

Structure-preserving algorithms with uniform error bound and long-time energy conservation for highly oscillatory Hamiltonian systems

Bin Wang* Yaolin Jiang †

February 8, 2021

Abstract

Structure-preserving algorithms and algorithms with uniform error bound have constituted two interesting classes of numerical methods. In this paper, we blend these two kinds of methods for solving nonlinear Hamiltonian systems with highly oscillatory solution, and the blended algorithms inherit and respect the advantage of each method. Two kinds of algorithms are presented to preserve the symplecticity and energy of the Hamiltonian systems, respectively. Moreover, the proposed algorithms are shown to have uniform error bound for the highly oscillatory structure. A numerical experiment is carried out to support the theoretical results established in this paper by showing the performance of the blended algorithms.

Keywords: Hamiltonian system, highly oscillatory solution, symplectic algorithms, energy-preserving algorithms, uniform error bound, long-time conservation

MSC: 65L05, 65L20, 65L70, 65P10.

1 Introduction

It is known that nonlinear Hamiltonian systems are ubiquitous in science and engineering applications. In numerical simulation of evolutionary problems, one of the most difficult problems is to deal with highly oscillatory problems, since they cannot be solved efficiently using conventional methods. The crucial point is that standard methods need a very small stepsize and hence a long runtime to reach an acceptable accuracy [16]. In this paper we are concerned with efficient algorithms for the following damped second-order differential equation

$$\ddot{x}(t) = \frac{1}{\varepsilon} \tilde{B} \dot{x}(t) + F(x(t)), \quad x(0) = x_0, \quad \dot{x}(0) = \dot{x}_0, \quad t \in [0, T], \quad (1)$$

where $x(t) \in \mathbb{R}^d$, \tilde{B} is a $d \times d$ skew symmetric matrix, and $F(x) = -\nabla_x U(x)$ is the negative gradient of a real-valued function $U(x)$ whose second derivatives are continuous. In this work, we focus on

*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shannxi 710049, P.R.China. E-mail: wangbinmaths@xjtu.edu.cn

†School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shannxi 710049, P.R.China. E-mail: yljjiang@mail.xjtu.edu.cn

the case where $0 < \varepsilon \ll 1$. This implies that the solution of this dynamic is *highly oscillatory*. For the dimension d , it is required that $d \geq 2$ since \tilde{B} is a zero matrix once $d = 1$, and then the system (1) reduces to a second-order ODE $\ddot{x}(t) = F(x(t))$ without highly oscillatory solutions. Denote by $v = \dot{x}$ and then the energy of this dynamic is given by

$$E(x, v) = \frac{1}{2} |v|^2 + U(x), \quad (2)$$

which is exactly conserved along the solutions, i.e.

$$E(x(t), v(t)) = E(x(0), v(0)) \text{ for any } t \in [0, T].$$

We note further that with $p = v - \frac{1}{2\varepsilon} \tilde{B}x$, the equation (1) can be transformed into a Hamiltonian system with the non-separable Hamiltonian

$$H(x, p) = \frac{1}{2} \left| p + \frac{1}{2\varepsilon} \tilde{B}x \right|^2 + U(x). \quad (3)$$

Hamiltonian systems with highly oscillatory solutions frequently occur in physics and engineering such as charged-particle dynamics, Vlasov equations, classical and quantum mechanics, and molecular dynamics. Their numerical computation contains numerous enduring challenges. In the recent few decades, *geometric numerical integration* also called as structure-preserving algorithm for differential equations has received more and more attention. This kind of algorithms is designed to respect the structural invariants and geometry of the considered system. This idea has been used by many researchers to derive different structure-preserving algorithms (see, e.g. [6, 9, 16, 28]). For the Hamiltonian system (3), there are two remarkable features: the symplecticity of its flow and the conservation of the Hamiltonian. Consequently, for a numerical algorithm, these two features should be respected as much as possible in the spirit of geometric numerical integration.

One typical example of system (1) is Vlasov equations or charged-particle dynamics in a strong and uniform magnetic field, which has been studied by many researchers. In practice, the numerical methods used to treat this system can be summarized in the following three categories.

a) The primitive numerical methods usually depend on the knowledge of certain other characteristics of the solution besides high-frequency oscillation and structure preservation such as the Boris method [1] as well as its further researches [11, 22]. This method does not perform well for highly oscillatory systems and cannot preserve any structure of the system.

b) Some recent methods are devoted to the structure preservation such as the volume-preserving algorithms [18], symplectic methods [17, 24], symmetric methods [12] and energy-preserving methods [2, 20, 21, 23]. In [13], the long-time near-conservation property of a variational integrator was analyzed under the condition $0 < \varepsilon \ll 1$. Very recently, some integrators with large stepsize and their long term behaviour were studied in [14] for charged-particle dynamics. All of these methods can preserve or nearly preserve some structure of the considered system. However, these methods mentioned above do not pay attention to the high-frequency oscillation, and then the convergence of these methods is not uniformly accurate for ε . Their error constant usually increases when ε decreases.

c) Accuracy is often an important consideration for highly oscillatory systems over long-time intervals. Some new methods with uniform accuracy for ε have been proposed and analysed recently. The authors in [15] improved asymptotic behaviour of the Boris method and derived a filtered Boris algorithm under a maximal ordering scaling. Some multiscale schemes have been proposed such

as the asymptotic preserving schemes [7, 8] and the uniformly accurate schemes [3, 4]. Although these powerful numerical methods have very good performance in accuracy, structure (nearly) preservation usually cannot be achieved.

Based on the above points, a natural question to ask is whether one can design a numerical method for (1) such that it has uniform error bound for ε and can exactly preserve some structure simultaneously. A kind of energy-preserving method without convergent analysis was given in [25]. It will be shown in this paper that this method has a uniform error bound which has not been studied in [25]. Very recently, the authors in [27] presented some splitting methods with first-order uniform error bound in x and energy or volume preservation. However, only first-order methods are proposed there and higher-order ones with energy or other structure preservation have not been investigated. A numerical method combining high-order uniform error bound and structure preservation has more challenges and importance.

In this paper, we will derive two kinds of algorithms to preserve the symplecticity and energy, respectively. For symplectic algorithms, their near energy conservation over long times will be analysed. Moreover, all the structure-preserving algorithms will be shown to have second-order uniform error bound for $0 < \varepsilon \ll 1$ in x . Meanwhile, an algorithm with first-order uniform error bound in both x and v will be proposed. The remainder of this paper is organised as follows. In Section 2, we formulate two kinds of algorithms. The main results of these algorithms are given in Section 3 and a numerical experiment is carried out there to numerically show the performance of the algorithms. The proofs of the main results are presented in Sections 4-6 one by one. The last section includes some concluding remarks.

2 Numerical algorithms

Before deriving effective algorithms for the system (1), we first present the implicit expression of its exact solution as follows.

Theorem 2.1 (See [15].) *The exact solution of system (1) can be expressed as*

$$\begin{aligned} x(t_n + h) &= x(t_n) + h\varphi_1(h\Omega)v(t_n) + h^2 \int_0^1 (1-\tau)\varphi_1((1-\tau)h\Omega)F(x(t_n + h\tau))d\tau, \\ v(t_n + h) &= \varphi_0(h\Omega)v(t_n) + h \int_0^1 \varphi_0((1-\tau)h\Omega)F(x(t_n + h\tau))d\tau, \end{aligned} \quad (4)$$

where $\Omega = \frac{1}{\varepsilon}\tilde{B}$, h is a stepsize, $t_n = nh$ and the φ -functions are defined by (see [19])

$$\varphi_0(z) = e^z, \quad \varphi_k(z) = \int_0^1 e^{(1-\sigma)z} \frac{\sigma^{k-1}}{(k-1)!} d\sigma, \quad k = 1, 2, \dots$$

In what follows, we present two kinds of algorithms which will correspond to symplectic algorithms and energy-preserving algorithms, respectively.

Algorithm 2.2 *By denoting the numerical solution $x_n \approx x(t_n)$, $v_n \approx v(t_n)$ with $n = 0, 1, \dots$, an s -stage adaptive exponential algorithm applied with stepsize h is defined by:*

$$\begin{aligned} X_i &= x_n + c_i h \varphi_1(c_i h \Omega) v_n + h^2 \sum_{j=1}^s \alpha_{ij} (h \Omega) F(X_j), \quad i = 1, 2, \dots, s, \\ x_{n+1} &= x_n + h \varphi_1(h \Omega) v_n + h^2 \sum_{i=1}^s \beta_i (h \Omega) F(X_i), \\ v_{n+1} &= \varphi_0(h \Omega) v_n + h \sum_{i=1}^s \gamma_i (h \Omega) F(X_i), \end{aligned} \quad (5)$$

where $\alpha_{ij}(h\Omega), \beta_i(h\Omega), \gamma_i(h\Omega)$ are bounded functions of $h\Omega$. As some practical examples, we present five explicit algorithms.

For constant $F \equiv F_0 \in \mathbb{R}^3$, the variation-of-constants formula (4) reads

$$\begin{aligned} x(t_n + h) &= x(t_n) + h\varphi_1(h\Omega)v(t_n) + h^2\varphi_2(h\Omega)F_0, \\ v(t_n + h) &= \varphi_0(h\Omega)v(t_n) + h\varphi_1(h\Omega)F_0. \end{aligned}$$

Based on this, we consider the following algorithm

$$\begin{aligned} x_{n+1} &= x_n + h\varphi_1(h\Omega)v_n + h^2\varphi_2(h\Omega)F(x_n), \\ v_{n+1} &= \varphi_0(h\Omega)v_n + h\varphi_1(h\Omega)F(x_n). \end{aligned}$$

which means that in (5)

$$s = 1, \quad c_1 = 0, \quad \alpha_{11} = 0, \quad \beta_1 = \varphi_2(h\Omega), \quad \gamma_1 = \varphi_1(h\Omega).$$

This method is referred to M1. This method can be verified to be non-symmetric. We modify it to be a symmetric method by considering

$$\begin{aligned} s = 2, \quad c_1 = 0, \quad c_2 = 1, \quad \alpha_{11} = \alpha_{12} = \alpha_{22} = 0, \\ \beta_1 = \varphi_2(h\Omega), \quad \beta_2 = 0, \quad \gamma_1 = \frac{\varphi_2(h\Omega)}{\varphi_1(-h\Omega)}, \quad \gamma_2 = \frac{2e^{h\Omega}\varphi_2(-h\Omega)}{\varphi_1(h\Omega)}, \end{aligned}$$

and denote this method as M2.

It is noted that the next three methods are formulated based on the conditions (18) of symplecticity given below. The coefficients are obtained by considering the s -stage adaptive exponential algorithm (5) with the coefficients for $i = 1, \dots, s$, $j = 1, \dots, i$,

$$\alpha_{ij} = a_{ij}(c_i - c_j)\varphi_1((c_i - c_j)h\Omega), \quad \beta_i = b_i(1 - c_i)\varphi_1((1 - c_i)h\Omega), \quad \gamma_i = b_i e^{(1-c_i)h\Omega}, \quad (6)$$

where $c = (c_1, \dots, c_s)$, $b = (b_1, \dots, b_s)$ and $A = (a_{ij})_{s \times s}$ are coefficients of an s -stage diagonal implicit RK method. It can be checked easily that if this RK method is chosen as a symplectic method, then the corresponding coefficients (6) satisfy the conditions (18). We omit the details of calculations for brevity. We first consider

$$s = 1, \quad c_1 = \frac{1}{2}, \quad b_1 = 1.$$

The adaptive exponential algorithm whose coefficients are given by this choice and (6) is denoted by SM1. For $s = 2$, choosing

$$c_1 = 0, \quad c_2 = 1, \quad a_{21} = \frac{1}{2}, \quad \beta_1 = \frac{1}{2}, \quad b_2 = 1$$

yields another method, which is called as SM2. If we consider

$$c_1 = \frac{1}{4}, \quad c_2 = \frac{3}{4}, \quad a_{21} = \frac{1}{2}, \quad b_1 = b_2 = \frac{1}{2},$$

then the corresponding method is referred to SM3.

The following algorithm is devoted to the energy-preserving methods which are designed based on the variation-of-constants formula (4) and the idea of continuous-stage methods.

Algorithm 2.3 An s -degree continuous-stage adaptive exponential algorithm applied with stepsize h is defined by

$$\begin{aligned} X_\tau &= x_n + hC_\tau(h\Omega)v_n + h^2 \int_0^1 A_{\tau\sigma}(h\Omega)F(X_\sigma)d\sigma, \quad 0 \leq \tau \leq 1, \\ x_{n+1} &= x_n + h\varphi_1(h\Omega)v_n + h^2 \int_0^1 \bar{B}_\tau(h\Omega)F(X_\tau)d\tau, \\ v_{n+1} &= \varphi_0(h\Omega)v_n + h \int_0^1 B_\tau(h\Omega)F(X_\tau)d\tau, \end{aligned} \tag{7}$$

where X_τ is a polynomial of degree s with respect to τ satisfying $X_0 = x_n$, $X_1 = x_{n+1}$. C_τ , \bar{B}_τ , B_τ and $A_{\tau,\sigma}$ are polynomials which depend on $h\Omega$. The $C_\tau(h\Omega)$ satisfies $C_{c_i}(h\Omega) = c_i\varphi_1(c_i h\Omega)$, where c_i for $i = 1, \dots, s+1$ are the fitting nodes, and one of them is required to be one.

As an illustrative example, we consider $s = 1$, $c_1 = 0$, $c_2 = 1$ and choose

$$C_\tau = (1 - \tau)I + \tau\varphi_1(h\Omega), \quad A_{\tau\sigma} = \tau\varphi_2(h\Omega), \quad \bar{B}_\tau = \varphi_2(h\Omega), \quad B_\tau = \varphi_1(h\Omega).$$

This obtained algorithm can be rewritten as

$$\begin{aligned} x_{n+1} &= x_n + h\varphi_1(h\Omega)v_n + h^2\varphi_2(h\Omega) \int_0^1 F(x_n + \sigma(x_{n+1} - x_n))d\sigma, \\ v_{n+1} &= \varphi_0(h\Omega)v_n + h\varphi_1(h\Omega) \int_0^1 F(x_n + \sigma(x_{n+1} - x_n))d\sigma, \end{aligned} \tag{8}$$

which is denoted by *EM1*.

