Chapter 14
Long-Time Momentum and Actions Behaviour of Energy-Preserving Methods for Wave Equations

Wave equations have physically very important properties which should be respected by numerical schemes in order to predict correctly the solution over a long-time period. In this chapter, the long-time behaviour of momentum and actions for energy-preserving methods are analysed in detail for semilinear wave equations.

14.1 Introduction

The main theme of this chapter is the long-time behaviour of energy-preserving (EP) methods when applied to the following one-dimensional semilinear wave equation (see [1–3])

$$\partial_t^2 u - \partial_x^2 u + \rho u + g(u) = 0, \quad -\pi \leqslant x \leqslant \pi, \quad t > 0, \tag{14.1}$$

where g is a nonlinear and smooth real function with $g(0) = g'(0) = 0$ and ρ is a positive number. Following the Refs. [1–3], we assume that the initial values $u(\cdot, 0)$ and $\partial_t u(\cdot, 0)$ for this equation are bounded by a small parameter ε, which provides small initial data in appropriate Sobolev norms. Here, we consider 2π-periodic boundary condition $u(x, t) = u(x + 2\pi, t)$ for (14.1).

As is known, several important quantities are conserved by the solution of (14.1). Firstly, the total energy

$$H(u, v) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Big(\frac{1}{2} \big(v^2 + (\partial_x u)^2 + \rho u^2 \big) + U(u) \Big) \mathrm{d}x$$

X. Wu, B. Wang, *Geometric Integrators for Differential Equations with Highly Oscillatory Solutions*, https://doi.org/10.1007/978-981-16-0147-7_14

is exactly preserved along the solution, where $v = \partial_t u$ and $U(u)$ is the potential such that $U'(u) = g(u)$. Secondly, the solution of (14.1) also conserves the momentum

$$K(u, v) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \partial_x u(x) v(x) \mathrm{d}x.$$

Thirdly, the harmonic actions

$$I_j(u, v) = \frac{\omega_j}{2} |u_j|^2 + \frac{1}{2\omega_j} |v_j|^2, \ j \in \mathbb{Z}$$

are conserved for the linear wave equation, i.e., $g(u) \equiv 0$, where $\omega_j = \sqrt{\rho + j^2}$ for $j \in \mathbb{Z}$. In the nonlinear case, it has been proved in [2, 4] that, for smooth and small initial data and for almost all values of $\rho > 0$, the actions $I_j(u, v)$ remain constant up to small deviations over a long-time period.

In the past decades it has become increasingly important to design numerical integrators for wave equations aiming at respecting qualitative properties of the solution (see, e.g. [5–14]). Among others, long-time conservation properties of numerical methods when applied to wave equations have been well studied [1–3, 15, 16]. All these analyses are achieved by the technique of modulated Fourier expansions, which was developed by Hairer and Lubich in [17] and has been frequently used in the long-term analysis (see, e.g. [18–22]). On the other hand, as an important kind of method, energy-preserving (EP) methods have also been the subject of many investigations for wave equations. EP methods can exactly preserve the energy of the system under consideration. Concerning some examples of this topic, we refer the readers to [23–31]. Unfortunately, it seems that the study of the long-time behaviour of EP methods in other structure-preserving aspects is quite inadequate for wave equations in the literature, e.g. the numerical conservation of momentum and actions. This chapter focuses on this point.

14.2 Full Discretisation

This section presents a full discretisation for solving the semilinear wave equation (14.1). We begin with a spectral semidiscretisation in space introduced in [1, 3], and then use EP methods in time.

14.2.1 Spectral Semidiscretisation in Space

We here choose equidistant collocation points $x_k = k\pi/M$, $k = -M, -M + 1, \cdots, M - 1$, for the pseudospectral semidiscretisation in space and consider a pair of real-valued trigonometric polynomials as an approximation for the solution

of (14.1)

$$u^M(x,t) = \sum_{|j|\leqslant M}{}' q_j(t)\mathrm{e}^{\mathrm{i}jx}, \quad v^M(x,t) = \sum_{|j|\leqslant M}{}' p_j(t)\mathrm{e}^{\mathrm{i}jx}, \quad \mathrm{i}=\sqrt{-1}, \tag{14.2}$$

where $p_j(t) = \dfrac{\mathrm{d}}{\mathrm{d}t}q_j(t)$ and the prime indicates that the first and last terms in the summation are taken with the factor $1/2$. We collect all the q_j in a $2M$-periodic coefficient vector $q(t) = (q_j(t))$, which is a solution of the $2M$-dimensional system of oscillatory ODEs

$$\frac{\mathrm{d}^2 q}{\mathrm{d}t^2} + \Omega^2 q = f(q), \tag{14.3}$$

where $f(q) = -\mathscr{F}_{2M} g(\mathscr{F}_{2M}^{-1} q)$, Ω is diagonal with entries ω_j, and $\mathscr{F}_{2M}$ denotes the discrete Fourier transform $(\mathscr{F}_{2M} w)_j = \dfrac{1}{2M}\sum\limits_{k=-M}^{M-1} w_k \mathrm{e}^{-\mathrm{i}jx_k}$ for $|j| \leqslant M$. It is noted that the system (14.3) is a finite-dimensional complex Hamiltonian system with the energy

$$H_M(p,q) = \frac{1}{2}\sum_{|j|\leqslant M}{}' \left(|p_j|^2 + \omega_j^2 |q_j|^2\right) + V(q), \tag{14.4}$$

where $V(q) = \dfrac{1}{2M}\sum\limits_{k=-M}^{M-1} U((\mathscr{F}_{2M}^{-1} q)_k)$. Accordingly, the actions (for $|j| \leqslant M$) and the momentum of (14.3) are respectively given by

$$I_j(p,q) = \frac{\omega_j}{2}|q_j|^2 + \frac{1}{2\omega_j}|p_j|^2, \quad K(p,q) = -\sum_{|j|\leqslant M}{}'' \mathrm{i}j q_{-j} p_j, \quad \mathrm{i}=\sqrt{-1},$$

where the double prime indicates that the first and last terms in the summation are taken with the factor $1/4$. We are interested only in real approximation (14.2) throughout this chapter, and hence it holds that $q_{-j} = \bar{q}_j$, $p_{-j} = \bar{p}_j$ and $I_{-j} = I_j$.

It is important to note that the energy (14.4) is exactly preserved along the solution of (14.3). For the momentum and actions in the semidiscretisation, the following results have been proved in [3].

Theorem 14.1 (See [3]) *Under the non-resonance condition* (14.10) *and the Assumption* (14.7) *which are stated in Sect. 14.3.1, it holds that*

$$\sum_{l=0}^{M} \omega_l^{2s+1} \frac{|I_l(p(t), q(t)) - I_l(p(0), q(0))|}{\varepsilon^2} \leqslant C\varepsilon,$$
$$\frac{|K(p(t), q(t)) - K(p(0), q(0))|}{\varepsilon^2} \leqslant Ct\varepsilon M^{-s+1},$$

where $0 \leqslant t \leqslant \varepsilon^{-N+1}$ *and the constant* C *is independent of* ε, M, h *and* t.

