

# Journal Pre-proof

Energy-preserving exponential integrators of arbitrarily high order for conservative or dissipative systems with highly oscillatory solutions

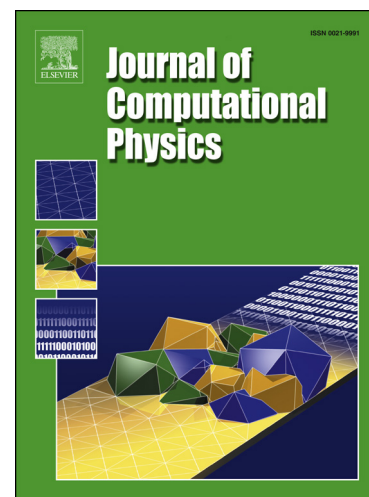
Lijie Mei, Li Huang and Xinyuan Wu

PII: S0021-9991(21)00324-7

DOI: <https://doi.org/10.1016/j.jcp.2021.110429>

Reference: YJCPH 110429

To appear in: *Journal of Computational Physics*



Please cite this article as: L. Mei, L. Huang and X. Wu, Energy-preserving exponential integrators of arbitrarily high order for conservative or dissipative systems with highly oscillatory solutions, *Journal of Computational Physics*, 110429, doi: <https://doi.org/10.1016/j.jcp.2021.110429>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier.

**Highlights**

- A general framework to design arbitrarily high order energy-preserving exponential integrators for first-order skew-gradient systems is presented.
- The fourth-order energy-preserving exponential integrator is constructed and tested in detail.
- The remarkable numerical performance demonstrates the high efficiency and good energy preservation of the proposed integrator.

# Energy-preserving exponential integrators of arbitrarily high order for conservative or dissipative systems with highly oscillatory solutions

Lijie Mei<sup>a</sup>, Li Huang<sup>b</sup>, Xinyuan Wu<sup>c,d,\*</sup>

<sup>a</sup>*School of Mathematics and Computer Science, Shangrao Normal University, Shangrao 334001, P.R.China*

<sup>b</sup>*School of Physics and Electronic Information, Shangrao Normal University, Shangrao 334001, P.R.China*

<sup>c</sup>*Department of Mathematics, Nanjing University, Nanjing 210093, P.R.China*

<sup>d</sup>*School of Mathematical Sciences, Qufu Normal University, Qufu 273165, P.R.China*

---

## Abstract

Taking into account the limited accuracy of the energy-preserving exponential integrator of order two [Li & Wu, SIAM J. Sci. Comput. 38 (2016) A1876-A1895] for conservative or dissipative systems with highly oscillatory solutions, this paper is devoted to presenting a uniform framework to design energy-preserving exponential integrators of arbitrarily high order based on the modifying integrator theory. To this end, we first show that the second-order energy-preserving exponential integrator is a B-series method. Using the adapted substitution law, we then prove that there exist arbitrary order energy-preserving exponential integrators and show how to design arbitrarily high-order integrators by finding the truncated modified differential equations. As an example, the fourth-order energy-preserving exponential integrator is constructed in detail. The stability and convergence of the proposed integrators are analyzed as well. Finally, numerical experiments are accompanied, including both ODEs and PDEs, and the numerical results demonstrate the remarkable superiority over the existing energy-preserving integrators for highly oscillatory systems in the literature.

**Keywords:** Energy-preserving exponential integrators, highly oscillatory systems, modified differential equations, B-series

---

Mathematics Subject Classification (2010): 65L04, 65L05, 65L06, 65P10

## 1. Introduction

It is known that traditional numerical methods need a very small stepsize and hence a long runtime to achieve an acceptable accuracy for differential equations with highly oscillatory solutions. Structure-preserving integrators have been demonstrated to possess great superiority over the traditional numerical integrators in dealing with problems of certain mathematical or physical structures. The best known structure-preserving integrators are symplectic methods for Hamiltonian systems [18, 21, 36, 37], and energy-preserving schemes for linear-gradient systems with Lyapunov functions or first integrals [4, 5, 17, 20, 29, 30, 35, 40, 41, 43]. Besides, exponential integrators were also proposed as a new class of structure-preserving methods, specially for efficiently dealing with stiff or highly oscillatory problems [1, 3, 23, 24, 27], since they can preserve the stiffness or the high-frequency oscillation of the original problem. Recent researches related to exponential integrators cover both aspects for first-order stiff problems (see, e.g. [3, 37]) and for second-order highly oscillatory problems (see, e.g. [36, 41]), including nonlinear Klein–Gordon equations. In particular, for certain hyperbolic PDEs, such as the “good” Boussinesq equation, quite a few exponential integrators with Fourier pseudo-spectral method for space discretization are investigated in the literature [11, 39, 42, 50, 52]. Readers are referred to [24, 31, 32, 45, 46, 49] for more details on exponential integrators.

Recently, the authors of [9, 33, 38] have shown that the energy-preserving methods can also preserve the correct monotonic decrease of energy for dissipative systems, except for their exact energy preservation for conservative systems. Given that energy-preserving schemes for Hamiltonian wave equations are easily to prove the stability, Li & Wu

---

\*Corresponding author

Email addresses: bxhanm@126.com (Lijie Mei), hjnh1dtyf@126.com (Li Huang), xywu@nju.edu.cn (Xinyuan Wu)

[29] proposed the second-order energy-preserving exponential integrator for the skew-gradient system by combining ideas of exponential integrators and energy-preserving methods. The initial-value problem of skew-gradient systems is expressed by

$$\begin{cases} y'(t) = Q(Ay + \nabla U(y)), & t \in [t_0, T_{end}], \\ y(t_0) = y_0, \end{cases} \quad (1)$$

where  $Q \in \mathbb{R}^{d \times d}$  is a skew symmetric or negative semidefinite matrix,  $A \in \mathbb{R}^{d \times d}$  is a symmetric matrix that possesses the dominant stiffness or high-frequency oscillation of the system with  $\|A\|_2 \gg 1$ , and  $U(y)$  is a smooth nonlinear potential function satisfying  $\|A\|_2 \gg \|Hess(U)\|_2$ . System (1) is evidently conservative provided that  $Q$  is skew-symmetric, while it is a dissipative system when  $Q$  is negative semidefinite (see [29] for details).

Here, we remark that the second-order energy-preserving exponential integrator in [29] is identical to the energy-preserving adapted average vector field (AAVF) integrator in [44] when applied to the second-order highly oscillatory Hamiltonian system. Very recently, Jiang *et al.* [25] proposed a linearly implicit energy-preserving exponential integrator for the nonlinear Klein–Gordon equation by utilizing the scalar auxiliary variable (SAV) approach. By applying symplectic exponential Runge–Kutta methods [37] to the reformulated nonlinear Schrödinger equation with the SAV transformation, Cui *et al.* [16] showed that the numerical schemes can preserve the modified mass and energy.

However, it should be noted that the proposed energy-preserving exponential integrator in [29] only has the limited second-order accuracy that cannot meet the requirement of high precision especially for stiff or highly oscillatory problems. The numerical schemes obtained in [16, 25] can only preserve the reformulated quadratic energy, but cannot preserve the original Hamiltonian energy as claimed in [25]. That is, the approach to constructing arbitrarily high-order exponential integrators that can preserve the original energy for first-order systems is still absent and not reported yet in the literature at least to our knowledge. These facts motivate us to make an intensive study of a general approach to constructing arbitrarily high-order energy-preserving exponential integrators with higher efficiency. To achieve this goal, we will study the *modifying integrator* of the second-order energy-preserving integrator [29] which is shown to be a B-series method (see [10]) in this paper. This strategy has been used by Chartier *et al.* in [13] and is very useful.

Finally, two points are worthy of being emphasized. The first point is that once applied to dissipative systems, the energy-preserving methods do not preserve the energy but preserve the decrease of the energy. Hence, it may be better to use “energy-stable integrators” instead of “energy-preserving integrators” for dissipative systems. Throughout this paper, we just confirm to the traditional use of energy-preserving integrators for both conservative and dissipative systems, but bear in mind that the same integrator displays different energy behaviors for different systems. The second point is that, in spite of the implicit energy-preserving methods mentioned above, there exist explicit energy-preserving methods, such as high-order explicit RK methods with the projection technique that can preserve the quadratic invariant [7, 8]. Recently, by using the invariant energy quadratization approach and reformulating Hamiltonian ODEs [51] or PDEs [26], explicit high-order energy-preserving methods that preserve the quadratic energy of the reformulated system can be constructed. It is noted that special techniques such as the projection or the reformulation of the original system are adopted in these explicit methods. In this sense, the study on “pure” energy-preserving methods that can preserve the energy of general Hamiltonian systems without special treatments for the numerical solutions of the original system is still desirable.

This paper is organised as follows. In Section 2, we present some fundamental results on exponential integrators, including the formulation of the second-order energy-preserving exponential integrator and the adapted substitution law. In Section 3, we first derive the fourth-order energy-preserving exponential integrator with the help of the results established in Section 2. Then, arbitrarily high-order energy-preserving exponential integrators are formulated uniformly. Section 4 is concerned with the analyses of the convergence of the fixed-point iteration, the stability and convergence of the proposed integrators. Numerical experiments are conducted in Section 5, and the numerical results greatly support our theoretical analysis. That is, our energy-preserving exponential integrators are more efficient in many settings than existing methods in the literature for the computation of highly oscillatory problems. Conclusions are drawn in the last section. Throughout this paper, we admit that  $\|Q\|_2 = 1$  because the case  $\|Q\|_2 \neq 1$  can be normalized by the changing  $Q \rightarrow Q/\|Q\|_2$ ,  $A \rightarrow \|Q\|_2 \cdot A$  and  $U \rightarrow \|Q\|_2 \cdot U$ . We use the uniform notation  $\|\cdot\|$  instead of  $\|\cdot\|_2$  since the two-norm is used throughout this paper.

## 2. Preliminaries

Some preliminaries concerning energy-preserving exponential integrators are needed. For convenience, we rewrite (1) as follows

$$\begin{cases} y'(t) = My + f(y), & t \in [t_0, T_{end}], \\ y(t_0) = y_0, \end{cases} \quad (2)$$

where  $M = QA$  and  $f(y) = Q\nabla U(y)$ . Then the second-order exponential averaged vector field (EAVF) method for (2) is given by

$$\begin{aligned} y_{n+1} &= \exp(hM)y_n + h\varphi(hM)Q \int_0^1 \nabla U((1-\xi)y_n + \xi y_{n+1})d\xi \\ &= \exp(hM)y_n + h\varphi(hM) \int_0^1 f((1-\xi)y_n + \xi y_{n+1})d\xi, \end{aligned} \quad (3)$$

with the stepsize  $h$ . Here, the scalar function is defined as

$$\varphi(0) = 1, \quad \varphi(z) = \frac{\exp(z) - 1}{z} = \sum_{j=0}^{\infty} \frac{z^j}{(j+1)!}. \quad (4)$$

Note that the EAVF method is just the energy-preserving exponential integrator emphatically proposed in [29]. In fact, the average vector field  $\int_0^1 \nabla U((1-\xi)y_n + \xi y_{n+1})d\xi$  can also be replaced by other discrete gradients, such as the midpoint discrete gradient and the coordinate increment discrete gradient [35], and then the new integrator retains the energy preservation property as well. In this paper, we still focus on the EAVF method due to its wide application and easy manipulation in practice.

Let  $\tilde{f}(y) = My + f(y)$  in (2), and then we denote the EAVF integrator (3) and the exact phase flow of (2) by  $\Phi_{\tilde{f}}^h(y)$  and  $\varphi_{\tilde{f}}^h(y)$ , respectively. Using the idea of *modifying integrator* [13], we attempt to find the modified differential equation

$$y' = \tilde{g}(y) = My + g(y) = My + f(y) + hf_2(y) + h^2f_3(y) + \cdots, \quad (5)$$

from which the exact solution of (2) can be obtained by applying the EAVF integrator (3) to the system (5), namely,  $\Phi_{\tilde{g}}^h(y) = \varphi_{\tilde{f}}^h(y)$ . Then the  $r$ th-order modifying integrator  $\Phi_{\tilde{g}^{[r]}}^h(y)$  can be obtained by applying the EAVF integrator (3) to the  $r$ th-order truncated modified vector field  $\tilde{g}^{[r]}(y)$ :

$$\tilde{g}^{[r]}(y) = My + f(y) + hf_2(y) + h^2f_3(y) + \cdots + h^{r-1}f_r(y). \quad (6)$$

The result from [13] guarantees that  $\Phi_{\tilde{g}^{[r]}}^h(y)$  possesses the same energy preservation as the EAVF integrator (3).

To establish the general formula of the modified differential equation (5) conveniently, the substitution law based on B-series for exponential methods is very important, as stated in [12]. Unfortunately, however, we note that the substitution law in [12] cannot be directly applied to exponential integrators. Hence, we adapt the substitution law in [12] by modifying the key definition regarding the skeleton of a partition as follows. We present some auxiliary definitions and lemmas in Appendix A (see also in [1]), with which the omitted proof of the following Theorem 2.3 can be accomplished similarly to that in [12].

**Definition 2.1.** (See [1, 12, 21].) For a mapping  $a : T \cup \{\emptyset\} \rightarrow \mathbb{R}$ , a formal series of the form

$$B_f(a, y) = a(\emptyset)y + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) \mathcal{F}_f(\tau)y, \quad (7)$$

is called a B-series.

**Definition 2.2. (Skeleton of a partition).** The skeleton  $\chi(p^\tau) \in T$  of a partition  $p^\tau \in \mathcal{P}(\tau)$  of a bi-colored tree  $\tau \in T$  is the tree obtained by replacing each tree of  $P(p^\tau)$  in  $p^\tau$  with a colored single vertex and dashed edges with solid ones. The color of the single vertex is determined by the tree in  $P(p^\tau)$ , namely, the color is white provided that all the vertices of the tree are white vertices (i.e., the tree belongs to  $T_0$ ), otherwise, the color is black.

**Theorem 2.3.** Let  $a, b : T \cup \{\emptyset\} \rightarrow R$  be two mappings with  $b(\emptyset) = 0$ . Given a vector field  $\tilde{f}(y) = My + f(y) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , consider the ( $h$ -dependent) vector field  $\tilde{g}(y) = My + g(y) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  defined by

$$h\tilde{g}(y) = B_f(b, y). \quad (8)$$

Then there exists a mapping  $b \star a : T \cup \{\emptyset\} \rightarrow R$  satisfying

$$B_g(a, y) = B_f(b \star a, y), \quad (9)$$

and  $b \star a$  is defined by

$$b \star a(\emptyset) = a(\emptyset), \quad \forall \tau \in T, \quad b \star a(\tau) = \sum_{p^\tau \in \mathcal{P}(\tau)} a(\chi(p^\tau)) \prod_{\delta \in P(p^\tau)} b(\delta). \quad (10)$$

**Remark 2.4.** Although they share the seemingly same form, the adapted substitution law in Theorem 2.3 contains more details than that in [12–14], since the set of rooted trees considered in [12–14] is just the subset of bi-colored trees in this paper. Therefore, Theorem 2.3 can be naturally thought of as an extension of that in [12–14].

The following two theorems are important, which show that the exact solution of (2) and the second-order EAVF integrator (3) can be expressed by B-series.