Remark 2.4 It is noted that *EM1* has been given in [25] and it was shown to be energy-preserving. However, its convergence has not been studied there. In this paper, we will analyse the convergence of each algorithm. It will be shown that *M1* has a first-order uniform error bound in both x and v and the others are of order two and have a uniform convergence in x for $0 < \varepsilon \ll 1$. In contrast, many classical methods such as Euler methods, Runge-Kutta (-Nyström) methods often show non-uniform error bounds in both x and v , where the error constant is usually proportional to $1/\varepsilon^k$ for some $k > 0$.

Remark 2.5 It is remarked that the following integrators for solving (1) has been given in [15]

$$\begin{aligned} x_{n+1} &= x_n + h\varphi_1(h\Omega)v_n + \frac{1}{2}h^2\Psi(h\Omega)F_n, \\ v_{n+1} &= \varphi_0(h\Omega)v_n + \frac{1}{2}h(\Psi_0(h\Omega)F_n + \Psi_1(h\Omega)F_{n+1}), \end{aligned}$$

where $F_n = F(x_n)$ and Ψ, Ψ_0, Ψ_1 are matrix-valued and bounded functions of $h\Omega$ satisfying $\Psi(0) = \Psi_0(0) = \Psi_1(0) = 1$. However, only convergence is researched there and the structure preservation such as symplecticity or energy conservation has not been discussed.

3 Main results and a numerical test

3.1 Main results

The main results of this paper are given by the following four theorems. The first three theorems are about structure preservations and the last one concerns uniform error bound.

Theorem 3.1 (Symplecticity of SM1-SM3.) Consider the methods SM1-SM3 where $p_{n+1} = v_{n+1} - \frac{1}{2\varepsilon} \tilde{B}x_{n+1}$. In this case, for the non-separable Hamiltonian (3), the map $(x_n, p_n) \rightarrow (x_{n+1}, p_{n+1})$ determined by these methods is symplectic.

Theorem 3.2 (Energy preservation of EM1 [25].) The method EM1 preserves the energy (2) exactly, i.e. $E(x_{n+1}, v_{n+1}) = E(x_n, v_n)$, $n = 0, 1, \dots$.

Theorem 3.3 (Long time energy conservation of M2 and SM1-SM3.) Consider the following assumptions.

- It is assumed that the initial values x_0 and $v_0 := \dot{x}_0$ are bounded such that the energy E is bounded independently of ε along the solution.
- Suppose that the considered numerical solution stays in a compact set.
- A lower bound on the stepsize $h/\varepsilon \geq c_0 > 0$ is required.
- Assume that the numerical non-resonance condition is true

$$|\sin(\frac{h}{2}(k \cdot \tilde{\Omega}))| \geq c\sqrt{h} \text{ for } k \in \mathbb{Z}^l \setminus \mathcal{M} \text{ with } |k| \leq N$$

for some $N \geq 2$ and $c > 0$. The notations used here are referred to the last part of Section 5.

For the methods M2 and SM1-SM3, it holds that

$$|E(x_n, v_n) - E(x_0, v_0)| \leq Ch \tag{9}$$

for $0 \leq nh \leq h^{-N+1}$. The constant C is independent of n, h, ε , but depends on N, T and the constants in the assumptions.

Remark 3.4 It is noted that M1 does not have the above energy conservation property. The reason is that it is not a symmetric method. It will be seen from the proof given in Section 5 that symmetry plays an important role in the analysis.

Theorem 3.5 (Convergence.) For the methods M1-M2 and the energy-preserving method EM1, under the condition that $h \leq C_1\varepsilon$, the global errors are bounded by

$$\text{M1: } |x_n - x(t_n)| + |v_n - v(t_n)| \lesssim h, \tag{10a}$$

$$\text{M2 and EM1: } |x_n - x(t_n)| \lesssim h^2, |v_n - v(t_n)| \lesssim h^2/\varepsilon, \tag{10b}$$

where $0 < nh \leq T$. Here we denote $A \lesssim B$ for $A \leq CB$ with a generic constant $C > 0$ independent of h or n or ε but depends on T and C_1 .

For the symplectic methods SM1-SM3, under the conditions of Theorem 3.3, the global errors are

$$\text{SM1-SM3: } |x_n - x(t_n)| \lesssim h^2, |v_n - v(t_n)| \lesssim h^2/\varepsilon, \tag{11}$$

where the error constants are independent of n, h, ε , but depend on T and the constants in the assumptions of Theorem 3.3.

Methods	Symplecticity	Symmetry	Energy property	Convergence
M1	No	No	No	$ x_n - x(t_n) \lesssim h$ $ v_n - v(t_n) \lesssim h$
M2	No	Yes	Near conservation	$ x_n - x(t_n) \lesssim h^2$ $ v_n - v(t_n) \lesssim h^2/\varepsilon$
SM1	Yes	Yes	Near conservation	$ x_n - x(t_n) \lesssim h^2$ $ v_n - v(t_n) \lesssim h^2/\varepsilon$
SM2	Yes	Yes	Near conservation	$ x_n - x(t_n) \lesssim h^2$ $ v_n - v(t_n) \lesssim h^2/\varepsilon$
SM3	Yes	Yes	Near conservation	$ x_n - x(t_n) \lesssim h^2$ $ v_n - v(t_n) \lesssim h^2/\varepsilon$
EM1	No	Yes	Exact conservation	$ x_n - x(t_n) \lesssim h^2$ $ v_n - v(t_n) \lesssim h^2/\varepsilon$

Table 1: Properties of the obtained methods.

For the six methods presented in this paper, concerning the symmetry [16], it is easy to check that all of them except M1 are symmetric. Their properties are summarized in Table 1. The main observation of the paper is that M1 has a first uniform (in ε) error bound in both x and v and symmetric or symplectic or energy-preserving methods show second-order uniform error bound in x . Moreover, SM1 can preserve the energy exactly and symplectic or symmetric methods have a good near conservations of energy over long times. All of these observations will be numerically illustrated by a test given below.

3.2 A numerical test

As an illustrative numerical experiment, we consider the charged particle system of [11] with an additional factor $1/\varepsilon$ and a constant magnetic field. The system can be expressed by (1) with $d = 3$, where the potential $U(x) = x_1^3 - x_2^3 + x_1^4/5 + x_2^4 + x_3^4$ and $\tilde{B} = \begin{pmatrix} 0 & 0.2 & 0.2 \\ -0.2 & 0 & 1 \\ -0.2 & -1 & 0 \end{pmatrix}$. The initial values are chosen as $x(0) = (0.6, 1, -1)^\top$ and $v(0) = (-1, 0.5, 0.6)^\top$.

3.2.1 Energy conservation

We take $\varepsilon = 0.05, 0.005$ and apply our six methods as well as the symplectic Euler method (denoted by SE) to this problem on $[0, 100000]$ with $h = \varepsilon$. The standard fixed point iteration is used for EM1 and we set 10^{-16} as the error tolerance and 10 as the maximum number of iterations. The relative errors $ERR := (E(x_n, v_n) - E(x_0, v_0))/E(x_0, v_0)$ of the energy are displayed in Figs. 1-3. We do not show the figure if the error is too large. According to these results, we have the following observations. M2 and SM1-SM3 (Figs. 1-2) have near energy conservation over long times, EM1 preserves the energy very well (Fig. 3) but SE and M1 show a bad energy conservation (Fig. 1).

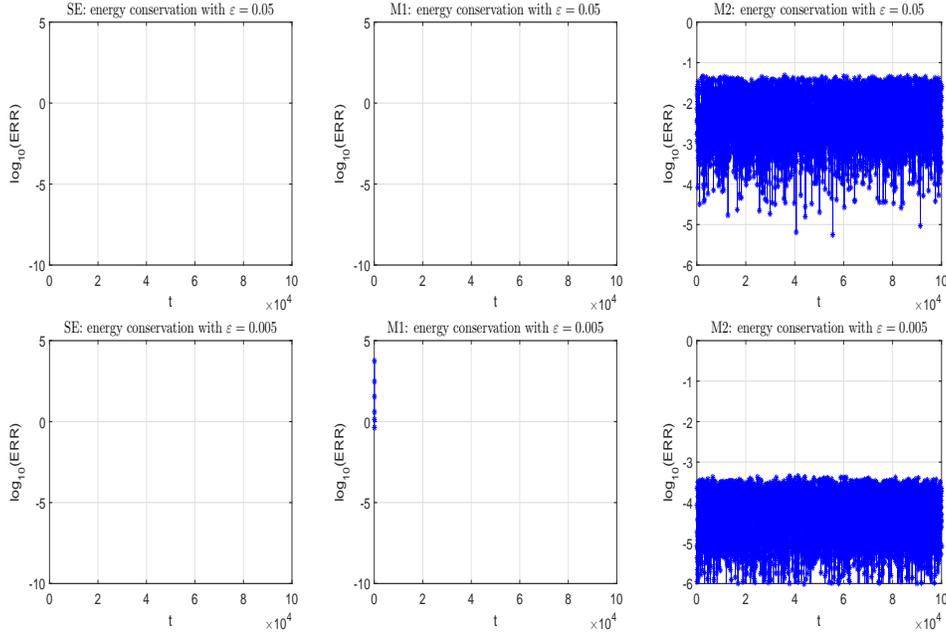


Figure 1: The relative energy errors (ERR) against t for SE and M1-M2.

3.2.2 Convergence

For displaying the results of convergence, the problem is solved on $[0, 1]$ with $h = 1/2^i$ for $i = 6, \dots, 12$. The global errors $errx := \frac{|x_n - x(t_n)|}{|x(t_n)|}$, $errv := \frac{|v_n - v(t_n)|}{|v(t_n)|}$ for different ϵ are shown in Figs. 4-6, respectively. It is noted that we use the result of standard ODE45 method in MATLAB with an absolute and relative tolerance equal to 10^{-12} as the true solution. It follows from these results that M1 has a uniform first-order convergence in both x and v as stated by (10a). The other methods only have a uniform second-order error bounds in x as given in Theorem 3.5.

3.2.3 Efficiency

In order to illustrate the efficiency of the proposed methods, we solve this system till $T = 10$. The efficiency of each method (the error $err2$ versus the CPU time) is displayed in Fig. 7. This test is conducted by MATLAB on a laptop ThinkPad (CPU: Intel (R) Core (TM) i7-10510U CPU @ 2.30 GHz, Memory: 8 GB, Os: Microsoft Windows 10 with 64bit). It can be observed that the computational cost of the new methods is cheap compared with the symplectic Euler method (Fig. 7).

3.2.4 Resonance instability

Finally, we show the resonance instability of the proposed methods. This is done by fixing $\epsilon = 1/2^{10}$ and showing the errors at $T = 1$ against h/ϵ in Fig. 8. It can be observed that M1 gives a very good result but other methods have a good behavior for values of h/ϵ except integral multiples of

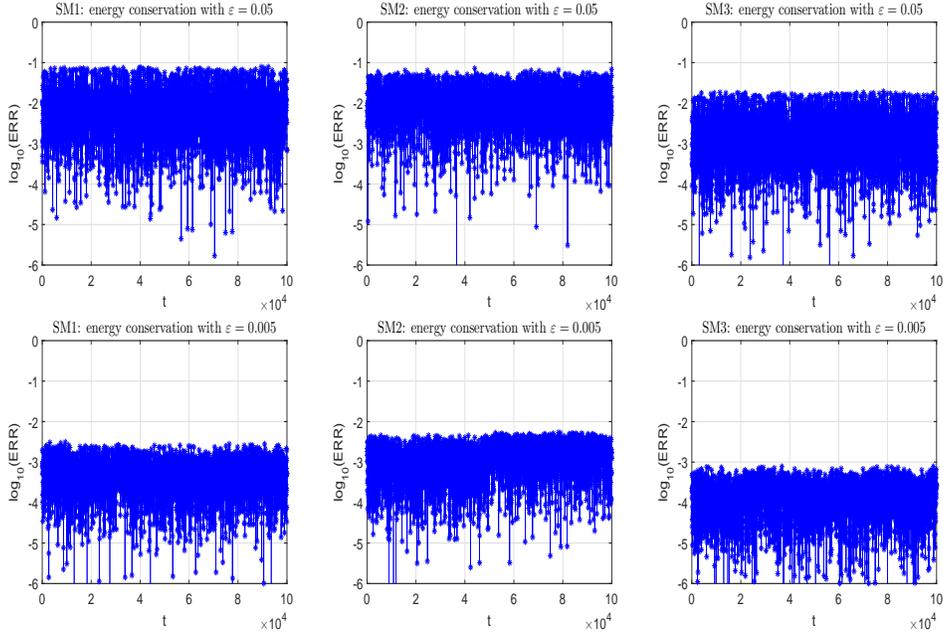


Figure 2: The relative energy errors (ERR) against t for symplectic SM1-SM3.

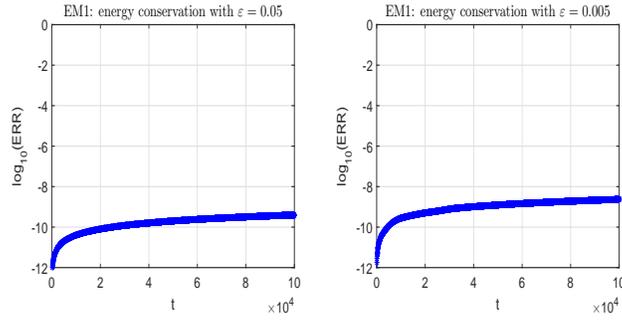


Figure 3: The relative energy errors (ERR) against t for energy-preserving EM1.

π . SM3 shows a not uniform result close to 4π , other methods M2, SM1, SM2 and EM1 close to even multiples of π . This means that SM3 appears more robust near stepsize resonances and other methods behave very similar away from stepsize resonances.

In the following three sections, we will prove Theorems 3.1, 3.3-3.5, respectively. In each proof, we will firstly consider $d = 3$ for brevity and then show that how to extend the analysis to other d with some necessary modifications.

