Journal of Scientific Computing (2023) 95:66
https://doi.org/10.1007/s10915-023-02178-6

Structure-Preserving Algorithms with Uniform Error Bound and Long-time Energy Conservation for Highly Oscillatory Hamiltonian Systems

Bin Wang[1] · **Yaolin Jiang**[1]

Received: 17 August 2022 / Revised: 13 February 2023 / Accepted: 9 March 2023 /
Published online: 17 April 2023

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 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. Long time energy conservation is analysed for symplectic algorithms and the proposed algorithms are shown to have uniform error bound in the position for the highly oscillatory structure. Moreover, some methods with uniform error bound in the position and in the velocity are derived and analysed. Two numerical experiments are carried out to support all the theoretical results established in this paper by showing the performance of the blended algorithms.

Keywords Nonlinear Hamiltonian systems · Highly oscillatory systems · Symplectic algorithms · Energy-preserving algorithms · Uniform error bound · Long-time conservation

Mathematics Subject Classification 65L05 · 65P10 · 65L20 · 65L70

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 [25]. In this paper we are concerned with efficient algorithms for the following highly oscillatory second-order

✉ Yaolin Jiang
yljiang@mail.xjtu.edu.cn

Bin Wang
wangbinmaths@xjtu.edu.cn

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China

differential equation

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

where $x(t) \in \mathbb{R}^d$, $\tilde{B}$ is a $d \times d$ skew symmetric matrix, and it is assumed that $F(x)$ is a smooth function and is the negative gradient of a real-valued function $U(x)$. In this work, we focus on the case where $0 < \varepsilon \ll 1$. Under this case, the solution of this dynamic is *highly oscillatory*. We note that with $p = \dot{x} - \frac{1}{2\varepsilon}\tilde{B}x$, the Eq. (1) can be transformed into a Hamiltonian system with the non-separable Hamiltonian

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

It immediately follows that the energy of (1) is given by

$$E(x, v) = \frac{1}{2}|v|^2 + U(x), \tag{3}$$

with $v = \dot{x}$. This energy 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 denote in this paper by $|\cdot|$ the Euclidean norm. If the system (1) is two dimensional and the matrix $\tilde{B}$ appeared there depends on x, it has been shown in [5] that $x(t) - x(0) = \mathcal{O}(\varepsilon)$. Therefore, all the methods presented in this paper can be extended to the two dimensional system (1) with a matrix $\tilde{B}(x)$ by rewriting (1) as

$$\ddot{x}(t) = \frac{1}{\varepsilon}\tilde{B}(x(0))\dot{x}(t) + F(x(t)) + \frac{\tilde{B}(x(t)) - \tilde{B}(x(0))}{\varepsilon}\dot{x}(t).$$

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 [2, 5, 6, 15, 20–22, 24, 29, 30, 36, 40]. 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. [14, 17, 25, 44]). For the Hamiltonian system (2), 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.

The numerical computation of highly oscillatory system contains numerous enduring challenges. In practice, the numerical methods used to treat (1) are focus on charged-particle dynamics and can be summarized in the following three categories.

a) The primitive numerical methods usually depend on the approximation of the solution besides high-frequency oscillation and structure preservation characteristics such as the Boris method [1] as well as its further researches [20, 34]. 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 [27], symplectic methods [26, 37, 38, 41], symmetric methods [21] and energy-preserving methods [2–4, 35]. In [22], 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 [23] 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 [24] 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 [15, 16] and the uniformly accurate schemes [5, 6, 18]. Although these powerful numerical methods have very good performance in accuracy, structure 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 [40]. It will be shown in this paper that this method has a uniform error bound which has not been studied in [40]. In a recent paper [7], a kind of uniformly accurate methods has been demonstrated numerically to have a near energy conservation but without theoretical analysis. Very recently, the authors in [42] 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. More recently, geometric two-scale integrators have been developed in [43] but the methods only have near conservations. A numerical method combining 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, some algorithms with first-order or higher-order uniform error bound in both x and v will be proposed. Compared with the existing analysis techniques, the main differences and contributions in the proof involve in two aspects. a) We transform the system and the methods, and then the symplecticity and long time energy conservation are shown for the transformed methods. These transformations make the analysis be more ingenious and simpler. b) For the convergence analysis, to make good use of the scheme of methods, two different techniques named as the re-scaled technique and modulated Fourier expansion are taken. According to the scheme of methods, we choose the suitable analytical technique and make the necessary modifications to make the proof go smoothly.

The remainder of this paper is organised as follows. In Sect. 2, we formulate three kinds of algorithms. The main results of these algorithms are given in Sect. 3 and two numerical experiments are carried out there to numerically show the performance of the algorithms. The proofs of the main results are presented in Sects. 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 [24].) *The exact solution of system (1) can be expressed as*

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

where $\Omega = \frac{1}{\varepsilon}\tilde{B}$, *h is a stepsize,* $t_n = nh$ *and the* φ*-functions are defined by (see [28])*

$$\varphi_0(z) = e^z, \quad \varphi_n(z) = \int_0^1 e^{(1-\sigma)z}\frac{\sigma^{n-1}}{(n-1)!}d\sigma, \quad n = 1, 2, \ldots.$$

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* $n = 0, 1, \ldots,$ *an s-stage adaptive exponential algorithm applied with stepsize h is defined by:*

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

where $\alpha_{ij}(h\Omega)$, $\beta_i(h\Omega)$, $\gamma_i(h\Omega)$ *are matrix-valued functions of* $h\Omega$.

As some practical examples, we present three explicit algorithms based on the conditions (27) of symplecticity given below. The coefficients are obtained by considering the s-stage adaptive exponential algorithm (5) with the coefficients for $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, i$,

$$\begin{aligned} &\alpha_{ij} = a_{ij}(c_i - c_j)\varphi_1((c_i - c_j)h\Omega), \ \beta_i = b_i(1 - c_i)\varphi_1((1 - c_i)h\Omega), \\ &\gamma_i = b_i\varphi_0((1 - c_i)h\Omega), \end{aligned} \tag{6}$$

where $(c_1, c_2, \ldots, c_s)$, $(b_1, b_2, \ldots, b_s)$ *and* $(a_{ij})_{s\times s}$ *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 symplectic conditions (27) given below. We omit the details of calculations for brevity. We first consider*

$$\textbf{\textit{Symplectic Method 1 (SM1)}}: \ s = 1, \ c_1 = \frac{1}{2}, \ b_1 = 1.$$

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

$$\textbf{\textit{Symplectic Method 2 (SM2)}}: \ c_1 = 0, \ c_2 = 1, \ a_{21} = \frac{1}{2}, \ b_1 = \frac{1}{2}, \ b_2 = 1$$

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

$$\textbf{\textit{Symplectic Method 3 (SM3)}}: \ c_1 = \frac{1}{4}, \ c_2 = \frac{3}{4}, \ a_{21} = \frac{1}{2}, \ b_1 = b_2 = \frac{1}{2},$$

then the corresponding method is referred to SM3.

Remark 2.3 It is remarked that the following integrator for solving (1) has been given in [24]

$$\begin{aligned} x_{n+1} &= x_n + h\varphi_1(h\Omega)v_n + \tfrac{1}{2}h^2\Psi(h\Omega)F_n, \\ v_{n+1} &= \varphi_0(h\Omega)v_n + \tfrac{1}{2}h\big(\Psi_0(h\Omega)F_n + \Psi_1(h\Omega)F_{n+1}\big), \end{aligned} \tag{7}$$

where $F_n = F(x_n)$ and Ψ, Ψ_0, Ψ_1 are matrix-valued functions of $h\Omega$ satisfying $\Psi(0) = \Psi_0(0) = \Psi_1(0) = 1$. For this scheme, convergence is researched but the structure preservation such as symplecticity or energy conservation has not been discussed. It is noted that Algorithm 2.2 given by (5) contains (7) and thence the results of SM1-SM3 also hold for (7). In this paper, we will study not only the convergence but also the symplecticity and long time energy conservation for (5). It will be shown that SM1-SM3 are all symplectic and have a good near conservation of energy over long times, which are also true for the algorithm (7) presented in [24].

We also note that it will be shown in this paper that SM1-SM3 are all second order and they have similar long time conservations. Higher-order methods can be derived but their long term analysis is very complicated and challenging. We hope to make some progress on this aspect in our future work.

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

$$\begin{aligned} X_\tau &= x_n + hC_\tau(h\Omega)v_n + h^2\int_0^1 A_{\tau\sigma}(h\Omega)F(X_\sigma)d\sigma, \quad 0 \le \tau \le 1, \\ x_{n+1} &= x_n + h\varphi_1(h\Omega)v_n + h^2\int_0^1 \bar{B}_\tau(h\Omega)F(X_\tau)d\tau, \\ v_{n+1} &= \varphi_0(h\Omega)v_n + h\int_0^1 B_\tau(h\Omega)F(X_\tau)d\tau, \end{aligned} \tag{8}$$

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, $\bar{B}_\tau$, 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 $i = 1, 2, \ldots, s+1$ 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

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

This obtained algorithm can be rewritten as

$$\begin{aligned} &\textbf{\textit{Energy-preserving}} \quad x_{n+1} = x_n + h\varphi_1(h\Omega)v_n + h^2\varphi_2(h\Omega)\textstyle\int_0^1 F\big(x_n + \sigma(x_{n+1} - x_n)\big)d\sigma, \\ &\textbf{\textit{Method 1 (EM1):}} \quad v_{n+1} = \varphi_0(h\Omega)v_n + h\varphi_1(h\Omega)\textstyle\int_0^1 F\big(x_n + \sigma(x_{n+1} - x_n)\big)d\sigma, \end{aligned} \tag{9}$$

which is denoted by EM1.

Remark 2.5 It is noted that EM1 of Algorithm 2.4 has been given in [40] 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 some methods have a first-order or higher-order uniform error bound in both x and v and the others have a second-order uniform convergence in x for $0 < \varepsilon \ll 1$. In contrast, many classical methods such as Euler method, 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 $k > 0$.

We remark that the above four methods will be shown to have uniform error bound only in x. In order to obtain some methods with the same uniform error bound in both x and v, we derive the following algorithm.

Algorithm 2.6 *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

$$\begin{aligned} \textbf{\textit{Method 1 (M1):}}\ \ x_{n+1} &= x_n + h\varphi_1(h\Omega)v_n + h^2\varphi_2(h\Omega)F(x_n), \\ v_{n+1} &= \varphi_0(h\Omega)v_n + h\varphi_1(h\Omega)F(x_n). \end{aligned}$$

This method is referred to M1.

By the simple M1 and parareal algorithms (see [31]), we formulate some higher order methods. For solving the second-order system (1), the parareal algorithm uses two propagators: the fine propagator $\mathcal{F}$ and the coarse propagator $\mathcal{G}$, where classically $\mathcal{F}$ uses a small (fine) time step δt and $\mathcal{G}$ a large (coarse) time step Δt. In this paper, we consider M1 with $\Delta t = h$ as the coarse propagator and denote this propagator by

$$[x_{n+1}; v_{n+1}] = \mathcal{G}_{t_n}^{t_{n+1}}([x_n; v_n]) := \begin{pmatrix} x_n + h\varphi_1(h\Omega)v_n + h^2\varphi_2(h\Omega)F(x_n) \\ \varphi_0(h\Omega)v_n + h\varphi_1(h\Omega)F(x_n) \end{pmatrix}. \tag{10}$$

For the fine propagator $\mathcal{F}$, we also choose M1 with a small time step $0 < \delta t < h$ and refer to it as

$$[x_{n+1}; v_{n+1}] = \mathcal{F}_{t_n}^{t_{n+1}}([x_n; v_n]) = \mathcal{F}_{t_{n+1}-\delta t}^{t_{n+1}} \circ \cdots \circ \mathcal{F}_{t_n}^{t_n+\delta t}([x_n; v_n]), \tag{11}$$

where

$$\mathcal{F}_{t_n}^{t_n+\delta t}([x_n; v_n]) := \begin{pmatrix} x_n + \delta t\varphi_1(\delta t\Omega)v_n + \delta t^2\varphi_2(\delta t\Omega)F(x_n) \\ \varphi_0(\delta t\Omega)v_n + \delta t\varphi_1(\delta t\Omega)F(x_n) \end{pmatrix}.$$

For solving (1), the parareal algorithms compute for iteration index $k = 0, 1, \ldots$ and $n = 0, 1, \ldots, \frac{T}{h} - 1$, and with $[x_0^k, v_0^k] = [x_0, v_0]$

$$\textbf{\textit{Parareal Method (PM):}}\ [x_{n+1}^{k+1}; v_{n+1}^{k+1}] = \mathcal{G}_{t_n}^{t_{n+1}}([x_n^{k+1}; v_n^{k+1}]) + \mathcal{F}_{t_n}^{t_{n+1}}([x_n^k; v_n^k]) - \mathcal{G}_{t_n}^{t_{n+1}}([x_n^k; v_n^k]). \tag{12}$$

The initial guess $\{[x_n^0; v_n^0]\}_{n\geq 1}$ can be random or generated by the $\mathcal{G}$-propagator. We shall refer to (12) by PM. As two examples, we choose $k = 1, \delta t = h^2$ and $k = 2, \delta t = h^3$ and denote them by parareal method 1 (PM1) and parareal method 2 (PM2), respectively.

Remark 2.7 It is noted that M1 is a kind of exponential integrators and compared with the methods given in [43], it has simple scheme. The aim of presenting M1 is to derive higher-order methods with uniform error bounds and simple scheme. For the higher-order uniformly accurate methods, more complicated formulations are needed and we refer to [8, 18, 43] for some recent work on this topic.

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 of the methods for the Hamiltonian system (2) and the last one concerns uniform error bound for the format (1).

Theorem 3.1 *(**Symplecticity of SM1-SM3**) Consider the methods SM1-SM3 of Algorithm 2.2 where* $p_{n+1} = v_{n+1} - \frac{1}{2\varepsilon}\tilde{B}x_{n+1}$. *In this case, for the non-separable Hamiltonian (2), the map* $(x_n, p_n) \to (x_{n+1}, p_{n+1})$ *determined by these methods is symplectic, i.e.,*

$$dx_{n+1} \wedge dp_{n+1} = dx_n \wedge dp_n \quad for \ \ n = 0, 1, \ldots.$$

Theorem 3.2 *(**Energy preservation of EM1** [40].) The method EM1 of Algorithm 2.4 preserves the energy (3) exactly, i.e.*

$$E(x_{n+1}, v_{n+1}) = E(x_n, v_n) \quad for \ \ n = 0, 1, \ldots.$$

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

- *It is assumed that the initial values* x_0 *and* $v_0 := \dot{x}_0$ *are bounded independently of* ε.
- *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.
- *After diagonalization of* $\tilde{B}/\varepsilon$, *denote the obtained diagonal matrix by* $i\tilde{\Omega}$. *Consider the notations* $k = (k_1, k_2, \ldots, k_l)$, $\varpi = (\varpi_1, \varpi_2, \ldots, \varpi_l) := (\text{diagonal elements of } \tilde{\Omega})$, $k \cdot \varpi = k_1\varpi_1 + k_2\varpi_2 + \ldots + k_l\varpi_l$ *and the resonance module*

$$\mathcal{M} = \{k \in \mathbb{Z}^l : \ k \cdot \varpi = 0\}. \tag{13}$$

  Assume that the numerical non-resonance condition is true

$$|\sin(\frac{h}{2}(k \cdot \varpi))| \geq c\sqrt{h} \ \text{ for } \ k \in \mathbb{Z}^l \backslash \mathcal{M} \ \text{ with } \ \sum_{j=1}^{l} |k_j| \leq N$$

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

For the symplectic methods SM1-SM3 of Algorithm 2.2, it holds that

$$|E(x_n, v_n) - E(x_0, v_0)| \leq Ch \tag{14}$$

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 and PM1-PM2 do not have the above energy conservation property. The reason is that they are not symplectic and symmetric methods. It will be seen from the proof given in Sect. 5 that symplecticity and symmetry play an important role in the analysis.