**Theorem 2.5.** (See [1].) The exact solution  $y(t_0 + h)$  of (2) can be expressed by a B-series:

$$y(t_0 + h) = B_f(e, y_0) = e(\emptyset)y_0 + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} e(\tau) \mathcal{F}_f(\tau)(y_0), \quad (11)$$

where the mapping  $e$  is defined as follows:

- (i)  $e(\emptyset) = e(\bullet) = e(\circ) = 1$ ;
- (ii)  $e(\tau) = \frac{1}{|\tau|} e(\tau_1), \quad \tau = W_+(\tau_1)$ ;
- (iii)  $e(\tau) = \frac{1}{|\tau|} e(\tau_1) \cdots e(\tau_k), \quad \tau = B_+(\tau_1, \dots, \tau_k)$ .

**Theorem 2.6.** The second-order EAVF integrator (3) can be expanded into a B-series:

$$\Phi_f^h(y) = B_f(a, y) = a(\emptyset)y + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) \mathcal{F}_f(\tau)(y), \quad (12)$$

where  $T$  is the set of bi-colored trees and the mapping  $a$  is defined by

- (i)  $a(\emptyset) = a(\bullet) = a(\circ) = 1$ ;
- (ii)  $a(\tau) = \frac{1}{|\tau|!}, \quad \tau \in T_0$ ;
- (iii)  $a(\tau) = \frac{a(\tau_1) \cdots a(\tau_k)}{(k+1)(j+1)!}, \quad \tau = W_+^j(B_+(\tau_1, \dots, \tau_k)) \in T \setminus T_0$ .

*Proof.* Let

$$\tilde{y} = \Phi_f^h(y) = \exp(hM)y + h\varphi(hM) \int_0^1 f((1-\xi)y + \xi\tilde{y})d\xi. \quad (13)$$

As is known,  $\tilde{y}$  can be expanded into a series of  $h$  at the point  $y$ . Then,  $\tilde{y}$  can be expanded into a B-series defined on the bi-colored trees  $T$ , and we write it as

$$\tilde{y} = B_f(a, y) = y + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) \mathcal{F}_f(\tau)(y). \quad (14)$$

Furthermore,  $(1 - \xi)y + \xi\tilde{y}$  can be expressed by the following B-series

$$(1 - \xi)y + \xi\tilde{y} = y + \xi(\tilde{y} - y) = y + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} c(\tau) \mathcal{F}_f(\tau)(y), \quad (15)$$

where  $c(\tau) = \xi a(\tau)$ .

It follows from Lemma 2.1 of Ref. [2] or Lemma 1.9 in Chapter III of Ref. [21] that

$$hf((1 - \xi)y + \xi\tilde{y}) = \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} c'(\tau) \mathcal{F}_f(\tau)(y), \quad (16)$$

where  $c'(\bullet) = 1$ ,  $c'(B_+(\tau_1, \dots, \tau_k)) = c(\tau_1) \cdots c(\tau_k) = \xi^k a(\tau_1) \cdots a(\tau_k)$  and  $c'(\tau) = 0$  for  $\tau \in T_w \cup \emptyset$ . In consequence, we have

$$h \int_0^1 f((1 - \xi)y + \xi\tilde{y}) d\xi = \int_0^1 \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} c'(\tau) \mathcal{F}_f(\tau)(y) d\xi = \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a'(\tau) \mathcal{F}_f(\tau)(y), \quad (17)$$

where  $a'(\bullet) = 1$ ,  $a'(B_+(\tau_1, \dots, \tau_k)) = \int_0^1 c'(B_+(\tau_1, \dots, \tau_k)) d\xi = \frac{a(\tau_1) \cdots a(\tau_k)}{k+1}$ , and  $a'(\tau) = 0$  for  $\tau \in T_w \cup \emptyset$ .

It is clear that

$$\exp(hM) = \sum_{j=0}^{\infty} \frac{h^j M^j}{j!}, \quad \varphi(hM) = \sum_{j=0}^{\infty} \frac{h^j M^j}{(j+1)!}. \quad (18)$$

It then follows from (13), (17) and (18) that

$$\begin{aligned} B_f(a, y) &= \sum_{j=0}^{\infty} \frac{h^j M^j \cdot y}{j!} + \sum_{j=0}^{\infty} \frac{h^j M^j}{(j+1)!} \sum_{\tau \in T_b} \frac{h^{|\tau|}}{\sigma(\tau)} a'(\tau) \mathcal{F}_f(\tau)(y) \\ &= y + \sum_{\tau \in T_0} \frac{h^{|\tau|}}{|\tau|!} \mathcal{F}_f(\tau)(y) + \sum_{\substack{\tilde{\tau} = W_+^j(\tau) \in T \setminus T_0, \\ j \geq 0, \tau \in T_b}} \frac{h^{|\tilde{\tau}|}}{\sigma(\tau)} \cdot \frac{a'(\tau)}{(j+1)!} \mathcal{F}_f(\tilde{\tau})(y). \end{aligned} \quad (19)$$

The last equality follows from the definition of elementary differential in Appendix A. Noting that  $\sigma(\tau) = 1$  holds for  $\tau \in T_0$ , and  $\sigma(\tilde{\tau}) = \sigma(\tau)$  holds for  $\tilde{\tau} = W_+^j(\tau)$ , we then obtain the required property from (19) for the mapping  $a$ .  $\square$

### 3. Energy-preserving exponential integrators of arbitrarily high order

This section concerns energy-preserving exponential integrators of arbitrarily high order for the system (2) and uses the adapted substitution law established in the previous section.

Suppose that  $y'(t) = \tilde{g}(y) = My + g(y)$  is the modified differential equation, from which the exact solution of (2) can be achieved by applying the EAVF integrator (3), that is,  $\Phi_g^h(y) = \varphi_{\tilde{f}}^h(y)$ . Then  $\Phi_g^h(y)$  can be expanded into a B-series, i.e.,  $\Phi_g^h(y) = B_g(a, y)$ . Using  $\varphi_{\tilde{f}}^h(y) = B_f(e, y)$  by Theorem 2.5, we formally obtain that  $B_g(a, y) = B_f(e, y)$ . If  $\tilde{g}(y)$  is also defined as a B-series, namely,  $h\tilde{g}(y) = B_f(b, y)$ , then it follows from Theorem 2.3 that

$$B_f(b \star a, y) = B_g(a, y) = B_f(e, y). \quad (20)$$

We can derive the mapping  $b$  by solving the equations  $b \star a(\tau) = e(\tau)$  for all  $\tau \in T$  since the two mappings  $a$  and  $e$  have been known from Theorem 2.5 and Theorem 2.6. Consequently, the modified vector field  $\tilde{g}(y)$  is obtained as  $\tilde{g}(y) = \frac{1}{h} B_f(b, y)$ . Then, a truncation in appropriate terms for  $\tilde{g}(y)$  shown in (6) will provide high-order exponential integrators which preserve the energy.

We list the explicit formulae determined by (20) in Table 1, where the order of considered bi-colored trees is up to 3. We also list  $a(\tau)$  and  $e(\tau)$  of bi-colored trees of order  $\leq 3$  in Table 2. By solving the equations in Table 1, we can recursively obtain the value of  $b(\tau)$ , which are also listed in Table 2. It is noted that because the EAVF integrator is symmetric,  $\tilde{g}(y)$  will be in even powers of  $h$  (Theorem 2.2 in Page 342 in [21], see also in [13]), that is,  $f_j(y) = 0$  in (6) once  $j$  is even. This means that  $b(\tau) = 0$  holds for all trees of even order.

$b \star a(\emptyset) = a(\emptyset) = e(\emptyset)$
$b \star a(\bullet) = a(\bullet)b(\bullet) = e(\bullet)$
$b \star a(\circ) = a(\circ)b(\circ) = e(\circ)$
$b \star a(\uparrow) = a(\bullet)b(\uparrow) + a(\uparrow)b(\bullet)^2 = e(\uparrow)$
$b \star a(\updownarrow) = a(\bullet)b(\updownarrow) + a(\updownarrow)b(\circ)b(\bullet) = e(\updownarrow)$
$b \star a(\downarrow) = a(\bullet)b(\downarrow) + a(\downarrow)b(\circ)b(\bullet) = e(\downarrow)$
$b \star a(\circlearrowleft) = a(\circ)b(\circlearrowleft) + a(\circlearrowleft)b(\circ)^2 = e(\circlearrowleft)$
$b \star a(\circlearrowright) = a(\bullet)b(\circlearrowright) + 2a(\uparrow)b(\bullet)b(\uparrow) + a(\circlearrowright)b(\bullet)^3 = e(\circlearrowright)$
$b \star a(\circlearrowleft\circlearrowright) = a(\bullet)b(\circlearrowleft\circlearrowright) + a(\uparrow)b(\bullet)b(\updownarrow) + a(\updownarrow)b(\circ)b(\updownarrow) + a(\circlearrowleft\circlearrowright)b(\circ)b(\bullet)^2 = e(\circlearrowleft\circlearrowright)$
$b \star a(\circlearrowright\circlearrowleft) = a(\bullet)b(\circlearrowright\circlearrowleft) + 2a(\updownarrow)b(\circ)b(\updownarrow) + a(\circlearrowright\circlearrowleft)b(\bullet)b(\circ)^2 = e(\circlearrowright\circlearrowleft)$
$b \star a(\circlearrowleft\circlearrowright\circlearrowleft) = a(\bullet)b(\circlearrowleft\circlearrowright\circlearrowleft) + 2a(\updownarrow)b(\bullet)b(\updownarrow) + a(\circlearrowleft\circlearrowright\circlearrowleft)b(\bullet)^3 = e(\circlearrowleft\circlearrowright\circlearrowleft)$
$b \star a(\circlearrowright\circlearrowleft\circlearrowright) = a(\bullet)b(\circlearrowright\circlearrowleft\circlearrowright) + a(\updownarrow)b(\circ)b(\updownarrow) + a(\updownarrow)b(\bullet)b(\updownarrow) + a(\circlearrowright\circlearrowleft\circlearrowright)b(\circ)b(\bullet)^2 = e(\circlearrowright\circlearrowleft\circlearrowright)$
$b \star a(\circlearrowleft\circlearrowright\circlearrowright) = a(\bullet)b(\circlearrowleft\circlearrowright\circlearrowright) + a(\updownarrow)b(\bullet)b(\updownarrow) + a(\updownarrow)b(\circ)b(\updownarrow) + a(\circlearrowleft\circlearrowright\circlearrowright)b(\circ)b(\bullet)^2 = e(\circlearrowleft\circlearrowright\circlearrowright)$
$b \star a(\circlearrowright\circlearrowleft\circlearrowleft) = a(\bullet)b(\circlearrowright\circlearrowleft\circlearrowleft) + a(\updownarrow)b(\circ)b(\updownarrow) + a(\updownarrow)b(\bullet)b(\updownarrow) + a(\circlearrowright\circlearrowleft\circlearrowleft)b(\circ)b(\bullet)^2 = e(\circlearrowright\circlearrowleft\circlearrowleft)$
$b \star a(\circlearrowleft\circlearrowright\circlearrowright\circlearrowright) = a(\bullet)b(\circlearrowleft\circlearrowright\circlearrowright\circlearrowright) + a(\updownarrow)b(\bullet)b(\updownarrow) + a(\updownarrow)b(\circ)b(\updownarrow) + a(\circlearrowleft\circlearrowright\circlearrowright\circlearrowright)b(\circ)^2b(\bullet) = e(\circlearrowleft\circlearrowright\circlearrowright\circlearrowright)$
$b \star a(\circlearrowright\circlearrowleft\circlearrowright\circlearrowright) = a(\bullet)b(\circlearrowright\circlearrowleft\circlearrowright\circlearrowright) + a(\updownarrow)b(\bullet)b(\updownarrow) + a(\updownarrow)b(\circ)b(\updownarrow) + a(\circlearrowright\circlearrowleft\circlearrowright\circlearrowright)b(\circ)^2b(\bullet) = e(\circlearrowright\circlearrowleft\circlearrowright\circlearrowright)$
$b \star a(\circlearrowleft\circlearrowright\circlearrowleft\circlearrowright) = a(\bullet)b(\circlearrowleft\circlearrowright\circlearrowleft\circlearrowright) + a(\updownarrow)b(\bullet)b(\updownarrow) + a(\updownarrow)b(\circ)b(\updownarrow) + a(\circlearrowleft\circlearrowright\circlearrowleft\circlearrowright)b(\circ)^2b(\bullet) = e(\circlearrowleft\circlearrowright\circlearrowleft\circlearrowright)$
$b \star a(\circlearrowright\circlearrowleft\circlearrowleft\circlearrowleft) = a(\bullet)b(\circlearrowright\circlearrowleft\circlearrowleft\circlearrowleft) + 2a(\updownarrow)b(\circ)b(\updownarrow) + a(\circlearrowright\circlearrowleft\circlearrowleft\circlearrowleft)b(\circ)^3 = e(\circlearrowright\circlearrowleft\circlearrowleft\circlearrowleft)$

Table 1: The formulae of the adapted substitution law for bi-colored trees of order  $\leq 3$ .

According to the coefficients of  $b(\tau)$  obtained in Table 2, the fourth-order truncation  $\tilde{g}^{[4]}(y)$  of  $\tilde{g}(y)$  is expressed as follows:

$$\begin{aligned}
\tilde{g}^{[4]}(y) &= My + f(y) - \frac{h^2}{2} \left( \mathcal{F}_f(\circlearrowright) + \mathcal{F}_f(\circlearrowleft\circlearrowright) + \mathcal{F}_f(\circlearrowright\circlearrowleft) + \mathcal{F}_f(\circlearrowleft\circlearrowright\circlearrowleft) + \mathcal{F}_f(\circlearrowright\circlearrowleft\circlearrowright) + \mathcal{F}_f(\circlearrowleft\circlearrowright\circlearrowright\circlearrowright) \right) \\
&= \left( E_0 - \frac{h^2}{12} (MF + FM + FF) \right) Q (Ay + \nabla U(y)) \\
&= \tilde{Q} (Ay + \nabla U(y)),
\end{aligned} \tag{21}$$

where  $E_0$  is the identity matrix of the same dimension as  $M$ ,  $F = \frac{\partial f}{\partial y}$  is the Jacobian of  $f(y)$  evaluated at  $y$ , and  $\tilde{Q} = \left( E - \frac{h^2}{12} (MF + FM + FF) \right) Q$ . Applying the EAVF rule to the system  $y'(t) = \tilde{g}^{[4]}(y) = \tilde{Q} (Ay + \nabla U(y))$ , we obtain the following numerical scheme

$$y_{n+1} = \exp(h\tilde{M})y_n + h\varphi(h\tilde{M})\tilde{Q} \int_0^1 \nabla U((1-\xi)y_n + \xi y_{n+1})d\xi, \tag{22}$$

where  $\tilde{M} = \tilde{Q}A$ . Note that  $\tilde{Q}$  and  $\tilde{M}$  are involved with  $y$  that is evaluated at  $(y_n + y_{n+1})/2$ . The scheme (22) is denoted by EAVF4. It is important to see that the new EAVF integrator (22) reduces to the fourth-order AVF integrator derived in [28, 40], when  $M \rightarrow 0$ .