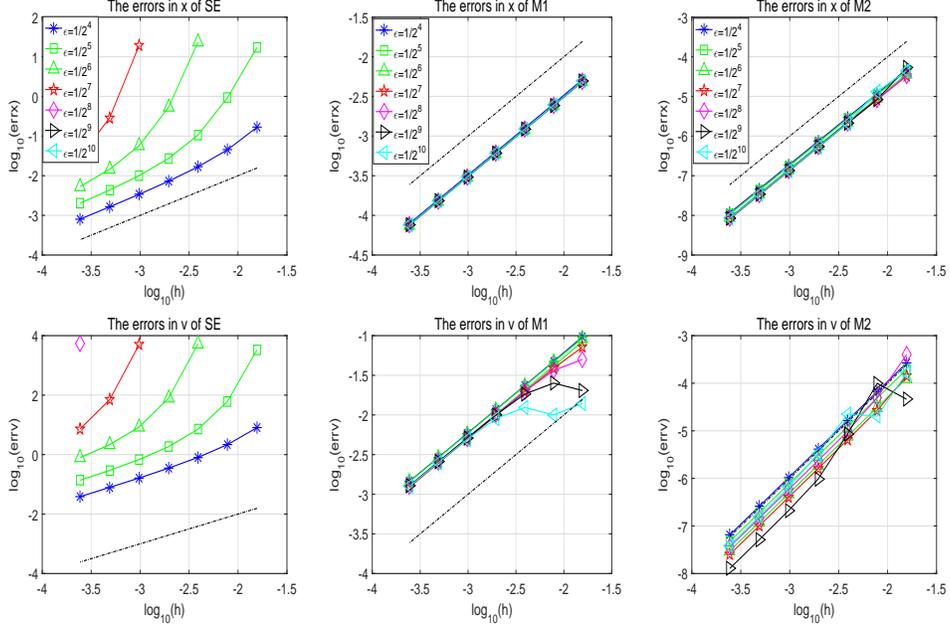


Figure 4: The errors in x (errx) and v (errv) against h for SE and M1-M2 (the slope of the dotted line for SE and M1 is one and for M2 is two).

4 Proof of symplecticity (Theorem 3.1)

• Transformed system and methods.

Due to the skew-symmetric matrix \tilde{B} , it is clear that there exists a unitary matrix P and a diagonal matrix Λ such that $\tilde{B} = P\Lambda P^H$, where $\Lambda = \text{diag}(-\|\tilde{B}\|i, 0, \|\tilde{B}\|i)$. With the linear change of variable

$$\tilde{x}(t) = P^H x(t), \quad \tilde{v}(t) = P^H v(t), \quad (12)$$

the system (1) can be rewritten as

$$\frac{d}{dt} \begin{pmatrix} \tilde{x} \\ \tilde{v} \end{pmatrix} = \begin{pmatrix} 0 & I \\ 0 & \tilde{\Omega}i \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{v} \end{pmatrix} + \begin{pmatrix} 0 \\ \tilde{F}(\tilde{x}) \end{pmatrix}, \quad \begin{pmatrix} \tilde{x}_0 \\ \tilde{v}_0 \end{pmatrix} = \begin{pmatrix} P^H x_0 \\ P^H \dot{x}_0 \end{pmatrix}, \quad (13)$$

where $\tilde{\Omega} = \text{diag}(-\tilde{\omega}, 0, \tilde{\omega})$, $\tilde{\omega} = \frac{\|\tilde{B}\|}{\epsilon}$, and $\tilde{F}(\tilde{x}) = P^H F(P\tilde{x}) = -\nabla_{\tilde{x}} U(P\tilde{x})$. In this paper, we denote the vector x by $x = (x^{-1}, x^0, x^1)^\top$ and the same notation is used for all the vectors in \mathbb{R}^3 or \mathbb{C}^3 . According to (12) and the property of the unitary matrix P , one has that

$$\tilde{x}^{-1} = \overline{(\tilde{x}^1)}, \quad \tilde{v}^{-1} = \overline{(\tilde{v}^1)}, \quad \tilde{x}^0, \tilde{v}^0 \in \mathbb{R}. \quad (14)$$

The energy of this transformed system (13) is given by

$$E(x, v) = \frac{1}{2} |P\tilde{v}|^2 + U(P\tilde{x}) = \frac{1}{2} |\tilde{v}|^2 + U(P\tilde{x}) := \tilde{E}(\tilde{x}, \tilde{v}).$$

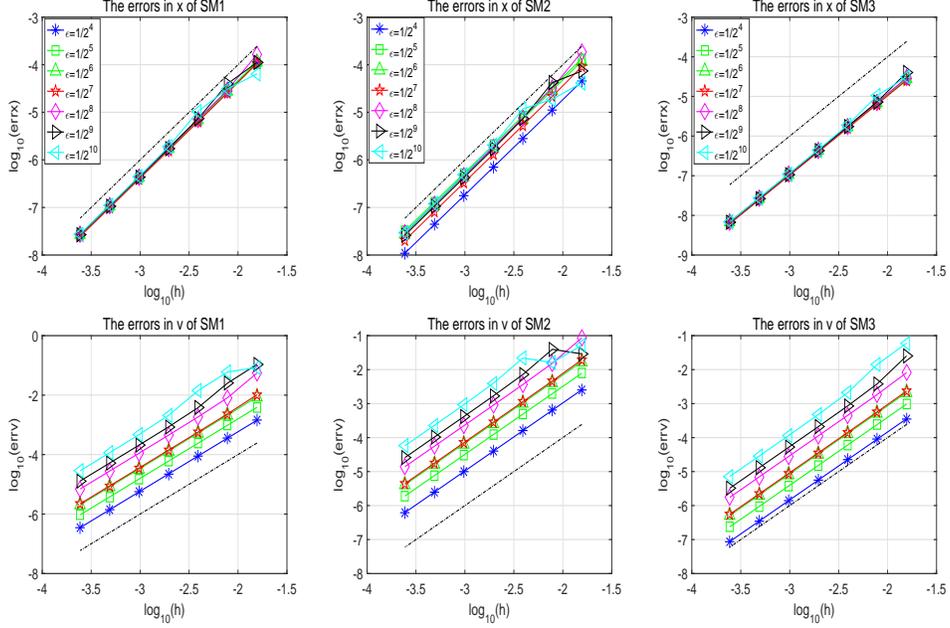


Figure 5: The errors in x (errx) and v (errv) against h for symplectic SM1-SM3 (the slope of the dotted line is two).

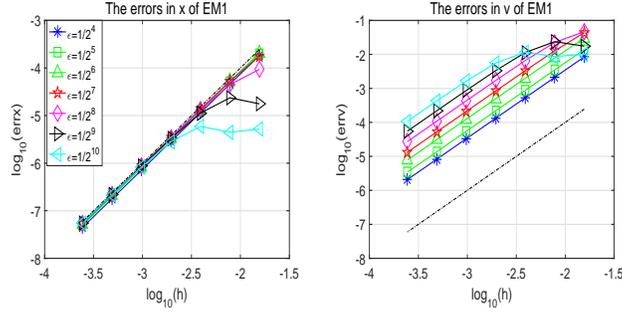


Figure 6: The errors in x (errx) and v (errv) against h for energy-preserving EM1 (the slope of the dotted line is two).

For this transformed system, we can modify the schemes of SM1-SM3 accordingly. For example, the scheme (5) has a transformed form for (13)

$$\begin{aligned}
 \tilde{X}_i &= \tilde{x}_n + c_i h \varphi_1(c_i h \tilde{\Omega} i) \tilde{v}_n + h^2 \sum_{j=1}^s \alpha_{ij}(h \tilde{\Omega} i) \tilde{F}(\tilde{X}_j), \quad i = 1, 2, \dots, s, \\
 \tilde{x}_{n+1} &= \tilde{x}_n + h \varphi_1(h \tilde{\Omega} i) \tilde{v}_n + h^2 \sum_{i=1}^s \beta_i(h \tilde{\Omega} i) \tilde{F}(\tilde{X}_i), \\
 \tilde{v}_{n+1} &= \varphi_0(h \tilde{\Omega} i) \tilde{v}_n + h \sum_{i=1}^s \gamma_i(h \tilde{\Omega} i) \tilde{F}(\tilde{X}_i).
 \end{aligned} \tag{15}$$

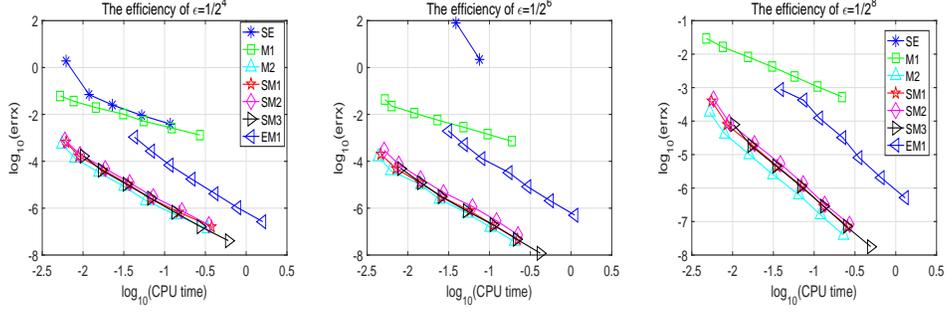


Figure 7: The uniform errors (err2) against CPU time.

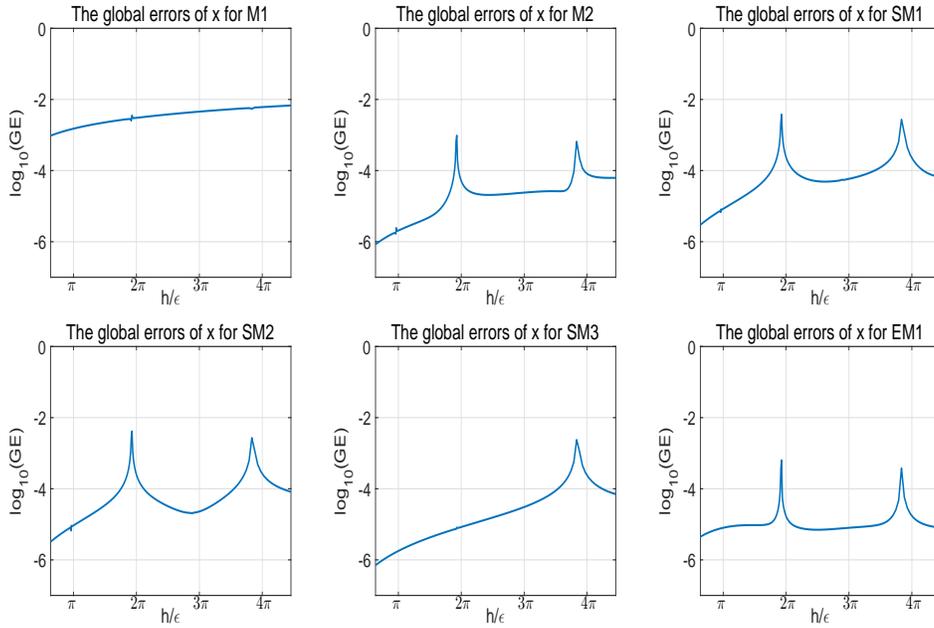


Figure 8: The global errors (GE) of x against h/ϵ .

Denote the transformed method (15) by

$$\begin{aligned}
 \tilde{X}_i^J &= \tilde{x}_n^J + c_i h \varphi_1(c_i h \tilde{\Omega}^J i) \tilde{v}_n^J + h^2 \sum_{j=1}^s \alpha_{ij}(h \tilde{\Omega}^J i) \tilde{F}_j^J, \quad i = 1, 2, \dots, s, \\
 \tilde{x}_{n+1}^J &= \tilde{x}_n^J + h \varphi_1(h \tilde{\Omega}^J i) \tilde{v}_n^J + h^2 \sum_{i=1}^s \beta_i(h \tilde{\Omega}^J i) \tilde{F}_i^J, \\
 \tilde{v}_{n+1}^J &= e^{h \tilde{\Omega}^J i} \tilde{v}_n^J + h \sum_{i=1}^s \gamma_i(h \tilde{\Omega}^J i) \tilde{F}_i^J,
 \end{aligned} \tag{16}$$

where the superscript index J for $J = -1, 0, 1$ denotes the $(J+2)$ th entry of a vector or a matrix and \tilde{F}_i^J denotes the $(J+2)$ th entry of $\tilde{F}(\tilde{X}_i)$. With the notation of differential 2-form, we need to

prove that (see [16])

$$\sum_{J=-1}^1 dx_{n+1}^J \wedge dp_{n+1}^J = \sum_{J=-1}^1 dx_n^J \wedge dp_n^J.$$

We compute

$$\begin{aligned} \sum_{J=-1}^1 dx_{n+1}^J \wedge dp_{n+1}^J &= \sum_{J=-1}^1 d\bar{x}_{n+1}^J \wedge d\bar{p}_{n+1}^J = \sum_{J=-1}^1 d(\bar{P}\bar{x}_{n+1})^J \wedge d(P\bar{p}_{n+1})^J \\ &= \sum_{J=-1}^1 \left(\sum_{i=-1}^1 (\bar{P}_{J+2, i+2} d\bar{x}_{n+1}^i) \right) \wedge \left(\sum_{k=-1}^1 (P_{J+2, k+2} d\bar{p}_{n+1}^k) \right) \\ &= \sum_{J=-1}^1 \sum_{i=-1}^1 \sum_{k=-1}^1 \bar{P}_{J+2, i+2} P_{J+2, k+2} (d\bar{x}_{n+1}^i \wedge d\bar{p}_{n+1}^k) \\ &= \sum_{i=-1}^1 d\bar{x}_{n+1}^i \wedge d\bar{p}_{n+1}^i = \sum_{J=-1}^1 d\bar{x}_{n+1}^J \wedge d\bar{p}_{n+1}^J, \end{aligned}$$

where $P^H P = I$ is used here. Similarly, one has $\sum_{J=-1}^1 dx_n^J \wedge dp_n^J = \sum_{J=-1}^1 d\bar{x}_n^J \wedge d\bar{p}_n^J$. Thus we only need to prove $\sum_{J=-1}^1 d\bar{x}_{n+1}^J \wedge d\bar{p}_{n+1}^J = \sum_{J=-1}^1 d\bar{x}_n^J \wedge d\bar{p}_n^J$, i.e.

$$\begin{aligned} &\sum_{J=-1}^1 d\bar{x}_{n+1}^J \wedge d\bar{v}_{n+1}^J - \frac{1}{2} \sum_{J=-1}^1 d\bar{x}_{n+1}^J \wedge d(\tilde{\Omega}^J i\bar{x}_{n+1}^J) \\ &= \sum_{J=-1}^1 d\bar{x}_n^J \wedge d\bar{v}_n^J - \frac{1}{2} \sum_{J=-1}^1 d\bar{x}_n^J \wedge d(\tilde{\Omega}^J i\bar{x}_n^J). \end{aligned} \tag{17}$$

• **Symplecticity of the transformed methods.**

In this part, we will prove that the result (17) is true if the following conditions are satisfied

$$\begin{aligned} \gamma_j(K) - K\beta_j(K) &= d_j I, \quad d_j \in \mathbb{C}, \\ \gamma_j(K)[\bar{\varphi}_1(K) - c_j \bar{\varphi}_1(c_j K)] &= \beta_j(K)[e^{-K} + K\bar{\varphi}_1(K) - c_j K\bar{\varphi}_1(c_j K)], \\ \bar{\beta}_i(K)\gamma_j(K) - \frac{1}{2}K\bar{\beta}_i(K)\beta_j(K) - \bar{\alpha}_{ji}(K)[\gamma_j(K) - K\beta_j(K)] \\ &= \beta_j(K)\bar{\gamma}_i(K) + \frac{1}{2}K\beta_j(K)\bar{\beta}_i(K) - \alpha_{ij}(K)[\bar{\gamma}_i(K) + K\bar{\beta}_i(K)], \end{aligned} \tag{18}$$

where $i, j = 1, 2, \dots, s$, and $K = h\tilde{\Omega}i$. Here $\bar{\varphi}_1$ denotes the conjugate of φ_1 and the same notation is used for other functions.

