Numerical Algorithms (2024) 96:1295–1329
https://doi.org/10.1007/s11075-024-01781-4

ORIGINAL PAPER

High-order linearly implicit exponential integrators conserving quadratic invariants with application to scalar auxiliary variable approach

Shun Sato[1]

Received: 21 August 2023 / Accepted: 11 February 2024 / Published online: 29 February 2024

Abstract
This paper proposes a framework for constructing high-order linearly implicit exponential integrators that conserve a quadratic invariant. This is then applied to the scalar auxiliary variable (SAV) approach. Quadratic invariants are significant objects that are present in various physical equations and also in computationally efficient conservative schemes for general invariants. For instance, the SAV approach converts the invariant into a quadratic form by introducing scalar auxiliary variables, which have been intensively studied in recent years. In this vein, Sato et al. (Appl. Numer. Math. **187**, 71-88 2023) proposed high-order linearly implicit schemes that conserve a quadratic invariant. In this study, it is shown that their method can be effectively merged with the Lawson transformation, a technique commonly utilized in the construction of exponential integrators. It is also demonstrated that combining the constructed exponential integrators and the SAV approach yields schemes that are computationally less expensive. Specifically, the main part of the computational cost is the product of several matrix exponentials and vectors, which are parallelizable. Moreover, we conduct some mathematical analyses on the proposed schemes.

Keywords Ordinary differential equations · Quadratic invariants · Geometric numerical integration · Exponential integrators · Scalar auxiliary variable approach

Mathematics Subject Classification (2010) 65L05 · 65M06 · 65P10

✉ Shun Sato
shun@mist.i.u-tokyo.ac.jp

[1] Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo-ku, Tokyo 1138656, Japan

1 Introduction

In this paper, we consider ordinary differential equations (ODEs)

$$\dot{y} = S(y)\nabla V(y), \tag{1}$$

on a finite-dimensional real Hilbert space $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$, where $y\colon [0,T) \to \mathcal{X}$ is a dependent variable, $S\colon \mathcal{X} \to \mathcal{L}(\mathcal{X})$ is a "skew-symmetric matrix function" ($\mathcal{L}(\mathcal{X})$ denotes the space of linear operators on $\mathcal{X}$), i.e., $\langle x, S(z)y\rangle_{\mathcal{X}} = -\langle S(z)x, y\rangle_{\mathcal{X}}$ holds for all $x, y, z \in \mathcal{X}$, $V\colon \mathcal{X} \to \mathbb{R}$ is a differentiable function, and $\nabla V : \mathcal{X} \to \mathcal{X}$ denotes the gradient of V with respect to the inner product of $\mathcal{X}$. Although the methods in this paper can be applied even when $\mathcal{X}$ is infinite-dimensional, we limit ourselves to finite-dimensional cases for the sake of mathematical analysis (see Remark 4).

The class of ODEs in the form (1) includes many examples: Hamiltonian systems, Poisson systems, and spatial discretization of variational partial differential equations (PDEs) (see, e.g., [1]). These ODEs satisfy the conservation law with respect to V:

$$\frac{\mathrm{d}}{\mathrm{d}t}V(y(t)) = \langle \nabla V(y(t)), \dot{y}(t)\rangle_{\mathcal{X}} = \langle \nabla V(y(t)), S(y(t))\nabla V(y(t))\rangle_{\mathcal{X}} = 0,$$

where the last equality holds due to the skew-symmetry of $S(y(t))$.

Since this conservation law is an important property of the differential equation (1), numerical methods that preserve it have been studied in the literature, for example, the discrete gradient method [2–4] for the gradient ODEs and the discrete variational derivative method [5, 6] (see also [7]) for variational PDEs. In addition, Cohen and Hairer [8] proposed a high-order extension of the discrete gradient method.

Since these conservative numerical schemes are fully implicit, several techniques have been devised to reduce computational cost. For example, Besse [9] and Zhang, Pérez-García, and Vázquez [10] proposed linearly implicit conservative schemes for the nonlinear Schrödinger equation. For polynomial invariants, Matsuo and Furihata [11] proposed multistep linearly implicit DVDM (see also Dahlby and Owren [12]). Recently, Yang and Han [13] proposed the invariant energy quadratization (IEQ) approach, and Shen et al. [14] proposed the scalar auxiliary variable (SAV) approach (see also [15] and the references therein). The SAV approach employs scalar auxiliary variables to convert the invariant into a quadratic form, which is then preserved by a linearly implicit scheme (see Sect. 2.4 for details).

Although the above computationally inexpensive methods differ, they have in common that they attribute the invariant to a quadratic function in some way. Given this, Sato et al. [16] recently proposed a framework for constructing high-order and linearly implicit schemes conserving a quadratic invariant.

Exponential integrators (cf. [17]) are efficient numerical methods for solving semilinear differential equations, and combining it with geometric numerical integration has also been studied. Celledoni et al. [18] deal with L^2 norm conservation for the Schrödinger equation and derive the condition to preserve it in exponential Runge–Kutta methods. Mei et al. [19] deal with (1) with $V(y) = \frac{1}{2}\langle y, y\rangle_{\mathcal{X}}$ and construct conservative exponential integrators. Some researchers deal with the equation in the

form $\dot{y} = J(Ly + \nabla E(y))$, where J is skew-symmetric, L is symmetric, and E is a function: Li and Wu [20] propose second-order exponential discrete gradient schemes; Mei et al. [21] discuss how to design high-order conservative schemes based on Li and Wu [20], modified differential equation, and order condition in terms of B-series; and Li [22] proposes a multistep linearly implicit conservative exponential scheme based on polarization.

Combinations of exponential integrators and the SAV approach have also been investigated [23–26]. In particular, Jiang et al. [27] propose high-order linearly implicit structure-preserving exponential integrators for the nonlinear Schrödinger equation based on the SAV approach and the Lawson transformation [28]. They also mention that the same method can be applied to some general differential equations, and numerically confirm the superiority of their method. However, no theoretical guarantee of high accuracy is given. In this paper, we reveal that linearly implicit high-order conservative schemes proposed by Sato et al. [16] can be combined with exponential integrators (Sect. 3). The resulting scheme, although linearly implicit, is computationally not very cheap, since it includes matrix exponentials in the coefficient matrix. However, when combined with the SAV approach, its structure can be exploited to provide a computationally efficient implementation (Sect. 4). Specifically, the main part of the computational cost is a few products of matrix exponential functions and vectors with the size of the dimension of the given differential equation. Furthermore, these products can be computed in parallel. Theoretical guarantees such as accuracy are also provided, albeit being limited to finite dimensions. (The relationship with [27] is discussed in Remark 6.)

Here, we note a limitation of the proposed method: it cannot be applied to dissipative systems. In many cases, if a conservative numerical method can be constructed, a dissipative numerical method, one that replicates the dissipation law, can be constructed correspondingly: [16] can also be applied to dissipative cases. However, this is not the case for the proposed method in the present paper. The proposed method is essentially driven by the fact that under a mild assumption, the quadratic invariants are invariant by the Lawson transformation (see Lemma 4), and the extension to dissipative cases is nontrivial (see Remark 2).

The remainder of the paper is organized as follows. In Sect. 2, we briefly review the canonical Runge–Kutta methods, linearly implicit high-order conservative schemes proposed by Sato et al. [16], the Lawson transformation, and the SAV approach. The contents in Sects. 3 and 4 are already described above. The proposed schemes are numerically examined in Sect. 5.

2 Preliminaries

2.1 Runge–Kutta methods and quadratic invariants

For the autonomous system

$$\dot{y} = f(y),$$

general Runge–Kutta methods that compute an approximation $y_1 \approx y(h)$ from $y_0 = y(0)$ can be written as

$$\begin{cases} Y_i = y_0 + h \sum_{j\in[s]} a_{ij} f(Y_j) & \left(i \in [s]\right), \\ y_1 = y_0 + h \sum_{i\in[s]} b_i f(Y_i), \end{cases}$$

where $[s] := \{1, 2, \ldots, s\}$. Throughout this paper, the fixed step size is considered, but as is common in the context of conservative methods, variable step sizes can also be employed without difficulty.

A subclass of Runge–Kutta methods preserves all quadratic invariants.

Proposition 1 (Cooper [29]) *Runge–Kutta methods satisfying*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \qquad i, j \in [s] \tag{2}$$

automatically preserve all quadratic invariants.

The Runge–Kutta methods satisfying (2) are said to be *canonical* (see [30] for details on canonical Runge–Kutta methods). Note that a canonical Runge–Kutta method must be implicit, yet diagonally implicit Runge–Kutta methods can be canonical.

For example, the second-order Gauss, fourth-order Gauss, and third-order diagonally implicit canonical Runge–Kutta methods are as follows:

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}, \qquad \begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}, \qquad \begin{array}{c|ccc} \frac{\alpha}{2} & \frac{\alpha}{2} & 0 & 0 \\ \frac{3}{2}\alpha & \alpha & \frac{\alpha}{2} & 0 \\ \frac{1}{2} + \alpha & \alpha & \alpha & \frac{1}{2} - \alpha \\ \hline & \alpha & \alpha & 1 - 2\alpha \end{array},$$

where $\alpha = \frac{1}{3}\left(2 + \frac{1}{2^{1/3}} + 2^{1/3}\right)$.

2.2 High-order linearly implicit schemes that conserve quadratic invariants

In this section, we briefly review the high-order linearly implicit schemes for the gradient system (1) with a quadratic invariant V proposed by [16].

Definition 1

Step 0 Prepare $Y_i^{(0)} \approx y(c_i h)$ *and set* $k = 1$.
Step 1 Solve the linear equation system

$$Y_i^{(k)} = y_0 + h \sum_{j\in[s]} a_{ij} S\left(Y_j^{(k-1)}\right) \nabla V\left(Y_j^{(k)}\right) \qquad \left(i \in [s]\right) \tag{3}$$

to obtain $Y_i^{(k)}$*'s. If some criteria hold, go to Step 2. Otherwise, set* $k = k + 1$ *and repeat Step 1.*

Step 2 Output

$$y_1^{(k)} = y_0 + h \sum_{j \in [s]} b_j S\left(Y_j^{(k-1)}\right) \nabla V\left(Y_j^{(k)}\right).$$

Theorem 2 *Suppose that A and b satisfy (2), and V is quadratic. Then, the solution $y_1^{(k)}$ of* Definition 1 *satisfies $V\left(y_1^{(k)}\right) = V(y_0)$ for any $h > 0$.*

Theorem 3 *Assume the following conditions:*

(A1) $Y_i^{(0)}$ satisfies $\left\|Y_i^{(0)} - y(c_i h)\right\|_{\mathcal{X}} \le Ch^q$ for each $i \in [s]$.
(A2) $S\colon \mathcal{X} \to \mathcal{L}(\mathcal{X})$ is Lipschitz continuous.
(A3) The base Runge–Kutta method is of order p.

Then, the numerical solution $y_1^{(k)}$ of the scheme defined by Definition 1 *satisfies*

$$\left\|y_1^{(k)} - y(h)\right\|_{\mathcal{X}} \le C' h^{\min\{p, q+k-1\}+1}$$

for sufficiently small $h > 0$, i.e., the scheme with k iteration is of order $\min\{p, q+k-1\}$. Here, C' is a constant depending only on the exact solution y, k, S, V, A, b, and the constant C in (A1).

Remark 1 *In view of* Theorem 3, *$k = p - q + 1$ is a natural choice as the criterion in Step 1 of* Definition 1. *In addition, there are several other possibilities. As shown in* [16, Theorem 4.1], *$Y_i^{(k)}$ linearly coverges to the corresponding inner stage of the base Runge–Kutta method. Based on this fact, one can choose k so that the update $\left\|Y_i^{(k)} - Y_i^{(k-1)}\right\|_{\mathcal{X}}$ is sufficiently small. Another possibility is to choose k so that $\left\|y_1^{(k)} - y_1^{(k-1)}\right\|_{\mathcal{X}}$, an approximation of the local error, is sufficiently small (note that, when $k \le p - q + 1$, $y_1^{(k)}$ and $y_1^{(k-1)}$ are approximations of $y(h)$ with different order of accuracy, which are known to be useful to approximate the local error). However, in the numerical experiments in* Sect. 5, *we simply choose some fixed k to verify the theoretical order of accuracy.*

2.3 Lawson transformation

We briefly review the Lawson transformation [28] which is useful to construct exponential integrators. For the ODE in the form

$$\dot{y} = My + f(y), \tag{4}$$

we consider the transformation $w(t) = \exp(-tM)y(t)$ to obtain

$$\dot{w} = \exp(-tM) f\left(\exp(tM)w(t)\right), \tag{5}$$

where exp denotes the matrix exponential function. It is known that, when M has large eigenvalues, the transformed system (5) is easier to solve numerically than the original ODE. Therefore, a good approach to solve the system (4) is to apply a numerical

method to (5) and use the inverse transformation to return to the original variable. If a Runge–Kutta method is adopted as the numerical method, the resulting numerical method is written as

$$\begin{cases} Y_i = \exp(c_i hM) y_0 + h \sum_{j\in[s]} a_{ij} \exp((c_i - c_j)hM) f(Y_j) \qquad (i \in [s]), \\ y_1 = \exp(hM) y_0 + h \sum_{i\in[s]} b_i \exp\left((1 - c_i)hM\right) f\left(Y_i\right). \end{cases} \tag{6}$$

Hereafter, the scheme is designated the *Lawson method*.

2.4 Scalar auxiliary variable approach

In this section, we review the scalar auxiliary variable (SAV) approach for the Hamiltonian system

$$\dot{u} = J \nabla H(u) \tag{7}$$

on a finite-dimensional real Hilbert space $(\mathcal{V}, \langle\cdot,\cdot\rangle)$, where $u\colon [0, T) \to \mathcal{V}$ is a dependent variable, $J \in \mathcal{L}(\mathcal{V})$ is a (constant) skew-symmetric linear operator, and $H\colon \mathcal{V} \to \mathbb{R}$ is written in the form

$$H(u) = \frac{1}{2}\langle Lu, u\rangle + E(u),$$

where $L \in \mathcal{L}(\mathcal{V})$ is symmetric.

Here, for brevity, we assume that the function $E\colon \mathcal{V} \to \mathbb{R}$ is bounded from below, i.e., $\alpha := -\inf E(u) + \epsilon$ is finite (see [15] for unbounded cases). Using this assumption, we introduce a scalar auxiliary variable $r := \sqrt{E(u) + \alpha}$. Then, (7) is rewritten as

$$\begin{cases} \dot{u} = J\left(Lu + 2r\phi(u)\right), \\ \dot{r} = \langle \phi(u), \dot{u}\rangle, \end{cases} \tag{8}$$

where $\phi(u) := \nabla E(u) / \left(2\sqrt{E(u) + \alpha}\right)$.