**Theorem 3.1.** *If  $Q$  is skew-symmetric, then the EAVF integrator (22) is symmetric and of order four, and it can preserve the Hamiltonian  $H(y) = \frac{1}{2}y^\top Ay + U(y)$  corresponding to (1), i.e.,  $H(y_{n+1}) = H(y_n)$ .*


















$\tau$	$e(\tau)$	$a(\tau)$	$b(\tau)$	$\tau$	$e(\tau)$	$a(\tau)$	$b(\tau)$	$\tau$	$e(\tau)$	$a(\tau)$	$b(\tau)$
$\emptyset$	1	1	0		$\frac{1}{2}$	$\frac{1}{2}$	0		$\frac{1}{6}$	$\frac{1}{4}$	$-\frac{1}{12}$
$\bullet$	1	1	1		$\frac{1}{3}$	$\frac{1}{3}$	0		$\frac{1}{6}$	$\frac{1}{4}$	$-\frac{1}{12}$
$\circ$	1	1	1		$\frac{1}{3}$	$\frac{1}{3}$	0		$\frac{1}{6}$	$\frac{1}{4}$	$-\frac{1}{12}$
	$\frac{1}{2}$	$\frac{1}{2}$	0		$\frac{1}{3}$	$\frac{1}{3}$	0		$\frac{1}{6}$	$\frac{1}{4}$	$-\frac{1}{12}$
	$\frac{1}{2}$	$\frac{1}{2}$	0		$\frac{1}{6}$	$\frac{1}{4}$	$-\frac{1}{12}$		$\frac{1}{6}$	$\frac{1}{6}$	0
	$\frac{1}{2}$	$\frac{1}{2}$	0		$\frac{1}{6}$	$\frac{1}{4}$	$-\frac{1}{12}$		$\frac{1}{6}$	$\frac{1}{6}$	0

Table 2: Coefficients  $a(\tau)$ ,  $e(\tau)$  and  $b(\tau)$  for bi-colored trees of order  $\leq 3$ .

*Proof.* Exchanging  $y_{n+1} \leftrightarrow y_n$  and  $h \leftrightarrow -h$  in the EAVF integrator (22) yields

$$y_n = \exp(-h\tilde{M})y_{n+1} - h\varphi(-h\tilde{M})\tilde{Q} \int_0^1 \nabla U((1-\xi)y_{n+1} + \xi y_n) d\xi. \quad (23)$$

It follows from the equality  $\exp(h\tilde{M}) \cdot \varphi(-h\tilde{M}) = \varphi(h\tilde{M})$  that

$$\exp(h\tilde{M})y_n = y_{n+1} - h\varphi(h\tilde{M})\tilde{Q} \int_0^1 \nabla U((1-\xi)y_{n+1} + \xi y_n) d\xi, \quad (24)$$

which is exactly the same as (22) on noting that  $\int_0^1 \nabla U((1-\xi)y_{n+1} + \xi y_n) d\xi = \int_0^1 \nabla U((1-\xi)y_n + \xi y_{n+1}) d\xi$ . This shows the symmetry of the EAVF integrator (22).

It can be observed that  $-\frac{h^3}{12}(MF + FM + FF)(My + f(y))$  is the dominant part of the local truncation error for the second-order EAVF integrator (3), i.e.,

$$B_f(e, y) - B_f(a, y) = -\frac{h^3}{12}(MF + FM + FF)(My + f(y)) + O(h^5).$$

We can conclude from the derivation of (22) that the integrator (22) is at least of order three. It follows from the symmetry of (22) that this integrator is at least of order four. Furthermore, the corresponding AVF method to (22) once  $M \rightarrow \mathbf{0}$  is of order four. Thus, the EAVF integrator (22) is exactly of order four.

Following Theorem 2.3 in [29], to prove the energy-preserving property of (22), we just need to show the same skew-symmetry of  $\tilde{Q}$  as  $Q$ . Note that  $F = \frac{\partial f}{\partial y}$ ,  $f(y) = Q\nabla U(y)$ . Hence,  $F = Q\mathcal{H}$ , where  $\mathcal{H}$  is the Hessian of  $U(y)$  and thus symmetric. This gives

$$\tilde{Q} = \left(E_0 - \frac{h^2}{12}(MF + FM + FF)\right)Q = Q - \frac{h^2}{12}(QAQ\mathcal{H}Q + Q\mathcal{H}QAQ + Q\mathcal{H}Q\mathcal{H}Q),$$

which shows the skew-symmetry of  $\tilde{Q}$  combining with the symmetry of  $A$  and  $\mathcal{H}$ . This completes the proof.  $\square$

On the basis of the expression of the fourth-order truncated modified vector field  $\tilde{g}^{[4]}(y)$  in (21) and the corresponding fourth-order energy-preserving exponential integrator in (22), the following theorem uniformly states the formulation of the higher-order truncated modified vector field  $\tilde{g}^{[2k]}(y)$  ( $k \geq 3$ ) and the corresponding higher-order energy-preserving exponential integrator. Note that Theorem 3.2 is intuitively true from the aspect of both the generating function theory [18] and the modifying integrator theory [13], and its proof involving only technical approaches is thus not presented in the main text. Readers who are interested in such content can refer to Appendix B.

**Theorem 3.2.** *The modified vector field  $\tilde{g}(y)$  of the EAVF integrator (3) can be formulated as*

$$\tilde{g}(y) = (E_0 + h^2 E_1 + h^4 E_2 + \cdots + h^{2k} E_k + \cdots) Q (Ay + \nabla U(y)) = \bar{Q} (Ay + \nabla U(y)), \quad (25)$$

where  $E_0$  is the identity matrix,  $E_k$  ( $k \geq 1$ ) is determined by (20),  $\bar{Q} = \lim_{k \rightarrow \infty} \bar{Q}_k$ , and

$$\bar{Q}_k = (E_0 + h^2 E_1 + h^4 E_2 + \cdots + h^{2k} E_k) Q = \left( \sum_{j=0}^k h^{2j} E_j \right) Q, \quad (26)$$

is skew-symmetric for all  $k \geq 0$ . Then, the exponential integrator corresponding to the  $2k$ th-order truncated modified vector field

$$\tilde{g}^{[2k]}(y) = \bar{Q}_{k-1}(y) (Ay + \nabla U(y)), \quad k \geq 1,$$

reads

$$y_{n+1} = \exp(h \bar{M}_{k-1}) y_n + h \varphi(h \bar{M}_{k-1}) \bar{Q}_{k-1} \left( \frac{y_n + y_{n+1}}{2} \right) \int_0^1 \nabla U((1-\xi)y_n + \xi y_{n+1}) d\xi, \quad (27)$$

where  $\bar{M}_{k-1} = \bar{Q}_{k-1} \left( \frac{y_n + y_{n+1}}{2} \right) A$  and  $\bar{Q}_{k-1} \left( \frac{y_n + y_{n+1}}{2} \right)$  means evaluating  $\bar{Q}_{k-1}$  at the point  $\frac{y_n + y_{n+1}}{2}$ . The numerical scheme is symmetric and of order  $2k$  for the system (2), and consequently preserves the energy, i.e.,  $H(y_{n+1}) = H(y_n)$ .

**Remark 3.3.** *It seems that we only construct EAVF integrators of even order by (27) in Theorem 3.2. In fact, if  $\bar{Q}_{k-1}$  and  $\bar{M}_{k-1}$  are evaluated at the point  $y_n$ , then the symmetry of the EAVF integrator (27) will be lost. This results in the fact that the order conditions of (27) are satisfied up to bi-colored trees of only order  $2k-1$ , which implies that (27) is of order  $2k-1$ . A simple discussion on this issue can also be referred to [40]. That is, EAVF integrators of arbitrary (odd or even) order can be constructed according to Theorem 3.2. However, taking into account the nearly same formulation of  $2k$ th-order and  $(2k-1)$ th-order integrators, we naturally prefer to the  $2k$ th-order integrator due to its higher accuracy. This is the main reason that we consider only the EAVF integrators of even order in this paper.*

Here, we remark that although in this section, we only derive the fourth-order EAVF integrator (22) from the adapted substitution law in Theorem 2.3 based on the second-order EAVF integrator (3), EAVF integrators of higher order can be successively obtained in a similar way by considering higher-order bi-colored trees in Table 1. The construction of the sixth-order AVF integrator in [28] can be referred as an analogy under a simpler case of rooted trees.

Before the end of this section, we remark that the adapted substitution law in Theorem 2.3 along with the modified definition for skeleton in Definition 2.2 is applicable to not only the EAVF integrators considered in this paper, but also other exponential integrators for (1), such as the exponential Runge–Kutta methods and the general linear methods [20]. As stated in [19, 21], besides the construction of high-order methods based on modified differential equations, the backward error analysis of these exponential integrators also needs the use of the substitution law. Therefore, the adapted substitution law proposed in this paper possesses broad applications in numerical analysis for exponential integrators.

#### 4. Stability and convergence

In this section, we are concerned with the stability and convergence of the EAVF integrator (27). Throughout this section, we use the notation  $\mathcal{A} \lesssim \mathcal{B}$  to represent that there exists a constant  $C > 0$  independent of  $h$  such that  $\mathcal{A} \leq \mathcal{B}(1 + Ch)$  for small  $h$ . This means that the difference between  $\mathcal{A}$  and  $\mathcal{B}$  will be very close to zero for sufficiently small  $h$ . The following assumption will be admitted throughout the analysis in this section.

**Assumption 4.1.** *It is assumed that the solution  $y(t)$  of (2) is located in the closed ball*

$$B(y_0, R) = \{y \in \mathbb{R}^d : \|y - y_0\| \leq R\},$$

for  $t \in [t_0, T_{\text{end}}]$ , and  $\|y\| \leq B$  holds for all  $y \in B(y_0, R)$ . In addition, the function  $\nabla U$  is bounded by a constant  $\lambda$  and satisfies the Lipchitz condition, i.e.,

$$\|\nabla U(y)\| \leq \lambda, \quad \|\nabla U(y) - \nabla U(z)\| \leq L\|y - z\|,$$

for all  $y, z \in B(y_0, R)$ . We finally assume that  $\|\varphi(hM^*)\| \leq C$  holds for all occurring matrix  $M^*$ .

Considering the similar expressions of the second-order EAVF integrator (3) and high-order EAVF integrators (27), we are hopeful of obtaining the convergence of the fixed-point iteration for (27) as well as for (3) (see [29]). We will show this result based on the following lemma.

**Lemma 4.2.** Let  $\bar{E}_k = E_0 + h^2 E_1 + h^4 E_4 + \cdots + h^{2k} E_k = \sum_{j=0}^k h^{2j} E_j$ . Suppose that there exists a constant  $C_k \geq 0$  such that the inequality

$$\|\bar{E}_k(y) - \bar{E}_k(z)\| \leq h^2 C_k \|y - z\|, \quad (28)$$

holds for all  $y, z \in B(y_0, R)$ . We then have the following estimations

$$\|\bar{Q}_k(y) - \bar{Q}_k(z)\| \leq h^2 C_k \|y - z\|, \quad (29)$$

$$\|\bar{M}_k(y) - \bar{M}_k(z)\| \leq h^2 C_k \|A\| \cdot \|y - z\|, \quad (30)$$

$$\|\exp(h\bar{M}_k(y)) - \exp(h\bar{M}_k(z))\| \lesssim h^3 C_k \|A\| \cdot \|y - z\|, \quad (31)$$

$$\|\exp(h\bar{M}_k(y))\| \lesssim 1 + h\|A\|, \quad (32)$$

$$\|\varphi(h\bar{M}_k(y)) - \varphi(h\bar{M}_k(z))\| \lesssim \frac{1}{2} h^3 C_k \|A\| \cdot \|y - z\|. \quad (33)$$

*Proof.* On noting the fact that  $\bar{Q}_k = \bar{E}_k Q$ ,  $\bar{M}_k = \bar{E}_k Q A$  and  $\|Q\| = 1$ , the inequalities (29) and (30) directly follow from (28). With the notation  $\lesssim$ , we naturally have that  $\|\bar{E}_k(y)\| \lesssim 1$ ,  $\|\bar{Q}_k(y)\| \lesssim 1$  and  $\|\bar{M}_k(y)\| \lesssim \|A\|$ . Then, the inequalities (31) and (32) directly follow from the definition (18) of the matrix exponential.

Since

$$\varphi(h\bar{M}_k(y)) - \varphi(h\bar{M}_k(z)) = \int_0^1 \exp((1-\theta)h\bar{M}_k(y)) - \exp((1-\theta)h\bar{M}_k(z)) d\theta,$$

the inequality (33) immediately follows from (31). This completes the proof.  $\square$

**Remark 4.3.** It is noted that, the estimations in Lemma 4.2 hold for general cases of  $M$ . However, for special matrix  $M$ , some estimations can be improved more exactly. For instance, if  $M$  is skew symmetric or symmetric negative semidefinite, it can be derived that  $\|\exp(hM)\| \leq 1$  and  $\|\varphi(hM)\| \leq 1$ . In addition, it follows from (29) and the formulation of  $\bar{E}_k$  that  $C_0 = 0$  and  $C_k = O(\|A\|)$  for large  $\|A\|$  on noticing that  $E_1 = -\frac{1}{12}(QAQH Q + QHQ A Q + QHQHQ)$ . Here, we remark that an estimation to  $\|\varphi(h\bar{M}_k(y))\|$  similar to (32) is not presented but the upper bound  $C$  is given in Assumption 4.1 just because  $\|\varphi(h\bar{M}_k(y))\| \leq C$  is used in [29]. Therefore, it is more convenient to make a comparison under this setting. Finally, it is worth mentioning that the result in Theorem 4.4 coincides with [29] once  $k = 0$ .

**Theorem 4.4.** (Convergence of the fixed-point iteration) Under Assumption 4.1, if

$$0 < h \leq \hat{h}_1 \lesssim \frac{1}{\Gamma}, \quad (34)$$

where  $\Gamma = \max\{\|A\|, C_k, \frac{1}{2}CL + \frac{1}{2}B + \frac{1}{4}\lambda + \frac{1}{2}\lambda C\}$ , then the fixed-point iteration determined by the EAVF integrator (27) is convergent.