In view of the definition of differential 2-form (see [16]), it can be proved that $\overline{d\bar{x}_n^J \wedge d\bar{v}_n^J} = d\bar{x}_n^J \wedge d\bar{v}_n^J$ and $d\bar{x}_n^J \wedge d\bar{x}_n^J \in i\mathbb{R}$. In the light of the scheme (16) and the fact that any exterior

product \wedge appearing here is real, it is obtained that

$$\begin{aligned}
& d\tilde{x}_{n+1}^J \wedge d\tilde{v}_{n+1}^J - \frac{1}{2} d\tilde{x}_{n+1}^J \wedge d(\tilde{\Omega}^J i\tilde{x}_{n+1}^J) = d\tilde{x}_n^J \wedge d\tilde{v}_n^J - \frac{1}{2} d\tilde{x}_n^J \wedge d(\tilde{\Omega}^J i\tilde{x}_n^J) \\
& + h \sum_{j=1}^s [\gamma_j(K^J) - K^J \beta_j(K^J)] d\tilde{x}_n^J \wedge d\tilde{F}_j^J + [he^{K^J} \bar{\varphi}_1(K^J) - \frac{1}{2} h^2 \tilde{\Omega}^J i\bar{\varphi}_1(K^J) \varphi_1(K^J)] d\tilde{v}_n^J \wedge d\tilde{v}_n^J \\
& + h^2 \sum_{j=1}^s [\bar{\varphi}_1(K^J) \gamma_j(K^J) - \beta_j(K^J) e^{-K^J} - h\tilde{\Omega}^J i\bar{\varphi}_1(K^J) \beta_j(K^J)] d\tilde{v}_n^J \wedge d\tilde{F}_j^J \\
& + h^3 \sum_{i,j=1}^s [\bar{\beta}_i(K^J) \gamma_j(K^J) - \frac{1}{2} h\tilde{\Omega}^J i\bar{\beta}_i(K^J) \beta_j(K^J)] d\tilde{F}_i^J \wedge d\tilde{F}_j^J,
\end{aligned} \tag{19}$$

where the fact that $e^{K^J} - h\tilde{\Omega}^J i\varphi_1(K^J) = I$ is used here.

On the other hand, from the first s equalities of (16), it follows that

$$d\tilde{x}_n^J = d\tilde{X}_i^J - c_i h \varphi_1(c_i K^J) d\tilde{v}_n^J - h^2 \sum_{j=1}^s \alpha_{ij}(K^J) d\tilde{F}_j^J, \quad i = 1, 2, \dots, s.$$

We then obtain

$$d\tilde{x}_n^J \wedge d\tilde{F}_j^J = d\tilde{X}_j^J \wedge d\tilde{F}_j^J - c_j h \bar{\varphi}_1(c_j K^J) d\tilde{v}_n^J \wedge d\tilde{F}_j^J - h^2 \sum_{i=1}^s \bar{\alpha}_{ji}(K^J) d\tilde{F}_i^J \wedge d\tilde{F}_j^J, \quad j = 1, 2, \dots, s.$$

Inserting this into (19) and summing over all J yields

$$\begin{aligned}
\sum_{J=-1}^1 d\tilde{x}_{n+1}^J \wedge d\tilde{v}_{n+1}^J - \frac{1}{2} \sum_{J=-1}^1 d\tilde{x}_{n+1}^J \wedge d(\tilde{\Omega}^J i\tilde{x}_{n+1}^J) &= \sum_{J=-1}^1 d\tilde{x}_n^J \wedge d\tilde{v}_n^J - \frac{1}{2} \sum_{J=-1}^1 d\tilde{x}_n^J \wedge d(\tilde{\Omega}^J i\tilde{x}_n^J) \\
&+ h \sum_{j=1}^s \sum_{J=-1}^1 [\gamma_j(K^J) - K^J \beta_j(K^J)] d\tilde{X}_j^J \wedge d\tilde{F}_j^J
\end{aligned} \tag{20a}$$

$$+ h \sum_{J=-1}^1 [e^{K^J} \bar{\varphi}_1(K^J) - \frac{1}{2} K^J \bar{\varphi}_1(K^J) \varphi_1(K^J)] d\tilde{v}_n^J \wedge d\tilde{v}_n^J \tag{20b}$$

$$+ h^2 \sum_{j=1}^s \sum_{J=-1}^1 \left[\bar{\varphi}_1(K^J) \gamma_j(K^J) - \beta_j(K^J) e^{-K^J} - K^J \bar{\varphi}_1(K^J) \beta_j(K^J) \right] d\tilde{v}_n^J \wedge d\tilde{F}_j^J \tag{20c}$$

$$- c_j \bar{\varphi}_1(c_j K^J) [\gamma_j(K^J) - K^J \beta_j(K^J)] d\tilde{v}_n^J \wedge d\tilde{F}_j^J \tag{20d}$$

$$+ h^3 \sum_{i,j=1}^s \sum_{J=-1}^1 \left[\bar{\beta}_i(K^J) \gamma_j(K^J) - \frac{1}{2} h\tilde{\Omega}^J i\bar{\beta}_i(K^J) \beta_j(K^J) \right] d\tilde{F}_i^J \wedge d\tilde{F}_j^J \tag{20e}$$

$$- \bar{\alpha}_{ji}(K^J) [\gamma_j(K^J) - K^J \beta_j(K^J)] d\tilde{F}_i^J \wedge d\tilde{F}_j^J. \tag{20f}$$

◦ Prove that (20a) = 0.

Based on the first s conditions of (18), $\tilde{F}(\tilde{x}) = -\nabla_{\tilde{x}}U(P\tilde{x})$ and (17), it can be verified that $d\tilde{X}_j^J \wedge d\tilde{F}_j^J = dX_j^J \wedge dF_j^J$. Thus, one has

$$\begin{aligned} & \sum_{J=-1}^1 [\gamma_j(K^J) - K^J \beta_j(K^J)] d\tilde{X}_j^J \wedge d\tilde{F}_j^J \\ = & d_j \sum_{J=-1}^1 d\tilde{X}_j^J \wedge d\tilde{F}_j^J = d_j \sum_{J=-1}^1 dX_j^J \wedge dF_j^J = -d_j \sum_{J=-1}^1 dF_j^J \wedge dX_j^J \\ = & -d_j \sum_{J=-1}^1 \left(\frac{\partial F_j^J(X_j)}{\partial x^I} dX_j^I \right) \wedge dX_j^J = -d_j \sum_{J,I=-1}^1 \left(-\frac{\partial^2 U(Px)}{\partial x^J \partial x^I} \right) dX_j^I \wedge dX_j^J = 0. \end{aligned}$$

◦ Prove that (20b) = 0.

Using the property of \tilde{v}_n , we have

$$d\tilde{v}_n^{-1} \wedge d\tilde{v}_n^{-1} = -d\tilde{v}_n^{-1} \wedge d\tilde{v}_n^{-1}, \quad d\tilde{v}_n^0 \wedge d\tilde{v}_n^0 = 0,$$

and

$$e^{K^1} \bar{\varphi}_1(K^1) - \frac{1}{2} K^1 \bar{\varphi}_1(K^1) \varphi_1(K^1) = e^{K^{-1}} \bar{\varphi}_1(K^{-1}) - \frac{1}{2} K^{-1} \bar{\varphi}_1(K^{-1}) \varphi_1(K^{-1}).$$

Therefore, it follows that

$$\sum_{J=-1}^1 [e^{K^J} \bar{\varphi}_1(K^J) - \frac{1}{2} K^J \bar{\varphi}_1(K^J) \varphi_1(K^J)] d\tilde{v}_n^J \wedge d\tilde{v}_n^J = 0.$$

◦ Prove that (20c)-(20f) = 0.

In the light of all the identities after the previous s ones in (18), the last two terms (20c)-(20f) vanish.

The results stated above leads to (17). Then, it can be verified straightforwardly that the coefficients of SM1-SM3 satisfy (18). Therefore, these methods are symplectic.

• **Extension of the proof to other d .**

For a general $d \geq 2$, since \tilde{B} is skew-symmetric, there exists a unitary matrix P and a diagonal matrix Λ such that $\tilde{B} = P\Lambda P^H$, where

$$\Lambda = \begin{cases} \text{diag}(-\tilde{\omega}_l i, \dots, -\tilde{\omega}_1 i, 0, \tilde{\omega}_1 i, \dots, \tilde{\omega}_l i), & d = 2l + 1, \\ \text{diag}(-\tilde{\omega}_l i, \dots, -\tilde{\omega}_1 i, \tilde{\omega}_1 i, \dots, \tilde{\omega}_l i), & d = 2l. \end{cases} \quad (21)$$

For both cases, the above proof can be extended without any difficulty.

5 Proof of long-time energy conservation (Theorem 3.3)

In this section, we will show the long time near-conservation of energy along SM2 algorithm. We first derive modulated Fourier expansion (see, e.g. [10, 13, 15, 26]) with sufficient many terms for SM2. Then one almost-invariant of the expansion is studied and based on which the long-time near conservation is confirmed. The proof of other methods can be given by modifying the operators $\mathcal{L}(hD), \hat{\mathcal{L}}(hD)$ (24) and following the way given below.

• **Reformulation of SM2.**

Using symmetry, the algorithm SM2 can be expressed in a two-step form

$$\begin{cases} x_{n+1} - 2x_n + x_{n-1} = h(\varphi_1(h\Omega) - \varphi_1(-h\Omega))v_n + \frac{1}{2}h^2(\varphi_1(h\Omega) + \varphi_1(-h\Omega))F_n, \\ x_{n+1} - x_{n-1} = h(\varphi_1(h\Omega) + \varphi_1(-h\Omega))v_n + \frac{1}{2}h^2(\varphi_1(h\Omega) - \varphi_1(-h\Omega))F_n, \end{cases} \quad (22)$$

with $F_n := F(x_n)$, which yields that

$$\alpha(h\Omega) \frac{x_{n+1} - 2x_n + x_{n-1}}{h^2} = \beta(h\Omega)\Omega \frac{x_{n+1} - x_{n-1}}{2h} + \gamma(h\Omega)F_n,$$

where $\alpha(\xi) = \frac{\xi}{\varphi_1(\xi) - \varphi_1(-\xi)}$, $\beta(\xi) = \frac{2}{\varphi_1(\xi) + \varphi_1(-\xi)}$, $\gamma(\xi) = \xi \frac{2\varphi_1(\xi)\varphi_1(-\xi)}{\varphi_1^2(\xi) - \varphi_1^2(-\xi)}$.

For the transformed system (13), it becomes

$$\tilde{\alpha}(h\tilde{\Omega}) \frac{\tilde{x}_{n+1} - 2\tilde{x}_n + \tilde{x}_{n-1}}{h^2} = \tilde{\beta}(h\tilde{\Omega})i\tilde{\Omega} \frac{\tilde{x}_{n+1} - \tilde{x}_{n-1}}{2h} + \tilde{\gamma}(h\tilde{\Omega})\tilde{F}_n, \quad (23)$$

where the coefficient functions are given by $\tilde{\alpha}(\xi) = \frac{1}{\text{sinc}^2(\frac{\xi}{2})}$, $\tilde{\beta}(\xi) = \frac{1}{\text{sinc}(\xi)}$, $\tilde{\gamma}(\xi) = \xi \csc(\xi)$ with $\text{sinc}(\xi) = \sin(\xi)/\xi$.