14.2.2 EP Methods in Time

It is known that among typical EP integrators is the average vector field (AVF) method (see [32]). Unfortunately, however, it has been pointed out in Chap. 1 that the AVF method cannot efficiently solve the highly oscillatory system (14.3) (see also [33, 34]) since the AVF method is not oscillation preserving. Moreover, the integral appearing in the AVF formula is dependent on the frequency matrix Ω. This fact leads to the following definition.

Definition 14.1 (See [33, 34]) For efficiently solving the oscillatory system (14.3), the *adapted average vector field* (AAVF) method has the form

$$\begin{cases} q_{n+1} = \phi_0(V)q_n + h\phi_1(V)p_n + h^2\phi_2(V)\int_0^1 f((1-\sigma)q_n + \sigma q_{n+1})\mathrm{d}\sigma, \\ p_{n+1} = -h\Omega^2\phi_1(V)q_n + \phi_0(V)p_n + h\phi_1(V)\int_0^1 f((1-\sigma)q_n + \sigma q_{n+1})\mathrm{d}\sigma, \end{cases} \tag{14.5}$$

where h is the stepsize, and

$$\phi_l(V) := \sum_{k=0}^{\infty} \frac{(-1)^k V^k}{(2k+l)!}, \quad l = 0, 1, 2 \tag{14.6}$$

are matrix-valued functions of $V = h^2\Omega^2$.

According to (14.6), it is clear that

$$\phi_0(V) = \cos(h\Omega), \quad \phi_1(V) = \sin(h\Omega)(h\Omega)^{-1}, \quad \phi_2(V) = (I - \cos(h\Omega))(h\Omega)^{-2}.$$

It is interesting to note that as $V \to 0$ the method (14.5) reduces to the well-known AVF method. The following properties of the AAVF method have been shown in [33, 34].

Theorem 14.2 (See [33, 34]) *The AAVF method is symmetric and exactly preserves the energy* (14.4)*, which means that*

$$H_M(p_{n+1}, q_{n+1}) = H_M(p_n, q_n) \quad for \quad n = 0, 1, \cdots.$$

Theorem 14.2 ensures that the energy-preserving AAVF method does not exclude symmetry structure, and, as is known, preserving the energy and symmetry of the system simultaneously at the discrete level is important for geometric integrators.

14.3 Main Result and Numerical Experiment

In what follows, we shall use the following notations (see [1]). We denote

$$|k| = (|k_l|)_{l=0}^{M}, \quad \|k\| = \sum_{l=0}^{M} |k_l|, \quad k \cdot \boldsymbol{\omega} = \sum_{l=0}^{M} k_l \omega_l, \quad \boldsymbol{\omega}^{\sigma|k|} = \prod_{l=0}^{M} \omega_l^{\sigma|k_l|}.$$

for sequences of integers $k = (k_l)_{l=0}^{M}$, $\boldsymbol{\omega} = (\omega_l)_{l=0}^{M}$ and a real number σ. We also denote by $\langle j \rangle$ the unit coordinate vector $(0, \cdots, 0, 1, 0, \cdots, 0)^{\mathsf{T}}$ with 1 in the j-th entry 1 and 0 elsewhere. For $s \in \mathbb{R}^+$, the space of $2M$-periodic sequences $q = (q_j)$ endowed with the weighted norm $\|q\|_s = \Big(\sum_{|j| \leqslant M}'' \omega_j^{2s} |q_j|^2 \Big)^{1/2}$ is denoted by H^s. Furthermore, we set

$$[[k]] = \begin{cases} (\|k\| + 1)/2, & k \neq \mathbf{0}, \\ 3/2, & k = \mathbf{0}. \end{cases}$$

14.3.1 Main Result

In this subsection we first present the main result of this chapter, which will be illustrated by numerical experiments. The following assumptions (see [1]) are needed for the main result.

Assumption 14.1 It is assumed that the initial values of (14.3) are bounded by

$$\big(\|q(0)\|_{s+1}^2 + \|p(0)\|_s^2 \big)^{1/2} \leqslant \varepsilon \tag{14.7}$$

with a small parameter $\varepsilon > 0$.

Assumption 14.2 The following non-resonance condition holds for a given step-size h:

$$\left|\sin\left(\frac{h}{2}(\omega_j - k\cdot\boldsymbol{\omega})\right)\cdot\sin\left(\frac{h}{2}(\omega_j + k\cdot\boldsymbol{\omega})\right)\right| \geqslant \varepsilon^{1/2}h^2(\omega_j + |k\cdot\boldsymbol{\omega}|). \tag{14.8}$$

If this condition is not true, we define a set of near-resonant indices

$$\mathscr{R}_{\varepsilon,h} = \{(j,k) : |j| \leqslant M,\ \|k\| \leqslant 2N,\ \ k \neq \pm\langle j\rangle,\ \text{not satisfying (14.8)}\}, \tag{14.9}$$

where $N \geqslant 1$ is the truncation number of the expansion (14.15) which will be presented in the next section. Moreover, we assume that there exist $\sigma > 0$ and a constant C_0 such that

$$\sup_{(j,k)\in\mathscr{R}_{\varepsilon,h}} \frac{\omega_j^\sigma}{\boldsymbol{\omega}^{\sigma|k|}}\varepsilon^{\|k\|/2} \leqslant C_0\varepsilon^N, \tag{14.10}$$

for the set $\mathscr{R}_{\varepsilon,h}$.

Assumption 14.3 Assume that the following numerical non-resonance condition

$$|\sin(h\omega_j)| \geqslant h\varepsilon^{1/2} \ \text{ for } \ |j| \leqslant M, \tag{14.11}$$

is satisfied.

Assumption 14.4 Suppose that, for a positive constant $c > 0$, another non-resonance condition

$$\begin{aligned}&\left|\sin\left(\frac{h}{2}(\omega_j - k\cdot\boldsymbol{\omega})\right)\cdot\sin\left(\frac{h}{2}(\omega_j + k\cdot\boldsymbol{\omega})\right)\right| \geqslant ch^2|2\phi_2(h^2\omega_j^2)|\\ &\text{for } (j,k) \text{ of the form } j = j_1 + j_2 \text{ and } k = \pm\langle j_1\rangle \pm \langle j_2\rangle,\end{aligned} \tag{14.12}$$

is also fulfilled, which leads to improved conservation estimates.

The following theorem represents the main result of this chapter.

