u^{\text{new}} + \lambda\Phi$
$u_\gamma^n = u^{\text{old}} + \gamma(u^{\text{new}} - u^{\text{old}})$
$\eta \equiv \text{const}$
$u^{\text{old}}$
$\eta \equiv c_1 + \gamma(c_2 - c_1)$
$u_\gamma^n = u^{\text{old}} + \gamma(u^{\text{new}} - u^{\text{old}})$
$u_\lambda^n = u^{\text{new}} + \lambda\Phi$
$u^{\text{new}}$
$\eta \equiv c_1$
$\eta \equiv c_2 < c_1$
$u^{\text{old}}$
$u^{\text{new}}$
$u_\lambda^n = u^{\text{new}} + \lambda\Phi$
$u_\gamma^n = u^{\text{old}} + \gamma(u^{\text{new}} - u^{\text{old}})$
$\eta \equiv c_1 + \gamma(c_2 - c_1)$
$\eta \equiv c_1$
$\eta \equiv c_2 < c_1$

**(a)** Conservative problem, anti-dissipative method.

**(b)** Conservative problem, dissipative method.

**(c)** Dissipative problem, anti-dissipative method.

**(d)** Dissipative problem, dissipative method.

**Fig. 1** Visualization of the orthogonal projection and relaxation approaches for conservative/dissipative problems and anti-dissipative/dissipative time integration methods for $m = 1$ in (12)

distance $\|u^{\text{new}} - u_\lambda^n\|$. Often, simplified Newton iterations are used in such projection methods [25, Section IV.4].

The orthogonal projection and relaxation modifications of time integration schemes are visualized in Fig. 1 for $m = 1$. For conservative problems, both relaxation and orthogonal projection yield results on the same level set of the invariant $\eta$. If $\eta$ is convex and the baseline method is anti-dissipative, the relaxation approach decreases the actual time step $\gamma \Delta t$, i.e. $\gamma < 1$. If the baseline scheme is dissipative, relaxation yields larger effective time steps $\gamma \Delta t$, i.e. $\gamma > 1$.

For dissipative problems, the corrected numerical solutions of orthogonal projection and relaxation methods are in general on different level sets of $\eta$. For both anti-dissipative and dissipative baseline schemes, the value of $\eta(u_\gamma^n)$ is closer to $\eta^{\text{new}}$ than $\eta(u^{\text{new}})$. This is in accordance with the adaptation of the time (7b): while $\eta^{\text{new}}$ is an approximation at time $t^{\text{new}}$, $\eta(u_\gamma^n)$ is an approximation at $t_\gamma^n$.

***Remark 2.5*** Based on the sketches shown in Fig. 1, one can expect that it is also possible to choose $\widetilde{\gamma}$ such that $\eta(u^n_{\widetilde{\gamma}}) = \eta^{\mathrm{new}}$ for dissipative problems. At least for convex problems visualized there, there will be two solutions $\widetilde{\gamma}$ for sufficiently small time steps $\Delta t$. One of these solutions is near unity and the other one is closer to zero.

In order to solve $\eta(u^n_{\widetilde{\gamma}}) = \eta^{\mathrm{new}}$, $\widetilde{\gamma}$ must deviate more from unity than a solution $\gamma$ of $\eta(u^n_\gamma) = \mathscr{H}(t^n_\gamma)$, cf. (11). Since $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$ for a $p$th order baseline scheme as will be shown below, $\widetilde{\gamma}$ cannot be closer to unity than $\mathscr{O}(\Delta t^{p-1})$. In the following, we will prove

$$u^n_\gamma = u(t^{\mathrm{new}}) + \mathscr{O}(\Delta t^{p-1}) \quad \text{and} \quad u^n_\gamma = u(t^n_\gamma) + \mathscr{O}(\Delta t^{p+1}). \tag{16}$$

Hence, $\eta^{\mathrm{new}}$ is only an $\mathscr{O}(\Delta t^{p-1})$ approximation at $t^n_\gamma$ and cannot be better at $t^n_{\widetilde{\gamma}}$. Thus, solving $\eta(u^n_{\widetilde{\gamma}}) = \eta^{\mathrm{new}}$ will lead to some order reduction and is not pursued further.

Following the development of the relaxation approach for Runge–Kutta methods [31,53], the accuracy and suitability of general relaxation time integration methods is studied in three steps. Firstly, accuracy of the relaxed approximation (13) is studied given assumptions on the relaxation parameter $\gamma \approx 1$. Secondly, existence and accuracy of a suitable relaxation parameter $\gamma$ satisfying (11) is studied at first for convex $\eta$ and then for general $\eta$. Finally, methods for computing $\eta^{\mathrm{new}}$ are given.

### 2.1 Accuracy of the solution *u* for relaxation methods

Before proving the accuracy of $u^n_\gamma$, we introduce the following

**Lemma 2.6** *For a smooth function $\varphi$, define*

$$\varphi^{\mathrm{old}} = \sum_{i=0}^{m-1} \nu_i \varphi(u^{n-m+i}), \tag{17}$$

*where $\nu_i$ and $m$ are as in (12). Then, $\varphi^{\mathrm{old}} = \varphi(u^{\mathrm{old}}) + \mathscr{O}(\Delta t^2)$.*

***Proof*** Consider the expansions

$$\varphi^{\mathrm{old}} = \sum_{i=0}^{m-1} \nu_i \varphi(u^{n-m+i}) = \sum_{i=0}^{m-1} \nu_i \Big( \varphi(u^{n-m}) + \varphi'(u^{n-m})(u^{n-m+i} - u^{n-m}) \Big) + \mathscr{O}(\Delta t^2) \tag{18}$$

and

$$\varphi(u^{\mathrm{old}}) = \varphi\left( \sum_{i=0}^{m-1} \nu_i u^{n-m+i} \right) = \varphi(u^{n-m}) + \varphi'(u^{n-m}) \left( \sum_{i=0}^{m-1} \nu_i u^{n-m+i} - u^{n-m} \right) + \mathscr{O}(\Delta t^2). \tag{19}$$

Because of $\sum_i \nu_i = 1$, we have $\varphi^{\mathrm{old}} - \varphi(u^{\mathrm{old}}) = \mathscr{O}(\Delta t^2)$. □

Note that in general

$$\begin{aligned} u^{\mathrm{old}} = \sum_{i=0}^{m-1} \nu_i u^{n-m+i} &= \sum_{i=0}^{m-1} \nu_i u(t^{n-m+i}) + \mathscr{O}(\Delta t^{p+1}) \\ &\neq u\Bigl(\sum_{i=0}^{m-1} \nu_i t^{n-m+i}\Bigr) + \mathscr{O}(\Delta t^{p+1}) = u(t^{\mathrm{old}}) + \mathscr{O}(\Delta t^{p+1}), \end{aligned} \tag{20}$$

except for $m = 1$ or similarly special choices of $\nu = (\nu_i)_i$. In general, $u^{\mathrm{old}} = u(t^{\mathrm{old}}) + \mathscr{O}(\Delta t^2)$.

**Lemma 2.7** *If the method (5) is of order $p \geq 2$ and $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$, the relaxation solution (13) is also of order $p$.*

***Proof*** Use the accuracy of the baseline method (5) and apply Lemma 2.6 to get the expansion

$$\begin{aligned} u^n_\gamma &= u^{\mathrm{new}} + (\gamma - 1)(u^{\mathrm{new}} - u^{\mathrm{old}}) \\ &= u(t^{\mathrm{new}}) + (\gamma - 1)\bigl(u(t^{\mathrm{new}}) - u(t^{\mathrm{old}})\bigr) + \mathscr{O}(\Delta t^{p+1}) + \mathscr{O}\bigl((\gamma - 1)\Delta t^2\bigr) \\ &= u(t^{\mathrm{new}}) + u'(t^{\mathrm{new}})(\gamma - 1)(t^{\mathrm{new}} - t^{\mathrm{old}}) + \mathscr{O}(\Delta t^{p+1}) + \mathscr{O}\bigl((\gamma - 1)\Delta t^2\bigr). \end{aligned} \tag{21}$$

Subtracting the Taylor expansion

$$\begin{aligned} u(t^n_\gamma) &= u(t^{\mathrm{new}} + (\gamma - 1)(t^{\mathrm{new}} - t^{\mathrm{old}})) \\ &= u(t^{\mathrm{new}}) + u'(t^{\mathrm{new}})(\gamma - 1)(t^{\mathrm{new}} - t^{\mathrm{old}}) + \mathscr{O}\bigl((\gamma - 1)^2 \Delta t^2\bigr) \end{aligned} \tag{22}$$

results in the estimate

$$u^n_\gamma - u(t^n_\gamma) = \mathscr{O}(\Delta t^{p+1}) + \mathscr{O}\bigl((\gamma - 1)^2 \Delta t^2\bigr) + \mathscr{O}\bigl((\gamma - 1)\Delta t^2\bigr) = \mathscr{O}(\Delta t^{p+1}). \tag{23}$$

□

***Remark 2.8*** This basic argument (for $m = 1$, i.e. $u^{\mathrm{old}} = u^{n-1}$) proves also the accuracy of relaxation Runge–Kutta methods if $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$; cf. [31,53]. In particular, it simplifies the proof of [31, Theorem 2.7] by avoiding the usual Runge–Kutta order conditions. In addition, the result holds also for more general schemes such as the class of general linear methods [8] or (modified) Patankar–Runge–Kutta methods [5].

### 2.2 Existence and accuracy of $\gamma$ for relaxation methods for convex entropies

**Lemma 2.9** *Consider a relaxation method (13) based on a time integration method of order $p \geq 2$.*

*If $\eta$ is a convex entropy for the ODE (1), $\Delta t$ is sufficiently small, and $\eta''(u^{\mathrm{old}})\bigl(f(u^{\mathrm{old}}), f(u^{\mathrm{old}})\bigr) \neq 0$, then there is a unique $\gamma > 0$ that solves (13). This $\gamma$ satisfies $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$.*

***Proof*** The function $r$ defined in a neighborhood of $[0, 1]$ by

$$\begin{aligned} r(\gamma) &= \eta(u^n_\gamma) - \mathscr{H}(t^n_\gamma) \\ &= \eta\big(u^{\mathrm{old}} + \gamma(u^{\mathrm{new}} - u^{\mathrm{old}})\big) - \eta^{\mathrm{old}} - \gamma\big(\eta^{\mathrm{new}} - \eta^{\mathrm{old}}\big) \end{aligned} \tag{24}$$

is convex, since $\eta$ is convex. Moreover,

$$r(0) = \eta(u^{\mathrm{old}}) - \eta^{\mathrm{old}} = \eta\left(\sum_{i=0}^{m-1} \nu_i u^{n-m+i}\right) - \sum_{i=0}^{m-1} \nu_i \eta(u^{n-m+i}) \le 0, \tag{25}$$

since $u^{\mathrm{old}}$ is a convex combination and $\eta$ is convex. Furthermore,

$$\begin{aligned} r'(0) &= \eta'(u^{\mathrm{old}})(u^{\mathrm{new}} - u^{\mathrm{old}}) - \eta^{\mathrm{new}} + \eta^{\mathrm{old}} \\ &= \eta^{\mathrm{old}} - \eta(u^{\mathrm{new}}) + \eta'(u^{\mathrm{old}})(u^{\mathrm{new}} - u^{\mathrm{old}}) + \mathscr{O}(\Delta t^{p+1}). \end{aligned} \tag{26}$$