The system (8) has a quadratic modified invariant $V(u, r) = \frac{1}{2}\langle Lu, u\rangle + r^2 - \alpha$. Since quadratic invariants are much easier to preserve in numerical schemes than general invariants, this property enables us to construct computationally efficient conservative schemes (see, e.g., [14]).

A comprehensive way to construct conservative schemes based on the SAV approach is given in [15]; the reformulated system (8) can be rewritten for it to be regarded as a special case of (1). To this end, the inner product space $\mathcal{X}$ is defined by $\mathcal{X} = \mathcal{V} \times \mathbb{R}$, and the associated inner product is defined as $\langle (x_1, r_1), (x_2, r_2)\rangle_{\mathcal{X}} = \langle x_1, x_2\rangle + r_1 r_2$. Then, the system (8) can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u \\ r \end{bmatrix} = \begin{bmatrix} I & \phi(u) \\ & 1 \end{bmatrix}^* \begin{bmatrix} J & \\ & 0 \end{bmatrix} \begin{bmatrix} I & \phi(u) \\ & 1 \end{bmatrix} \nabla V(u, r), \tag{9}$$

where the superscript $*$ denotes the adjoint.

Since the scheme in Sect. 2.2 can be applied to this system, such reformulation enables the construction of high-order linearly implicit schemes that conserve the modified invariant V. However, high-order schemes require solving large linear equation systems (see Appendix 4 for details).

3 Exponential Runge–Kutta methods conserving quadratic invariants

Let us consider the system

$$\dot{y} = My + S(y)\nabla V(y), \tag{10}$$

where $S\colon \mathcal{X} \to \mathcal{L}(\mathcal{X})$ is a skew-symmetric matrix function, and V is a quadratic function ($V(y) = \frac{1}{2}\langle y, Qy\rangle_{\mathcal{X}}$). We further assume that V is also an invariant of the linear part $\dot{y} = My$. In Sect. 3.1, we show that the Lawson method based on the canonical Runge–Kutta method conserves the quadratic invariant V. Then, Sect. 3.2 shows that the combination of the Lawson transformation and the scheme defined by Definition 1 yields a linearly implicit exponential integrators that conserves the quadratic invariant V. Although the linearly implicit exponential integrators are not always computationally efficient, but as we will see in Sect. 4, it works very well with the SAV approach.

3.1 Lawson transformation and quadratic invariants

The following lemma, which is a slight extension of [20, Lemma 2.2], is crucial in constructing conservative schemes (see Remark 2).

Lemma 4 *Let $V\colon \mathcal{X} \to \mathbb{R}$ be a quadratic function, i.e., $V(y) = \frac{1}{2}\langle y, Qy\rangle_{\mathcal{X}}$, where $Q \in \mathcal{L}(\mathcal{X})$ is symmetric. Then, V is an invariant of the linear ODE $\dot{y} = My$ if and only if $(\exp(tM))^* Q \exp(tM) = Q$ holds for any $t \in \mathbb{R}$.*

Proof Since $y(t) = \exp(tM)y(0)$ holds for a solution of the linear ODE $\dot{y} = My$, $V(y(t))$ can be written as

$$\begin{aligned} V(y(t)) &= \frac{1}{2}\langle \exp(tM)y(0), Q\exp(tM)y(0)\rangle_{\mathcal{X}} \\ &= \frac{1}{2}\big\langle y(0), (\exp(tM))^* Q\exp(tM)y(0)\big\rangle_{\mathcal{X}}. \end{aligned}$$

Therefore, V is an invarint, i.e., $V(y(t)) = V(y(0))$ holds for any $t \in \mathbb{R}$ and initial value $y(0)$, if and only if $(\exp(tM))^* Q \exp(tM) = Q$ holds for any $t \in \mathbb{R}$. □

Assumption that V is an invariant of the linear ODE $\dot{y} = My$ may seem restrictive, but there are many examples satisfying this assumption. For example, since the difference operators often commute with each other, ODEs obtained by a finite difference discretization of PDEs often satisfy this assumption. Moreover, as shown in Sect. 4, ODEs obtained by the SAV approach satisfy this assumption.

Remark 2 *In* [20]*, it is assumed that "$M = SQ$ holds with a skew-symmetric matrix S" instead of "V is an invariant." The latter condition is weaker in the sense that, even when V is an invariant of the linear ODE $\dot{y} = My$, we cannot conclude the existence of a skew-symmetric matrix S that satisfies $M = SQ$; the pair*

$$Q = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 0 \end{bmatrix}, \qquad M = \begin{bmatrix} & 1 & \\ -1 & & \\ & & 1 \end{bmatrix}$$

is a counterexample.

The lemma in [20] *also deals with dissipative cases.* Lemma 4 *can also be extended to the dissipative cases. Under the setting of* Lemma 4*, if V is a weak Lyapunov function, then $(\exp(tM))^* Q \exp(tM) \le Q$ holds for any $t > 0$. However, extending the following theorem to dissipative systems is nontrivial and will be the subject of future research.*

By using the lemma above, we show a sufficient condition for conserving quadratic invariants.

Theorem 5 *Suppose that V is a quadratic invariant of the semilinear ODE* (4)*. Assume that V is also an invariant of the linear part $\dot{y} = My$. In addition, we assume (A, b) satisfies* (2)*. Then, the solution y_1 of the Lawson method* (6) *satisfies $V(y_1) = V(y_0)$.*

Proof Using the Lawson transformation $w(t) = \exp(-tM)\, y(t)$ and Lemma 4, we see

$$\begin{aligned} V(y(t)) &= \frac{1}{2}\langle y(t), Qy(t)\rangle_{\mathcal{X}} \\ &= \frac{1}{2}\langle \exp(tM)\, y(t), Q \exp(tM)\rangle_{\mathcal{X}} y(t) \\ &= \frac{1}{2}\langle w(t), Qw(t)\rangle_{\mathcal{X}}. \end{aligned}$$

Since this implies the transformed ODE (5) also has a quadratic invariant in the form $\frac{1}{2}\langle w, Qw\rangle_{\mathcal{X}}$, a canonical RK method applied to the system (5) preserves the invariant (Proposition 1). □

3.2 Linearly implicit exponential integrators conserving quadratic invariants

As shown in the previous section, a class of Lawson methods automatically preserves quadratic invariants under the additional assumption that the linear part also preserves the invariants. However, since all canonical RK methods are implicit, the resulting scheme must be fully implicit. Therefore, in this section, we construct linearly implicit schemes conserving quadratic invariants. To this end, we apply the schemes introduced in Sect. 2.2 instead of the Lawson methods.

For (10), the Lawson transformation $w(t) = \exp(-tM)\, y(t)$ gives

$$\dot{w} = \exp(-tM) S\, (\exp(tM) w(t))\, Q \exp(tM) w(t).$$

Using Lemma 4 and $(\exp(tM))^{-1} = \exp(-tM)$, we can further rewrite the equation as

$$\dot{w} = \exp(-tM)S\left(\exp(tM)w(t)\right)\left(\exp(-tM)\right)^* \nabla V(w(t)).$$

Since the map $w \mapsto \exp(-tM)S\left(\exp(tM)w\right)\left(\exp(-tM)\right)^*$ is a skew-symmetric matrix function, we can apply the scheme in Sect. 2.2. The inverse transformation reads the following scheme.

Definition 2

Step 0 Prepare $Y_i^{(0)} \approx y(c_i h)$ *and set* $k = 1$.
Step 1 Solve the linear equation system

$$Y_i^{(k)} = \exp(c_i hM)y_0 + h\sum_{j\in[s]} a_{ij}\exp\left((c_i - c_j)hM\right) S\left(Y_j^{(k-1)}\right)\nabla V\left(Y_j^{(k)}\right) \quad (i \in [s]) \tag{11}$$

to obtain $Y_i^{(k)}$*'s. If some criteria hold, go to Step 2. Otherwise, set* $k = k + 1$ *and repeat Step 1.*
Step 2 Output

$$y_1^{(k)} = \exp(hM)y_0 + h\sum_{j\in[s]} b_j\exp\left((1 - c_j)hM\right) S\left(Y_j^{(k-1)}\right)\nabla V\left(Y_j^{(k)}\right).$$

Similar to Theorem 5, the discrete conservation law for the scheme above can be proved.

Theorem 6 *Suppose that* V *is a quadratic invariant of the semilinear ODE* (4)*. Assume that* V *is also an invariant of the linear part* $\dot{y} = My$*. In addition, we assume* (A, b) *satisfies* (2)*. Then, the solution* $y_1^{(k)}$ *of the scheme defined by* Definition 2 *satisfies* $V\left(y_1^{(k)}\right) = V(y_0)$.

Moreover, Theorem 3 implies the following theorem showing the accuracy of the scheme defined by Definition 2. Although the proof of Theorem 7 is similar to that for Theorem 3, we present the complete proof here for the reader's convenience.

Theorem 7 *Assume the following conditions:*

(A1) $Y_i^{(0)}$ *satisfies* $\left\|Y_i^{(0)} - y(c_i h)\right\|_{\mathcal{X}} \le Ch^q$ *for each* $i \in [s]$.
(A2) $S\colon \mathcal{X} \to \mathcal{L}(\mathcal{X})$ *is* L_S*-Lipschitz continuous.*
(A3) The base Runge–Kutta method is of order p.

Then, the numerical solution $y_1^{(k)}$ *of the scheme defined by* Definition 2 *satisfies*

$$\left\|y_1^{(k)} - y(h)\right\|_{\mathcal{X}} \le C' h^{\min\{p, q+k-1\}+1}$$

for sufficiently small $h > 0$*, i.e., the scheme with* k *iteration is of order* $\min\{p, q+k-1\}$*. Here,* C' *is a constant depending only on the exact solution* y*,* k*,* M*,* S*,* V*,* A*,* b*, and the constant* C *in (A1).*

Proof Due to the assumption (A1), there exists a function $y^{(0)}\colon [0,h] \to \mathcal{X}$ satisfying $y^{(0)}(c_i h) = Y_i^{(0)}$ and $\sup_{t\in[0,h]} \left\| y^{(0)}(t) - y(t) \right\|_{\mathcal{X}} \le Ch^q$. Then, the scheme defined by Definition 2 can be regarded as the usual Runge–Kutta method corresponding to A, b for the system

$$\begin{cases} \dot{w}^{(1)}(t) = \exp(-tM) S\left(\exp(tM) w^{(0)}(t)\right) \left(\exp(-tM)\right)^* \nabla V(w^{(1)}(t)), \\ \dot{w}^{(2)}(t) = \exp(-tM) S\left(\exp(tM) w^{(1)}(t)\right) \left(\exp(-tM)\right)^* \nabla V(w^{(2)}(t)), \\ \quad \vdots \\ \dot{w}^{(k)}(t) = \exp(-tM) S\left(\exp(tM) w^{(k-1)}(t)\right) \left(\exp(-tM)\right)^* \nabla V(w^{(k)}(t)), \end{cases}$$

where $w^{(0)}(t) = \exp(-tM) y^{(0)}(t)$. Otherwise expressed, the scheme can be regarded as the usual Lawson method for the system

$$\begin{cases} \dot{y}^{(1)}(t) = My^{(1)} + S\left(y^{(0)}(t)\right) \nabla V(y^{(1)}(t)), \\ \dot{y}^{(2)}(t) = My^{(2)} + S\left(y^{(1)}(t)\right) \nabla V(y^{(2)}(t)), \\ \quad \vdots \\ \dot{y}^{(k)}(t) = My^{(k)} + S\left(y^{(k-1)}(t)\right) \nabla V(y^{(k)}(t)), \end{cases}$$

where $y^{(i)}(t) = \exp(tM) w^{(i)}(t)$ for $i = 1, \ldots, k$. Therefore, it is sufficient to prove $\left\| y^{(k)}(h) - y(h) \right\|_{\mathcal{X}} \le C^{(k)} h^{q+k}$.

To this end, we prove

$$\sup_{t\in[0,h]} \left\| y^{(j)}(t) - y(t) \right\|_{\mathcal{X}} \le C^{(j)} h^{q+j} \qquad (j = 1, \ldots, k)$$

by induction, where $C^{(j)} := \left(2L_S C_y\right)^j C$ ($C_y := \sup_{t\in[0,h]} \left\| \nabla V(y(t)) \right\|_{\mathcal{X}}$). We assume that $\sup_{t\in[0,h]} \left\| y^{(j-1)}(t) - y(t) \right\|_{\mathcal{X}} \le C^{(j-1)} h^{q+j-1}$ holds so that

$$\begin{aligned} C_S^{(j-1)} &:= \sup_{t\in[0,h]} \left\| S\left(y^{(j-1)}(t)\right) \right\|_{\mathcal{X}} \\ &\le \left\| S(y_0) \right\|_{\mathcal{X}} + L_S \left(C^{(j-1)} h^{q+j-1} + \sup_{t\in[0,h]} \left\| y(t) - y_0 \right\|_{\mathcal{X}} \right) \\ &< \infty, \end{aligned}$$

where $\|S(y_0)\|_{\mathcal{X}}$ denotes the operator norm of $S(y_0)$ with respect to the inner product of $\mathcal{X}$. Then, we see

$$\begin{aligned}
&\sup_{t\in[0,h]} \left\| y^{(j)}(t) - y(t) \right\|_{\mathcal{X}} \\
&= \sup_{t\in[0,h]} \left\| \int_0^t \left(My^{(j)}(r) + S\left(y^{(j-1)}(r)\right)\nabla V\left(y^{(j)}(r)\right) - My(r) - S(y(r))\nabla V\left(y(r)\right)\right) \mathrm{d}r \right\|_{\mathcal{X}} \\
&\le h\|M\|_{\mathcal{X}} \sup_{t\in[0,h]} \left\| y^{(j)}(t) - y(t) \right\|_{\mathcal{X}} \\
&\quad + \sup_{t\in[0,h]} \left\| \int_0^t \left(S\left(y^{(j-1)}(r)\right)\nabla V\left(y^{(j)}(r)\right) - S\left(y^{(j-1)}(r)\right)\nabla V\left(y(r)\right)\right) \mathrm{d}r \right\|_{\mathcal{X}} \\
&\quad + \sup_{t\in[0,h]} \left\| \int_0^t \left(S\left(y^{(j-1)}(r)\right)\nabla V\left(y(r)\right) - S(y(r))\nabla V\left(y(r)\right)\right) \mathrm{d}r \right\|_{\mathcal{X}} \\
&\le h\left(\|M\|_{\mathcal{X}} + C_S^{(j-1)}\|Q\|_{\mathcal{X}}\right) \sup_{t\in[0,h]} \left\| y^{(j)}(t) - y(t) \right\|_{\mathcal{X}} + L_S C^{(j-1)} C_y h^{q+j}.
\end{aligned}$$

Since we assume that h is sufficiently small, in particular, $h \le (2\|M\|_{\mathcal{X}} + 2M_S^{(j-1)}\|Q\|_{\mathcal{X}})^{-1}$ holds, we see

$$\sup_{t\in[0,h]} \left\| y^{(j)}(t) - y(t) \right\|_{\mathcal{X}} \le 2L_S C^{(j-1)} C_y h^{q+j}.$$

Thus, by induction, we obtain $\sup_{t\in[0,h]} \left\| y^{(j)}(t) - y(t) \right\|_{\mathcal{X}} \le C^{(j)} h^{q+j}$ $(j = 1, \ldots, k)$. Since it implies $\left\| y^{(k)}(h) - y(h) \right\|_{\mathcal{X}} \le C^{(k)} h^{q+k}$, the theorem holds. □

Remark 3 *In the above and subsequent theorems, we assume the global Lipschitz continuity of S for the sake of brevity. However, this can be relaxed to the local Lipschitz continuity.*

Remark 4 *As can be seen from the proof above, even if $\mathcal{X}$ is infinite-dimensional, the same theorem holds if $\|S(y_0)\|_{\mathcal{X}}$, $\|M\|_{\mathcal{X}}$, and $\|Q\|_{\mathcal{X}}$ are bounded.*

4 Application to the SAV approach

The scheme defined in Definition 2 generally requires solving linear equations that involve matrix exponential functions, which is computationally somewhat more expensive than the usual linearly implicit schemes from [16]. However, combining it to the SAV system (9) produces a scheme that is computationally extremely inexpensive. Specifically, the main part of each iteration is the product of a matrix exponential function and a vector $O(s)$ times, which can be computed in parallel. The computational cost of the scheme can be the same level as that of explicit exponential integrators (see Appendix 3). When an efficient implementation of the matrix exponential function is available, the proposed scheme in this section overwhelms the linearly implicit scheme defined in Definition 1 in terms of computational efficiency (see Fig. 2).