Theorem 14.3 *We define the following modified momentum and actions, respectively*

$$\hat{I}_j(p,q) = \frac{\cos\left(\frac{1}{2}h\omega_j\right)}{\operatorname{sinc}\left(\frac{1}{2}h\omega_j\right)}I_j(p,q), \quad \hat{K}(p,q) = -\sum_{|j|\leqslant M}{}'' \mathrm{i}j\frac{\cos\left(\frac{1}{2}h\omega_j\right)}{\operatorname{sinc}\left(\frac{1}{2}h\omega_j\right)}q_{-j}p_j,$$

and choose the stepsize h such that

$$\left| \frac{\cos\left(\frac{1}{2}h\omega_j\right)}{\operatorname{sinc}\left(\frac{1}{2}h\omega_j\right)} \right| \leqslant C_1 \qquad for \quad |j| \leqslant M. \tag{14.13}$$

Then under the conditions of Assumptions 14.1–14.4 with $s \geqslant \sigma + 1$, for the AAVF method (14.5) and $0 \leqslant t = nh \leqslant \varepsilon^{-N+1}$, the following near-conservation estimates of the modified momentum and actions

$$\sum_{l=0}^{M} \omega_l^{2s+1} \frac{|\hat{I}_l(p_n, q_n) - \hat{I}_l(p_0, q_0)|}{\varepsilon^2} \leqslant C\varepsilon,$$

$$\frac{|\hat{K}(p_n, q_n) - \hat{K}(p_0, q_0)|}{\varepsilon^2} \leqslant C(\varepsilon + M^{-s} + \varepsilon t M^{-s+1})$$

hold with a constant C, depending on s, N, C_0 and C_1, but not on ε, M, h and time t. If (14.12) is not satisfied, then the bound $C\varepsilon$ is weakened to $C\varepsilon^{1/2}$.

The proof of this theorem will be shown in detail in Sect. 14.4 based on the technique of multi-frequency modulated Fourier expansions. It is remarked that the above result for the AAVF method with the integral is also true for the AAVF method with a suitable quadrature rule instead of the integral, and this point will be explained briefly in Sect. 14.5.

An interesting study of the long-time behaviour of a symmetric and symplectic trigonometric integrator for solving wave equations was made by Cohen et al. in [1], and it was shown that this integrator has a near-conservation of energy, momentum and actions in numerical discretisations. However, it is noted that the method studied in [1] cannot preserve the energy (14.4) exactly. Fortunately, it follows from Theorems 14.2 and 14.3 that the AAVF method not only preserves the energy (14.4) exactly but also has a near-conservation of modified momentum and actions over long terms.

Remark 14.1 Theorem 14.3 claims that the AAVF method has a near-conservation of a modified momentum and modified actions over long terms. We here remark that we have tried to prove long-time conservation for natural discretisations. However, after the whole procedure of the proof using modulated Fourier expansion, it turns out that artificial coefficients $\cos(h\omega_j)/\sin(h\omega_j/2)$ form part of each term of the summation of the natural discretisation. Therefore, we only obtain the conservation of the modified momentum and modified actions. Similar results have also been shown in some other publications. For example, the authors in [19] proved long-time conservation of modified energy and modified action for the Störmer-Verlet method and in [35], conservation of the modified energy and modified magnetic moment were shown for a variational integrator. In both publications, long-time

conservation of natural invariants was not given. We also note that although the result cannot be obtained for the momentum K and actions I_j, K and I_j are no longer exactly conserved quantities in the semidiscretisation, which can be seen from Theorem 14.1. Moreover, it will be shown in the next subsection that, in comparison with the near-conservation of K and I_j, the modified momentum and modified actions are preserved rather well by the AAVF method. This supports the result of Theorem 14.3.

14.3.2 Numerical Experiments

In what follows, we implement two numerical experiments to show the behaviour of the AAVF method. Since the AAVF method is implicit, iteration solutions are needed. Here, we use fixed-point iteration in practical computation. We set 10^{-16} as the error tolerance and 100 as the maximum number of iterates.

Problem 14.1 Consider the semilinear wave equation (14.1), where $\rho = 0.5$ and $g(u) = -u^2$. The initial conditions are given by (see [1])

$$u(x,0) = 0.1\left(\frac{x}{\pi} - 1\right)^3\left(\frac{x}{\pi} + 1\right)^2, \quad \partial_t u(x,0) = 0.01\frac{x}{\pi}\left(\frac{x}{\pi} - 1\right)\left(\frac{x}{\pi} + 1\right)^2,$$

for $-\pi \leqslant x \leqslant \pi$. We carry out the spatial discretisation[1] with the dimension $2M = 2^7$ and apply the midpoint rule to the integral[2] appearing in the AAVF formula (14.5), which yields

$$\begin{cases} q_{n+1} = \phi_0(V)q_n + h\phi_1(V)p_n + h^2\phi_2(V)f((q_n + q_{n+1})/2), \\ p_{n+1} = -h\Omega^2\phi_1(V)q_n + \phi_0(V)p_n + h\phi_1(V)f((q_n + q_{n+1})/2). \end{cases} \tag{14.14}$$

It is easily verified that the assumption (14.7) holds for $s = 2$ with $\varepsilon \approx 0.1$. We solve this problem with the stepsize $h = 0.05$ on $[0, 10000]$, and the relative errors of momentum/modified momentum and actions/modified actions against t are shown

[1] It is noted that for wave equations, the spatial discretisation with the dimension $2M = 2^7$ has been considered in [1, 8, 36] and it worked well in those publications. That is the reason why we use the spatial discretisation with $2M = 2^7$ here.

[2] From the analysis of Sect. 14.5, it follows that the main result is still true for the AAVF method with some quadrature rule.

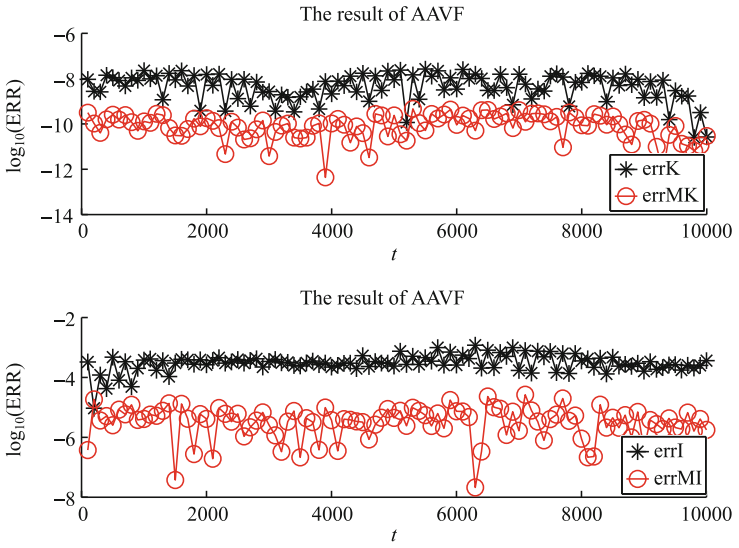

Fig. 14.1 The logarithm of the errors against t

in Fig. 14.1. We here adopt the following notations:

$$\text{errK} = \frac{|K(p_n, q_n) - K(p_0, q_0)|}{|K(p_0, q_0)|}, \qquad \text{errMK} = \frac{|\hat{K}(p_n, q_n) - \hat{K}(p_0, q_0)|}{|\hat{K}(p_0, q_0)|},$$

$$\text{errI} = \frac{\sum_{l=0}^{M} \omega_l^5 |I_l(p_n, q_n) - I_l(p_0, q_0)|}{\sum_{l=0}^{M} \omega_l^5 |I_l(p_0, q_0)|}, \qquad \text{errMI} = \frac{\sum_{l=0}^{M} \omega_l^5 |\hat{I}_l(p_n, q_n) - \hat{I}_l(p_0, q_0)|}{\sum_{l=0}^{M} \omega_l^5 |\hat{I}_l(p_0, q_0)|}.$$

It follows from Fig. 14.1 that the modified momentum and modified actions are better conserved than the momentum and actions, which supports the results given in Theorem 14.3.

