Journal of Computational Physics 295 (2015) 209–229

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

An efficient time advancing strategy for energy-preserving simulations

F. Capuano [a,b], G. Coppola [a,*], L. de Luca [a]

[a] Dipartimento di Ingegneria Industriale (DII), Università di Napoli "Federico II", Napoli, 80125, Italy
[b] Centro Italiano Ricerche Aerospaziali (CIRA), Capua, 81043, Italy

ARTICLE INFO

Article history:
Received 16 August 2014
Received in revised form 16 March 2015
Accepted 21 March 2015
Available online 13 April 2015

Keywords:
Runge–Kutta
Burgers' equation
Energy conservation
Skew-symmetric form

ABSTRACT

Energy-conserving numerical methods are widely employed within the broad area of convection-dominated systems. Semi-discrete conservation of energy is usually obtained by adopting the so-called skew-symmetric splitting of the non-linear convective term, defined as a suitable average of the divergence and advective forms. Although generally allowing global conservation of kinetic energy, it has the drawback of being roughly twice as expensive as standard divergence or advective forms alone. In this paper, a general theoretical framework has been developed to derive an efficient time-advancement strategy in the context of explicit Runge–Kutta schemes. The novel technique retains the conservation properties of skew-symmetric-based discretizations at a reduced computational cost. It is found that optimal energy conservation can be achieved by properly constructed Runge–Kutta methods in which only divergence and advective forms for the convective term are used. As a consequence, a considerable improvement in computational efficiency over existing practices is achieved. The overall procedure has proved to be able to produce new schemes with a specified order of accuracy on both solution and energy. The effectiveness of the method as well as the asymptotic behavior of the schemes is demonstrated by numerical simulation of Burgers' equation.

1. Introduction

Over the past few decades, substantial research efforts have been devoted to the construction of numerical methods mimicking fundamental properties of the underlying mathematical/physical system. These so-called *physics-compatible* discretizations have gained increasing popularity over the years, especially in numerical simulations of turbulent flows [1]. In this last context, energy-preserving numerical methods are usually the preferred choice, as they provide a natural stability bound over long-time integration. Moreover, being free of numerical diffusion, they ensure that the energy cascade is not artificially contaminated, in fully-resolved computations, and that the contribution of subgrid-scale motions is entirely modeled, in under-resolved cases [2].

Finite difference and spectral energy-conserving schemes found in literature are often built upon the skew-symmetric splitting of the convective term, which is defined as an average of the divergence and the advective forms [3,4]. When coupled to operators satisfying a discrete summation-by-parts rule (i.e., centered finite-difference and spectral schemes), skew-symmetric methods ensure semidiscrete global conservation of energy for incompressible flows in the inviscid limit,

* Corresponding author.
E-mail address: gcoppola@unina.it (G. Coppola).

http://dx.doi.org/10.1016/j.jcp.2015.03.070
0021-9991/

and prevent spurious production or dissipation of kinetic energy by convection for compressible flows [5]. The skew-symmetric splitting yielded relatively stable simulations for both incompressible [6] and compressible [7] turbulence, and proved to be also beneficial in reducing the amplitude of aliasing errors [8].

Despite its remarkable features, one relevant drawback of the skew-symmetric form is that its computation is roughly twice as expensive as standard divergence or advective forms alone [9]. Moreover, full energy conservation (i.e., in time as well as in space) requires the use of costly implicit time-stepping methods, otherwise a (typically dissipative) error is introduced regardless of the spatial scheme. Finally, fully-conservative methods are not always advantageous as they can display aliasing issues for under-resolved computations without numerical regularization [10]. A trade-off between cost-effectiveness and conservation properties is thus warranted.

The additional expense related to the use of the skew-symmetric form has been mentioned by many authors (e.g. [11]). Most of the attempts appeared so far to achieve cost-effective implementations are simply based on using the advective and divergence form at alternate time steps [9,12]. In this paper, a novel time-advancement strategy that mimics the conservation properties of skew-symmetric-based schemes at a reduced computational cost is presented. It is found that optimal energy-conservation properties can be achieved by properly constructed Runge–Kutta schemes in which a different form (advective or divergence) for the convective term is adopted at each stage. This splitting strategy is able to reproduce the effects of the skew-symmetric form on energy conservation, up to a specific order of accuracy. The main achievement is that, since the method is based only on advective and divergence forms, it can be considerably faster than skew-symmetric-based techniques.

The analysis is conducted by considering the inviscid Burgers' equation, which is a well-known prototype equation of considerable physical and mathematical interest. The reason for this choice stems from the fact that, while the general idea can be extended to more complete models, the detailed analysis on convergence and orders of accuracy requires some analytical developments which are, in the first instance, more neatly conducted on a model equation. In this respect, the Burgers' equation can be considered as the simplest partial differential equation reproducing some of the peculiar features of nonlinear convective transport terms present in more realistic models (e.g., the Navier–Stokes equations). Hence, besides the intrinsic value of the application of the present theory to Burgers' equation, the present analysis is intended also as a first step towards the application to more complex systems.

The paper is organized as follows. In Section 2, the discrete energy conservation properties of both spatial and temporal discretizations are reviewed. In Section 3, a first, simple approach to obtain a cost-effective energy-conserving method is analyzed. The theoretical development for energy-preserving Runge–Kutta schemes is presented in Section 4. Results are shown in Section 5. In Section 6, some indications are given for the extension of the main idea to the incompressible Navier–Stokes equations. Finally, Section 7 presents a summary.

2. Conservation properties of semi-discretized equations

2.1. Spatial discretization

Energy-conservation properties of spatial discretizations of nonlinear convective terms can be investigated by considering the inviscid Burgers' equation, which can be formally written as

$$\partial_t u + \mathcal{N}(u) = 0, \tag{1}$$

where the non-linear convective term $\mathcal{N}(u)$ can be expressed in one of the equivalent forms $u\partial_x u$ (*advective*), $\partial_x u^2/2$ (*divergence*) or $(\partial_x u^2 + u\partial_x u)/3$ (*skew-symmetric*). When posed on an interval $[a, b]$ with periodic boundary conditions, Eq. (1) has solutions for which all the moments are conserved:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \frac{u^n}{n}\,dx = \int_a^b u^{n-1}\frac{\partial u}{\partial t}\,dx = -\int_a^b u^n \frac{\partial u}{\partial x}\,dx = -\int_{u(a)}^{u(b)} u^n\,du = 0.$$

Specifically, the total momentum and the total energy of the solution (which are obtained in the cases $n = 1$ and $n = 2$, respectively) remain fixed to their initial values during time evolution.

Although in the continuous case the different forms of the convective term are mathematically equivalent, their spatially discretized counterparts can behave very differently, especially in terms of energy-preserving features and aliasing errors. This is due to the fact that discrete operators generally are not guaranteed to correctly reproduce the numerical equivalents of integration by parts and differentiation chain rule (cf. [4]). The beneficial conservation and aliasing properties of the skew-symmetric form have long been recognized by many authors [4,9,13], while there is much more debate about the other two formulations. The present analysis examines the energy-conserving behavior of the spatially discretized version of Eq. (1) when the different forms are employed.

To pursue the scope, the semi-discretized version of the Burgers' equation is considered, which can be expressed by introducing the matrix $\mathbf{C}(\mathbf{u})$ as

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} + \mathbf{C}(\mathbf{u})\mathbf{u} = 0. \tag{2}$$

In this equation $\mathbf{u}$ is the vector of the nodal values of u, i.e. $u_i(t) = u(x_i, t)$ and $\mathbf{C}(\mathbf{u})$ can assume one of the following forms, $\mathbf{C}^{\text{adv}} = \mathbf{UD}$, $\mathbf{C}^{\text{div}} = \frac{1}{2}\mathbf{DU}$ or $\mathbf{C}^{\text{skw}} = (\mathbf{DU} + \mathbf{UD})/3$, where $\mathbf{U} = \text{diag}(\mathbf{u})$ and $\mathbf{D}$ is the derivative matrix associated with the spatial discretization. Hereinafter, the analysis will be conducted by considering uniform meshes and periodic boundary conditions. For spectral methods and centered finite difference schemes, both explicit and compact, the derivative matrix turns out to be skew-symmetric.

The conservation properties of the discretized equation (2) can be inferred by considering the induced equations for the evolution of the discrete counterpart of integral, i.e. the scalar product. By premultiplying Eq. (2) by $\mathbf{1}^T$, where $\mathbf{1}$ is the column vector of all ones (discrete integrator on uniform mesh), it can be shown that, for skew-symmetric derivative matrices, the total momentum p is conserved, for both $\mathbf{C}^{\text{adv}}$ and $\mathbf{C}^{\text{div}}$. From this result one easily concludes that every discretization employing a linear convex combination of the advective and divergence forms (e.g. the skew-symmetric form) conserves total momentum.

As regards global energy conservation, the relevant scalar product is $\mathbf{u}^T\mathbf{u} = \|\mathbf{u}\|^2$, for which the evolution equation reads:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{u}\|^2 = -2\mathbf{u}^T\mathbf{C}(\mathbf{u})\mathbf{u}. \tag{3}$$

Since the time derivative of the energy of any solution of the semi-discretized equation (2) can be expressed as a quadratic form associated to the matrix $\mathbf{C}$, a sufficient condition for the semi-discretized equation to be energy-conserving is that the matrix $\mathbf{C}$ is skew-symmetric for every $\mathbf{u}$. Similar considerations have been recently employed for the construction of difference operators which are optimized by forcing the reproduction of crucial symmetry properties of the underlying differential operator, in spite of minimizing local truncation error [14,15].

From this analysis one immediately concludes that the semi-discretized equation (2) obtained by employing divergence or advective forms is not globally energy-conserving, since the associated matrices $\mathbf{C}^{\text{div}}$ and $\mathbf{C}^{\text{adv}}$ are in general not skew-symmetric. The averaged skew-symmetric form, on the other hand, is naturally energy-conserving in cases in which the derivative matrix $\mathbf{D}$ is skew-symmetric. This property is also readily seen to be equivalent to the numerical analogue of integration by parts.

The analysis presented above can be completed by characterizing the different errors on energy conservation introduced by the divergence and advective forms. By evaluating Eq. (3) in the two cases one obtains:

$$\frac{\mathrm{d}E_{\text{adv}}}{\mathrm{d}t} = -\mathbf{u}^T\mathbf{UDu} \tag{4}$$

$$\frac{\mathrm{d}E_{\text{div}}}{\mathrm{d}t} = -\frac{1}{2}\mathbf{u}^T\mathbf{DUu} = \frac{1}{2}\mathbf{u}^T\mathbf{UDu} \tag{5}$$

where $E = \|\mathbf{u}^2\|/2$. From Eqs. (4) and (5) it follows that:

$$\frac{\mathrm{d}E_{\text{div}}}{\mathrm{d}t} = -\frac{1}{2}\frac{\mathrm{d}E_{\text{adv}}}{\mathrm{d}t}. \tag{6}$$

Eq. (6) shows that the energy derivatives for the divergence and convective forms have opposite signs and that their relation is such that an average of the two forms with weights 2 and 1 (which defines the skew-symmetric form) furnishes exact conservation of energy:

$$\frac{\mathrm{d}E_{\text{skw}}}{\mathrm{d}t} = \frac{2\frac{\mathrm{d}E_{\text{adv}}}{\mathrm{d}t} + \frac{\mathrm{d}E_{\text{div}}}{\mathrm{d}t}}{3} = 0. \tag{7}$$

2.2. Time-advancement

In contrast to spatial discretizations, energy conservation properties of time-integration algorithms are much less discussed in literature. It is generally believed that the errors due to the spatial discretization are much larger than those coming from time-advancement, especially in numerical simulation of turbulent flows. In such cases, the time-step is usually dictated by accuracy and not by stability restrictions, leading to particularly small energy errors. Common choices for time-integration methods are the multi-step Adams–Bashforth or the multi-stage Runge–Kutta schemes. In the present paragraph, the attention will be focused on the latter.