Define the operators

$$\mathcal{L}(hD) = \frac{1}{2h \text{sinc}(h\tilde{\Omega})}(e^{hD} - e^{-hD}), \quad \hat{\mathcal{L}}(hD) = \tilde{\alpha}(h\tilde{\Omega}) \frac{e^{hD} - 2 + e^{-hD}}{h^2} - \tilde{\beta}(h\tilde{\Omega})i\tilde{\Omega} \frac{e^{hD} - e^{-hD}}{2h}, \quad (24)$$

where D is the differential operator. The Taylor expansions of the operator $\mathcal{L}(hD)$ are

$$\begin{aligned} \mathcal{L}(hD) &= \tilde{\Omega} \csc(h\tilde{\Omega})(hD) + \frac{1}{6}\tilde{\Omega} \csc(h\tilde{\Omega})(hD)^3 + \dots, \\ \mathcal{L}(hD + ih\tilde{\omega}) &= \text{idia}\left(\tilde{\omega}, \frac{\sin(h\tilde{\omega})}{h}, \tilde{\omega}\right) + \text{diag}\left(\tilde{\omega} \cot(h\tilde{\omega}), \frac{\cos(h\tilde{\omega})}{h}, \tilde{\omega} \cot(h\tilde{\omega})\right)(hD) + \dots, \\ \mathcal{L}(hD - ih\tilde{\omega}) &= -\text{idia}\left(\tilde{\omega}, \frac{\sin(h\tilde{\omega})}{h}, \tilde{\omega}\right) + \text{diag}\left(\tilde{\omega} \cot(h\tilde{\omega}), \frac{\cos(h\tilde{\omega})}{h}, \tilde{\omega} \cot(h\tilde{\omega})\right)(hD) + \dots, \\ \mathcal{L}(hD + ikh\tilde{\omega}) &= i \sin(kh\tilde{\omega})\tilde{\Omega} \csc(h\tilde{\Omega}) + \cos(kh\tilde{\omega})\tilde{\Omega} \csc(h\tilde{\Omega}) + \dots, \quad \text{where } |k| > 1. \end{aligned}$$

The operator $\hat{\mathcal{L}}(hD)$ can be expressed in its Taylor expansion as

$$\begin{aligned} \hat{\mathcal{L}}(hD) &= -\tilde{\Omega}^2 \csc(h\tilde{\Omega})(ihD) - \frac{1}{4}\tilde{\Omega}^2 \csc^2\left(\frac{1}{2}h\tilde{\Omega}\right)(ihD)^2 + \dots, \\ \hat{\mathcal{L}}(hD + ih\tilde{\omega}) &= \text{diag}\left(-2\tilde{\omega}^2, \frac{2(\cos(h\tilde{\omega}) - 1)}{h^2}, 0\right) \\ &\quad + \text{diag}\left(\tilde{\omega}^2(2 \cot(h\tilde{\omega}) + \csc(h\tilde{\omega})), \frac{2 \sin(h\tilde{\omega})}{h^2}, \tilde{\omega}^2 \csc(h\tilde{\Omega})\right)(ihD) + \dots, \\ \hat{\mathcal{L}}(hD - ih\tilde{\omega}) &= \text{diag}\left(0, \frac{2(\cos(h\tilde{\omega}) - 1)}{h^2}, -2\tilde{\omega}^2\right) \\ &\quad - \text{diag}\left(\tilde{\omega}^2 \csc(h\tilde{\Omega}), \frac{2 \sin(h\tilde{\omega})}{h^2}, \tilde{\omega}^2(2 \cot(h\tilde{\omega}) + \csc(h\tilde{\omega}))\right)(ihD) + \dots, \\ \hat{\mathcal{L}}(hD + ikh\tilde{\omega}) &= 2 \sin\left(\frac{1}{2}hk\tilde{\omega}\right)\tilde{\Omega}^2 \csc\left(\frac{1}{2}h\tilde{\Omega}\right) \csc(h\tilde{\Omega}) \sin\left(\frac{1}{2}h(\tilde{\Omega} - k\tilde{\omega}I)\right) \\ &\quad - \sin\left(\frac{1}{2}h(\tilde{\Omega} - 2k\tilde{\omega})\right)\tilde{\Omega}^2 \csc\left(\frac{1}{2}h\tilde{\Omega}\right) \csc(h\tilde{\Omega})(ihD) + \dots, \quad \text{where } |k| > 1. \end{aligned}$$

• **Modulated Fourier expansion.**

We first present the modulated Fourier expansion of the numerical result \tilde{x}_n and \tilde{v}_n for solving the transformed system (13).

We will look for smooth coefficient functions $\tilde{\zeta}_k$ and $\tilde{\eta}_k$ such that for $t = nh$, the functions

$$\tilde{x}_h(t) = \sum_{|k| < N} e^{ik\tilde{\omega}t} \tilde{\zeta}_k(t) + \tilde{R}_{h,N}(t), \quad \tilde{v}_h(t) = \sum_{|k| < N} e^{ik\tilde{\omega}t} \tilde{\eta}_k(t) + \tilde{S}_{h,N}(t) \quad (25)$$

yield a small defect \tilde{R}, \tilde{S} when they are inserted into the numerical scheme (23).

◦ Construction of the coefficients functions.

Inserting the first expansion of (25) into the two-step form (23), expanding the nonlinear function into its Taylor series and comparing the coefficients of $e^{ik\tilde{\omega}t}$, we obtain

$$\begin{aligned} \hat{\mathcal{L}}(hD)\tilde{\zeta}_0 &= \tilde{\gamma}(h\tilde{\Omega}) \left(\tilde{F}(\tilde{\zeta}_0) + \sum_{s(\alpha)=0} \frac{1}{m!} \tilde{F}^{(m)}(\tilde{\zeta}_0)(\tilde{\zeta})_\alpha \right), \\ \hat{\mathcal{L}}(hD + ihk\tilde{\omega})\tilde{\zeta}_k &= \tilde{\gamma}(h\tilde{\Omega}) \sum_{s(\alpha)=k} \frac{1}{m!} \tilde{F}^{(m)}(\tilde{\zeta}_0)(\tilde{\zeta})_\alpha, \quad |k| > 0, \end{aligned} \quad (26)$$

where the sum ranges over $m \geq 0$, $s(\alpha) = \sum_{j=1}^m \alpha_j$ with $\alpha = (\alpha_1, \dots, \alpha_m)$ and $0 < |\alpha_i| < N$, and $(\tilde{\zeta})_\alpha$ is an abbreviation for $(\tilde{\zeta}_{\alpha_1}, \dots, \tilde{\zeta}_{\alpha_m})$. This formula gives the modulation system for the coefficients $\tilde{\zeta}_k$ of the modulated Fourier expansion. Choosing the dominating terms and considering the Taylor expansion of $\hat{\mathcal{L}}$ given above, the following ansatz of $\tilde{\zeta}_k$ can be obtained:

$$\begin{aligned} \dot{\tilde{\zeta}}_0^{\pm 1} &= \frac{-h^2 \tilde{\omega} A(h\tilde{\omega})}{8i \sin^2(\frac{1}{2}h\tilde{\omega})} (\mathcal{G}^{\pm 10}(\cdot) + \dots), & \ddot{\tilde{\zeta}}_0^0 &= \mathcal{G}^{00}(\cdot) + \dots, \\ \tilde{\zeta}_1^{-1} &= \frac{h^3 \tilde{\omega} A(h\tilde{\omega})}{-16 \sin^2(\frac{1}{2}h\tilde{\Omega}) \sin(h\tilde{\omega})} (\mathcal{F}_1^{-10}(\cdot) + \dots), & \tilde{\zeta}_1^0 &= \frac{h^2}{-4 \sin^2(h\tilde{\omega}/2)} (\mathcal{F}_1^{00}(\cdot) + \dots), \\ \dot{\tilde{\zeta}}_1^1 &= \frac{h^2 \tilde{\omega} A(h\tilde{\omega})}{8i \sin^2(\frac{1}{2}h\tilde{\omega})} (\mathcal{F}_1^{10}(\cdot) + \dots), & \dot{\tilde{\zeta}}_1^{-1} &= \frac{h^2 \tilde{\omega} A(h\tilde{\omega})}{-8i \sin^2(\frac{1}{2}h\tilde{\omega})} (\mathcal{F}_1^{-10}(\cdot) + \dots), \\ \tilde{\zeta}_{-1}^0 &= \frac{h^2}{-4 \sin^2(h\tilde{\omega}/2)} (\mathcal{F}_{-1}^{00}(\cdot) + \dots), & \tilde{\zeta}_{-1}^1 &= \frac{h^3 \tilde{\omega} A(h\tilde{\omega})}{-16 \sin^2(\frac{1}{2}h\tilde{\Omega}) \sin(h\tilde{\omega})} (\mathcal{F}_{-1}^{10}(\cdot) + \dots), \\ \tilde{\zeta}_k &= \frac{h^3 \tilde{\Omega} A(h\tilde{\Omega})}{16 \sin(\frac{1}{2}h\tilde{\Omega}) \sin(\frac{1}{2}h(\tilde{\Omega} - k\tilde{\omega}I)) \sin(\frac{1}{2}hk\tilde{\omega}I)} (\mathcal{F}_k^0(\cdot) + \dots) \quad \text{for } |k| > 1, \end{aligned} \quad (27)$$

where the dots stand for power series in \sqrt{h} and $A(h\tilde{\omega}) = 2 \text{sinc}^2(\frac{1}{2}h\tilde{\omega})$. In this paper we truncate the ansatz after the $\mathcal{O}(h^{N+1})$ terms. On the basis of the second formula of (22), one has

$$\begin{aligned} \tilde{v}_n &= \frac{1}{h(\varphi_1(ih\tilde{\Omega}) + \varphi_1(-ih\tilde{\Omega}))} (\tilde{x}_{n+1} - \tilde{x}_{n-1}) - \frac{1}{2} h^2 \frac{\varphi_1(ih\tilde{\Omega}) - \varphi_1(-ih\tilde{\Omega})}{h(\varphi_1(ih\tilde{\Omega}) + \varphi_1(-ih\tilde{\Omega}))} \tilde{F}(\tilde{x}_n) \\ &= \frac{1}{2h \text{sinc}(h\tilde{\Omega})} (\tilde{x}_{n+1} - \tilde{x}_{n-1}) - \frac{1}{2} ih \tan(\frac{h}{2}\tilde{\Omega}) \tilde{F}(\tilde{x}_n). \end{aligned} \quad (28)$$

Inserting (25) into (28), expanding the nonlinear function into its Taylor series and comparing the coefficients of $e^{ik\tilde{\omega}t}$, one arrives

$$\begin{aligned} \tilde{\eta}_0 &= \mathcal{L}(hD)\tilde{\zeta}_0 - \frac{1}{2} ih \tan(\frac{h}{2}\tilde{\Omega}) \left(\tilde{F}(\tilde{\zeta}_0) + \sum_{s(\alpha)=0} \frac{1}{m!} \tilde{F}^{(m)}(\tilde{\zeta}_0)(\tilde{\zeta})_\alpha \right), \\ \tilde{\eta}_k &= \mathcal{L}(hD + ihk\tilde{\omega})\tilde{\zeta}_k - \frac{1}{2} ih \tan(\frac{h}{2}\tilde{\Omega}) \sum_{s(\alpha)=k} \frac{1}{m!} \tilde{F}^{(m)}(\tilde{\zeta}_0)(\tilde{\zeta})_\alpha, \quad |k| > 0. \end{aligned} \quad (29)$$

This formula gives the modulation system for the coefficients $\tilde{\eta}_k$ of the modulated Fourier expansion by the Taylor expansion of \mathcal{L} and by choosing the dominating terms.

◦ Initial values.

For the first-order and second-order differential equations appeared in (27), initial values are needed and we derive them as follows.

According to the conditions $\tilde{x}_h(0) = \tilde{x}_0$ and $\tilde{v}_h(0) = \tilde{v}_0$, we have

$$\begin{aligned}\tilde{x}_0^0 &= \tilde{\zeta}_0^0(0) + \mathcal{O}(\tilde{\omega}^{-1}), & \tilde{x}_0^{\pm 1} &= \tilde{\zeta}_0^{\pm 1}(0) + \mathcal{O}(\tilde{\omega}^{-1}), \\ \tilde{v}_0^0 &= \tilde{\eta}_0^0(0) + \mathcal{O}(\tilde{\omega}^{-1}) = \dot{\tilde{\zeta}}_0^0(0) + \mathcal{O}(\tilde{\omega}^{-1}), \\ \tilde{v}_0^1 &= \tilde{\eta}_0^1(0) + \tilde{\eta}_1^1(0) + \mathcal{O}(\tilde{\omega}^{-1}) = \dot{\tilde{\zeta}}_0^1(0) + i\tilde{\omega}\tilde{\zeta}_1^1(0) + \mathcal{O}(\tilde{\omega}^{-1}), \\ \tilde{v}_0^{-1} &= \tilde{\eta}_0^{-1}(0) + \tilde{\eta}_{-1}^{-1}(0) + \mathcal{O}(\tilde{\omega}^{-1}) = \dot{\tilde{\zeta}}_0^{-1}(0) - i\tilde{\omega}\tilde{\zeta}_{-1}^{-1}(0) + \mathcal{O}(\tilde{\omega}^{-1}).\end{aligned}\tag{30}$$

Thus the initial values $\tilde{\zeta}_0^0(0) = \mathcal{O}(1)$ and $\dot{\tilde{\zeta}}_0^0(0) = \mathcal{O}(1)$ can be derived by considering the first and third formulae, respectively. According to the second equation of (30), one gets the initial value $\tilde{\zeta}_0^{\pm 1}(0) = \mathcal{O}(1)$. It follows from the fourth formula that $\tilde{\zeta}_1^1(0) = \frac{1}{i\tilde{\omega}}(\tilde{v}_0^1 - \dot{\tilde{\zeta}}_0^1(0) + \mathcal{O}(h)) = \mathcal{O}(\tilde{\omega}^{-1})$, and likewise one has $\tilde{\zeta}_{-1}^{-1}(0) = \mathcal{O}(\tilde{\omega}^{-1})$.

◦ Bounds of the coefficient functions.

With the ansatz (27), we achieve the bounds

$$\begin{aligned}\dot{\tilde{\zeta}}_0^{\pm 1} &= \mathcal{O}(h), & \ddot{\tilde{\zeta}}_0^0 &= \mathcal{O}(1), & \dot{\tilde{\zeta}}_1^1 &= \mathcal{O}(h), & \dot{\tilde{\zeta}}_{-1}^{-1} &= \mathcal{O}(h), \\ \tilde{\zeta}_1^{-1} &= \mathcal{O}(h^{\frac{5}{2}}), & \tilde{\zeta}_1^0 &= \mathcal{O}(h^2), & \tilde{\zeta}_{-1}^0 &= \mathcal{O}(h^2), & \tilde{\zeta}_{-1}^1 &= \mathcal{O}(h^{\frac{5}{2}}).\end{aligned}$$

According to the initial values stated above, the bounds

$$\tilde{\zeta}_0^{\pm 1} = \mathcal{O}(1), \quad \tilde{\zeta}_0^0 = \mathcal{O}(1), \quad \tilde{\zeta}_1^1 = \mathcal{O}(h), \quad \tilde{\zeta}_{-1}^{-1} = \mathcal{O}(h),$$

are obtained. Moreover, we have the following results for coefficient functions $\tilde{\eta}$

$$\begin{aligned}\tilde{\eta}_0^0 &= \dot{\tilde{\zeta}}_0^0 + \mathcal{O}(h), & \tilde{\eta}_0^{\pm 1} &= \frac{h\tilde{\omega}}{\sin(h\tilde{\omega})}\dot{\tilde{\zeta}}_0^{\pm 1} + \mathcal{O}(h), \\ \tilde{\eta}_{\pm 1}^0 &= i\tilde{\omega}\operatorname{sinc}(h\tilde{\omega})\tilde{\zeta}_0^{\pm 1} + \mathcal{O}(h), & \tilde{\eta}_1^{\pm 1} &= i\tilde{\omega}\tilde{\zeta}_1^{\pm 1} + \mathcal{O}(h), & \tilde{\eta}_{-1}^{\pm 1} &= -i\tilde{\omega}\tilde{\zeta}_{-1}^{\pm 1} + \mathcal{O}(h).\end{aligned}\tag{31}$$

A further result is true

$$\tilde{\zeta}_k = \mathcal{O}(h^{|k|+1}), \quad \tilde{\eta}_k = \mathcal{O}(h^{|k|}) \quad \text{for } |k| > 1.$$

◦ Defect.