Because of

$$\sum_{i=0}^{m-1} \nu_i \eta'(u^{n-m+i}) = \eta'(u^{\mathrm{old}}) + \mathscr{O}(\Delta t^2), \tag{27}$$

cf. Lemma 2.6, we have

$$\begin{aligned} r'(0) &= -\frac{1}{2}\eta''(u^{\mathrm{old}})(u^{\mathrm{new}} - u^{\mathrm{old}}, u^{\mathrm{new}} - u^{\mathrm{old}}) + \mathscr{O}(\Delta t^3) \\ &= -\frac{1}{2}(t^{\mathrm{new}} - t^{\mathrm{old}})^2 \eta''(u^{\mathrm{old}})(f(u^{\mathrm{old}}), f(u^{\mathrm{old}})) + \mathscr{O}(\Delta t^3) < 0 \end{aligned} \tag{28}$$

for sufficiently small $\Delta t > 0$. Here, the accuracy of the estimate (10) has been used in the second line. Similarly,

$$\begin{aligned} r'(1) &= \eta'(u^{\mathrm{new}})(u^{\mathrm{new}} - u^{\mathrm{old}}) - \eta^{\mathrm{new}} + \eta^{\mathrm{old}} \\ &= \eta^{\mathrm{old}} - \eta(u^{\mathrm{new}}) - \eta'(u^{\mathrm{new}})(u^{\mathrm{old}} - u^{\mathrm{new}}) + \mathscr{O}(\Delta t^{p+1}) \\ &= \frac{1}{2}\eta''(u^{\mathrm{new}})(u^{\mathrm{old}} - u^{\mathrm{new}}, u^{\mathrm{old}} - u^{\mathrm{new}}) + \mathscr{O}(\Delta t^3) \\ &= \frac{1}{2}(t^{\mathrm{new}} - t^{\mathrm{old}})^2 \eta''(u^{\mathrm{new}})(f(u^{\mathrm{old}}), f(u^{\mathrm{old}})) + \mathscr{O}(\Delta t^3) > 0 \end{aligned} \tag{29}$$

for sufficiently small $\Delta t > 0$. Hence, $r$ has a unique positive root $\gamma$.

Because of the accuracy of the baseline scheme, $r(1) = \mathscr{O}(\Delta t^{p+1})$. Using (29), $r'(1) = c\Delta t^2 + \mathscr{O}(\Delta t^3)$ with $c > 0$. Hence, the root $\gamma$ of $r$ satisfies $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$. □

***Remark 2.10*** Instead of the condition $\eta''(u^{\mathrm{old}})\big(f(u^{\mathrm{old}}), f(u^{\mathrm{old}})\big) \neq 0$, similar conditions using other step/stage values can be used by performing the Taylor expansions around them.

***Remark 2.11*** As can be seen from Fig. 1, choosing $\gamma$ slightly smaller than the root of $r(\gamma) = 0$ is a way to introduce additional dissipation.

***Remark 2.12*** If the convex entropy is a squared inner-product norm, i.e., if $\eta(u) = \frac{1}{2}\|u\|^2$ in a Hilbert space, the relaxation parameter can be calculated explicitly as

$$\gamma = \begin{cases} \frac{-b+\sqrt{b^2-4ac}}{2c}, & a \neq 0, \\ \frac{-b}{c}, & a = 0, \end{cases} \tag{30}$$

where

$$\begin{gathered} a = \eta(u^{\text{old}}) - \eta^{\text{old}} \leq 0, \qquad b = \left\langle u^{\text{old}},\, u^{\text{new}} - u^{\text{old}} \right\rangle - \eta^{\text{new}} + \eta^{\text{old}}, \\ c = \eta(u^{\text{new}} - u^{\text{old}}) \geq 0. \end{gathered} \tag{31}$$

**Theorem 2.13** *Consider a relaxation method* (13) *based on a time integration method of order $p \geq 2$.*

*If $\eta$ is a convex entropy for the ODE* (1)*, $\Delta t$ is sufficiently small, and $\eta''(u^{\text{old}})\big(f(u^{\text{old}}), f(u^{\text{old}})\big) \neq 0$, then there is a unique $\gamma > 0$ that satisfies the relaxation condition* (11) *and the resulting relaxation method is of order $p$.*

***Proof*** Combine Lemmas 2.7 and 2.9. □

### 2.3 Existence and accuracy of $\gamma$ for relaxation methods for general functionals

**Theorem 2.14** *Consider a relaxation method* (13) *based on a time integration method of order $p \geq 2$. If $\eta$ is a general (i.e. not necessarily convex) smooth functional of* (1)*, $\Delta t$ is sufficiently small, and*

$$\eta'(u^{\text{new}}) \frac{u^{\text{new}} - u^{\text{old}}}{\|u^{\text{new}} - u^{\text{old}}\|} = c\Delta t + \mathcal{O}(\Delta t^2), \quad \textit{with } c \neq 0, \tag{32}$$

*then there is a unique $\gamma = 1 + \mathcal{O}(\Delta t^{p-1})$ that satisfies the relaxation condition* (11) *and the resulting relaxation method is of order $p$.*

***Proof*** Following [10, Theorem 2], this proof is based on the implicit function theorem, e.g., the version of [54, Section VIII.2].

Given an initial condition $u^0$ and a time step $\Delta t$, approximate solutions $u^{n-k+i} \approx u(t^{n-k+i})$ and $u^{\text{new}} \approx u(t^{\text{new}})$ are computed, e.g., via a (relaxation) LMM and a suitable starting procedure. In this setting, the proof of [10, Theorem 2] can be adapted, similarly to [53, Proposition 2.18], yielding a unique solution $\gamma = 1 + \mathcal{O}(\Delta t^{p-1})$. Because of Lemma 2.7, the relaxation method is of order $p$. □

***Remark 2.15*** For one-step methods studied in [31,50,53], the natural choice of $m, \nu$ in (12) is $m = 1$ and $\nu_i = \delta_{i,0}$, where $\delta_{i,0}$ is the Kronecker delta. Then, $r(0) = 0$ in (24).

## 3 Suitable estimates of $\eta$ for relaxation methods

For dissipative systems, the relaxation approach requires an estimate of the entropy at $t^{\text{new}}$ satisfying

$$\eta^{\text{new}} \le \eta^{\text{old}} \tag{33}$$

in order to ensure that (3) is satisfied if $m \le k$. We first review the approach of [53] for Runge–Kutta methods with positive weights, showing how it fits naturally into the approach we have just described. We then discuss how to obtain a suitable estimate for multistep methods with positive coefficients, and for more general methods.

### 3.1 Runge–Kutta methods with positive quadrature weights

A Runge–Kutta method with $s$ stages takes the form [8,26]

$$y^i = u^{n-1} + \Delta t \sum_{j=1}^{s} a_{ij}\, f(t^{n-1} + c_j \Delta t, y^j), \qquad i \in \{1, \ldots, s\}, \tag{34a}$$

$$u^{\text{new}} = u^{n-1} + \Delta t \sum_{i=1}^{s} b_i\, f(t^{n-1} + c_i \Delta t, y^i). \tag{34b}$$

Given the Runge–Kutta stage values, a natural estimate $\eta^{\text{new}}$ is given by using the quadrature rule of the Runge–Kutta method itself:

$$\eta^{\text{new}} = \eta(u^{n-1}) + \Delta t \sum_{i=1}^{s} b_i (\eta' f)(y^i). \tag{35}$$

This corresponds to the approach used in [9,15,31,53]. The inequality (33) is guaranteed if the weights $b_i \ge 0$.

***Remark 3.1*** The adaptation of the time $t^n_\gamma$ and this choice of the estimate appear to be very natural. Indeed, considering the augmented system

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} t \\ u(t) \\ \eta(u(t)) \end{pmatrix} = \begin{pmatrix} 1 \\ f(u(t)) \\ (\eta' f)(u(t)) \end{pmatrix}, \tag{36}$$

the update formula (with $\sum_i b_i = 1$)

$$\begin{pmatrix} t^n_\gamma \\ u^n_\gamma \\ \eta(u^n_\gamma) \end{pmatrix} = \begin{pmatrix} t^{n-1} \\ u^{n-1} \\ \eta(u^{n-1}) \end{pmatrix} + \gamma \Delta t \sum_{i=1}^{s} b_i \begin{pmatrix} 1 \\ f(y^i) \\ (\eta' f)(y^i) \end{pmatrix} \tag{37}$$

is a natural discretization of (36), where the relaxation parameter $\gamma$ is introduced to enforce the consistent evolution of $\eta$.

***Remark 3.2*** In the context of dissipative PDEs such as second-order parabolic ones, (35) can provide important (spatial) gradient estimates of the stage/step values by bounding $\eta' f$.

### 3.2 Linear multistep methods with positive coefficients

A linear multistep method can be written in the form

$$u^{\text{new}} = \sum_{i=0}^{k-1} \alpha_i^n u^{n-k+i} + \Delta t \sum_{i=0}^{k} \beta_i^n f^{n-k+i}. \tag{38}$$

The coefficients $\alpha_i^n, \beta_i^n$ have a time index because they depend on the sequence of step sizes. If all coefficients $\alpha_i^n, \beta_i^n$ are non-negative, the method itself can be used to obtain a suitable estimate $\eta^{\text{new}}$. Indeed, considering again the augmented system (36), a high-order estimate is obtained as

$$\eta^{\text{new}} = \sum_{i\geq 0} \left( \alpha_i^n \eta(u^{n-k+i}) + \Delta t \beta_i^n (\eta' f)(u^{n-k+i}) \right) = \eta(u^{\text{new}}) + \mathscr{O}(\Delta t^{p+1}). \tag{39}$$

Since $\sum_i \alpha_i^n = 1$ for any consistent method and $\beta_i^n (\eta' f)(u^{n-k+i}) \leq 0$, the dissipation condition (33) is guaranteed. Note in particular that strong stability preserving (SSP) LMMs have non-negative coefficients [22, Chapter 8] and can be used in this manner.

### 3.3 General time integration methods

The approach of the previous two subsections relies on non-negativity of the coefficients $b_i$ and $\alpha_i, \beta_i$, respectively. For methods with negative coefficients, we can follow the technique developed in [10]. In this approach, an estimate $\eta^{\text{new}}$ is obtained by interpolation (continuous/dense output) and a positive quadrature rule. Indeed, consider a quadrature rule of order at least $p$ with nodes $\tau_i \in [t^{n-m}, t^{\text{new}}]$ and positive weights $w_i$, e.g., a Gauß quadrature. Compute an interpolant $y(\tau_i)$ of order at least $p-1$ at the nodes and use

$$\eta^{\text{new}} = \eta^{n-m} + \sum_i w_i (\eta' f)(y(\tau_i)). \tag{40}$$

Because $w_i > 0$, the estimate is guaranteed to satisfy (33) for dissipative systems.