4.1 Simple case (9)

The SAV system (9) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u \\ r \end{bmatrix} = \begin{bmatrix} J & J\phi(u) \\ -(J\phi(u))^* & \langle \phi(u), J\phi(u)\rangle \end{bmatrix}\begin{bmatrix} Lu \\ 2r \end{bmatrix}.$$

Recall that $J \in \mathcal{L}(\mathcal{V})$, $L \in \mathcal{L}(\mathcal{V})$, $\phi\colon \mathcal{V} \to \mathcal{V}$ is a function defined as $\phi(u) = \nabla E(u)/\left(2\sqrt{E(u)+\alpha}\right)$, and this system has the quadratic invariant $V(u,r) = \frac{1}{2}\langle Lu, u\rangle + r^2 - \alpha$. Since the skew-symmetry of J implies $\langle \phi(u), J\phi(u)\rangle = 0$, it can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u \\ r \end{bmatrix} = \begin{bmatrix} JL & \\ & 0 \end{bmatrix}\begin{bmatrix} u \\ r \end{bmatrix} + \begin{bmatrix} & J\phi(u) \\ -(J\phi(u))^* & \end{bmatrix}\nabla V(u,r), \tag{12}$$

which is a special case of (10) with

$$y = \begin{bmatrix} u \\ r \end{bmatrix}, \quad M = \begin{bmatrix} JL & \\ & 0 \end{bmatrix}, \quad S(u) = \begin{bmatrix} & J\phi(u) \\ -(J\phi(u))^* & \end{bmatrix}, \quad Q = \begin{bmatrix} L & \\ & 2 \end{bmatrix}.$$

Note that the linear part $\dot{y} = My$ preserves V:

$$\frac{\mathrm{d}}{\mathrm{d}t}V(y(t)) = \langle My, \dot{y}\rangle_{\mathcal{X}} = \langle Lu, JLu\rangle = 0.$$

Therefore, we can apply the scheme defined by Definition 2 for (12). In this case, the linear equation system (11) reads

$$\begin{aligned}\begin{bmatrix} U_i^{(k)} \\ R_i^{(k)} \end{bmatrix} &= \begin{bmatrix} \exp(c_i h J L) & \\ & 1 \end{bmatrix}\begin{bmatrix} u_0 \\ r_0 \end{bmatrix} \\ &\quad + h\sum_{j\in[s]} a_{ij}\begin{bmatrix} \exp\left((c_i - c_j)hJL\right) & \\ & 1 \end{bmatrix}\begin{bmatrix} & J\phi\left(U_j^{(k-1)}\right) \\ -\left(J\phi\left(U_j^{(k-1)}\right)\right)^* & \end{bmatrix}\begin{bmatrix} LU_j^{(k)} \\ 2R_j^{(k)} \end{bmatrix},\end{aligned}$$

where $U_i^{(k)} \in \mathcal{V}$ and $R_i^{(k)} \in \mathbb{R}$ are the inner stages with respect to u and r, respectively. This linear equation, expressed element by element, is as follows:

$$\begin{aligned} U_i^{(k)} &= \exp(c_i h J L)\, u_0 + 2h\sum_{j\in[s]} a_{ij} R_j^{(k)} \exp\left((c_i - c_j)hJL\right) J\phi\left(U_j^{(k-1)}\right), \\ R_i^{(k)} &= r_0 - h\sum_{j\in[s]} a_{ij}\left\langle J\phi\left(U_j^{(k-1)}\right), LU_j^{(k)}\right\rangle. \end{aligned}$$

By substituting the first equation into the second equation, we obtain a linear equation

$$R_i^{(k)} = r_0 - h\sum_{\ell\in[s]} a_{i\ell}\left\langle J\phi\left(U_\ell^{(k-1)}\right), L\exp\left(c_\ell hJL\right)u_0\right\rangle - 2h^2\sum_{j\in[s]}\left(\sum_{\ell\in[s]} a_{i\ell}a_{\ell j}\left\langle J\phi\left(U_\ell^{(k-1)}\right), L\exp\left((c_\ell - c_j)hJL\right)J\phi\left(U_j^{(k-1)}\right)\right\rangle\right) R_j^{(k)}$$

only on $\{R_i^{(k)}\}_{i=1}^s$. To reduce the number of computations of the matrix exponential function, we use Lemma 4, i.e., the relation $L\exp(tJL) = (\exp(-tJL))^* L$, and obtain

$$R_i^{(k)} = r_0 + h\sum_{\ell\in[s]} a_{i\ell}\left\langle\psi_\ell^{(k-1)}, Lu_0\right\rangle - 2h^2\sum_{j\in[s]}\left(\sum_{\ell\in[s]} a_{i\ell}a_{\ell j}\left\langle\psi_\ell^{(k-1)}, L\psi_j^{(k-1)}\right\rangle\right) R_j^{(k)},$$

where $\psi_j^{(k-1)} := \exp\left(-c_j hJL\right)J\phi\left(U_j^{(k-1)}\right)$. By introducing $R^{(k)} = \left[R_1^{(k)}\ \ldots\ R_s^{(k)}\right]^{\mathrm{T}}$, the linear equation can be simplified to

$$\left(I_s + 2h^2 A(A\circ\Psi)\right)R^{(k)} = r_0\mathbf{1}_s - hA\nu, \tag{13}$$

where $I_s \in \mathbb{R}^{s\times s}$ denotes the identity matrix, $A = (a_{ij}) \in \mathbb{R}^{s\times s}$, $\circ$ denotes the Hadamard product, $\Psi \in \mathbb{R}^{s\times s}$ is defined by $\Psi_{i,j} = \left\langle\psi_i^{(k-1)}, L\psi_j^{(k-1)}\right\rangle$, $\mathbf{1}_s \in \mathbb{R}^s$ denotes all one vector of size s, and $\nu \in \mathbb{R}^s$ is defined by $\nu_i = \left\langle\psi_i^{(k-1)}, Lu_0\right\rangle$.

Another note on the implementation should address the fact that $r_1^{(k)}$ can be computed without $\left\{U_i^{(k)}\right\}_{i=1}^s$:

$$r_1^{(k)} = r_0 + h\sum_{i\in[s]} b_i\left\langle J\phi\left(U_i^{(k-1)}\right), LU_i^{(k)}\right\rangle = r_0 + \sum_{i,j\in[s]} b_i\omega_{ij}\left(R_j^{(k)} - r_0\right),$$

where ω_{ij} denotes the (i,j) element of the inverse matrix of A. Consequently, we propose the following scheme.

Definition 3

Step 0 Prepare $U_i^{(0)} \approx u(c_i h)$ *and set* $k = 1$.
Step 1 Compute $\{\psi_i^{(k-1)}\}_{i=1}^s$ *by*

$$\psi_i^{(k-1)} = \exp\left(-c_i hJL\right)J\phi\left(U_i^{(k-1)}\right). \tag{14}$$

Solve the linear equation (13) to obtain $R^{(k)}$. If some criteria hold, go to Step 2. Otherwise, compute $\{U_i^{(k)}\}_{i=1}^s$ by

$$U_i^{(k)} = \exp(c_i h J L)\left(u_0 + 2h \sum_{j\in[s]} a_{ij} R_j^{(k)} \psi_j^{(k-1)}\right), \tag{15}$$

set $k = k+1$ and repeat Step 1.

Step 2 Output

$$u_1^{(k)} = \exp(hJL)\left(u_0 + 2h \sum_{j\in[s]} b_j R_j^{(k)} \psi_j^{(k-1)}\right),$$
$$r_1^{(k)} = r_0 + \sum_{i,j\in[s]} b_i \omega_{ij}\left(R_j^{(k)} - r_0\right).$$

Note that, in many cases, the computation of (14) and (15) is the most computationally expensive part. However, efficient algorithms for computing the matrix exponential and vector products are known. Moreover, since the computation of $\psi_1^{(k-1)}, \psi_2^{(k-1)}, \ldots, \psi_s^{(k-1)}$ by using (14) can be done independently, they can be trivially parallelized. Similarly, the computation of $U_i^{(k)}$'s by (15) can be parallelized. See Appendix 3 for more details on the computational cost.

The scheme defined by Definition 3 has a unique solution if and only if the linear equation (13) has a unique solution, which is satisfied for small step size h. The step size restriction depends largely on the nature of J and L (see Remark 5).

Theorem 8 *Suppose that the step size h satisfies*

$$h \max\left\{1, \exp\left(\frac{c_{\max} h}{2} \lambda_{\max}(LJ - JL)\right)\right\} < \frac{1}{\|A\|_2 \|J\| \sqrt{2\|L\|} \sup_i \left\|\phi\left(U_i^{(k-1)}\right)\right\|}, \tag{16}$$

where $\lambda_{\max}(LJ - JL)$ denotes the maximum eigenvalue of $LJ - JL$, $c_{\max} = \max_i c_i$, and $\|A\|_2$ denotes the spectral norm of $A \in \mathbb{R}^{s\times s}$. Then, the linear (13) has a unique solution.

Proof The linear (13) has a unique solution if and only if the matrix $I_s + 2h^2 A\,(A \circ \Psi)$ is nonsingular. Then, $\left\|2h^2 A\,(A \circ \Psi)\right\|_2 < 1$ is a sufficient condition for the solution's unique existence.

Since we have $\|A \circ \Psi\|_2 \leq \|A\|_2 \|\Psi\|_2$ (cf. [31]), we evaluate the norm $\|\Psi\|_2$:

$$\begin{aligned}
\|\Psi\|_2 &= \sup_{x \in \mathbb{R}^s} \frac{\sum_{i,j \in [s]} x_i x_j \left\langle \psi_i^{(k-1)}, L\psi_j^{(k-1)} \right\rangle}{\|x\|_2^2} \\
&\leq \|L\| \sup_{x \in \mathbb{R}^s} \frac{\sum_{i,j \in [s]} x_i x_j \left\| \psi_i^{(k-1)} \right\| \left\| \psi_j^{(k-1)} \right\|}{\|x\|_2^2} \\
&= \|L\| \max_i \left\| \psi_i^{(k-1)} \right\|^2 \\
&= \|L\| \max_i \left\| \exp(-c_i h J L) J \phi \left(U_i^{(k-1)} \right) \right\|^2 \\
&\leq \|L\| \|J\|^2 \left(\sup_{t \in [0, c_{\max} h]} \left\| \exp\left(-tJL\right) \right\| \right)^2 \max_i \left\| \phi \left(U_i^{(k-1)} \right) \right\|^2.
\end{aligned}$$

Note that $\|\exp(t(-JL))\| \leq \exp(t\omega(-JL))$ holds for any $t \geq 0$ (cf. [32]), where $\omega(-JL)$ denotes the numerical abscissa of the matrix $-JL$. Since $\omega(-JL) = \lambda_{\max}\left(\frac{1}{2}(-JL + (-JL)^*)\right) = \lambda_{\max}\left(\frac{1}{2}(LJ - JL)\right)$ holds, we have

$$\sup_{t \in [0, c_{\max} h]} \|\exp(-tJL)\|^2 \leq \max\{1, \exp(c_{\max} h \lambda_{\max}(LJ - JL))\},$$

which implies the theorem. □

Remark 5 *The step size restriction* (16) *may seem severe. Of course, this is true when* $\lambda_{\max}(LJ - JL)$ *is positive, and the above argument only shows the unique existence for very small step sizes. However, when* L *and* J *commute, we have* $\lambda_{\max}(LJ - JL) = 0$*, and the step size restriction is rather mild. The commutativity of* L *and* J *is sometimes assumed in the literature* (cf. [33], see also Sect. 5)*.*

Theorem 6 implies the following theorem showing the discrete conservation law with respect to the modified invariant V.

Theorem 9 (Conservation law) *The solution* $\left(u_1^{(k)}, r_1^{(k)}\right)$ *of the scheme defined by* Definition 3 *satisfies* $V\left(u_1^{(k)}, r_1^{(k)}\right) = V(u_0, r_0)$.

Theorem 7 implies the following theorem showing the accuracy of the scheme defined by Definition 3.

Theorem 10 (Accuracy) *Assume the following conditions:*

(A1) $U_i^{(0)}$ *satisfies* $\left\|U_i^{(0)} - u(c_i h)\right\| \leq Ch^q$ *for each* $i \in [s]$.
(A2) $\phi\colon \mathcal{V} \to \mathcal{V}$ *is Lipschitz continuous.*
(A3) The base Runge–Kutta method is of order p.