Table 1 Properties of the obtained methods

Methods	Symplecticity	Symmetry	Energy property	Convergence
Symplectic Method (SM1)	Yes	Yes	Near conservation	h^2 in x and $\frac{h^2}{\varepsilon}$ in v
Symplectic Method (SM2)	Yes	Yes	Near conservation	h^2 in x and $\frac{h^2}{\varepsilon}$ in v
Symplectic Method (SM3)	Yes	Yes	Near conservation	h^2 in x and $\frac{h^2}{\varepsilon}$ in v
Energy-preserving Method (EM1)	No	Yes	Exact conservation	h^2 in x and $\frac{h^2}{\varepsilon}$ in v
Method (M1)	No	No	No	h in both x and v
Parareal Method (PM1)	No	No	No	h^2 in both x and v
Parareal Method (PM2)	No	No	No	h^3 in both x and v

Remark 3.5 It is noted that in Theorem 3.3, a lower bound on the stepsize $h \geq c_0\varepsilon$ is presented. At first glance, it seems that this contradicts with the fact that a small stepsize is needed in the methods for highly oscillatory problems. From Theorem 3.6 given blew, it follows that the symplectic methods SM1-SM3 have the accuracy $\mathcal{O}(h^2)$ in x and $\mathcal{O}(h^2/\varepsilon)$ in v. Thus if one only concerns the accuracy in x, large stepsizes can be used for the symplectic methods and Theorem 3.3 shows that under this case, the methods still have a long-time near energy conservation. If a small stepsize is used to keep a good accuracy in both x and v, the energy behaviour of the symplectic methods can be derived with the help of backward error analysis (Chap. IX of [25]).

Theorem 3.6 ***(Convergence.)*** *Assume that the initial value of* (1) *is uniformly bounded w.r.t* ε, $F(x)$ *is smooth and uniformly bounded for all* ε, *and the solution of* (1) *stays in a uniformly bounded set. For the methods M1 and PM of Algorithm 2.6, the global errors are bounded by*

$$M1: \quad |x_n - x(t_n)| \lesssim h, \quad |v_n - v(t_n)| \lesssim h, \tag{15a}$$

$$PM: \quad \left|x_n^k - x(t_n)\right| \lesssim h^{k+1} + \delta t, \quad \left|v_n^k - v(t_n)\right| \lesssim h^{k+1} + \delta t, \tag{15b}$$

where $0 < nh \leq T$. *The convergence of the energy-preserving method EM1 of Algorithm 2.4 is*

$$EM1: \ |x_n - x(t_n)| \lesssim h^2, \ |v_n - v(t_n)| \lesssim h^2/\varepsilon. \tag{16}$$

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 the bound of* F_x. *For the symplectic methods SM1-SM3 of Algorithm 2.2, under the conditions of Theorem 3.3,*

$$SM1\text{-}SM3: \quad |x_n - x(t_n)| \lesssim h^2, \ |v_n - v(t_n)| \lesssim h^2/\varepsilon, \tag{17}$$

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

In the Sects. 4–6, we will prove Theorems 3.1, 3.3-3.6, respectively. 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.

Remark 3.7 For the seven methods presented in this paper, concerning the symmetry [25], it is easy to check that all of them except M1 and PM1-PM2 are symmetric. Their properties

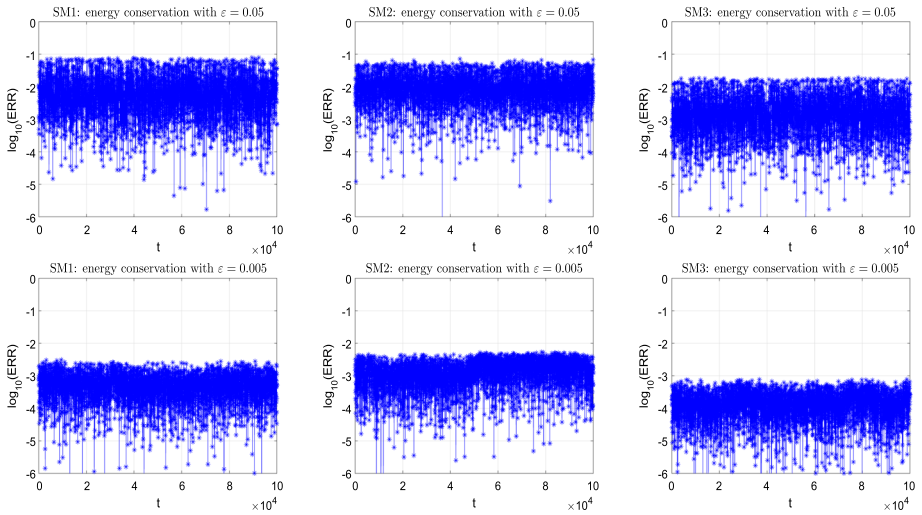

Fig. 1 The relative energy errors (ERR) against t for our symplectic SM1-SM3

are summarized in Table 1. All of these observations will be numerically illustrated by a test given below. Ones can choose the appropriate algorithm according to their interest.

- If uniform (in ε) error bound is needed in both x and v, M1 and PM1-PM2 are most appropriate and these three algorithms provide different uniform accuracy from the first order to the third order.
- If ones only focus on uniform error bound in x, the symplectic or energy-preserving methods are good choices. EM1 can preserve the energy exactly but it is implicit. Symplectic methods SM1-SM3 have similar numerical behaviour. They are explicit, preserve the symplecticity and have a good near conservation of energy over long times. Ones can choose the preferred method depending on their demands.

3.2 Numerical Tests

In this part, we carry out two numerical experiments to show the performance of the derived methods.

Problem 1 As an illustrative numerical experiment, we consider the charged particle system of [20] 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 $x(0) = (0.6, 1, -1)^\intercal$ and $v(0) = (-1, 0.5, 0.6)^\intercal$. It is noted here that we use the four-point Gauss-Legendre's quadrature to the integral involved in the numerical flow EM1.

Energy conservation We take $\varepsilon = 0.05, 0.005$ and apply our seven methods as well as the symplectic Euler method (denoted by SE), the exponential Euler method (denoted by EE) [28] and the explicit exponential Runge–Kutta method (denoted by ERK) of order two [28] 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

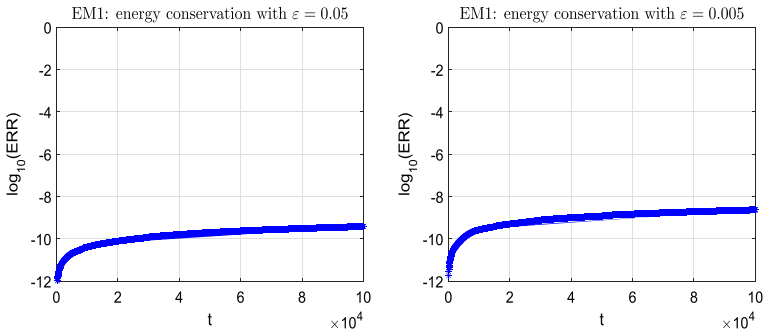

Fig. 2 The relative energy errors (ERR) against t for our energy-preserving EM1

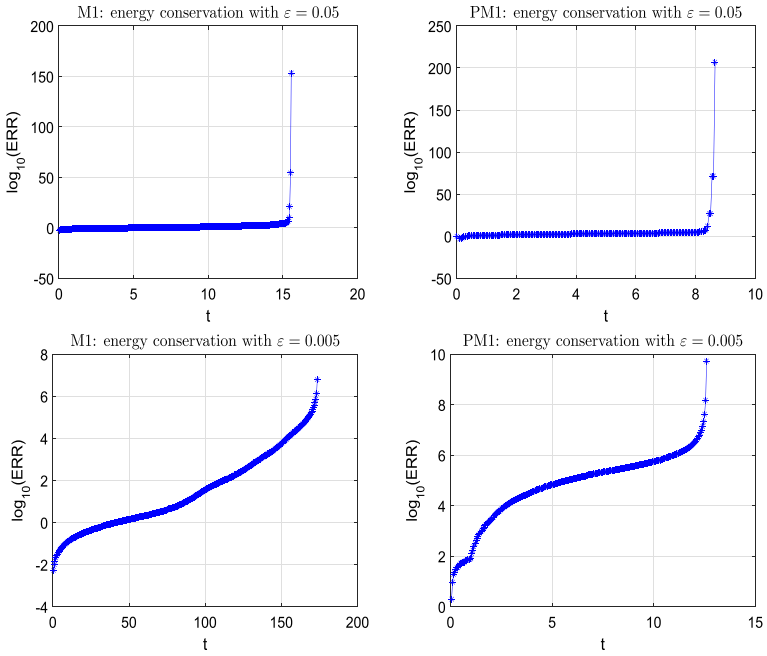

Fig. 3 The relative energy errors (ERR) against t for our M1 and PM1

iterations. For the other methods, they are explicit and the iteration is not needed. 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. 1–4. According to these results, we have the following observations. SM1-SM3 (Fig. 1) have near energy conservation over long times, EM1 preserves the energy very well (Fig. 2) but others show a bad energy conservation (Figs. 3–4). We do not show the result for PM2 since it has a similar behaviour as PM1.

Convergence For displaying the results of convergence, the problem is solved on $[0, 1]$ and the global errors $errx := \frac{|x_n - x(t_n)|}{|x(t_n)|}$, $errv := \frac{|v_n - v(t_n)|}{|v(t_n)|}$ of each method for different ε are shown in Figs. 5–8. It is noted that we use the result of HOODESolver given in [32] as the true solution. It follows from these results that M1 and PM1-PM2 have uniform convergence in both x and v (Fig. 5) and the other our methods have a uniform second-order error bound

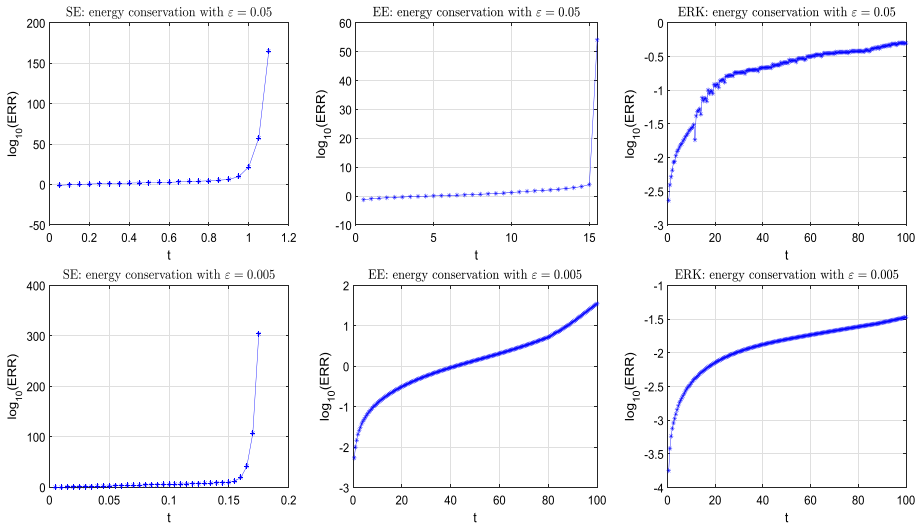

Fig. 4 The relative energy errors (ERR) against t for the existing methods SE, EE and ERK

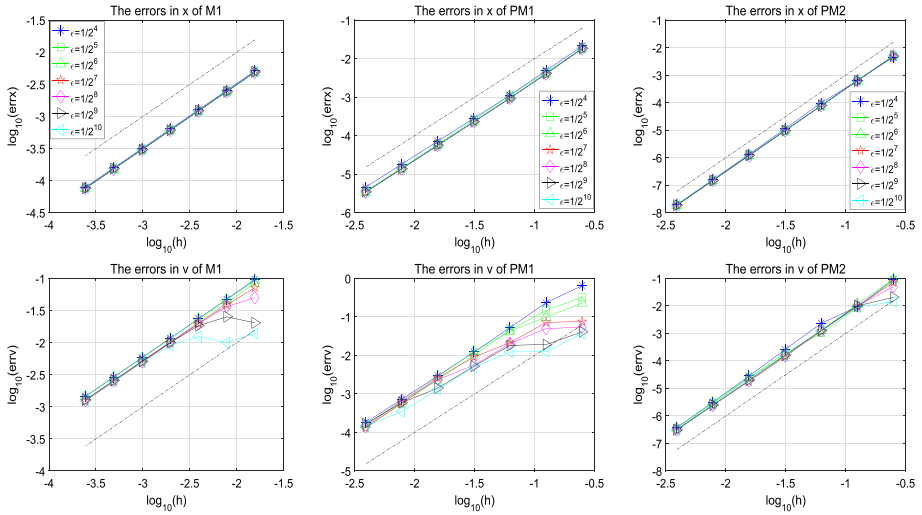

Fig. 5 The errors in x (errx) and v (errv) against h for our M1 and PM1-PM2 (the slope of the dotted line for M1 is one, for PM1 two and for PM2 three)

in x (Figs. 6–7), which agree with the results stated in Theorem 3.6. The existing method SE behaves very poor in the accuracy of both x and v and the exponential integrators EE, ERK have a uniform error bound in x (Fig. 8).

Resonance instability Finally, we show the resonance instability of the proposed methods. This is done by fixing $\varepsilon = 1/2^{10}$ and showing the errors at $T = 1$ against h/ε in Fig. 9. It can be observed that PM1 gives a very poor result[1], M1 shows very well but other methods have a good behavior for values of h/ε except integral multiples of π. SM3 shows a not

[1] PM2 and SE also behave very badly and we omit their numerical results for brevity.

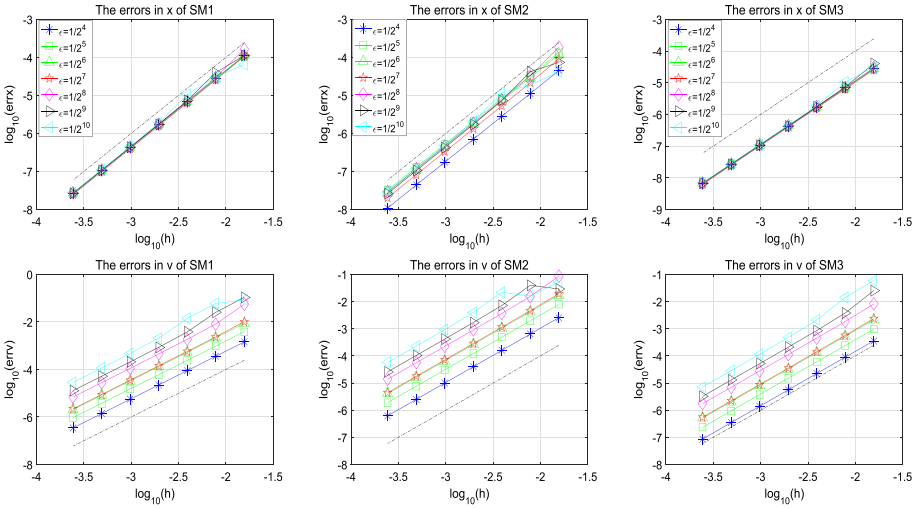

Fig. 6 The errors in x (errx) and v (errv) against h for our symplectic SM1-SM3 (the slope of the dotted line is two)

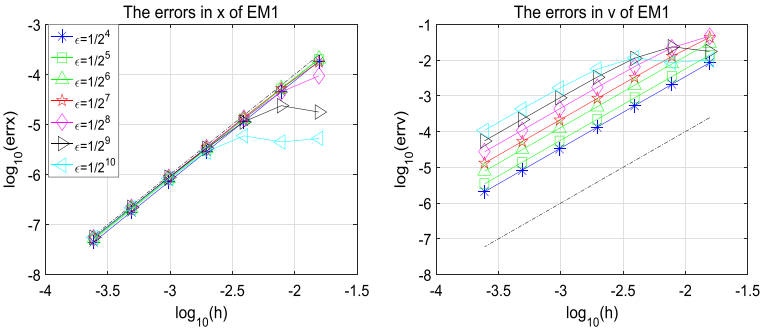

Fig. 7 The errors in x (errx) and v (errv) against h for our energy-preserving EM1 (the slope of the dotted line is two)

uniform result close to 4π, other methods EM1 and SM1-SM2 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.

Problem 2 The second test is devoted the system (1) with a large dimension $d = 32$, where the nonlinear function $F(x) = -\sin(x)$ and $\tilde{B}$ is chosen as a skew-symmetric tridiagonal matrix with $\tilde{B}_{j,j+1} = j/d$ for $j = 1, 2, \ldots, d-1$. We consider the initial values $x(0) = (0.1, 0.1, \ldots, 0.1)^\intercal$ and $v(0) = (0.2, 0.2, \ldots, 0.2)^\intercal$. This problem is firstly solved with $\varepsilon = 0.05$ and $h = 0.1$. The relative energy errors are presented in Fig. 10. PM2 has a similar behaviour as PM1 and thus we do not show its result for brevity. Then we integrate this problem on $[0, 1]$ and the global errors in x and v are displayed in Figs. 11–14. All the numerical observations remain the same as before.