*Proof.* Denote  $\bar{E}_k^{[1]} = \bar{E}_k(\frac{y+z_1}{2})$  and  $\bar{E}_k^{[2]} = \bar{E}_k(\frac{y+z_2}{2})$ . It then follows from Lemma 4.2 that

$$\begin{aligned} \|\bar{E}_k^{[1]} - \bar{E}_k^{[2]}\| &\leq \frac{1}{2} h^2 C_k \|z_1 - z_2\|, \quad \|\bar{Q}_k^{[1]} - \bar{Q}_k^{[2]}\| \leq \frac{1}{2} h^2 C_k \|z_1 - z_2\|, \quad \|\bar{M}_k^{[1]} - \bar{M}_k^{[2]}\| \leq \frac{1}{2} h^2 C_k \|A\| \cdot \|z_1 - z_2\|, \\ \|\exp(h\bar{M}_k^{[1]}) - \exp(h\bar{M}_k^{[2]})\| &\lesssim \frac{1}{2} h^3 C_k \|A\| \cdot \|z_1 - z_2\|, \quad \|\varphi(h\bar{M}_k^{[1]}) - \varphi(h\bar{M}_k^{[2]})\| \lesssim \frac{1}{4} h^3 C_k \|A\| \cdot \|z_1 - z_2\|. \end{aligned}$$

Keeping it in mind that  $\|\bar{Q}_k\| \lesssim 1$ , we further obtain

$$\|\varphi(h\bar{M}_k^{[1]})\bar{Q}_k^{[1]} - \varphi(h\bar{M}_k^{[2]})\bar{Q}_k^{[2]}\| \lesssim \frac{1}{4} h^3 C_k \|A\| \cdot \|z_1 - z_2\| + \frac{1}{2} h^2 C_k \|z_1 - z_2\|.$$

Let  $\Psi(z) = \exp(h\bar{M}_k)y + h\varphi(h\bar{M}_k)\bar{Q}_k \int_0^1 \nabla U((1-\xi)y + \xi z)d\xi$ . Using Assumption 4.1 and the above inequalities yields

$$\begin{aligned} \|\Psi(z_1) - \Psi(z_2)\| &\leq \|\exp(h\bar{M}_k^{[1]}) - \exp(h\bar{M}_k^{[2]})\| \cdot \|y\| + \left\| h[\varphi(h\bar{M}_k^{[1]})\bar{Q}_k^{[1]} - \varphi(h\bar{M}_k^{[2]})\bar{Q}_k^{[2]}] \int_0^1 \nabla U((1-\xi)y + \xi z_2)d\xi \right\| \\ &\quad + \left\| h\varphi(h\bar{M}_k^{[1]})\bar{Q}_k^{[1]} \cdot \int_0^1 [\nabla U((1-\xi)y + \xi z_1) - \nabla U((1-\xi)y + \xi z_2)]d\xi \right\| \\ &\lesssim \left( \frac{1}{2}h^3 BC_k \|A\| + \frac{1}{4}h^4 \lambda C_k \|A\| + \frac{1}{2}h^3 \lambda CC_k + \frac{1}{2}hCL \right) \|z_1 - z_2\|. \end{aligned}$$

Let  $\rho = \frac{1}{2}hCL + \frac{1}{2}h^3 BC_k \|A\| + \frac{1}{4}h^4 \lambda C_k \|A\| + \frac{1}{2}h^3 \lambda CC_k$ . Since  $\|A\| \gg 1$  and  $h < 1$  due to the condition (34), we further confirm that  $h^2 C_k \leq 1$ ,  $h^2 C_k \|A\| \leq 1$  and  $h^3 C_k \|A\| \leq 1$ . This shows that  $\rho \leq \frac{1}{2}hCL + \frac{1}{2}hB + \frac{1}{4}h\lambda + \frac{1}{2}h\lambda C$ . Under the condition (34), the fixed-point iteration converges by the contraction mapping theorem and the fact that  $\rho < 1$ .  $\square$

**Theorem 4.5.** (Convergence of the numerical solution) Under Assumption 4.1, if the stepsize  $h$  is sufficiently small such that  $h \leq \frac{1}{2\Gamma}$ , then the numerical solutions  $y_n$  obtained by the EAVF integrator (27) satisfy the following estimation

$$\|y(t_n) - y_n\| \lesssim \mathcal{M}h^{2k+2}, \quad (35)$$

where the constant  $\mathcal{M}$  is independent of  $h$ .

*Proof.* We first denote  $\bar{E}_k^{[3]} = \bar{E}_k(\frac{y(t_n)+y(t_{n+1})}{2})$ ,  $\bar{E}_k^{[4]} = \bar{E}_k(\frac{y_n+y_{n+1}}{2})$ , and the global error  $e_n = y(t_n) - y_n$ . It then follows from Lemma 4.2 that

$$\begin{aligned} \|\bar{E}_k^{[3]} - \bar{E}_k^{[4]}\| &\leq \frac{1}{2}h^2 C_k (\|e_n\| + \|e_{n+1}\|), \quad \|\bar{Q}_k^{[3]} - \bar{Q}_k^{[4]}\| \leq \frac{1}{2}h^2 C_k (\|e_n\| + \|e_{n+1}\|), \\ \|\bar{M}_k^{[3]} - \bar{M}_k^{[4]}\| &\leq \frac{1}{2}h^2 C_k \|A\| (\|e_n\| + \|e_{n+1}\|), \quad \|\exp(h\bar{M}_k^{[3]}) - \exp(h\bar{M}_k^{[4]})\| \lesssim \frac{1}{2}h^3 C_k \|A\| (\|e_n\| + \|e_{n+1}\|), \\ \|\varphi(h\bar{M}_k^{[3]})\bar{Q}_k^{[3]} - \varphi(h\bar{M}_k^{[4]})\bar{Q}_k^{[4]}\| &\lesssim \left( \frac{1}{4}h^3 C_k \|A\| + \frac{1}{2}h^2 CC_k \right) (\|e_n\| + \|e_{n+1}\|). \end{aligned}$$

Denote the local truncation error  $\delta_{n+1}$  by

$$\delta_{n+1} = y(t_{n+1}) - \left[ \exp(h\bar{M}_k^{[3]})y(t_n) + h\varphi(h\bar{M}_k^{[3]})\bar{Q}_k^{[3]} \int_0^1 \nabla U((1-\xi)y(t_n) + \xi y(t_{n+1}))d\xi \right],$$

that is,

$$y(t_{n+1}) = \exp(h\bar{M}_k^{[3]})y(t_n) + h\varphi(h\bar{M}_k^{[3]})\bar{Q}_k^{[3]} \int_0^1 \nabla U((1-\xi)y(t_n) + \xi y(t_{n+1}))d\xi + \delta_{n+1}. \quad (36)$$

It follows from the construction of the EAVF integrator (27) that the order conditions of (27) are satisfied for bi-colored trees up to order  $2k+2$  according to the modifying integrator theory [13]. That is

$$\delta_{n+1} \leq \Lambda h^{2k+3}, \quad (37)$$

where  $\Lambda$  is a constant independent of  $h$ , but may depend on  $\|A\|$ . Moreover, numerical solutions satisfy

$$y_{n+1} = \exp(h\bar{M}_k^{[4]})y_n + h\varphi(h\bar{M}_k^{[4]})\bar{Q}_k^{[4]} \int_0^1 \nabla U((1-\xi)y_n + \xi y_{n+1})d\xi. \quad (38)$$

Then, subtracting (38) from (36) yields

$$\begin{aligned} e_{n+1} &= (\exp(h\bar{M}_k^{[3]}) - \exp(h\bar{M}_k^{[4]}))y(t_n) + \exp(h\bar{M}_k^{[4]})(y(t_n) - y_n) \\ &\quad + h[\varphi(h\bar{M}_k^{[3]})\bar{Q}_k^{[3]} - \varphi(h\bar{M}_k^{[4]})\bar{Q}_k^{[4]}] \int_0^1 \nabla U((1-\xi)y(t_n) + \xi y(t_{n+1}))d\xi \\ &\quad + h\varphi(h\bar{M}_k^{[4]})\bar{Q}_k^{[4]} \int_0^1 [\nabla U((1-\xi)y(t_n) + \xi y(t_{n+1})) - \nabla U((1-\xi)y_n + \xi y_{n+1})]d\xi + \delta_{n+1}. \end{aligned} \quad (39)$$

Under Assumption 4.1 and the above estimations, the equation (39) gives

$$\begin{aligned} \|e_{n+1}\| &\lesssim (1 + h\|A\|)\|e_n\| + \frac{1}{2}h^3 BC_k\|A\|(\|e_n\| + \|e_{n+1}\|) + \left(\frac{1}{4}h^4 \lambda C_k\|A\| + \frac{1}{2}h^3 \lambda CC_k\right)(\|e_n\| + \|e_{n+1}\|) \\ &\quad + \frac{1}{2}hCL(\|e_n\| + \|e_{n+1}\|) + \Lambda h^{2k+3}, \end{aligned}$$

which further yields

$$\begin{aligned} &\left(1 - \frac{1}{2}hCL - \frac{1}{2}h^3 BC_k\|A\| - \frac{1}{4}h^4 \lambda C_k\|A\| - \frac{1}{2}h^3 \lambda CC_k\right)\|e_{n+1}\| \\ &\lesssim (1 + h\|A\| + \frac{1}{2}hCL + \frac{1}{2}h^3 BC_k\|A\| + \frac{1}{4}h^4 \lambda C_k\|A\| + \frac{1}{2}h^3 \lambda CC_k)\|e_n\| + \Lambda h^{2k+3}. \end{aligned} \quad (40)$$

We denote  $\Theta = \frac{1}{2}CL + \frac{1}{2}h^2 BC_k\|A\| + \frac{1}{4}h^3 \lambda C_k\|A\| + \frac{1}{2}h^2 \lambda CC_k$ . For sufficiently small  $h$ , it follows from the inequality (40) that

$$\|e_{n+1}\| \lesssim \left(1 + \frac{\|A\| + 2\Theta}{1 - h\Theta} \cdot h\right)\|e_n\| + \frac{\Lambda h^{2k+3}}{1 - h\Theta}.$$

With the setting  $e_0 = 0$ , using the discrete Gronwall inequality gives

$$\|e_n\| \lesssim \exp\left(\frac{\|A\| + 2\Theta}{1 - h\Theta} (T_{end} - t_0)\right) \frac{\Lambda h^{2k+2}}{\|A\| + 2\Theta}. \quad (41)$$

Similarly to Theorem 4.4, if the stepsize  $h$  satisfies  $h \lesssim \frac{1}{2\Gamma}$ , we have

$$\|e_{n+1}\| \lesssim \mathcal{M} h^{2k+2},$$

where  $\mathcal{M} = \exp(2(\|A\| + 2\mathcal{G})(T_{end} - t_0))\Lambda/\|A\|$  and  $\mathcal{G} = \frac{1}{2}B + \frac{1}{2}CL + \frac{1}{4}\lambda + \frac{1}{2}\lambda C$ . The constant  $\mathcal{M}$  is evidently independent of  $h$ . This completes the proof.  $\square$

**Theorem 4.6.** (Stability) Under Assumption 4.1, if the stepsize  $h$  is sufficiently small such that  $h \leq \frac{1}{2\Gamma}$ , then the numerical solutions  $\tilde{y}_n$  and  $y_n$  obtained by the EAVF integrator (27) respectively with the initial conditions  $\tilde{y}_0$  and  $y_0$  satisfy

$$\|\tilde{y}_n - y_n\| \lesssim \Pi \|\tilde{y}_0 - y_0\|, \quad (42)$$

where  $\Pi = \exp(2(\|A\| + 2\mathcal{G})(T_{end} - t_0))$  and  $\mathcal{G} = \frac{1}{2}B + \frac{1}{2}CL + \frac{1}{4}\lambda + \frac{1}{2}\lambda C$  are independent of  $h$ .

*Proof.* The proof is similar to that of Theorem 4.5 by only dropping off the local truncation error term  $\Lambda h^{2k+3}$ . Hence the details are omitted here.  $\square$

**Remark 4.7.** We note that once the upper bound of  $h$  is decreased, the constants  $\mathcal{M}$  in (35) and  $\Pi$  in (42) can be decreased as well, though they are bounded by  $\mathcal{M} \geq \frac{\exp((\|A\| + CL)(T_{end} - t_0))\Lambda}{\|A\| + CL}$  and  $\Pi \geq \exp(\|A\| + CL)(T_{end} - t_0)$ .

## 5. Numerical experiments

In this section, we carry out numerical experiments with different problems. The numerical results will show the high accuracy and good energy preservation of the proposed fourth-order EAVF integrator (22) in comparison with some effective energy-preserving methods appeared in the literature. In our experiments, numerical methods are selected as follows to make a comparison:

- AVF2: the second-order average vector field method [15, 35];
- AVF4: the fourth-order average vector field method [28];
- EAVF2: the second-order exponential average vector field method [29];

- EAVF4: the fourth-order exponential average vector field method (22) proposed in this paper;
- CRK4: the fourth-order continuous-stage Runge–Kutta method [20].

All these methods are energy preserving, and EAVF2 and EAVF4 will reduce to AVF2 and AVF4 once  $M \rightarrow 0$ .

It is clear that all these energy-preserving integrators are implicit. Hence, iterative solutions are required in the implementation of these methods. As analyzed in Section 4, we use the fixed-point iteration for all these methods. The tolerance error for the iteration is set as  $10^{-14}$ , and this means that the iteration will be stopped once the error between two successive approximations is smaller than  $10^{-14}$ . Throughout the experiment, we use the numerical solution computed by high-order methods with a sufficiently small stepsize as the reference solution. The global errors of energy (GHE) are evaluated by

$$GHE = |H(y(t_n)) - H(y_n)|.$$

In our numerical experiments, the efficiency curves mean the log-log plots of the errors. In addition, a quadrature formula is needed to easily calculate the integral  $\int_0^1 \nabla U((1-\tau)y_{n+1} + \tau y_n) d\tau$  for all the five energy-preserving integrators. To preserve the symmetry of these methods and provide high accuracy, we select the 5-point Gauss-Legendre quadrature formula  $(b_i, c_i)_{i=1}^5$ . For large-scale problems, we point out that the matrix-valued functions  $\varphi(hM)$  and  $\exp(hM)$  may be evaluated by the Krylov subspace method [22] because of its fast convergence.

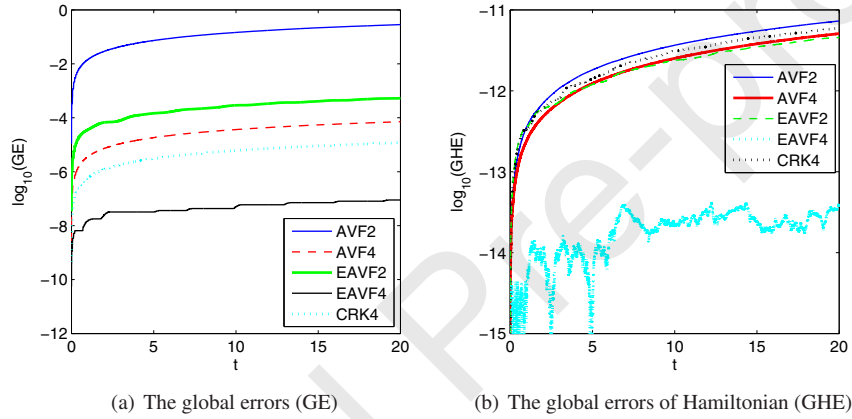


Figure 1: Results for Problem 1 with the stepsize  $h = 1/10^3$  and  $\omega = 50$ .