We next show the efficiency of the AAVF method in comparison with some other methods. To this end, we consider the classical Störmer-Verlet formula (denoted by SV), Gautschi's method of order two (denoted by GM1s2) given in [17] and the two-stage diagonally implicit symplectic Runge–Kutta method of order three (denoted by RK2s3) presented in [37]. With regard to Gautschi's method, its coefficient functions are chosen as $\phi(\xi) = 1$ and $\psi(\xi) = (\sin(\xi)/\xi)^2$. The long-time behaviour of this method has been shown in [17], and the non-resonance conditions given in [1] are satisfied for this method. We first solve the system on $[0, 10]$ with $h = 0.2/2^j$ for $j = 2, 3, 4, 5$, and the errors $\text{GE} = \left(\|q_n - q\|_3^2 + \|p_n - p\|_2^2 \right)^{1/2}$ measured at the final time against the CPU time are presented in Fig. 14.2a. We then integrate the

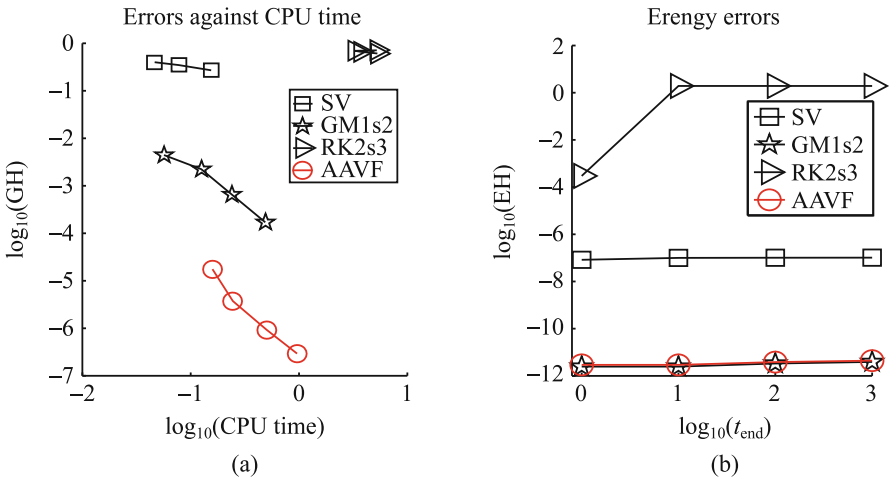

Fig. 14.2 (**a**) The logarithm of the errors against the logarithm of CPU time. (**b**) The logarithm of the energy errors against the logarithm of time

problem on $[0, t_{\text{end}}]$ with $h = 0.01$ and $t_{\text{end}} = 10^j$ for $j = 0, 1, 2, 3$. The errors of the semidiscrete energy conservation are presented in Fig. 14.2b. It can be observed from Fig. 14.2 that the AAVF method shows good overall efficiency.

Problem 14.2 Consider the semilinear Klein–Gordon equation

$$\begin{cases} \partial_t^2 u - a^2 \partial_x^2 u = bu^3 - au, \quad -\pi \leqslant x \leqslant \pi, \quad u(-\pi, t) = u(\pi, t), \quad 0 \leqslant t \leqslant T, \\ u(x, 0) = \sqrt{\dfrac{2a}{b}} \operatorname{sech}(\lambda x), \quad u_t(x, 0) = c\lambda \sqrt{\dfrac{2a}{b}} \operatorname{sech}(\lambda x) \tanh(\lambda x) \end{cases}$$

where $\lambda = \sqrt{\dfrac{a}{a^2 - c^2}}$ and $a, b, a^2 - c^2 > 0$. The exact solution is

$$u(x, t) = \sqrt{\frac{2a}{b}} \operatorname{sech}(\lambda(x - ct)).$$

The choice of parameters $a = 1, b = 0.01, c = 0.25$ makes this problem fit into the form (14.1).

Likewise, the spatial variable is discretised with the dimension $2M = 2^7$, and it can be verified that the assumption (14.7) is true for $s = 1$ with $\varepsilon \approx 0.015$. This problem is solved on $[0, 10000]$ with $h = 0.05$, and the relative errors of momentum/modified momentum and actions/modified actions against t are shown in Fig. 14.3.

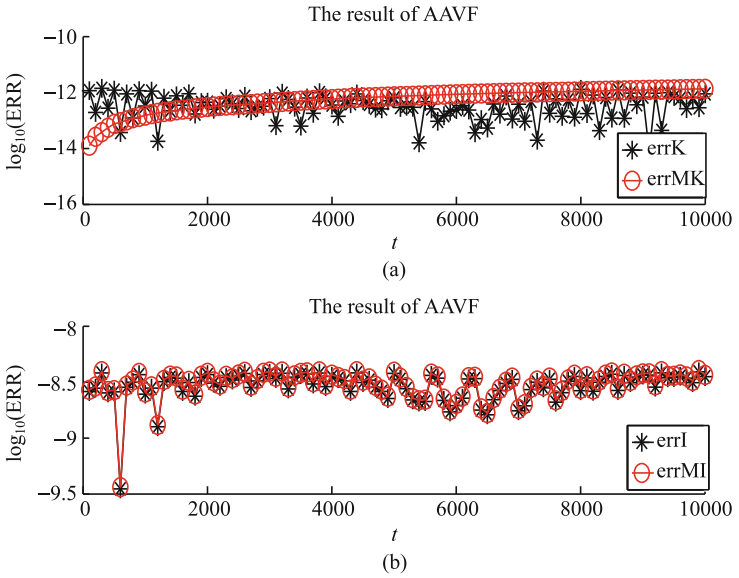

Fig. 14.3 The logarithm of the errors against t

We then apply the AAVF method as well as the methods SV, GM1s2, and RK2s3 to the problem on $[0, 100]$ with $h = 0.2/2^j$ for $j = 0, 1, 2, 3$. The errors measured at the final time against the CPU time are given in Fig. 14.4a. Finally we solve the problem on $[0, t_{\text{end}}]$ with $h = 0.01$ and $t_{\text{end}} = 10^j$ for $j = 0, 1, 2, 3$, and present the errors of the semidiscrete energy conservation in Fig. 14.4b. Here, it is remarked that for this problem, the conservation of modified momentum and modified actions seems to be similar to those for the natural discretisations of momentum and actions. The reason is that, for some problems, it can be checked that the modified momentum and modified actions are very close to the natural ones of the considered system. Apart from this, according to Fig. 14.4, it is clear that Gautschi's method behaves at least as well as AAVF since both methods behave similarly with respect to the conservation of invariants, but Gautschi's method is explicit while AAVF is implicit although both methods are of order two.