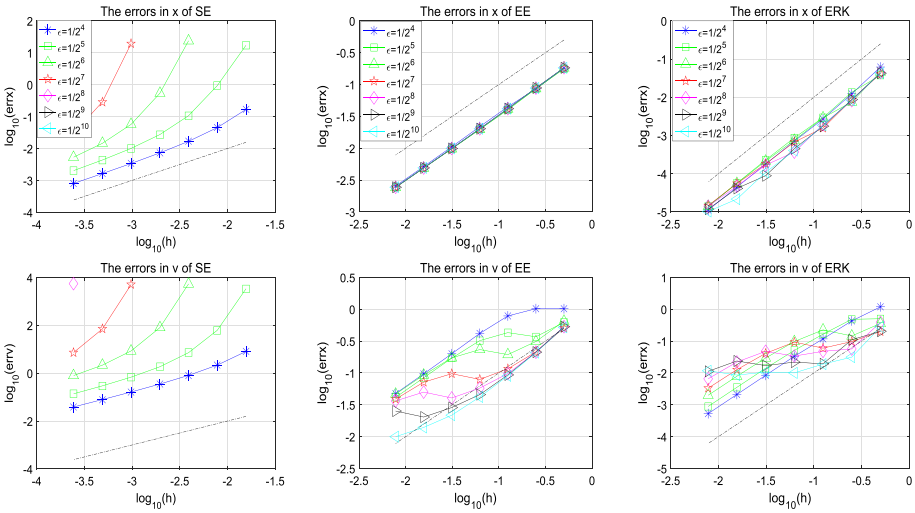

Fig. 8 The errors in x (errx) and v (errv) against h for the existing methods SE, EE and ERK (the slope of the dotted line for SE, EE is one and for ERK is two)

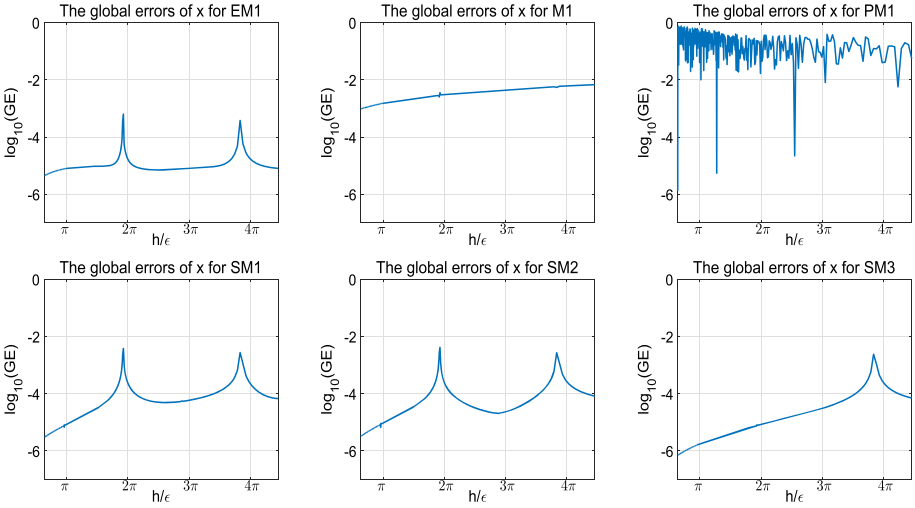

Fig. 9 The global errors (GE) of x against h/ε

4 Analysis on Symplecticity (Theorem 3.1)

In this section, we will prove the symplecticity of SM1-SM3 given in Theorem 3.1.

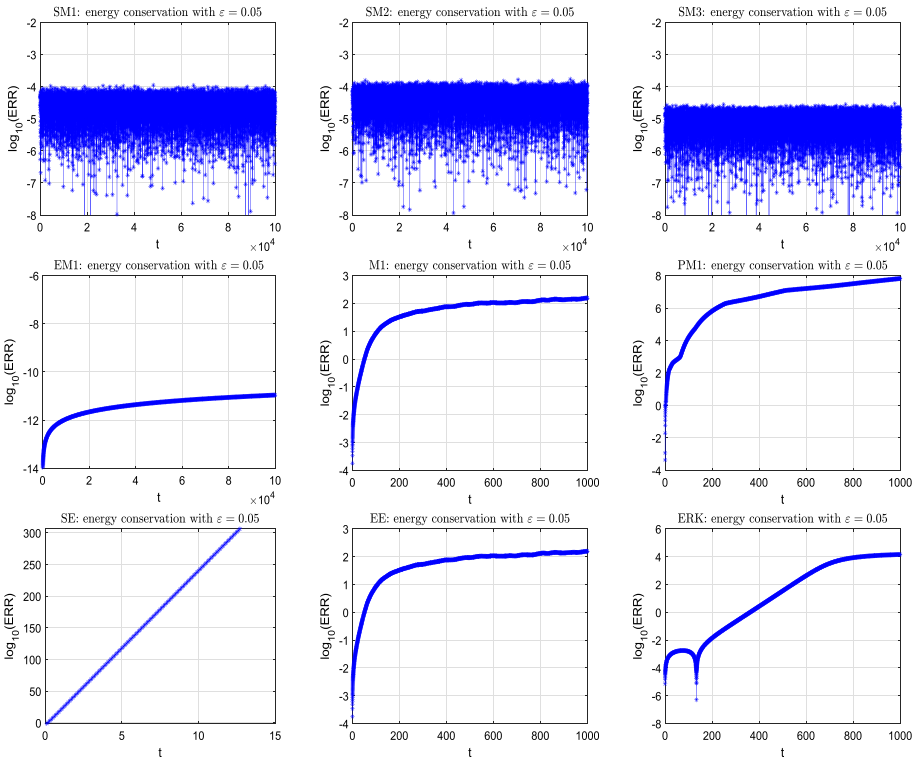

Fig. 10 The relative energy errors (ERR) against t for all the methods

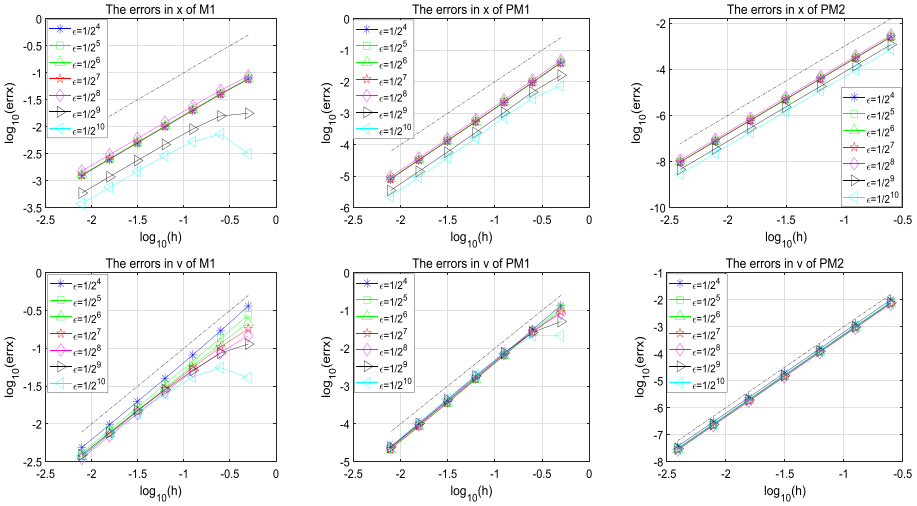

Fig. 11 The errors in x (errx) and v (errv) against h for our M1 and PM1-PM2 (the slope of the dotted line for M1 is one, for PM1 two and for PM2 three)

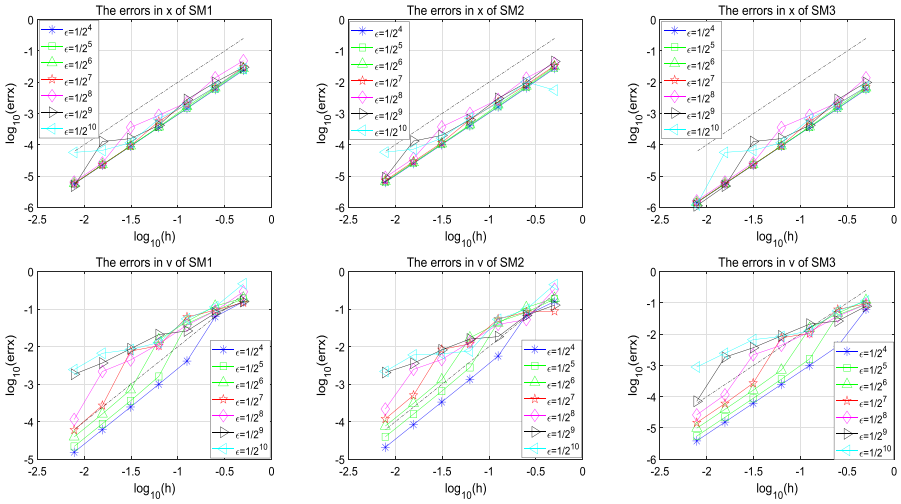

Fig. 12 The errors in x (errx) and v (errv) against h for our symplectic SM1-SM3 (the slope of the dotted line is two)

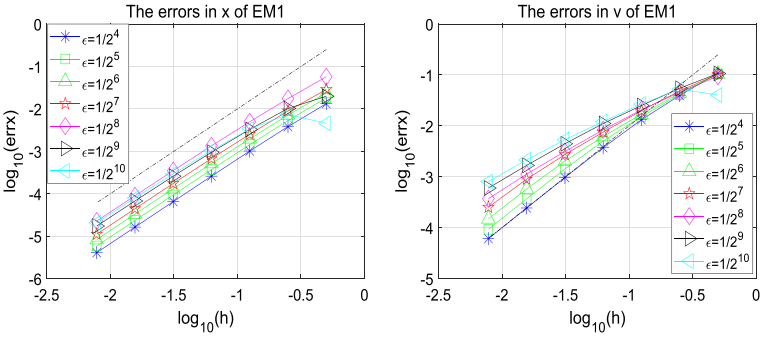

Fig. 13 The errors in x (errx) and v (errv) against h for our energy-preserving EM1 (the slope of the dotted line is two)

4.1 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 $\tilde{\Omega}$ such that

$$\frac{\tilde{B}}{\varepsilon} = P\mathrm{i}\tilde{\Omega}P^{\mathrm{H}},$$

where

$$\tilde{\Omega} = \begin{cases} \mathrm{diag}(-\tilde{\omega}_l, -\tilde{\omega}_{l-1}, \ldots, -\tilde{\omega}_1, 0, \tilde{\omega}_1, \ldots, \tilde{\omega}_{l-1}, \tilde{\omega}_l), & d = 2l+1, \\ \mathrm{diag}(-\tilde{\omega}_l, -\tilde{\omega}_{l-1} \ldots, -\tilde{\omega}_1, \tilde{\omega}_1, \ldots, \tilde{\omega}_{l-1}, \tilde{\omega}_l), & d = 2l, \end{cases} \tag{18}$$

with the integer $l \geq 1$. With the linear change of variable

$$\tilde{x}(t) = P^{\mathrm{H}}x(t), \quad \tilde{v}(t) = P^{\mathrm{H}}v(t), \tag{19}$$

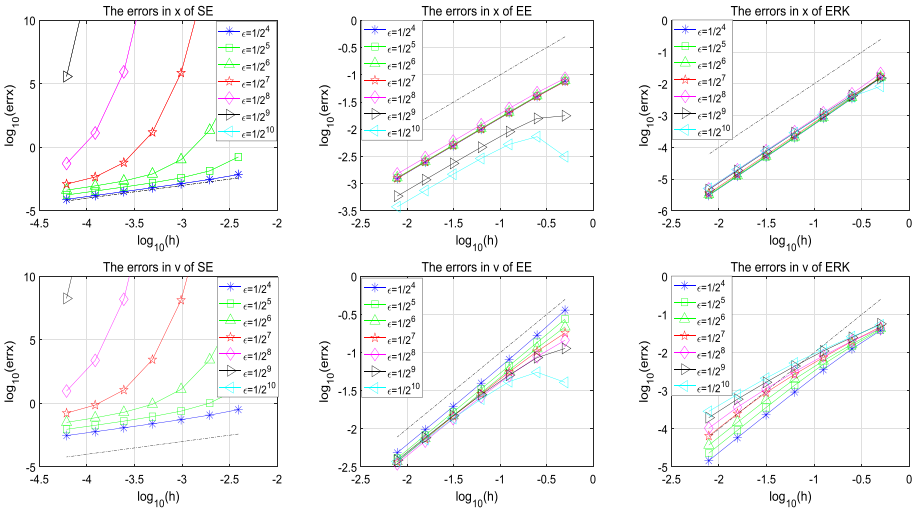

Fig. 14 The errors in x (errx) and v (errv) against h for the existing methods SE, EE and ERK (the slope of the dotted line for SE, EE is one and for ERK is two)

the system (1) can be rewritten as

$$\frac{d}{dt}\begin{pmatrix}\tilde{x}\\ \tilde{v}\end{pmatrix} = \begin{pmatrix}0 & I\\ 0 & \tilde{\Omega}\mathrm{i}\end{pmatrix}\begin{pmatrix}\tilde{x}\\ \tilde{v}\end{pmatrix} + \begin{pmatrix}0\\ \tilde{F}(\tilde{x})\end{pmatrix}, \quad \begin{pmatrix}\tilde{x}_0\\ \tilde{v}_0\end{pmatrix} = \begin{pmatrix}P^{\mathrm{H}}x_0\\ P^{\mathrm{H}}\dot{x}_0\end{pmatrix}, \tag{20}$$

where $\tilde{F}(\tilde{x}) = P^{\mathrm{H}}F(P\tilde{x}) = -\nabla_{\tilde{x}}U(P\tilde{x})$. In the rest parts of this paper, we denote the vector x by

$$x = \begin{cases}(x^{-l}, x^{-l+1}, \ldots, x^{-1}, x^0, x^1, \ldots, x^{l-1}, x^l)^{\intercal}, & d = 2l+1,\\ (x^{-l}, x^{-l+1}, \ldots, x^{-1}, x^1, \ldots, x^{l-1}, x^l)^{\intercal}, & d = 2l,\end{cases} \tag{21}$$

and the same notation is used for all the vectors in $\mathbb{R}^d$ or $\mathbb{C}^d$ and for the diagonal matrix in $\mathbb{R}^{d\times d}$ or $\mathbb{C}^{d\times d}$. For example, $\tilde{\Omega}^{-l}$ is referred to $-\tilde{\omega}_l$. According to (19) and the property of the unitary matrix P, one has that for $k = 1, 2, \ldots, l$

$$\tilde{x}^{-k} = \overline{(\tilde{x}^k)}, \quad \tilde{v}^{-k} = \overline{(\tilde{v}^k)}, \quad \tilde{x}^0,\ \tilde{v}^0 \in \mathbb{R} \ \text{ if they exist.} \tag{22}$$

The energy of this transformed system (20) is given by

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

For this transformed system, we can modify the schemes of SM1-SM3 accordingly. For example, the scheme (5) has a transformed form for (20)

$$\begin{aligned}\tilde{X}_i &= \tilde{x}_n + c_i h\varphi_1(c_i h\tilde{\Omega}\mathrm{i})\tilde{v}_n + h^2\sum_{j=1}^{s}\alpha_{ij}(h\tilde{\Omega}\mathrm{i})\tilde{F}(\tilde{X}_j), \quad i = 1, 2, \ldots, s,\\ \tilde{x}_{n+1} &= \tilde{x}_n + h\varphi_1(h\tilde{\Omega}\mathrm{i})\tilde{v}_n + h^2\sum_{i=1}^{s}\beta_i(h\tilde{\Omega}\mathrm{i})\tilde{F}(\tilde{X}_i),\\ \tilde{v}_{n+1} &= \varphi_0(h\tilde{\Omega}\mathrm{i})\tilde{v}_n + h\sum_{i=1}^{s}\gamma_i(h\tilde{\Omega}\mathrm{i})\tilde{F}(\tilde{X}_i).\end{aligned} \tag{23}$$

We summarise the relationships as follows:

$$\begin{array}{ccc} \text{System (1)} & \xleftrightarrow{P^H \text{or} P} & \text{Transformed System (20)} \\ \text{Method} \uparrow & & \uparrow \text{Method} \\ \text{SM1-SM3} & \xleftrightarrow{P^H \text{or} P} & \text{Transformed SM1-SM3} \end{array} \tag{24}$$

Denote the transformed method (23) by

$$\begin{aligned} \tilde{X}_i^J &= \tilde{x}_n^J + c_i h\varphi_1(c_i h\tilde{\Omega}^J \mathrm{i})\tilde{v}_n^J + h^2 \sum_{j=1}^{s} \alpha_{ij}(h\tilde{\Omega}^J \mathrm{i})\tilde{F}_j^J, \quad i = 1, 2, \ldots, s, \\ \tilde{x}_{n+1}^J &= \tilde{x}_n^J + h\varphi_1(h\tilde{\Omega}^J \mathrm{i})\tilde{v}_n^J + h^2 \sum_{i=1}^{s} \beta_i(h\tilde{\Omega}^J \mathrm{i})\tilde{F}_i^J, \\ \tilde{v}_{n+1}^J &= e^{h\tilde{\Omega}^J \mathrm{i}}\tilde{v}_n^J + h \sum_{i=1}^{s} \gamma_i(h\tilde{\Omega}^J \mathrm{i})\tilde{F}_i^J, \end{aligned} \tag{25}$$

where the superscript index J for $J = -l, -l+1, \ldots, l$ denotes the J-th entry of a vector or a matrix and $\tilde{F}_i^J$ denotes the J-th entry of $\tilde{F}(\tilde{X}_i)$ as the scheme of (21). It is noted that when $d = 2l$, $J = 0$ does not exist. With the notation of differential 2-form, we need to prove that (see [25])

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

We compute

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

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

$$\begin{aligned} &\sum_{J=-l}^{l} d\bar{\tilde{x}}_{n+1}^J \wedge d\tilde{v}_{n+1}^J - \tfrac{1}{2} \sum_{J=-l}^{l} d\bar{\tilde{x}}_{n+1}^J \wedge d(\tilde{\Omega}^J \mathrm{i}\tilde{x}_{n+1}^J) \\ &\quad = \sum_{J=-l}^{l} d\bar{\tilde{x}}_n^J \wedge d\tilde{v}_n^J - \tfrac{1}{2} \sum_{J=-l}^{l} d\bar{\tilde{x}}_n^J \wedge d(\tilde{\Omega}^J \mathrm{i}\tilde{x}_n^J). \end{aligned} \tag{26}$$

4.2 Symplecticity of the Transformed Methods

In this part, we will prove the following lemma.