Define

$$\begin{aligned}\delta_1(t+h) &= \tilde{x}_h(t+h) - \tilde{x}_h(t) - h\varphi_1(ih\tilde{\Omega})\tilde{v}_h(t) - \frac{1}{2}h^2\varphi_1(ih\tilde{\Omega})\tilde{F}(\tilde{x}_h(t)), \\ \delta_2(t+h) &= \tilde{v}_h(t+h) - e^{ih\tilde{\Omega}}\tilde{v}_h(t) - \frac{1}{2}h\varphi_0(ih\tilde{\Omega})\tilde{F}(\tilde{x}_h(t)) - \frac{1}{2}h\tilde{F}(\tilde{x}_h(t+h))\end{aligned}$$

for $t = nh$. Considering the two-step formulation, it is clear that $\delta_1(t+h) + \delta_1(t-h) = \mathcal{O}(h^4)$. According to the choice for the initial values, we obtain $\delta_1(0) = \mathcal{O}(h^{N+2})$. Therefore, one has

$\delta_1(t) = \mathcal{O}(h^{N+2}) + \mathcal{O}(th^{N+1})$. Using this result and (28), we have $\delta_2 = \mathcal{O}(h^N)$. Then let $\tilde{R}_n = \tilde{x}_n - \tilde{x}_h(t)$ and $\tilde{S}_n = \tilde{v}_n - \tilde{v}_h(t)$. With the scheme of SM2, the error recursion is obtained as follows:

$$\begin{pmatrix} \tilde{R}_{n+1} \\ \tilde{S}_{n+1} \end{pmatrix} = \begin{pmatrix} I & h\varphi_1(ih\tilde{\Omega}) \\ 0 & e^{ih\tilde{\Omega}} \end{pmatrix} \begin{pmatrix} \tilde{R}_n \\ \tilde{S}_n \end{pmatrix} + \frac{1}{2}h \begin{pmatrix} h\varphi_1\Gamma_n\tilde{R}_n \\ \varphi_0\Gamma_n\tilde{R}_n + \Gamma_{n+1}\tilde{R}_{n+1} \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix},$$

where $\Gamma_n := \int_0^1 \tilde{F}_x(\tilde{x}_n + \tau\tilde{R}_n)d\tau$. Solving this recursion and the application of a discrete Gronwall inequality gives

$$\tilde{R}_{h,N}(t) = \mathcal{O}(t^2h^N), \quad \tilde{S}_{h,N}(t) = \mathcal{O}(t^2h^N/\varepsilon).$$

Using the relationships shown in (??), the numerical solution of SM2 admits the following modulated Fourier expansion

$$x_n = \sum_{|k|<N} e^{ik\tilde{\omega}t} \zeta_k(t) + \mathcal{O}(t^2h^N), \quad v_n = \sum_{|k|<N} e^{ik\tilde{\omega}t} \eta_k(t) + \mathcal{O}(t^2h^N/\varepsilon),$$

where $\zeta_k = P\tilde{\zeta}_k$ and $\eta_k = P\tilde{\eta}_k$. Moreover, we have $\zeta_{-k} = \overline{\zeta_k}$ and $\eta_{-k} = \overline{\eta_k}$.

• **An almost-invariant.**

Denote

$$\vec{\zeta} = (\tilde{\zeta}_{-N+1}, \dots, \tilde{\zeta}_{-1}, \tilde{\zeta}_0, \tilde{\zeta}_1, \dots, \tilde{\zeta}_{N-1}).$$

An almost-invariant of the modulated Fourier expansion (25) is given as follows.

According to the above analysis, it is deduced that

$$\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD)\tilde{x}_h = \tilde{F}(\tilde{x}_h) + \mathcal{O}(h^N),$$

where we use the denotations $\tilde{x}_h = \sum_{|k|<N} \tilde{x}_{h,k}$ with $\tilde{x}_{h,k} = e^{ik\tilde{\omega}t} \tilde{\zeta}_k$. Multiplication of this result with P yields

$$\begin{aligned} P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD)P^H P\tilde{x}_h &= P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD)P^H x_h \\ &= P\tilde{F}(\tilde{x}_h) + \mathcal{O}(h^{N+2}) = F(x_h) + \mathcal{O}(h^N), \end{aligned}$$

where $x_h = \sum_{|k|<N} x_{h,k}$ with $x_{h,k} = e^{ik\tilde{\omega}t} \zeta_k$. Rewrite the equation in terms of $x_{h,k}$ and then one has

$$P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD)P^H x_{h,k} = -\nabla_{x_{-k}} \mathcal{U}(\vec{x}) + \mathcal{O}(h^N),$$

where $\mathcal{U}(\vec{x})$ is defined as

$$\mathcal{U}(\vec{x}) = U(x_{h,0}) + \sum_{s(\alpha)=0} \frac{1}{m!} U^{(m)}(x_{h,0})(x_h)_\alpha$$

with $\vec{x} = (x_{h,-N+1}, \dots, x_{h,-1}, x_{h,0}, x_{h,1}, \dots, x_{h,N-1})$. Multiplying this equation with $(\dot{x}_{h,-k})^\top$ and summing up yields

$$\sum_{|k|<N} (\dot{x}_{h,-k})^\top P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD)P^H x_{h,k} + \frac{d}{dt} \mathcal{U}(\vec{x}) = \mathcal{O}(h^N).$$

Denoting $\vec{\zeta} = (\zeta_{-N+1}, \dots, \zeta_{-1}, \zeta_0, \zeta_1, \dots, \zeta_{N-1})$ and switching to the quantities ζ^k , we obtain

$$\begin{aligned}
\mathcal{O}(h^N) &= \sum_{|k| < N} (\dot{\zeta}_{-k} - ik\tilde{\omega}\zeta_{-k})^\top P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + ihk\tilde{\omega})P^H\zeta_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}) \\
&= \sum_{|k| < N} (\dot{\tilde{\zeta}}_k - ik\tilde{\omega}\tilde{\zeta}_k)^\top P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + ihk\tilde{\omega})P^H\zeta_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}) \\
&= \sum_{|k| < N} (\dot{\tilde{\zeta}}_k - ik\tilde{\omega}\tilde{\zeta}_k)^\top P^H P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + ihk\tilde{\omega})P^H\tilde{\zeta}_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}) \\
&= \sum_{|k| < N} (\dot{\tilde{\zeta}}_k - ik\tilde{\omega}\tilde{\zeta}_k)^\top \tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + ihk\tilde{\omega})\tilde{\zeta}_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}).
\end{aligned} \tag{32}$$

In what follows, we show that the right-hand side of (32) is the total derivative of an expression that depends only on $\tilde{\zeta}_k$ and derivatives thereof. Consider $k = 0$ and then it follows that

$$\hat{\mathcal{L}}(hD)\tilde{\zeta}_0 = ihM_1\dot{\tilde{\zeta}}_0 + h^2M_2\ddot{\tilde{\zeta}}_0 + ih^3M_3\ddot{\tilde{\zeta}}_0 + \dots, \text{ where } M_k \in \mathbb{R}^{3 \times 3} \text{ for } k = 1, 2, \dots$$

By the formulae given on p. 508 of [16], we know that $\text{Re}(\dot{\tilde{\zeta}}_0)^\top \hat{\mathcal{L}}(hD)\tilde{\zeta}_0$ is a total derivative. For $k \neq 0$, in the light of

$$\hat{\mathcal{L}}(hD + ihk\tilde{\omega})\tilde{\zeta}_k = N_0\tilde{\zeta}_k + ihN_1\dot{\tilde{\zeta}}_k + h^2N_2\ddot{\tilde{\zeta}}_k + ih^3N_3\ddot{\tilde{\zeta}}_k + \dots, \text{ where } N_k \in \mathbb{R}^{3 \times 3} \text{ for } k = 0, 1, \dots,$$

it is easy to check that $\text{Re}(\dot{\tilde{\zeta}}_k)^\top \tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + ihk\tilde{\omega})\tilde{\zeta}_k$ and $\text{Re}(ik\tilde{\omega}\tilde{\zeta}_k)^\top \tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + ihk\tilde{\omega})\tilde{\zeta}_k$ are both total derivatives. Therefore, there exists a function \mathcal{E} such that $\frac{d}{dt}\mathcal{E}[\vec{\zeta}](t) = \mathcal{O}(h^N)$. It follows from an integration that

$$\mathcal{E}[\vec{\zeta}](t) = \mathcal{E}[\vec{\zeta}](0) + \mathcal{O}(th^N). \tag{33}$$

On the basis of the previous analysis, the construction of \mathcal{E} is derived as follows

$$\begin{aligned}
\mathcal{E}[\vec{\zeta}](t_n) &= \frac{1}{2}(\dot{\tilde{\zeta}}_0)^\top \frac{2 \text{sinc}(\frac{1}{2}h\tilde{\Omega})}{\varphi_1(ih\tilde{\Omega})\varphi_1(-ih\tilde{\Omega}) + \varphi_1(-ih\tilde{\Omega})\varphi_1(ih\tilde{\Omega})} \dot{\tilde{\zeta}}_0 \\
&\quad + \frac{1}{2} \frac{\tilde{\omega}}{h} h\tilde{\omega} \frac{2 \text{sinc}^2(\frac{1}{2}h\tilde{\omega})}{\varphi_1(ih\tilde{\omega})\varphi_1(-ih\tilde{\omega}) + \varphi_1(-ih\tilde{\omega})\varphi_1(ih\tilde{\omega})} (|\tilde{\zeta}_1^1|^2 + |\tilde{\zeta}_{-1}^{-1}|^2) + \mathcal{U}(\vec{\zeta}) + \mathcal{O}(h^2) \\
&= \frac{1}{2} |\dot{\tilde{\zeta}}_0|^2 + \frac{1}{2} \tilde{\omega}^2 (|\tilde{\zeta}_1^1|^2 + |\tilde{\zeta}_{-1}^{-1}|^2) + U(P^H\tilde{\zeta}^0) + \mathcal{O}(h).
\end{aligned}$$

• **Long-time near-conservation.**

Considering the result of \mathcal{E} and the relationship (48) between $\tilde{\zeta}$ and $\tilde{\eta}$, we obtain

$$\begin{aligned}
\mathcal{E}[\vec{\zeta}](t_n) &= \frac{1}{2} |\dot{\tilde{\zeta}}_0|^2 + \frac{1}{2} \tilde{\omega}^2 (|\tilde{\zeta}_1^1|^2 + |\tilde{\zeta}_{-1}^{-1}|^2) + U(P^H\tilde{\zeta}^0) + \mathcal{O}(h) \\
&= \frac{1}{2} |\tilde{\eta}_0^0|^2 + \frac{1}{2} (|\tilde{\eta}_1^1|^2 + |\tilde{\eta}_{-1}^{-1}|^2) + U(P^H\tilde{\zeta}^0) + \mathcal{O}(h).
\end{aligned} \tag{34}$$

We are now in a position to show the long-time conservations of SM2.

In terms of the bounds of the coefficient functions, one arrives at

$$E(x_n, v_n) = \tilde{E}(\tilde{x}_n, \tilde{v}_n) = \frac{1}{2} (|\tilde{\eta}_0^0|^2 + |\tilde{\eta}_1^1|^2 + |\tilde{\eta}_{-1}^{-1}|^2) + U(P^H \tilde{\zeta}^0) + \mathcal{O}(h). \quad (35)$$

A comparison between (34) and (35) gives $\mathcal{E}[\tilde{\zeta}^{\vec{\zeta}}](t_n) = E(x_n, v_n) + \mathcal{O}(h)$. Based on (33) and this result, the statement (9) is easily obtained by considering $nh^N \leq 1$ and

$$\begin{aligned} E(x_n, v_n) &= \mathcal{E}[\tilde{\zeta}^{\vec{\zeta}}](t_n) + \mathcal{O}(h) = \mathcal{E}[\tilde{\zeta}^{\vec{\zeta}}](t_{n-1}) + \mathcal{O}(h^{N+1}) + \mathcal{O}(h) \\ &= \mathcal{E}[\tilde{\zeta}^{\vec{\zeta}}](t_{n-2}) + 2\mathcal{O}(h^{N+1}) + \mathcal{O}(h) = \dots \\ &= \mathcal{E}[\tilde{\zeta}^{\vec{\zeta}}](t_0) + n\mathcal{O}(h^{N+1}) + \mathcal{O}(h) = E(x_0, v_0) + \mathcal{O}(h). \end{aligned}$$

• **Extension of the proof to other d .**

According to the scheme (21) of Λ , the operators $\mathcal{L}(hD)$ and $\hat{\mathcal{L}}(hD)$ as well as their properties can be changed accordingly. The modulated Fourier expansions are modified as

$$\tilde{x}_h(t) = \sum_{k \in \mathcal{N}^*} e^{i(k \cdot \tilde{\Omega})t} \tilde{\zeta}_k(t) + \tilde{R}_{h,N}(t), \quad \tilde{v}_h(t) = \sum_{k \in \mathcal{N}^*} e^{i(k \cdot \tilde{\Omega})t} \tilde{\eta}_k(t) + \tilde{S}_{h,N}(t),$$

where $k = (k_1, \dots, k_l)$, $\tilde{\Omega} = (\tilde{\omega}_1, \dots, \tilde{\omega}_l)$, $k \cdot \tilde{\Omega} = k_1 \tilde{\omega}_1 + \dots + k_l \tilde{\omega}_l$. The set \mathcal{N}^* is defined as follows. For the resonance module $\mathcal{M} = \{k \in \mathbb{Z}^l : k \cdot \tilde{\Omega} = 0\}$, we let \mathcal{K} be a set of representatives of the equivalence classes in $\mathbb{Z}^l \setminus \mathcal{M}$ which are chosen such that for each $k \in \mathcal{K}$ the sum $|k|$ is minimal in the equivalence class $[k] = k + \mathcal{M}$, and that with $k \in \mathcal{K}$, also $-k \in \mathcal{K}$. We denote, for the positive integer N , $\mathcal{N} = \{k \in \mathcal{K} : |k| \leq N\}$, $\mathcal{N}^* = \mathcal{N} \setminus \{(0, \dots, 0)\}$. Then the almost-invariant can be modified accordingly and the long-time near conservation can be proved.