The application of a generic Runge–Kutta scheme to the semi-discrete equation (2) reads:

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \Delta t \sum_{i=1}^{s} b_i \mathbf{C}(\mathbf{u}_i)\mathbf{u}_i \tag{8}$$

$$\mathbf{u}_i = \mathbf{u}^n - \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{C}(\mathbf{u}_j)\mathbf{u}_j, \tag{9}$$

where s is the number of stages. The coefficients a_{ij} and b_i are often arranged into the so-called *Butcher array* [16]:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \dots & \dots & a_{ss} \\ \hline & b_1 & \dots & \dots & b_s \end{array}$$

where $c_i = \sum_j a_{ij}$. Since the semi-discretized Burgers' equation constitutes a system of *autonomous* ordinary differential equations, the c_i coefficients will not be displayed hereinafter. Nonetheless, these coefficients might still come into play whenever time-dependent source terms or boundary conditions are considered. Eqs. (8)–(9) can be manipulated to give an expression for the evolution of energy in a single time step of the Runge–Kutta procedure. By defining $\Delta E = E^{n+1} - E^n$, Eqs. (8)–(9) easily lead to the relation [17]:

$$\frac{\Delta E}{\Delta t} = \underbrace{-\sum_{i=1}^{s} b_i \mathbf{u}_i^T \mathbf{C}(\mathbf{u}_i)\mathbf{u}_i}_{I} \underbrace{-\frac{\Delta t}{2}\sum_{i,j=1}^{s}\left(b_i a_{ij} + b_j a_{ji} - b_i b_j\right)\mathbf{u}_i^T \mathbf{C}^T(\mathbf{u}_i)\mathbf{C}\left(\mathbf{u}_j\right)\mathbf{u}_j}_{II}. \tag{10}$$

The two terms in the right-hand side of Eq. (10) can be defined as

I) *Spatial* error: $\sum_{i=1}^{s} b_i \mathbf{u}_i^T \mathbf{C}(\mathbf{u}_i)\mathbf{u}_i$;

II) *Temporal* error: $\frac{\Delta t}{2}\sum_{i,j=1}^{s}\left(b_i a_{ij} + b_j a_{ji} - b_i b_j\right)\mathbf{u}_i^T \mathbf{C}^T(\mathbf{u}_i)\mathbf{C}\left(\mathbf{u}_j\right)\mathbf{u}_j$.

The first one is composed by a linear combination of s different terms, each having the usual structure of a quadratic form for the convection matrix $\mathbf{C}$. Each term is identically zero if a skew-symmetric form is adopted for $\mathbf{C}$, hence the name *spatial* error appears to be appropriate.

The second quantity causing a variation of energy has a more complex structure and does not vanish in general, even in the case of skew-symmetric operators $\mathbf{C}$. It can be nullified only for suitably chosen Runge–Kutta integrators, and this justifies the term *temporal* for this part of the error. It is well known that for the temporal error to be zero one has to employ so-called *symplectic* methods, for which $b_i a_{ij} + b_j a_{ji} - b_i b_j = 0$, a constraint which cannot be satisfied by explicit methods [17].

2.3. Computational cost

The above considerations can be summarized as follows. If one employs a spatial discretization in which the (computationally more expensive) skew-symmetric form for the convective operator is adopted, the spatial error on energy conservation is absent. The temporal error due to Runge–Kutta integration is however still present and it can be shown to vanish, as Δt is reduced, at least with the same order of accuracy of the Runge–Kutta scheme employed. If one wants to completely nullify the temporal error, implicit (symplectic) methods have to be employed, thus further increasing the computational cost per time step. The employment of a non-energy-conserving spatial discretization (e.g. in divergence or advective forms), on the other hand, produces a zeroth order error on energy conservation (due to the spatial part of the error) independently of the accuracy of the Runge–Kutta procedure employed. This situation cannot be alleviated by the time integration procedure, since the error is completely due to the spatial discretization and would still be present in an exact integration of the system of ODEs.

On the other hand, the number of floating point operations required to perform a complete time-advancement step depends heavily on the form adopted for the non-linear term. The practical implementation of the skew-symmetric form in a finite-difference code is easily shown to be roughly twice as expensive as standard divergence or advective forms alone. In the framework of incompressible Navier–Stokes equations, for instance, the splitting form requires 18 derivatives evaluations, while the advective or divergence forms take 9 derivatives [7]. In 1D, both numbers are reduced by a factor of 9. Although the other modules of the overall solution algorithm can take a certain part of the total CPU time (e.g., solution of pressure equation, computation of viscous terms, etc.), the net cost increase due to the use of the skew-symmetric splitting can be noteworthy, especially for explicit time-marching algorithms or in spectral codes. By using the cost metric presented in Appendix A, it can be shown that in order to advance in time the convective term with a standard RK4 scheme and a second-order central difference scheme, 52 operations per node are required for the skew-symmetric form, while 28 operations are needed for divergence or advective forms. The difference increases as the spatial order of accuracy is increased. The objective of the present work is to investigate time-advancing strategies that are able to retain the beneficial properties of the skew-symmetric form at a reduced computational cost.

3. Alternating time advancing strategy

In Section 2, the favorable energy-conservation properties of the skew-symmetric form have been commented. Perhaps, the most useful information stemming from this analysis lies in the observation that the skew-symmetric splitting averages the energy errors introduced by the divergence and advective forms, so that they completely cancel out (cf. Eq. (7)). This interpretation naturally suggests that it is reasonable to expect a significant error reduction by an algorithm in which the advective and divergence forms are used at alternate time steps. Clearly, for infinitely small time-steps, the beneficial properties of the skew-symmetric form would be exactly recovered.

Similar approaches have been investigated by some authors in the past, but the technique does not seem to be consolidated and well understood. In the framework of nonlinear convective terms approximations, the alternating approach was employed, for instance, by Kerr [12] for a scalar equation, and later analyzed numerically by Zang [9] for the incompressible Navier–Stokes equations, showing well-behaved performances. In both cases, the alternating strategy was applied without any theoretical analysis of accuracy, and only a qualitative comparison among the various methods was given. In the following, as a prelude to the new method, an analysis of the energy conservation properties of alternating schemes with reference to Burgers' equation is presented.

By defining the scalar function $F(\mathbf{u}(t)) = -\mathbf{u}^T\mathbf{U}\mathbf{D}\mathbf{u}$, Eqs. (4) and (5) can be rewritten as:

$$\frac{\mathrm{d}E_{\mathrm{adv}}}{\mathrm{d}t} = F \tag{11}$$

$$\frac{\mathrm{d}E_{\mathrm{div}}}{\mathrm{d}t} = -\frac{1}{2}F. \tag{12}$$

Eqs. (11) and (12) can be used to evaluate the energy produced over a small time step Δt by exact time integration of Eq. (3). By employing a Taylor series expansion for E, one has in general:

$$E(t+\Delta t) - E(t) = \alpha F(\mathbf{u}(t))\,\Delta t + \alpha \left.\frac{\mathrm{d}F}{\mathrm{d}t}\right|_t \frac{\Delta t^2}{2} + \alpha \left.\frac{\mathrm{d}^2F}{\mathrm{d}t^2}\right|_t \frac{\Delta t^3}{6} + O(\Delta t^4) \tag{13}$$

with $\alpha = 1$ in the advective case and $\alpha = -1/2$ in the divergence case.

Eq. (13) represents the *spatial* energy conservation error occurring over a single time-step. The spatial error can be either entirely canceled by employing a skew-symmetric form or *minimized* — up to a certain order of accuracy — by alternating a suitable sequence of convective and divergence forms at each time step.

The variation in the energy E produced over a certain number of time increments can be calculated (as a Taylor series expansion) by simply summing the increments obtained at each time step. If an alternating procedure is employed, the correct increments relative to the form employed at each time step have to be considered. For an arbitrary number n of alternating steps, the energy variation is given by the following expression:

$$\Delta E = E^n - E^0 = \sum_{i=0}^{n-1} \alpha_i F_0 \Delta t + \sum_{i=0}^{n-1} (2i+1)\alpha_i \left.\frac{\mathrm{d}F}{\mathrm{d}t}\right|_0 \frac{\Delta t^2}{2} + \sum_{i=0}^{n-1} (3i(i+1)+1)\alpha_i \left.\frac{\mathrm{d}^2F}{\mathrm{d}t^2}\right|_0 \frac{\Delta t^3}{6} + O(\Delta t^4), \tag{14}$$

where the coefficient α has been indexed with the suffix i to highlight that a different form for the convective term is employed at each time step. Since the procedure has to be consistent with the skew-symmetric form as Δt tends to zero, one should consider an alternating protocol in which at least three integration steps are included for the minimal sequence which is repeated during time integration. For such a sequence, the first order condition $\alpha_0 + \alpha_1 + \alpha_2 = 0$ imposes the consistency condition that among the three integration time steps, two upgrades have to be performed with the divergence form and one with the advective form. By employing the notation that the letters A and D are used to denote advective and divergence forms respectively, any one of the sequences DDA, ADD and DAD produces an integration which is of first order with respect to energy conservation. The second order condition, when imposed on sequences already satisfying the first order condition, is equivalent to the relation $\alpha_2 + 2\alpha_3 = 0$ which is satisfied only by the values $\alpha_1 = 1$ and $\alpha_2 = -1/2$, corresponding to the sequence DAD. This last sequence is the only possible minimal sequence of three time steps producing a second order scheme for the conservation of energy, provided that the time integration procedure in each time step is performed at least with second order accuracy. No third-order accuracy can be achieved with a three-step alternating protocol. In order to reach higher accuracy, it can be shown that one has to consider sequences involving at least 18 alternating forms.

In Fig. 1, numerical results are presented concerning the time-evolution of energy for various sequences. Results are obtained by integrating the inviscid Burgers' equation with periodic boundary conditions on the domain $[0, 1]$, discretized by $N = 100$ equidistant mesh points, with initial conditions $u_0(x) = \sin(\pi x)$ and $\Delta t = 10^{-4}$. A fourth-order explicit central scheme is used for spatial integration, while for time-integration an implicit mid-point method is adopted. The latter belongs to the class of so-called *symplectic* integrators, hence it does not introduce any error on energy conservation [17]. Its use allows to closely resemble the theoretical analysis of Eqs. (13)–(14), which is based on exact time-integration of energy. Fig. 1 shows that the energy associated to the divergence and advective forms soon bifurcates from the initial state. A similar behavior is followed by their single alternance (AD), which does not satisfy the consistency relation; it is interesting to note,

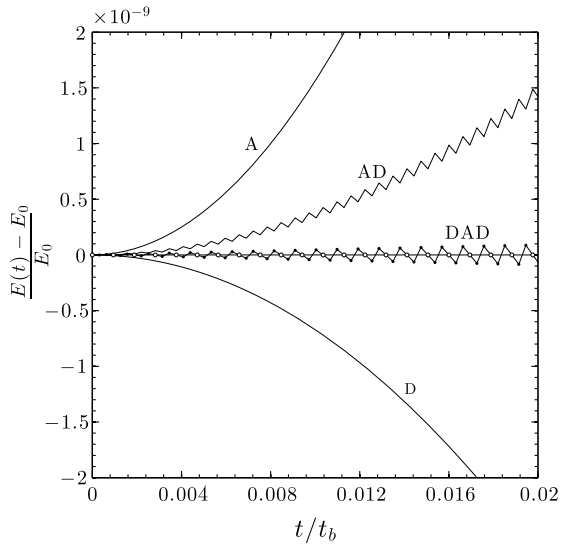

Fig. 1. Time-evolution of normalized energy for Burgers' equation integrated by employing alternating advective and divergence forms arranged in various sequences.

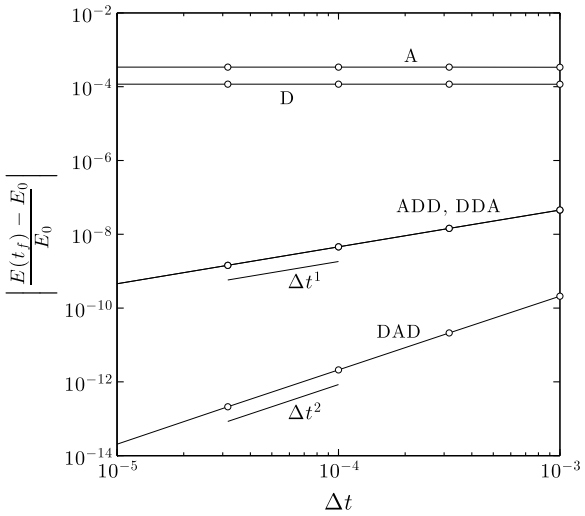

Fig. 2. Convergence of the relative error on energy conservation for different sequences of alternating divergence and advective forms.

however, that its error value lies in between the A and D curves. This is due to the fact that the errors introduced by advective and divergence forms have opposite signs, although they are not equal in magnitude. As a consequence, at each time step the alternation globally produces an error whose magnitude lies between that of the errors given by divergence and advective forms alone.

On the contrary, the second-order DAD sequence oscillates around the initial conditions, assuming a very accurate value at the end of the sequence. Note that the two first order sequences DDA and ADD and the second order sequence DAD are obtained one by the other by cyclically permuting the forms. This implies that the curve labeled with DAD actually displays the errors produced by all these three sequences. The empty circles on the curve denote the error measured at the end of the sequence DAD, while dots denote the error measured at the end of the other two sequences. Fig. 2 shows the time-step convergence of the relative error on energy conservation at $t_f = t_b/2$, where t_b is the break-time at which characteristic lines intersect. The plot fully confirms the analysis by displaying the correct scalings predicted by theory.