Lemma 4.1 *The result* (26) *is true if the following conditions are satisfied*

$$\begin{aligned}
&\gamma_j(K) - K\beta_j(K) = d_j I, \quad d_j \in \mathbb{C},\\
&\gamma_j(K)[\bar{\varphi}_1(K) - c_j\bar{\varphi}_1(c_jK)] = \beta_j(K)[e^{-K} + K\bar{\varphi}_1(K) - c_jK\bar{\varphi}_1(c_jK)],\\
&\bar{\beta}_i(K)\gamma_j(K) - \frac{1}{2}K\bar{\beta}_i(K)\beta_j(K) - \bar{\alpha}_{ji}(K)[\gamma_j(K) - K\beta_j(K)]\\
&\quad = \beta_j(K)\bar{\gamma}_i(K) + \frac{1}{2}K\beta_j(K)\bar{\beta}_i(K) - \alpha_{ij}(K)[\bar{\gamma}_i(K) + K\bar{\beta}_i(K)],
\end{aligned} \tag{27}$$

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

Proof In view of the definition of differential 2-form (see [25]), it can be proved that for $J = -l, -l+1, \ldots, l$,

$$\overline{d\tilde{\bar{x}}_n^J \wedge d\tilde{v}_n^J} = d\tilde{x}_n^J \wedge d\tilde{\bar{v}}_n^J \text{ and } d\tilde{\bar{x}}_n^J \wedge d\tilde{x}_n^J \in \mathrm{i}\mathbb{R}.$$

In the light of the scheme (25) and the fact that any exterior product $\wedge$ appearing here is real, it is obtained that

$$\begin{aligned}
&d\tilde{\bar{x}}_{n+1}^J \wedge d\tilde{v}_{n+1}^J - \frac{1}{2}d\tilde{\bar{x}}_{n+1}^J \wedge d(\tilde{\Omega}^J\mathrm{i}\tilde{x}_{n+1}^J) = d\tilde{\bar{x}}_n^J \wedge d\tilde{v}_n^J - \frac{1}{2}d\tilde{\bar{x}}_n^J \wedge d(\tilde{\Omega}^J\mathrm{i}\tilde{x}_n^J)\\
&+h\sum_{j=1}^{s}[\gamma_j(K^J) - K^J\beta_j(K^J)]d\tilde{\bar{x}}_n^J \wedge d\tilde{F}_j^J\\
&+[he^{K^J}\bar{\varphi}_1(K^J) - \frac{1}{2}h^2\tilde{\Omega}^J\mathrm{i}\bar{\varphi}_1(K^J)\varphi_1(K^J)]d\tilde{\bar{v}}_n^J \wedge d\tilde{v}_n^J\\
&+h^2\sum_{j=1}^{s}[\bar{\varphi}_1(K^J)\gamma_j(K^J) - \beta_j(K^J)e^{-K^J} - h\tilde{\Omega}^J\mathrm{i}\bar{\varphi}_1(K^J)\beta_j(K^J)]d\tilde{\bar{v}}_n^J \wedge d\tilde{F}_j^J\\
&+h^3\sum_{i,j=1}^{s}[\bar{\beta}_i(K^J)\gamma_j(K^J) - \frac{1}{2}h\tilde{\Omega}^J\mathrm{i}\bar{\beta}_i(K^J)\beta_j(K^J)]d\tilde{\bar{F}}_i^J \wedge d\tilde{F}_j^J,
\end{aligned} \tag{28}$$

where the fact that $e^{K^J} - h\tilde{\Omega}^J\mathrm{i}\varphi_1(K^J) = I$ is used here. On the other hand, from the first s equalities of (25), it follows that

$$d\tilde{x}_n^J = d\tilde{X}_i^J - c_ih\varphi_1(c_iK^J)d\tilde{v}_n^J - h^2\sum_{j=1}^{s}\alpha_{ij}(K^J)d\tilde{F}_j^J$$

for $i = 1, 2, \ldots, s$. We then obtain for $j = 1, 2, \ldots, s$

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

Inserting this into (28) and summing over all J yields

$$\begin{aligned}
&\sum_{J=-l}^{l} d\bar{\tilde{x}}_{n+1}^J \wedge d\tilde{v}_{n+1}^J - \frac{1}{2}\sum_{J=-l}^{l} d\bar{\tilde{x}}_{n+1}^J \wedge d(\tilde{\Omega}^J \mathrm{i}\tilde{x}_{n+1}^J)\\
=\ &\sum_{J=-l}^{l} d\bar{\tilde{x}}_{n}^J \wedge d\tilde{v}_{n}^J - \frac{1}{2}\sum_{J=-l}^{l} d\bar{\tilde{x}}_{n}^J \wedge d(\tilde{\Omega}^J \mathrm{i}\tilde{x}_{n}^J)\\
&+h\sum_{j=1}^{s}\sum_{J=-l}^{l}[\gamma_j(K^J) - K^J\beta_j(K^J)]d\bar{\tilde{X}}_j^J \wedge d\tilde{F}_j^J \qquad (29a)\\
&+h\sum_{J=-l}^{l}[e^{K^J}\bar{\varphi}_1(K^J) - \frac{1}{2}K^J\bar{\varphi}_1(K^J)\varphi_1(K^J)]d\bar{\tilde{v}}_n^J \wedge d\tilde{v}_n^J \qquad (29b)\\
&+h^2\sum_{j=1}^{s}\sum_{J=-l}^{l}\Big[\bar{\varphi}_1(K^J)\gamma_j(K^J) - \beta_j(K^J)e^{-K^J} - K^J\bar{\varphi}_1(K^J)\beta_j(K^J) \qquad (29c)\\
&\qquad - c_j\bar{\varphi}_1(c_jK^J)[\gamma_j(K^J) - K^J\beta_j(K^J)]\Big]d\bar{\tilde{v}}_n^J \wedge d\tilde{F}_j^J \qquad (29d)\\
&+h^3\sum_{i,j=1}^{s}\sum_{J=-l}^{l}\Big[\bar{\beta}_i(K^J)\gamma_j(K^J) - \frac{1}{2}h\tilde{\Omega}^J\mathrm{i}\bar{\beta}_i(K^J)\beta_j(K^J) \qquad (29e)\\
&\qquad - \bar{\alpha}_{ji}(K^J)[\gamma_j(K^J) - K^J\beta_j(K^J)]\Big]d\bar{\tilde{F}}_i^J \wedge d\tilde{F}_j^J. \qquad (29f)
\end{aligned}$$

∘ Prove that (29a)=0.

Based on the first s conditions of (27), $\tilde{F}(\tilde{x}) = -\nabla_{\tilde{x}}U(P\tilde{x})$ and (26), it can be verified that $d\bar{\tilde{X}}_j^J \wedge d\tilde{F}_j^J = dX_j^J \wedge dF_j^J$. Thus, one has

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

∘ Prove that (29b)=0.

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

$$d\bar{\tilde{v}}_n^{-J} \wedge d\tilde{v}_n^{-J} = -d\bar{\tilde{v}}_n^J \wedge d\tilde{v}_n^J,\ \ d\bar{\tilde{v}}_n^0 \wedge d\tilde{v}_n^0 = 0,$$

and

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

Therefore, it follows that

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

∘ Prove that (29c)-(29f)= 0.

In the light of all the identities after the previous s ones in (27), the last two terms (29c)-(29f) vanish.

The results stated above lead to (26). □

Based on the results of Lemma 4.1, it can be verified straightforwardly that the coefficients of SM1-SM3 satisfy (27). Therefore, these methods are symplectic.

5 Analysis on Long-time Energy Conservation (Theorem 3.3)

In this section, we will show the long time near-conservation of energy presented in Theorem 3.3 along SM2 algorithm. We first derive modulated Fourier expansion (see, e.g. [10, 19, 22, 24]) 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 $\mathcal{L}(hD)$, $\hat{\mathcal{L}}(hD)$ (32) and following the way given below. We skip this proof for brevity.

5.1 Reformulation of SM2

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

$$\begin{cases} x_{n+1} - 2x_n + x_{n-1} = h(\varphi_1(h\Omega) - \varphi_1(-h\Omega))v_n + \frac{1}{2}h^2(\varphi_1(h\Omega) + \varphi_1(-h\Omega))F_n, \\ x_{n+1} - x_{n-1} = h(\varphi_1(h\Omega) + \varphi_1(-h\Omega))v_n + \frac{1}{2}h^2(\varphi_1(h\Omega) - \varphi_1(-h\Omega))F_n, \end{cases} \tag{30}$$

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

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

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 (20), it becomes

$$\tilde{\alpha}(h\tilde{\Omega})\tfrac{\tilde{x}_{n+1}-2\tilde{x}_n+\tilde{x}_{n-1}}{h^2} = \tilde{\beta}(h\tilde{\Omega})\mathrm{i}\tilde{\Omega}\tfrac{\tilde{x}_{n+1}-\tilde{x}_{n-1}}{2h} + \tilde{\gamma}(h\tilde{\Omega})\tilde{F}_n, \tag{31}$$

where the coefficient functions are given by $\tilde{\alpha}(\xi) = \frac{1}{\mathrm{sinc}^2(\frac{\xi}{2})}$, $\tilde{\beta}(\xi) = \frac{1}{\mathrm{sinc}(\xi)}$, $\tilde{\gamma}(\xi) = \xi\,\mathrm{csc}(\xi)$ with $\mathrm{sinc}(\xi) = \sin(\xi)/\xi$. Based on (31), we define the operator

$$\hat{\mathcal{L}}(hD) = \tilde{\alpha}(h\tilde{\Omega})\tfrac{e^{hD}-2+e^{-hD}}{h^2} - \tilde{\beta}(h\tilde{\Omega})\mathrm{i}\tilde{\Omega}\tfrac{e^{hD}-e^{-hD}}{2h}, \tag{32}$$

where D is the differential operator.

Before we derive the modulated Fourier expansion for SM2, we need the following notations. We collect the diagonal elements of $\tilde{\Omega}$ (18) in the vector

$$\varpi = \begin{cases} (-\tilde{\omega}_l, \ldots, -\tilde{\omega}_1, 0, \tilde{\omega}_1, \ldots, \tilde{\omega}_l)^\intercal, & d = 2l+1, \\ (-\tilde{\omega}_l, \ldots, -\tilde{\omega}_1, \tilde{\omega}_1, \ldots, \tilde{\omega}_l)^\intercal, & d = 2l. \end{cases}$$

It is noted that $\varpi = \mathcal{O}(1/\varepsilon)$. In this paper, the notation k is used to describe a vector in $\mathbb{R}^d$ and as stated in the previous section, its components are denoted by

$$k = \begin{cases} (k^{-l}, k^{-l+1}, \ldots, k^{-1}, k^0, k^1, \ldots, k^{l-1}, k^l)^\intercal, & d = 2l+1, \\ (k^{-l}, k^{-l+1}, \ldots, k^{-1}, k^1, \ldots, k^{l-1}, k^l)^\intercal, & d = 2l, \end{cases}$$

and the same notation is used for all the vectors with the same dimension as k. For example, ϖ^l is referred to $\tilde{\omega}_l$. In this paper, we also use the notations

$$|k| = \textstyle\sum_{j=-l}^{l} |k^j|, \quad k \cdot \varpi = \sum_{j=-l}^{l} k^j \varpi^j,$$

and the resonance module $\mathcal{M} = \{k \in \mathbb{Z}^D : k \cdot \varpi = 0\}$. Denote $\langle j \rangle$ by the unit coordinate vector $(0, \ldots, 0, 1, 0, \ldots, 0)^\intercal \in \mathbb{R}^d$ with the only entry 1 at the j-th position. The set $\mathcal{N}^*$ is defined as follows. For the resonance module $\mathcal{M}$, we let $\mathcal{K}$ be a set of representatives of the equivalence classes in $\mathbb{Z}^l \backslash \mathcal{M}$ 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, $\mathcal{N} = \{k \in \mathcal{K} : |k| \le N\}$, $\mathcal{N}^* = \mathcal{N} \backslash \{(0, \ldots, 0)\}$.

5.2 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 (20). In the rest parts of this paper, we consider $d = 2l + 1$ and all the analysis can be modified to $d = 2l$ without any difficulty.

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

$$\tilde{x}_h(t) = \sum_{k \in \mathcal{N}^*} \mathrm{e}^{\mathrm{i}(k \cdot \varpi)t} \tilde{\zeta}_k(t) + \tilde{R}_{h,N}(t), \quad \tilde{v}_h(t) = \sum_{k \in \mathcal{N}^*} \mathrm{e}^{\mathrm{i}(k \cdot \varpi)t} \tilde{\eta}_k(t) + \tilde{S}_{h,N}(t), \tag{33}$$

yield a small defect $\tilde{R}$, $\tilde{S}$ when they are inserted into the numerical scheme (31).

Lemma 5.1 *Under the conditions of Theorem 3.3 and for $t = nh$, the numerical results $\tilde{x}_n$ and $\tilde{v}_n$ produced by (31) satisfy*

$$\tilde{x}_n = \sum_{k \in \mathcal{N}^*} \mathrm{e}^{\mathrm{i}(k \cdot \varpi)t} \tilde{\zeta}_k(t) + \tilde{R}_{h,N}(t), \quad \tilde{v}_n = \sum_{k \in \mathcal{N}^*} \mathrm{e}^{\mathrm{i}(k \cdot \varpi)t} \tilde{\eta}_k(t) + \tilde{S}_{h,N}(t). \tag{34}$$

The coefficient functions $\tilde{\zeta}_k$ have the bounds

$$\begin{array}{llllll} \tilde{\zeta}^{\pm j}_{\langle 0 \rangle} = \mathcal{O}(1), & \tilde{\zeta}^{0}_{\langle 0 \rangle} = \mathcal{O}(1), & \tilde{\zeta}^{j}_{\langle j \rangle} = \mathcal{O}(h), & \tilde{\zeta}^{-j}_{\langle -j \rangle} = \mathcal{O}(h), \\ \tilde{\zeta}^{-j}_{\langle j \rangle} = \mathcal{O}(h^{\frac{5}{2}}), & \tilde{\zeta}^{0}_{\langle j \rangle} = \mathcal{O}(h^2), & \tilde{\zeta}^{0}_{\langle -j \rangle} = \mathcal{O}(h^2), & \tilde{\zeta}^{j}_{\langle -j \rangle} = \mathcal{O}(h^{\frac{5}{2}}), \end{array} \tag{35}$$

and we have the following results for the coefficient functions $\tilde{\eta}_k$

$$\begin{array}{ll} \tilde{\eta}^{0}_{\langle 0 \rangle} = \dot{\tilde{\zeta}}^{0}_{\langle 0 \rangle} + \mathcal{O}(h), & \tilde{\eta}^{\pm j}_{\langle 0 \rangle} = \frac{h\tilde{\omega}_j}{\sin(h\tilde{\omega}_j)} \dot{\tilde{\zeta}}^{\pm j}_{\langle 0 \rangle} + \mathcal{O}(h), \\ \tilde{\eta}^{0}_{\langle \pm j \rangle} = \mathrm{i}\tilde{\omega}_j \operatorname{sinc}(h\tilde{\omega}_j) \tilde{\zeta}^{0}_{\langle \pm j \rangle} + \mathcal{O}(h), & \tilde{\eta}^{j}_{\langle j \rangle} = \mathrm{i}\tilde{\omega}_j \tilde{\zeta}^{j}_{\langle j \rangle} + \mathcal{O}(h), \ \tilde{\eta}^{-j}_{\langle -j \rangle} = -\mathrm{i}\tilde{\omega}_j \tilde{\zeta}^{-j}_{\langle -j \rangle} + \mathcal{O}(h). \end{array} \tag{36}$$

A further result is true

$$\tilde{\zeta}_k = \mathcal{O}(h^{|k|+1}), \quad \tilde{\eta}_k = \mathcal{O}(h^{|k|}) \quad \text{for} \ \ |k| > 1. \tag{37}$$

The remainders in (34) are bounded by

$$\tilde{R}_{h,N}(t) = \mathcal{O}(t^2 h^N), \qquad \tilde{S}_{h,N}(t) = \mathcal{O}(t^2 h^N / \varepsilon). \tag{38}$$

The constants symbolised by the notation $\mathcal{O}$ are independent of h, ε, but depend on c_0, c appeared in the conditions of Theorem 3.3.