6 Proof of convergence (Theorem 3.5)

In this section, we discuss the convergence of the algorithms. The proof will be firstly given for M1-M2 and EM1 by using the averaging technique and then presented for SM1-SM3 by using modulated Fourier expansion.

6.1 Proof for M1-M2 and EM1

The proof will be given for EM1 and it can be adapted to M1-M2 easily.

• **Re-scaled system and methods.**

In order to establish the uniform error bounds, the strategy developed in [5, 27] will be used in the proof. This means that the time re-scaling $\tau := t/\varepsilon$ is considered and $\dot{q}(\tau) = \varepsilon \dot{x}(t)$, $\dot{w}(\tau) = \varepsilon \dot{v}(t)$, where the notations $q(\tau) := x(t)$, $w(\tau) := v(t)$ are used. Then the convergent analysis will be given for the following long-time problem

$$\begin{aligned} \dot{q}(\tau) &= \varepsilon w(\tau), \quad \dot{w}(\tau) = \tilde{B}w(\tau) + \varepsilon F(q(\tau)), \quad \tau \in [0, \frac{T}{\varepsilon}], \\ q(0) &= q_0 := x_0, \quad w(0) = w_0 := v_0, \end{aligned} \quad (36)$$

where \dot{q} (resp. \dot{w}) is referred to the derivative of q (resp. w) with respect to τ . The solution of this system satisfies $\|q\|_{L^\infty(0,T/\varepsilon)} + \|w\|_{L^\infty(0,T/\varepsilon)} \lesssim 1$ and for solving (36), the method EM1 becomes

$$\begin{aligned} q_{n+1} &= q_n + \varepsilon \Delta\tau \varphi_1(\Delta\tau \tilde{B}) w_n + \frac{\Delta\tau^2 \varepsilon^2}{2} \varphi_2(\Delta\tau \tilde{B}) \int_0^1 F(\rho q_n + (1-\rho)q_{n+1}) d\rho, \\ w_{n+1} &= e^{\Delta\tau \tilde{B}} w_n + \Delta\tau \varepsilon \varphi_1(\Delta\tau \tilde{B}) \int_0^1 F(\rho q_n + (1-\rho)q_{n+1}) d\rho, \quad 0 \leq n < \frac{T}{\varepsilon \Delta\tau}. \end{aligned} \quad (37)$$

where $\Delta\tau$ is the time step $\Delta\tau = \tau_{n+1} - \tau_n$ and $q_n \approx q(\tau_n)$, $w_n \approx w(\tau_n)$ is the numerical solution.

• **Local truncation errors.**

Based on (37), the local truncation errors ξ_n^q and ξ_n^w for $0 \leq n < \frac{T}{\varepsilon \Delta\tau}$ are defined as

$$\begin{aligned} q(\tau_{n+1}) &= q(\tau_n) + \Delta\tau \varepsilon \varphi_1(\Delta\tau \tilde{B}) w(\tau_n) \\ &\quad + \frac{\Delta\tau^2 \varepsilon^2}{2} \varphi_2(\Delta\tau \tilde{B}) \int_0^1 F(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})) d\rho + \xi_n^q, \\ w(\tau_{n+1}) &= e^{\Delta\tau \tilde{B}} w(\tau_n) + \Delta\tau \varepsilon \varphi_1(\Delta\tau \tilde{B}) \int_0^1 F(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})) d\rho + \xi_n^w. \end{aligned} \quad (38)$$

By this result and the variation-of-constants formula of (36), we compute

$$\begin{aligned} \xi_n^w &= \varepsilon \Delta\tau \int_0^1 e^{(1-\sigma)\Delta\tau \tilde{B}} F(q(\tau_n + \Delta\tau\sigma)) d\sigma \\ &\quad - \varepsilon \Delta\tau \varphi_1(\Delta\tau \tilde{B}) \int_0^1 F(q(\tau_n) + \sigma(q(\tau_{n+1}) - q(\tau_n))) d\sigma \\ &= \varepsilon \sum_{j=0}^1 \Delta\tau^{j+1} \varphi_{j+1}(h\tilde{B}) \hat{F}^{(j)}(\tau_n) - \varepsilon \Delta\tau \varphi_1(\Delta\tau \tilde{B}) F(q(\tau_n)) \\ &\quad - \varepsilon^2 \Delta\tau^2 \varphi_1(\Delta\tau \tilde{B}) \int_0^1 [\sigma \frac{\partial F}{\partial q}(q(\tau_n)) w(\tau_n)] d\sigma + \mathcal{O}(\varepsilon^2 \Delta\tau^3) \\ &= \varepsilon \Delta\tau^2 \varphi_2(\Delta\tau \tilde{B}) \hat{F}^{(1)}(\tau_n) - \frac{1}{2} \varepsilon^2 \Delta\tau^2 \varphi_1(\Delta\tau \tilde{B}) \frac{\partial F}{\partial q}(q(\tau_n)) w(\tau_n) + \mathcal{O}(\varepsilon^2 \Delta\tau^3), \end{aligned}$$

where $\hat{F}(\xi) = F(q(\xi))$ and $\hat{F}^{(j)}$ denotes the j th derivative of \hat{F} with respect to τ . By this definition, it follows that

$$\hat{F}^{(1)}(\tau_n) = \frac{\partial F}{\partial q}(q(\tau_n)) \frac{dq}{d\tau}(\tau_n) = \frac{\partial F}{\partial q}(q(\tau_n)) \varepsilon w(\tau_n).$$

Consequently, the local error becomes

$$\xi_n^w = \varepsilon^2 \Delta\tau^2 (\varphi_2(\Delta\tau \tilde{B}) - \frac{1}{2} \varphi_1(\Delta\tau \tilde{B})) \frac{\partial F}{\partial q}(q(\tau_n)) w(\tau_n) + \mathcal{O}(\varepsilon^2 \Delta\tau^3) = \mathcal{O}(\varepsilon^2 \Delta\tau^3), \quad (39)$$

where the result $\varphi_2(\Delta\tau \tilde{B}) - \frac{1}{2} \varphi_1(\Delta\tau \tilde{B}) = \mathcal{O}(\Delta\tau)$ is used here. Similarly, we obtain

$$\xi_n^q = \mathcal{O}(\varepsilon^3 \Delta\tau^3). \quad (40)$$

Remark 6.1 *It is noted that for M1, the local truncation errors are*

$$\xi_n^w = \mathcal{O}(\varepsilon^2 \Delta\tau^2), \quad \xi_n^q = \mathcal{O}(\varepsilon^2 \Delta\tau^2). \quad (41)$$

• **Error bound.**

In this part, we will first prove the boundedness of EM1: there exists a generic constant $\Delta\tau_0 > 0$ independent of ε and n , such that for $0 < \Delta\tau \leq \Delta\tau_0$, the following inequalities are true:

$$|q_n| \leq \|q\|_{L^\infty(0, T/\varepsilon)} + 1, \quad |w_n| \leq \|w\|_{L^\infty(0, T/\varepsilon)} + 1, \quad 0 \leq n \leq \frac{T}{\varepsilon\Delta\tau}. \quad (42)$$

For $n = 0$, (42) is obviously true. Then we assume (42) is true up to some $0 \leq m < \frac{T}{\varepsilon\Delta\tau}$, and we shall show that (42) holds for $m + 1$.

Define the error of the scheme

$$e_n^q := q(\tau_n) - q_n, \quad e_n^w := w(\tau_n) - w_n, \quad 0 \leq n < \frac{T}{\varepsilon\Delta\tau}.$$

For $n \leq m$, subtracting (38) from the scheme (37) leads to

$$e_{n+1}^q = e_n^q + \Delta\tau\varepsilon\varphi_1(\Delta\tau\tilde{B})e_n^w + \eta_n^q + \xi_n^q, \quad e_{n+1}^w = e^{\Delta\tau\tilde{B}}e_n^w + \eta_n^w + \xi_n^w, \quad 0 \leq n \leq m, \quad (43)$$

where we use the following notations

$$\begin{aligned} \eta_n^q &= \frac{\Delta\tau^2\varepsilon^2}{2}\varphi_2(\Delta\tau\tilde{B}) \int_0^1 [F(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})) - F(\rho q_n + (1-\rho)q_{n+1})] d\rho, \\ \eta_n^w &= \Delta\tau\varepsilon\varphi_1(\Delta\tau\tilde{B}) \int_0^1 [F(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})) - F(\rho q_n + (1-\rho)q_{n+1})] d\rho. \end{aligned}$$

From the induction assumption of the boundedness, it follows that

$$|\eta_n^q| \lesssim \Delta\tau^2\varepsilon^2 (|e_n^q| + |e_{n+1}^q|), \quad |\eta_n^w| \lesssim \Delta\tau\varepsilon (|e_n^q| + |e_{n+1}^q|), \quad 0 \leq n < m. \quad (44)$$

Taking the absolute value (Euclidean norm) on both sides of (43) and using (44), we have

$$|e_{n+1}^q| + |e_{n+1}^w| - |e_n^q| - |e_n^w| \lesssim \Delta\tau\varepsilon (|e_n^w| + |e_n^q| + |e_{n+1}^q|) + |\xi_n^q| + |\xi_n^w|, \quad 0 \leq n \leq m.$$

Summing them up for $0 \leq n \leq m$ gives

$$|e_{m+1}^q| + |e_{m+1}^w| \lesssim \Delta\tau\varepsilon \sum_{n=0}^m (|e_n^w| + |e_n^q| + |e_{n+1}^q|) + \sum_{n=0}^m (|\xi_n^q| + |\xi_n^w|).$$

In the light of the truncation errors in (39) and the fact that $m\Delta\tau\varepsilon \lesssim 1$, one has

$$|e_{m+1}^q| + |e_{m+1}^w| \lesssim \Delta\tau\varepsilon \sum_{n=0}^m (|e_n^w| + |e_n^q| + |e_{n+1}^q|) + \varepsilon\Delta\tau^2,$$

and then by Gronwall's inequality arrives at

$$|e_{m+1}^q| + |e_{m+1}^w| \lesssim \varepsilon\Delta\tau^2, \quad 0 \leq m < \frac{T}{\varepsilon\Delta\tau}. \quad (45)$$

Meanwhile, concerning

$$|q_{m+1}| \leq |q(\tau_{m+1})| + |e_{m+1}^q|, \quad |w_{m+1}| \leq |w(\tau_{m+1})| + |e_{m+1}^w|,$$

there exists a generic constant $\Delta\tau_0 > 0$ independent of ε and m , such that for $0 < \Delta\tau \leq \Delta\tau_0$, (42) holds for $m + 1$. This completes the induction.

Now we rewrite (43) as

$$e_{n+1}^q = e_n^q + \Delta\tau\varphi_1(\Delta\tau\tilde{B})(\varepsilon e_n^w) + \eta_n^q + \xi_n^q, \quad (\varepsilon e_{n+1}^w) = e^{\Delta\tau\tilde{B}}(\varepsilon e_n^w) + \varepsilon\eta_n^w + \varepsilon\xi_n^w.$$

Following the same way as stated above, it is arrived that

$$|e_{m+1}^q| + |\varepsilon e_{m+1}^w| \lesssim \varepsilon^2 \Delta\tau^2, \quad 0 \leq m < \frac{T}{\varepsilon\Delta\tau}. \quad (46)$$

Remark 6.2 We note that for M1, the global error given in (45) becomes

$$|e_{m+1}^q| + |e_{m+1}^w| \lesssim \varepsilon\Delta\tau, \quad 0 \leq m < \frac{T}{\varepsilon\Delta\tau},$$

which proves the result (10a) of M1.

• **Proof of the results for the methods applied to (1).**

By considering the grids in the t variable and τ variable, it is obtained that $h = \varepsilon\Delta\tau$. This shows that for the original system (1) and the re-scaled system (36), $x(t_n) = q(\tau_n)$ and $v(t_n) = w(\tau_n)$. Moreover, by comparing (8) with (37), we know that the numerical solution x_n, v_n of (8) is identical to q_n, w_n of (37). Therefore, the result (46) yields the uniform error bound in x given in (10b) and also shows the non-uniform error in v of (10b).

6.2 Proof for SM1-SM3

For SM1-SM3, the above proof cannot be applied since their local truncation errors will lose a factor of ε in (39) and (40). This motivates us to consider modulated Fourier expansions (see, e.g. [10, 13, 15, 16, 25]) for analysis in this part. The proof will be briefly shown for SM2 and it can be modified for the other two methods easily.