In Fig. 3 the time evolution of the normalized energy is reported in the case in which an 18-forms, third order, sequence is adopted. The figure reports a time history in which both the positive and negative errors on energy erratically accumulate during the evolution around a zero mean value. This plot, however, displays also a regular pattern in which the energy approaches zero (to plotting accuracy) after each group of 18 time steps (empty circles on the curve). This behavior is a confirmation that the chosen sequence of forms computes an energy, at the end of the cycle, which is a higher order accurate function of the exact (constant) energy of the system. An accurate inspection of the plot shows also that there is

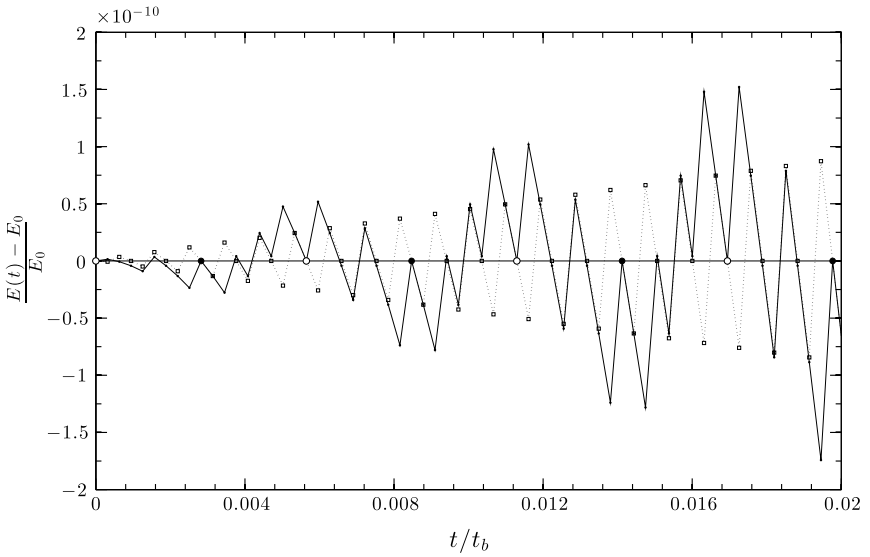

Fig. 3. Time evolution of the normalized energy for the 18-forms, third order sequence ADDDADDDADDADADADD (solid line). Dashed line is the second order sequence DAD, which is plotted for comparison.

another regular pattern on the main curve (highlighted by black dots) for which the error goes near zero at the end of each 9 time steps computation. The explanation of this regular behavior consists in the fact that the subsequence constituted by the first 9 forms actually turns out to be a second order sequence of forms for energy conservation. An analogous plot relative to a longer time integration would show that, as time goes on, the black dots deviate from the constant mean value more rapidly than the empty circles.

4. Alternating Runge–Kutta schemes

The alternating procedure analyzed in the previous section, while being highly attractive for its simplicity and easiness of implementation, has some drawbacks, mainly associated to the oscillatory character of the energy evolution. The theoretically predicted scaling of the energy conservation, as the time increments are reduced, is obtained only when measured at the end of each sequence of alternating forms, when the different errors of advective and divergence discretizations are globally compensated. At intermediate stages, the energy of the solution usually deviates from its mean value in an oscillatory fashion, with an amplitude which either diverges or reduces as a first order function of the time increments. This oscillatory character of the energy can possibly constitute a source of amplification of errors when the solution is coupled to other nonlinear convective transport phenomenologies, as in the case of the Navier–Stokes equations. Another rigidity of the procedure is that the theoretical development presented in the previous section is based on the hypothesis that all the time increments within a given sequence are equal. This can constitute a difficulty when a variable time step algorithm is employed, since the time increments adjustment can be performed only at the end of each sequence, if one wants to save the maximum order of accuracy for the conservation of energy.

A more elegant and compact method can be developed by coupling the alternating procedure to a multi-stage, one-step time integration method. The time integration strategy presented in this section employs a Runge–Kutta scheme and aims to reproduce the skew-symmetric form behavior by adopting a suitable sequence of advective and divergence forms within the stages of a Runge–Kutta scheme, in order to globally compensate the errors relative to a single time step.

The method developed overcomes many of the disadvantages of the "external" alternating procedure, since the solution is computed at the end of the various stages of the Runge–Kutta upgrade, when the alternating sequence has been completed. The energy evolution is free of any oscillating behavior and the time step adjustment can be made exactly in the same manner as it is usually made for a standard Runge–Kutta scheme. Moreover, the highly underdetermined structure of the system of equations for the coefficients of the Runge–Kutta scheme usually allows one to have more degrees of freedom, which can be employed for the achievement of different targets.

4.1. R–K schemes

The proposed procedure is based on a modified Runge–Kutta time advancement, which can be expressed, for Burgers' equation, as:

$$\mathbf{u}^{n+1} = \mathbf{u}^n - \Delta t \sum_{i=1}^{s} b_i \mathbf{C}_i(\mathbf{u}_i)\mathbf{u}_i \tag{15}$$

$$\mathbf{u}_i = \mathbf{u}^n - \Delta t \sum_{j=1}^{s} a_{ij} \mathbf{C}_j(\mathbf{u}_j)\mathbf{u}_j. \tag{16}$$

Eqs. (15)–(16) differ from the standard form of Eqs. (8)–(9) for the fact that the discretized convective operator $\mathbf{C}$ is indexed by a suffix, meaning that a different form (i.e. divergence or advective) can be used for its evaluation at each stage. For instance, in the case of a three-stage Runge–Kutta scheme, the convective forms $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ employed within the stages of the time advancement procedure can be arbitrarily chosen among the possible sequences of advective and divergence forms. The coefficients a_{ij} and b_i are usually set in such a way that the maximum formal order of accuracy is obtained for the convergence of the solution. This procedure is based on the matching of the different terms arising in the Taylor expansions of both the right hand side of Eqs. (15)–(16) and of the difference $\mathbf{u}^{n+1} - \mathbf{u}^n$, evaluated exactly by taking into account the right hand side of the original system of ODEs. The result is a (usually strongly underdetermined) system of nonlinear equations, whose solutions are expressed as families of schemes with equal order of accuracy parametrized by one or more constants. In the present approach this procedure is complemented by the additional requirement that a formal order of accuracy is achieved also for the conservation of energy. The optimization is performed on both the coefficient values and on the possible sequences of the $\mathbf{C}_i$'s and is based on the expression of the energy produced in a single time step by the Runge–Kutta procedure as a Taylor series.

4.2. Energy evolution

The energy error occurring over a single time-step for the modified Runge–Kutta procedure can be expressed as:

$$\frac{\Delta E}{\Delta t} = -\sum_{i=1}^{s} b_i \mathbf{u}_i^T \mathbf{C}_i(\mathbf{u}_i)\mathbf{u}_i - \frac{\Delta t}{2}\sum_{i,j=1}^{s}\left(b_i a_{ij} + b_j a_{ji} - b_i b_j\right)\mathbf{u}_i^T \mathbf{C}_i^T(\mathbf{u}_i)\mathbf{C}_j(\mathbf{u}_j)\mathbf{u}_j. \tag{17}$$

The idea of adopting different discretized forms inside the stages of a Runge–Kutta procedure presents new possibilities for obtaining cost-effective energy-preserving algorithms. In general, the alternating procedure associated to a given sequence of forms produces separately both spatial and temporal errors. In this case, however, there is the possibility of conserving energy up to a certain order of accuracy if the mixed (spatial and temporal) errors are weighted is such a way that they globally compensate, at least asymptotically as Δt tends towards zero, with a certain accuracy. The advantage presented by the alternating procedure is of course the low cost evaluation of the convective term at each stage.

The starting point of the analysis is the evaluation of the energy error as a Taylor series expansion in the time step increment Δt. This expression can be obtained by recursively substituting Eq. (16) into the right hand side of Eq. (16) itself and in Eq. (17). The key ingredient of the procedure will be, as shown below, the linearity of the $\mathbf{C}_i(\mathbf{u}_i)$ as functions of $\mathbf{u}_i$. By expressing the terms $\mathbf{u}_j$ at the right hand side of Eq. (16) through Eq. (16) itself and by employing the exact relation:

$$\mathbf{C}_j(\mathbf{u}_j) = \mathbf{C}_j(\mathbf{u}^n) - \Delta t \sum_k a_{jk}\mathbf{C}_j(\mathbf{C}_k(\mathbf{u}_k)\mathbf{u}_k) \tag{18}$$

one easily obtains:

$$\mathbf{u}_i = \mathbf{u} - \Delta t\left(\sum_j a_{ij}\mathbf{C}_j\mathbf{u}\right) + \Delta t^2\left[\sum_{j,k} a_{ij}a_{jk}\left(\mathbf{C}_j\mathbf{C}_k + \mathbf{C}_{jk}\right)\mathbf{u}\right] + O\left(\Delta t^3\right) \tag{19}$$

where the summations are extended to the number of stages s and the following conventions have been adopted: $\mathbf{u} = \mathbf{u}^n$; $\mathbf{C}_j = \mathbf{C}_j(\mathbf{u}^n)$; $\mathbf{C}_{jk} = \mathbf{C}_j(\mathbf{C}_k(\mathbf{u}^n)\mathbf{u}^n)$. In these last two relations, and in the following ones involving the matrices $\mathbf{C}_i$, parentheses denote functional dependence. In all the other cases in which the functional dependence of the matrices $\mathbf{C}_j$ is not specified, it is assumed that the quantities are evaluated at $\mathbf{u}^n$.

Eq. (19) can be employed in order to express $\mathbf{C}_i(\mathbf{u}_i)$ as a Taylor series in Δt through Eq. (18):

$$\mathbf{C}_i(\mathbf{u}_i) = \mathbf{C}_i - \Delta t\sum_j a_{ij}\mathbf{C}_{ij} + \Delta t^2\sum_{i,j} a_{ij}a_{jk}\left(\mathbf{C}_{i,jk} + \mathbf{C}_{ijk}\right) + O\left(\Delta t^3\right) \tag{20}$$

where the further definitions have been introduced: $\mathbf{C}_{i,jk} = \mathbf{C}_i(\mathbf{C}_j\mathbf{C}_k\mathbf{u})$ and $\mathbf{C}_{ijk} = \mathbf{C}_i(\mathbf{C}_{jk}\mathbf{u})$.

By substituting Eqs. (19)–(20) into Eq. (17) one obtains the expression for the energy variation ΔE as a Taylor series. The spatial part of the error is:

$$\begin{aligned}
-\sum_i b_i \mathbf{u}_i^T \mathbf{C}_i(\mathbf{u}_i)\mathbf{u}_i = -\mathbf{u}^T\left(\sum_i b_i\mathbf{C}_i\right)\mathbf{u} + \Delta t\,\mathbf{u}^T\left[\sum_{i,j} b_i a_{ij}\left(\mathbf{C}_i\mathbf{C}_j + \mathbf{C}_{ij} + \mathbf{C}_j^T\mathbf{C}_i\right)\right]\mathbf{u} - \\
- \Delta t^2\mathbf{u}^T\sum_{i,j,k} b_i a_{ij}\left[a_{jk}\left(\mathbf{C}_i\mathbf{C}_j\mathbf{C}_k + \mathbf{C}_i\mathbf{C}_{jk} + \mathbf{C}_{i,jk} + \mathbf{C}_{ijk} + \mathbf{C}_k^T\mathbf{C}_j^T\mathbf{C}_i + \mathbf{C}_{jk}^T\mathbf{C}_i\right) + \right. \\
\left. + a_{ik}\left(\mathbf{C}_{ij}\mathbf{C}_k + \mathbf{C}_j^T\mathbf{C}_i\mathbf{C}_k + \mathbf{C}_j^T\mathbf{C}_{ik}\right)\right]\mathbf{u} + O(\Delta t^3).
\end{aligned}$$

The temporal part turns out to be:

$$-\frac{\Delta t}{2}\sum_{ij} g_{ij}\mathbf{u}_i^T\mathbf{C}_i^T\left(\mathbf{u}_i\right)\mathbf{C}_j\left(\mathbf{u}_j\right)\mathbf{u}_j = -\frac{\Delta t}{2}\mathbf{u}^T\left[\sum_{i,j} g_{ij}\mathbf{C}_i^T\mathbf{C}_j\right]\mathbf{u} + \frac{\Delta t^2}{2}\mathbf{u}^T\left[\sum_{i,j,k} g_{ij}a_{jk}\mathbf{F}_{jk}^i + g_{ij}a_{ik}\mathbf{F}_{ik}^j\right]\mathbf{u} + O\left(\Delta t^3\right)$$

where $g_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j$ and $\mathbf{F}_{jk}^i = \mathbf{C}_i^T\mathbf{C}_{jk} + \mathbf{C}_i^T\mathbf{C}_j\mathbf{C}_k$. The sum of the two contributions can be rearranged to give:

$$\begin{aligned}\frac{\Delta E}{\Delta t} = &-\mathbf{u}^T\left[\sum_i b_i\mathbf{C}_i\right]\mathbf{u} + \frac{\Delta t}{2}\mathbf{u}^T\left[\sum_{ij} 2b_i a_{ij}\left(\mathbf{C}_i\mathbf{C}_j + \mathbf{C}_{ij}\right) + g_{ij}\mathbf{C}_i^T\mathbf{C}_j\right]\mathbf{u} - \\ &- \frac{\Delta t^2}{2}\mathbf{u}^T\left\{\sum_{ijk} 2b_i a_{ij}\left[a_{jk}\left(\mathbf{C}_i\mathbf{C}_j\mathbf{C}_k + \mathbf{C}_i\mathbf{C}_{jk} + \mathbf{C}_{i,jk} + \mathbf{C}_{ijk}\right) + a_{ik}\left(\mathbf{C}_{ij}\mathbf{C}_k\right)\right] + \right. \\ &\left. + g'_{ij}\left(\mathbf{F}_{jk}^i a_{jk} + \mathbf{F}_{ik}^j a_{ik}\right)\right\}\mathbf{u} + O\left(\Delta t^3\right)\end{aligned} \tag{21}$$

where $g'_{ij} = b_i a_{ij} - b_j a_{ji} + b_i b_j$.