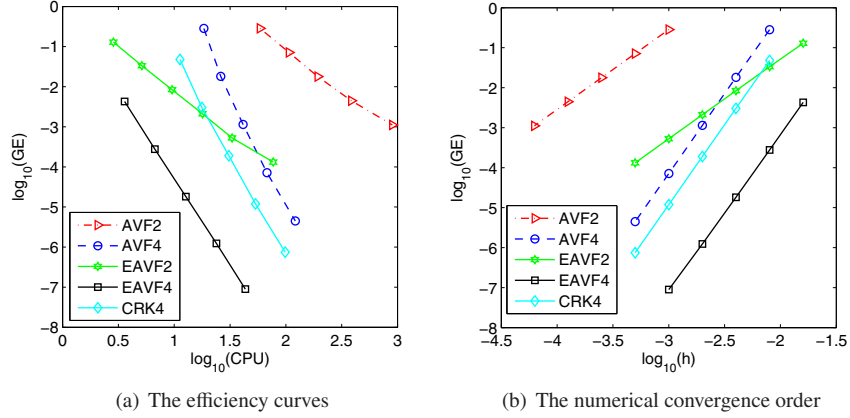
	AVF2	AVF4	EAVF2	EAVF4	CRK4
Convergence order	1.9993	3.9894	1.9890	3.8909	3.9937
Average iteration number	9	9	3	3	10

Table 3: Results for Problem 1: the average iteration number of the fixed-point iteration.

**Problem 1.** We first consider the notable Fermi-Pasta-Ulam problem (see, e.g. [21]), whose Hamiltonian is given by

$$H(y, x) = \frac{1}{2} \sum_{i=1}^{2m} y_i^2 + \frac{\omega^2}{2} \sum_{i=1}^m x_{m+i}^2 + \frac{1}{4} \left( (x_1 - x_{m+1})^4 + \sum_{i=1}^{m-1} (x_{i+1} - x_{m+i+1} - x_i - x_{m+i})^4 + (x_m + x_{2m})^4 \right), \quad (43)$$

where  $x_i$  represents a scaled displacement of the  $i$ th stiff spring,  $x_{m+i}$  is a scaled expansion (or compression) of the  $i$ th stiff spring, and  $y_i, y_{m+i}$  are their velocities (or momenta). By setting  $z = (y^\top, x^\top)^\top$ ,  $U(x) = \frac{1}{4}((x_1 - x_{m+1})^4 +$

Figure 2: Results for Problem 1 with  $\omega = 50$ .

$\sum_{i=1}^{m-1} (x_{i+1} - x_{m+i+1} - x_i - x_{m+i})^4 + (x_m + x_{2m})^4$ ,  $f = \left( -(\nabla_x U(x))^T, 0^T \right)^T$ ,  $A_s = \begin{pmatrix} 0_{m \times m} & 0_{m \times m} \\ 0_{m \times m} & \omega^2 I_{m \times m} \end{pmatrix}$ , and  $M = \begin{pmatrix} 0 & -A_s \\ I & 0 \end{pmatrix}$ , we can express its differential equation in the form of (2) as follows

$$z'(t) = Mz + f(z).$$

In this experiment, we choose

$$x_1(0) = 1, y_1(0) = 1, x_4(0) = \frac{1}{\omega}, y_1(0) = 1,$$

and zero for the remaining initial values. Other parameters are fixed as  $m = 3$ ,  $\omega = 50$  and the integration time  $T_{end} = 20$ . The global errors (GE) of the solution, and the global errors of Hamiltonian for each numerical scheme with the stepsize  $h = 1/10^3$  are shown in Fig. 1. It can be observed from Fig. 1 that the new scheme EAVF4 given by (22) in this paper shows good numerical behaviour in both global errors and energy preservation. Fig. 2 (a) exhibits the efficiency curves, which indicates that the energy-preserving exponential integrators have higher efficiency than other existing energy-preserving integrators in the literature, and EAVF4 possesses the best efficiency among the five integrators. Fig. 2 (b) shows the numerical convergence order for each method by varying the stepsize  $h$ . We further list the convergence order by fitting the slope of the data in Fig. 2(b) and the average iteration number of the fixed-point iteration in each step with  $h = 1/10^3$  in Table 3. The results listed in Table 3 are highly in accordance with our theoretical analyses that EAVF4 is of order four, and that the exponential integrators converge much faster than the classical integrators in dealing with highly oscillatory problems.

**Problem 2.** Consider the averaged system in wind-induced oscillation (see, e.g. [29, 34])

$$\begin{aligned} \dot{x}_1 &= -\zeta x_1 - \lambda x_2 + x_1 x_2, \\ \dot{x}_2 &= \lambda x_1 - \zeta x_2 + \frac{1}{2}(x_1^2 - x_2^2), \end{aligned} \tag{44}$$

where  $\zeta \geq 0$  is a damping factor and  $\lambda$  is a detuning parameter. By denoting  $\zeta = \rho \cos(\theta)$ ,  $\lambda = \rho \sin(\theta)$ ,  $\rho \geq 0$  and  $0 \leq \theta \leq \pi/2$ , this system can be written in the linear-gradient form  $\dot{x} = Q\nabla H$  with

$$Q = \begin{pmatrix} -\cos(\theta) & -\sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix},$$

and



L. Mei, L. Huang and X. Wu

$$H(x_1, x_2) = \frac{1}{2}\rho(x_1^2 + x_2^2) - \frac{1}{2}\sin(\theta)\left(x_1x_2^2 - \frac{x_1^3}{3}\right) + \frac{1}{2}\sin(\theta)\left(\frac{x_2^3}{3} - x_1^2x_2\right).$$

We can also write the equation in the form of (2) as follows:

$$\frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = M \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + Q \nabla U,$$

where  $M = \rho Q$  and  $U(x_1, x_2) = -\frac{1}{2}\sin(\theta)\left(x_1x_2^2 - \frac{x_1^3}{3}\right) + \frac{1}{2}\sin(\theta)\left(\frac{x_2^3}{3} - x_1^2x_2\right)$ . It is noted that  $H(x_1, x_2)$  is either the Lyapunov function in the dissipative case provided  $\cos(\theta) > 0$ , or the first integral in the conservative case provided  $\cos(\theta) = 0$ .

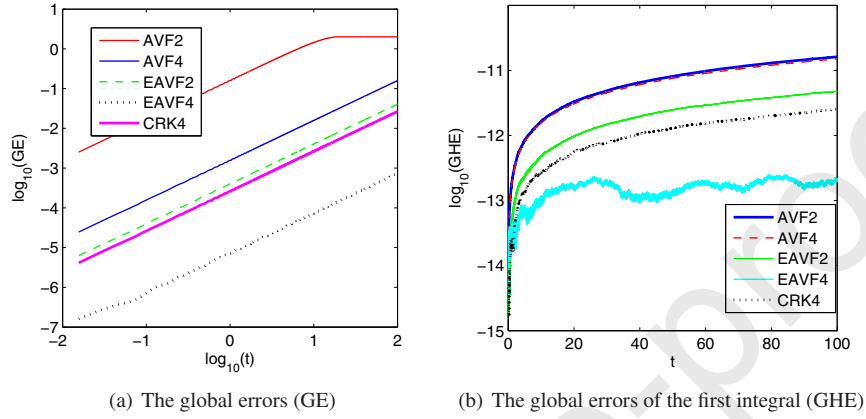


Figure 3: Results for Problem 2: the conservative case of  $\theta = \pi/2$  with the stepsize  $h = 1/2^6$  and  $\rho = 20$ .

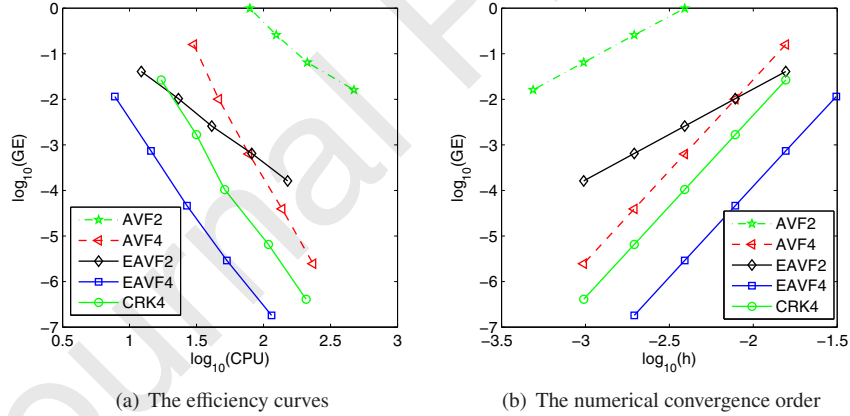


Figure 4: Results for Problem 2: the conservative case of  $\theta = \pi/2$  with  $\rho = 20$ .

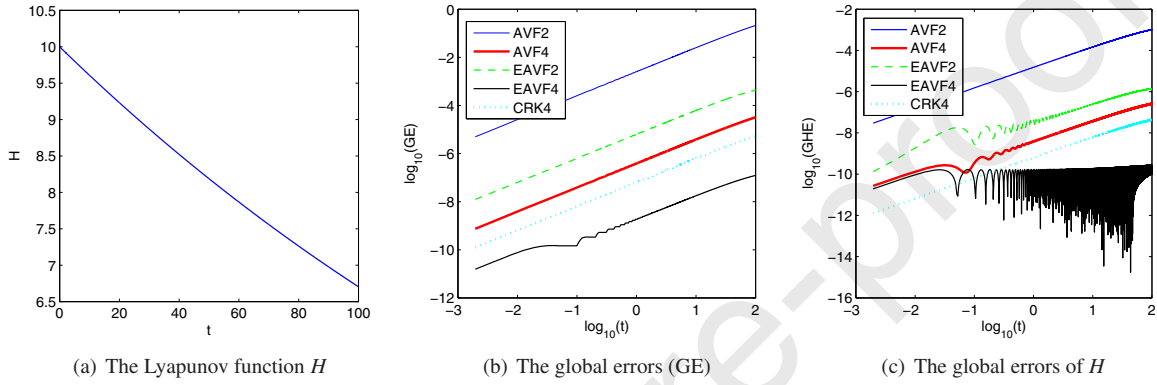
In this experiment, we set  $x_1(0) = 0$ ,  $x_2(0) = 1$  and the integration time  $T_{end} = 100$ . Fig. 3 presents both the global errors of the solution, and the global errors of the first integral in the conservative case of  $\theta = \pi/2$  with the stepsize  $h = 1/2^6$ . This demonstrates the better performance of the EAVF4 integrator. The efficiency curves and the numerical convergence order of these five methods are shown in Fig. 4, in which EAVF4 again shows the best numerical behaviour. The detailed results are further listed in Table 4, and these demonstrate the comparable convergence of the fixed-point iteration of EAVF4 and EAVF2. It can be observed from Table 4 that the convergence rate of both EAVF4 and EAVF2 is much faster than that of other integrators. Furthermore, the numerical convergence



	AVF2	AVF4	EAVF2	EAVF4	CRK4
Convergence order	1.9798	3.99948	1.9942	3.9885	3.9981
Average iteration number	19	19	7	7	17

Table 4: Results for Problem 2: The average iteration number of the fixed-point iteration.

orders corresponding to Fig. 4 (b) in this table also coincide with their theoretical orders. Moreover, we also conduct the numerical experiment for the dissipative case in Fig. 5, where  $\theta = \pi/2 - 10^{-4}$  and  $h = 1/2^9$ . Fig. 5 (a) displays the evolution of the Lyapunov function  $H$  against time  $t$ , where  $H$  strictly decreases along with time. It can be observed from Fig. 5 that besides the higher accuracy of EAVF4, it also preserves the dissipation of the Lyapunov function best among all the five energy-preserving integrators.

Figure 5: Results for Problem 2: the dissipative case of  $\theta = \pi/2 - 10^{-4}$  with the stepsize  $h = 1/2^9$  and  $\rho = 20$ .

**Problem 3.** We finally consider the sine-Gordon equation (see, e.g. Chap. 4 in [49]) with the periodic boundary conditions

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} - \sin u, & -5 \leq x \leq 5, t \geq 0, \\ u(-5, t) = u(5, t). \end{cases} \quad (45)$$

A semi-discretisation on the spatial variable with the second-order symmetric difference gives the following differential equations in time

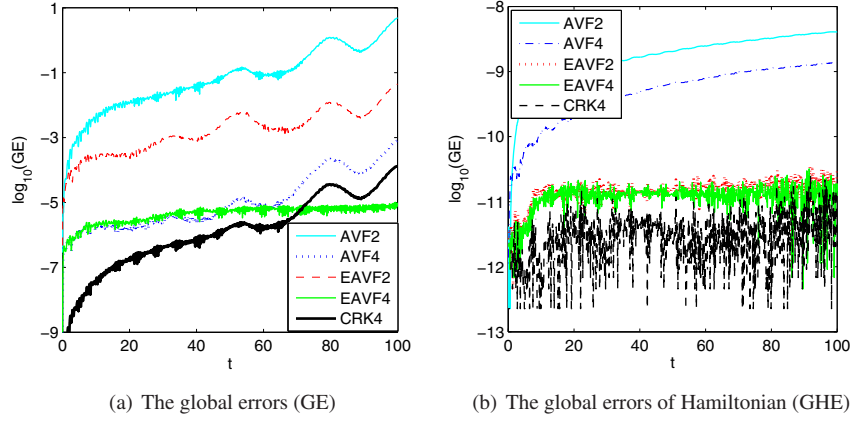
$$\frac{d}{dt} \begin{pmatrix} U' \\ U \end{pmatrix} = \begin{pmatrix} \mathbf{0} & M \\ I & \mathbf{0} \end{pmatrix} \begin{pmatrix} U' \\ U \end{pmatrix} + \begin{pmatrix} F(U) \\ 0 \end{pmatrix}, \quad (46)$$

where  $U(t) = (u_1(t), \dots, u_N(t))^T$  with  $u_i(t) \approx u(x_i, t)$ ,  $x_i = -5 + i\Delta x$  for  $i = 1, \dots, N$ ,  $\Delta x = 10/N$ , and

$$M = \frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & & -1 \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 & -1 \\ -1 & & & -1 & 2 \end{pmatrix}, \quad F(U) = -\sin(U) = -(\sin u_1, \dots, \sin u_N)^T.$$

The corresponding Hamiltonian is given by

$$H(U', U) = \frac{1}{2} U'^T U' + \frac{1}{2} U^T M U - (\cos u_1 + \dots + \cos u_N).$$

Figure 6: Results for Problem 3 with the stepsize  $h = 1/200$ .

For this problem, we take the initial conditions as

$$U(0) = (\pi)_{i=1}^N, \quad U'(0) = \sqrt{N}(0.01 + \sin(\frac{2\pi i}{N}))_{i=1}^N,$$

with  $N = 64$ , the integration time  $T_{end} = 100$  and the stepsize  $h = 1/200$ . We plot the global errors of the solution, and the global error of Hamiltonian in Fig. 6. It can be observed from Fig. 6 (a) that EAVF4 finally provides the best accuracy, even though it does not perform the best at the beginning. Fig. 6 (b) shows that EAVF4 preserves the Hamiltonian as well as EAVF2 and CRK4, and it is better than AVF2 and AVF4, for the sine-Gordon equation.

It is important to note that the numerical simulations in this paper include the important equations: the well-known multi-scale and highly oscillatory FPU problem, the oscillatory conservative or dissipative systems, and a nonlinear hyperbolic equation, the sine-Gordon equation, which was attracted a lot of attention in the 1970s due to the presence of soliton solutions. The numerical results presented above for highly oscillatory conservative or dissipative systems demonstrate the remarkable superiority of our new high-order energy-preserving exponential integrators in comparison with other energy-preserving schemes in the scientific literature.