Proof The proof is presented by using the technique of the modulated Fourier expansion [19, 25] and it involves the construction of the series and the analysis of the coefficients and the truncation. For brevity, we present the key steps here and the details are referred to some related works [21–24, 42, 43].

• **Proof of** (34). Inserting the first expansion of (33) into the two-step form (31), expanding the nonlinear function into its Taylor series and comparing the coefficients of $\mathrm{e}^{\mathrm{i}(k \cdot \varpi)t}$, we

obtain

$$\begin{aligned}\hat{\mathcal{L}}(hD)\tilde{\zeta}_{\langle 0\rangle} &= \tilde{\gamma}(h\tilde{\Omega})\Big(\tilde{F}(\tilde{\zeta}_{\langle 0\rangle}) + \sum_{s(\alpha)=0}\frac{1}{m!}\tilde{F}^{(m)}(\tilde{\zeta}_{\langle 0\rangle})(\tilde{\zeta})_\alpha\Big),\\ \hat{\mathcal{L}}(hD+\mathrm{i}h(k\cdot\varpi))\tilde{\zeta}_k &= \tilde{\gamma}(h\tilde{\Omega})\sum_{s(\alpha)=k}\frac{1}{m!}\tilde{F}^{(m)}(\tilde{\zeta}_{\langle 0\rangle})(\tilde{\zeta})_\alpha,\quad |k|>0,\end{aligned} \tag{39}$$

where the sum ranges over $m \geq 0$, $s(\alpha) = \sum_{j=1}^{m}\alpha_j$ with $\alpha = (\alpha_1,\ldots,\alpha_m)$ and $0 < |\alpha_i| < N$, and $(\tilde{\zeta})_\alpha$ is an abbreviation for $(\tilde{\zeta}_{\alpha_1},\ldots,\tilde{\zeta}_{\alpha_m})$. 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 $\hat{\mathcal{L}}$

$$\begin{aligned}&\hat{\mathcal{L}}(hD) = -\tilde{\Omega}^2\csc(h\tilde{\Omega})(\mathrm{i}hD) - \frac{1}{4}\tilde{\Omega}^2\csc^2\big(\frac{1}{2}h\tilde{\Omega}\big)(\mathrm{i}hD)^2+\ldots,\\ &\hat{\mathcal{L}}(hD+\mathrm{i}h(k\cdot\varpi)) = 2\sin\big(\frac{1}{2}h(k\cdot\varpi)\big)\tilde{\Omega}^2\csc\big(\frac{1}{2}h\tilde{\Omega}\big)\csc(h\tilde{\Omega})\sin\big(\frac{1}{2}h(\tilde{\Omega}-(k\cdot\varpi)I)\big)\\ &\quad-\sin\big(\frac{1}{2}h(\tilde{\Omega}-2(k\cdot\varpi))\big)\tilde{\Omega}^2\csc\big(\frac{1}{2}h\tilde{\Omega}\big)\csc(h\tilde{\Omega})(\mathrm{i}hD)+\ldots,\\ &\hat{\mathcal{L}}(hD+\mathrm{i}h(\langle\pm j\rangle\cdot\varpi))^{\pm j} = \pm\tilde{\omega}_j^2\csc(h\tilde{\omega}_j)(\mathrm{i}hD) - \frac{1}{4}\tilde{\omega}_j^2\csc^2(h\tilde{\omega}_j/2)(\mathrm{i}hD)^2+\ldots,\end{aligned}$$

the following ansatz of $\tilde{\zeta}_k$ can be obtained:

$$\begin{aligned}&\dot{\tilde{\zeta}}^{\pm j}_{\langle 0\rangle} = \tfrac{-h^2\tilde{\omega}_j A(h\tilde{\omega}_j)}{8\mathrm{i}\sin^2(\frac{1}{2}h\tilde{\omega}_j)}\big(\mathcal{F}^{\pm j0}(\cdot)+\ldots\big), && \ddot{\tilde{\zeta}}^{0}_{\langle 0\rangle} = \mathcal{F}^{00}(\cdot)+\ldots,\\ &\dot{\tilde{\zeta}}^{j}_{\langle j\rangle} = \tfrac{h^2\tilde{\omega}_j A(h\tilde{\omega}_j)}{8\mathrm{i}\sin^2(\frac{1}{2}h\tilde{\omega}_j)}\big(\mathcal{F}^{j0}_{j}(\cdot)+\ldots\big), && \dot{\tilde{\zeta}}^{-j}_{\langle -j\rangle} = \tfrac{h^2\tilde{\omega}_j A(h\tilde{\omega}_j)}{-8\mathrm{i}\sin^2(\frac{1}{2}h\tilde{\omega}_j)}\big(\mathcal{F}^{-j0}_{-j}(\cdot)+\ldots\big),\\ &\tilde{\zeta}_k = \tfrac{h^3\tilde{\Omega}A(h\tilde{\Omega})}{16\sin(\frac{1}{2}h\tilde{\Omega})\sin(\frac{1}{2}h(\tilde{\Omega}-(k\cdot\varpi)I))\sin(\frac{1}{2}h(k\cdot\varpi)I)}\ \big(\mathcal{F}^{0}_{k}(\cdot)+\ldots\big)\ \text{for}\ |k|>1,\end{aligned} \tag{40}$$

where $j = 1,2,\ldots,l$, $A(x) = 2\,\mathrm{sinc}^2(\frac{1}{2}x)$, all the $\mathcal{F}$ and so on are formal series, and the dots stand for power series in $\sqrt{h}$. In this paper we truncate the ansatz after the $\mathcal{O}(h^{N+1})$ terms. On the basis of the second formula of (30), one has

$$\begin{aligned}\tilde{v}_n &= \tfrac{1}{h(\varphi_1(\mathrm{i}h\tilde{\Omega})+\varphi_1(-\mathrm{i}h\tilde{\Omega}))}(\tilde{x}_{n+1}-\tilde{x}_{n-1}) - \tfrac{1}{2}h^2\tfrac{\varphi_1(\mathrm{i}h\tilde{\Omega})-\varphi_1(-\mathrm{i}h\tilde{\Omega})}{h(\varphi_1(\mathrm{i}h\tilde{\Omega})+\varphi_1(-\mathrm{i}h\tilde{\Omega}))}\tilde{F}(\tilde{x}_n)\\ &= \tfrac{1}{2h\,\mathrm{sinc}(h\tilde{\Omega})}(\tilde{x}_{n+1}-\tilde{x}_{n-1}) - \tfrac{1}{2}\mathrm{i}h\tan(\tfrac{h}{2}\tilde{\Omega})\tilde{F}(\tilde{x}_n).\end{aligned} \tag{41}$$

Inserting (33) into (41), expanding the nonlinear function into its Taylor series and comparing the coefficients of $\mathrm{e}^{\mathrm{i}(k\cdot\varpi)t}$, one arrives

$$\begin{aligned}\tilde{\eta}_{\langle 0\rangle} &= \mathcal{L}(hD)\tilde{\zeta}_{\langle 0\rangle} - \frac{1}{2}\mathrm{i}h\tan(\frac{h}{2}\tilde{\Omega})\Big(\tilde{F}(\tilde{\zeta}_{\langle 0\rangle}) + \sum_{s(\alpha)=0}\frac{1}{m!}\tilde{F}^{(m)}(\tilde{\zeta}_{\langle 0\rangle})(\tilde{\zeta})_\alpha\Big),\\ \tilde{\eta}_k &= \mathcal{L}(hD+\mathrm{i}h(k\cdot\varpi))\tilde{\zeta}_k - \frac{1}{2}\mathrm{i}h\tan(\frac{h}{2}\tilde{\Omega})\sum_{s(\alpha)=k}\frac{1}{m!}\tilde{F}^{(m)}(\tilde{\zeta}_{\langle 0\rangle})(\tilde{\zeta})_\alpha,\quad |k|>0.\end{aligned} \tag{42}$$

where

$$\mathcal{L}(hD) = \frac{1}{2h\,\mathrm{sinc}(h\tilde{\Omega})}(e^{hD}-e^{-hD}).$$

In a same way as [21–24, 42, 43], this formula gives the modulation system for the coefficients $\tilde{\eta}_k$ of the modulated Fourier expansion by choosing the dominating terms and by the Taylor

expansion of $\mathcal{L}$

$$\begin{aligned}
&\mathcal{L}(hD) = \tilde{\Omega}\csc(h\tilde{\Omega})(hD) + \frac{1}{6}\tilde{\Omega}\csc(h\tilde{\Omega})(hD)^3 + \ldots,\\
&\mathcal{L}(hD + \mathrm{i}h(\langle \pm j\rangle \cdot \varpi))^{\pm j} = \pm \mathrm{i}\tilde{\omega}_j + \tilde{\omega}_j \cot(h\tilde{\omega}_j)(hD) + \ldots,\\
&\mathcal{L}(hD + \mathrm{i}h(k \cdot \varpi)) = \mathrm{i}\sin(h(k \cdot \varpi))\tilde{\Omega}\csc(h\tilde{\Omega}) + \cos(h(k \cdot \varpi))\tilde{\Omega}\csc(h\tilde{\Omega}) + \ldots.
\end{aligned}$$

Under the above analysis, the construction of $\tilde{\zeta}_k$ and $\tilde{\eta}_k$ is presented and this proves (34).

• **Proof of** (35)-(37). For the first-order and second-order differential equations appeared in (40), 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

$$\begin{aligned}
&\tilde{x}_0^0 = \tilde{\zeta}_{\langle 0\rangle}^0(0) + \mathcal{O}(\varepsilon), \quad \tilde{x}_0^{\pm j} = \tilde{\zeta}_{\langle 0\rangle}^{\pm j}(0) + \mathcal{O}(\varepsilon),\\
&\tilde{v}_0^0 = \tilde{\eta}_{\langle 0\rangle}^0(0) + \mathcal{O}(\varepsilon) = \dot{\tilde{\zeta}}_{\langle 0\rangle}^0(0) + \mathcal{O}(\varepsilon),\\
&\tilde{v}_0^j = \tilde{\eta}_{\langle 0\rangle}^j(0) + \tilde{\eta}_{\langle j\rangle}^j(0) + \mathcal{O}(\varepsilon) = \dot{\tilde{\zeta}}_{\langle 0\rangle}^j(0) + \mathrm{i}\tilde{\omega}_j\tilde{\zeta}_{\langle j\rangle}^j(0) + \mathcal{O}(\varepsilon),\\
&\tilde{v}_0^{-j} = \tilde{\eta}_{\langle 0\rangle}^{-j}(0) + \tilde{\eta}_{\langle -j\rangle}^{-j}(0) + \mathcal{O}(\varepsilon) = \dot{\tilde{\zeta}}_{\langle 0\rangle}^{-j}(0) - \mathrm{i}\tilde{\omega}_j\tilde{\zeta}_{\langle -j\rangle}^{-j}(0) + \mathcal{O}(\varepsilon).
\end{aligned} \tag{43}$$

Thus the initial values $\tilde{\zeta}_{\langle 0\rangle}^0(0) = \mathcal{O}(1)$ and $\dot{\tilde{\zeta}}_{\langle 0\rangle}^0(0) = \mathcal{O}(1)$ can be derived by considering the first and third formulae, respectively. According to the second equation of (43), one gets the initial value $\tilde{\zeta}_{\langle 0\rangle}^{\pm j}(0) = \mathcal{O}(1)$. It follows from the fourth formula that $\tilde{\zeta}_{\langle j\rangle}^j(0) = \frac{1}{\mathrm{i}\tilde{\omega}_j}\big(\tilde{v}_{\langle 0\rangle}^j - \dot{\tilde{\zeta}}_{\langle 0\rangle}^j(0) + \mathcal{O}(\varepsilon)\big) = \mathcal{O}(\varepsilon)$, and likewise one has $\tilde{\zeta}_{\langle -j\rangle}^{-j}(0) = \mathcal{O}(\varepsilon)$.

With the ansatz (40), we achieve the bounds

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

According to the initial values stated above, the bounds

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

are obtained. Moreover, based on the above bounds and the relationship (42), the coefficient functions $\tilde{\eta}_k$ are bounded as (36). With these results and considering (40) and (42) for $|k| > 1$, a further result (37) is deduced by the same arguments in [23, 24, 42, 43].

• **Proof of** (38). Define

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

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 (41), 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:

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

where $\Gamma_n := \int_0^1 \tilde{F}_x(\tilde{x}_n + \tau \tilde{R}_n) d\tau$. Similar to [23, 24, 42, 43], solving this recursion and the application of a discrete Gronwall inequality gives (38). □

Using the relationship shown in (24), the modulated Fourier expansion of the method SM2 is immediately obtained.

Lemma 5.2 *The numerical results x_n and v_n produced by SM2 admits the following modulated Fourier expansion*

$$x_n = \sum_{k\in\mathcal{N}^*} \mathrm{e}^{\mathrm{i}(k\cdot\varpi)t}\zeta_k(t) + \mathcal{O}(t^2h^N), \quad v_n = \sum_{k\in\mathcal{N}^*} \mathrm{e}^{\mathrm{i}(k\cdot\varpi)t}\eta_k(t) + \mathcal{O}(t^2h^N/\varepsilon),$$

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

5.3 An Almost-Invariant

Denote $\vec{\tilde{\zeta}} = \left(\tilde{\zeta}_k\right)_{k\in\mathcal{N}^*}$. An almost-invariant of the modulated Fourier expansion (33) is given as follows.