Eq. (21) constitutes the basic relation on which the optimized Runge–Kutta schemes can be constructed. The procedure is based on the principle that suitable choices for the coefficients b_i and a_{ij} can nullify the successive terms in the series expansion of ΔE, thus identifying "optimal" schemes for what concerns conservation of energy.

From a practical point of view, the determination of the coefficients is obtained by explicitly imposing the conditions that the various terms at the right hand side of Eq. (21) vanish, independently of $\mathbf{u}$. These constraints conduct to the determination of algebraic nonlinear equations involving the coefficients b_i and a_{ij} which can be coupled to classical order conditions to give a global system for the determination of new Runge–Kutta schemes. The number and the structure of the nonlinear equations related to energy conservation depend on several parameters, such as the number of stages, the chosen sequence of $\mathbf{C}_i$'s and the details of the spatial discretization. In general, each of the operators involving the product or the composition of two or more matrices $\mathbf{C}_i$ (i.e. each of the terms $\mathbf{C}_i\mathbf{C}_j$, $\mathbf{C}_{ij}$, $\mathbf{C}_i^T\mathbf{C}_j$, ...) has to be considered as an independent function of the state vector $\mathbf{u}$. In these general cases the number of constraint to be imposed to the coefficients dramatically grows with both the number of stages and the order of accuracy, since the number of independent groups to be nullified grows linearly with the number of stages and more than linearly with the order of accuracy. In such situations the global system of equations quickly becomes overdetermined, especially in the more appealing case of explicit schemes. The key observation that strongly simplifies the structure of the needed constraints is that when the "physically compatible" relation $\mathbf{D}^T = -\mathbf{D}$ for the discrete derivative operator can be assumed, the advective and divergence discrete forms are related by $\left(\mathbf{C}^{\mathrm{adv}}\right)^T = -2\mathbf{C}^{\mathrm{div}}$ and the various products or compositions of matrices $\mathbf{C}_i$ are no more a set of independent functions of $\mathbf{u}$. In fact, the various terms at the right hand side of Eq. (21) can be grouped as a linear combination of fewer independent terms. For instance, each of the $3s$ terms $\mathbf{C}_i\mathbf{C}_j$, $\mathbf{C}_i(\mathbf{C}_j\mathbf{u})$ and $\mathbf{C}_i^T\mathbf{C}_j$ can be recast into one of the three basic forms $\mathbf{C}^{\mathrm{adv}}\mathbf{C}^{\mathrm{adv}}$, $\mathbf{C}^{\mathrm{adv}}\mathbf{C}^{\mathrm{div}}$ and $\mathbf{C}^{\mathrm{div}}\mathbf{C}^{\mathrm{adv}}$. A similar procedure can be applied to the constant term and to the higher order contributions.

In what follows, the analysis will be limited to Runge–Kutta schemes with up to four stages and to the requirement of a maximum of second-order on energy conservation. The argument will not be treated exhaustively since the manipulation of third-order terms on energy conservation and of higher-order Runge–Kutta schemes would require involved calculations which add little to the exposition of the main idea behind the technique, which is the principal motivation of this paper.

5. Results

5.1. Two and three-stage R–K methods

The analysis is firstly conducted for the simple case of two-stage explicit Runge–Kutta methods. The free parameters for these schemes are b_1, b_2 and a_{12} and the maximum order conditions which can be satisfied are that of a second order scheme: $b_1 + b_2 = 1$ and $b_2 a_{21} = 1/2$. By imposing these two constraints one obtains the well known one-parameter family $b_1 = 1 - \theta$, $b_2 = \theta$ and $a_{12} = 1/2\theta$, with $\theta \neq 0$. The free parameter θ can be fixed by requiring that also the first order condition on energy conservation is satisfied by b_1 and b_2, for each given alternation of forms. In the case of two stages schemes this condition reduces to: $\mathbf{u}^T\left(b_1\mathbf{C}_1 + b_2\mathbf{C}_2\right)\mathbf{u} = 0$. For a standard Runge–Kutta method in which $\mathbf{C}_1 = \mathbf{C}_2$, this term cannot vanish, except for the case in which the skew-symmetric form is employed, since the condition to be satisfied by b_1 and b_2 would be $b_1 + b_2 = 0$, which is incompatible with the first order condition $b_1 + b_2 = 1$. In the case in which an alternation of forms is employed, one has different conditions on the b_i's, one for each alternation sequence. If the first stage is computed by adopting the advective form and the second the divergence form (sequence AD) one has the condition

$$\mathbf{u}^T\left(b_1\mathbb{A} + b_2\mathbb{D}\right)\mathbf{u} = 0, \tag{22}$$

where the more compact notations $\mathbb{A} = \mathbf{C}^{\mathrm{adv}} = \mathbf{UD}$ and $\mathbb{D} = \mathbf{C}^{\mathrm{div}} = \mathbf{DU}/2$ have been employed. This equation can be simplified by noting that in the case in which the derivative operator is a skew-symmetric matrix, the relation $\mathbb{A}^T = -2\mathbb{D}$ is easily

Table 1
2 stages optimized Runge–Kutta methods.

AD-2S1E(2)	DA-2S1E(2)
$\begin{array}{c\|cc} 0 & & \\ \frac{3}{4} & 0 & \\ \hline & \frac{1}{3} & \frac{2}{3} \end{array}$	$\begin{array}{c\|cc} 0 & & \\ \frac{3}{2} & 0 & \\ \hline & \frac{2}{3} & \frac{1}{3} \end{array}$

derived, which in turn implies: $\mathbf{u}^T\mathbb{A}\mathbf{u} = -2\mathbf{u}^T\mathbb{D}\mathbf{u}$. It will be useful in what follows to express this circumstance by introducing an equivalence relation between the various operators. In this context, two matrix operators (which are functions of the vector $\mathbf{u}$) will be "equivalent" if their quadratic forms associated to the vector $\mathbf{u}$ are equal for each value of $\mathbf{u}$. Related to this definition, the following notation will be used $\mathbf{A}(\mathbf{u}) \sim \mathbf{B}(\mathbf{u})$ if $\mathbf{u}^T\mathbf{A}(\mathbf{u})\,\mathbf{u} = \mathbf{u}^T\mathbf{B}(\mathbf{u})\,\mathbf{u}$ for all $\mathbf{u}$. With this notation, one has $\mathbb{A} \sim -2\mathbb{D}$ and the operator $\mathbb{D}$ can be factored out in Eq. (22), which is hence satisfied, independently of $\mathbf{u}$, if and only if $b_2 = 2b_1$. This last condition fixes the parameters of the Runge–Kutta scheme to the values $b_1 = 1/3$, $b_2 = 2/3$ and $a_{12} = 3/4$, which has second order accuracy on the convergence of the solution and first order accuracy on the conservation of energy. In what follows, the following synthetic notation will be employed. Each method is accompanied by an acronym in which the alternating sequence is indicated together with the theoretical order of accuracy on solution (S) and on energy-conservation (E). The number of derivatives computations required per time-step is also reported in bracket. This quantity is used here as a simple and intuitive cost metric in order to compare the performances of the various schemes. In this notation the derived scheme is referred to as AD-2S1E(2). The dual scheme DA-2S1E(2) can be easily obtained with the same procedure, which leads to $b_1 = 2b_2$. The coefficients of the schemes are summarized in Table 1.

Optimized three stage methods can be derived by applying a similar procedure. The number of free parameters for the case of explicit three stages schemes rises to six ($s(s+1)/2$), while the number of conditions to be satisfied for the case of an order 3 scheme is four. Hence, the family of classical third-order, three-stage schemes is at best parametrized by two constants, although two one-parameter families of third order schemes also exist (cf. Butcher [16]). The degrees of freedom constituted by the free parameters can be employed, for example, to guarantee a first order accuracy on conservation of energy in the cases in which alternating divergence and advective forms are used. The number of possible sequences in which the same form is not repeated three times is six. For each of these sequences, a relation between the constants b_i has to be satisfied for first order conservation of energy. This additional constraint can then be imposed to the coefficients satisfying third order relations to obtain 3S1E(3) schemes, as it has been done for two stage schemes. This procedure, however, although giving a complete picture of all the possible solutions, would be a little cumbersome, since each of the six admissible sequences should be separately considered for each of the three families of third order schemes. Here, a treatment is presented in which, by renouncing to the complete analysis of all the possible solutions, the full system of order relations (on the solution and on the energy) is directly solved in the general case, without discussing the many exceptional cases. The treatment can be developed at once for all the possible sequences by observing that in all cases the equation related to first order conservation of energy: $\mathbf{u}^T(b_1\mathbf{C}_1 + b_2\mathbf{C}_2 + b_3\mathbf{C}_3)\,\mathbf{u} = 0$, leads to the relation

$$\alpha_1 b_1 + \alpha_2 b_2 + \alpha_3 b_3 = 0 \tag{23}$$

where $\alpha_i = 1$ or $\alpha_i = -1/2$ in the cases $\mathbf{C}_i = \mathbb{A}$ or $\mathbf{C}_i = \mathbb{D}$, respectively. For the determination of 3S1E(3) families of schemes, Eq. (23) has to be coupled to the classical relations for a third order Runge–Kutta scheme:

$$\sum_i b_i = 1 \tag{24}$$

$$\sum_i b_i \sum_j a_{ij} = \frac{1}{2} \tag{25}$$

$$\sum_i b_i \left(\sum_j a_{ij}\right)^2 = \frac{1}{3} \tag{26}$$

$$\sum_i b_i \sum_j a_{ij} \sum_k a_{jk} = \frac{1}{6}. \tag{27}$$

The general solution for this system can be obtained in two steps: first, derive b_1, b_2 and b_3 from the linear equations (24) and (23) as a one-parameter family of coefficients. Second, solve for a_{21}, a_{31} and a_{32} from the nonlinear system (25)–(27) by considering b_2 and b_3 as parameters, yielding

$$\begin{aligned} a_{21} &= \frac{1 - 2b_3c_\pm}{2b_2} \\ a_{32} &= \frac{1}{6b_3a_{21}} \\ a_{31} &= c_\pm - a_{32} \end{aligned} \tag{28}$$

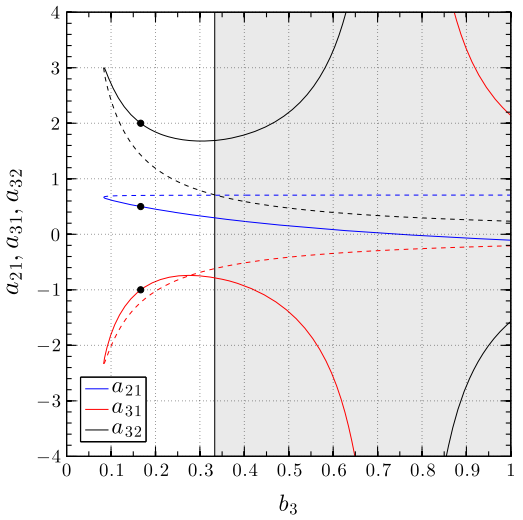

Fig. 4. Coefficients a_{21}, a_{31} and a_{32} as functions of the parameter b_3 for the two families associated with the 3S1E schemes for the sequence ADA.

where

$$c_{\pm} = \frac{1}{2\,(b_2+b_3)}\left(1 \pm \sqrt{1 - \frac{(b_2+b_3)\,(3-4b_2)}{2b_3}}\right) \tag{29}$$

which is real for $(b_2+b_3)(3-4b_2)/2b_3 \leq 1$. Eqs. (28)–(29) furnish two families of one-parameter schemes for each sequence of alternating forms inside the three stages, provided that b_1, b_2 and $a_{12} \neq 0$ and that the coefficients b_2 and b_3 lie in the range of real values for $c_{\pm}$.