## 6. Concluding remarks and discussion

As is known, differential equations with highly oscillatory solutions cannot be solved efficiently using conventional methods. Geometric integrators for oscillatory differential equations have become increasingly important in recent decades (see, e.g. [18, 21, 45, 47, 48]). This paper made an effort to present a general approach to analysing and constructing energy-preserving exponential integrators of a higher order, motivated by the second-order EAVF integrator for conservative or dissipative oscillatory problems. To this end, we first showed that the EAVF integrator is a B-series method and adapted the substitution law to exponential integrators. Then, with the adapted substitution law it was proved that there exist arbitrarily high-order energy-preserving exponential integrators which can be uniformly formulated. In particular, the construction of the fourth-order EAVF integrator were showed in detail as an example. Furthermore, we proved the stability and convergence of the integrators proposed in this paper. Finally, the numerical experiments were carried out and the numerical results soundly confirmed our theoretical analysis that the proposed fourth-order EAVF integrator can preserve the Hamiltonian energy well and has higher efficiency than the existing energy-preserving non-exponential integrators of the same order in the literature.

Again, it should be emphasized that higher-order EAVF integrators can be derived in the light of the approach proposed and developed in this paper.

## Acknowledgement

The authors sincerely thank the anonymous reviewers for their valuable suggestions, which helped improve this manuscript. This research is supported in part by the National Natural Science Foundation of China under Grants 11801377, 11903022 and 11671200 and 11861053, the Natural Science Foundation of Jiangxi Province under Grants 20192BCBL23030 and 20192ACBL21053, and the Natural Science Foundation of Jiangsu Province under Grant BK20150934.

## References

### References

- [1] H. Berland, B. Owren, B. Skaflestad, B-series and order conditions for exponential integrators, *SIAM J. Numer. Anal.* 43 (2005) 1715-1727.
- [2] H. Berland, B. Owren, B. Skaflestad, B-series and order conditions for exponential integrators, Technical report 5/04, The Norwegian University of Science and Technology, Trondheim, Norway, 2004, <http://www.math.ntnu.no/preprint/numerics/2004/N5-2004.ps>.
- [3] A. Bhatt, B.E. Moore, Structure-preserving exponential Runge-Kutta methods, *SIAM J. Sci. Comput.* 39 (2017) A593-A612.
- [4] L. Brugnano, F. Iavernaro, D. Trigiante, Hamiltonian boundary value methods (energy preserving discrete line integral methods), *JNAIAM. J. Numer. Anal. Ind. Appl. Math.* 5 (2010) 17-37.
- [5] L. Brugnano, J.I. Montijano, L. Rández, On the effectiveness of spectral methods for the numerical solution of multi-frequency highly oscillatory Hamiltonian problems, *Numer. Algor.* 81 (2019) 345-376.
- [6] N. Del Buono, C. Mastroserio, Explicit methods based on a class of four stage fourth order Runge-Kutta methods for preserving quadratic laws, *J. Comput. Appl. Math.* 140 (2002) 231-243.
- [7] M. Calvo, D. Hernández-Abreu, J.I. Montijano, L. Rández, On the preservation of invariants by explicit Runge-Kutta methods, *SIAM J. Sci. Comput.* 28 (2002) 868-885.
- [8] M. Calvo, M.P. Labarta, J.I. Montijano, L. Rández, Runge-Kutta projection methods with low dispersion and dissipation errors, *Adv. Comput. Math.* 41 (2015) 231-251.
- [9] E. Celledoni, V. Grimm, R.I. McLachlan, D.I. McLaren, D. O’Neale, B. Owren, G.R.W. Quispel, Preserving energy resp. dissipation in numerical PDEs using the “Average Vector Field” method, *J. Comput. Phys.* 231 (2012) 6770-6789.
- [10] E. Celledoni, R.I. McLachlan, B. Owren, G.R.W. Quispel, Energy-preserving integrators and the structure of B-series, *Found. Comput. Math.* 10 (2010) 673-693.
- [11] K. Cheng, W. Feng, S. Gottlieb, C. Wang, A Fourier pseudospectral method for the “good” Boussinesq equation with second-order temporal accuracy, *Numer. Meth. Partial D. E.* 31 (2015) 202-224.
- [12] P. Chartier, E. Hairer, G. Vilmart, A substitution law for B-series vector fields, Technical Report 5498, INRIA, France, 2005.
- [13] P. Chartier, E. Hairer, G. Vilmart, Numerical integrators based on modified differential equations, *Math. Comput.* 76 (2007) 1941-1953.
- [14] P. Chartier, E. Hairer, G. Vilmart, Composing and substituting S-series and B-series of integrators and vector fields, preprint, [www.irisa.fr/irisa/fichiers/algebraic.pdf](http://www.irisa.fr/irisa/fichiers/algebraic.pdf), (2008).
- [15] J.L. Cieřliński, Improving the accuracy of the AVF method, *J. Comput. Appl. Math.* 259 (2014) 233-243.
- [16] J. Cui, Z. Xu, Y. Wang, C. Jiang, Mass- and energy-preserving exponential Runge-Kutta methods for the nonlinear Schrödinger equation, *Appl. Math. Lett.* 112 (2021) 106770.
- [17] M. Dahlby, B. Owren, A general framework for deriving integral preserving numerical methods for PDEs, *SIAM J. Sci. Comput.* 33 (2011) 2318-2340.
- [18] K. Feng, M. Qin, *Symplectic Geometric Algorithms for Hamiltonian Systems*. Springer, Berlin, 2010.
- [19] E. Hairer, Backward analysis of numerical integrators and symplectic methods, *Ann. Numer. Math.* 1 (1994) 107-132.
- [20] E. Hairer, Energy-preserving variant of collocation methods, *JNAIAM. J. Numer. Anal. Ind. Appl. Math.* 5 (2010) 73-84.
- [21] E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 2006.
- [22] M. Hochbruck, C. Lubich, H. Selhofer, Exponential integrators for large systems of differential equations, *SIAM J. Sci. Comput.* 19 (1998) 1552-1574.
- [23] M. Hochbruck, A. Ostermann, Explicit exponential Runge-Kutta methods for semilinear parabolic problems, *SIAM J. Numer. Anal.* 43 (2005) 1069-1090.
- [24] M. Hochbruck, A. Ostermann, Exponential integrators, *Acta Numer.* 19 (2010) 209-286.
- [25] C. Jiang, Y. Wang, W. Cai, A linearly implicit energy-preserving exponential integrator for the nonlinear Klein-Gordon equation, *J. Comput. Phys.* 419 (2020) 109690.
- [26] C. Jiang, Y. Wang, Y. Gong, Explicit high-order energy-preserving methods for general Hamiltonian partial differential equations, *J. Comput. Appl. Math.* 388 (2021) 113298.
- [27] A.-K. Kassam, L.N. Trefethen, Fourth-order time stepping for stiff PDEs, *SIAM J. Sci. Comput.* 26 (2005) 1214-1233.
- [28] H.C. Li, Y.S. Wang, M.Z. Qin, A sixth order averaged vector field method, *J. Comput. Math.* 34 (2015) 479-498.
- [29] Y.W. Li, X. Wu, Exponential integrators preserving first integrals or Lyapunov functions for conservative or dissipative systems, *SIAM J. Sci. Comput.* 38 (2016) A1876-A1895.
- [30] Y.W. Li, X. Wu, Functionally fitted energy-preserving methods for solving oscillatory nonlinear Hamiltonian systems, *SIAM J. Numer. Anal.* 54 (2016) 2036-2059.
- [31] C. Liu, X. Wu, Arbitrarily high-order time-stepping schemes based on the operator spectrum theory for high-dimensional nonlinear Klein-Gordon equations, *J. Comput. Phys.* 340 (2017) 243-275.

- [32] C. Liu, X. Wu, The boundness of the operator-valued functions for multidimensional nonlinear wave equations with applications, *Appl. Math. Lett.* 74 (2017) 60-67.
- [33] K. Liu, W. Shi, X. Wu, A linearly-fitted conservative (dissipative) scheme for efficiently solving conservative (dissipative) nonlinear wave PDEs, *J Comput. Math.* 35 (2017) 780-800.
- [34] R.I. McLachlan, G.R.W. Quispel, N. Robidoux, Unified approach to Hamiltonian systems, poisson systems, gradient systems, and systems with Lyapunov functions or first integrals, *Phys. Rev. Lett.* 81 (1998) 2399-2403.
- [35] R.I. McLachlan, G.R.W. Quispel, N. Robidoux, Geometric integration using discrete gradients, *Philos. Trans. R. Soc. A* 357 (1999) 1021-1045.
- [36] L. Mei, X. Wu, The construction of arbitrary order ERKN methods based on group theory for solving oscillatory Hamiltonian systems with applications, *J. Comput. Phys.* 323 (2016) 171-190.
- [37] L. Mei, X. Wu, Symplectic exponential Runge-Kutta methods for solving nonlinear Hamiltonian systems, *J. Comput. Phys.* 338 (2017) 567-584.
- [38] Y. Miyatake, T. Matsuo, A general framework for finding energy dissipative/conservative  $H^1$ -Galerkin schemes and their underlying  $H^1$ -weak forms for nonlinear evolution equations, *BIT Numer. Math.* 54 (2014) 1119-1154.
- [39] A. Ostermann, C. Su, Two exponential-type integrators for the “good” Boussinesq equation, *Numer. Math.* 143 (2019) 683-712.
- [40] G.R.W. Quispel, D.I. McLaren, A new class of energy-preserving numerical integration methods, *J. Phys. A: Math. Theor.* 41 (2008) 045206.
- [41] W. Shi, K. Liu, X. Wu, C. Liu, An energy-preserving algorithm for nonlinear Hamiltonian wave equations with Neumann boundary conditions, *Calcolo* 54 (2017) 1379-1402.
- [42] C. Su, W. Yao, A Deuffhard-type exponential integrator Fourier pseudo-spectral method for the “good” Boussinesq equation, *J. Sci. Comput.* 83 (2020) 4.
- [43] L. Wang, W. Chen, C. Wang, An energy-conserving second order numerical scheme for nonlinear hyperbolic equation with an exponential nonlinear term, *J. Comput. Appl. Math.* 280 (2015) 347-366.
- [44] B. Wang, X. Wu, A new high precision energy-preserving integrator for system of oscillatory second-order differential equations, *Phys. Lett. A* 376 (2012) 1185-1190.
- [45] X. Wu, K. Liu, W. Shi, *Structure-Preserving Algorithms for Oscillatory Differential Equations II*, Springer-Verlag, Berlin, Heidelberg, 2015.
- [46] X. Wu, B. Wang, *Recent Developments in Structure-Preserving Algorithms for Oscillatory Differential Equations*, Springer Nature Singapore Pte Ltd (2018).
- [47] X. Wu, B. Wang, *Geometric Integrators for Differential Equations with Highly Oscillatory Solutions*, Science Press and Springer nature Singapore Pte Ltd. 2020.
- [48] X. Wu, B. Wang, L. Mei, Oscillation-preserving algorithms for efficiently solving highly oscillatory second-order ODEs, *Numer. Algor.* 86 (2021) 693-727.
- [49] X. Wu, X. You, B. Wang, *Structure-Preserving Algorithms for Oscillatory Differential Equations*, Springer-Verlag, Berlin, Heidelberg, 2013.
- [50] C. Zhang, J. Huang, C. Wang, X. Yue, On the operator splitting and integral equation preconditioned deferred correction methods for the “good” Boussinesq equation, *J. Sci. Comput.* 75 (2018) 687-712.
- [51] H. Zhang, X. Qian, J. Yan, S. Song, Highly efficient invariant-conserving explicit Runge-Kutta schemes for the nonlinear Hamiltonian differential equations, *J. Comput. Phys.* 418 (2020) 109598.
- [52] C. Zhang, H. Wang, J. Huang, C. Wang, X. Yue, A second order operator splitting numerical scheme for the “good” Boussinesq equation, *Appl. Numer. Math.* 119 (2017) 179-193.

### Appendix A: Auxiliary definition and lemma

Before presenting the detailed proof of Theorem 3.2, we introduce the following auxiliary definitions and lemmas.

**Definition 6.1.** (See, e.g. [1].) The set of bi-colored trees  $T$  is recursively defined as:

- (i) black node  $\bullet \in T$ , white node  $\circ \in T$ ;
- (ii) if  $\tau_1, \dots, \tau_m \in T$ , then the tree  $B_+(\tau_1, \dots, \tau_m)$  connecting the roots of  $\tau_i$  to a new black node is still in  $T$ , i.e.,  $B_+(\tau_1, \dots, \tau_m) \in T$ ;
- (iii) if  $\tau \in T$ , then the tree  $W_+(\tau)$  connecting the root of  $\tau$  to a new white node is still in  $T$ , i.e.,  $W_+(\tau) \in T$ .

With this definition, we introduce the following denotations:  $B_+(\emptyset) = \bullet$ ,  $W_+(\emptyset) = W_+^1(\emptyset) = \circ$ ,  $W_+^0(\tau) = \tau$  and  $W_+^m(\tau) = \overbrace{W_+(\dots W_+(\tau))}^{m\text{-fold}}$ . Suppose that  $T_b$  and  $T_w$  denote the set of trees with black root and the set of trees with white root, respectively. We then have  $T = T_b \cup T_w$  and

$$T \cup \emptyset = \bigcup_{m \geq 0} W_+^m(T_b \cup \emptyset), \quad T_w = \bigcup_{m \geq 1} W_+^m(T_b \cup \emptyset). \quad (47)$$

In particular, we denote

$$T_w^k = \bigcup_{m \geq k} W_+^m(T_b \cup \emptyset).$$

It is obvious that  $T_w^1 = T_w$  and  $T_w^i \subset T_w^j$  once  $i \geq j$ . In addition,  $T_0$  is used as the set of trees with only white nodes. Thus, for any  $\tau \in T_0$ , this is a tall tree because white node has at most one child according to Definition 6.1. Moreover, we call that  $\tau$  has  $m$  branches if  $\tau = B_+(\tau_1, \dots, \tau_m)$ , and one branch if  $\tau = W_+(\tau_1)$ .

**Definition 6.2.** (See, e.g. [1].) The elementary differential corresponding to the vector field  $\tilde{f}(y) = My + f(y)$  is a vector-valued function  $\mathcal{F}_f : \mathbb{R}^d \rightarrow \mathbb{R}^d$  recursively defined as:

- (i)  $\mathcal{F}_f(\bullet)(y) = f(y)$ ,  $\mathcal{F}_f(\circ)(y) = My$ ;
- (ii) if  $\tau = B_+(\tau_1, \dots, \tau_m)$ , then  $\mathcal{F}_f(\tau)(y) = f^{(m)}(y)(\mathcal{F}_f(\tau_1)(y), \dots, \mathcal{F}_f(\tau_m)(y))$ ;
- (iii) if  $\tau = W_+(\tau_0)$ , then  $\mathcal{F}_f(\tau)(y) = M\mathcal{F}_f(\tau_0)(y)$ .