Lemma 5.3 *There exists a function $\mathcal{E}[\vec{\tilde{\zeta}}](t)$ such that*

$$\mathcal{E}[\vec{\tilde{\zeta}}](t) = \mathcal{E}[\vec{\tilde{\zeta}}](0) + \mathcal{O}(th^N). \tag{44}$$

Meanwhile, this function has the form

$$\mathcal{E}[\vec{\tilde{\zeta}}](t) = \frac{1}{2}\left|\dot{\tilde{\zeta}}^0_{\langle 0\rangle}(t)\right|^2 + \frac{1}{2}\sum_{j=1}^{l}\tilde{\omega}_j{}^2\left(\left|\tilde{\zeta}^j_{\langle j\rangle}(t)\right|^2 + \left|\tilde{\zeta}^{-j}_{\langle -j\rangle}(t)\right|^2\right) + U(P^{\mathrm{H}}\tilde{\zeta}_{\langle 0\rangle}(t)) + \mathcal{O}(h).$$

Proof According to the analysis of Sect. 5.2, 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 $\tilde{x}_h = \sum_{k\in\mathcal{N}^*}\tilde{x}_{h,k}$ with $\tilde{x}_{h,k} = \mathrm{e}^{\mathrm{i}(k\cdot\varpi)t}\tilde{\zeta}_k$. Multiplication of this result with P yields

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

where $x_h = \sum_{k\in\mathcal{N}^*} x_{h,k}$ with $x_{h,k} = \mathrm{e}^{\mathrm{i}(k\cdot\varpi)t}\zeta_k$. Rewrite the equation in terms of $x_{h,k}$ and then one has

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

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 $\vec{x} = \left(x_{h,k}\right)_{k\in\mathcal{N}^*}$. Multiplying this equation with $\left(\dot{x}_{h,-k}\right)^{\intercal}$ and summing up yields

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

Denoting $\vec{\zeta} = (\zeta_k)_{k\in\mathcal{N}^*}$ and switching to the quantities ζ^k, we obtain

$$\begin{aligned}\mathcal{O}(h^N) &= \sum_{k\in\mathcal{N}^*} \big(\dot{\zeta}_{-k} - \mathrm{i}(k\cdot\varpi)\zeta_{-k}\big)^{\intercal} P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + \mathrm{i}h(k\cdot\varpi))P^{\mathrm{H}}\zeta_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}) \\ &= \sum_{k\in\mathcal{N}^*} \big(\dot{\overline{\zeta}}_k - \mathrm{i}(k\cdot\varpi)\overline{\zeta_k}\big)^{\intercal} P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + \mathrm{i}h(k\cdot\varpi))P^{\mathrm{H}}\zeta_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}) \\ &= \sum_{k\in\mathcal{N}^*} \big(\dot{\overline{\tilde{\zeta}}}_k - \mathrm{i}(k\cdot\varpi)\overline{\tilde{\zeta}_k}\big)^{\intercal} P^{\mathrm{H}}P\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + \mathrm{i}h(k\cdot\varpi))P^{\mathrm{H}}P\tilde{\zeta}_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}) \\ &= \sum_{k\in\mathcal{N}^*} \big(\dot{\overline{\tilde{\zeta}}}_k - \mathrm{i}(k\cdot\varpi)\overline{\tilde{\zeta}_k}\big)^{\intercal} \tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + \mathrm{i}h(k\cdot\varpi))\tilde{\zeta}_k + \frac{d}{dt}\mathcal{U}(\vec{\zeta}).\end{aligned} \tag{45}$$

In what follows, we show that the right-hand side of (45) is the total derivative of an expression that depends only on $\tilde{\zeta}_k$ and derivatives thereof. Consider $k = \langle 0\rangle$ and then it follows that

$$\hat{\mathcal{L}}(hD)\tilde{\zeta}_{\langle 0\rangle} = \mathrm{i}hM_1\dot{\tilde{\zeta}}_{\langle 0\rangle} + h^2M_2\ddot{\tilde{\zeta}}_{\langle 0\rangle} + \mathrm{i}h^3M_3\dddot{\tilde{\zeta}}_{\langle 0\rangle} + \ldots, \text{ where } M_n \in \mathbb{R}^{d\times d} \text{ for } n = 1, 2, \ldots.$$

By the formulae given on p. 508 of [25], we know that $\mathrm{Re}\big(\dot{\overline{\tilde{\zeta}}}_{\langle 0\rangle}\big)^{\intercal}\hat{\mathcal{L}}(hD)\tilde{\zeta}_{\langle 0\rangle}$ is a total derivative. For $k \neq \langle 0\rangle$, in the light of

$$\hat{\mathcal{L}}(hD + \mathrm{i}h(k\cdot\varpi))\tilde{\zeta}_k = N_0\tilde{\zeta}_k + \mathrm{i}hN_1\dot{\tilde{\zeta}}_k + h^2N_2\ddot{\tilde{\zeta}}_k + \mathrm{i}h^3N_3\dddot{\tilde{\zeta}}_k + \ldots,$$

where $N_n \in \mathbb{R}^{d\times d}$ for $n = 0, 1, \ldots$, it is easy to check that $\mathrm{Re}\big(\dot{\overline{\tilde{\zeta}}}_k\big)^{\intercal}\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + \mathrm{i}h(k\cdot\varpi))\tilde{\zeta}_k$ and $\mathrm{Re}\big(\mathrm{i}(k\cdot\varpi)\overline{\tilde{\zeta}_k}\big)^{\intercal}\tilde{\gamma}^{-1}(h\tilde{\Omega})\hat{\mathcal{L}}(hD + \mathrm{i}h(k\cdot\varpi))\tilde{\zeta}_k$ are both total derivatives. Therefore, there exists a function $\mathcal{E}$ such that $\frac{d}{dt}\mathcal{E}[\vec{\tilde{\zeta}}](t) = \mathcal{O}(h^N)$. It follows from an integration that (44) holds.

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

$$\begin{aligned}\mathcal{E}[\vec{\tilde{\zeta}}](t) &= \frac{1}{2}(\dot{\overline{\tilde{\zeta}}}_{\langle 0\rangle}(t))^{\intercal}\frac{2\,\mathrm{sinc}(\frac{1}{2}h\tilde{\Omega})}{\varphi_1(\mathrm{i}h\tilde{\Omega})\varphi_1(-\mathrm{i}h\tilde{\Omega}) + \varphi_1(-\mathrm{i}h\tilde{\Omega})\varphi_1(\mathrm{i}h\tilde{\Omega})}\dot{\tilde{\zeta}}_{\langle 0\rangle}(t) + \mathcal{U}(\vec{\zeta}(t)) + \mathcal{O}(h^2) \\ &\quad + \frac{1}{2}\sum_{j=1}^{l}\frac{\tilde{\omega}_j}{h}h\tilde{\omega}_j\frac{2\,\mathrm{sinc}^2(\frac{1}{2}h\tilde{\omega}_j)}{\varphi_1(\mathrm{i}h\tilde{\omega}_j)\varphi_1(-\mathrm{i}h\tilde{\omega}_j) + \varphi_1(-\mathrm{i}h\tilde{\omega}_j)\varphi_1(\mathrm{i}h\tilde{\omega}_j)}\big(\big|\tilde{\zeta}^{j}_{\langle j\rangle}(t)\big|^2 + \big|\tilde{\zeta}^{-j}_{\langle -j\rangle}(t)\big|^2\big) \\ &= \frac{1}{2}\big|\dot{\tilde{\zeta}}^{0}_{\langle 0\rangle}(t)\big|^2 + \frac{1}{2}\sum_{j=1}^{l}\tilde{\omega}_j^2\big(\big|\tilde{\zeta}^{j}_{\langle j\rangle}(t)\big|^2 + \big|\tilde{\zeta}^{-j}_{\langle -j\rangle}(t)\big|^2\big) + U(P^{\mathrm{H}}\tilde{\zeta}_{\langle 0\rangle}(t)) + \mathcal{O}(h).\end{aligned}$$

□

5.4 Long-Time Near-conservation

Considering the result of $\mathcal{E}$ and the relationship (36) between $\tilde{\zeta}$ and $\tilde{\eta}$, we obtain

$$\begin{aligned}\mathcal{E}[\vec{\tilde{\zeta}}](t_n) &= \frac{1}{2}\big|\dot{\tilde{\zeta}}^{0}_{\langle 0\rangle}(t_n)\big|^2 + \frac{1}{2}\sum_{j=1}^{l}\tilde{\omega}_j^2\big(\big|\tilde{\zeta}^{j}_{\langle j\rangle}(t_n)\big|^2 + \big|\tilde{\zeta}^{-j}_{\langle -j\rangle}(t_n)\big|^2\big) + U(P^{\mathrm{H}}\tilde{\zeta}_{\langle 0\rangle}(t_n)) + \mathcal{O}(h) \\ &= \frac{1}{2}\big|\tilde{\eta}^{0}_{\langle 0\rangle}(t_n)\big|^2 + \frac{1}{2}\sum_{j=1}^{l}\big(\big|\tilde{\eta}^{j}_{\langle j\rangle}(t_n)\big|^2 + \big|\tilde{\eta}^{-j}_{\langle -j\rangle}(t_n)\big|^2\big) + U(P^{\mathrm{H}}\tilde{\zeta}_{\langle 0\rangle}(t_n)) + \mathcal{O}(h).\end{aligned} \tag{46}$$

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

$$E(x_n, v_n) = \tilde{E}(\tilde{x}_n, \tilde{v}_n) = \frac{1}{2}\left|\tilde{\eta}^0_{\langle 0\rangle}(t_n)\right|^2 + \frac{1}{2}\sum_{j=1}^{l}\left(\left|\tilde{\eta}^j_{\langle j\rangle}(t_n)\right|^2 + \left|\tilde{\eta}^{-j}_{\langle -j\rangle}(t_n)\right|^2\right) + U(P^{\mathrm{H}}\tilde{\zeta}_{\langle 0\rangle}(t_n)) + \mathcal{O}(h). \tag{47}$$

A comparison between (46) and (47) gives $\mathcal{E}[\vec{\tilde{\zeta}}](t_n) = E(x_n, v_n) + \mathcal{O}(h)$. Based on (44) and this result, the statement (14) is easily obtained by considering $nh^N \leq 1$ and

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

6 Analysis on Convergence (Theorem 3.6)

In this section, we discuss the convergence of the algorithms stated in Theorem 3.6. The proof will be firstly given for EM1, M1 and PM by using the averaging technique and then presented for SM1-SM3 by using modulated Fourier expansion.

6.1 Proof for EM1, M1 and PM

6.1.1 Re-scaled System and Methods

In order to establish the uniform error bounds, the strategy developed in [9, 42] 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

$$\begin{aligned} &\dot{q}(\tau) = \varepsilon w(\tau), \quad \dot{w}(\tau) = \tilde{B}w(\tau) + \varepsilon F(q(\tau)), \ \tau \in [0, \frac{T}{\varepsilon}], \\ &q(0) = q_0 := x_0, \ w(0) = w_0 := v_0, \end{aligned} \tag{48}$$

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 (48), the method EM1 becomes

$$\begin{aligned} q_{n+1} =& q_n + \varepsilon\Delta\tau\varphi_1(\Delta\tau\tilde{B})w_n + \frac{\Delta\tau^2\varepsilon^2}{2}\varphi_2(\Delta\tau\tilde{B})\int_0^1 F\big(\rho q_n + (1-\rho)q_{n+1}\big)d\rho, \\ w_{n+1} =& \mathrm{e}^{\Delta\tau\tilde{B}}w_n + \Delta\tau\varepsilon\varphi_1(\Delta\tau\tilde{B})\int_0^1 F\big(\rho q_n + (1-\rho)q_{n+1}\big)d\rho, \quad 0 \leq n < \frac{T}{\varepsilon\Delta\tau}. \end{aligned} \tag{49}$$

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. The variation-of-constants formula of (48) reads

$$\begin{aligned} q(\tau_n + \Delta\tau) &= q(\tau_n) + \varepsilon\Delta\tau\varphi_1(\Delta\tau\Omega)w(\tau_n) \\ &\quad + \varepsilon^2\Delta\tau^2\int_0^1 (1-\rho)\varphi_1((1-\rho)\Delta\tau\Omega)F(q(\tau_n+\rho\Delta\tau))d\rho, \\ w(\tau_n + \Delta\tau) &= \varphi_0(\Delta\tau\Omega)w(\tau_n) + \varepsilon\Delta\tau\int_0^1 \varphi_0((1-\rho)\Delta\tau\Omega)F(q(\tau_n+\rho\Delta\tau))d\rho. \end{aligned} \tag{50}$$

6.1.2 Local Truncation Errors

Based on (49), the local truncation errors ξ_n^q and ξ_n^w of EM1 for $0 \le n < \frac{T}{\varepsilon\Delta\tau}$ are defined as

$$\begin{aligned} q(\tau_{n+1}) =& q(\tau_n) + \Delta\tau\varepsilon\varphi_1(\Delta\tau\tilde{B})w(\tau_n) \\ &+ \frac{\Delta\tau^2\varepsilon^2}{2}\varphi_2(\Delta\tau\tilde{B})\int_0^1 F\left(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})\right)d\rho + \xi_n^q, \\ w(\tau_{n+1}) =& \mathrm{e}^{\Delta\tau\tilde{B}}w(\tau_n) + \Delta\tau\varepsilon\varphi_1(\Delta\tau\tilde{B})\int_0^1 F\left(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})\right)d\rho + \xi_n^w. \end{aligned} \tag{51}$$

By this result and the variation-of-constants formula of (48), we compute

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

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

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

Consequently, the local error becomes

$$\xi_n^w = \ \varepsilon^2\Delta\tau^2\big(\varphi_2(\Delta\tau\tilde{B}) - \frac{1}{2}\varphi_1(\Delta\tau\tilde{B})\big)\frac{\partial F}{\partial q}(q(\tau_n))w(\tau_n) + \mathcal{O}(\varepsilon^2\Delta\tau^3) = \mathcal{O}(\varepsilon^2\Delta\tau^3), \tag{52}$$

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

$$\xi_n^q = \mathcal{O}(\varepsilon^3\Delta\tau^3). \tag{53}$$

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

$$\xi_n^w = \mathcal{O}(\varepsilon^2\Delta\tau^2),\ \ \xi_n^q = \mathcal{O}(\varepsilon^2\Delta\tau^2). \tag{54}$$

6.1.3 Error Bounds

Define the error of the considered scheme

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

We first prove the error bounds of re-scaled EM1 and M1.

Lemma 6.2 *The convergence of re-scaled EM1 is given by*

$$|e_{n+1}^q| + |\varepsilon e_{n+1}^w| \lesssim \varepsilon^2\Delta\tau^2, \quad 0 \le n < \frac{T}{\varepsilon\Delta\tau}. \tag{55}$$

For the re-scaled M1, its global error becomes

$$|e_{n+1}^q| + |e_{n+1}^w| \lesssim \varepsilon\Delta\tau, \quad 0 \le n < \frac{T}{\varepsilon\Delta\tau}. \tag{56}$$

Proof 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 $0 < \Delta\tau \le \Delta\tau_0$, the following inequalities are true:

$$|q_n| \le \|q\|_{L^\infty(0,T/\varepsilon)} + 1, \quad |w_n| \le \|w\|_{L^\infty(0,T/\varepsilon)} + 1, \quad 0 \le n \le \frac{T}{\varepsilon\Delta\tau}. \tag{57}$$

For $n = 0$, (57) is obviously true. Then we assume that (57) is true up to some $0 \le m < \frac{T}{\varepsilon\Delta\tau}$, and we shall show that (57) holds for $m + 1$.

For $n \le m$, subtracting (51) from the scheme (49) leads to

$$e_{n+1}^q = e_n^q + \Delta\tau\varepsilon\varphi_1(\Delta\tau\tilde{B})e_n^w + \eta_n^q + \xi_n^q, \; e_{n+1}^w = \mathrm{e}^{\Delta\tau\tilde{B}}e_n^w + \eta_n^w + \xi_n^w, \; 0 \le n \le m, \tag{58}$$

where we use the following notations

$$\eta_n^q = \frac{\Delta\tau^2\varepsilon^2}{2}\varphi_2(\Delta\tau\tilde{B})\int_0^1 \left[F\left(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})\right) - F\left(\rho q_n + (1-\rho)q_{n+1}\right)\right] d\rho,$$

$$\eta_n^w = \Delta\tau\varepsilon\varphi_1(\Delta\tau\tilde{B})\int_0^1 \left[F\left(\rho q(\tau_n) + (1-\rho)q(\tau_{n+1})\right) - F\left(\rho q_n + (1-\rho)q_{n+1}\right)\right] d\rho.$$

From the induction assumption of the boundedness, it follows that

$$|\eta_n^q| \lesssim \Delta\tau^2\varepsilon^2\left(|e_n^q| + |e_{n+1}^q|\right), \quad |\eta_n^w| \lesssim \Delta\tau\varepsilon\left(|e_n^q| + |e_{n+1}^q|\right), \quad 0 \le n < m. \tag{59}$$

Taking the absolute value (Euclidean norm) on both sides of (58) and using (59), we have

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

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

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

In the light of the truncation errors in (52) and the fact that $m\Delta\tau\varepsilon \lesssim 1$, one has

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

and then by Gronwall's inequality arrives at

$$|e^q_{m+1}| + |e^w_{m+1}| \lesssim \varepsilon \Delta\tau^2, \quad 0 \le m < \frac{T}{\varepsilon\Delta\tau}. \tag{60}$$

Meanwhile, concerning

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

there exists a generic constant $\Delta\tau_0 > 0$ independent of ε and m, such that for $0 < \Delta\tau \le \Delta\tau_0$, (57) holds for $m+1$. This completes the induction.

We rewrite (58) as

$$e^q_{n+1} = e^q_n + \Delta\tau\varphi_1(\Delta\tau\tilde{B})(\varepsilon e^w_n) + \eta^q_n + \xi^q_n,\ (\varepsilon e^w_{n+1}) = \mathrm{e}^{\Delta\tau\tilde{B}}(\varepsilon e^w_n) + \varepsilon\eta^w_n + \varepsilon\xi^w_n.$$

Following the same way as stated above, (55) is arrived. We note that for M1, the global error given in (60) becomes (56).