14.4 The Proof of the Main Result

This section concerns the proof of Theorem 14.3. We first present the outline of the proof and then show the key points one by one since the proof is a bit long.

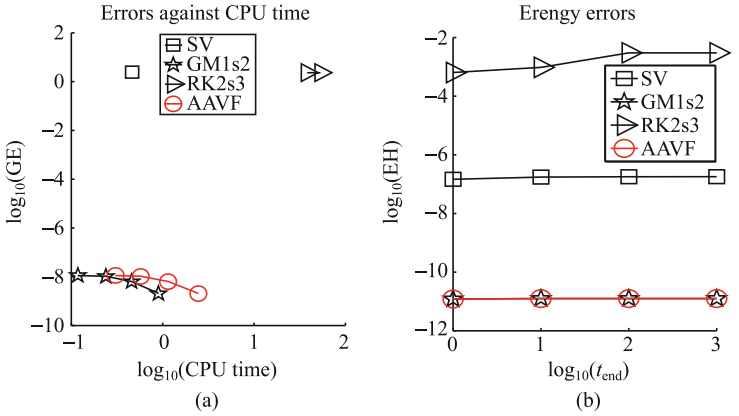

Fig. 14.4 (**a**) The logarithm of the errors against the logarithm of CPU time. (**b**) The logarithm of the energy errors against the logarithm of time

14.4.1 The Outline of the Proof

The proof relies on a careful study of a modulated Fourier expansion of the AAVF method (14.5). It is assumed that the conditions of Theorem 14.3 are true. For the numerical solution (p_n, q_n), determined by (14.5), we will consider the following truncated multi-frequency modulated Fourier expansion (with N from (14.9))

$$\tilde{q}(t) = \sum_{\|k\| \leqslant 2N} \mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t} \zeta^k(\varepsilon t), \quad \tilde{p}(t) = \sum_{\|k\| \leqslant 2N} \mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t} \eta^k(\varepsilon t), \tag{14.15}$$

where $t = nh$ and $\zeta_{-j}^{-k} = \overline{\zeta_j^k}$, $\eta_{-j}^{-k} = \overline{\eta_j^k}$. For this modulated Fourier expansion, the following key points will be addressed one by one in the rest of this section.

- Formal modulation equations for the modulation functions are derived in Sect. 14.4.2.
- We consider an iterative construction of the functions using reverse Picard iteration in Sect. 14.4.3.
- We then work with a more convenient rescaling and study the estimation of nonlinear terms in Sect. 14.4.4.
- Abstract reformulation of the iteration is presented in Sect. 14.4.5.
- We control the size of the numerical solution by studying the bounds of modulation functions in Sect. 14.4.6.
- The bound of the defect is estimated in Sect. 14.4.7.
- We study the difference between the numerical solution and its modulated Fourier expansion in Sect. 14.4.8.
- We show two invariants of the modulation system and establish their relationship with the modified momentum and modified actions in Sect. 14.4.9.

- Finally, the previous results that are valid only on a short time interval are extended to a long-time interval in Sect. 14.4.10.

It is noted that the above procedure is a standard approach to the study of the long-time behaviour of numerical methods of Hamiltonian partial differential equations by using modulated Fourier expansions (see, e.g. [1–3, 15, 16]). Although the proof presented here closely follows these previous publications, there are novel modifications adapted to the AAVF method in each part. The differences in the analysis arise due to the implicitness of the AAVF method and the integral appearing in the method.

Throughout the proof, denote by C a generic constant which is independent of ε, M, h and $t = nh$. The following lemma, presented in [2], will be needed in the analysis of this chapter.

Lemma 14.1 (See [2]) *For $s > 1/2$, one has $\sum_{\|k\| \leqslant K} \boldsymbol{\omega}^{-2s|k|} \leqslant C_{K,s} \leqslant \infty$. For $s > 1/2$ and $m \geqslant 2$, it is true that*

$$\sup_{\|k\| \leqslant K} \sum_{k^1 + \cdots + k^m = k} \frac{\boldsymbol{\omega}^{-2s(|k^1| + \cdots + |k^m|)}}{\boldsymbol{\omega}^{-2s|k|}} \leqslant C_{m,K,s} < \infty,$$

where the sum is taken over $(k^1, \cdots, k^m)$ satisfying $\|k^i\| \leqslant K$. For $s \geqslant 1$, it is further true that $\sup_{\|k\| \leqslant K} \dfrac{\sum_{l \geqslant 0} |k_l| \, \omega_l^{2s+1}}{\boldsymbol{\omega}^{2s|k|}(1 + |k \cdot \boldsymbol{\omega}|)} \leqslant C_{K,s} < \infty.$

14.4.2 Modulation Equations

We commence from the formulation of the modulation equations for the modulated functions. To this end, we first define five operators by

$$\begin{aligned}
L_1^k &:= \mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{\varepsilon h D} - 2\cos(h\Omega) + \mathrm{e}^{-\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{-\varepsilon h D},\\
L_2^k &:= \mathrm{e}^{\frac{1}{2}\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{\frac{1}{2}\varepsilon h D} + \mathrm{e}^{-\frac{1}{2}\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{-\frac{1}{2}\varepsilon h D},\\
L_3^k &:= (\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{\varepsilon h D} - 1)(\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{\varepsilon h D} + 1)^{-1},\\
L_4^k(\sigma) &:= (1-\sigma)\mathrm{e}^{-\frac{1}{2}\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{-\frac{h}{2}\varepsilon D} + \sigma\mathrm{e}^{\frac{1}{2}\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{\frac{h}{2}\varepsilon D},\\
L^k &:= (L_2^k)^{-1} L_1^k,
\end{aligned}$$

where D is the differential operator (see [20]). Then the following results for these operators are essential in the analysis.

Proposition 14.1 *The operator L^k can be expressed in Taylor expansions as follows:*

$$\begin{aligned} L^{\pm\langle j\rangle}(hD)\alpha_j^{\pm\langle j\rangle}(\varepsilon t) &= \pm 2i\varepsilon h s_{\langle j\rangle}\dot{\alpha}_j^{\pm\langle j\rangle}(\varepsilon t) + \frac{1}{2}\varepsilon^2 h^2 \sec\left(\frac{1}{2}h\omega_j\right)\ddot{\alpha}_j^{\pm\langle j\rangle}(\varepsilon t) + \cdots, \\ L^k(hD)\alpha_j^k(\varepsilon t) &= 2\frac{s_{\langle j\rangle+k}s_{\langle j\rangle-k}}{c_k}\alpha_j^k(\varepsilon t) + i\varepsilon h\frac{s_k(1+c_{\langle j\rangle+k}c_{\langle j\rangle-k})}{c_k^2}\dot{\alpha}_j^k(\varepsilon t) + \cdots, \end{aligned} \tag{14.16}$$

for $|j| > 0$ and $k \neq \pm\langle j\rangle$, where $s_k = \sin\left(\frac{h}{2}(k\cdot\boldsymbol{\omega})\right)$ and $c_k = \cos\left(\frac{h}{2}(k\cdot\boldsymbol{\omega})\right)$. The Taylor expansions of L_3^k are of the forms

$$L_3^k\alpha_j^k(\varepsilon t) = i\tan\left(\frac{1}{2}h(k\cdot\boldsymbol{\omega})\right)\alpha_j^k(\varepsilon t) + \frac{h\varepsilon}{1+c_{2k}}\dot{\alpha}_j^k(\varepsilon t) + \cdots,$$

for $|j| > 0$ and $\|k\| \leqslant 2N$. Moreover, for the operator $L_4^k(\sigma)$ with $\|k\| \leqslant 2N$, we have

$$L_4^k\left(\frac{1}{2}\right) = \cos\left(\frac{h(k\cdot\boldsymbol{\omega})}{2}\right) + \frac{1}{2}\sin\left(\frac{h(k\cdot\boldsymbol{\omega})}{2}\right)(ih\varepsilon D) + \cdots.$$