As an example, for the sequence $\mathbf{C}_1 = \mathbb{A}$, $\mathbf{C}_2 = \mathbb{D}$, $\mathbf{C}_3 = \mathbb{A}$, (i.e. for the sequence ADA), first order conservation of energy leads to the relation: $b_1 - b_2/2 + b_3 = 0$ which, coupled to Eq. (24) has the solution $b_1 = 1/3 - \theta$, $b_2 = 2/3$ and $b_3 = \theta$. In Fig. 4 the coefficients a_{21}, a_{31} and a_{32} of the two families associated with this sequence are reported. The dots on the curves relative to the first family of coefficients are the values of the coefficients of the classical Kutta scheme ($b_1 = b_3 = 1/6$, $b_2 = 2/3$, $a_{21} = 1/2$, $a_{31} = -1$, $a_{32} = 2$), which is a member of this class of schemes and hence has a first order accuracy on energy conservation when employed in conjunction with the ADA alternation of forms. In Fig. 4, the range of values of b_3 leading to $b_1 < 0$ are shaded, as in Runge–Kutta methods the coefficients b_i are normally taken positive for stability reasons [16].

The treatment of the second order term in the energy expression needs some additional care. Eq. (21) shows that it is composed by the sum of several terms, each being proportional to a quadratic form involving one of the operators $\mathbf{C}_i\mathbf{C}_j$, $\mathbf{C}_i^T\mathbf{C}_j$, $\mathbf{C}_{ij}$. The first two operators, constituted by the product of convective matrices, are equivalent in all cases to one of the operators $\mathbb{AA}$, $\mathbb{DD}$, $\mathbb{AD}$ and $\mathbb{DA}$ multiplied by a scalar. Actually, since $\mathbf{u}^T\mathbb{AA}\mathbf{u} = 4\mathbf{u}^T\mathbb{DD}\mathbf{u}$, one has $\mathbb{AA} \sim 4\mathbb{DD}$ and the number of independent forms reduces to three. Hence, each of the products between two convective matrices $\mathbf{C}_i\mathbf{C}_j$ or $\mathbf{C}_i^T\mathbf{C}_j$ is equivalent to one of the basic operators $\mathbb{AA}$, $\mathbb{AD}$ and $\mathbb{DA}$, multiplied by a scalar. It can be easily shown, although not completely evident at first, that also the composite terms $\mathbf{C}_{ij}$ are always equivalent to one of the three forms $\mathbb{AA}, \mathbb{AD}$ and $\mathbb{DA}$. This fact can be shown by observing that for every vector $\mathbf{u}$ and for every matrix $\mathbf{B}$ the following relation holds:

$$\mathrm{diag}\,(\mathbf{Bu})\,\mathbf{u} = \mathbf{UBu}$$

where, as usual, $\mathbf{U} = \mathrm{diag}\,(\mathbf{u})$. This relation can be employed in order to express the forms $\mathbf{C}_{ij}$ into one of the basic forms $\mathbb{AA}, \mathbb{AD}$ and $\mathbb{DA}$. As an example, for the terms $\mathbb{D}\,(\mathbb{A}\mathbf{u})$ and $\mathbb{A}\,(\mathbb{A}\mathbf{u})$ (parentheses denoting functional dependence) one has:

$$\mathbf{u}^T\mathbb{D}\,(\mathbb{A}\mathbf{u})\,\mathbf{u} = \mathbf{u}^T\frac{1}{2}\mathbf{D}\mathrm{diag}\,(\mathbb{A}\mathbf{u})\,\mathbf{u} = \mathbf{u}^T\frac{1}{2}\mathbf{DU}\mathbb{A}\mathbf{u} = \mathbf{u}^T\mathbb{DA}\mathbf{u}$$

$$\mathbf{u}^T\mathbb{A}\,(\mathbb{A}\mathbf{u})\,\mathbf{u} = \mathbf{u}^T\mathrm{diag}\,(\mathbb{A}\mathbf{u})\,\mathbf{Du} = -\mathbf{u}^T\mathbf{D}\mathrm{diag}\,(\mathbb{A}\mathbf{u})\,\mathbf{u} = -\mathbf{u}^T\mathbf{DU}\mathbb{A}\mathbf{u} = -2\mathbf{u}^T\mathbb{DA}\mathbf{u}.$$

By acting in a similar way on the other possible terms $\mathbf{C}_{ij}$ one finally obtains the following equivalences:

$$\mathbb{D}\,(\mathbb{A}\mathbf{u}) \sim \mathbb{DA}$$

$$\mathbb{A}\,(\mathbb{A}\mathbf{u}) \sim -2\mathbb{DA}$$

$$\mathbb{A}\,(\mathbb{D}\mathbf{u}) \sim -\frac{1}{2}\mathbb{AA}$$

$$\mathbb{D}\,(\mathbb{D}\mathbf{u}) \sim \mathbb{DD} \sim \frac{1}{4}\mathbb{AA}.$$

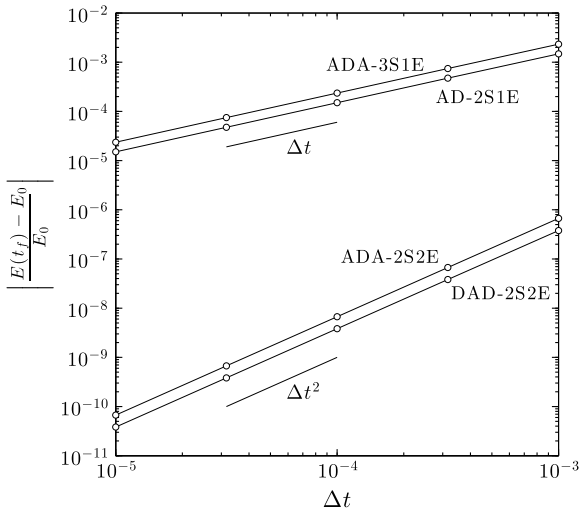

Fig. 5. Convergence of the relative error on energy conservation as a function of the time step Δt for different Runge–Kutta schemes with 2 and 3 stages.

These relations can be employed in the evaluation of the second order term in Eq. (21), and permit to group the various terms into three independent ones, associated to the forms $\mathbb{AA}$, $\mathbb{AD}$ and $\mathbb{DA}$, independentely of the number of stages. Hence, the number of equations associated to the fulfillment of second order accuracy on energy conservation is three. The form of these equations depends, of course, on the particular alternation sequence. These three relations can be coupled to first and second order equations on the accuracy of the solution (two equations) and to the first order equation on the accuracy of energy conservation. The result is a system of six nonlinear equations in the six unknowns b_1, b_2, b_3, a_{21}, a_{31} and a_{32} for the determination of explicit three stage 2S2E(3) Runge–Kutta schemes. Due to nonlinearity, it is difficult to ascertain if this system has in general a solution and, in that case, if it is unique. Among all the possible sequences of forms for the case of explicit three stage methods, it is found that only in the case of the two sequences ADA and DAD a solution can be found. In all the other cases the system has no solution. Moreover, the solution, when available, is not unique.

The sequence ADA produces the system:

$$
\begin{aligned}
& b_1 + b_2 + b_3 = 1 && \text{1st order on the solution} \\
& b_2 a_{21} + b_3 (a_{31} + a_{32}) = \frac{1}{2} && \text{2nd order on the solution} \\
& b_1 - \frac{b_2}{2} + b_3 = 0 && \text{1st order on the energy} \\
& \left.\begin{aligned} & b_2^2 - b_3 a_{32} = 0 \\ & 2 - b_3 (a_{32} - 2a_{31}) + b_2 b_3 + b_1 b_2 = 0 \\ & b_1^2 - 2b_2 a_{21} + 2b_3 a_{31} + 2b_1 b_3 + b_3^2 = 0 \end{aligned}\right\} && \text{2nd order on the energy}
\end{aligned}
$$

whose solution is easily found by successive substitution and can be expressed as the one-parameter family of schemes (ADA-2S2E(3)):

$$
b_1 = \frac{1}{3} - \theta, \quad b_2 = \frac{2}{3}, \quad b_3 = \theta \quad a_{21} = \frac{1}{3}, \quad a_{31} = \frac{1}{6\theta}, \quad a_{32} = \frac{1}{9\theta} \tag{30}
$$

which is valid for $b_3 \neq 0$. The analogous family of DAD-2S2E(3) schemes is:

$$
b_1 = \frac{2}{3} - \theta, \quad b_2 = \frac{1}{3}, \quad b_3 = \theta \quad a_{21} = \frac{1}{3}, \quad a_{31} = \frac{1}{3\theta}, \quad a_{32} = \frac{1}{18\theta}. \tag{31}
$$

It can be readily seen that none of these schemes can satisfy Eqs. (26)–(27) for a particular value of θ, and hence three stage 3S2E(3) schemes cannot be obtained.

In Fig. 5 the convergence of the relative error on energy conservation for different Runge–Kutta schemes with 2 and 3 stages is reported. The curve labeled AD-2S1E is relative to the 2 stage Runge–Kutta scheme reported in Table 1, while the curve labeled ADA-3S1E is the classical Kutta scheme with ADA alternation of forms. The curves displaying second order convergence and labeled ADA-2S2E and DAD-2S2E are relative to the schemes whose coefficients are given in Eqs. (30) and (31) respectively. The plot shows that all the schemes display the correct scaling with time step Δt.

5.2. Four-stage R–K methods

Four stage explicit Runge–Kutta methods are characterized by ten coefficients, which can be determined (at least as families of values) by imposing the required order of accuracy on the solution in the form of nonlinear constraints. The standard procedure is to obtain the families of coefficients by imposing the eight nonlinear conditions necessary to obtain 4th order accuracy. The complete solution to this system is quite cumbersome to reproduce here, and is usually given as a two parameter family of coefficients, or, in some exceptional cases, as a number of one-parameter families. An extensive treatment of this problem, together with the steps to be carried out in order to solve the nonlinear systems of conditions, can be found in the classical book by Butcher [16]. The undertaken approach will be again to complement the classical procedure by introducing the possibility of alternating the forms inside the various stages of the Runge–Kutta method. Then, the nonlinear constraints on the coefficients associated to conservation of energy (for a given order of accuracy) will be taken into account. The complete treatment of all the particular cases which can arise in this procedure is out of the scope of the present work. The variety of nonlinear conditions and the number of particular cases exceptionally grow as one introduces all the possible sequences of alternating forms. In this paragraph, it will be simply illustrated how the general program can be carried out, and some representative examples will be given.