In what follows, we derive the error bound for re-scaled PM.

Lemma 6.3 *For the error bound of re-scaled PM, it satisfies*

$$\left|[q^j_{n+1}; w^j_{n+1}] - [q(\tau_{n+1}); w(\tau_{n+1})]\right| \le C\big((\varepsilon\Delta\tau)^{j+1} + \varepsilon\delta\tau\big)(1 + |[q_0; w_0]|), \tag{61}$$

where we denote by $C > 0$ a generic constant independent of $\Delta\tau$ or $\delta\tau$ or n or ε.

Proof Firstly, for the coarse propagator (10), we compute

$$\begin{aligned}
&\left|\mathcal{G}^{\tau_n+\Delta\tau}_{\tau_n}([q; w]) - \mathcal{G}^{\tau_n+\Delta\tau}_{\tau_n}([\tilde{q}; \tilde{w}])\right| \\
=&\left|\begin{pmatrix} I & \varepsilon\Delta\tau\phi_1(\Delta\tau\tilde{B}) \\ 0 & \phi_0(\Delta\tau\tilde{B}) \end{pmatrix}([q; w] - [\tilde{q}; \tilde{w}]) + \begin{pmatrix} \varepsilon^2\Delta\tau^2\phi_2(\Delta\tau\tilde{B})\big(F(q) - F(\tilde{q})\big) \\ \varepsilon\Delta\tau\phi_1(\Delta\tau\tilde{B})\big(F(q) - F(\tilde{q})\big) \end{pmatrix}\right| \\
\le&(1 + \varepsilon\Delta\tau)\,|[q; w] - [\tilde{q}; \tilde{w}]| + (\varepsilon^2\Delta\tau^2/2 + \varepsilon\Delta\tau)L\,|q - \tilde{q}|\,,
\end{aligned}$$

with $\|F_q\| \le L$, which gives

$$\left|\mathcal{G}^{\tau_n+\Delta\tau}_{\tau_n}([q; w]) - \mathcal{G}^{\tau_n+\Delta\tau}_{\tau_n}([\tilde{q}; \tilde{w}])\right| \le (1 + C\varepsilon\Delta\tau)\,|[q; w] - [\tilde{q}; \tilde{w}]|\,. \tag{62}$$

Similarly, we have the same result for the fine coarse propagator (11)

$$\left|\mathcal{F}^{\tau_n+\delta\tau}_{\tau_n}([q; w]) - \mathcal{F}^{\tau_n+\delta\tau}_{\tau_n}([\tilde{q}; \tilde{w}])\right| \le (1 + C\varepsilon\delta\tau)\,|[q; w] - [\tilde{q}; \tilde{w}]|\,.$$

Then by the induction, it is immediately derived that

$$\begin{aligned}
&\left|\mathcal{F}^{\tau_n+m\delta\tau}_{\tau_n+(m-1)\delta\tau} \circ \ldots \circ \mathcal{F}^{\tau_n+\delta\tau}_{\tau_n}([q; w]) - \mathcal{F}^{\tau_n+m\delta\tau}_{\tau_n+(m-1)\delta\tau} \circ \ldots \circ \mathcal{F}^{\tau_n+\delta\tau}_{\tau_n}([\tilde{q}; \tilde{w}])\right| \\
\le&\ (1 + C\varepsilon\delta\tau)^m\,|[q; w] - [\tilde{q}; \tilde{w}]|\,, \quad m = \tfrac{\Delta\tau}{\delta\tau}.
\end{aligned}$$

On the other hand, it can be seen that $(1 + C\varepsilon\delta\tau)^{\frac{\Delta\tau}{\delta\tau}} = 1 + (C\varepsilon\delta\tau)\frac{\Delta\tau}{\delta\tau} + \mathcal{O}(\Delta\tau) \le C$. Thus $\left|\mathcal{F}^{\tau_n+\Delta\tau}_{\tau_n}([q; w]) - \mathcal{F}^{\tau_n+\Delta\tau}_{\tau_n}([\tilde{q}; \tilde{w}])\right| \le C\,|[q; w] - [\tilde{q}; \tilde{w}]|$. Letting $[\tilde{q}; \tilde{w}] = [0; 0]$ in this result and (62) yields the boundedness

$$\left|\mathcal{G}^{\tau_n+\Delta\tau}_{\tau_n}([q; w])\right| \le C\,|[q; w]|\,, \quad \left|\mathcal{F}^{\tau_n+\Delta\tau}_{\tau_n}([q; w])\right| \le C\,|[q; w]|\,.$$

For the exact solution of (48), it is assumed that there exists a propagator $\mathcal{E}$ such that $[q(\tau); w(\tau)] = \mathcal{E}_0^\tau[q(0); w(0)]$. Moreover, based on (50), there exists a constant $C > 0$, such that

$$\left|\mathcal{E}_0^\tau[q(0); w(0)]\right| \leq C\left|[q(0); w(0)]\right|, \qquad \tau \in [0, \frac{T}{\varepsilon}]. \tag{63}$$

Define

$$\begin{aligned} e_n^{\mathcal{G}}([q; w]) &= \mathcal{G}_{\tau_n}^{\tau_n+\Delta\tau}([q; w]) - \mathcal{E}_{\tau_n}^{\tau_n+\Delta\tau}([q; w]), \\ e_n^{\mathcal{F}}([q; w]) &= \mathcal{F}_{\tau_n}^{\tau_n+\Delta\tau}([q; w]) - \mathcal{E}_{\tau_n}^{\tau_n+\Delta\tau}([q; w]). \end{aligned}$$

Using the same argument as stated in the part of local truncation errors, it is obtained immediately that there exists a constant $C > 0$, such that,

$$\begin{aligned} \left|e_n^{\mathcal{G}}([q; w]) - e_n^{\mathcal{G}}([\tilde{q}; \tilde{w}])\right| &\leq C\varepsilon^2\Delta\tau^2\left|[q; w] - [\tilde{q}; \tilde{w}]\right|, \\ \left|e_n^{\mathcal{F}}([q; w]) - e_n^{\mathcal{F}}([\tilde{q}; \tilde{w}])\right| &\leq C\varepsilon^2\Delta\tau\delta\tau\left|[q; w] - [\tilde{q}; \tilde{w}]\right|. \end{aligned} \tag{64}$$

Letting $[\tilde{q}; \tilde{w}] = [0; 0]$ gives

$$\left|e_n^{\mathcal{G}}([q; w])\right| \leq C\varepsilon^2\Delta\tau^2\left|[q; w]\right|, \quad \left|e_n^{\mathcal{F}}([q; w])\right| \leq C\varepsilon^2\Delta\tau\delta\tau\left|[q; w]\right|. \tag{65}$$

Now we are in a position to prove the statement (61) by induction over $j \geq 0$. In the light of the result (56) of M1, this statement is obvious for $j = 0$. In what follows, it is assumed that (61) is true for j and we will prove it for $j + 1$. From the definition of the method, it follows that

$$\begin{aligned} [q_{n+1}^{j+1}; w_{n+1}^{j+1}] - [q(\tau_{n+1}); w(\tau_{n+1})] &= \mathcal{G}_{\tau_n}^{\tau_{n+1}}([q_n^{j+1}; w_n^{j+1}]) + \mathcal{F}_{\tau_n}^{\tau_{n+1}}([q_n^j; w_n^j]) \\ &\quad - \mathcal{G}_{\tau_n}^{\tau_{n+1}}([q_n^j; w_n^j]) - \mathcal{E}_{\tau_n}^{\tau_{n+1}}([q(\tau_n); w(\tau_n)]). \end{aligned}$$

In order to prove the result, we split this form as follows

$$\begin{aligned} &[q_{n+1}^{j+1}; w_{n+1}^{j+1}] - [q(\tau_{n+1}); w(\tau_{n+1})] \\ &= \mathcal{G}_{\tau_n}^{\tau_{n+1}}([q_n^{j+1}; w_n^{j+1}]) - \mathcal{G}_{\tau_n}^{\tau_{n+1}}([q(\tau_n); w(\tau_n)]) + (\mathcal{E}_{\tau_n}^{\tau_{n+1}} - \mathcal{G}_{\tau_n}^{\tau_{n+1}})([q_n^j; w_n^j]) \\ &- (\mathcal{E}_{\tau_n}^{\tau_{n+1}} - \mathcal{G}_{\tau_n}^{\tau_{n+1}})([q(\tau_n); w(\tau_n)]) - (\mathcal{E}_{\tau_n}^{\tau_{n+1}} - \mathcal{F}_{\tau_n}^{\tau_{n+1}})([q_n^j; w_n^j]) \\ &+ (\mathcal{E}_{\tau_n}^{\tau_{n+1}} - \mathcal{F}_{\tau_n}^{\tau_{n+1}})([q(\tau_n); w(\tau_n)]) - (\mathcal{E}_{\tau_n}^{\tau_{n+1}} - \mathcal{F}_{\tau_n}^{\tau_{n+1}})([q(\tau_n); w(\tau_n)]) \\ &= \mathcal{G}_{\tau_n}^{\tau_{n+1}}([q_n^{j+1}; w_n^{j+1}]) - \mathcal{G}_{\tau_n}^{\tau_{n+1}}([q(\tau_n); w(\tau_n)]) - e_n^{\mathcal{G}}([q_n^j; w_n^j]) + e_n^{\mathcal{G}}([q(\tau_n); w(\tau_n)]) \\ &+ e_n^{\mathcal{F}}([q_n^j; w_n^j]) - e_n^{\mathcal{F}}([q(\tau_n); w(\tau_n)]) + e_n^{\mathcal{F}}([q(\tau_n); w(\tau_n)]). \end{aligned}$$

By triangular inequality and the results (62)-(65) stated above, it is obtained that

$$\begin{aligned} &\left|[q_{n+1}^{j+1}; w_{n+1}^{j+1}] - [q(\tau_{n+1}); w(\tau_{n+1})]\right| \\ &\leq \left|\mathcal{G}_{\tau_n}^{\tau_{n+1}}([q_n^{j+1}; w_n^{j+1}]) - \mathcal{G}_{\tau_n}^{\tau_{n+1}}([q(\tau_n); w(\tau_n)])\right| + \left|e_n^{\mathcal{G}}([q_n^j; w_n^j]) - e_n^{\mathcal{G}}([q(\tau_n); w(\tau_n)])\right| \\ &+ \left|e_n^{\mathcal{F}}([q_n^j; w_n^j]) - e_n^{\mathcal{F}}([q(\tau_n); w(\tau_n)])\right| + \left|e_n^{\mathcal{F}}([q(\tau_n); w(\tau_n)])\right| \\ &\leq (1 + C\varepsilon\Delta\tau)\left|[q_n^{j+1}; w_n^{j+1}] - [q(\tau_n); w(\tau_n)]\right| + C\Delta\tau^2\varepsilon^2\left|[q_n^j; w_n^j] - [q(\tau_n); w(\tau_n)]\right| \\ &+ C\Delta\tau\delta\tau\varepsilon^2\left|[q_n^j; w_n^j] - [q(\tau_n); w(\tau_n)]\right| + C\Delta\tau\delta\tau\varepsilon^2\left|[q(\tau_n); w(\tau_n)]\right|. \end{aligned}$$

The boundedness of the exact solution as well as the induction hypothesis (61) allow to get

$$\begin{aligned}&\left|[q_{n+1}^{j+1}; w_{n+1}^{j+1}] - [q(\tau_{n+1}); w(\tau_{n+1})]\right| \le (1 + C\varepsilon\Delta\tau)\left|[q_n^{j+1}; w_n^{j+1}] - [q(\tau_n); w(\tau_n)]\right| \\ &+ C\Delta\tau(\Delta\tau + \delta\tau)\big((\varepsilon\Delta\tau)^{j+1} + \varepsilon\delta\tau\big)\varepsilon^2(1 + |[q_0; w_0]|) + C\Delta\tau\delta\tau\varepsilon^2\,|[q_0; w_0]|\,.\end{aligned}$$

As long as $\varepsilon(\Delta\tau + \varepsilon^j\Delta\tau^{j+1} + \delta\tau)(1 + |[q_0; w_0]|) \le 1$, we have

$$\begin{aligned}&\left|[q_{n+1}^{j+1}; w_{n+1}^{j+1}] - [q(\tau_{n+1}); w(\tau_{n+1})]\right| \\ &\le (1 + C\varepsilon\Delta\tau)\left|[q_n^{j+1}; w_n^{j+1}] - [q(\tau_n); w(\tau_n)]\right| \\ &\quad + C\varepsilon\Delta\tau\big((\varepsilon\Delta\tau)^{j+2} + \varepsilon\delta\tau\big)(1 + |[q_0; w_0]|),\end{aligned}$$

from which we get (61) for $j + 1$ from the discrete Gronwall lemma. □

6.1.4 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 for the original system (1) and the re-scaled system (48), $x(t_n) = q(\tau_n)$ and $v(t_n) = w(\tau_n)$. Moreover, by comparing (9) with (49), we know that the numerical solution x_n, v_n of (9) is identical to q_n, w_n of (49). Therefore, the result (55) yields the uniform error bound in x given in (16) and also shows the non-uniform error in v of (16). The results (15a) of M1 and (15b) of PM are direct results of (56) and (61), respectively.

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 (52) and (53). This motivates us to consider modulated Fourier expansions (see, e.g. [19, 22, 24, 25, 40]) for analysis in this part. The proof will be briefly shown for SM2 and it can be modified for the other two methods easily. Since the result is given under a lower bound on the stepsize, here the truncation error of modulated Fourier expansion is measured under h.

6.2.1 Decomposition of the Numerical Solution

Now we turn back to the SM2 given in (31) and consider its modulated Fourier expansion (33). In order to derive the convergence, we need to explicitly present the results of $\tilde{\zeta}_k$ and $\tilde{\eta}_k$ with $|k| \le 1$. In the light of (39) and the properties of $\hat{\mathcal{L}}(hD)$, we obtain

$$\begin{aligned}\dot{\tilde{\zeta}}_{\langle 0\rangle}^{\pm j} &= \frac{-h^2\tilde{\omega}_j A(h\tilde{\omega}_j)}{8\mathrm{i}\sin^2(\frac{1}{2}h\tilde{\omega}_j)}\Big(\tilde{F}(\tilde{\zeta}_{\langle 0\rangle}) + \tilde{F}''(\tilde{\zeta}_{\langle 0\rangle})(\tilde{\zeta}_{\langle j\rangle}, \tilde{\zeta}_{\langle -j\rangle}) + \dots\Big)^{\pm j},\\ \ddot{\tilde{\zeta}}_{\langle 0\rangle}^{0} &= \Big(\tilde{F}(\tilde{\zeta}_{\langle 0\rangle}) + \tilde{F}''(\tilde{\zeta}_{\langle 0\rangle})(\tilde{\zeta}_{\langle j\rangle}, \tilde{\zeta}_{\langle -j\rangle}) + \dots\Big)^{0},\\ \tilde{\zeta}_{\langle j\rangle}^{-j} &= \frac{h^3\tilde{\omega}_j A(h\tilde{\omega}_j)}{-16\sin^2(\frac{1}{2}h\tilde{\omega}_j)\sin(h\tilde{\omega}_j)}\big(\tilde{F}'(\tilde{\zeta}_{\langle 0\rangle})\tilde{\zeta}_{\langle j\rangle} + \dots\big)^{-j},\\ \tilde{\zeta}_{\langle j\rangle}^{0} &= \frac{h^2}{-4\sin^2(h\tilde{\omega}_j/2)}\big(\tilde{F}'(\tilde{\zeta}_{\langle 0\rangle})\tilde{\zeta}_{\langle j\rangle} + \dots\big)^{0},\end{aligned}$$

$$
\begin{aligned}
\dot{\tilde{\zeta}}^{j}_{\langle j\rangle} &= \frac{h^2\tilde{\omega}_j A(h\tilde{\omega}_j)}{8\mathrm{i}\sin^2(\frac{1}{2}h\tilde{\omega}_j)}\big(\tilde{F}'(\tilde{\zeta}_{\langle 0\rangle})\tilde{\zeta}_{\langle j\rangle}+\ldots\big)^{j},\\
\dot{\tilde{\zeta}}^{-j}_{\langle -j\rangle} &= \frac{h^2\tilde{\omega}_j A(h\tilde{\omega}_j)}{-8\mathrm{i}\sin^2(\frac{1}{2}h\tilde{\omega}_j)}\big(\tilde{F}'(\tilde{\zeta}_{\langle 0\rangle})\tilde{\zeta}_{\langle -j\rangle}+\ldots\big)^{-j},\\
\tilde{\zeta}^{0}_{\langle -j\rangle} &= \frac{h^2}{-4\sin^2(h\tilde{\omega}_j/2)}\big(\tilde{F}'(\tilde{\zeta}_{\langle 0\rangle})\tilde{\zeta}_{\langle -j\rangle}+\ldots\big)^{0},\\
\tilde{\zeta}^{j}_{\langle -j\rangle} &= \frac{h^3\tilde{\omega}_j A(h\tilde{\omega}_j)}{-16\sin^2(\frac{1}{2}h\tilde{\omega}_j)\sin(h\tilde{\omega}_j)}\big(\tilde{F}'(\tilde{\zeta}_{\langle 0\rangle})\tilde{\zeta}_{\langle -j\rangle}+\ldots\big)^{j}.
\end{aligned}
\tag{66}
$$

Then the following results

$$
\begin{aligned}
&\tilde{\eta}^{0}_{\langle 0\rangle}=\dot{\tilde{\zeta}}^{0}_{\langle 0\rangle}+\mathcal{O}(h),\ \tilde{\eta}^{\pm j}_{\langle 0\rangle}=\tfrac{h\tilde{\omega}_j}{\sin(h\tilde{\omega}_j)}\dot{\tilde{\zeta}}^{\pm j}_{\langle 0\rangle}+\mathcal{O}(h),\ \tilde{\eta}^{0}_{\langle \pm j\rangle}=\mathrm{i}\tilde{\omega}_j\,\mathrm{sinc}(h\tilde{\omega}_j)\tilde{\zeta}^{0}_{\langle \pm j\rangle}+\mathcal{O}(h),\\
&\tilde{\eta}^{j}_{\langle j\rangle}=\mathrm{i}\tilde{\omega}_j\tilde{\zeta}^{j}_{\langle j\rangle}+\mathcal{O}\Big(h\,\Big|\mathrm{i}\tan(\tfrac{h}{2}\tilde{\omega}_j)\Big|\Big),\qquad\qquad \tilde{\eta}^{-j}_{\langle -j\rangle}=-\mathrm{i}\tilde{\omega}_j\tilde{\zeta}^{-j}_{\langle -j\rangle}+\mathcal{O}\Big(h\,\Big|\mathrm{i}\tan(\tfrac{h}{2}\tilde{\omega}_j)\Big|\Big)
\end{aligned}
\tag{67}
$$

are easily arrived by considering (42) as well as the property of $\mathcal{L}(hD)$.