In contrast to the previous method, this approach requires additional evaluations of $\eta'$ and $f$ at the intermediate values and is thus more costly. On the other hand, it can be applied to general time integration methods and does not require any special structure (besides a continuous output formula). In particular, it is not limited to Runge–Kutta or linear multistep methods.

***Remark 3.3*** This approach is particularly interesting for linear multistep methods, since these schemes are often defined naturally in terms of (interpolating) polynomials [2,38].

## 4 Relaxation linear multistep methods with fixed coefficients

Since orthogonal projection does not inherently involve any change in the step size, it can be used with a fixed step size. In contrast, relaxation methods necessarily introduce variation in the step size, due to the parameter $\gamma$, even if the intended $\Delta t$ at each step is constant. To make use of Theorems 2.13 and 2.14 to guarantee high-order accuracy, multistep relaxation methods must be implemented in a way that takes into account the variation in the step size. On the other hand, since $\gamma$ is very close to unity, this step-size variation is quite small. Thus it is interesting to know what accuracy may be obtained by relaxation methods using the new time step (7b) but implemented with the coefficients of the corresponding fixed step size method. We will write simply $\alpha_i$, $\beta_i$ (without superscripts) to refer to the coefficients of a fixed step size LMM.

**Theorem 4.1** *Given a $k$-step LMM (38) of order $p \geq 1$ with coefficients $(\alpha, \beta)$, suppose that the method is used with step sizes $\Delta t_{n-k+j} = \gamma_j \Delta t_n$, where $\gamma_j - 1 = \mathscr{O}(\Delta t^q)$ but the coefficients are not adapted. Then the one-step error is $\mathscr{O}(\Delta t^{\min(p+1,q+1)})$. Furthermore, for Adams methods the one-step error is $\mathscr{O}(\Delta t^{\min(p+1,q+2)})$.*

***Proof*** The one-step error takes the form [35, p. 133], [24, Eqn. (2.16)]

$$E = \sum_{\ell=0}^{\infty} \Delta t^{\ell} u^{(\ell)}(t^{\text{new}}) C_{\ell}(\alpha, \beta, \gamma) \tag{41}$$

where $C_0 = \sum_j \alpha_j - 1$ and for $\ell > 0$

$$C_{\ell}(\alpha, \beta, \gamma) = \sum_{j=0}^{k-1} \left( \Omega_j^{\ell} \alpha_j + \ell \Omega_j^{\ell-1} \beta_j \right) - \Omega_k^{\ell} \tag{42}$$

with

$$\Omega_j = \sum_{i=1}^{j} \frac{\Delta t_{n-k+i}}{\Delta t_n} = \sum_{i=1}^{j} \gamma_j = j + \sum_{i=1}^{j} (\gamma_j - 1) = j + \delta_j, \tag{43}$$

where $\delta_j = \mathscr{O}(\Delta t^q)$. Since the LMM has order $p$, its coefficients satisfy the fixed step size order conditions $C_0 = 0$ and

$$C_{\ell}^{\text{unif}} := C_{\ell}(\alpha, \beta, \mathbb{1}) = \sum_{j=0}^{k-1} \left( j^{\ell} \alpha_j + \ell j^{\ell-1} \beta_j \right) - k^{\ell} = 0, \quad \ell \in \{1, 2, \ldots, p\}. \tag{44}$$

Subtracting (44) from (42) shows that $C_\ell = \mathcal{O}(\Delta t^q)$ for $1 \le \ell \le p$. Substitution into (41) gives the result stated in the first part of the theorem. For the second result, observe that

$$C_1(\alpha, \beta, \gamma) = \sum_{j=0}^{k-1} \big((j + \delta_j)\alpha_j + \beta_j\big) - k - \delta_k \tag{45}$$

$$= C_1^{\text{unif}} + \sum_{j=0}^{k-1} (\delta_j \alpha_j) - \delta_k = \sum_{j=0}^{k-1} (\delta_j \alpha_j) - \delta_k. \tag{46}$$

Since $\gamma_k = 1$, we have $\delta_k = \delta_{k-1}$. For Adams methods we have $\alpha_{k-1} = 1$, while $\alpha_j = 0$ for $j \neq k-1$, so $C_1(\alpha, \beta, \gamma) = 0$. □

Note that in the theorem we have analyzed the accuracy of $u^{\text{new}}$ rather than that of $u^n_\gamma$, but the proof of Lemma 2.7 shows that $u^n_\gamma$ will have the same order of accuracy. According to Lemma 2.9, the assumption in the theorem is fulfilled for relaxation methods with $q = p - 1$. This suggests that (for non-Adams methods) the local error in the first step will be one order worse than the design order of the method. But then for the next step Lemma 2.9 shows that $\gamma - 1$ will also be one order worse. In this manner, the local error can grow by one order at each step, until very quickly all accuracy is lost and/or a suitable solution $\gamma > 0$ cannot be found. We see that Adams methods are free from this problem. A cure for this problem for another classes of LMMs is discussed next.

Consider a linear multistep method (38) with non-negative coefficients $\alpha_i$ and take the special choice $\nu_i = \alpha_i$ for computing the old values in (12). This yields

$$\begin{pmatrix} t^{\text{new}} \\ u^{\text{new}} \end{pmatrix} - \begin{pmatrix} t^{\text{old}} \\ u^{\text{old}} \end{pmatrix} = \Delta t \sum_{i=0}^{k} \beta_i \begin{pmatrix} 1 \\ f^{n-k+i} \end{pmatrix}. \tag{47}$$

Instead of scaling the time step, this relaxation method can be interpreted as scaling the right hand side of the augmented ODE (36) by introducing the pseudotime $\tau$ with constant time steps $\Delta\tau$ and solving

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \begin{pmatrix} t(\tau) \\ u(\tau) \\ \eta(u(\tau)) \end{pmatrix} = \Gamma(\tau) \begin{pmatrix} 1 \\ f(u(\tau)) \\ (\eta' f)(u(\tau)) \end{pmatrix}, \qquad \frac{\mathrm{d}}{\mathrm{d}\tau} \Gamma(\tau) = \gamma(\tau), \tag{48}$$

where $\gamma(\tau)$ is the relaxation parameter. If $\Gamma$ is continuous and bounded away from zero, this new augmented ODE (48) results from (36) by the variable transformation

$$t(0) = 0, \qquad \frac{\mathrm{d}t(\tau)}{\mathrm{d}\tau} = \Gamma(\tau). \tag{49}$$

Using this interpretation (and choice of $\nu_i$), relaxation LMMs with $\alpha_i \ge 0$ can also be used with fixed coefficients.

**Theorem 4.2** *Consider a fixed coefficient linear multistep method* (38) *of order $p \geq 2$ with non-negative coefficients $\alpha_i \geq 0$ and the corresponding relaxation method* (13) *with $\nu_i = \alpha_i$. If $\eta$ is a smooth functional of* (1)*, $\Delta t$ is sufficiently small, and the non-degeneracy condition of Theorem 2.13 (if $\eta$ is convex) or Theorem 2.14 (for general $\eta$) is satisfied, then there is a solution $\gamma$ of* (13) *and the resulting relaxation method is of order $p$.*

***Proof*** Since the fixed step size method is of order $p \geq 2$, Theorem 2.13 (if $\eta$ is convex) or Theorem 2.14 (for general $\eta$) can be applied. Hence, there is a solution $\gamma(\tau) = 1 + \mathcal{O}(\Delta\tau^{p-1})$ of (13).

Because of the special choice of $\nu_i = \alpha_i$, the sequence of the relaxation solutions is a $p$th order approximation to the solution of (48). Hence, it is also a $p$th order approximation of the solution of (1). □

***Remark 4.3*** Because of the scaling by $\Gamma(\tau)$ in (48), the relaxation parameter $\gamma$ may be further than expected from unity. Indeed, $\gamma(\tau) = 1 + \mathcal{O}(\Delta\tau^{p-1})$ yields $\Gamma(\tau^{\mathrm{final}}) = 1 + \mathcal{O}(\Delta\tau^{p-2})$, since $\mathcal{O}(\Delta\tau^{-1})$ time steps have to be used. Hence, the observed alteration of the physical time $t$ is $\max_\tau |\Gamma(\tau)\gamma(\tau) - 1| = \mathcal{O}(\Delta\tau^{p-2})$.

***Remark 4.4*** A relaxation LMM (13) with adapted coefficients is typically more accurate than the same relaxation LMM with fixed coefficients, in particular for second-order methods. Indeed, the factor $\Gamma(\tau)$ in the underlying ODE (48) grows as $\Gamma(\tau) = 1 + \mathcal{O}(\Delta\tau^{p-2})$. Hence, it does not decrease in size if $\Delta\tau$ is reduced for $p = 2$. Since an increase of $\Gamma(\tau)$ results in an increase of the Lipschitz constant of the underlying ODE (48), it is reasonable to expect bigger errors compared to a relaxation method with adapted coefficients solving (36).

## 5 Stability and accuracy properties of relaxation methods

Since the relaxation parameter $\gamma = 1 + \mathcal{O}(\Delta t^{p-1})$ introduces only a small variation of the baseline time integration scheme, basic stability properties are often not lost, similarly to the case of relaxation Runge–Kutta methods [31, Section 3]. Furthermore, applying relaxation can increase the accuracy of baseline schemes.

The incremental direction technique (IDT) approach is basically the relaxation approach without adapting the new time to (7b). For Runge–Kutta methods, this results in a slight loss of the order of accuracy [9,31,53], i.e. the IDT method is of order $p-1$. For some LMMs and conservative/dissipative problems, IDT versions result in $\gamma = 1+\mathcal{O}(\Delta t^{p-2})$ and an order of accuracy $p-2$. However, there are also dissipative problems where IDT versions fail because no solution for $\gamma$ can be found.

### 5.1 Zero-stability of relaxation linear multistep methods

In general, proving even zero-stability of LMMs with variable step sizes is not easy [62]. Typically, stability of variable step size LMMs is implied if the corresponding fixed step size method has certain stability properties and the step sizes do not vary too much, cf. e.g., [26, Theorems III.5.5 and III.5.7] or [2]. For relaxation LMMs, the step sizes are chosen via the relaxation coefficient $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$. Hence, the step sizes vary as

$$\frac{\Delta t_n}{\Delta t_{n-1}} = \frac{\gamma_n \Delta t}{\gamma_{n-1} \Delta t} = 1 + \mathscr{O}(\Delta t^{p-1}). \tag{50}$$

Thus, under the usual conditions, relaxation LMMs are stable if the time step is small enough.

### 5.2 Strong stability preserving methods

SSP methods with SSP coefficient $\mathscr{C} > 0$ guarantee a given convex stability property under a time step restriction $\Delta t \leq \mathscr{C} \Delta t_{\mathrm{FE}}$ whenever the explicit Euler method satisfies the same convex stability property under the time step restriction $\Delta t \leq \Delta t_{\mathrm{FE}}$; cf. [22] and the references cited therein.