Then, the numerical solution $\left(u_1^{(k)}, r_1^{(k)}\right)$ *of the scheme defined by Definition 3 satisfies*

$$\left\|u_1^{(k)} - u(h)\right\| \le C' h^{\min\{p,q+k-1\}+1},$$

$$\left|r_1^{(k)} - r(h)\right| \le C' h^{\min\{p,q+k-1\}+1}$$

for sufficiently small $h > 0$*, i.e., the scheme with* k *iteration is of order* $\min\{p, q+k-1\}$*. Here,* C' *is a constant depending only on the exact solution* u*,* k*,* J*,* L*,* ϕ*,* A*,* b*, and the constant* C *in (A1).*

Proof It is sufficient to prove

$$\left\|\begin{bmatrix} & J\phi(u_1) \\ -(J\phi(u_1))^* & \end{bmatrix} - \begin{bmatrix} & J\phi(u_2) \\ -(J\phi(u_2))^* & \end{bmatrix}\right\|_{\mathcal{X}} = \|J\phi(u_1) - J\phi(u_2)\|. \tag{17}$$

In general, for any $v \in \mathcal{V}$, we see

$$\left\|\begin{bmatrix} & v \\ -(v)^* & \end{bmatrix}\right\|_{\mathcal{X}} = \sup_{w\in\mathcal{V},\rho\in\mathbb{R}} \frac{\sqrt{\rho^2\|v\|^2 + (\langle v, w\rangle)^2}}{\sqrt{\|w\|^2+\rho^2}} = \|v\| \sup_{w\in\mathcal{V},\rho\in\mathbb{R}} \frac{\sqrt{\rho^2+\|w\|^2}}{\sqrt{\|w\|^2+\rho^2}} = \|v\|,$$

which proves (17). □

Remark 6 *The scheme defined by* Definition 3 *includes the schemes proposed in* [27] *as special cases. In* [27]*, they focus on the cases* $U_i^{(0)}$ *computed by the extrapolation or* $U_i^{(0)} = u_0$*. In addition, the intended implementation is also somewhat different: for example, their algorithm requires* $ks(s-1)$ *computations of the product of the matrix exponential functions and vectors, while it is reduced to* $(2k-1)s$ *in* Definition 3 *by introducing* $\psi^{(k-1)}$*.*

Moreover, the consequences of Theorem 10 *are consistent with the accuracy confirmed numerically in* [27]*. For example, from numerical experiments in* [27, Remark 3.6]*, the authors predict that the scheme has fourth-order accuracy when the base RK method is a fourth-order Gauss method,* $U_i^{(0)} = u_0$ *(i.e.,* $q = 1$*), and* $k = 4$*.*

4.2 Multiple scalar auxiliary variables

As shown in [15], when E is unbounded, two scalar auxiliary variables are needed. The scheme described in the previous section can be extended to this case.

Let us consider the Hamiltonian system (7) with

$$H(u) = \frac{1}{2}\langle Lu, u\rangle + E_{\mathrm{L}}(u) - E_{\mathrm{U}}(u),$$

where $E_{\mathrm{X}}\colon \mathcal{V} \to \mathbb{R}$ are bounded from below, i.e., $\alpha_{\mathrm{X}} := -\inf E_{\mathrm{X}}(u) + \epsilon_{\mathrm{X}}$ is finite ($\mathrm{X} \in \{\mathrm{L}, \mathrm{U}\}$). Then, by introducing $r_{\mathrm{X}} := \sqrt{E_{\mathrm{X}}(u) + \alpha_{\mathrm{X}}}$, $\phi_{\mathrm{X}}(u) := \nabla E_{\mathrm{X}}(u) / \left(2\sqrt{E_{\mathrm{X}}(u) + \alpha_{\mathrm{X}}}\right)$, and $V(u, r_{\mathrm{L}}, r_{\mathrm{U}}) = \frac{1}{2}\langle Lu, u\rangle + r_{\mathrm{L}}^2 - r_{\mathrm{U}}^2$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u \\ r_{\mathrm{L}} \\ r_{\mathrm{U}} \end{bmatrix} = \begin{bmatrix} JL & & \\ & 0 & \\ & & 0 \end{bmatrix}\begin{bmatrix} u \\ r_{\mathrm{L}} \\ r_{\mathrm{U}} \end{bmatrix} + \begin{bmatrix} & J\phi_{\mathrm{L}} & J\phi_{\mathrm{U}} \\ -(J\phi_{\mathrm{L}})^* & 0 & \langle \phi_{\mathrm{L}}, J\phi_{\mathrm{U}}\rangle \\ -(J\phi_{\mathrm{U}})^* & -\langle \phi_{\mathrm{L}}, J\phi_{\mathrm{U}}\rangle & 0 \end{bmatrix} \nabla V(u, r_{\mathrm{L}}, r_{\mathrm{U}}), \tag{18}$$

which is a special case of (10) ($\phi_{\mathrm{X}}(u)$ is abbreviated to ϕ_{X}).

Therefore, we can apply the scheme defined by Definition 2. In addition, by introducing $\psi_i^{\mathrm{X}} := \exp(-c_i h J L)\, J\phi_{\mathrm{X}}\left(U_i^{(k-1)}\right)$, the techniques to reduce the computational cost in Sect. 4.1 can also be used in this case (see Appendix 1 for details). Then, the linear equation system with respect to $R_{\mathrm{L}}^{(k)} = \left[R_{\mathrm{L},1}^{(k)} \ \dots \ R_{\mathrm{L},s}^{(k)}\right]^{\mathrm{T}}$ and $R_{\mathrm{U}}^{(k)} = \left[R_{\mathrm{U},1}^{(k)} \ \dots \ R_{\mathrm{U},s}^{(k)}\right]^{\mathrm{T}}$ can be written as

$$\begin{bmatrix} I_s + 2h^2 A\left(A \circ \Psi^{\mathrm{LL}}\right) & 2hA\Phi - 2h^2 A\left(A \circ \Psi^{\mathrm{LU}}\right) \\ 2hA\Phi + 2h^2 A\left(A \circ \Psi^{\mathrm{UL}}\right) & I_s - 2h^2 A\left(A \circ \Psi^{\mathrm{UU}}\right) \end{bmatrix}\begin{bmatrix} R_{\mathrm{L}}^{(k)} \\ R_{\mathrm{U}}^{(k)} \end{bmatrix} = \begin{bmatrix} r_{\mathrm{L},0}\mathbf{1}_s - hA\nu^{\mathrm{L}} \\ r_{\mathrm{U},0}\mathbf{1}_s - hA\nu^{\mathrm{U}} \end{bmatrix} \tag{19}$$

where $\Psi^{\mathrm{LL}}, \Psi^{\mathrm{LU}}, \Psi^{\mathrm{UL}}, \Psi^{\mathrm{UU}} \in \mathbb{R}^{s\times s}$ are defined by $\Psi_{i,j}^{\mathrm{XY}} = \left\langle \psi_i^{\mathrm{X}}, L\psi_j^{\mathrm{Y}}\right\rangle$ for $\mathrm{X}, \mathrm{Y} \in \{\mathrm{L}, \mathrm{U}\}$ (note that $\Psi^{\mathrm{LU}} = \left(\Psi^{\mathrm{LU}}\right)^{\mathrm{T}}$ holds), $\Phi \in \mathbb{R}^{s\times s}$ is a diagonal matrix defined by $\Phi_{ii} = \left\langle \phi_{\mathrm{L}}\left(U_i^{(k-1)}\right), J\phi_{\mathrm{U}}\left(U_i^{(k-1)}\right)\right\rangle$, and $\nu^{\mathrm{L}}, \nu^{\mathrm{U}} \in \mathbb{R}^s$ are defined by $\nu_i^{\mathrm{X}} = \left\langle \psi_i^{\mathrm{X}}, Lu_0\right\rangle$. Consequently, we obtain the following scheme:

Definition 4

Step 0 *Prepare* $U_i^{(0)} \approx u(c_i h)$ *and set* $k = 1$.
Step 1 *Compute* $\{\psi_i^{\mathrm{X}}\}_{i=1}^s$ *by*

$$\psi_i^{\mathrm{X}} = \exp(-c_i h J L)\, J\phi_{\mathrm{X}}\left(U_i^{(k-1)}\right).$$

Solve the linear equation system (19) *to obtain* $R_{\mathrm{L}}^{(k)}$ *and* $R_{\mathrm{U}}^{(k)}$. *If some criteria hold, go to Step 2. Otherwise, compute* $\{U_i^{(k)}\}_{i=1}^s$ *by*

$$U_i^{(k)} = \exp(c_i h J L)\left(u_0 + 2h\sum_{j\in[s]} a_{ij}\left(R_{\mathrm{L},j}^{(k)}\psi_j^{\mathrm{L}} - R_{\mathrm{U},j}^{(k)}\psi_j^{\mathrm{U}}\right)\right),$$

set $k = k + 1$ *and repeat Step 1.*

Step 2 Output

$$u_1^{(k)} = \exp(hJL)\left(u_0 + 2h\sum_{j\in[s]} b_j\left(R_{\mathrm{L},j}^{(k)}\psi_j^{\mathrm{L}} - R_{\mathrm{U},j}^{(k)}\psi_j^{\mathrm{U}}\right)\right),$$
$$r_{\mathrm{X},1}^{(k)} = r_{\mathrm{X},0} + \sum_{i,j\in[s]} b_i\omega_{ij}\left(R_{\mathrm{X},j}^{(k)} - r_0\right).$$

The scheme defined by Definition 4 has properties similar to the scheme defined by Definition 3. Here, we only provide the results, and the proof is given in Appendix 2. In particular, for the uniqueness and existence, only the case $LJ = JL$ is presented here, while the general case is presented in Theorem 14.

Theorem 11 *Suppose that $LJ = JL$ holds. If h satisfies*

$$h < \frac{\sqrt{1+\frac{4\|L\|}{C_\phi}}-1}{4\|A\|_2\|J\|\|L\|}, \qquad C_\phi := \max_{i\in[s],\,\mathrm{X}\in\{\mathrm{L},\mathrm{U}\}}\left\|\phi_{\mathrm{X}}\left(U_i^{(k-1)}\right)\right\|^2,$$

the linear equation (19) *has a unique solution.*

Theorem 12 (Conservation law) *The solution $\left(u_1^{(k)}, r_{\mathrm{L},1}^{(k)}, r_{\mathrm{U},1}^{(k)}\right)$ of the scheme defined by* Definition 4 *satisfies $V\left(u_1^{(k)}, r_{\mathrm{L},1}^{(k)}, r_{\mathrm{U},1}^{(k)}\right) = V(u_0, r_{\mathrm{L},0}, r_{\mathrm{U},0})$.*

Theorem 13 (Accuracy) *Assume the following conditions:*

(A1) $U_i^{(0)}$ satisfies $\left\|U_i^{(0)} - u(c_i h)\right\| \le Ch^q$ for each $i \in [s]$.
(A2) $\phi_{\mathrm{X}}\colon \mathcal{V} \to \mathcal{V}$ is Lipschitz continuous and bounded.
(A3) The base Runge–Kutta method is of order p.

Then, the numerical solution $\left(u_1^{(k)}, r_{\mathrm{L},1}^{(k)}, r_{\mathrm{U},1}^{(k)}\right)$ of the scheme defined by Definition 4 *satisfies*

$$\left\|u_1^{(k)} - u(h)\right\| \le C'h^{\min\{p,q+k-1\}+1},$$
$$\left|r_{\mathrm{L},1}^{(k)} - r_{\mathrm{L}}(h)\right| \le C'h^{\min\{p,q+k-1\}+1},$$
$$\left|r_{\mathrm{U},1}^{(k)} - r_{\mathrm{U}}(h)\right| \le C'h^{\min\{p,q+k-1\}+1}$$

for a sufficiently small $h > 0$, i.e., the scheme with k iteration is of order $\min\{p, q + k - 1\}$. Here, C' is a constant depending only on the exact solution u, k, J, L, ϕ_{X}, A, b, and the constant C in (A1).

5 Numerical experiments

We employ the sixth-order Gauss method as the base RK method. The initial approximation $U_i^{(0)} \approx u(c_i h)$ is computed by the Lawson transformation and the continuous explicit RK (CERK) method with order $1, 2, \ldots, 5$ ($q = 2, 3, \ldots, 6$). All numerical experiments are performed in Julia (Version 1.8.0) on a PC with Apple M1 Ultra and 128GB RAM. The numerical experiments in this paper are not parallelized. We leave it to future work.

5.1 Modified Korteweg–de Vries equation

Let us consider the modified Korteweg–de Vries (mKdV) equation (on $\mathbb{S} := \mathbb{R}/L\mathbb{Z}$):

$$u_t = -\partial_x \frac{\delta \mathcal{H}}{\delta u}, \qquad \mathcal{H}(u) = \int \left(-\frac{1}{2}(u_x)^2 + \frac{1}{2}u^4 \right) \mathrm{d}x.$$

We employ an energy-preserving spatial discretization

$$\dot{u}_k = -\delta_x \left(\delta_x^2 u_k + 2(u_k)^3 \right), \quad H(u) = -\frac{1}{2} \sum_{k=1}^{N} (\delta_x u_k)^2 \, \Delta x + \frac{1}{2} \sum_{k=1}^{N} (u_k)^4 \Delta x, \quad (20)$$

where δ_x denotes the Fourier-spectral difference operator (see, e.g., [34] for details on the difference operator). The inner product is defined as $\langle v, w \rangle = \sum_{k=1}^{N} v_k w_k \Delta x$. Then, $\nabla H(u) = \delta_x^2 u_k + 2(u_k)^3$ holds.

Then, $E(u) := \frac{1}{2} \sum_{k=1}^{N} (u_k)^4 \Delta x$ is bounded from below so that we can use the scheme defined by Definition 3. Note that the product of the matrix exponential function $e^{t\delta_x^3}$ and a vector can be computed by the FFT. In numerical experiments, we use the exact solution $\mathrm{dn}(x - (2 - m)t \mid m)$ with the spatial period $L = 2K(m)$ and temporal period $T = L/|2 - m|$, where dn is one of the Jacobi elliptic functions and K denotes the complete elliptic integral of the first kind. We choose the parameters $m = 0.1$ ($L \approx 3.225$, $T \approx 1.697$) and $N = 16$.

Figure 1 summarizes the relative errors. They decrease along the reference lines drawn based on Theorem 10. Figure 2 compares the computational cost of the proposed scheme with the scheme defined by [16] (see Appendix 4 for details) and an explicit Lawson method (corresponds to a seven stage sixth-order explicit Runge–Kutta method [35]). Note that all schemes have sixth order. The proposed scheme is more efficient than the scheme defined by [16]. However, the proposed scheme is less efficient than the explicit Lawson method for such a short time interval. As shown in the next example, the proposed scheme is more efficient than the explicit Lawson method for a long time interval (this advantage is typical in the comparison between conservative and non-conservative schemes).