The analysis starts with 4S1E(4) schemes, which can be obtained by imposing the first order condition on energy conservation on the families of schemes satisfying fourth order conditions on the solution. As examples, some of the special one-parameter classes of schemes identified by Kutta are analyzed, which are reported in the book by Butcher [16] and labeled as case I–case V. The first special case identified by Kutta (case I), for which a family of fourth order schemes is obtained and given by the following Butcher array:

case I

0				
$1-\theta$	0			
$\frac{6\theta^2(1-2\theta)}{\gamma(\theta)}$	$\frac{6\theta^2}{\gamma(\theta)}$	0		
$\frac{12\theta^3-24\theta^2+17\theta-4}{2(1-\theta)\mu(\theta)}$	$\frac{\theta(1-2\theta)}{2(1-\theta)\mu(\theta)}$	$\frac{1-\theta}{\mu(\theta)}$	0	
$\frac{\mu(\theta)}{\gamma(\theta)}$	$\frac{1}{\gamma(\theta)}$	$\frac{1}{\gamma(\theta)}$	$\frac{\mu(\theta)}{\gamma(\theta)}$	

where $\gamma(\theta)=12\theta(1-\theta)$, $\mu(\theta)=6\theta-1-6\theta^2$ and $a_{31}+a_{32}=\theta$. This family produces fourth order schemes provided $a_{21}\notin\left\{0,1/2,1/2\pm\sqrt{3}/6,1\right\}$. Given a generic sequence of forms, identified as usual by the sequence of α_i, the first order condition on energy conservation gives: $(\alpha_1+\alpha_4)\mu(\theta)+(\alpha_2+\alpha_3)=0$ which conducts to the requirement that the function $\mu(\theta)$ assumes specific values in correspondence of each sequence of α_i:

$$\mu(\theta)=-\frac{(\alpha_2+\alpha_3)}{(\alpha_1+\alpha_4)}.$$

The results of this analysis is that the sequences AAAD and DAAA produce two 4S1E(4) schemes, corresponding to $\theta=(1\pm\sqrt{3})/2$, as well as the sequences AADA and ADAA, which also produce two 4S1E(4) schemes corresponding to $\theta=(2\pm\sqrt{2})/4$. All the other sequences produce zeroth order schemes on conservation of energy. The case II family of schemes:

case II

0				
θ	0			
$\frac{1}{2}-\frac{1}{8\theta}$	$\frac{1}{8\theta}$	0		
$\frac{1}{2\theta}-1$	$-\frac{1}{2\theta}$	2	0	
$\frac{1}{6}$	0	$\frac{2}{3}$	$\frac{1}{6}$	

is a 4S1E(4) family, independently of θ, for the two sequences ADDA and AADA. The cases III, IV and V:

case III

0				
$\frac{1}{2}$	0			
$-\frac{1}{12\theta}$	$\frac{1}{12\theta}$	0		
$-\frac{1}{2}-6\theta$	$\frac{3}{2}$	6θ	0	
$\frac{1}{6}-\theta$	$\frac{2}{3}$	θ	$\frac{1}{6}$	

case IV

0				
1	0			
$\frac{3}{8}$	$\frac{1}{8}$	0		
$1-\frac{1}{4\theta}$	$-\frac{1}{12\theta}$	$\frac{1}{3\theta}$	0	
$\frac{1}{6}$	$\frac{1}{6}-\theta$	$\frac{2}{3}$	θ	

case V

0				
$\frac{1}{2}$	0			
$\frac{1}{2}-\frac{1}{6\theta}$	$\frac{1}{6\theta}$	0		
0	$1-3\theta$	3θ	0	
$\frac{1}{6}$	$\frac{2}{3}-\theta$	θ	$\frac{1}{6}$	

are 4S1E(4) schemes with the sequences ADAA, AADA and ADDA respectively. Note that class V schemes are of particular interest, since the 'classical' Runge–Kutta scheme belongs to this class for the case $\theta = 1/3$. The analysis shows that when implemented with a sequence ADDA the 'classical' Runge–Kutta scheme has a first order conservation of energy.

In addition to 4S1E(4) schemes already illustrated, the ten coefficients arising in four stages Runge–Kutta methods can be fixed by imposing, for each of the possible sequences of alternating forms, a third order accuracy on the solution (four equations) together with a second order accuracy on energy conservation (four equations). The solution of the eight-equation nonlinear system (when available) will produce 3S2E(4) schemes and, since the system of equations is overdetermined, in general one expects that also families of schemes, having one or more parameters, will appear. The nonlinear systems associated to the 14 possible sequences of alternating forms all have four common equations, given by the third order constraints for a four stage method:

$$
\begin{aligned}
&b_1 + b_2 + b_3 + b_4 = 1 \\
&b_2 a_{21} + b_3 (a_{31} + a_{32}) + b_4 (a_{41} + a_{42} + a_{43}) = \frac{1}{2} \\
&b_2 a_{21}^2 + b_3 (a_{31} + a_{32})^2 + b_4 (a_{41} + a_{42} + a_{43})^2 = \frac{1}{3} \\
&a_{21} (b_3 a_{32} + b_4 a_{42}) + b_4 a_{43} (a_{31} + a_{32}) = \frac{1}{6}.
\end{aligned}
\tag{32}
$$

In addition to these equations there is the first order energy conservation constraint:

$$\alpha_1 b_1 + \alpha_2 b_2 + \alpha_3 b_3 + \alpha_4 b_4 = 0$$

where, as usual, the coefficients α_i depend on the particular sequence of forms, and three equations for second order conservation of energy, whose structure depends on the particular sequence. In Appendix A the four equations associated to second order conservation of energy are reported for all the possible sequences of alternating forms. Each of these nonlinear systems can be separately studied in order to obtain families of 3S2E(4) schemes for every given sequence. The complete characterization of all the possible solutions for each alternating sequence is, again, out of the scopes of the present study. However, some general considerations, together with some particular solutions, will be presented, as it has been done for previous cases. In many circumstances symbolic nonlinear solvers can also be employed to simplify the task of obtaining solutions.

The structure of the nonlinear system is such that in many cases significative simplifications can be obtained by firstly considering the equations for first order accuracy on both solution and energy. These two equations involve only the b_i's and immediately furnish the numerical value of one of the unknowns (when three forms are equal within the sequence) or of the sum of two unknowns. In some circumstances this information alone can be sufficient in order to conclude that the nonlinear system has no solutions, as in the cases of the sequences AAAD and AADD. In general, it is found that in many cases a next useful step is the derivation of the b_i's as parametric functions of the a_{ij}'s by solving the linear equations in the b_i's. At this point, the remaining fully nonlinear equations can be attacked. It results that, among the possible 14 series, the six sequences starting with AA or DD have no solution, together with the two sequences ADDD and DAAA. The remaining six sequences have solutions which are parametrized by one or more constants. Such families of schemes are usually quite difficult to express compactly, but representative examples can be obtained by fixing one or more constants.

As an example, the case of the sequence ADDA is considered, for which the complete nonlinear system for the coefficients is given by Eqs. (32), for third order accuracy on the solution, and by equations:

$$
\begin{aligned}
&b_1 - \frac{b_2}{2} - \frac{b_3}{2} + b_4 = 0 && \text{1st order on the energy} \\
&\left.
\begin{aligned}
&\frac{9}{8} b_3 a_{32} + b_4 (2a_{41} - a_{42} - a_{43}) - (b_1 + b_4)(b_2 + b_3) = 0 \\
&b_3 a_{32} - 4 b_4 (a_{42} + a_{43}) + (b_2 + b_3)^2 = 0 \\
&4 b_2 a_{21} + b_3 (4a_{31} - \frac{a_{32}}{2}) - 4 b_4 a_{41} - 2 (b_1 + b_4)^2 = 0
\end{aligned}
\right\} && \text{2nd order on the energy}
\end{aligned}
\tag{33}
$$

for second order accuracy on the energy. By making the assumption $a_{32} = 0$ and $b_2 = 0$ or $b_3 = 0$, after some manipulations one obtains the two families:

$$
\begin{array}{c|cccc}
0 & & & \\
\frac{25\theta - 42}{75\theta - 28} & 0 & & \\
\frac{1}{3} & 0 & 0 & \\
\frac{14}{25} & \frac{28}{75} - \theta & \theta & 0 \\
\hline
\frac{1}{28} & 0 & \frac{2}{3} & \frac{25}{84}
\end{array}
\qquad
\begin{array}{c|cccc}
0 & & & \\
\frac{1}{3} & 0 & & \\
\frac{75\theta + 98}{225\theta} & 0 & 0 & \\
\frac{14}{25} & \frac{28}{75} - \theta & \theta & 0 \\
\hline
\frac{1}{28} & \frac{2}{3} & 0 & \frac{25}{84}
\end{array}
$$

The choice of the parameter $\theta = 0$ for the first family or $\theta = 28/75$ for the second is particularly convenient from a computational point of view and conducts to the schemes:

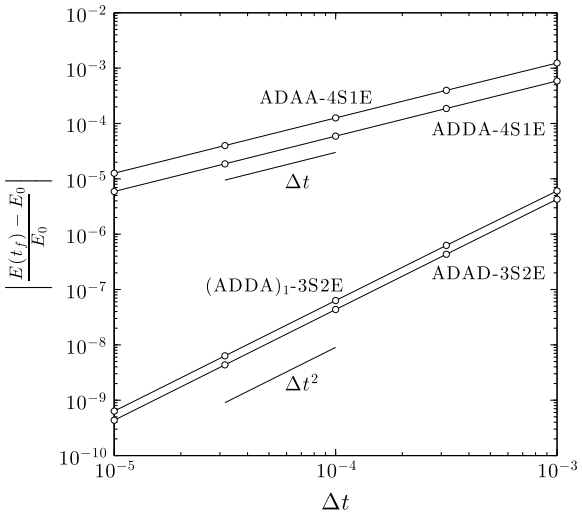

Fig. 6. Convergence of the relative error on energy conservation as a function of the time step Δt for different R–K schemes with 4 stages.

$(\mathrm{ADDA})_1$-3S2E(4)

0				
$\frac{3}{2}$	0			
$\frac{1}{3}$	0	0		
$\frac{14}{25}$	$\frac{28}{75}$	0	0	
$\frac{1}{28}$	0	$\frac{2}{3}$	$\frac{25}{84}$	

$(\mathrm{ADDA})_2$-3S2E(4)

0				
$\frac{1}{3}$	0			
$\frac{3}{2}$	0	0		
$\frac{14}{25}$	0	$\frac{28}{75}$	0	
$\frac{1}{28}$	$\frac{2}{3}$	0	$\frac{25}{84}$	

For the other sequences, solutions to the corresponding nonlinear system are in general quite difficult to obtain analytically. In all these cases, however, a symbolic solver is usually able to derive one or more families of solutions typically parametrized by two constants. From these expressions one can optimize the scheme by imposing additional requirements. An example of the result of this procedure is the scheme:

ADAD-3S2E(4)

0				
$\frac{1}{3}$	0			
$\frac{14}{25}$	$\frac{37+1/3}{100}$	0		
0	0	$\frac{1}{3}$	0	
$\frac{177+11/7}{5\times10^3}$	$\frac{1}{4}$	$\frac{25}{84}$	$\frac{5}{12}$	

which has been obtained with the aid of a symbolic manipulator and by requiring a maximum number of zeros in the Butcher array for the sequence ADAD.

The convergence of the relative error on energy conservation for this last scheme is reported in Fig. 6, where also the convergence properties of various four stage methods presented in this section are included. The scheme ADAA-4S1E is the fourth order RK scheme labeled as case III, for the value of the parameter $\theta = 1/12$. The scheme ADDA-4S1E is the 'classical' fourth order Runge–Kutta scheme belonging to case V for $\theta = 1/3$. In all cases the theoretically predicted scaling is recovered.

It is worth noting that many of the schemes presented in this section as well as in Section 5.1 can be further optimized to achieve multiple purposes, which were not taken into consideration in the present paper. For instance, the remaining degrees of freedom of the proposed 3S2E methods can be exploited to optimize the dispersion and dissipation properties in wavenumber space [18]. Moreover, low-storage implementations have not been taken into account. Note also that for some methods, $c_i > 1$ occurs (cf. DA-2S1E(2) and $(\mathrm{ADDA})_{1,2}$-3S2E(4) schemes). In such cases, care must be taken when these methods are used in conjunction with time-dependent terms [16].

5.3. Computational efficiency analysis

The performances of the various schemes derived in Section 5.2 have been assessed by means of a comparison in terms of computational efficiency, i.e., on an Error–Cost plane. This analysis is also meant to guide the reader through the selection of the best methods.

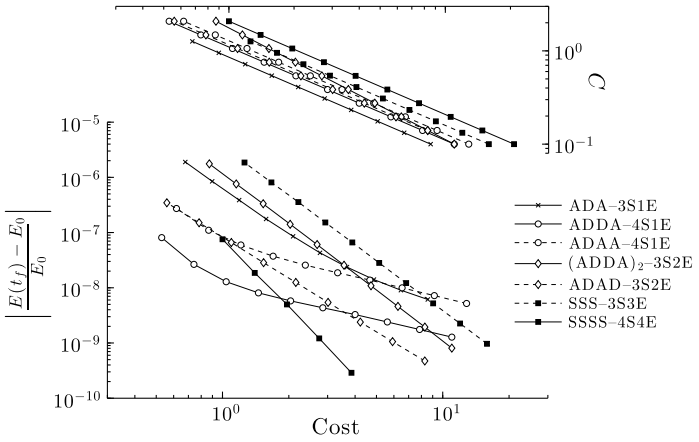

Fig. 7. Comparison in terms of computational efficiency between classical and novel schemes.

Several choices can be made regarding the error. In accordance with the aim of the paper, the kinetic energy conservation error is considered here. As a representative unit time, $t_f = t_b/2$ is chosen, where t_b is the break-time at which characteristic lines intersect for the inviscid Burgers' equation. The computational cost is calculated by means of the cost-metric presented in Appendix A.