More precisely, the elementary differential corresponding to the vector field  $\tilde{f}(y) = My + f(y)$  should be written as  $\mathcal{F}_{\tilde{f}}(\tau)(y)$  instead of  $\mathcal{F}_f(\tau)(y)$ . Here, however, we use the later for simplicity, because there is no confusion between  $\mathcal{F}_{\tilde{f}}(\tau)(y)$  and  $\mathcal{F}_f(\tau)(y)$ . Likewise, we also use the denotation  $\mathcal{F}_g(\tau)(y)$  for the case of the vector field  $\tilde{g}(y) = \tilde{M}y + g(y)$ . According to Definition 6.2, we have the following similar result:  $\mathcal{F}_g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ : (i)  $\mathcal{F}_g(\bullet)(y) = g(y)$ ,  $\mathcal{F}_g(\circ)(y) = \tilde{M}y$ ; (ii) if  $\tau = B_+(\tau_1, \dots, \tau_m)$ , then  $\mathcal{F}_g(\tau)(y) = g^{(m)}(y)(\mathcal{F}_g(\tau_1)(y), \dots, \mathcal{F}_g(\tau_m)(y))$ ; (iii) if  $\tau = W_+(\tau_0)$ , then  $\mathcal{F}_g(\tau)(y) = \tilde{M}\mathcal{F}_g(\tau_0)(y)$ . Furthermore, we will simplify  $f^{(m)}(y)(\mathcal{F}_f(\tau_1)(y), \dots, \mathcal{F}_f(\tau_m)(y))$  as  $f^{(m)}(\mathcal{F}_f(\tau_1), \dots, \mathcal{F}_f(\tau_m))$  by omitting  $y$ , if there is no any confusion.

**Definition 6.3.** (Ordered bi-colored trees). The set  $OT$  of ordered bi-colored trees is recursively defined by

$$\bullet \in OT, \circ \in OT, \quad (\omega_1, \dots, \omega_m) \in OT \text{ for all } \omega_1, \dots, \omega_m \in OT, \\ \text{and } W_+(\omega) \in OT \text{ for all } \omega \in OT.$$

Similarly to the rooted trees [12, 21], a bi-colored tree  $\tau \in T$  can be considered as an equivalent class of ordered bi-colored trees by neglecting the ordering. We denote the class of ordered bi-colored trees by  $\tau = \overline{\omega}$ . Then, any function  $\psi$  defined on  $T$  can be extended to  $OT$  by setting  $\psi(\omega) = \psi(\overline{\omega})$ . Moreover, for all  $\tau \in T$ , we can choose such an ordered bi-colored tree  $\omega(\tau) \in OT$  that  $\tau = \overline{\omega(\tau)}$ .

**Definition 6.4.** (*Partitions of a bi-colored tree*, see, e.g. [12, 13].) A partition  $p^\theta$  of an ordered bi-colored tree  $\theta \in OT$  is the ordered tree obtained from  $\theta$  by replacing some of its edges by dashed ones. We denote  $P(p^\theta) = \{s_1, \dots, s_k\}$  the list of subtrees  $s_i \in T$  obtained from  $p^\theta$  by removing dashed edges and by neglecting the ordering of each subtree. We denote  $\#(p^\theta) = k$  the number of  $s_i$ . The only subtree of  $s_i$  containing the root of  $\theta$  is expressed by  $R(p^\theta) \in T$ . We denote  $P^*(p^\theta) = \{s_1, \dots, s_k\} \setminus \{R(p^\theta)\}$  the list of  $s_i$  that do not contain the root of  $\theta$ . The set of all partitions  $p^\theta$  of  $\theta$  is denoted by  $\mathcal{P}(\theta)$ . Finally, for  $\tau \in T$  we put  $\mathcal{P}(\tau) = \mathcal{P}(\omega(\theta))$  where  $\omega(\theta) \in OT$  is given in Definition 6.3.

**Definition 6.5.** (*Admissible partitions* [12, 21]). A partition  $p^\tau \in \mathcal{P}(\tau)$  of a bi-colored tree  $\tau \in T$  is called admissible if  $\chi(p^\tau)$  is the bush tree. That is, any path from the root of  $p^\tau$  to any vertex of  $\chi(p^\tau)$  has at most one dashed edge. The set of admissible partitions  $p^\tau$  of  $\tau$  is denoted by  $\mathcal{AP}(\tau)$ .

**Definition 6.6.** (See, e.g. [12, 21]) For all trees  $\tau = B_+(\tau_1^{\mu_1}, \dots, \tau_n^{\mu_n})$  or  $\tau = W_+(\tau_1) \in T$  in  $T$ , we recursively define the following mappings:

$$|\tau| = 1 + \sum_{k=1}^n \mu_k |\tau_k| \quad (\text{the order}), \quad (48)$$

$$\sigma(\tau) = \prod_{k=1}^n \mu_k! \sigma(\tau_k)^{\mu_k} \quad (\text{the symmetry}), \quad (49)$$

$$\gamma(\tau) = |\tau| \prod_{k=1}^n \gamma(\tau_k)^{\mu_k} \quad (\text{the density}), \quad (50)$$

$$\beta(\tau) = \frac{m!}{\mu_1! \cdots \mu_n!}, \quad (51)$$

$$\nu(\tau) = \beta(\tau) \prod_{k=1}^n \nu(\tau_k)^{\mu_k}, \quad (52)$$

where  $\mu_i$  denotes the multiplicity of  $\tau_i$ , and  $m = \sum_{k=1}^n \mu_k$  denotes the exact number of branches of  $\tau$ .

**Definition 6.7.** Suppose that  $\theta \in OT$  and  $\delta_1, \dots, \delta_m \in OT$ , we then define a new ordered bi-colored tree  $\theta \circ_{(\gamma_1, \dots, \gamma_m)}^k (\delta_1, \dots, \delta_m)$ , which means selecting  $m$ -tuples  $(\gamma_1, \dots, \gamma_m)$  of possible vertices of  $\theta$ , and then attaching the  $m$  roots of  $\delta_i$  to the  $m$  vertices  $\gamma_i$  with  $m$  new branches following the  $k$ -th way.

Note that  $\gamma_i$  for  $i = 1, \dots, s$  are not required to be distinct, that is, some (even all) of them could be the same vertex. Since  $\theta \circ_{(\gamma_1, \dots, \gamma_m)}^k (\delta_1, \dots, \delta_m) \in OT$ , the possible vertices of  $\theta$  should be restricted by that if  $\gamma_j$  is the white vertex, then it must be the leaf of  $\theta$  and occurs once in  $\gamma_1, \dots, \gamma_m$ . This attaching operation can be done in  $\eta(\gamma_1, \dots, \gamma_m)$  different ways, depending on the number of upwards leaving branches on each vertex  $\gamma_i$ . More precisely,  $\eta(\gamma_1, \dots, \gamma_m) = \prod_{k=1}^m (\lambda_i + 1)$ , where  $\lambda_i$  denotes the number of upward leaving branches on the vertex  $\gamma_i$ .

**Lemma 6.8.** For each  $\tau \in T$ , the value of  $\nu(\tau)$  represents the number of ordered trees  $\omega \in OT$  such that  $\bar{\omega} = \tau$ . Therefore, given a real function  $\psi$  on bi-colored trees  $T$ , we have

$$\begin{aligned} \psi(\tau) &= \sum_{\substack{\omega \in OT \\ \bar{\omega} = \tau}} \frac{1}{\nu(\omega)} \psi(\omega), \\ \sum_{\tau \in T} \psi(\tau) &= \sum_{\omega \in OT} \frac{1}{\nu(\omega)} \psi(\omega). \end{aligned}$$

**Lemma 6.9.** For each  $\tau \in T$ , the value of  $\beta(\tau)$  represents the number of ordered tuples  $(\tau_1, \dots, \tau_m)$  of unordered trees  $\tau_i \in T$  such that  $\tau = W_+^k B_+(\tau_1, \dots, \tau_m)$ . Therefore, given a real function  $\psi$  on bi-colored trees  $T$ , we have

$$\begin{aligned} \psi(\tau) &= \sum_{\substack{\tau_1, \dots, \tau_m \in OT \\ W_+^k B_+(\tau_1, \dots, \tau_m) = \tau}} \frac{1}{\beta(B_+(\tau_1, \dots, \tau_m))} \psi(W_+^k B_+(\tau_1, \dots, \tau_m)), \\ \sum_{\tau \in T} \psi(\tau) &= \sum_{k \geq 0} \sum_{m \geq 0} \sum_{\tau_1, \dots, \tau_m \in T} \frac{1}{\beta(B_+(\tau_1, \dots, \tau_m))} \psi(W_+^k B_+(\tau_1, \dots, \tau_m)). \end{aligned}$$



Lemma 6.8 and Lemma 6.9 are adapted from those in [12] to the bi-colored trees considered in this paper. The proof is similar, and then we omit the details here.

**Lemma 6.10.** *For any  $\tau \in T$ , let  $\kappa(\tau) = \nu(\tau)\sigma(\tau)$ . If  $\omega = \theta \circ_{(\gamma_1, \dots, \gamma_m)}^k (\delta_1, \dots, \delta_m) \in OT$  and  $\delta = (\delta_1, \dots, \delta_m)$ , then we have*

$$\kappa(\omega) = \eta(\gamma_1, \dots, \gamma_m) \kappa(\theta) \frac{\kappa(\delta)}{m!}. \quad (53)$$

*Proof.* Suppose that  $\gamma$  is an arbitrary vertex of  $\tau$ . We denote  $\tau_\gamma$  as the upwards leaving branch of  $\tau$ , whose root is just the vertex  $\gamma$ . Replacing the branch  $\tau_\gamma$  with a new tree  $\tilde{\tau}_\gamma$  in  $\tau$ , we then obtain a new tree  $\tilde{\tau}$ . By induction, it follows from (49) and (52) that

$$\kappa(\tilde{\tau}) = \kappa(\tau) \cdot \frac{\kappa(\tilde{\tau}_\gamma)}{\kappa(\tau_\gamma)}.$$

Likewise, suppose that  $\tau = B_+(\tau_1, \dots, \tau_n) \in T_b$ , then using (49) and (52), we can obtain that

$$\kappa(\tau) = n! \prod_{k=1}^n \kappa(\tau_k).$$

Let  $\delta' \in T$  be an arbitrary tree. Taking account of the Butcher product (see [21]) of  $\tau$  and  $\delta'$ , i.e.,  $\tilde{\tau} = B_+(\tau_1, \dots, \tau_m, \delta')$ , we consequently have that

$$\kappa(\tilde{\tau}) = (n+1)! \kappa(\delta') \prod_{k=1}^n \kappa(\tau_k).$$

This means that

$$\frac{\kappa(\tilde{\tau})}{\kappa(\tau)} = (n+1) \kappa(\delta').$$

In addition, for the special case of  $\tau = \circ$  and  $\tilde{\tau} = W_+(\delta')$ , it can be concluded that  $\frac{\kappa(\tilde{\tau})}{\kappa(\tau)} = (0+1) \kappa(\delta')$  also holds because  $\tau = \circ$  exactly has 0 branch. Remember that for  $\omega = \theta \circ_{(\gamma_1, \dots, \gamma_m)}^k (\delta_1, \dots, \delta_m)$  if  $\gamma_j$  is the white vertex, then it must be the leaf of  $\theta$  and occurs once in  $\gamma_1, \dots, \gamma_m$ .

We next prove (53) for  $\omega = \theta \circ_{(\gamma_1, \dots, \gamma_m)}^k (\delta_1, \dots, \delta_m)$ . Let  $\omega_1 = \theta \circ_{\gamma_1}^k (\delta_1)$ , and  $\omega_{j+1} = \omega_j \circ_{\gamma_j}^k (\delta_j)$  for  $j = 1, \dots, m-1$ . It can be consequently obtained that  $\omega_m = \omega$ . Suppose that there are  $\lambda_i$  upwards leaving branches on the vertex  $\gamma_i$ . On the basis of the two above results, we can derive that  $\kappa(\omega_1) = (\lambda_1 + 1) \kappa(\delta_1) \kappa(\theta)$  and  $\kappa(\omega_{j+1}) = (\lambda_j + 1) \kappa(\delta_{j+1}) \kappa(\omega_j)$  for  $j = 1, \dots, m-1$ . This gives  $\kappa(\omega) = \kappa(\theta) \prod_{k=1}^m (\lambda_k + 1) \kappa(\delta_k)$ . Hence,  $\kappa(\omega) = \eta(\gamma_1, \dots, \gamma_m) \kappa(\theta) \prod_{k=1}^m \kappa(\delta_k)$  on account of  $\eta(\gamma_1, \dots, \gamma_m) = \prod_{k=1}^m (\lambda_k + 1)$ .  $\square$

**Definition 6.11.** *Suppose that  $\tau \in T$  is a bi-colored tree. Changing the color of only one leaf of  $\tau$  (from black to white or from white to black) yields another tree  $\tau'$ . We then call  $\tau'$  the one-leaf color-changing tree corresponding to  $\tau$ . Conversely,  $\tau$  is also the one-leaf color-changing tree corresponding to  $\tau'$ . We denote this relation by  $\tau \stackrel{1}{\rightleftharpoons} \tau'$ . Recursively, we define  $\tau_0$  as the  $k$ -leaf color-changing tree corresponding to  $\tau$ , if  $\tau_0$  can be obtained by successively changing the color of  $k$  ( $1 \leq k \leq |\tau|$ ) leaves of  $\tau$  (from black to white or from white to black) while one leaf is allowed to be changed at most once. Similarly, we denote the relation between  $\tau_0$  and  $\tau$  by  $\tau \stackrel{k}{\rightleftharpoons} \tau_0$ . In particular, for the first-order bi-colored trees we have that  $\bullet \stackrel{1}{\rightleftharpoons} \circ$ . For convenience, we denote  $\tau \stackrel{0}{\rightleftharpoons} \tau$  for arbitrary  $\tau \in T$ , which means that  $\tau$  can be obtained by changing the color of 0 leaf of  $\tau$ .*

**Lemma 6.12.** *Suppose that the mappings  $e : T \cup \{\emptyset\} \rightarrow \mathbb{R}$  and  $a : T \cup \{\emptyset\} \rightarrow \mathbb{R}$  are defined in Theorem 2.5 and Theorem 2.6, respectively. Then, the equations*

$$a(\tau) = a(\tau'), \quad (54)$$

and

$$e(\tau) = e(\tau'), \quad (55)$$

hold for all bi-colored trees  $\tau, \tau' \in T$  and positive integer  $\lambda$  that  $\tau \stackrel{\lambda}{\rightleftharpoons} \tau'$ .

*Proof.* The two equations (54) and (55) can be proved in a similar way. Hence, we only present the details about the proof for (54).

It can be confirmed from Definition 6.11 that there must exist a set of trees  $\tau_j$  for  $j = 1, 2, \dots, \lambda - 1$  that  $\tau \stackrel{1}{\rightleftharpoons} \tau_1$ ,  $\tau_{\lambda-1} \stackrel{1}{\rightleftharpoons} \tau'$  and  $\tau_j \stackrel{1}{\rightleftharpoons} \tau_{j+1}$  for  $j = 1, 2, \dots, \lambda - 2$  once  $\tau \stackrel{\lambda}{\rightleftharpoons} \tau'$  and  $\lambda \geq 2$ . With this observation, it suffices to only consider the case of  $\lambda = 1$ , i.e.,  $\tau \stackrel{1}{\rightleftharpoons} \tau'$ .

The proof will be conducted by induction on the order of  $\tau$ , i.e.  $|\tau|$ . For the two first-order trees  $\bullet$  and  $\circ$  that  $\bullet \stackrel{1}{\rightleftharpoons} \circ$ , Theorem 2.6 directly supports the equality  $a(\bullet) = a(\circ)$ .