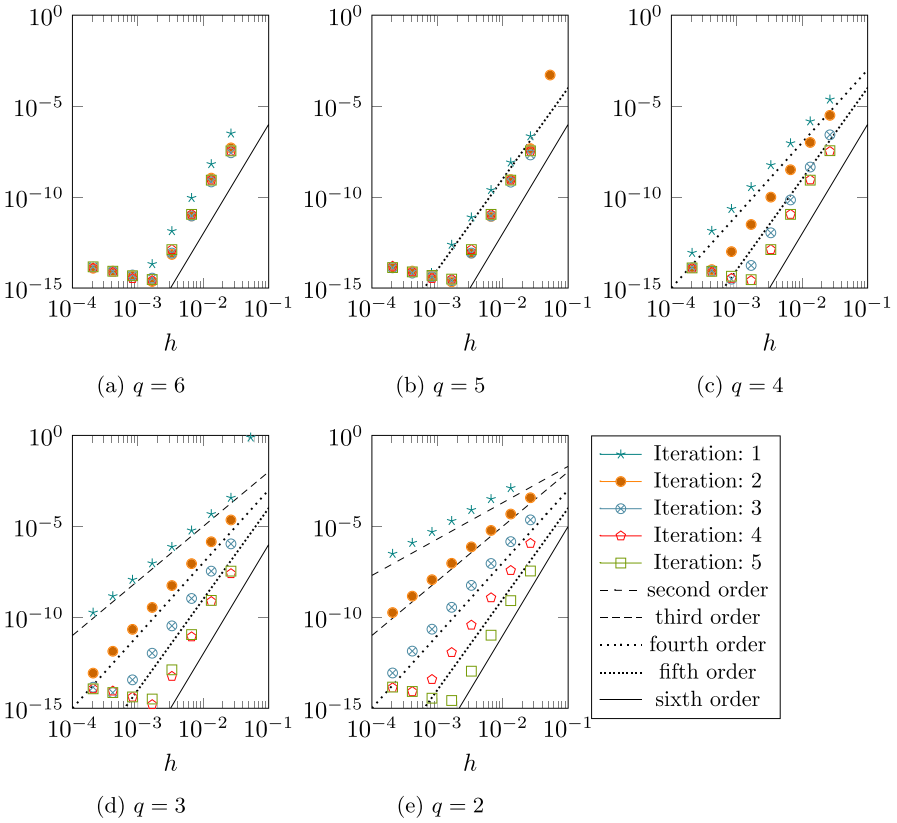


Fig. 1 Relative errors of the numerical solution for the mKdV equation

5.2 Korteweg–de Vries equation

Let us consider the Korteweg–de Vries (KdV) equation (on $\mathbb{S} := \mathbb{R}/L\mathbb{Z}$):

$$u_t = \partial_x \frac{\delta \mathcal{H}}{\delta u}, \qquad \mathcal{H}(u) = \int \left(\frac{1}{2}(u_x)^2 - u^3 \right) \mathrm{d}x.$$

We employ an energy-preserving spatial discretization

$$\dot{u}_k = \delta_x \left(-\delta_x^2 u_k - 3(u_k)^2 \right), \quad H(u) = \frac{1}{2} \sum_{k=1}^{N} (\delta_x u_k)^2 \, \Delta x + \sum_{k=1}^{N} (u_k)^3 \Delta x. \tag{21}$$

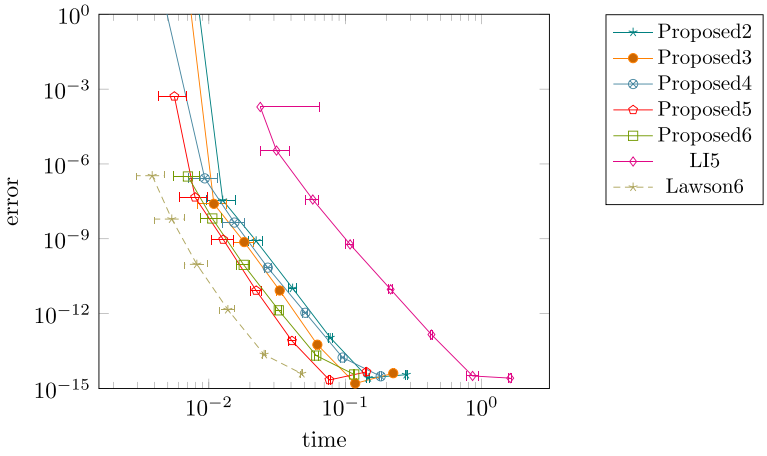


Fig. 2 Comparison of the computational efficiency of several schemes for the mKdV equation. In the legend, "proposed$i$" denotes the sixth-order scheme defined by Definition 3 with $q = i$ and $k = 6-i$, "LI5" denotes the sixth-order scheme defined by [16] (1) with $q = 5$ and $k = 2$, and "Lawson6" denotes a seven stage sixth-order explicit Lawson method. Note that, the proposed schemes and the scheme defined by [16] are applied to the SAV systems (12) and (9), respectively, and the explicit Lawson method is directly applied to the system (20). Each scheme is computed 30 times, and the mean values and the standard deviations of the computational times are shown

Since $\sum_{k=1}^{N} (u_k)^3 \Delta x$ is unbounded, we employ the scheme defined by Definition 4. We define

$$E_{\mathrm{L}}(u) := \sum_{k=1}^{N} \left((u_k)^4 - (u_k)^3 \right) \Delta x, \qquad E_{\mathrm{U}}(u) := \sum_{k=1}^{N} (u_k)^4 \, \Delta x$$

similarly to [15]. In the numerical experiments below, we use the exact solution $2m\,(\mathrm{cn}(x - ct \mid m))^2$, where $c = 4(2m - 1)$ and cn is a Jacobi elliptic function. We choose the parameters $m = 0.1$ ($L = 2K(m) \approx 3.225$, $T = L/|c| \approx 1.008$) and $N = 64$.

Figure 3 summarizes the relative errors. They decrease along the reference lines.

To confirm the long-term behavior of the proposed scheme, we conducted the numerical experiments over 32 periods. Figure 4 shows the evolution of relative errors of the invariant H, modified invariant V, and $I(u) := \frac{1}{2} \sum (u_k)^2 \, \Delta x$ corresponding to the another invariant $\mathcal{I}(u) := \int \left(\frac{1}{2}u^2\right) \mathrm{d}x$ of the KdV equation.

As for the modified invariant V, as expected, it is very well preserved. The relative errors for the original invariants H and I are also small.

There, the proposed scheme ($q = 6$ and $k = 1$) with $h = T/64 \approx 0.01575$ is applied to the SAV system (12), and the explicit Lawson method with $h = T/512 \approx 0.001968$ is directly applied to the system (21). We employ the small step size for the Lawson method because numerical solutions computed with $h = T/256$ diverged. Even in the step size employed in the figure, the behavior of the relative errors of H suggests that it is expected to diverge after a slightly long run. In the current setup,

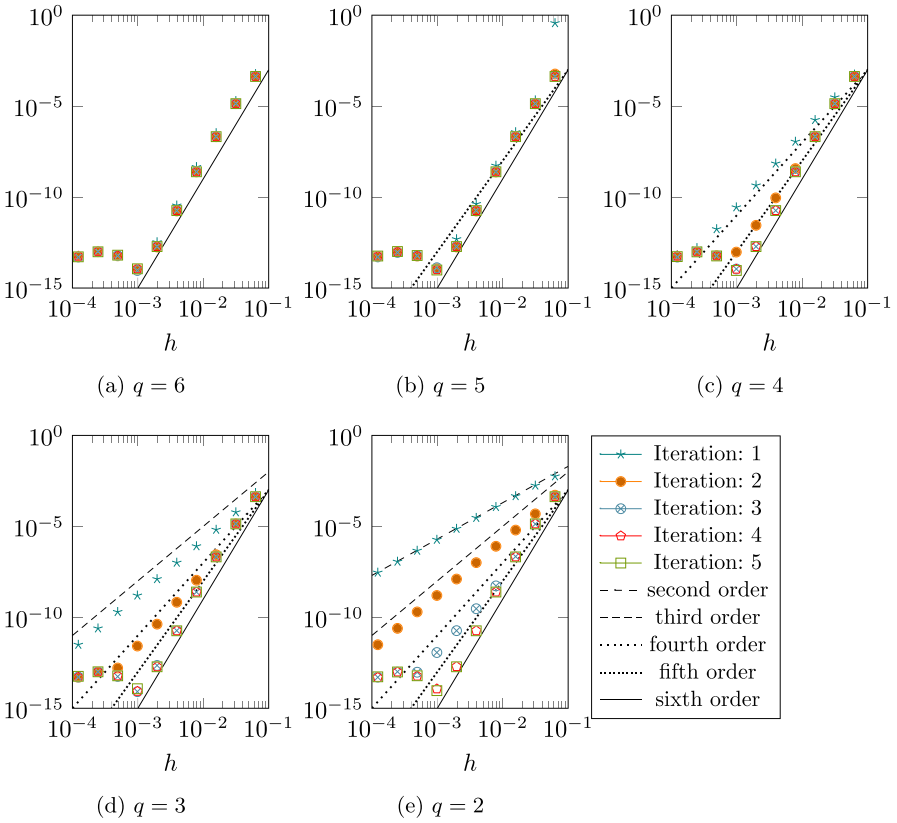


Fig. 3 Relative errors of the numerical solution for the KdV equation

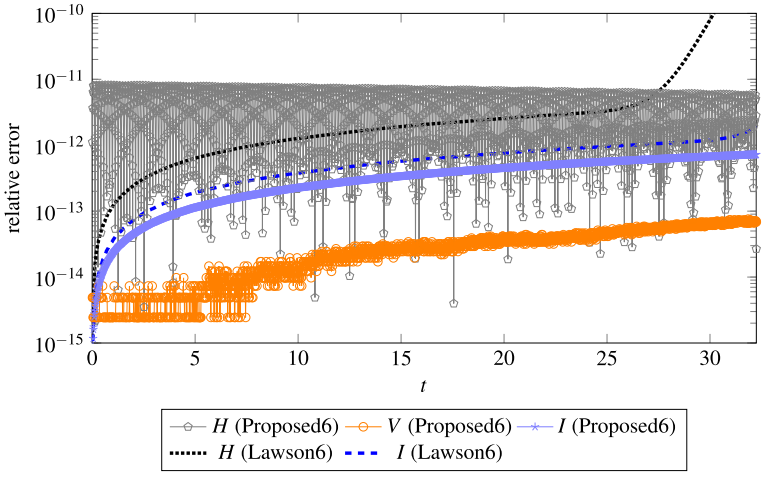


Fig. 4 Evolution of the relative errors of invariants of the KdV equation. In the figure, H, V, and I denote the invariant, the modified invariant, and the other invariant, respectively. The proposed scheme employed the step size $h = T/64 \approx 0.01575$, and the Lawson method employed the step size $h = T/512 \approx 0.001968$

the relative errors of numerical solutions itself at the final time are approximately 5.39×10^{-7} for the proposed scheme and 8.33×10^{-7} for the Lawson method. The computational time of the proposed scheme is 0.678 s, and that of the Lawson method is 0.882 s (they are mean values of 30 computations, with a standard deviation of about 0.01 for both methods).

5.3 Sine-Gordon equation

Let us consider the sine-Gordon (sG) equation (on $\mathbb{S} := \mathbb{R}/L\mathbb{Z}$):

$$u_{tt} - u_{xx} + \sin u = 0.$$

By introducing a new variable $v := u_t$, we obtain the following first-order system:

$$\frac{\partial}{\partial t}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} \frac{\delta \mathcal{H}}{\delta u} \\ \frac{\delta \mathcal{H}}{\delta v} \end{bmatrix}, \qquad \mathcal{H}(u, v) = \int \left(\frac{1}{2}(u_x)^2 + \frac{1}{2}v^2 - \cos u \right) \mathrm{d}x.$$

We employ an energy-preserving spatial discretization

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u_k \\ v_k \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} v_k \\ -\delta_x^2 u_k + \sin u_k \end{bmatrix}, \\ H(u, v) &= \sum_{k=1}^{N} \left(\frac{1}{2}(\delta_x u_k)^2 + \frac{1}{2}(v_k)^2 - \cos u_k \right) \Delta x. \end{aligned}$$

Since $E(u) = \sum_{k=1}^{N} \cos u_k \Delta x$ is bounded, we employ the scheme defined by Definition 3.

In this case, since we employ the Fourier spectral difference δ_x which can be diagonalized by the Fourier transform matrix F, i.e., $\delta_x = F^* (\mathrm{i}\Lambda) F$ ($\Lambda \in \mathbb{R}^{N \times N}$ is a diagonal matrix), we have

$$\begin{aligned} (JL)^{2m} &= \begin{bmatrix} F & \\ & F \end{bmatrix}^* \begin{bmatrix} (-1)^m \Lambda^{2m} & \\ & (-1)^m \Lambda^{2m} \end{bmatrix} \begin{bmatrix} F & \\ & F \end{bmatrix}, \\ (JL)^{2m+1} &= \begin{bmatrix} F & \\ & F \end{bmatrix}^* \begin{bmatrix} & (-1)^m \Lambda^{2m} \\ (-1)^{m+1} \Lambda^{2m+2} & \end{bmatrix} \begin{bmatrix} F & \\ & F \end{bmatrix} \end{aligned}$$

for all $m = 0, 1, \ldots$. Therefore, we have

$$\exp(tJL) = \begin{bmatrix} F & \\ & F \end{bmatrix}^* \begin{bmatrix} \cos(t\Lambda) & f_{\mathrm{sG}}(t,\Lambda) \\ -\Lambda \sin(t\Lambda) & \cos(t\Lambda) \end{bmatrix} \begin{bmatrix} F & \\ & F \end{bmatrix},$$

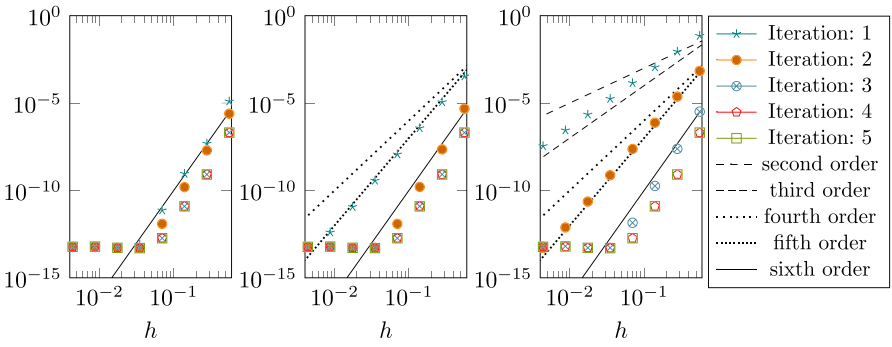
where $f_{\mathrm{sG}}(t,\Lambda)$ is a diagonal matrix defined by $(f_{\mathrm{sG}}(t,\Lambda))_{ii} = \Lambda_{ii}^{-1} \sin(t\Lambda_{ii})$ if $\Lambda_{ii} \neq 0$ and $(f_{\mathrm{sG}}(t,\Lambda))_{ii} = t$ otherwise.