The results are reported in the graph of Fig. 7. The plot was constructed by analyzing the energy errors in a range of Courant numbers $C = u\Delta t/\Delta x$, in particular $0.1 < C < C_{\max}$, where $C_{\max}$ is the maximum bound allowed by linear stability for each scheme. The cost is then obtained by multiplying the number of floating-point operations by the number of time-steps needed to integrate the inviscid Burgers' equation from t_0 to t_f. The left-end of each curve corresponds to the simulation run at the maximum Courant number. In the upper portion of the graph, the corresponding Courant number dependence is also reported for convenience. One 3S1E, two 4S1E and two 3S2E alternating schemes are compared to three-stage and four-stage Runge–Kutta methods in fully skew-symmetric form. For the latter, the classical Kutta and RK4 methods are considered, respectively. The computational cost for all the curves is normalized with respect to the cost of the RK4 in skew-symmetric form run at the maximum Courant number. The results were obtained by integrating the inviscid Burgers' equation with periodic boundary conditions on the domain $[0, 1]$, discretized by $N = 100$ equidistant mesh points, with initial condition $u_0(x) = \sin(\pi x)$. A fourth-order explicit central scheme was used for spatial integration. Varying the number of mesh points did not significantly affect the qualitative picture in terms of efficiency comparison among the various schemes.

Several conclusions can be drawn from the graph. For the smallest error levels, the classical RK4 in skew-symmetric form is the most efficient scheme, due to the high order of accuracy, together with the particularly compact Butcher tableau. However, there is a very large range of energy errors in which several alternating schemes are the most efficient ones. For such methods, the plot shows that the possible computational time saving can be up to 50%, although this value can be rather case dependent. In particular, the ADDA-4S1E, i.e. the alternating version of the classical RK4 scheme, results as the most efficient method. This result is particularly remarkable, since many computational codes built upon the skew-symmetric RK4 can be easily recast in a more efficient procedure by simply switching the SSSS sequence to the alternating ADDA one. By using the same arguments, it can also be concluded that the three-stage Kutta method in skew-symmetric form can be profitably substituted by its alternating counterpart. It is worth reminding that for classical skew-symmetric spatial discretizations, the global energy error is constituted only by the temporal part, while for the new methods it is a mix of spatial and temporal contributions.

Further insight can be gained from the following considerations. The crossover point in efficiency is the result of a balance between two effects: the reduction in computational cost of the alternating procedure and the higher formal order of accuracy on energy conservation of the skew-symmetric schemes. At lower values of C the formal order of accuracy prevails, while at higher values the errors of the two approaches become comparable and the maximum time saving is reached.

In the case previously discussed, the crossover point in efficiency occurs for Courant numbers greater than ≈ 0.4. It has to be underlined that this point can shift somewhere to the left or right depending on various parameters, such as the final integration time or the initial condition. However, additional numerical tests (not reported here) show that the main trends are retained in a variety of situations. The fact that the better performances of the novel schemes occur in the range of moderate values of the Courant number is advantageous, since this is the range of practical interest for many applications.

As an alternative, the performances of the ADAD-3S2E scheme are also particularly good. In addition, one has to consider that 3S2E schemes have two remaining degrees of freedom, that can be further exploited to achieve specific targets, depending on the problem under study.

6. Extensions to incompressible Navier–Stokes equations

As already stressed in the Introduction section, the proposed ideas may be applied to more realistic systems. When looking at such cases, one should first of all highlight that the developed technique rests on the assumption that $\mathbf{C}(\mathbf{u})$ is a linear function of $\mathbf{u}$. Hence, a straightforward extension can be attempted within the class of nonlinear convection models which can be expressed as the product of a linear function of the velocity field and the velocity field itself. This class is still wide and includes, most remarkably, the incompressible Navier–Stokes model.

In the case of 2D and 3D incompressible flows, the divergence free condition assures an equivalence between the continuous advective and divergence forms. The analysis can thus be conducted by following a procedure similar to that employed for the scalar 1D Burgers' equation.

An extension of the proposed technique to incompressible Navier–Stokes equations has to firstly take into account the layout of the variables on the grid. In a staggered system, the divergence free condition renders the divergence and skew symmetric forms discretely equivalent. When turning to regular or collocated grid systems this equivalence is not assured, even in the case of divergence-free velocity fields. A straightforward application of the novel procedure can hence be at first tested within a regular grid system (i.e. a grid in which both velocity components and pressure are stored at the same points). In this case, an analogous underlying structure of the discrete operators occurs and a similar time saving for the computation of the convective term can be obtained by employing only the more economical advective and divergence forms.

The overall time saving can be influenced by many other computational modules, whose relative importance is strongly case dependent. The procedure is expected to be particularly advantageous in cases in which the computation of the nonlinear term takes a relevant part of the whole algorithm. For instance, spectral methods are often used in conjunction with the skew-symmetric formulation to perform stable long-time integrations of homogeneous isotropic turbulence. In such cases, the solution of the pressure equation is practically costless and the viscous term is often treated analytically. Most of the CPU time is spent to compute the non-linear term, and a regular-type arrangement of flow variables is usually employed. Therefore, a noteworthy time-saving is expected by adopting the proposed method.

The main points of the extension to Navier–Stokes equations are outlined as follows:

1. The semidiscretization of the momentum equation contains a nonlinear term of the form $\mathbf{C}(\mathbf{u})\mathbf{u}$ where the matrix $\mathbf{C}(\mathbf{u})$ is a block-diagonal operator. Each block is relative to one spatial component of $\mathbf{u}$ and contains the derivative matrices along the three spatial directions.
2. As in the case of 1D Burgers' equation, there is a simple relation between $\mathbf{C}^{\mathrm{adv}}$ and $\mathbf{C}^{\mathrm{div}}$ which in this case is $(\mathbf{C}^{\mathrm{adv}})^T = -\mathbf{C}^{\mathrm{div}}$. This immediately implies that the energy variations due to divergence and advective forms are equal in magnitude and have opposite sign.
3. An equation for the energy error introduced over a single time step advancement of the modified RK procedure, analogue to Eq. (21), can be written for the case of inviscid, incompressible NS equations. Note that in this case the pressure gradient affects the convective operators appearing in Eq. (21).
4. A new set of nonlinear relations involving the coefficients a_{ij} and b_i can be derived by imposing the vanishing of successive terms in the expansion. Since the relation between $\mathbf{C}^{\mathrm{adv}}$ and $\mathbf{C}^{\mathrm{div}}$ in NS equation is different from that for the Burgers' equation, different systems of equation take place, leading to different optimized RK schemes.

As in the 1D Burgers' case, the redundant degrees of freedom can be exploited to optimize the dispersion and dissipation properties in wavenumber space, as well as to improve the accuracy of the pressure, as in [19].

7. Conclusions

A novel technique for cost-effective energy preserving simulations has been developed. The method consists of a time-integration strategy in which the costly skew-symmetric form is profitably split into sequences of divergence and advective forms within the stages of a properly constructed Runge–Kutta (RK) scheme. A general framework for the design of the alternating RK schemes has been established, based on a theoretical analysis of the energy error occurring for the fully-discrete problem. It is found that one and three additional constraints are necessary to achieve first- and second-order accuracy on energy conservation, respectively. The procedure has proven to be able to produce new time-saving methods with a specified order of accuracy on both solution and energy conservation. Where possible, the remaining degrees of freedom for the choice of coefficients can be properly exploited for further purposes, e.g., to minimize dissipation and dispersion errors, or to achieve a computationally efficient Butcher tableau.

An alternative procedure has also been analyzed, in which the splitting is carried *outside* the RK scheme, i.e., the divergence and advective forms are used at alternate time steps in a given sequence. Although this procedure can lead (formally) to an arbitrary order of accuracy on energy conservation, it produces an oscillatory pattern in the time-evolution of energy, in contrast to alternating RK schemes.

The performances as well as the formal order of accuracy of the new methods have been demonstrated systematically by numerical tests on the Burgers equation. The computational cost of the various schemes has been assessed by a suitably constructed cost-metric which accommodates both space discretization schemes and Runge–Kutta coefficients. On equal or

comparable performances, the new schemes can save up to 50% of CPU time for the time-advancement of the non-linear term, in comparison with standard Runge–Kutta methods.

In particular, the ADDA-4S1E, i.e. the alternating version of the classical RK4 scheme, results as one of the most efficient methods in a variety of situations.

The methods proposed in the paper have been derived for a 1D equation on uniform mesh and in the simple case of periodic boundary conditions. In this last hypothesis the derivative operator is typically skew-symmetric and this property has been used in various parts of the derivation. However, a more general formulation including the treatment of general BCs could be accomplished in the framework of SBP operators [20,21], that generalize the symmetry properties of the derivative matrices to include boundary terms. The extension of the main technique to more complex systems, e.g., to incompressible Navier–Stokes equations, has been outlined in the paper and will be the subject of future work.

Appendix A. Cost analysis

In a one-dimensional setting, an explicit central difference discretization of the non-linear convective term reads, for the ith grid point:

$$\partial_x \frac{u^2}{2}\bigg|_i \approx \sum_{l=1}^{L} w_l \left(u_{i+l}^2 - u_{i-l}^2\right) \tag{A.1}$$

$$u\partial_x u|_i \approx u_i \sum_{l=1}^{L} w_l \left(u_{i+l} - u_{i-l}\right), \tag{A.2}$$

in case of divergence and advective form, respectively. In (A.1) and (A.2), L determines the size of the computational stencil and thus the accuracy of the approximation (i.e., $2L$), while it is implicitly assumed that the coefficients of the scheme already take into account the grid spacing. On a grid of N elements, the number of floating-point operations required to calculate the above terms (*spatial* operations) is equal for the divergence and advective terms. It involves N multiplications to evaluate the products u^2 or $u \cdot \partial u$, plus L multiplications and L sums per node to calculate the difference formula, yielding a total of

$$O_{\mathrm{a,d}}^{sp} = N(1 + 2L). \tag{A.3}$$

It is assumed that additions and multiplications take roughly the same computational effort. The skew-symmetric form requires the evaluation of advective and divergence forms separately, together with a linear combination of these quantities, for a total of:

$$O_{\mathrm{s}}^{sp} = 2N(1 + 2L) + 3N. \tag{A.4}$$

If an explicit s-stage Runge–Kutta scheme is adopted for time-advancement:

$$\begin{cases} u^{n+1} = u^n - \Delta t \sum_{k=1}^{s} b_k \mathcal{N}(U^k) \\ U^k = u^n - \Delta t \sum_{j=1}^{k-1} a_{kj} \mathcal{N}(U^j), \end{cases} \tag{A.5}$$

then the number of floating-point operations contains, assuming that Δt is included within the coefficients:

- sO^{sp} evaluations of the non-linear term;
- $\hat{b}N$ sums and $\hat{b}N$ multiplications for the final update, where $\hat{b}$ is the number of non-zero b_k coefficients;
- $\hat{a}N$ multiplications between the non-zero (and non-unity) coefficients a_{kj} and the terms $\mathcal{N}(U^k)$;
- $(s-1)N$ sums between u^n and the other terms, within the stages, except for the first stage;
- $(\hat{a}-1)N$ sums between the remaining terms, within the stages.

The *total* (i.e. spatial and temporal) number of operations is then

$$O_{\mathrm{a,d}}^{tot} = 2N(s + Ls + \hat{a} + \hat{b} - 1), \tag{A.6}$$

for advective or divergence forms, while for the skew-symmetric form is

$$O_{\mathrm{s}}^{tot} = 2N(3s + 2Ls + \hat{a} + \hat{b} - 1). \tag{A.7}$$

By comparing Eqs. (A.6) and (A.7), an average factor of order 2 can be established between the approaches.

The developed cost-metric has been validated against the measured CPU times obtained with a Fortran 90 program written by the authors. Two reference Runge–Kutta schemes have been considered for validation, namely the three-stage Kutta scheme (RK3) and the classical RK4, for both the convective/divergence and skew-symmetric cases, and for a second- and fourth-order difference scheme. The measured CPU times have been averaged over 10 runs. Various parameters have been varied during the numerical experiments (e.g., the number of nodes, the size of the arrays, the number of time-steps, etc.)

Table 2
Comparison between cost-metric estimates and measured CPU times for two Runge–Kutta schemes and various spatial discretizations. The CPU times have been normalized with respect to the RK3 scheme with second-order central difference in advective/divergence form. For convenience, the number of operations is reported for $N = 1$.