If the relaxation parameter $\gamma \in [0, 1]$, then $u^n_\gamma$ is a convex combination of $u^{\mathrm{new}}$ and $u^{\mathrm{old}}$. Hence, all convex stability properties satisfied by these two values are retained. However, if $\gamma > 1$, $u^n_\gamma$ is not a convex combination and the SSP property with the same SSP coefficient $\mathscr{C}$ can be lost, cf. [31, Theorem 3.3 and Table 1], where also more detailed investigations of the SSP property of relaxation Runge–Kutta methods can be found. Specifically, therein it was proved that the SSP coefficient of a relaxation RK method is not smaller than that of the original RK method if

$$0 \leq \gamma \leq \frac{-1}{R(-\mathscr{C}) - 1} \geq 1, \tag{51}$$

where $R$ is the stability function of the explicit SSP Runge–Kutta method (34). We also have the following

**Theorem 5.1** *Consider an explicit SSP Runge–Kutta method* (34) *with SSP coefficient* $\mathscr{C}$. *The corresponding relaxation RK method* (13) *with* $m = 1$ *has SSP coefficient* $\mathscr{C}_\gamma = \mathscr{C} + \mathscr{O}(\Delta t^{p-1})$.

***Proof*** The stability function $R$ of an $s$-stage Runge–Kutta method with Butcher coefficients $A, b$ is

$$R(z) = 1 + z b^T (\mathrm{I} - zA)^{-1} \mathbb{1}, \tag{52}$$

where $\mathbb{1} = (1, \dots, 1)^T \in \mathbb{R}^s$. The stability function of the relaxation method is given by $R_\gamma(z) = 1 + \gamma(R(z) - 1)$. The relaxation method is SSP with SSP coefficient $\geq \mathscr{C}_\gamma$ if $R_\gamma(-\mathscr{C}_\gamma) \geq 0$, cf. [31, Lemma 3.2 and Theorem 3.3]. Because $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$, this implies $R_\gamma(-\mathscr{C}) = R(-\mathscr{C}) + \mathscr{O}(\Delta t^{p-1})$. □

Let us turn now to linear multistep methods. To maintain the SSP property when the step size is not fixed, we require that any coefficients of the fixed step size method that are equal to zero remain exactly zero when the step size is varied:

$$\alpha_i = 0 \implies \alpha_i^n = 0, \tag{53a}$$

$$\beta_i = 0 \implies \beta_i^n = 0. \tag{53b}$$

This assumption is satisfied by the methods of [24,38]:

**Theorem 5.2** *Consider an explicit SSP LMM* (38) *with SSP coefficient* $\mathscr{C}$ *and satisfying* (53)*. If* $m$ *and* $\nu_i$ *are chosen such that* $\nu_{i-m} > 0 \implies \alpha_{i-k} > 0$*, then for small enough* $\Delta t$ *the relaxation LMM* (13) *is SSP with an SSP coefficient* $\mathscr{C}_\gamma = \mathscr{C} + \mathscr{O}(\Delta t^{p-1})$ *and* $\mathscr{C}_\gamma \geq \mathscr{C}$ *if* $\gamma \in [0, 1]$*. Furthermore, if* $m = k$ *and* $\nu_i = \alpha_i^n$ *then* $\mathscr{C}_\gamma = \mathscr{C}/\gamma$*.*

***Proof*** Here, we use the definition of the SSP coefficient given as $\mathscr{C}_n$ in [24], which can be written as

$$\mathscr{C} = \begin{cases} \max\{r \in [0, \infty) | \forall i : \alpha_i - r\beta_i \geq 0\}, & \text{if } \forall i : \alpha_i, \beta_i \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{54}$$

The relaxation LMM can be interpreted as an LMM with parameters

$$\alpha_{i,\gamma} = \nu_i(1-\gamma) + \gamma\alpha_i, \quad \text{and} \quad \beta_{i,\gamma} = \gamma\beta_i. \tag{55}$$

Using $\gamma = 1 + \mathscr{O}(\Delta t^{p-1})$, $\alpha_{i,\gamma} = \alpha_i + \mathscr{O}(\Delta t^{p-1})$ and $\beta_{i,\gamma} = \beta_i + \mathscr{O}(\Delta t^{p-1})$. For small enough $\Delta t$ these coefficients are also non-negative. Hence, $\mathscr{C}_\gamma = \mathscr{C} + \mathscr{O}(\Delta t^{p-1})$.

Meanwhile, if $m = k$ and $\nu_i = \alpha_i$ then the relaxation LMM can be written as a standard LMM but with coefficients $(\alpha, \gamma\beta)$. Hence the SSP coefficient is $\mathscr{C}/\gamma$. □

Choosing a variable step size method that does not satisfy the implication $\nu_{i-m} > 0 \implies \alpha_{i-k} > 0$ can lead to loss of the SSP property. For instance, the second-order three-step method with variable step size of [24] is given by

$$u^{\text{new}} = \frac{\Omega_2^2 - 1}{\Omega_2^2}\left(u^{n-1} + \frac{\Omega_2}{\Omega_2 - 1}\Delta t f(u^{n-1})\right) + \frac{1}{\Omega_2^2}u^{n-3}, \tag{56}$$

where $\Omega_2 = (t^{n-1} - t^{n-3})/\Delta t$. Taking e.g., $m = 2$ and $\nu_0 \neq 0$ generates a term $(1-\gamma)u^{n-2}$ in the relaxation solution, which destroys the SSP property when $\gamma > 1$.

In general, orthogonal projection methods can violate SSP properties. Indeed, linear functionals are convex and linear invariants are preserved by the explicit Euler method (as well as by all Runge–Kutta, linear multistep, and general linear methods). Since orthogonal projection methods do not conserve linear invariants in general, the corresponding convex stability property is lost.

***Example 5.3*** Consider the two-step second-order SSP Runge–Kutta method SSPRK(2,2) given by

$$\begin{aligned} y^1 &= u^{n-1}, \\ y^2 &= u^{n-1} + \Delta t f(y^1), \\ u^{\text{new}} &= u^{n-1} + \frac{1}{2}\Delta t\left(f(y^1) + f(y^2)\right). \end{aligned} \tag{57}$$

Solutions $u$ of the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} u(t) = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix} u(t), \quad u(0) = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \tag{58}$$

have a constant energy $\eta(u) = \frac{1}{2}\|u\|^2$ and total mass $\mathscr{M}(u) = \sum_i u_i$. The first step of the orthogonal projection SSPRK(2,2) method results in the total mass

$$\mathscr{M}(u_\lambda^1) = -\frac{\sqrt{2}}{\sqrt{2 + 3\Delta t^4}} > \mathscr{M}(u^0). \tag{59}$$

Hence, the convex stability property related to the total mass is violated. In contrast, the relaxation SSPRK(2,2) method preserves the total mass.

## 6 Numerical results

Here, the following classes of linear multistep methods (38) are considered. If not stated otherwise, the estimate $\eta^{\text{new}}$ is obtained using a dense output formula and Gauß quadrature using one ($k = 2$) or two ($k \in \{3, 4\}$) nodes.

- Adams($k$)
  The $k$-step explicit Adams methods (also known as Adams–Bashforth methods) are based on the formula $u^n = u^{n-1} + \int_{t^{n-1}}^{t^n} \mathscr{P}_f$, where $\mathscr{P}_f$ is the polynomial interpolating $f(u^{n-1}), \dots, f(u^{n-k})$, see [3] and [26, Section III.1]. These methods can be used with variable step sizes. A natural dense output at an intermediate value $\tau_i$ is generated by evaluating the integral with upper limit $\tau_i$ instead of $t^n$.
- Nyström($k$)AS
  The $k$-step Nyström methods are based on the formula $u^n = u^{n-2} + \int_{t^{n-2}}^{t^n} \mathscr{P}_f$, where $\mathscr{P}_f$ is again the polynomial interpolating $f(u^{n-1}), \dots, f(u^{n-k})$, see [40] and [26, Section III.1]. Based on these constant step size Nyström methods, an extension to variable step sizes that is equipped with a dense output formula has been proposed by Arévalo and Söderlind [2] and will be denoted as Nyström($k$)AS.
- eBDF($k$), eBDF($k$)AS
  The family of extrapolated backward difference formula (eBDF) methods is based on the formula $\mathscr{P}_u'(t^n) = \mathscr{P}_f(t^n)$, where $\mathscr{P}_u$ and $\mathscr{P}_f$ are polynomials that interpolate the previous step values $u^{n-i}$ and step derivatives $f(u^{n-i})$, respectively, cf. [55]. Based on the constant step size eBDF methods, an extension to variable step sizes that is equipped with a dense output formula has been proposed by Arévalo and Söderlind [2] and will be denoted as eBDF($k$)AS.
- SSP($k$, $p$), SSP($k$, $p$)AS
  Second and third order accurate variable step size SSP LMMs have been proposed in [24]. The estimate of the evolution of $\eta$ can either be based on the evolution predicted by the SSP method itself (39) or on the quadrature (40) using the dense

output formula of [38], which is based on the framework of [2]. If the latter option is chosen, the method is denoted as SSP($k$, $p$)AS.

- EDC($i$,$j$)
  The explicit difference correction (EDC) methods of [1] are extended to variable step sizes and equipped with a dense output using the approach of [2].
- BDF($k$)
  The family of backward difference formula (BDF) methods is based on the formula $p_u'(t^n) = f(\mathscr{P}_u(t^n))$, where $\mathscr{P}_u$ is a polynomial that interpolates the step values $u^n, u^{n-1}, \dots, u^{n-k}$, cf. [12] and [26, Section III.1].

If not stated otherwise, relaxation LMMs have been adapted to the new step sizes using $m = 1$ and $\nu_i = \delta_{i,0}$ in (12). We have checked that the results using different choices of $\nu_i$ are similar. Since Theorem 4.2 can be applied to Adams methods, Nyström methods, and SSP LMMs, we have also tested the corresponding fixed coefficient version of these schemes.

***Remark 6.1*** The modified leapfrog method of [56] is the relaxation Nyström(2)AS method for conservative inner-product norms with $m = 2$ and $\nu_i = \delta_{i,0}$ in (12) and fixed step sizes, cf. Theorem 4.2.

We have implemented the relaxation methods used in this article in Python, using SciPy [69] to solve the scalar non-quadratic equations for the relaxation parameter $\gamma$. Matplotlib [29] has been used to generate the plots. The source code for all numerical examples is available online [49].

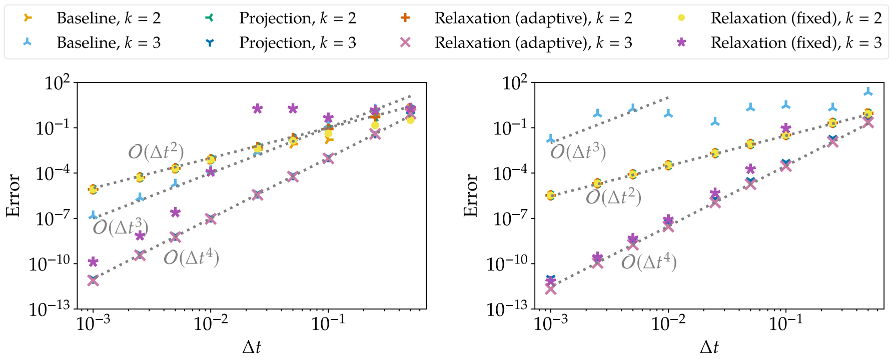
### 6.1 Nonlinear oscillator

For the nonlinear oscillator

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = \|u(t)\|^{-2} \begin{pmatrix} -u_2(t) \\ u_1(t) \end{pmatrix}, \quad u^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{60}$$

of [45,48], the energy $\eta(u) = \frac{1}{2}\|u\|^2$ is conserved. We choose this test problem to demonstrate the convergence properties of relaxation methods in a simple setting where the relaxation parameter can also be computed explicitly by solving a quadratic equation. We use some common explicit methods here because this problem is not stiff.