In numerical experiments, we use the exact solution [36]

$$u(t,x) = 4 \arctan\left(\gamma^2 \mathrm{cn}(\beta_1 x \mid k_1)\mathrm{cn}(\beta_2 t \mid k_2)\right),$$

where $k_1^2 = \frac{\gamma^2}{1+\gamma^2}\left(1 + \frac{1}{\beta_1(1+\gamma^2)}\right)$, $k_2^2 = \frac{\gamma^2}{1+\gamma^2}\left(1 - \frac{1}{\beta_2(1+\gamma^2)}\right)$, $\beta_2^2 = \beta_1^2 + \frac{1-\gamma^2}{1+\gamma^2}$, β_1 and γ are parameters (we choose $\gamma = 0.1$ and $\beta_1 = 1$ in the numerical experiments below). This solution has the spatial period $L = 4K(k_1^2)$ and the temporal period $T = 4K(k_2^2)$. In the numerical experiment, we choose $N = 16$.

Figure 5 summarizes the relative errors. The errors achieved higher orders of convergence than expected by Theorem 10. From the numerical results, the order of accuracy seems to be $\min\{p, q + 2k - 1\}$. In fact, this can be proved by using the specific form of the sine-Gordon equation (see Appendix 5). There, we show that the order of accuracy with respect to u is $\min\{p, q + 2k - 1\}$, while that with respect to v and r is $\min\{p, q + 2k - 2\}$.



(a) $q = 6$ (b) $q = 4$ (c) $q = 2$

Fig. 5 Relative errors of the numerical solution with respect to u for the sG equation

Appendix 1. Derivation of the scheme defined by Definition 4

When we apply the scheme defined by (2) to the equation (18), the linear equation system (11) reads

$$
\begin{aligned}
U_i^{(k)} &= \exp\left(c_i h J L\right) u_0 \\
&\quad + 2h \sum_{j\in[s]} a_{ij} \exp\left((c_i - c_j) h J L\right) \left(R_{\mathrm{L},j}^{(k)} J\phi_{\mathrm{L}}\left(U_j^{(k-1)}\right) - R_{\mathrm{U},j}^{(k)} J\phi_{\mathrm{U}}\left(U_j^{(k-1)}\right)\right), \\
R_{\mathrm{L},i}^{(k)} &= r_{\mathrm{L},0} - h \sum_{j\in[s]} a_{ij} \left\langle J\phi_{\mathrm{L}}\left(U_j^{(k-1)}\right), L U_j^{(k)} \right\rangle \\
&\quad - 2h \sum_{j\in[s]} a_{ij} R_{\mathrm{U},j}^{(k)} \left\langle \phi_{\mathrm{L}}\left(U_j^{(k-1)}\right), J\phi_{\mathrm{U}}\left(U_j^{(k-1)}\right)\right\rangle, \\
R_{\mathrm{U},i}^{(k)} &= r_{\mathrm{U},0} - h \sum_{j\in[s]} a_{ij} \left\langle J\phi_{\mathrm{U}}\left(U_j^{(k-1)}\right), L U_j^{(k)} \right\rangle \\
&\quad - 2h \sum_{j\in[s]} a_{ij} R_{\mathrm{L},j}^{(k)} \left\langle \phi_{\mathrm{L}}\left(U_j^{(k-1)}\right), J\phi_{\mathrm{U}}\left(U_j^{(k-1)}\right)\right\rangle.
\end{aligned}
$$

By introducing $\psi_i^{\mathrm{X}} = \exp\left(-c_i h J L\right) J\phi_{\mathrm{X}}\left(U_i^{(k-1)}\right)$, we see

$$
\begin{aligned}
\left\langle J\phi_{\mathrm{X}}\left(U_i^{(k-1)}\right), L U_i^{(k)} \right\rangle &= \left\langle \psi_i^{\mathrm{X}}, \left(\exp\left(c_i h J L\right)\right)^* L U_i^{(k)} \right\rangle \\
&= \left\langle \psi_i^{\mathrm{X}}, L \exp\left(-c_i h J L\right) U_i^{(k)} \right\rangle \\
&= \left\langle \psi_i^{\mathrm{X}}, L \left(u_0 + 2h \sum_{j\in[s]} a_{ij} \left(R_{\mathrm{L},j}^{(k)} \psi_j^{L} - R_{\mathrm{U},j}^{(k)} \psi_j^{U} \right) \right) \right\rangle \\
&= \langle \psi_i^{\mathrm{X}}, L u_0 \rangle + 2h \sum_{j\in[s]} a_{ij} \left(R_{\mathrm{L},j}^{(k)} \langle \psi_i^{\mathrm{X}}, L\psi_j^{\mathrm{L}} \rangle - R_{\mathrm{U},j}^{(k)} \langle \psi_i^{\mathrm{X}}, L\psi_j^{\mathrm{U}} \rangle \right).
\end{aligned}
$$

Therefore, we obtain the linear equation system

$$
\begin{aligned}
R_{\mathrm{L},i}^{(k)} &= r_{\mathrm{L},0} - h \sum_{\ell\in[s]} a_{i\ell} \langle \psi_\ell^{\mathrm{L}}, L u_0 \rangle - 2h^2 \sum_{j\in[s]} \left(\sum_{\ell\in[s]} a_{i\ell} a_{\ell j} \langle \psi_\ell^{\mathrm{L}}, L\psi_j^{\mathrm{L}} \rangle \right) R_{\mathrm{L},j}^{(k)} \\
&\quad - 2h \sum_{j\in[s]} \left(a_{ij} \langle \phi_{\mathrm{L}}\left(U_j^{(k-1)}\right), J\phi_{\mathrm{U}}\left(U_j^{(k-1)}\right) s \rangle - h \sum_{\ell\in[s]} a_{i\ell} a_{\ell j} \langle \psi_\ell^{\mathrm{L}}, L\psi_j^{\mathrm{U}} \rangle \right) R_{\mathrm{U},j}^{(k)}, \\
R_{\mathrm{U},i}^{(k)} &= r_{\mathrm{U},0} - h \sum_{\ell\in[s]} a_{i\ell} \langle \psi_\ell^{\mathrm{U}}, L u_0 \rangle + 2h^2 \sum_{j\in[s]} \left(\sum_{\ell\in[s]} a_{i\ell} a_{\ell j} \langle \psi_\ell^{\mathrm{U}}, L\psi_j^{\mathrm{U}} \rangle \right) R_{\mathrm{U},j}^{(k)} \\
&\quad - 2h \sum_{j\in[s]} \left(a_{ij} \langle \phi_{\mathrm{L}}\left(U_j^{(k-1)}\right), J\phi_{\mathrm{U}}\left(U_j^{(k-1)}\right) \rangle + h \sum_{\ell\in[s]} a_{i\ell} a_{\ell j} \langle \psi_\ell^{\mathrm{U}}, L\psi_j^{\mathrm{L}} \rangle \right) R_{\mathrm{L},j}^{(k)}.
\end{aligned}
$$

Appendix 2. Proofs of theorems in 4.2

Theorem 11 is a corollary of the following theorem.

Theorem 14 *Suppose that the step size h satisfies*

$$h + 2h^2\|A\|_2\|J\|\|L\| \max\left\{1, \exp\left(\frac{c_{\max}h}{2}\lambda_{\max}(LJ - JL)\right)\right\} < \frac{1}{2\|A\|_2\|J\|C_\phi}.$$

Then, the linear equation (19) has a unique solution.

Proof Since the coefficient matrix of the linear equation (19) can be written as

$$I_{2s} + 2h\begin{bmatrix} & A\Phi \\ A\Phi & \end{bmatrix} + 2h^2\begin{bmatrix} A\left(A\circ\Psi^{\mathrm{LL}}\right) & -A\left(A\circ\Psi^{\mathrm{LU}}\right) \\ A\left(A\circ\Psi^{\mathrm{UL}}\right) & -A\left(A\circ\Psi^{\mathrm{UU}}\right) \end{bmatrix},$$

it is sufficient to evaluate the spectral norms of the second and the third terms. The second term can be evaluated as

$$\left\|\begin{bmatrix} & A\Phi \\ A\Phi & \end{bmatrix}\right\|_2 = \|A\Phi\|_2 \le \|A\|_2 \max_i \left|\left\langle\phi_{\mathrm{L}}\left(U_i^{(k-1)}\right), J\phi_{\mathrm{U}}\left(U_i^{(k-1)}\right)\right\rangle\right| \le \|A\|_2\|J\|C_\phi.$$

The third term can be evaluated as

$$\begin{aligned}\left\|\begin{bmatrix} A\left(A\circ\Psi^{\mathrm{LL}}\right) & -A\left(A\circ\Psi^{\mathrm{LU}}\right) \\ A\left(A\circ\Psi^{\mathrm{UL}}\right) & -A\left(A\circ\Psi^{\mathrm{UU}}\right) \end{bmatrix}\right\| &= \left\|\begin{bmatrix} A & \\ & A \end{bmatrix}\left(\begin{bmatrix} A & -A \\ A & -A \end{bmatrix}\circ\begin{bmatrix} \Psi^{\mathrm{LL}} & \Psi^{\mathrm{LU}} \\ \Psi^{\mathrm{UL}} & \Psi^{\mathrm{UU}} \end{bmatrix}\right)\right\|_2 \\ &\le \|A\|_2\,(2\|A\|_2)\left\|\begin{bmatrix} \Psi^{\mathrm{LL}} & \Psi^{\mathrm{LU}} \\ \Psi^{\mathrm{UL}} & \Psi^{\mathrm{UU}} \end{bmatrix}\right\|_2,\end{aligned}$$

and

$$\left\|\begin{bmatrix} \Psi^{\mathrm{LL}} & \Psi^{\mathrm{LU}} \\ \Psi^{\mathrm{UL}} & \Psi^{\mathrm{UU}} \end{bmatrix}\right\|_2 \le \|L\|\|J\|^2C_\phi \max\left\{1, \exp\left(\frac{c_{\max}h}{2}\lambda_{\max}(LJ - JL)\right)\right\}$$

holds, which can be shown in a manner similar to the proof of Theorem 8. □

Next, we prove Theorem 13.

Proof of Theorem 13 It is sufficient to prove that the skew-symmetric matrix function

$$S_{\mathrm{MSAV}}(u) := \begin{bmatrix} & J\phi_{\mathrm{L}}(u) & J\phi_{\mathrm{U}}(u) \\ -\left(J\phi_{\mathrm{L}}(u)\right)^* & 0 & \langle\phi_{\mathrm{L}}(u), J\phi_{\mathrm{U}}(u)\rangle \\ -\left(J\phi_{\mathrm{U}}(u)\right)^* & -\langle\phi_{\mathrm{L}}(u), J\phi_{\mathrm{U}}(u)\rangle & 0 \end{bmatrix}$$

is Lipschitz continuous.

For any $v_1, v_2 \in \mathcal{V}$ and $\rho \in \mathbb{R}$, we have

$$\begin{aligned}\left\|\begin{bmatrix} & v_1 & v_2 \\ -v_1^* & & \rho \\ -v_2^* & -\rho & \end{bmatrix}\right\| &\le \left\|\begin{bmatrix} & v_1 & v_2 \\ -v_1^* & & \\ -v_2^* & & \end{bmatrix}\right\| + \left\|\begin{bmatrix} O & & \\ & & \rho \\ & -\rho & \end{bmatrix}\right\| \\ &\le \sqrt{\|v_1\|^2 + \|v_2\|^2} + |\rho| \\ &\le \|v_1\| + \|v_2\| + |\rho|.\end{aligned}$$

Therefore, we see

$$\begin{aligned}\|S_{\mathrm{MSAV}}(u_1) - S_{\mathrm{MSAV}}(u_2)\| &\le \|J\phi_{\mathrm{L}}(u_1) - J\phi_{\mathrm{L}}(u_2)\| + \|J\phi_{\mathrm{U}}(u_1) - J\phi_{\mathrm{U}}(u_2)\| \\ &\quad + |\langle \phi_{\mathrm{L}}(u_1), J\phi_{\mathrm{U}}(u_1)\rangle - \langle \phi_{\mathrm{L}}(u_2), J\phi_{\mathrm{U}}(u_2)\rangle| \\ &\le \|J\|\|\phi_{\mathrm{L}}(u_1) - \phi_{\mathrm{L}}(u_2)\| + \|J\|\|\phi_{\mathrm{U}}(u_1) - \phi_{\mathrm{U}}(u_2)\| \\ &\quad + \frac{1}{2}\left|\left\langle \phi_{\mathrm{L}}(u_1) - \phi_{\mathrm{L}}(u_2),\, J\left(\phi_{\mathrm{U}}(u_1) + \phi_{\mathrm{U}}(u_2)\right)\right\rangle\right| \\ &\quad + \frac{1}{2}\left|\left\langle \phi_{\mathrm{L}}(u_1) + \phi_{\mathrm{L}}(u_2),\, J\left(\phi_{\mathrm{U}}(u_1) - \phi_{\mathrm{U}}(u_2)\right)\right\rangle\right| \\ &\le \|J\|\left(1 + \frac{\|\phi_{\mathrm{U}}(u_1) + \phi_{\mathrm{U}}(u_2)\|}{2}\right)\|\phi_{\mathrm{L}}(u_1) - \phi_{\mathrm{L}}(u_2)\| \\ &\quad + \|J\|\left(1 + \frac{\|\phi_{\mathrm{L}}(u_1) + \phi_{\mathrm{L}}(u_2)\|}{2}\right)\|\phi_{\mathrm{U}}(u_1) - \phi_{\mathrm{U}}(u_2)\|,\end{aligned}$$

which shows the Lipschitz continuity of S_{MSAV} under the assumption (A2). □

Appendix 3. Computational efficiency of the scheme defined by Definition 3

In this section, we evaluate the computational efficiency of the scheme defined by Definition 3 in terms of the number of the computation of the product of matrix exponential and vector. To this end, we consider the explicit Lawson method and the four types of the initial approximation. Before evaluating the computational efficiency, we summarize the number of stages of the explicit Runge–Kutta and continuous explicit Runge–Kutta (CERK) methods achieving the order p in Table 1.

Let us list the four types of initial approximation. To achieve the order p, we consider the Gauss method with $s = \lceil p/2 \rceil$ stages as the base Runge–Kutta method. Then, to minimize the number of iterations in Definition 3, we choose $k = p - q + 1$. Recall that Steps 1 and 2 in Definition 3 require the computation of the product of matrix exponential and vector $(2k-1)s + 1$ times in sequential computation and $2k$ times in parallel computation. Below, N_{S} and N_{P} denote the number of the computation of the product of matrix exponential and vector in sequential and parallel computation, respectively.

Table 1 The number of stages of the explicit RK and CERK methods achieving the order p

Methods\order	1	2	3	4	5	6	7	8
Explicit RK	1	2	3	4	6	7	9	11
Continuous explicit RK	1	2	4	6	8	11	15	–

The written number is the minimum number of stages achieving the order p (cf. [37–39]) except for seventh-order CERK, where we refer to the 15 stage method by Verner [39]. (Eighth-order CERK is omitted since we do not use it later.)