Scheme	$\hat{a}$	$\hat{b}$	s	$L = 1$				$L = 2$			
				$O_{\mathrm{a,d}}^{tot}$	T_{cpu}	O_{s}^{tot}	T_{cpu}	$O_{\mathrm{a,d}}^{tot}$	T_{cpu}	O_{s}^{tot}	T_{cpu}
RK3	2	2	3	18	1.00	32	1.46	24	1.29	48	2.19
RK4	2	2	4	22	1.28	46	1.90	30	1.69	62	2.86

showing minor effect on the results. The comparison between the cost-metric and the CPU times is reported in Table 2. The agreement between the operation count and the actual CPU times is satisfactory, if one considers that the cost-metric does not take into account issues related to storage or access to array variables. The results may also depend slightly upon the optimization level of the compiler and on the architecture of the computer.

Appendix B. Nonlinear systems for 3S2E(4) schemes

AAAD

$$\begin{aligned}
&b_1 + b_2 + b_3 - \frac{b_4}{2} = 0\\
&2b_2a_{21} + 2b_3(a_{31} + a_{32}) - b_4\,(b_1 + b_2 + b_3) = 0\\
&\frac{b_4^2}{2} = 0\\
&2b_2a_{21} + 2b_3(a_{31} + a_{32}) - 2b_4(a_{41} + a_{42} + a_{43}) + (b_1 + b_2 + b_3)^2 = 0
\end{aligned} \tag{B.1}$$

AADA

$$\begin{aligned}
&b_1 + b_2 - \frac{b_3}{2} + b_4 = 0\\
&2b_2a_{21} + b_4\,(2a_{41} + 2a_{42} - a_{43}) - b_3\,(b_1 + b_2 + b_4) = 0\\
&\frac{b_3^2}{2} - 2b_4a_{43} = 0\\
&2b_2a_{21} - 2b_3(a_{31} + a_{32}) + 2b_4(a_{41} + a_{42}) + (b_1 + b_2 + b_4)^2 = 0
\end{aligned} \tag{B.2}$$

AADD

$$\begin{aligned}
&b_1 + b_2 - \frac{b_3}{2} - \frac{b_4}{2} = 0\\
&2b_2a_{21} + b_4a_{43} - (b_1 + b_2)\,(b_3 + b_4) = 0\\
&(b_3 + b_4)^2 = 0\\
&2b_2a_{21} - 2b_3(a_{31} + a_{32}) - 2b_4(a_{41} + a_{42}) + (b_1 + b_2)^2 = 0
\end{aligned} \tag{B.3}$$

ADAA

$$\begin{aligned}
&b_1 - \frac{b_2}{2} + b_3 + b_4 = 0\\
&b_3(2a_{31} - a_{32}) + b_4\,(2a_{41} - a_{42} + 2a_{43}) - b_2\,(b_1 + b_3 + b_4) = 0\\
&-\frac{b_3^2}{2} + 2b_3a_{32} + 2b_4a_{42} = 0\\
&2b_2a_{21} - 2b_3a_{31} - 2b_4(a_{41} - a_{43}) - (b_1 + b_3 + b_4)^2 = 0
\end{aligned} \tag{B.4}$$

ADAD

$$\begin{aligned}
&b_1 - \frac{b_2}{2} + b_3 - \frac{b_4}{2} = 0\\
&b_3(2a_{31} - a_{32}) + \frac{9}{8}b_4a_{42} - (b_1 + b_3)\,(b_2 + b_4) = 0\\
&4b_3a_{32} - b_4a_{42} - (b_2 + b_4)^2 = 0\\
&2b_2a_{21} - 2b_3a_{31} + 2b_4\left(a_{41} - \frac{a_{42}}{8} + a_{43}\right) - (b_1 + b_3)^2 = 0
\end{aligned} \tag{B.5}$$

DAAA

$$
\begin{aligned}
&-\frac{b_1}{2}+b_2+b_3+b_4=0\\
&b_2a_{21}+b_3(a_{31}-2a_{32})+b_4(a_{41}-2a_{42}-2a_{43})+b_1\,(b_1+b_3+b_4)=0\\
&2b_2a_{21}+2b_3a_{31}+2b_4a_{41}-\frac{b_1^2}{2}=0\\
&2b_3a_{32}+2b_4(a_{42}+a_{43})+(b_2+b_3+b_4)^2=0
\end{aligned}
\tag{B.6}
$$

DADA

$$
\begin{aligned}
&-\frac{b_1}{2}+b_2-\frac{b_3}{2}+b_4=0\\
&b_2a_{21}-b_3a_{31}+b_4(a_{41}-2a_{42}+a_{43})(b_1+b_3)(b_2+b_4)=0\\
&2b_2a_{21}+2b_4(a_{41}+a_{43})-\frac{1}{2}(b_1+b_3)^2=0\\
&2b_3a_{32}-2b_4a_{42}-(b_2+b_4)^2=0
\end{aligned}
\tag{B.7}
$$

DADD

$$
\begin{aligned}
&b_1-2b_2+b_3+b_4=0\\
&b_2a_{21}-b_3a_{31}-b_4(a_{41}+a_{43})+b_2\,(b_1+b_3+b_4)=0\\
&-2b_2a_{21}-\frac{1}{2}(b_1+b_3+b_4)^2=0\\
&2b_3a_{32}+2b_4a_{42}-b_2^2=0
\end{aligned}
\tag{B.8}
$$

DDAA

$$
\begin{aligned}
&b_1+b_2-2b_3-2b_4=0\\
&\frac{9}{8}b_2a_{21}-b_3(a_{31}+a_{32})-b_4(a_{41}+a_{42}-2a_{43})-(b_1+b_2)(b_3+b_4)=0\\
&b_2a_{21}-4b_3(a_{31}+a_{32})-4b_4(a_{41}+a_{42})+(b_1+b_2)^2=0\\
&\frac{1}{2}b_2a_{21}+4b_4a_{43}+2(b_3+b_4)^2=0
\end{aligned}
\tag{B.9}
$$

DDAD

$$
\begin{aligned}
&b_1+b_2-2b_3+b_4=0\\
&\frac{9}{8}b_2a_{21}-b_3(a_{31}+a_{32})+\frac{9}{8}b_4(a_{41}+a_{42})-b_3(b_1+b_2+b_4)=0\\
&b_2a_{21}-4b_3(a_{31}+a_{32})+b_4(a_{41}+a_{42})+(b_1+b_2+b_4)^2=0\\
&\frac{1}{2}b_2a_{21}+b_4(\frac{a_{41}}{2}+\frac{a_{42}}{2}-4a_{43})+2b_3^2=0
\end{aligned}
\tag{B.10}
$$

DDDA

$$
\begin{aligned}
&b_1+b_2+b_3-2b_4=0\\
&b_2a_{21}-b_3a_{31}-b_4(a_{41}+a_{43})+b_2\,(b_1+b_3+b_4)=0\\
&-2b_2a_{21}-\frac{1}{2}(b_1+b_3+b_4)^2=0\\
&2b_3a_{32}+2b_4a_{42}-b_2^2=0
\end{aligned}
\tag{B.11}
$$

ADDD

$$
\begin{aligned}
&-2b_1+b_2+b_3+b_4=0\\
&\frac{9}{8}b_3a_{32}+\frac{9}{8}b_4\,(a_{42}+a_{43})-b_1\,(b_2+b_3+b_4)=0\\
&b_3a_{32}+b_4\,(a_{42}+a_{43})+(b_2+b_3+b_4)^2=0\\
&4b_2a_{21}+b_3\left(4a_{31}-\frac{a_{32}}{2}\right)+b_4\left(4a_{41}-\frac{a_{42}}{2}-\frac{a_{43}}{2}\right)-2b_1^2=0
\end{aligned}
\tag{B.12}
$$

DAAD

$$
\begin{aligned}
&-b_1 + 2b_2 + 2b_3 - b_4 = 0 \\
&b_2 a_{21} + b_3 (a_{31} - 2a_{32}) - b_4 a_{41} + (b_1 + b_4)(b_2 + b_3) = 0 \\
&2b_2 a_{21} + 2b_3 a_{31} - \frac{1}{2}(b_1 + b_4)^2 = 0 \\
&2b_3 a_{32} + 2b_4 (a_{42} + a_{43}) + (b_2 + b_3)^2 = 0
\end{aligned}
\tag{B.13}
$$

ADDA

$$
\begin{aligned}
&2b_1 - b_2 - b_3 + 2b_4 = 0 \\
&\frac{9}{8} b_3 a_{32} + b_4 (2a_{41} - a_{42} - a_{43}) - (b_1 + b_4)(b_2 + b_3) = 0 \\
&b_3 a_{32} - 4b_4 (a_{42} + a_{43}) + (b_2 + b_3)^2 = 0 \\
&4b_2 a_{21} + b3 \left(4a_{31} - \frac{a_{32}}{2}\right) - 4b_4 4a_{41} - 2(b_1 + b_4)^2 = 0
\end{aligned}
\tag{B.14}
$$

References

[1] B. Koren, R. Abgrall, P. Bochev, J. Frank, B. Perot, Physics-compatible numerical methods, J. Comput. Phys. 257 (2013) 1039.
[2] R. Mittal, P. Moin, Suitability of upwind-biased finite difference schemes for large-eddy simulation of turbulent flows, J. Comput. Phys. 35 (1997) 1415–1417.
[3] W.J. Feiereisen, W.C. Reynolds, J.H. Ferziger, Numerical simulation of a compressible, homogeneous, turbulent shear flow, Tech. rep., Stanford University, 1981.
[4] A.G. Kravhcenko, P. Moin, On the effect of numerical errors in large eddy simulations of turbulent flows, J. Comput. Phys. 131 (1997) 310–322.
[5] A.E. Honein, P. Moin, Higher entropy conservation and numerical stability of compressible turbulence simulations, J. Comput. Phys. 201 (2) (2004) 531–545.
[6] N.N. Mansour, P. Moin, W.C. Reynolds, J.H. Ferziger, Improved methods for large eddy simulations of turbulence, Turbul. Shear Flows 1 (1979) 386–401.
[7] T. Dubois, J.A. Domaradzki, A. Honein, The subgrid-scale estimation model applied to large eddy simulations of compressible turbulence, Phys. Fluids 14 (2002) 1781.
[8] G.A. Blaisdell, E.T. Spyropoulos, J.H. Qin, The effect of the formulation of nonlinear terms on aliasing errors in spectral methods, Appl. Numer. Math. 21 (3) (1996) 207–219.
[9] T.A. Zang, On the rotation and skew-symmetric forms for incompressible flow simulations, Appl. Numer. Math. 7 (1991) 27–40.
[10] J.B. Perot, Discrete conservation properties of unstructured mesh schemes, Annu. Rev. Fluid Mech. 43 (2011) 299–318.
[11] G.E. Karniadakis, S. Sherwin, Spectral/hp Element Methods for Computational Fluid Dynamics, Oxford University Press, 2005.
[12] R.M. Kerr, Higher-order derivative correlations and the alignment of small-scale structures in isotropic numerical turbulence, J. Fluid Mech. 153 (1985) 31–58.
[13] Y. Morinishi, T.S. Lund, O.V. Vasilyev, P. Moin, Fully conservative higher order finite difference schemes for incompressible flows, J. Comput. Phys. 143 (1998) 90–124.
[14] R.W.C.P. Verstappen, A.E.P. Veldman, Symmetry-preserving discretization of turbulent flow, J. Comput. Phys. 187 (2003) 343–368.
[15] F. Trias, O. Lehmkuhl, A. Oliva, C. Pérez-Segarra, R. Verstappen, Symmetry-preserving discretization of Navier–Stokes equations on collocated unstructured grids, J. Comput. Phys. 258 (2014) 246–267.
[16] J. Butcher, Numerical Methods for Ordinary Differential Equations, Wiley, 2004.
[17] B. Sanderse, Energy-conserving Runge–Kutta methods for the incompressible Navier–Stokes equations, J. Comput. Phys. 233 (2013) 100–131.
[18] F.Q. Hu, M.Y. Hussaini, J.L. Manthey, Low-dissipation and low-dispersion Runge–Kutta schemes for computational acoustics, J. Comput. Phys. 124 (1996) 177–191.
[19] B. Sanderse, B. Koren, Accuracy analysis of explicit Runge–Kutta methods applied to the incompressible Navier–Stokes equations, J. Comput. Phys. 231 (2012) 3041–3063.
[20] D.C.D.R. Fernandez, J.E. Hicken, D.W. Zingg, Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations, Comput. Fluids 95 (2014) 171–196.
[21] M. Svard, J. Nordstrom, Review of summation-by-parts schemes for initial-boundary-value problems, J. Comput. Phys. 268 (2014) 17–38.