Theorem 14.4 (Modulation Equations) *The formal modulation equations of the modulated functions ζ^k are given by*

$$\begin{aligned} L^{\pm\langle j\rangle}\zeta_j^{\pm\langle j\rangle} =& -h^2\phi_2(h^2\omega_j^2)\sum_{m\geqslant 2}\frac{g^{(m)}(0)}{m!}\sum_{k^1+\cdots+k^m=\pm\langle j\rangle}\sum_{j_1+\cdots+j_m\equiv j\ mod\ 2M}{}' \\ &\cdot\int_0^1\left[\left(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\right)(t\varepsilon,\sigma)\right]\mathrm{d}\sigma, \\ L^k\zeta_j^k =& -h^2\phi_2(h^2\omega_j^2)\sum_{m\geqslant 2}\frac{g^{(m)}(0)}{m!}\sum_{k^1+\cdots+k^m=k}\sum_{j_1+\cdots+j_m\equiv j\ mod\ 2M}{}' \\ &\cdot\int_0^1\left[\left(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\right)(t\varepsilon,\sigma)\right]\mathrm{d}\sigma, \qquad \text{for } k\neq\pm\langle j\rangle, \end{aligned} \tag{14.17}$$

where L^k is defined by (14.16) and

$$\xi^k(\varepsilon t,\sigma) = L_4^k(\sigma)\zeta^k(\varepsilon t).$$

The modulation equations for η^k are determined by

$$\eta_j^{\pm\langle j\rangle} = \pm \mathrm{i}\omega_j \zeta_j^{\pm\langle j\rangle} + \mathscr{O}(h\varepsilon), \quad \eta_j^k = \frac{\tan\left(\frac{1}{2}h(k\cdot\boldsymbol{\omega})\right)}{\tan\left(\frac{1}{2}h\omega_j\right)}\mathrm{i}\omega_j\zeta_j^k + \mathscr{O}(h\varepsilon) \tag{14.18}$$

for $k \neq \pm\langle j\rangle$.

Proof The proof will be divided into two parts.

The first part is the proof of (14.17).

Using the symmetry of the AAVF method and the following property

$$\int_0^1 f((1-\sigma)q_n + \sigma q_{n-1})\mathrm{d}\sigma = \int_0^1 f((1-\sigma)q_{n-1} + \sigma q_n)\mathrm{d}\sigma,$$

leads to

$$\begin{aligned} &q_{n+1} - 2\cos(h\Omega)q_n + q_{n-1} \\ &= h^2\phi_2(V)\left[\int_0^1 f((1-\sigma)q_n + \sigma q_{n+1})\mathrm{d}\sigma + \int_0^1 f((1-\sigma)q_{n-1} + \sigma q_n)\mathrm{d}\sigma\right]. \end{aligned} \tag{14.19}$$

We then seek for a modulated Fourier expansion of the form

$$\tilde{q}_h(t+\frac{h}{2},\sigma) = \sum_{\|k\|\leqslant 2N} \mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})\left(t+\frac{h}{2}\right)}\xi^k\left(\varepsilon\left(t+\frac{h}{2}\right),\sigma\right)$$

for the term $(1-\sigma)q_n + \sigma q_{n+1}$ appearing in (14.19). This implies that

$$\begin{aligned} \xi^k\left(\varepsilon\left(t+\frac{h}{2}\right),\sigma\right) &= \left((1-\sigma)\mathrm{e}^{-\frac{1}{2}\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{-\frac{h}{2}\varepsilon D} + \sigma\mathrm{e}^{\frac{1}{2}\mathrm{i}(k\cdot\boldsymbol{\omega})h}\mathrm{e}^{\frac{h}{2}\varepsilon D}\right)\zeta^k\left(\varepsilon\left(t+\frac{h}{2}\right)\right) \\ &= L_4^k(\sigma)\zeta^k\left(\varepsilon\left(t+\frac{h}{2}\right)\right). \end{aligned} \tag{14.20}$$

Likewise, for $(1-\sigma)q_{n-1} + \sigma q_n$, we can obtain the following modulated Fourier expansion

$$\tilde{q}_h(t-\frac{h}{2},\sigma) = \sum_{\|k\|\leqslant 2N} \mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})(t-\frac{h}{2})}\xi^k\left(\varepsilon\left(t-\frac{h}{2}\right),\sigma\right)$$

where

$$\xi^k\left(\varepsilon\left(t-\frac{h}{2}\right),\sigma\right)=L_4^k(\sigma)\zeta^k\left(\varepsilon\left(t-\frac{h}{2}\right)\right). \tag{14.21}$$

Inserting the modulated Fourier expansions (14.15), (14.20), and (14.21) into (14.19) yields

$$\begin{aligned}&\tilde{q}(t+h)-2\cos(h\Omega)\tilde{q}(t)+\tilde{q}(t-h)\\ =&h^2\phi_2(V)\left[\int_0^1 f\left(\tilde{q}_h\left(t+\frac{h}{2},\sigma\right)\right)\mathrm{d}\sigma+\int_0^1 f\left(\tilde{q}_h\left(t-\frac{h}{2},\sigma\right)\right)\mathrm{d}\sigma\right],\end{aligned}$$

which can be formulated as

$$(\mathrm{e}^{\frac{1}{2}hD}+\mathrm{e}^{-\frac{1}{2}hD})^{-1}(\mathrm{e}^{hD}-2\cos(h\Omega)+\mathrm{e}^{-hD})\tilde{q}(t)=h^2\phi_2(V)\int_0^1 f(\tilde{q}_h(t,\sigma))\mathrm{d}\sigma. \tag{14.22}$$