- Proposed (I): the simplest initial approximation $U_i^{(0)} = u_0$. In this case, since $q = 1$ holds, we choose $k = p$. Then, $N_S = (2p-1)s + 1$ and $N_P = 2p$ hold.
- Proposed (II): an accurate initial approximation using continuous explicit Runge–Kutta methods with the Lawson transformation. In this case, since $q = p$ holds, we choose $k = 1$. Then, $N_S = 2(s^* - 1) + 2s + 1$ and $N_P = 2(s^* - 1) + 3$ hold, where s^* is the number of stages of the continuous explicit Runge–Kutta method achieving the order $p - 1$ (see Table 1).
- Proposed (III): initial approximation using extrapolation of inner stages in the previous step ($p \geq 2$). The inner stages of the Gauss method are $(s+1)$th order approximations, the order of approximation is at most $(s+1)$th order. Therefore, let us construct the extrapolation satisfying $q = s + 1$ and $k = p - s$. However, since $U_k^{(i)}$ is not computed at the last iteration in Definition 3, the situation is a bit complicated. First, when $p = 2$, we have $s = 1$ and $k = 1$. In this case, to construct a second-order approximation, we employ the extrapolation using the input u_0 and u_1 in the previous step. Therefore, $N_S = N_P = 2$ hold. Next, when $p = 3$, we have $s = 2$ and $k = 1$. In this case, to construct a third-order approximation, we employ the extrapolation using the input u_0, u_1, and one of the inner stages in the previous step. Therefore, $N_S = 4$ and $N_P = 3$ hold. Finally, when $p \geq 4$, we have $k \geq 2$. In this case, we can use $U_i^{(k-1)}$'s which are computed in the previous step. Therefore, $N_S = (2p - 2s - 1)s + 1$ and $N_P = 2(p - s)$ hold.
- Proposed (IV): initial approximation using extrapolation of several previous steps. In this case, by using sufficiently many previous steps, we can construct the extrapolation satisfying $q = p$ so that $k = 1$. Then, we have $N_S = s + 1$ and $N_P = 2$ hold.

We compare the computational efficiency of the proposed methods with that of the explicit Lawson method in Table 2. Note that the explicit Lawson method requires the product of the matrix exponential and vector $2s - 1$ times.

As shown in Table 2, the computational efficiency of the proposed method largely depends on the choice of the initial approximation. In particular, the extrapolation methods (III) and (IV) are much more efficient than the other methods. However, as reported in [27], the extrapolation method (III) causes the instability in the numerical solution. We expect worse stability for the extrapolation method (IV). In this sense, we believe the methods (I) and (II) (numerically tested in Sect. 5) are better to use in practice. As summarized in Table 2, the proposed method (I) is better in parallel computation, while the proposed method (II) is better in sequential computation.

Table 2 The number of the computation of the product of matrix exponential and vector

Methods\order	1	2	3	4	5	6	7	8
Explicit Lawson	1	3	5	7	11	13	17	21
Proposed (Is)	2	4	11	15	28	34	53	61
Proposed (Ip)	2	4	6	8	10	12	14	16
Proposed (IIs)	2	3	7	11	17	21	29	37
Proposed (IIp)	2	3	5	9	13	17	23	31
Proposed (IIIs)	–	2	4	7	10	16	21	29
Proposed (IIIp)	–	2	3	4	4	6	6	8
Proposed (IVs)	–	2	3	3	4	4	5	5
Proposed (IVp)	–	2	2	2	2	2	2	2

The symbol (Is) and (Ip) denote the proposed method (I) in sequential and parallel computation, respectively. The other symbols have the same meaning

Appendix 4. Application of the scheme by [16] to the SAV system (18)

In this section, we consider the scheme by [16] (i.e., Definition 1) applied to the SAV system (18).

In this case, the (3) reads

$$\begin{bmatrix} U_i^{(k)} \\ R_{\mathrm{L},i}^{(k)} \\ R_{\mathrm{U},i}^{(k)} \end{bmatrix} = \begin{bmatrix} u_0 \\ r_{\mathrm{L},0} \\ r_{\mathrm{U},0} \end{bmatrix} + h \sum_{j\in[s]} a_{ij} \begin{bmatrix} JLU_j^{(k)} + 2R_{\mathrm{L},j}^{(k)} J\hat{\phi}_{\mathrm{L},j} - 2R_{\mathrm{U},j}^{(k)} J\hat{\phi}_{\mathrm{U},j} \\ -\left\langle J\hat{\phi}_{\mathrm{L},j}, LU_j^{(k)} \right\rangle - 2R_{\mathrm{U},j} \left\langle \hat{\phi}_{\mathrm{L},j}, J\hat{\phi}_{\mathrm{U},j} \right\rangle \\ -\left\langle J\hat{\phi}_{\mathrm{U},j}, LU_j^{(k)} \right\rangle - 2R_{\mathrm{L},j} \left\langle \hat{\phi}_{\mathrm{L},j}, J\hat{\phi}_{\mathrm{U},j} \right\rangle \end{bmatrix}, \quad (22)$$

where $\hat{\phi}_{\mathrm{X},j}^{(k-1)} = \phi_{\mathrm{X}}\left(U_j^{(k-1)}\right)$ for $\mathrm{X} \in \{\mathrm{L}, \mathrm{U}\}$. Let $U^{(k)}$ denote the vector $\left[U_1^{(k)} \cdots U_s^{(k)}\right]^{\mathsf{T}}$. Then, the first equation in the above equation can be written as

$$(I_s \otimes I_{\mathcal{V}} - hA \otimes JL)U^{(k)} = \mathbf{1}_s \otimes u_0 + 2h \sum_{j\in[s]} A_j \otimes J\hat{\phi}_{\mathrm{L},j}^{(k-1)} - 2h \sum_{j\in[s]} A_j \otimes J\hat{\phi}_{\mathrm{U},j}^{(k-1)},$$

where $\otimes$ denotes the Kronecker product, $I_s \in \mathbb{R}^{s\times s}$ is the identity matrix, $I_{\mathcal{V}} \in \mathcal{L}(\mathcal{V})$ is the identity operator, $A \in \mathbb{R}^{s\times s}$ is the matrix with the entries a_{ij}, $A_j = \left[a_{1j} \cdots a_{sj}\right]^{\mathsf{T}} \in \mathbb{R}^s$ is the vector, and $\mathbf{1}_s \in \mathbb{R}^s$ is the vector with the entries 1. Since the linear equation in the form

$$(I_s \otimes I_{\mathcal{V}} - hA \otimes JL)x = w \otimes v$$

can be efficiently solved (cf. [40, IV.8]), it is computationally inexpensive to compute

$$\begin{bmatrix} \tilde{u}_1 \\ \vdots \\ \tilde{u}_s \end{bmatrix} = (I_s \otimes I_{\mathcal{V}} - hA \otimes JL)^{-1} (\mathbf{1}_s \otimes u_0), \tag{23}$$

$$\begin{bmatrix} \tilde{\phi}^{\mathrm{X}}_{j1} \\ \vdots \\ \tilde{\phi}^{\mathrm{X}}_{js} \end{bmatrix} = (I_s \otimes I_{\mathcal{V}} - hA \otimes JL)^{-1} \left(A_j \otimes J\hat{\phi}^{(k-1)}_{\mathrm{X},j} \right). \tag{24}$$

By using this, $U_i^{(k)}$ can be written as

$$U_i^{(k)} = \tilde{u}_i + 2h \sum_{j\in[s]} R^{(k)}_{\mathrm{L},j} \tilde{\phi}^{\mathrm{L}}_{ji} - 2h \sum_{j\in[s]} R^{(k)}_{\mathrm{U},j} \tilde{\phi}^{\mathrm{U}}_{ji}. \tag{25}$$

Then, by using the above equation, the second and the third equations in (22) can be simplified into

$$\begin{bmatrix} I_s + 2h^2 A\tilde{\Phi}_{\mathrm{LL}} & -2h^2 A\tilde{\Phi}_{LU} + 2hA\Phi \\ 2h^2 A\tilde{\Phi}_{\mathrm{UL}} + 2hA\Phi & I_s - 2h^2 A\tilde{\Phi}_{\mathrm{UU}} \end{bmatrix} \begin{bmatrix} R^{(k)}_{\mathrm{L}} \\ R^{(k)}_{\mathrm{U}} \end{bmatrix} = \begin{bmatrix} r_{\mathrm{L},0}\mathbf{1}_s \\ r_{\mathrm{U},0}\mathbf{1}_s \end{bmatrix} - h \begin{bmatrix} A\nu_{\mathrm{L}} \\ A\nu_{\mathrm{U}} \end{bmatrix}, \tag{26}$$

where $R^{(k)}_{\mathrm{X}} = \begin{bmatrix} R^{(k)}_{\mathrm{X},1} & \cdots & R^{(k)}_{\mathrm{X},s} \end{bmatrix}^{\mathsf{T}}$, $\tilde{\Phi}_{\mathrm{XY}} \in \mathbb{R}^{s\times s}$ is defined by $\left(\tilde{\Phi}_{\mathrm{XY}_{ij}}\right) = \left\langle L\hat{\phi}_{\mathrm{X},i}, \tilde{\phi}^{\mathrm{Y}}_{ji} \right\rangle$, $\Phi \in \mathbb{R}^{s\times s}$ is a diagonal matrix defined by $\Phi_{ii} = \left\langle \hat{\phi}_{\mathrm{L},i}, J\hat{\phi}_{\mathrm{U},i} \right\rangle$, $\nu_{\mathrm{X}} \in \mathbb{R}^s$ is defined by $(\nu_{\mathrm{X}})_i = \left\langle L\hat{\phi}_{\mathrm{X},i}, \tilde{u}_i \right\rangle$. Consequently, we obtain the following scheme:

Definition 5

Step 0 *Prepare* $U_i^{(0)} \approx u(c_i h)$ *and set* $k = 1$.

Step 1 *Compute* $\tilde{u}_i$ *and* $\tilde{\phi}^{\mathrm{X}}_{ij}$ *for* $i, j \in [s]$ *by* (23) *and* (24). *Solve the linear equation system* (26) *to obtain* $R^{(k)}_{\mathrm{L}}$ *and* $R^{(k)}_{\mathrm{U}}$. *Compute* $\{U_i^{(k)}\}_{i=1}^s$ *by* (25). *If some criteria hold, go to Step 2. Otherwise, set* $k = k + 1$ *and repeat Step 1.*

Step 2 *Output*

$$\begin{bmatrix} u_1 \\ r_{\mathrm{L},1} \\ r_{\mathrm{U},1} \end{bmatrix} = \begin{bmatrix} u_0^{(k)} \\ r^{(k)}_{\mathrm{L},0} \\ r^{(k)}_{\mathrm{U},0} \end{bmatrix} + h \sum_{j\in[s]} b_j \begin{bmatrix} JLU_j^{(k)} + 2R^{(k)}_{\mathrm{L},j} J\hat{\phi}_{\mathrm{L},j} - 2R^{(k)}_{\mathrm{U},j} J\hat{\phi}_{\mathrm{U},j} \\ -\langle J\hat{\phi}_{\mathrm{L},j}, LU_j^{(k)} \rangle - 2R_{\mathrm{U},j} \langle \hat{\phi}_{\mathrm{L},j}, J\hat{\phi}_{\mathrm{U},j} \rangle \\ -\langle J\hat{\phi}_{\mathrm{U},j}, LU_j^{(k)} \rangle - 2R_{\mathrm{L},j} \langle \hat{\phi}_{\mathrm{L},j}, J\hat{\phi}_{\mathrm{U},j} \rangle \end{bmatrix}.$$

Appendix 5. The error behavior of the proposed scheme for the sine-Gordon equation

In this section, we consider the ODE

$$\begin{cases} \dot{u}(t) = v(t), \\ \dot{v}(t) = \delta_x^2 u(t) + 2r(t)\phi\left(u(t)\right), \\ \dot{r}(t) = -\langle \phi\left(u(t)\right), v(t) \rangle \end{cases} \tag{27}$$

given by a spatial discretization of the sine-Gordon equation.

Theorem 15 *Let (u, v, r) be the solution of the ODE (27). Assume the following conditions:*

(A1) $U_i^{(0)}$ satisfies $\left\| U_i^{(0)} - u(c_i h) \right\| \leq Ch^q$ for each $i \in [s]$.
(A2) α in the definition of ϕ satisfies $\alpha \geq 2N\Delta x$.
(A3) The base Runge–Kutta method is of order p.

Then, the solution $(u_1^{(k)}, v_1^{(k)}, r_1^{(k)})$ of the scheme defined by Definition 3 *satisfies*

$$\begin{aligned} \left\| u_1^{(k)}(h) - u(h) \right\| &\leq C' h^{\min\{p, q+2k-1\}+1}, \\ \left\| v_1^{(k)}(h) - v(h) \right\| &\leq C' h^{\min\{p, q+2k-2\}+1}, \\ \left\| r_1^{(k)}(h) - r(h) \right\| &\leq C' h^{\min\{p, q+2k-2\}+1} \end{aligned}$$

for a sufficiently small step size $h > 0$, Here, C' is a constant depending only on the exact solution (u, v, r), k, N, A, b, and the constant C in (A1).

Proof In this case, the coupled system considered in the proof of Theorem 7 reads

$$\begin{cases} \dot{u}^{(i)}(t) = v^{(i)}(t), \\ \dot{v}^{(i)}(t) = \delta_x^2 u^{(i)}(t) + 2r^{(i)}(t)\phi\left(u^{(i-1)}(t)\right), \\ \dot{r}^{(i)}(t) = -\langle \phi\left(u^{(i-1)}(t)\right), v^{(i)}(t) \rangle \end{cases}$$

for $i = 1, 2, \ldots, k$. According to the proof of Theorem 7, it is sufficient to show that $\sup_{t\in[0,h]} \left\| u^{(i)}(t) - u(t) \right\| \leq C'' h^{q+2k}$, $\sup_{t\in[0,h]} \left\| v^{(i)}(t) - v(t) \right\| \leq C'' h^{q+2k-1}$ and $\sup_{t\in[0,h]} \left| r^{(i)}(t) - r(t) \right| \leq C'' h^{q+2k-1}$.