Results of a convergence study for this problem are visualized in Fig. 2. The Nyström($k$)AS, $k \in \{3, 4\}$, methods result in a large error and are not completely in the asymptotic regime, which could be attributed to their lack of a reasonable stability region [27, Section V.1]. However, applying projection or relaxation results in the expected order of accuracy. The Adams methods do not have similar problems and work well. All other explicit methods described above behave similarly to the Adams methods. For this test problem and formally odd-order relaxation and projection methods, there is a certain superconvergence phenomenon, increasing the experimental order of accuracy by one in accordance with the analysis given in the extended version of this article available on arXiv.org [51].



**(a)** Adams($k$) methods.

**(b)** Nyström($k$)AS methods.

**Fig. 2** Convergence study for linear multistep methods applied to the nonlinear oscillator (60) with final time $t = 20$

The fixed coefficient versions of Adams(3) and Nyström(3)AS result in larger errors than the corresponding versions with adapted coefficients. For smaller time steps $\Delta t$, they are even not in the asymptotic regime. This behavior is in accordance with the analysis of Sect. 4. The energy evolution of a representative example from this section is shown in Fig. 4a.

### 6.2 Kepler problem

The Kepler problem

$$\begin{gathered}\tfrac{\mathrm{d}}{\mathrm{d}t} q(t) = \tfrac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} q_1(t) \\ q_2(t) \end{pmatrix} = p(t), \quad \tfrac{\mathrm{d}}{\mathrm{d}t} p_i(t) = -\frac{q_i(t)}{|q(t)|^3}, \\ q(0) = \begin{pmatrix} 1-e \\ 0 \end{pmatrix}, \quad p(0) = \begin{pmatrix} 0 \\ \sqrt{(1-e)/(1+e)} \end{pmatrix},\end{gathered} \tag{61}$$

with eccentricity $e = 0.5$ is a Hamiltonian system

$$\frac{\mathrm{d}}{\mathrm{d}t} q(t) = \partial_p H\big(q(t), p(t)\big), \quad \frac{\mathrm{d}}{\mathrm{d}t} p(t) = -\partial_q H\big(q(t), p(t)\big), \tag{62}$$

with Hamiltonian

$$H(q, p) = \frac{1}{2}|p|^2 - \frac{1}{|q|}, \tag{63}$$

where the angular momentum

$$L(q, p) = q_1 p_2 - q_2 p_1 \tag{64}$$

is an additional conserved functional, cf. [57, Section 1.2.4]. We choose this test problem to demonstrate the convergence properties of relaxation methods in a more

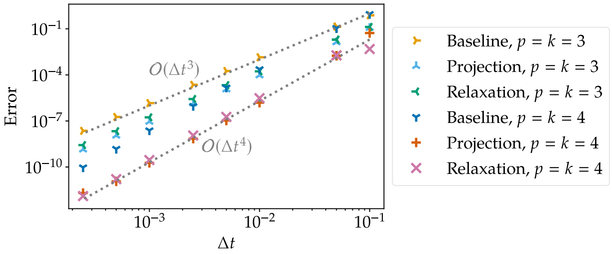
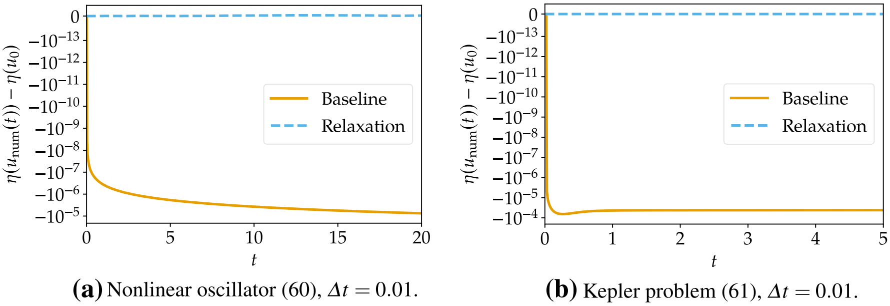


**Fig. 3** Convergence study for eBDF methods applied to the Kepler problem (61) with final time $t = 5$ and projection/relaxation methods conserving the energy. The results for methods conserving the angular momentum are very similar



**(a)** Nonlinear oscillator (60), $\Delta t = 0.01$.

**(b)** Kepler problem (61), $\Delta t = 0.01$.

**Fig. 4** Energy evolution of Adams(3) methods with and without relaxation (adapted time step and coefficients) for representative examples from Sects. 6.1 and 6.2

complex setting where the relaxation parameter is computed using a scalar root finding method. Since this problem is not stiff, we use explicit methods. We choose LMMs different from the ones used in Sect. 6.1 because we want to demonstrate the applicability of relaxation methods for a variety of schemes.

The baseline, projection, and relaxation variants of the explicit multistep methods described above converge with the expected order of accuracy for this problem if the energy or angular momentum is conserved by the projection/relaxation method. As examples, third- and fourth-order accurate eBDF methods yield the convergence results shown in Fig. 3. Clearly, both the projection and the relaxation methods reduce the error compared to the baseline schemes. The results for the other explicit methods described above are similar. The energy evolution of a representative example from this section is shown in Fig. 4b.

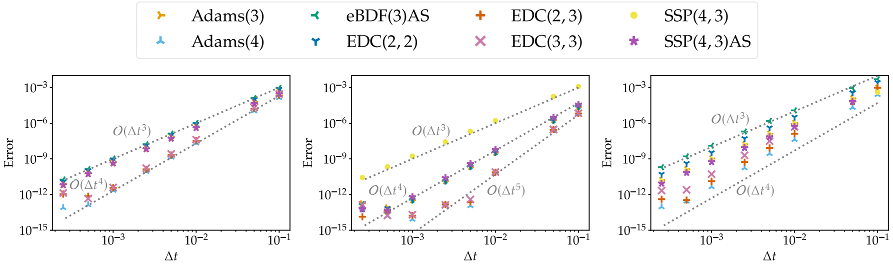
### 6.3 Dissipated exponential entropy

Consider the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} u(t) = -\exp(u(t)), \quad u^0 = 0.5, \tag{65}$$

with exponential entropy $\eta(u) = \exp(u)$, which is dissipated for the analytical solution

$$u(t) = -\log\bigl(\mathrm{e}^{-1/2} + t\bigr). \tag{66}$$



**(a)** Baseline methods. **(b)** Projection methods. **(c)** Relaxation methods.

**Fig. 5** Convergence study for linear multistep methods applied to the ODE (65) with dissipated exponential entropy, final time $t = 20$, and projection/relaxation methods based on the exponential entropy

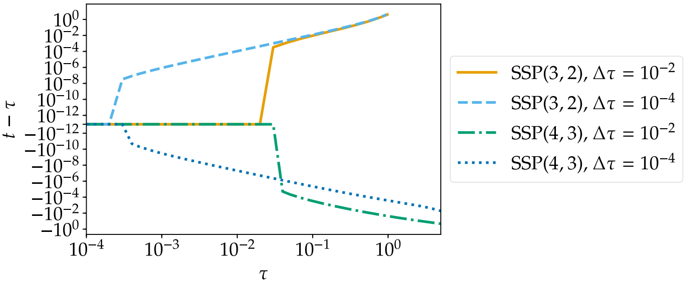


**Fig. 6** Difference of the physical time $t$ and the pseudotime $\tau$ for relaxation SSP LMMs with fixed step size $\Delta\tau$ applied to the ODE (65) with dissipated exponential entropy and final time $t = 5$

In contrast to the conservative problems described above, this problem is dissipative. We choose this test problem to demonstrate the convergence properties of relaxation methods also in this setting. Since the test problem is still not stiff, we choose a variety of explicit methods.

Again, the explicit multistep methods described above with or without projection/relaxation converge with the expected order of accuracy for this problem. As examples, third- and fourth-order accurate multistep methods yield the convergence results shown in Fig. 5. There is no significant difference between the two different estimates (39) and (40) for SSP(4, 3).

Results of fixed step size SSP LMMs with $\nu_i = \alpha_i$ applied to the ODE (65) with dissipated exponential entropy are shown in Fig. 6. Because of the exact starting procedure, $t = \tau$ for the first $k$ steps of a $k$-step method. Thereafter, $|t - \tau|$ increases in time. As discussed in Sect. 4, $\max_\tau |t - \tau| = \mathscr{O}(\Delta\tau^{p-2})$. This can also be observed for SSP(3, 2), where the final value of $t - \tau$ is independent of the time step, and SSP(4, 3), where the final value of $t - \tau$ decreases proportionally to $\Delta\tau$. Since $\max_\tau |t - \tau| = \mathscr{O}(1)$ for SSP(3, 2), the maximal effective relaxation parameter $\Gamma(\tau)\gamma(\tau)$ is also $\mathscr{O}(1)$.

If $\nu_i \neq \alpha_i$, e.g., if $m = 1$ and $\nu_i = \delta_{i,0}$ as usual for methods with adapted step sizes, no solution $\gamma > 0$ can be found for this problem and the SSP LMMs with fixed step sizes. The Adams(2) method with fixed step sizes applied to this problem works well if the final time is reduced to $t = 2.5$. For larger final times, the error of the numerical solutions grows because of the growth of $\Gamma(\tau)$. Then, the time step $\Delta\tau$ has to be reduced to get acceptable solutions. The Adams methods with adapted step sizes can be applied successfully to this problem with much larger values of the time step $\Delta t$, in accordance with the analysis of Sect. 4.