We next rewrite this equation by using the approach introduced in [3]. We begin with the following notation. For a 2π-periodic function $w(x)$, denote by $(\mathscr{Q}w)(x)$ the trigonometric interpolation polynomial to $w(x)$ at the points x_k. If $w(x)$ is of the form $w(x)=\sum_{j=-\infty}^{\infty} w_j\mathrm{e}^{\mathrm{i}jx}$, then we have that

$$(\mathscr{Q}w)(x)=\sum_{|j|\leqslant M}''\left(\sum_{l=-\infty}^{\infty} w_{j+2Ml}\right)\mathrm{e}^{\mathrm{i}jx},$$

where $x_k=\dfrac{k\pi}{M}$. For a $2M$-periodic coefficient sequence $q=(q_j)$, $(\mathscr{P}q)(x)$ is referred to the trigonometric polynomial with coefficients q_j, i.e.,

$$(\mathscr{P}q)(x)=\sum_{|j|\leqslant M}' q_j\mathrm{e}^{\mathrm{i}jx}.$$

With these new denotations, (14.22) is identical to

$$(\mathrm{e}^{\frac{1}{2}hD}+\mathrm{e}^{-\frac{1}{2}hD})^{-1}(\mathrm{e}^{hD}-2\cos(h\Omega)+\mathrm{e}^{-hD})\mathscr{P}\tilde{q}(t)=h^2\phi_2(V)\int_0^1 \mathscr{Q}g(\mathscr{P}\tilde{q}_h(t,\sigma))\mathrm{d}\sigma. \tag{14.23}$$

The Taylor expansion of the non-linearity $\mathscr{Q}g$ at 0 is given by[3]

$$\begin{aligned}
&\mathscr{Q}g(\mathscr{P}\tilde{q}_h(t,\sigma)) = \sum_{m\geqslant 2}\frac{g^{(m)}(0)}{m!}\mathscr{Q}(\mathscr{P}\tilde{q}_h(t,\sigma))^m \\
&= \sum_{m\geqslant 2}\frac{g^{(m)}(0)}{m!}\Bigg(\sum_{|j_1|\leqslant M}''\sum_{l=-\infty}^{\infty}\sum_{\|k^1\|\leqslant 2N}{}' \mathrm{e}^{\mathrm{i}(k^1\cdot\boldsymbol{\omega})t}\xi^{k^1}_{j_1+2Ml}(\tau,\sigma)\mathrm{e}^{\mathrm{i}j_1x}\Bigg) \\
&\quad\cdots\Bigg(\sum_{|j_m|\leqslant M}''\sum_{l=-\infty}^{\infty}\sum_{\|k^m\|\leqslant 2N}{}' \mathrm{e}^{\mathrm{i}(k^m\cdot\boldsymbol{\omega})t}\xi^{k^m}_{j_m+2Ml}(\tau,\sigma)\mathrm{e}^{\mathrm{i}j_mx}\Bigg) \\
&= \sum_{m\geqslant 2}\frac{g^{(m)}(0)}{m!}\sum_{|j|\leqslant M}''\sum_{j_1+\cdots+j_m\equiv j \bmod 2M}{}'\sum_{\|k^1\|\leqslant 2N,\cdots,\|k^m\|\leqslant 2N}(\xi^{k^1}_{j_1}\cdots\xi^{k^m}_{j_m})(\tau,\sigma) \\
&\quad \mathrm{e}^{\mathrm{i}((k^1+\cdots+k^m)\cdot\boldsymbol{\omega})t}\mathrm{e}^{\mathrm{i}jx},
\end{aligned}$$

where $\tau = h\varepsilon$ and the prime on the sum indicates that a factor $1/2$ is included in the appearance of $\xi^{k^i}_{j_i}$ with $j_i = \pm M$. Inserting this into (14.23), considering the j-th Fourier coefficient and comparing the coefficients of $\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t}$, we obtain (14.17).

On the other hand, we need to derive the initial values for $\dot{\zeta}_j^{\pm\langle j\rangle}$ appearing in (14.17). On noticing the fact that $\tilde{q}(0) = q(0)$, we obtain

$$\zeta_j^{\langle j\rangle}(0) + \zeta_j^{-\langle j\rangle}(0) = q_j(0) - \sum_{k\neq\pm\langle j\rangle}\zeta_j^k(0). \tag{14.24}$$

Furthermore, it follows from $\tilde{p}(0) = p(0)$ that

$$\eta_j^{\langle j\rangle}(0) + \eta_j^{-\langle j\rangle}(0) = p_j(0) - \sum_{k\neq\pm\langle j\rangle}\eta_j^k(0),$$

which results in

$$\begin{aligned}
&\mathrm{i}\omega_j(\zeta_j^{\langle j\rangle}(0) - \zeta_j^{-\langle j\rangle}(0)) = p_j(0) - \sum_{k\neq\pm\langle j\rangle}\eta_j^k(0) \\
&= p_j(0) - \sum_{k\neq\pm\langle j\rangle}\frac{\tan\left(\frac{1}{2}h(k\cdot\boldsymbol{\omega})\right)}{\tan\left(\frac{1}{2}h\omega_j\right)}\mathrm{i}\omega_j\zeta_j^k(0) + \mathscr{O}(h\varepsilon).
\end{aligned} \tag{14.25}$$

The formulae (14.24) and (14.25) determine the initial values for $\zeta_j^{\pm\langle j\rangle}$.

[3] It is noted that $g(0) = 0$ and $g'(0) = 0$ are used here.

We now turn to the second part, the proof of (14.18).

For the modulation equations of η^k, it follows from (14.5) that

$$q_{n+1} - q_n = \Omega^{-1} \tan\left(\frac{1}{2}h\Omega\right)(p_{n+1} + p_n). \tag{14.26}$$

According to the definition of L_3, this relation can be expressed as

$$L_3^k \zeta^k = \Omega^{-1} \tan\left(\frac{1}{2}h\Omega\right)\eta^k.$$

It then follows from the Taylor series of L_3^k that the relationship between η^k and ζ^k can be established by (14.18). The proof then is complete. □

14.4.3 Reverse Picard Iteration

In what follows, we consider the reverse Picard iteration (see [1, 3]) of the functions ζ^k such that after $4N$ iteration steps, the defects in (14.17), (14.24), and (14.25) are of magnitude $\mathscr{O}(\varepsilon^{N+1})$ in the H^s norm.

We here denote by $[\cdot]^{(n)}$ the nth iterate. For $k = \pm\langle j\rangle$ and under the condition (14.17), we design the iteration procedure as follows:

$$\pm 2\mathrm{i}s_{\langle j\rangle}h\varepsilon\big[\dot{\zeta}_j^{\pm\langle j\rangle}\big]^{(n+1)} = \Bigg[-h^2\phi_2(h^2\omega_j^2)\sum_{m\geqslant 2}\frac{g^{(m)}(0)}{m!}\sum_{k^1+\cdots+k^m=k}\sum_{j_1+\cdots+j_m\equiv j \bmod 2M}{}'$$
$$\cdot\int_0^1\Big[\big(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\big)(t\varepsilon,\sigma)\Big]\mathrm{d}\sigma - \Big(\frac{1}{2}\varepsilon^2h^2\sec(\frac{1}{2}h\omega_j)\ddot{\zeta}_j^{\pm\langle j\rangle} + \cdots\Big)\Bigg]^{(n)}. \tag{14.27}$$