First, we introduce several bounds with respect to ϕ. Since $E\left(u^{(i)}(t)\right) = \sum_{k=1}^N \cos\left(u_k^{(i)}(t)\right) \Delta x \geq -N\Delta x$ and $\left\| \sin\left(u^{(i)}(t)\right) \right\| = \sqrt{\sum_{k=1}^N \left| \sin\left(u_k^{(i)}\right) \right| \Delta x} \leq \sqrt{N\Delta x}$ hold, we see

$$\left\| \phi\left(u^{(i)}(t)\right) \right\| = \frac{1}{2\sqrt{E\left(u^{(i)}(t)\right) + \alpha}} \left\| \sin\left(u^{(i)}(t)\right) \right\| \leq \frac{\sqrt{N\Delta x}}{2\sqrt{\alpha - N\Delta x}},$$

which implies $\sup_{t\in[0,h]} \|\phi\left(u^{(i)}(t)\right)\| \le 1/2$ by choosing $\alpha = 2N\Delta x$. In addition, under the setting $\alpha = 2N\Delta x$, for any $v, w \in \mathbb{R}^N$, we see

$$
\begin{aligned}
&\|\phi(v) - \phi(w)\| \\
&= \left\| \frac{1}{2\sqrt{E(v)+\alpha}} \sin(v) - \frac{1}{2\sqrt{E(w)+\alpha}} \sin(w) \right\| \\
&\le \frac{1}{2\sqrt{E(v)+\alpha}} \|\sin(v) - \sin(w)\| + \left| \frac{1}{2\sqrt{E(v)+\alpha}} - \frac{1}{2\sqrt{E(w)+\alpha}} \right| \|\sin(w)\| \\
&\le \frac{1}{\sqrt{N\Delta x}} \|v - w\| + \frac{\sqrt{N\Delta x}}{2} \left| \frac{E(v) - E(w)}{\sqrt{(E(v)+\alpha)\,(E(w)+\alpha)}\left(\sqrt{E(v)+\alpha} + \sqrt{E(w)+\alpha}\right)} \right| \\
&\le \frac{1}{\sqrt{N\Delta x}} \|v - w\| + \frac{1}{4N\Delta x} \left| \sum_{k=1}^{N} (\cos v_k - \cos w_k)\, \Delta x \right| \\
&= \frac{1}{\sqrt{N\Delta x}} \|v - w\| + \frac{1}{2N\Delta x} \left| \sum_{k=1}^{N} \sin \frac{v_k + w_k}{2} \sin \frac{v_k - w_k}{2} \Delta x \right| \\
&\le \frac{1}{\sqrt{N\Delta x}} \|v - w\| + \frac{1}{2N\Delta x} \left\| \sin \frac{v+w}{2} \right\| \left\| \sin \frac{v-w}{2} \right\| \\
&\le \frac{5}{4\sqrt{N\Delta x}} \|v - w\|.
\end{aligned}
$$

Therefore, the map ϕ is Lipschitz continuous with the Lipschitz constant $L_\phi := 5/(4\sqrt{N\Delta x})$.

Let us introduce $E_u^{(i)} := \sup_{t\in[0,h]} \left\|u^{(i)}(t) - u(t)\right\|$, $E_v^{(i)} := \sup_{t\in[0,h]} \left\|v^{(i)}(t) - v(t)\right\|$ and $E_r^{(i)} := \sup_{t\in[0,h]} \left|r^{(i)}(t) - r(t)\right|$. Then, by introducing $C_v := \sup_{t\in[0,h]} \left\|v(t)\right\|$ and $C_r := \sup_{t\in[0,h]} \left|r(t)\right|$, we see

$$
\begin{aligned}
E_u^{(i)} &= \sup_{t\in[0,h]} \left\| \int_0^t \left(v^{(i)}(\tau) - v(\tau) \right) \mathrm{d}\tau \right\| \\
&\le h E_v^{(i)}, \\
E_v^{(i)} &= \sup_{t\in[0,h]} \left\| \int_0^t \left(\delta_x^2 u^{(i)}(\tau) + 2r^{(i)}(\tau)\phi\left(u^{(i-1)}(\tau)\right) - \delta_x^2 u(\tau) - 2r(\tau)\phi\left(u(\tau)\right) \right) \mathrm{d}\tau \right\| \\
&\le h \left\|\delta_x^2\right\| E_u^{(i)} + 2 \sup_{t\in[0,h]} \left\| \int_0^t \left(r^{(i)}(\tau) - r(\tau) \right) \phi\left(u^{(i-1)}(\tau)\right) \mathrm{d}\tau \right\| \\
&\qquad + 2 \sup_{t\in[0,h]} \left\| \int_0^t r(\tau) \left(\phi\left(u^{(i-1)}(\tau)\right) - \phi\left(u(\tau)\right) \right) \mathrm{d}\tau \right\| \\
&\le h \left\|\delta_x^2\right\| E_u^{(i)} + h E_r^{(i)} + 2h C_r L_\phi E_u^{(i-1)}, \\
E_r^{(i)} &= \sup_{t\in[0,h]} \left| \int_0^t \left(-\left\langle \phi\left(u^{(i-1)}(\tau)\right), v^{(i)}(\tau) \right\rangle + \left\langle \phi\left(u(\tau)\right), v(\tau) \right\rangle \right) \mathrm{d}\tau \right| \\
&\le \frac{h}{2} E_v^{(i)} + h C_v L_\phi E_u^{(i-1)}.
\end{aligned}
\tag{28}
$$

The inequalities on $E_v^{(i)}$ and $E_r^{(i)}$ imply

$$\left(1-\frac{h^2}{2}\right) E_v^{(i)} \le h\|\delta_x^2\| E_u^{(i)} + hL_\phi\,(2C_r + hC_v)\, E_u^{(i-1)}.$$

Therefore, when $h \le 1$, we see

$$E_v^{(i)} \le 2h\left\|\delta_x^2\right\| E_u^{(i)} + 2hL_\phi\,(2C_r + hC_v)\, E_u^{(i-1)}. \tag{29}$$

This inequality and $E_u^{(i)} \le hE_v^{(i)}$ imply

$$\left(1-2h\left\|\delta_x^2\right\|\right) E_u^{(i)} \le 2h^2 L_\phi\,(2C_r + hC_v)\, E_u^{(i-1)}.$$

Therefore, when $h \le \left(4\left\|\delta_x^2\right\|\right)^{-1}$, we see

$$E_u^{(i)} \le 4h^2 L_\phi\,(2C_r + hC_v)\, E_u^{(i-1)},$$

which implies $E_u^{(i)} \le C^{(k)}h^{q+2k}$, where $C^{(k)} = \left(4L_\phi\,(2C_r + C_v)\right)^k$. This fact and inequalities (29) and (28) show the theorem. □

As indicated in the proof above, the same argument holds true for the nonlinear Klein–Gordon equation with a general potential function when ϕ is Lipschitz continuous.

Acknowledgements We would like to thank Editage (www.editage.com) for English language editing. The author would also like to thank the anonymous referees for their helpful comments.

Author contribution S.S. wrote the whole manuscript.

Funding Open Access funding provided by The University of Tokyo. The author is supported by a Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) (No. 20H01822) and a JSPS Grant-in-Aid for Early-Career Scientists (No. 22K13955).

Data Availability The code and data that support the findings of this paper are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The author declares no competing interests.

References

1. Celledoni, E., Grimm, V., McLachlan, R.I., McLaren, D., O'Neale, D., Owren, B., Quispel, G.R.: Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method. J. Comp. Phys. **231**, 6770–6789 (2012)
2. Gonzalez, O.: Time integration and discrete Hamiltonian systems. J. Nonlinear Sci. **6**, 449–467 (1996)
3. McLachlan, R.I., Quispel, G.R.W., Robidoux, N.: Unified approach to Hamiltonian systems, Poisson systems, gradient systems with Lyapunov functions or first integrals. Phys. Rev. Lett. **81**, 2399–2403 (1998)
4. McLachlan, R.I., Quispel, G.R.W., Robidoux, N.: Geometric integration using discrete gradients. Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci. **357**, 1021–1045 (1999)
5. Furihata, D.: Finite difference schemes for $\partial u/\partial t = (\partial/\partial x)^{\alpha}\delta G/\delta u$ that inherit energy conservation or dissipation property. J. Comput. Phys. **156**(1), 181–205 (1999). https://doi.org/10.1006/jcph.1999.6377
6. Furihata, D., Mori, M.: General derivation of finite difference schemes by means of a discrete variation (in Japanese). Trans. Japan Soc. Indust. Appl. **8**(3), 317–340 (1998)
7. Furihata, D., Matsuo, T.: Discrete variational derivative method-a structure-preserving numerical method for partial differential equations. CRC Press, Boca Raton (2011)
8. Cohen, D., Hairer, E.: Linear energy-preserving integrators for Poisson systems. BIT **51**(1), 91–101 (2011). https://doi.org/10.1007/s10543-011-0310-z
9. Besse, C.: A relaxation scheme for the nonlinear Schrödinger equation. SIAM J. Numer. Anal. **42**(3), 934–952 (2004). https://doi.org/10.1137/S0036142901396521
10. Zhang, F., Pérez-García, V.M., Vázquez, L.: Numerical simulation of nonlinear Schrödinger systems: a new conservative scheme. Appl. Math. Comput. **71**(2–3), 165–177 (1995)
11. Matsuo, T., Furihata, D.: Dissipative or conservative finite-difference schemes for complex-valued nonlinear partial differential equations. J. Comput. Phys. **171**(2), 425–447 (2001). https://doi.org/10.1006/jcph.2001.6775
12. Dahlby, M., Owren, B.: A general framework for deriving integral preserving numerical methods for PDEs. SIAM J. Sci. Comput. **33**(5), 2318–2340 (2011). https://doi.org/10.1137/100810174
13. Yang, X., Han, D.: Linearly first- and second-order, unconditionally energy stable schemes for the phase field crystal model. J. Comput. Phys. **330**, 1116–1134 (2017). https://doi.org/10.1016/j.jcp.2016.10.020
14. Shen, J., Xu, J., Yang, J.: The scalar auxiliary variable (SAV) approach for gradient flows. J. Comput. Phys. **353**, 407–416 (2018). https://doi.org/10.1016/j.jcp.2017.10.021
15. Kemmochi, T., Sato, S.: Scalar auxiliary variable approach for conservative/dissipative partial differential equations with unbounded energy functionals. BIT **62**, 903–930 (2022)
16. Sato, S., Miyatake, Y., Butcher, J.C.: High-order linearly implicit schemes conserving quadratic invariants. Appl. Numer. Math. **187**, 71–88 (2023). https://doi.org/10.1016/j.apnum.2023.02.005
17. Hochbruck, M., Ostermann, A.: Exponential integrators. Acta Numer. **19**, 209–286 (2010). https://doi.org/10.1017/S0962492910000048
18. Celledoni, E., Cohen, D., Owren, B.: Symmetric exponential integrators with an application to the cubic Schrödinger equation. Found. Comput. Math. **8**, 303–317 (2008)
19. Mei, L., Huang, L., Huang, S.: Exponential integrators with quadratic energy preservation for linear Poisson systems. J. Comput. Phys. **387**, 446–454 (2019)
20. Li, Y.-W., Wu, X.: Exponential integrators preserving first integrals or Lyapunov functions for conservative or dissipative systems. J. Sci. Comput. **38**(3), 1876–1895 (2016)
21. Mei, L., Huang, L., Wu, X.: Energy-preserving exponential integrators of arbitrarily high order for conservative or dissipative systems with highly oscillatory solutions. J. Comput. Phys. **442**, 110429 (2021). https://doi.org/10.1016/j.jcp.2021.110429
22. Li, L.: A new symmetric linearly implicit exponential integrator preserving polynomial invariants or Lyapunov functions for conservative or dissipative systems. J. Comput. Phys. **449**, 110800 (2022). https://doi.org/10.1016/j.jcp.2021.110800
23. Jiang, C., Wang, Y., Cai, W.: A linearly implicit energy-preserving exponential integrator for the nonlinear Klein-Gordon equation. J. Comput. Phys. **419**, 109690 (2020)
24. Cui, J., Xu, Z., Wang, Y., Jiang, C.: Mass- and energy-preserving exponential Runge-Kutta methods for the nonlinear Schrödinger equation. Appl. Math. Lett. **112**, 106770 (2021). https://doi.org/10.1016/j.aml.2020.106770

25. Xu, Z., Cai, W., Song, Y., Wang, Y.: Explicit high-order energy-preserving exponential time differencing method for nonlinear Hamiltonian PDEs. Appl. Math. Comput. **404**, 126208 (2021)
26. Fu, Y., Hu, D., Xu, Z.: High-order explicit conservative exponential integrator schemes for fractional Hamiltonian PDEs. Appl. Numer. Math. **172**, 315–331 (2022). https://doi.org/10.1016/j.apnum.2021.10.011
27. Jiang, C., Cui, J., Qian, X., Song, S.: High-order linearly implicit structure-preserving exponential integrators for the nonlinear Schrödinger equation. J. Sci. Comput. **90**(1), 66–27 (2022). https://doi.org/10.1007/s10915-021-01739-x
28. Lawson, J.D.: Generalized Runge-Kutta processes for stable systems with large Lipschitz constants. SIAM J. Numer. Anal. **4**(3), 372–380 (1967)
29. Cooper, G.J.: Stability of Runge-Kutta methods for trajectory problems. IMA J. Numer. Anal. **7**(1), 1–13 (1987). https://doi.org/10.1093/imanum/7.1.1
30. Butcher, J.C.: B-series—algebraic analysis of numerical methods. Springer Series in Computational Mathematics, vol. 55, p. 310. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70956-3
31. Horn, R.A., Johnson, C.R.: Topics in matrix analysis, p. 607. Cambridge University Press, Cambridge (1991). https://doi.org/10.1017/CBO9780511840371
32. Trefethen, L.N., Embree, M.: Spectra and pseudospectra, p. 606. Princeton University Press, Princeton, NJ (2005)
33. Shen, X., Leok, M.: Geometric exponential integrators. J. Comput. Phys. **382**, 27–42 (2019). https://doi.org/10.1016/j.jcp.2019.01.005
34. Fornberg, B.: A practical guide to pseudospectral methods. Cambridge Monographs on Applied and Computational Mathematics, vol. 1, p. 231. Cambridge University Press, Cambridge (1996). https://doi.org/10.1017/CBO9780511626357
35. Butcher, J.C.: On Runge-Kutta processes of high order. J. Austral. Math. Soc. **4**, 179–194 (1964)
36. Marchesoni, F.: Exact solutions of the sine-Gordon equation with periodic boundary conditions. Progr. Theoret. Phys. **77**(4), 813–824 (1987). https://doi.org/10.1143/PTP.77.813
37. Zhang, D.K.: Discovering new Runge-Kutta methods using unstructured numerical search (2019)
38. Owren, B., Zennaro, M.: Continuous explicit Runge-Kutta methods. In: Computational Ordinary Differential Equations (London, 1989). Inst. Math. Appl. Conf. Ser. New Ser., vol. 39, pp. 97–105. Oxford Univ. Press, New York, ??? (1992)
39. Verner, J.H.: Differentiable interpolants for high-order Runge-Kutta methods. SIAM J. Numer. Anal. **30**(5), 1446–1466 (1993). https://doi.org/10.1137/0730075
40. Hairer, E., Wanner, G.:Solving ordinary differential equations. II, Stiff and Differential-algebraic Problems. Springer Series in Computational Mathematics, vol. 14, p. 614. Springer, Berlin (2010). https://doi.org/10.1007/978-3-642-05221-7