### 6.4 Korteweg–de Vries equation

The Korteweg–de Vries (KdV) equation

$$\partial_t u(t,x) + \partial_x \frac{u(t,x)^2}{2} + \partial_x^3 u(t,x) = 0 \tag{67}$$

is well-known in the literature as a nonlinear PDE which admits soliton solutions of the form

$$u(t,x) = A \cosh\left(\sqrt{3A}(x - ct - \mu)/6\right)^{-2}, \quad c = A/3, \tag{68}$$

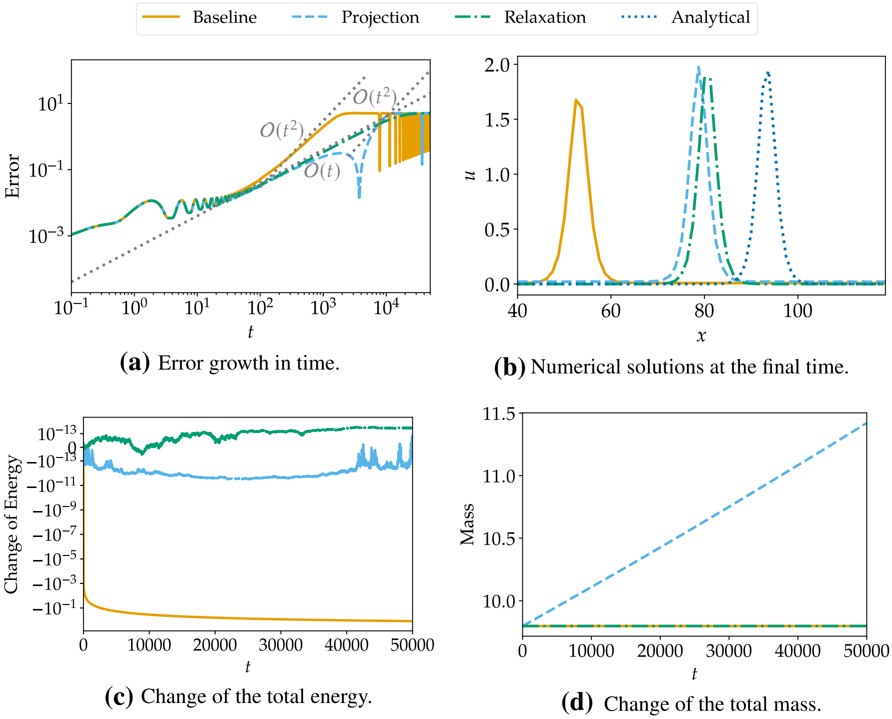
where $A \geq 0$ is the amplitude, $c$ the wave speed, and $\mu$ an arbitrary constant. The KdV equation possesses an infinite hierarchy of conserved integral functionals, including the mass (with density $u$) and the energy (with density $u^2$).

Numerical methods that conserve both the mass and the energy result in an asymptotic error growth that is only linear in time, while other methods will in general yield an asymptotically quadratic error growth [14]. If only the energy is conserved, the error is usually reduced at first and the quadratic error growth can be seen later than for methods that do not conserve the energy.

We choose this test problem because it is a stiff nonlinear problem. Hence, we use implicit methods to deal with the stiffness. In addition, this problem demonstrates improved qualitative properties of conservative schemes and the importance of preserving linear invariants. Moreover, this stiff problem demonstrates that important stability properties are not lost by introducing relaxation, even for robust, $A$-stable methods.

Here, we use the mass- and energy-conservative Fourier collocation semidiscretization described in [50] with $N = 64$ modes in an interval of length $L = 80$ for the amplitude $A = 2$. Integrating the resulting stiff ODE in time with the BDF(2) method and a time step $\Delta t = 0.1$ yields the results shown in Fig. 7. The error for both the projection and the relaxation grows linearly in time at first. For the projection method not conserving the total mass, the error starts to grow quadratically shortly before it saturates (since there is no overlap of the numerical solution and the analytical solution anymore).

We would like to point out that we considered a long-time integration for this stiff nonlinear PDE example. The (phase) error grows in time and will reach 100° for every time step at some time (which increases for decreasing step size $\Delta t$). In particular, having an error of 100° after more than 400 periods of the traveling wave solution does not indicate instabilities caused choosing the time step $\Delta t$ too big. Instead, this behavior is expected and occurs also for smaller time steps (possibly after more periods).



**(a)** Error growth in time.

**(b)** Numerical solutions at the final time.

**(c)** Change of the total energy.

**(d)** Change of the total mass.

**Fig. 7** Numerical solutions of the KdV equation (67) with final time $t = 5.0 \times 10^4$ and projection/relaxation methods conserving the energy applied to a mass- and energy-conservative Fourier collocation method. The baseline method is BDF2

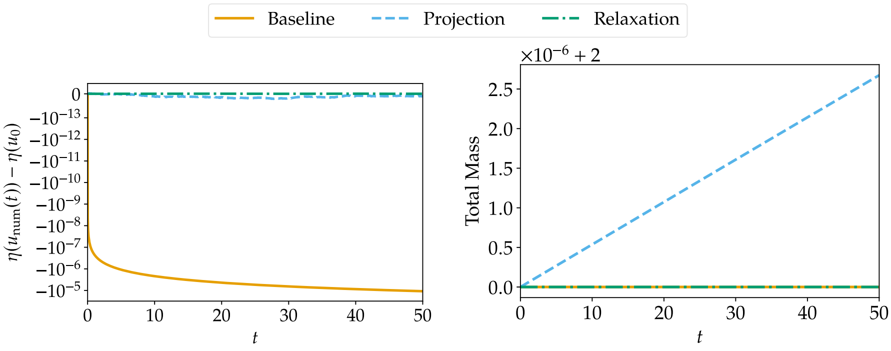
### 6.5 Compressible Euler equations

Here, we apply a second-order entropy-conservative finite difference method [68] using the entropy-conservative numerical flux of [43, Theorem 7.8] for the compressible Euler equations of an ideal gas in one space dimension. The initial condition

$$\varrho_0(x) = 1 + \frac{1}{2}\sin(\pi x), \quad v_0(x) = 1, \quad p_0(x) = 1, \tag{69}$$

where $\varrho$ is the density, $v$ the velocity, and $p$ the pressure, results in a smooth and entropy-conservative solution in the periodic domain $[0, 2]$. Integrating the entropy-conservative semidiscretization with $N = 100$ grid nodes in time with SSP(4, 3), where the starting values have been obtained with the relaxation version of the classical third-order, three-stage SSP Runge–Kutta method, yields the results shown in Fig. 8. Clearly, the baseline scheme is not entropy-conservative while the projection method does not conserve the total mass. In contrast, the relaxation method conserves both functionals.

We choose this test problem because it is a non-stiff nonlinear PDE problem where the preservation of linear invariants is particularly interesting. We choose SSP methods



**(a)** Total entropy change.

**(b)** Total mass.

**Fig. 8** Numerical solutions of the compressible Euler equations with final time $t = 50$ and time step $\Delta t = 0.1\Delta x$ using entropy-conservative finite differences and projection/relaxation versions of SSP(4, 3)

since these are often applied to computational fluid dynamics; results for other time integration methods are similar since the SSP property is not crucial here.

### 6.6 Burgers' equation

Solutions of Burgers' equation

$$\partial_t u(t,x) + \partial_x \frac{u(t,x)^2}{2} = 0, \quad u(0,x) = \exp(-30x^2), \tag{70}$$

in the periodic domain $[-1, 1]$ develop shocks in a finite time. Hence, energy-conservative methods are not appropriate. Here, we apply the same energy-dissipative semidiscretization used in [31] in the context of relaxation Runge–Kutta methods, which can be written as

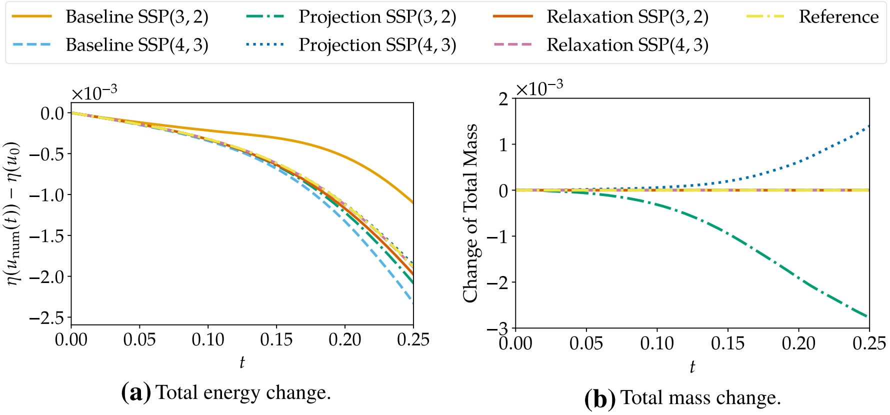
$$\frac{\mathrm{d}}{\mathrm{d}t} u_i(t) = -\frac{f^{\mathrm{num}}(u_i, u_{i+1}) - f^{\mathrm{num}}(u_{i-1}, u_i)}{\Delta x}. \tag{71}$$

The energy-dissipative numerical flux is obtained by adding some dissipation to the energy-conservative flux, resulting in

$$f^{\mathrm{num}}(u_-, u_+) = \frac{u_-^2 + u_- u_+ + u_+^2}{6} - \varepsilon(u_+ - u_-). \tag{72}$$

These semidiscretizations are integrated in time with SSP(3, 2) and SSP(4, 3), where the starting values have been obtained with the relaxation version of the classical third order, three-stage SSP Runge–Kutta method.

We choose this test problem because it is a non-stiff nonlinear PDE problem with decaying energy to demonstrate improved qualitative behavior also in this context. We choose SSP methods again since these are often applied to computational fluid dynamics problems.



**(a)** Total energy change.

**(b)** Total mass change.

**Fig. 9** Numerical solutions of Burgers' equation with final time $t = 0.25$ and time step $\Delta t = 0.2\Delta x$ using energy-dissipative finite differences and projection/relaxation versions of SSP(3, 2) and SSP(4, 3)

The changes of the total energy and mass of the numerical solutions and a semidiscrete reference solution are visualized in Fig. 9. The baseline schemes are either anti-dissipative [for SSPRK(3, 2)] or too dissipative [for SSPRK(4, 3)] compared to the reference solution, similarly to results for Runge–Kutta methods shown in [31]. In contrast, the energy dissipation of both the relaxation and the projection versions agrees very well with the reference solution. However, the projection schemes change the total mass while the relaxation methods conserve this invariant of the PDE (70).

Adding instead some dissipation to a semidiscretization based on the central numerical flux

$$f^{\mathrm{num}}(u_-, u_+) = \frac{u_-^2 + u_+^2}{4} - \varepsilon(u_+ - u_-) \tag{73}$$

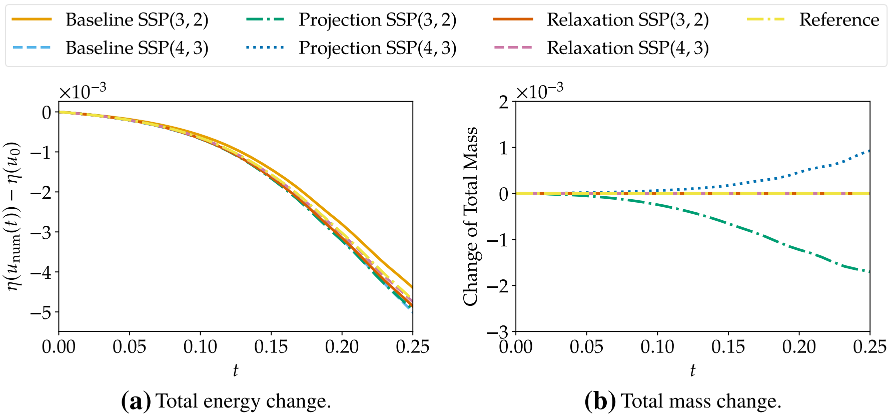
does not yield a provably energy-dissipative semidiscretization in general. However, the relaxation methods still improve the energy evolution as visualized in Fig. 10.