• **Decomposition of the numerical solution.** Now we turn back to the SM2 given in (23) and consider its modulated Fourier expansion (25). In order to derive the convergence, we need to explicitly present the results of $\tilde{\zeta}_k$ and $\tilde{\eta}_k$ with $|k| \leq 1$. In the light of (26) and the properties of $\hat{\mathcal{L}}(hD)$, we obtain

$$\begin{aligned} \dot{\tilde{\zeta}}_0^{\pm 1} &= \frac{-h^2\tilde{\omega}A(h\tilde{\omega})}{8i\sin^2(\frac{1}{2}h\tilde{\omega})} \left(\tilde{F}(\tilde{\zeta}_0) + \tilde{F}''(\tilde{\zeta}_0)(\tilde{\zeta}_1, \tilde{\zeta}_{-1}) \right)_{\pm 1}, & \ddot{\tilde{\zeta}}_0^0 &= \left(\tilde{F}(\tilde{\zeta}_0) + \tilde{F}''(\tilde{\zeta}_0)(\tilde{\zeta}_1, \tilde{\zeta}_{-1}) \right)_0, \\ \tilde{\zeta}_1^{-1} &= \frac{h^3\tilde{\omega}A(h\tilde{\omega})}{-16\sin^2(\frac{1}{2}h\tilde{\omega})\sin(h\tilde{\omega})} (\tilde{F}'(\tilde{\zeta}_0)\tilde{\zeta}_1)_{-1}, & \tilde{\zeta}_1^0 &= \frac{h^2}{-4\sin^2(h\tilde{\omega}/2)} (\tilde{F}'(\tilde{\zeta}_0)\tilde{\zeta}_1)_0, \\ \dot{\tilde{\zeta}}_1^1 &= \frac{h^2\tilde{\omega}A(h\tilde{\omega})}{8i\sin^2(\frac{1}{2}h\tilde{\omega})} (\tilde{F}'(\tilde{\zeta}_0)\tilde{\zeta}_1)_1, & \dot{\tilde{\zeta}}_{-1}^{-1} &= \frac{h^2\tilde{\omega}A(h\tilde{\omega})}{-8i\sin^2(\frac{1}{2}h\tilde{\omega})} (\tilde{F}'(\tilde{\zeta}_0)\tilde{\zeta}_{-1})_{-1}, \\ \tilde{\zeta}_{-1}^0 &= \frac{h^2}{-4\sin^2(h\tilde{\omega}/2)} (\tilde{F}'(\tilde{\zeta}_0)\tilde{\zeta}_{-1})_0, & \tilde{\zeta}_{-1}^1 &= \frac{h^3\tilde{\omega}A(h\tilde{\omega})}{-16\sin^2(\frac{1}{2}h\tilde{\omega})\sin(h\tilde{\omega})} (\tilde{F}'(\tilde{\zeta}_0)\tilde{\zeta}_{-1})_1. \end{aligned} \quad (47)$$

Then the following results

$$\begin{aligned} \tilde{\eta}_0^0 &= \dot{\tilde{\zeta}}_0^0 + \mathcal{O}(h), & \tilde{\eta}_0^{\pm 1} &= \frac{h\tilde{\omega}}{\sin(h\tilde{\omega})} \dot{\tilde{\zeta}}_0^{\pm 1} + \mathcal{O}(h), \\ \tilde{\eta}_{\pm 1}^0 &= i\tilde{\omega} \operatorname{sinc}(h\tilde{\omega}) \tilde{\zeta}_{\pm 1}^0 + \mathcal{O}(h), & \tilde{\eta}_1^{\pm 1} &= i\tilde{\omega} \tilde{\zeta}_1^{\pm 1} + \mathcal{O}\left(h \left| i \tan\left(\frac{h}{2}\tilde{\Omega}\right) \right|\right), \\ \tilde{\eta}_{-1}^{\pm 1} &= -i\tilde{\omega} \tilde{\zeta}_{-1}^{\pm 1} + \mathcal{O}\left(h \left| i \tan\left(\frac{h}{2}\tilde{\Omega}\right) \right|\right) \end{aligned} \quad (48)$$

are easily arrived by considering (29) as well as the property of $\mathcal{L}(hD)$.

• **Decomposition of the exact solution.** Following the result given in [15], the exact solution of (13) admits the following expansion

$$\tilde{x}(t) = \sum_{|k| \leq 1} e^{ik\tilde{\omega}t} \tilde{\mu}_k(t) + \tilde{d}_{\tilde{x}}(t), \quad \tilde{v}(t) = \sum_{|k| \leq 1} e^{ik\tilde{\omega}t} \tilde{\nu}_k(t) + \tilde{d}_{\tilde{v}}(t), \quad (49)$$

where the defects are bounded by $\tilde{d}_{\tilde{x}}(t) = \mathcal{O}(\tilde{\omega}^{-2})$, $\tilde{d}_{\tilde{v}}(t) = \mathcal{O}(\tilde{\omega}^{-2}/\varepsilon)$. The functions $\tilde{\mu}^k$ are given by

$$\begin{aligned} \dot{\tilde{\mu}}_0^{\pm 1} &= \frac{1}{\mp i\tilde{\omega}} (\tilde{F}_{\pm 1}(\tilde{\mu}^0) + \tilde{F}'_{\pm 1}(\tilde{\mu}^0)(\tilde{\mu}^1, \tilde{\mu}^{-1})), & \ddot{\tilde{\mu}}_0^0 &= \tilde{F}_0(\tilde{\mu}^0) + \tilde{F}_0''(\tilde{\mu}^0)(\tilde{\mu}^1, \tilde{\mu}^{-1}), \\ \tilde{\mu}_1^{-1} &= \frac{1}{-2\tilde{\omega}^2} \tilde{F}'_{-1}(\tilde{\mu}^0) \tilde{\mu}^1, & \tilde{\mu}_1^0 &= \frac{1}{-\tilde{\omega}^2} \tilde{F}'_0(\tilde{\mu}^0) \tilde{\mu}^1, \\ \dot{\tilde{\mu}}_1^1 &= \frac{1}{i\tilde{\omega}} \tilde{F}'_1(\tilde{\mu}^0) \tilde{\mu}^1, & \dot{\tilde{\mu}}_{-1}^{-1} &= \frac{1}{-i\tilde{\omega}} \tilde{F}'_{-1}(\tilde{\mu}^0) \tilde{\mu}^{-1}, \\ \tilde{\mu}_{-1}^0 &= \frac{1}{-2\tilde{\omega}^2} \tilde{F}'_0(\tilde{\mu}^0) \tilde{\mu}^{-1}, & \tilde{\mu}_{-1}^1 &= \frac{1}{-2\tilde{\omega}^2} \tilde{F}'_1(\tilde{\mu}^0) \tilde{\mu}^{-1}, \end{aligned} \quad (50)$$

and the functions $\tilde{\nu}^k$ are

$$\tilde{\nu}_0 = \dot{\tilde{\mu}}_0, \quad \tilde{\nu}_{\pm 1} = \pm i\tilde{\omega} \tilde{\mu}_{\pm 1} + \dot{\tilde{\mu}}_{\pm 1} = \pm i\tilde{\omega} \tilde{\mu}_{\pm 1} + \mathcal{O}(\tilde{\omega}^{-1}). \quad (51)$$

The initial values are the same as those of the numerical solutions.

• **Proof of the convergence.**

Looking closely to the equations (50) and (47), which determine the modulated Fourier expansion coefficients, it is obtained that $\tilde{x}^*(t) - \tilde{x}_h(t) = \mathcal{O}(h^2)$. Similarly, with (51) and (48), one has $\tilde{v}^*(t) - \tilde{v}_h(t) = \mathcal{O}(h^2)$. According to the above results and the defects of modulated Fourier expansions, we have the following diagram:

$$\begin{array}{ccc} \text{Exact solution } (\tilde{x}(nh), \tilde{v}(nh)) & & \text{Numerical solution } (\tilde{x}_n, \tilde{v}_n) \\ \downarrow (\mathcal{O}(h^2), \mathcal{O}(h^2/\varepsilon)) & & \downarrow (\mathcal{O}(h^2), \mathcal{O}(h^2/\varepsilon)) \\ \text{Modulated Fourier expansion } (\tilde{x}^*, \tilde{v}^*) & \xleftrightarrow{(\mathcal{O}(h^2), \mathcal{O}(h^2))} & \text{Modulated Fourier expansion } (\tilde{x}_h, \tilde{v}_h) \end{array}$$

The error bounds

$$\tilde{x}(nh) - \tilde{x}_n = \mathcal{O}(h^2), \quad \tilde{v}(nh) - \tilde{v}_n = \mathcal{O}(h^2/\varepsilon)$$

are immediately obtained on the basis of this diagram. This obviously yields

$$x(nh) - x_n = \mathcal{O}(h^2), \quad v(nh) - v_n = \mathcal{O}(h^2/\varepsilon)$$

and the proof is complete.

7 Conclusions

Structure-preserving algorithms constitute an interesting and important class of numerical methods. Furthermore, algorithms with uniformly errors of highly oscillatory systems have received a great deal of attention. In this paper, we have formulated and analysed some structure-preserving algorithms with uniform error bound for solving nonlinear highly oscillatory Hamiltonian systems. Two kinds of algorithms with uniform error bound were given to preserve the symplecticity and energy, respectively. All the theoretical results were supported by a numerical experiment and were proved in detail.

Last but not least, it is noted that all the algorithms and analysis are also suitable to the non-highly oscillatory system (1) with $\varepsilon = 1$. Meanwhile, there are some issues brought by this paper which can be researched further. For the system (1) with a matrix $\tilde{B}(x)$ depending on x , how to modify the methods and extend the analysis to get higher order uniformly accurate structure-preserving algorithms? This point will be considered in future. Another issue for future exploration is the extension and application of the methods presented in this paper to the Vlasov equations under strong magnetic field [3, 4].

Acknowledgements

This work was supported by the Natural Science Foundation of China (NSFC) under grant 11871393; International Science and Technology Cooperation Program of Shaanxi Key Research & Development Plan under grant 2019KWZ-08. The first author is grateful to Christian Lubich for his valuable comments on Theorem 3.3 as well as its proof, which was done in part at UNIVERSITÄT TÜBINGEN when the first author worked there as a postdoctoral researcher (2017-2019, supported by the Alexander von Humboldt Foundation).

References

- [1] J.P. BORIS, Relativistic plasma simulation-optimization of a hybrid code, Proceeding of Fourth Conference on Numerical Simulations of Plasmas, (1970), pp. 3-67.
- [2] L. BRUGNANO, F. IAVERNARO, R. ZHANG, Arbitrarily high-order energy-preserving methods for simulating the gyrocenter dynamics of charged particles, *J. Comput. Appl. Math.* 380 (2020), pp. 112994
- [3] PH. CHARTIER, N. CROUSEILLES, M. LEMOU, F. MÉHATS, X. ZHAO, Uniformly accurate methods for Vlasov equations with non-homogeneous strong magnetic field, *Math. Comp.* 88 (2019), pp. 2697-2736.
- [4] PH. CHARTIER, N. CROUSEILLES, M. LEMOU, F. MÉHATS, X. ZHAO, Uniformly accurate methods for three dimensional Vlasov equations under strong magnetic field with varying direction, *SIAM J. Sci. Comput.* 42 (2020), pp. B520-B547.
- [5] PH. CHARTIER, F. MÉHATS, M. THALHAMMER, Y. ZHANG, Improved error estimates for splitting methods applied to highly-oscillatory nonlinear Schrödinger equations, *Math. Comp.* 85 (2016), pp. 2863-2885.
- [6] K. FENG, M. QIN, *Symplectic Geometric algorithms for Hamiltonian systems*, Springer-Verlag, Berlin, Heidelberg, 2010.
- [7] F. FILBET, M. RODRIGUES, Asymptotically stable particle-in-cell methods for the Vlasov-Poisson system with a strong external magnetic field, *SIAM J. Numer. Anal.* 54 (2016), pp. 1120-1146.
- [8] F. FILBET, M. RODRIGUES, Asymptotically preserving particle-in-cell methods for inhomogeneous strongly magnetized plasmas, *SIAM J. Numer. Anal.* 55 (2017), pp. 2416-2443.

- [9] L. GAUCKLER, E. HAIRER, CH. LUBICH, Dynamics, numerical analysis, and some geometry. Proceedings of the International Congress of Mathematicians — Rio de Janeiro 2018. Vol. I. Plenary lectures, 453-485, World Sci. Publ., Hackensack, NJ, 2018.
- [10] E. HAIRER, CH. LUBICH, Long-time energy conservation of numerical methods for oscillatory differential equations, *SIAM J. Numer. Anal.*, 38 (2000), pp. 414–441.
- [11] E. HAIRER, CH. LUBICH, Energy behaviour of the Boris method for charged-particle dynamics, *BIT* 58 (2018), pp. 969-979.
- [12] E. HAIRER, CH. LUBICH, Symmetric multistep methods for charged-particle dynamics, *SMAI J. Comput. Math.* 3 (2017), pp. 205-218
- [13] E. HAIRER, CH. LUBICH, Long-term analysis of a variational integrator for charged-particle dynamics in a strong magnetic field, *Numer. Math.* 144 (2020), pp. 699-728.
- [14] E. HAIRER, CH. LUBICH, Y. SHI, Large-stepsize integrators for charged-particle dynamics over multiple time scales, <https://na.uni-tuebingen.de/preprints.shtml>
- [15] E. HAIRER, CH. LUBICH, B. WANG, A filtered Boris algorithm for charged-particle dynamics in a strong magnetic field, *Numer. Math.* 144 (2020), pp. 787-809.
- [16] E. HAIRER, CH. LUBICH, G. WANNER, Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer-Verlag, Berlin, Heidelberg, 2006
- [17] Y. HE, Z. ZHOU, Y. SUN, J. LIU, H. QIN, Explicit K-symplectic algorithms for charged particle dynamics, *Phys. Lett. A* 381 (2017), pp. 568-573
- [18] Y. HE, Y. SUN, J. LIU, H. QIN, Volume-preserving algorithms for charged particle dynamics, *J. Comput. Phys.* 281 (2015), pp. 135-147
- [19] M. HOCHBRUCK, A. OSTERMANN, Exponential integrators, *Acta Numer.* 19 (2010), pp. 209-286
- [20] T. LI, B. WANG, Efficient energy-preserving methods for charged-particle dynamics, *Appl. Math. Comput.* 361 (2019), pp. 703-714.
- [21] T. LI, B. WANG, Arbitrary-order energy-preserving methods for charged-particle dynamics, *Appl. Math. Lett.* 100 (2020), pp. 106050.
- [22] H. QIN, S. ZHANG, J. XIAO, J. LIU, Y. SUN, W.M. TANG, Why is Boris algorithm so good? *Physics of Plasmas*, 20 (2013), pp. 084503
- [23] L.F. RICKETSON, L. CHACÓN, An energy-conserving and asymptotic-preserving charged-particle orbit implicit time integrator for arbitrary electromagnetic fields, *J. Comput. Phys.* 418 (2020), pp. 109639
- [24] M. TAO, Explicit high-order symplectic integrators for charged particles in general electromagnetic fields, *J. Comput. Phys.* 327 (2016), pp. 245-251

- [25] B. WANG, Exponential energy-preserving methods for charged-particle dynamics in a strong and constant magnetic field, *J. Comput. Appl. Math.* 387 (2021), pp. 112617.
- [26] B. WANG, X. WU, A long-term numerical energy-preserving analysis of symmetric and or symplectic extended RKN integrators for efficiently solving highly oscillatory Hamiltonian systems, To appear in *BIT Numer. Math.* (2021)
- [27] B. WANG, X. ZHAO, Error estimates of some splitting schemes for charged-particle dynamics under strong magnetic field, arXiv:submit/3190341 (2020)
- [28] X. WU, B. WANG, *Recent Developments in Structure-Preserving Algorithms for Oscillatory Differential Equations*, Springer Nature Singapore Pte Ltd, 2018.