Suppose that the equality (54) holds for  $\tau, \tau' \in T$  of order  $|\tau| = |\tau'| \leq n$ , we will show (54) also holds for  $\tau, \tau' \in T$  of order  $|\tau| = |\tau'| = n + 1$ . According to the color of the root of  $\tau, \tau'$ , we respectively consider the two cases, i.e.,  $\tau, \tau' \in T_w$  with white root and  $\tau, \tau' \in T_b$  with black root.

For the case of  $\tau, \tau' \in T_w$  and  $|\tau| = |\tau'| = n + 1$ , there must exist two bi-colored trees  $\tau_0, \tau'_0 \in T$  and a positive integer  $k$  satisfying  $\tau = W_+^k(\tau_0)$ ,  $\tau' = W_+^k(\tau'_0)$ ,  $|\tau_0| = |\tau'_0| = n + 1 - k \leq n$ , and  $\tau_0 \stackrel{1}{\rightleftharpoons} \tau'_0$ . It follows from Theorem 2.6 that  $a(\tau) = \frac{a(\tau_0)}{(k+1)!}$  and  $a(\tau') = \frac{a(\tau'_0)}{(k+1)!}$ . Due to the induction hypothesis that  $a(\tau_0) = a(\tau'_0)$  because of  $\tau_0 \stackrel{1}{\rightleftharpoons} \tau'_0$ , the equation  $a(\tau) = a(\tau')$  holds naturally.

For the other case of  $\tau, \tau' \in T_b$  with  $m (\geq 1)$  branches and  $|\tau| = |\tau'| = n + 1$ , there must exist  $m + 1$  trees  $\tau_1, \dots, \tau_m$  and  $\tau'_m$  that  $\tau = B_+(\tau_1, \dots, \tau_{m-1}, \tau_m)$ ,  $\tau' = B_+(\tau_1, \dots, \tau_{m-1}, \tau'_m)$ , and  $\tau_m \stackrel{1}{\rightleftharpoons} \tau'_m$ . It then follows from Theorem 2.6 that  $a(\tau) = \frac{a(\tau_1) \cdots a(\tau_{m-1}) \cdot a(\tau_m)}{(m+1)!}$  and  $a(\tau') = \frac{a(\tau_1) \cdots a(\tau_{m-1}) \cdot a(\tau'_m)}{(m+1)!}$ . In addition, the fact  $\tau_m \stackrel{1}{\rightleftharpoons} \tau'_m$  and  $|\tau_m| = |\tau'_m| \leq n + 1 - m \leq n$  yields  $a(\tau_m) = a(\tau'_m)$ . Then, we can obtain that  $a(\tau) = a(\tau')$  also holds.

Clearly, the proof is complete by the basis and the inductive step.  $\square$

## Appendix B: Proof of Theorem 3.2

*Proof.* The key step to complete the proof of this theorem is that the modified vector field  $\tilde{g}(y)$  has the expression as that in (25). Then, with the formulation of  $\tilde{g}(y)$  in (25), the result on energy-preserving methods in [13] and Corollary IX. 5.4 in [21] can deduce that  $H(y) = \frac{1}{2}y^T A y + U(y)$  is also the first integral of the differential equation

$$y'(t) = \tilde{g}^{[2k]}(y) = \overline{Q}_k \cdot (A y + \nabla U(y)).$$

This implies that  $(H'(y))^T \tilde{g}(y) = 0$  holds for every possible  $y$ , i.e.,

$$(A y + \nabla U(y))^T \overline{Q}_k (A y + \nabla U(y)) = 0.$$

On account of the arbitrariness of  $y$ ,  $A$  and  $U$ , we then have that  $\overline{Q}_k$  must be skew-symmetric for all  $k \geq 0$ . The energy preservation of the high-order integrator (27) thus follows from the skew-symmetry of  $\overline{Q}_k$  and Theorem 2.3 in [29]. Moreover, the order of (27) directly follows from the result in [13], and the symmetry of (27) can be proved similarly to (22). With the above analysis, we pay our attention to proving the equation (25).

Suppose that  $\tau \in T$  is a tree of order  $|\tau| = n (> 1)$  whose all  $m (< n)$  leaves are black nodes. We introduce the set  $\Omega = \{\tau' \in T | \tau' \stackrel{k}{\rightleftharpoons} \tau, k = 0, 1, \dots, m\}$ . Obviously, the number of elements in  $\Omega$  is  $2^m$ . We number all the leaves by  $1, 2, \dots, m$  according to their positions in order from right to left. Denote the black node as 0 and the white node as 1. Arranging the  $m$  leaves according to their positions in order from right to left, then a tree  $\tau' \in \Omega$  will be represented by a sequence with  $m$  numbers containing only 0 and 1. In this way, we obtain a decimal number corresponding to  $\tau'$  once the sequence is consider as a binary number. That is, we have numbered all the  $2^m$  trees in  $\Omega$ . For example,  $\tau$  is numbered 0, while  $\tilde{\tau}$ , whose leaves are all white, i.e.,  $\tilde{\tau} \stackrel{m}{\rightleftharpoons} \tau$ , is numbered  $2^m - 1$ .

Denote  $\tau_j$  as the tree numbered  $j$  in  $\Omega$ . We now consider the  $k$ -th tree  $\tau_k$  ( $k = 0, 1, \dots, 2^{m-1} - 1$ ) and its counterpart  $\tau_{2^{m-1}+k}$ . Obviously, the only difference between  $\tau_k$  and  $\tau_{2^{m-1}+k}$  is that the leftmost leaf of  $\tau_k$  is black while it is white for  $\tau_{2^{m-1}+k}$ . That is, the relation  $\tau_k \stackrel{1}{\rightleftharpoons} \tau_{2^{m-1}+k}$  holds for all  $k = 0, 1, \dots, 2^{m-1} - 1$ . Then, all the  $2^m$  trees in  $\Omega$  can be separated into  $2^{m-1}$  pairs  $(\tau_k, \tau_{2^{m-1}+k})$ , and each tree in  $\Omega$  belongs to a unique pair  $(\tau_k, \tau_{2^{m-1}+k})$ . We next prove that the mapping  $b(\cdot)$  satisfies  $b(\tau_k) = b(\tau_{2^{m-1}+k})$  by induction on the order of the tree  $\tau$ .



It can be confirmed from Table 2 that  $b(\tau_k) = b(\tau_{2^{m-1}+k})$  holds for trees of order  $|\tau| \leq 3$ . Suppose that  $b(\tau_k) = b(\tau_{2^{m-1}+k})$  holds for all trees of  $|\tau| \leq n$  ( $n > 3$ ). We are then to prove that  $b(\tau_k) = b(\tau_{2^{m-1}+k})$  also holds for trees of order  $|\tau| = n + 1$ .

Consider the set  $\mathcal{P}(\tau_k)$  of all partitions  $p^{\tau_k}$  of  $\tau_k$  and its counterpart  $\mathcal{P}(\tau_{2^{m-1}+k})$ . Since the only difference between  $\tau_k$  and  $\tau_{2^{m-1}+k}$  is the color of their leftmost leaf, it follows from Definition 2.11 of [12] that each partition  $p^{\tau_k} \in \mathcal{P}(\tau_k)$  corresponds a unique partition  $p^{\tau_{2^{m-1}+k}} \in \mathcal{P}(\tau_{2^{m-1}+k})$  that their dashed edges in the partitions are in the same positions. Then, the subtree set  $P(p^{\tau_k}) = \{s_1, \dots, s_k, s_i \in T\}$  generated from the partition  $p^{\tau_k}$  and its counterpart  $P(p^{\tau_{2^{m-1}+k}}) = \{s'_1, \dots, s'_k, s'_i \in T\}$  must satisfy

$$s_1 \stackrel{1}{=} s'_1, \quad s_i = s'_i, \quad i = 2, \dots, k, \quad (56)$$

if  $s_i$  and  $s'_i$  are arranged appropriately. Moreover, Definition 2.2 leads to that the skeletons  $\chi(p^{\tau_k})$  of the partition  $p^{\tau_k}$  and  $\chi(p^{\tau_{2^{m-1}+k}})$  of the partition  $p^{\tau_{2^{m-1}+k}}$  satisfy either

$$\chi(p^{\tau_k}) = \chi(p^{\tau_{2^{m-1}+k}}), \quad (57)$$

or

$$\chi(p^{\tau_k}) \stackrel{1}{=} \chi(p^{\tau_{2^{m-1}+k}}). \quad (58)$$

Both (57) and (58) can yield that

$$a(\chi(p^{\tau_k})) = a(\chi(p^{\tau_{2^{m-1}+k}})), \quad (59)$$

by Lemma 6.12.

Denote  $\{\tau_k\}$  as the subtree set of the special partition of  $\tau_k$  that no dashed edge there exists. Then the partition  $p^{\tau_k} \in \mathcal{P}(\tau_k) \setminus \{\tau_k\}$  satisfies that the subtree set  $P(p^{\tau_k})$  contains as least two subtrees, thereby deducing that the subtree  $s_i$  in this set is of order no more than  $n$ , namely,  $|s_i| \leq n$ . A similar result can be obtained that all the subtrees in the set  $P(p^{\tau_{2^{m-1}+k}})$  of the partition  $p^{\tau_{2^{m-1}+k}} \in \mathcal{P}(\tau_{2^{m-1}+k}) \setminus \{\tau_{2^{m-1}+k}\}$  are of order no more than  $n$ , namely,  $|s'_i| \leq n$ . Consequently, it follows from the induction hypothesis and (56) that

$$b(s_i) = b(s'_i), \quad i = 1, \dots, k, \quad (60)$$

holds for the subtrees arranged as in (56). The simultaneous consideration of (59) and (60) yields that

$$a(\chi(p^{\tau_k})) \prod_{\delta \in P(p^{\tau_k})} b(\delta) = a(\chi(p^{\tau_{2^{m-1}+k}})) \prod_{\delta \in P(p^{\tau_{2^{m-1}+k}})} b(\delta) \quad (61)$$

holds for the partition  $p^{\tau_k} \in \mathcal{P}(\tau_k) \setminus \{\tau_k\}$  and its counterpart  $p^{\tau_{2^{m-1}+k}} \in \mathcal{P}(\tau_{2^{m-1}+k}) \setminus \{\tau_{2^{m-1}+k}\}$ . Summing over all the partitions  $p^{\tau_k}$  of  $\mathcal{P}(\tau_k) \setminus \{\tau_k\}$  in the left-hand side of (61) and that of  $\mathcal{P}(\tau_{2^{m-1}+k}) \setminus \{\tau_{2^{m-1}+k}\}$  in the right-hand side of (61) further gives

$$\sum_{p^{\tau_k} \in \mathcal{P}(\tau_k) \setminus \{\tau_k\}} a(\chi(p^{\tau_k})) \prod_{\delta \in P(p^{\tau_k})} b(\delta) = \sum_{p^{\tau_{2^{m-1}+k}} \in \mathcal{P}(\tau_{2^{m-1}+k}) \setminus \{\tau_{2^{m-1}+k}\}} a(\chi(p^{\tau_{2^{m-1}+k}})) \prod_{\delta \in P(p^{\tau_{2^{m-1}+k}})} b(\delta). \quad (62)$$

Since  $\tau_k \stackrel{1}{=} \tau_{2^{m-1}+k}$  and the mappings  $a(\cdot)$ ,  $b(\cdot)$  and  $e(\cdot)$  satisfy  $b \star a(\tau) = e(\tau)$  for all  $\tau \in T$ , using Theorem 6.12 we obtain

$$b \star a(\tau_k) = e(\tau_k) = e(\tau_{2^{m-1}+k}) = b \star a(\tau_{2^{m-1}+k}). \quad (63)$$

Note that the skeleton  $\chi(\{\tau_k\})$  corresponding to  $\{\tau_k\}$  must be the black node  $\bullet$ , and the skeleton  $\chi(\{\tau_{2^{m-1}+k}\})$  corresponding to  $\{\tau_{2^{m-1}+k}\}$  may be either the black node  $\bullet$  or the white node  $\circ$ . Taking the adapted substitution law (10) into consideration gives

$$b \star a(\tau_k) = a(\chi(\{\tau_k\})) \cdot b(\tau_k) + \sum_{p^{\tau_k} \in \mathcal{P}(\tau_k) \setminus \{\tau_k\}} a(\chi(p^{\tau_k})) \prod_{\delta \in P(p^{\tau_k})} b(\delta), \quad (64)$$

and

$$\begin{aligned} b \star a(\tau_{2^{m-1}+k}) &= a(\chi(\{\tau_{2^{m-1}+k}\})) \cdot b(\tau_{2^{m-1}+k}) \\ &+ \sum_{p^{\tau_{2^{m-1}+k}} \in \mathcal{P}(\tau_{2^{m-1}+k}) \setminus \{\tau_{2^{m-1}+k}\}} a(\chi(p^{\tau_{2^{m-1}+k}})) \prod_{\delta \in P(p^{\tau_{2^{m-1}+k}})} b(\delta). \end{aligned} \quad (65)$$

The relation  $a(\bullet) = a(\circ) = 1$  finally yields that  $b(\tau_k) = b(\tau_{2^{m-1}+k})$  along with the equations (62)–(65).

The basis and the inductive step prove that  $b(\tau_k) = b(\tau_{2^{m-1}+k})$  for the special pair  $(\tau_k, \tau_{2^{m-1}+k})$  of arbitrary order. As we have showed previously, all trees of order  $n > 1$  can be divided into at most  $n - 1$  sets  $\Omega_1, \Omega_2, \dots, \Omega_{n-1}$ , where  $\Omega_i$  has  $i$  leaves. Then, the  $2^i$  trees in  $\Omega_i$  can be separated into  $2^{i-1}$  pairs  $(\tau_k, \tau_{2^{i-1}+k})$  for  $k = 0, 1, \dots, 2^{i-1} - 1$ . As we have proved,  $b(\tau_k) = b(\tau_{2^{i-1}+k})$  indicates that the elementary differentials respectively corresponding to  $\tau_k$  and  $\tau_{2^{i-1}+k}$  have the same coefficient. It then follows from the definition of elementary differential in [1] that

$$\begin{aligned} b(\tau_k)\mathcal{F}_f(\tau_k) + b(\tau_{2^{i-1}+k})\mathcal{F}_f(\tau_{2^{i-1}+k}) &= c_k G_k(y)(My + f(y)) \\ &= c_k(G_k(y)Q)(Ay + \nabla U(y)), \end{aligned} \quad (66)$$

where  $c_k = b(\tau_k)$  and  $G_k(y)$  is a function depending only on  $\tau_k$ . Then, summing over all pairs  $(\tau_k, \tau_{2^{m-1}+k})$  of order  $n$  gives the formulation of  $\tilde{g}(y)$  in (25). The absence of odd power of  $h$  in (25) is due to the symmetry of the EAVF integrator (3). This completes the proof.  $\square$

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:



Corresponding author of the paper [YJCPH\_110429]: Xinyuan Wu  
17 May 2021

**All authors declare that no conflict of interest exists.**

Journal Pre-proof

**All authors should be listed and all authors declare that the descriptions are accurate and agreed by all authors.**