### 6.7 Linear advection with inflow

Solutions of the linear advection equation

$$\begin{aligned} \partial_t u(t,x) + \partial_x u(t,x) &= 0, && t \in (0,6),\ x \in (0,3), \\ u(0,x) &= 0, && x \in [0,3], \\ u(t,0) &= \sin(\pi t), && t \in [0,6], \end{aligned} \tag{74}$$

do neither conserve nor dissipate the energy $\frac{1}{2}\|u\|_{L^2}^2$ because of the boundary condition. Instead, the energy of the analytical solution increases till $t = 3$ to its final value 0.75 and stays constant thereafter. We choose this test problem because it results in a non-monotone behavior of the energy. We choose SSP methods again since these are often applied to computational fluid dynamics problems.



**(a)** Total energy change. **(b)** Total mass change.

**Fig. 10** Numerical solutions of Burgers' equation with final time $t = 0.25$ and time step $\Delta t = 0.2\Delta x$ using central finite differences with dissipation and projection/relaxation versions of SSP(3, 2) and SSP(4, 3)

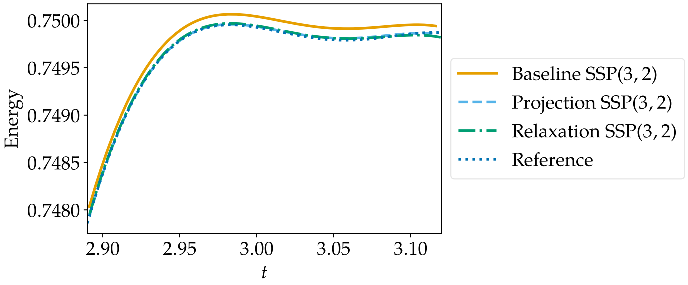


**Fig. 11** Energy of numerical solutions of the linear advection equation (74) computed with or without projection/relaxation methods and SSP(3, 2)

Using the classical second-order summation-by-parts operator with simultaneous approximation terms to impose the boundary condition weakly [66] on $N = 200$ uniformly spaced nodes yields an ODE with a similar behavior. As visualized in Fig. 11, the baseline SSP(3, 2) method results in an increase of the energy that is slightly bigger than that of the reference solution obtained by SSP(4, 3) with much smaller time steps. Instead, the energy variation enforced by the projection and relaxation methods is visually indistinguishable from the reference value. The slight variations of the energy for $t > 3$ are caused by the spatial semidiscretizations using a weak imposition of the boundary condition.

This example demonstrates that relaxation methods can also be useful for non-dissipative systems where energy or entropy estimates can still be obtained and are of interest. In [53], a similar lid-driven cavity flow with a heated wall for the Navier–Stokes equations has been solved with relaxation Runge–Kutta methods.

### 6.8 Some remarks on the costs of relaxation

For a quadratic or cubic entropy functional, the relaxation parameter can be computed explicitly. For more general entropies, it must be found via numerical iteration.

We have used SciPy [69] to solve the scalar equations for the relaxation parameter $\gamma^n$. Our implementations are based on functions written in pure Python and are not adapted to the specific problems. A detailed assessment of the computational cost of solving for the relaxation parameter $\gamma^n$ requires a more refined implementation and is beyond the scope of this work. Nevertheless, we give some preliminary discussion of the costs here, emphasizing that these numbers should be viewed as very crude upper bounds on the cost associated to relaxation.

For explicit time-stepping methods and very inexpensive right-hand sides, the costs of naive implementations of the relaxation approach are significant. For the example of the compressible Euler equations in Sect. 6.5, relaxation methods increase the runtime by a factor between two and three. For Burgers' equation as in Sect. 6.6, the total runtime increases by less than a factor of two. Although the relaxation technique increases the runtime significantly in this naive implementation, it is still cheaper than decreasing the time step to get basically the same results.

Relaxation methods were applied to large-scale PDE discretizations of the compressible Euler and Navier-Stokes equations in three space dimensions in [46,53]. In that context, the cost associated with solving the single scalar equation for the relaxation parameter $\gamma^n$ is much less than the cost of evaluating the right-hand side of the ODE during one time step.

When implicit time-stepping is used, the cost of relaxation is less significant. This was already discussed in [50] for the KdV equation. There, the energy-conserving methods decreased the runtime despite the added overhead of computing the relaxation or projection step. The reason for this is the improved accuracy and hence decreased costs to solve the nonlinear stage equations. The relaxation method benefited additionally from taking larger effective time steps since the relaxation parameter $\gamma^n > 1$.

## 7 Summary and conclusions

We have extended the framework of relaxation methods for the numerical solution of initial-value problems from Runge–Kutta methods to general time integration schemes with order of accuracy $p \geq 2$. By solving a single scalar algebraic equation per time step, the evolution in time of a given functional can be preserved. This includes functionals that are conserved or dissipated, as well as others for which estimates of the time evolution are available. For convex functionals, additional insights such as the possibility to add dissipation in time have been provided.

For certain classes of relaxation linear multistep methods, high-order accuracy is still attained even if a fixed-coefficient method is used. Nevertheless, we recommend the use of methods that correctly account for the step size variation, since such methods gave overall better results in numerical tests.

In contrast to orthogonal projection methods, relaxation methods preserve all linear invariants that are preserved by the baseline time integration scheme (which are all linear invariants of the ODE for general linear methods). This property can be very important, e.g., for conservation laws.

We have also studied the impact of the relaxation approach on other stability properties of time integration methods. In particular, zero stability and strong stability preserving properties of linear multistep and Runge–Kutta methods are not changed significantly.

While relaxation methods appear to provide good results in our numerical experiments, further practical experience on a wide range of problems is still needed to determine their general effectiveness. Other areas of ongoing research include the development of other means to estimate the change of a dissipated functional $\eta$ and development of a spatially localized relaxation approach for conservation laws with convex entropies.

**Acknowledgements** Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST). The project "Application-domain-specific highly reliable IT solutions" has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, and financed under the scheme Thematic Excellence Programme no. 2020-4.1.1-TKP2020 (National Challenges Subprogramme).

### Compliance with ethical standards

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

1. Arévalo, C., Führer, C., Söderlind, G.: Regular and singular $\beta$-blocking of difference corrected multistep methods for nonstiff index-2 DAEs. Appl. Numer. Math. **35**(4), 293–305 (2000). https://doi.org/10.1016/S0168-9274(99)00142-7
2. Arévalo, C., Söderlind, G.: Grid-independent construction of multistep methods. J. Comput. Math. **35**(5), 672–692 (2017). https://doi.org/10.4208/jcm.1611-m2015-0404
3. Bashforth, F., Adams, J.C.: An Attempt to Test the Theories of Capillary Action by Comparing the Theoretical and Measured Forms of Drops of Fluid with an Explanation of the Method of Integration Employed in Constructing the Tables Which Give the Theoretical Forms of Such Drops. Cambridge University Press, Cambridge (1883)
4. Boom, P.D., Zingg, D.W.: High-order implicit time-marching methods based on generalized summation-by-parts operators. SIAM J. Sci. Comput. **37**(6), A2682–A2709 (2015). https://doi.org/10.1137/15M1014917
5. Burchard, H., Deleersnijder, E., Meister, A.: A high-order conservative Patankar-type discretisation for stiff systems of production-destruction equations. Appl. Numer. Math. **47**(1), 1–30 (2003). https://doi.org/10.1016/S0168-9274(03)00101-6
6. Burrage, K., Butcher, J.C.: Stability criteria for implicit Runge–Kutta methods. SIAM J. Numer. Anal. **16**(1), 46–57 (1979). https://doi.org/10.1137/0716004
7. Burrage, K., Butcher, J.C.: Non-linear stability of a general class of differential equation methods. BIT Numer. Math. **20**(2), 185–203 (1980). https://doi.org/10.1007/BF01933191
8. Butcher, J.C.: Numerical Methods for Ordinary Differential Equations. John Wiley & Sons Ltd, Chichester (2016). 10.1002/9781119121534

9. Calvo, M., Hernández-Abreu, D., Montijano, J.I., Rández, L.: On the preservation of invariants by explicit Runge–Kutta methods. SIAM J. Sci. Comput. **28**(3), 868–885 (2006). https://doi.org/10.1137/04061979X
10. Calvo, M., Laburta, M., Montijano, J., Rández, L.: Projection methods preserving Lyapunov functions. BIT Numer. Math. **50**(2), 223–241 (2010). https://doi.org/10.1007/s10543-010-0259-3
11. Chan, J.: On discretely entropy conservative and entropy stable discontinuous Galerkin methods. J. Comput. Phys. **362**, 346–374 (2018). https://doi.org/10.1016/j.jcp.2018.02.033
12. Curtiss, C.F., Hirschfelder, J.O.: Integration of stiff equations. Proc. Natl. Acad. Sci. U. S. A. **38**(3), 235 (1952). https://doi.org/10.1073/pnas.38.3.235
13. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics. Springer-Verlag, Berlin (2010). https://doi.org/10.1007/978-3-642-04048-1
14. De Frutos, J., Sanz-Serna, J.M.: Accuracy and conservation properties in numerical integration: the case of the Korteweg–de Vries equation. Numer. Math. **75**(4), 421–445 (1997). https://doi.org/10.1007/s002110050247
15. Dekker, K., Verwer, J.G.: Stability of Runge–Kutta methods for stiff nonlinear differential equations, CWI Monographs, vol. 2. North-Holland, Amsterdam (1984)
16. Eich, E.: Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints. SIAM J. Numer. Anal. **30**(5), 1467–1482 (1993). https://doi.org/10.1137/0730076
17. Fisher, T.C., Carpenter, M.H.: High-order entropy stable finite difference schemes for nonlinear conservation laws: finite domains. J. Comput. Phys. **252**, 518–557 (2013). https://doi.org/10.1016/j.jcp.2013.06.014
18. Friedrich, L., Schnücke, G., Winters, A.R., Fernández, D.C.D.R., Gassner, G.J., Carpenter, M.H.: Entropy stable space-time discontinuous Galerkin schemes with summation-by-parts property for hyperbolic conservation laws. J. Sci. Comput. **80**(1), 175–222 (2019). https://doi.org/10.1007/s10915-019-00933-2. arxiv: 1808.08218 [math.NA]
19. Friedrich, L., Winters, A.R., Fernández, D.C.D.R., Gassner, G.J., Parsani, M., Carpenter, M.H.: An entropy stable h/p non-conforming discontinuous Galerkin method with the summation-by-parts property. J. Sci. Comput. **77**(2), 689–725 (2018). https://doi.org/10.1007/s10915-018-0733-7
20. Gear, C.W.: Maintaining solution invariants in the numerical solution of ODEs. SIAM J.Sci. Stat. Comput. **7**(3), 734–743 (1986). https://doi.org/10.1137/0907050
21. Glaubitz, J., Öffner, P., Ranocha, H., Sonar, T.: Artificial viscosity for correction procedure via reconstruction using summation-by-parts operators. In: C. Klingenberg, M. Westdickenberg (eds.) Theory, Numerics and Applications of Hyperbolic Problems II, In: Springer Proceedings in Mathematics and Statistics, vol. 237, pp. 363–375. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-91548-7_28
22. Gottlieb, S., Ketcheson, D.I., Shu, C.W.: Strong Stability Preserving Runge–Kutta and Multistep Time Discretizations. World Scientific, Singapore (2011)
23. Grimm, V., Quispel, G.: Geometric integration methods that preserve Lyapunov functions. BIT Numer. Math. **45**(4), 709–723 (2005). https://doi.org/10.1007/s10543-005-0034-z
24. Hadjimichael, Y., Ketcheson, D.I., Lóczi, L., Németh, A.: Strong stability preserving explicit linear multistep methods with variable step size. SIAM J. Numer. Anal. **54**(5), 2799–2832 (2016). https://doi.org/10.1137/15M101717X
25. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, Springer Series in Computational Mathematics, vol. 31. Springer-Verlag, Berlin (2006). https://doi.org/10.1007/3-540-30666-8
26. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems, Springer Series in Computational Mathematics, vol. 8. Springer-Verlag, Berlin (2008). https://doi.org/10.1007/978-3-540-78862-1
27. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, Springer Series in Computational Mathematics, vol. 14. Springer-Verlag, Berlin (2010). https://doi.org/10.1007/978-3-642-05221-7
28. Higueras, I.: Monotonicity for Runge–Kutta methods: inner product norms. J. Sci. Comput. **24**(1), 97–117 (2005). https://doi.org/10.1007/s10915-004-4789-1
29. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55