6.2.2 Decomposition of the Exact Solution

Following the result given in [24], the exact solution of (20) for $t=nh$ admits the following expansion

$$
\tilde{x}(t)=\sum_{k\in\mathcal{N}^*}\mathrm{e}^{\mathrm{i}(k\cdot\varpi)t}\tilde{\mu}_k(t)+\tilde{d}_{\tilde{x}}(t),\quad \tilde{v}(t)=\sum_{k\in\mathcal{N}^*}\mathrm{e}^{\mathrm{i}(k\cdot\varpi)t}\tilde{\nu}_k(t)+\tilde{d}_{\tilde{v}}(t),
\tag{68}
$$

where the defects are bounded by $\tilde{d}_{\tilde{x}}(t)=\mathcal{O}(\varepsilon^2)$, $\tilde{d}_{\tilde{v}}(t)=\mathcal{O}(\varepsilon)$. The functions $\tilde{\mu}_k$ with $|k|\le 1$ are given by

$$
\begin{aligned}
\dot{\tilde{\mu}}^{\pm j}_{\langle 0\rangle} &= \frac{1}{\mp\mathrm{i}\tilde{\omega}_j}\big(\tilde{F}(\tilde{\mu}_{\langle 0\rangle})+\tilde{F}''(\tilde{\mu}_{\langle 0\rangle})(\tilde{\mu}_{\langle j\rangle},\tilde{\mu}_{\langle -j\rangle})+\ldots\big)^{\pm j},\\
\ddot{\tilde{\mu}}^{0}_{\langle 0\rangle} &= \big(\tilde{F}(\tilde{\mu}_{\langle 0\rangle})+\tilde{F}_0''(\tilde{\mu}_{\langle 0\rangle})(\tilde{\mu}_{\langle j\rangle},\tilde{\mu}_{\langle -j\rangle})+\ldots\big)^{0},\\
\tilde{\mu}^{-j}_{\langle j\rangle} &= \frac{1}{-2\tilde{\omega}_j^2}\big(\tilde{F}'(\tilde{\mu}_{\langle 0\rangle})\tilde{\mu}_{\langle j\rangle}+\ldots\big)^{-j},\\
\tilde{\mu}^{0}_{\langle j\rangle} &= \frac{1}{-\tilde{\omega}_j^2}\big(\tilde{F}'(\tilde{\mu}_{\langle 0\rangle})\tilde{\mu}_{\langle j\rangle}+\ldots\big)^{0},\\
\dot{\tilde{\mu}}^{j}_{\langle j\rangle} &= \frac{1}{\mathrm{i}\tilde{\omega}_j}\big(\tilde{F}'(\tilde{\mu}_{\langle 0\rangle})\tilde{\mu}_{\langle j\rangle}+\ldots\big)^{j},\\
\dot{\tilde{\mu}}^{-j}_{\langle -j\rangle} &= \frac{1}{-\mathrm{i}\tilde{\omega}_j}\big(\tilde{F}'(\tilde{\mu}_{\langle 0\rangle})\tilde{\mu}_{\langle -j\rangle}+\ldots\big)^{-j},\\
\tilde{\mu}^{0}_{\langle -j\rangle} &= \frac{1}{-\tilde{\omega}_j^2}\big(\tilde{F}'(\tilde{\mu}_{\langle 0\rangle})\tilde{\mu}_{\langle -j\rangle}+\ldots\big)^{0},\\
\tilde{\mu}^{j}_{\langle -j\rangle} &= \frac{1}{-2\tilde{\omega}_j^2}\big(\tilde{F}'(\tilde{\mu}_{\langle 0\rangle})\tilde{\mu}_{\langle -j\rangle}+\ldots\big)^{j},
\end{aligned}
\tag{69}
$$

and the functions $\tilde{\nu}_k$ with $|k|\le 1$ are

$$
\tilde{\nu}_{\langle 0\rangle}=\dot{\tilde{\mu}}_{\langle 0\rangle},\quad \tilde{\nu}_{\langle \pm j\rangle}=\pm\mathrm{i}\tilde{\omega}_j\tilde{\mu}_{\langle \pm j\rangle}+\dot{\tilde{\mu}}_{\langle \pm j\rangle}=\pm\mathrm{i}\tilde{\omega}_j\tilde{\mu}_{\langle \pm j\rangle}+\mathcal{O}(\varepsilon).
\tag{70}
$$

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

6.2.3 Proof of the Convergence

Looking closely to the Eqs. (69) and (66), which determine the modulated Fourier expansion coefficients, it is obtained that $\tilde{x}^*(t) - \tilde{x}_h(t) = \mathcal{O}(h^2)$. Similarly, with (70) and (67), 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:

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

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 in x were given to preserve the symplecticity and energy, respectively. Moreover, some methods with uniform error in both x and v were derived. Long term energy conservation of symplectic methods were also discussed. All the theoretical results were supported by two numerical experiments 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 $\varepsilon = 1$. The algorithms are also applicable to the strongly damped Helmholtz-Duffing oscillator, strongly damped wave equation, eardrum oscillations, elasto-magnetic suspensions, and other physical phenomena [13, 33, 39]. Meanwhile, there are some issues brought by this paper which can be researched further. For the system (1) (not two dimensional) with a matrix $\tilde{B}(x)$ depending on x, how to get uniformly accurate structure-preserving algorithms? This point is challenging and 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 [5, 6, 11, 12].

Acknowledgements The authors sincerely thank the two anonymous reviewers for the very valuable comments and helpful suggestions. This work was supported by NSFC (12271426) and Key Research and Development Projects of Shaanxi Province (2023-YBSF-399). The first author is grateful to Christian Lubich for his valuable comments on the first version of Theorem 3.3 as well as its proof, which was done in part at UNIVERSITAT TÜBINGEN when he worked there as a postdoctoral researcher (2017-2019, supported by the Alexander von Humboldt Foundation).

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

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

References

1. Boris, J.P.: Relativistic plasma simulation-optimization of a hybrid code. In: Proceeding of Fourth Conference on Numerical Simulations of Plasmas, pp. 3–67 (1970)
2. Brugnano, L., Iavernaro, F., Zhang, R.: Arbitrarily high-order energy-preserving methods for simulating the gyrocenter dynamics of charged particles. J. Comput. Appl. Math. **380**, 112994 (2020)
3. Brugnano, L., Montijano, J.I., Rández, L.: High-order energy-conserving line integral methods for charged particle dynamics. J. Comput. Phys. **396**, 209–227 (2019)
4. Celledoni, E., Grimm, V., McLachlan, R.I., McLaren, D.I., O'Neale, D., Owren, B., Quispel, G.R.W.: Preserving energy resp. dissipation in numerical PDEs using the Average Vector Field method. J. Comput. Phys. **231**, 6770–6789 (2012)
5. Chartier, Ph., Crouseilles, N., Lemou, M., Méhats, F., Zhao, X.: Uniformly accurate methods for Vlasov equations with non-homogeneous strong magnetic field. Math. Comp. **88**, 2697–2736 (2019)
6. Chartier, Ph., Crouseilles, N., Lemou, M., Méhats, F., Zhao, X.: Uniformly accurate methods for three dimensional Vlasov equations under strong magnetic field with varying direction. SIAM J. Sci. Compt. **42**, B520–B547 (2020)
7. Chartier, Ph., Lemou, M., Méhats, F., Vilmart, G.: A new class of uniformly accurate methods for highly oscillatory evolution equations. Found. Comput. Math. **20**, 1–33 (2020)
8. Chartier, Ph., Lemou, M., Méhats, F., Zhao, X.: Derivative-free high-order uniformly accurate schemes for highly-oscillatory systems. IMA J. Numer. Anal. **42**, 1623–1644 (2022)
9. Chartier, Ph., Méhats, F., Thalhammer, M., Zhang, Y.: Improved error estimates for splitting methods applied to highly-oscillatory nonlinear Schrödinger equations. Math. Comp. **85**, 2863–2885 (2016)
10. Cohen, D., Hairer, E., Lubich, Ch.: Modulated Fourier expansions of highly oscillatory differential equations. Found. Comput. Math. **3**, 327–345 (2003)
11. Crouseilles, N., Hirstoaga, S., Zhao, X.: Multiscale Particle-in-Cell methods and comparisons for the long time two-dimensional Vlasov-Poisson equation with strong magnetic field. Comput. Phys. Comm. **222**, 136–151 (2018)
12. Crouseilles, N., Lemou, M., Méhats, F., Zhao, X.: Uniformly accurate Particle-In-Cell method for the long time solution of the two-dimensional Vlasov-Poisson equation with uniform strong magnetic field. J. Comput. Phys. **346**, 172–190 (2017)
13. Elías-Zúñiga, A.: Analytical solution of the damped Helmholtz-Duffing equation. Appl. Math. Lett. **25**, 2349–2353 (2012)
14. Feng, K., Qin, M.: Symplectic Geometric algorithms for Hamiltonian systems. Springer-Verlag, Berlin, Heidelberg (2010)
15. Filbet, F., Rodrigues, M.: Asymptotically stable particle-in-cell methods for the Vlasov-Poisson system with a strong external magnetic field. SIAM J. Numer. Anal. **54**, 1120–1146 (2016)
16. Filbet, F., Rodrigues, M.: Asymptotically preserving particle-in-cell methods for inhomogeneous strongly magnetized plasmas. SIAM J. Numer. Anal. **55**, 2416–2443 (2017)
17. Gauckler, L., Hairer, E., Lubich, CH.: Dynamics, numerical analysis, and some geometry. In: Proceedings of the International Congress of Mathematicians-Rio de Janeiro 2018. Plenary lectures,I, 453–485, World Sci. Publ., Hackensack, NJ (2018)
18. Grigori, L., Hirstoaga, S.A., Nguyen, V., Salomon, J.: Reduced model-based parareal simulations of oscillatory singularly perturbed ordinary differential equations. J. Comput. Phys. **436**, 110282 (2021)
19. Hairer, E., Lubich, Ch.: Long-time energy conservation of numerical methods for oscillatory differential equations. SIAM J. Numer. Anal. **38**, 414–441 (2000)
20. Hairer, E., Lubich, Ch.: Energy behaviour of the Boris method for charged-particle dynamics. BIT **58**, 969–979 (2018)
21. Hairer, E., Lubich, Ch.: Symmetric multistep methods for charged-particle dynamics. SMAI J. Comput. Math. **3**, 205–218 (2017)

22. Hairer, E., Lubich, Ch.: Long-term analysis of a variational integrator for charged-particle dynamics in a strong magnetic field. Numer. Math. **144**, 699–728 (2020)
23. Hairer, E., Lubich, Ch., Shi, Y.: Large-stepsize integrators for charged-particle dynamics over multiple time scales. Numer. Math. **151**, 659–691 (2022)
24. Hairer, E., Lubich, Ch., Wang, B.: A filtered Boris algorithm for charged-particle dynamics in a strong magnetic field. Numer. Math. **144**, 787–809 (2020)
25. Hairer, E., Lubich, Ch., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer-Verlag, Berlin, Heidelberg (2006)
26. He, Y., Zhou, Z., Sun, Y., Liu, J., Qin, H.: Explicit K-symplectic algorithms for charged particle dynamics. Phys. Lett. A **381**, 568–573 (2017)
27. He, Y., Sun, Y., Liu, J., Qin, H.: Volume-preserving algorithms for charged particle dynamics. J. Comput. Phys. **281**, 135–147 (2015)
28. Hochbruck, M., Ostermann, A.: Exponential integrators. Acta Numer. **19**, 209–286 (2010)
29. Iserles, A.: On the global error of discretization methods for highly-oscillatory ordinary differential equations. BIT **42**, 561–599 (2002)
30. Iserles, A., Nørsett, S.P.: From high oscillation to rapid approximation III: Multivariate expansions. IMA J. Num. Anal. **29**, 882–916 (2009)
31. Lions, J.-L., Maday, Y., Turinici, G.: A parareal in time discretization of PDE's, C. R. Acad. Sci. Ser. I Math. **332**, 661–668 (2001)
32. Mocquard, Y., Navaro, P., Crouseilles, N.: HOODESolver, jl: a Julia package for highly oscillatory problems. J. Open Source Softw. **61**, 3077 (2021)
33. Pata, V., Squassina, M.: On the strongly damped wave equation. Commun. Math. Phys. **253**, 511–533 (2005)
34. Qin, H., Zhang, S., Xiao, J., Liu, J., Sun, Y., Tang, W.M.: Why is Boris algorithm so good? Phys. Plasmas **20**, 084503 (2013)
35. Ricketson, L.F., Chacón, L.: An energy-conserving and asymptotic-preserving charged-particle orbit implicit time integrator for arbitrary electromagnetic fields. J. Comput. Phys. **418**, 109639 (2020)
36. Sanz-Serna, J.M.: Mollified impulse methods for highly-oscillatory differential equations. SIAM J. Numer. Anal. **46**, 1040–1059 (2008)
37. Sanz-Serna, J.M.: Symplectic Runge-Kutta schemes for adjoint equations, automatic differentiation, optimal control and more. SIAM Rev. **58**, 3–33 (2016)
38. Tao, M.: Explicit high-order symplectic integrators for charged particles in general electromagnetic fields. J. Comput. Phys. **327**, 245–251 (2016)
39. Thomee, V., Wahlbin, L.B.: Maximum-norm estimates for finite-element methods for a strongly damped wave equation. BIT **44**, 165–179 (2004)
40. Wang, B.: Exponential energy-preserving methods for charged-particle dynamics in a strong and constant magnetic field. J. Comput. Appl. Math. **387**, 112617 (2021)
41. Wang, B., Iserles, A., Wu, X.: Arbitrary-order trigonometric Fourier collocation methods for multi-frequency oscillatory systems. Found. Comput. Math. **16**, 151–181 (2016)
42. Wang, B., Zhao, X.: Error estimates of some splitting schemes for charged-particle dynamics under strong magnetic field. SIAM J. Numer. Anal. **59**, 2075–2105 (2021)
43. B. Wang, X. Zhao, Geometric two-scale integrators for highly oscillatory system: uniform accuracy and near conservations, SIAM J. Numer. Anal., Accepted for publication (2023)
44. Wu, X., Wang, B.: Geometric Integrators for Differential Equations with Highly Oscillatory Solutions. Springer, Singapore (2021)