30. Jüngel, A., Schuchnigg, S.: Entropy-dissipating semi-discrete Runge–Kutta schemes for nonlinear diffusion equations. Commun. Math. Sci. **15**(1), 27–53 (2017). https://doi.org/10.4310/CMS.2017.v15.n1.a2
31. Ketcheson, D.I.: Relaxation Runge–Kutta methods: conservation and stability for inner-product norms. SIAM J. Numer. Anal. **57**(6), 2850–2870 (2019). https://doi.org/10.1137/19M1263662. arxiv:1905.09847 [math.NA]
32. Kojima, H.: Invariants preserving schemes based on explicit Runge–Kutta methods. BIT Numer. Math. **56**(4), 1317–1337 (2016). https://doi.org/10.1007/s10543-016-0608-y
33. Laburta, M., Montijano, J.I., Rández, L., Calvo, M.: Numerical methods for non conservative perturbations of conservative problems. Comput. Phys. Commun. **187**, 72–82 (2015). https://doi.org/10.1016/j.cpc.2014.10.012
34. LeFloch, P.G., Mercier, J.M., Rohde, C.: Fully discrete, entropy conservative schemes of arbitrary order. SIAM J. Numer. Anal. **40**(5), 1968–1992 (2002). https://doi.org/10.1137/S003614290240069X
35. LeVeque, R.J.: Finite Difference Methods for Ordinary and Partial Differential Equations: steady-state and time-dependent problems. SIAM, Philadelphia (2007)
36. Lozano, C.: Entropy production by explicit Runge–Kutta schemes. J. Sci. Comput. **76**(1), 521–565 (2018). https://doi.org/10.1007/s10915-017-0627-0
37. Lozano, C.: Entropy production by implicit Runge–Kutta schemes. J.Sci. Comput. (2019). https://doi.org/10.1007/s10915-019-00914-5
38. Mohammadi, F., Arévalo, C., Führer, C.: A polynomial formulation of adaptive strong stability preserving multistep methods. SIAM J. Numer. Anal. **57**(1), 27–43 (2019). https://doi.org/10.1137/17M1158811
39. Nordström, J., La Cognata, C.: Energy stable boundary conditions for the nonlinear incompressible Navier–Stokes equations. Math. Comput. **88**(316), 665–690 (2019). https://doi.org/10.1090/mcom/3375
40. Nyström, E.J.: Über die numerische integration von differentialgleichungen. Acta Soc. Sci. Fennicae **50**(13), 1–56 (1925)
41. Öffner, P., Glaubitz, J., Ranocha, H.: Analysis of artificial dissipation of explicit and implicit time-integration methods. Int. J. Numer. Anal. Model. **17**(3), 332–349 (2020). arxiv:1609.02393 [math.NA]
42. Ranocha, H.: Comparison of some entropy conservative numerical fluxes for the Euler equations. J. Sci. Comput. **76**(1), 216–242 (2018). https://doi.org/10.1007/s10915-017-0618-1. arxiv:1701.02264 [math.NA]
43. Ranocha, H.: Generalised summation-by-parts operators and entropy stability of numerical methods for hyperbolic balance laws. Ph.D. thesis, TU Braunschweig (2018)
44. Ranocha, H.: Some notes on summation by parts time integration methods. Results Appl. Math. **1**, 100,004 (2019). https://doi.org/10.1016/j.rinam.2019.100004. arxiv:1901.08377 [math.NA]
45. Ranocha, H.: On strong stability of explicit Runge–Kutta methods for nonlinear semibounded operators. IMA Journal of Numerical Analysis (2020). https://doi.org/10.1093/imanum/drz070. arxiv:1811.11601 [math.NA]
46. Ranocha, H., Dalcin, L., Parsani, M.: Fully-discrete explicit locally entropy-stable schemes for the compressible Euler and Navier–Stokes equations. Comput. Math. Appl. **80**(5), 1343–1359 (2020). https://doi.org/10.1016/j.camwa.2020.06.016
47. Ranocha, H., Glaubitz, J., Öffner, P., Sonar, T.: Stability of artificial dissipation and modal filtering for flux reconstruction schemes using summation-by-parts operators. Appl. Numer. Math. **128**, 1–23 (2018). https://doi.org/10.1016/j.apnum.2018.01.019. arXiv:1606.00995 [math.NA] and arXiv:1606.01056 [math.NA]
48. Ranocha, H., Ketcheson, D.I.: Energy stability of explicit Runge-Kutta methods for non-autonomous or nonlinear problems (2020). Accepted in SIAM Journal on Numerical Analysis. arxiv:1909.13215 [math.NA]
49. Ranocha, H., Ketcheson, D.I.: Relaxation-LMM-notebooks. General relaxation methods for initial-value problems with application to multistep schemes. https://github.com/ranocha/Relaxation-LMM-notebooks (2020). https://doi.org/10.5281/zenodo.3697836
50. Ranocha, H., Ketcheson, D.I.: Relaxation Runge–Kutta methods for Hamiltonian problems. J. Sci. Comput. **84**(1), (2020). https://doi.org/10.1007/s10915-020-01277-y arxiv:2001.04826 [math.NA]
51. Ranocha, H., Lóczi, L., Ketcheson, D.I.: General relaxation methods for initial-value problems with application to multistep schemes (2020). arxiv:2003.03012 [math.NA]

52. Ranocha, H., Öffner, P.: $L_2$ stability of explicit Runge–Kutta schemes. J. Sci. Comput. **75**(2), 1040–1056 (2018). https://doi.org/10.1007/s10915-017-0595-4
53. Ranocha, H., Sayyari, M., Dalcin, L., Parsani, M., Ketcheson, D.I.: Relaxation Runge–Kutta methods: fully-discrete explicit entropy-stable schemes for the compressible Euler and Navier–Stokes equations. SIAM J. Sci. Comput. **42**(2), A612–A638 (2020). https://doi.org/10.1137/19M1263480. arxiv:1905.09129 [math.NA]
54. Rosenlicht, M.: Introduction to Analysis. Dover Publications Inc, New York (1986)
55. Ruuth, S.J., Hundsdorfer, W.: High-order linear multistep methods with general monotonicity and boundedness properties. J. Comput. Phys. **209**(1), 226–248 (2005). https://doi.org/10.1016/j.jcp.2005.02.029
56. Sanz-Serna, J.M.: An explicit finite-difference scheme with exact conservation properties. J. Comput. Phys. **47**(2), 199–210 (1982). https://doi.org/10.1016/0021-9991(82)90074-2
57. Sanz-Serna, J.M., Calvo, M.P.: Numerical Hamiltonian Problems, Applied Mathematics and Mathematical Computation, vol. 7. Chapman & Hall, London (1994)
58. Sanz-Serna, J.M., Manoranjan, V.: A method for the integration in time of certain partial differential equations. J. Comput. Phys. **52**(2), 273–289 (1983). https://doi.org/10.1016/0021-9991(83)90031-1
59. Shampine, L.F.: Conservation laws and the numerical solution of ODEs. Comput. Math. Appl. **12**(5–6), 1287–1296 (1986). https://doi.org/10.1016/0898-1221(86)90253-1
60. Shampine, L.F.: Conservation laws and the numerical solution of ODEs, II. Comput. Math. Appl. **38**(2), 61–72 (1999). https://doi.org/10.1016/S0898-1221(99)00183-2
61. Sjögreen, B., Yee, H.: High order entropy conservative central schemes for wide ranges of compressible gas dynamics and MHD flows. J. Comput. Phys. **364**, 153–185 (2018). https://doi.org/10.1016/j.jcp.2018.02.003
62. Söderlind, G., Fekete, I., Faragó, I.: On the zero-stability of multistep methods on smooth nonuniform grids. BIT Numer. Math. **58**(4), 1125–1143 (2018). https://doi.org/10.1007/s10543-018-0716-y
63. Sun, Z., Shu, C.W.: Stability of the fourth order Runge–Kutta method for time-dependent partial differential equations. Ann. Math. Sci. Appl. **2**(2), 255–284 (2017). https://doi.org/10.4310/AMSA.2017.v2.n2.a3
64. Sun, Z., Shu, C.W.: Enforcing strong stability of explicit Runge–Kutta methods with superviscosity (2019). arxiv:1912.11596 [math.NA]
65. Sun, Z., Shu, C.W.: Strong stability of explicit Runge–Kutta time discretizations. SIAM J. Numer. Anal. **57**(3), 1158–1182 (2019). https://doi.org/10.1137/18M122892X. arxiv:1811.10680 [math.NA]
66. Svärd, M., Nordström, J.: Review of summation-by-parts schemes for initial-boundary-value problems. J. Comput. Phys. **268**, 17–38 (2014). https://doi.org/10.1016/j.jcp.2014.02.031
67. Tadmor, E.: From semidiscrete to fully discrete: stability of Runge–Kutta schemes by the energy method II. In: Estep, D.J., Tavener, S. (eds.) Collected Lectures on the Preservation of Stability under Discretization, Proceedings in Applied Mathematics, vol. 109, pp. 25–49. Society for Industrial and Applied Mathematics, Philadelphia (2002)
68. Tadmor, E.: Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. Acta Numer. **12**, 451–512 (2003). https://doi.org/10.1017/S0962492902000156
69. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E.W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0 – fundamental algorithms for scientific computing in python (2019). arxiv:1907.10121 [cs.MS]
70. Zakerzadeh, H., Fjordholm, U.S.: High-order accurate, fully discrete entropy stable schemes for scalar conservation laws. IMA J. Numer. Anal. **36**(2), 633–654 (2016). https://doi.org/10.1093/imanum/drv020

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.