For $k \neq \pm\langle j\rangle$ and j subject to the non-resonant condition (14.8), the iteration procedure is of the form

$$2\frac{s_{\langle j\rangle+k}s_{\langle j\rangle-k}}{c_k}\big[\dot{\zeta}_j^k\big]^{(n+1)} = \Bigg[-h^2\phi_2(h^2\omega_j^2)\sum_{m\geqslant 2}\frac{g^{(m)}(0)}{m!}\sum_{k^1+\cdots+k^m=k}\sum_{j_1+\cdots+j_m\equiv j \bmod 2M}{}'$$
$$\cdot\int_0^1\Big[\big(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\big)(t\varepsilon,\sigma)\Big]\mathrm{d}\sigma - \Big(\mathrm{i}\varepsilon h\frac{s_k(1+c_{\langle j\rangle+k}c_{\langle j\rangle-k})}{c_k^2}\dot{\zeta}_j^k + \cdots\Big)\Bigg]^{(n)}, \tag{14.28}$$

where $\zeta_j^k = 0$ for $k \neq \pm\langle j\rangle$ in the near-resonant set $\mathscr{R}_{\varepsilon,h}$. For the initial values (14.24) and (14.25), the iteration procedure is given by

$$\left[\zeta_j^{\langle j\rangle}(0) + \zeta_j^{-\langle j\rangle}(0)\right]^{(n+1)} = \left[q_j(0) - \sum_{k\neq\pm\langle j\rangle} \zeta_j^k(0)\right]^{(n)},$$

$$\mathrm{i}\omega_j\left[\zeta_j^{\langle j\rangle}(0) - \zeta_j^{-\langle j\rangle}(0)\right]^{(n+1)} = \left[p_j(0) - \sum_{k\neq\pm\langle j\rangle} \frac{\tan\left(\frac{1}{2}h(k\cdot\boldsymbol{\omega})\right)}{\tan\left(\frac{1}{2}h\omega_j\right)}\mathrm{i}\omega_j\zeta_j^k(0) + \mathscr{O}(h\varepsilon)\right]^{(n)}. \tag{14.29}$$

It is assumed that $\|k\| \leqslant K := 2N$ and $\left\|k^i\right\| \leqslant K$ for $i = 1, \cdots, m$, in these iterations. We here remark that the procedure includes an initial value problem of first-order ODEs for $\zeta_j^{\pm\langle j\rangle}$ (for $|j| \leqslant M$) and algebraic equations for ζ_j^k with $k \neq \pm\langle j\rangle$ at each iteration step. The starting iterates ($n = 0$) are chosen as $\zeta_j^k(\tau) = 0$ for $k \neq \pm\langle j\rangle$, and $\zeta_j^{\pm\langle j\rangle}(\tau) = \zeta_j^{\pm\langle j\rangle}(0)$, where $\zeta_j^{\pm\langle j\rangle}(0)$ are determined by (14.29). Obviously, the iteration procedure is well defined.

14.4.4 Rescaling and Estimation of the Nonlinear Terms

In a similar way to Sect. 3.5 of [2] and Sect. 6.3 of [1], we next consider a more convenient rescaling

$$c\zeta_j^k = \frac{\boldsymbol{\omega}^{|k|}}{\varepsilon^{[[k]]}}\zeta_j^k, \quad c\zeta^k = \left(c\zeta_j^k\right)_{|j|\leqslant M} = \frac{\boldsymbol{\omega}^{|k|}}{\varepsilon^{[[k]]}}\zeta^k$$

in the space $H^s = (H^s)^{\mathscr{K}} = \{c\zeta = (c\zeta^k)_{k\in\mathscr{K}} : c\zeta^k \in H^s\}$. The norm of this space is defined by $|||c\zeta|||_s^2 = \sum_{k\in\mathscr{K}} \left\|c\zeta^k\right\|_s^2$, where the set $\mathscr{K}$ is given by $\mathscr{K} = \{k = (k_l)_{l=0}^M$ with integers $k_l : \|k\| \leqslant K\}$ with $K = 2N$. Likewise, we use the notation $c\xi^k \in \mathrm{H}^s$ having the same meaning.

With regard to the expression of the non-linearity for (14.17) in these rescaled variables, we define the nonlinear function $f = (f_j^k)$ by

$$f_j^k\left(c\xi(\tau)\right) = \frac{\boldsymbol{\omega}^{|k|}}{\varepsilon^{[[k]]}} \sum_{m=2}^{N} \frac{g^{(m)}(0)}{m!} \sum_{k^1+\cdots+k^m=k} \frac{\varepsilon^{[[k^1]]+\cdots+[[k^m]]}}{\boldsymbol{\omega}^{|k^1|+\cdots+|k^m|}}$$
$$\cdot \sum_{j_1+\cdots+j_m\equiv j \bmod 2M}{}' \int_0^1 \left(c\xi_{j_1}^{k^1}\cdot\cdots\cdot c\xi_{j_m}^{k^m}\right)(\tau,\sigma)\mathrm{d}\sigma.$$

Concerning this nonlinear function, we have the following bounds, which can be proved by using the similar arguments presented in [1, 2].

Proposition 14.2 (Estimation of the Nonlinear Terms) *It is true that*

$$\sum_{k\in\mathscr{K}}\left\|f^k(c\xi)\right\|_s^2\leqslant C\varepsilon P(|||c\tilde{\xi}|||_s^2),\quad \sum_{|j|\leqslant M}\left\|f^{\pm\langle j\rangle}(c\xi)\right\|_s^2\leqslant C\varepsilon^3 P_1(|||c\tilde{\xi}|||_s^2), \tag{14.30}$$

where $c\tilde{\xi}(\tau):=\sup_{0\leqslant\sigma\leqslant 1}\{c\xi(\tau,\sigma)\}$ *and* P *and* P_1 *are polynomials with coefficients bounded independently of* ε, h, *and* M.

Similarly, we can consider different rescaling

$$\hat{c}\zeta_j^k=\frac{\boldsymbol{\omega}^{s|k|}}{\varepsilon^{[[k]]}}\zeta_j^k,\quad \hat{c}\zeta^k=\left(\hat{c}\zeta_j^k\right)_{|j|\leqslant M}=\frac{\boldsymbol{\omega}^{s|k|}}{\varepsilon^{[[k]]}}\zeta^k \tag{14.31}$$

in $H^1=(H^1)^{\mathscr{K}}$ with norm $|||\hat{c}\zeta|||_1^2=\sum\limits_{\|k\|\leqslant K}\left\|\hat{c}\zeta^k\right\|_1^2$, where $\hat{f}_j^k$ is exactly the same as f_j^k, but with $\boldsymbol{\omega}^{|k|}$ replaced by $\boldsymbol{\omega}^{s|k|}$. We use similar notations $\hat{c}\xi^k\in\mathrm{H}^1$ and also obtain similar bounds

$$\sum_{k\in\mathscr{K}}\left\|\hat{f}^k(\hat{c}\xi)\right\|_1^2\leqslant C\varepsilon\hat{P}(|||\hat{c}\tilde{\xi}|||_1^2),\quad \sum_{|j|\leqslant M}\left\|\hat{f}^{\pm\langle j\rangle}(\hat{c}\xi)\right\|_1^2\leqslant C\varepsilon^3\hat{P}_1(|||\hat{c}\tilde{\xi}|||_1^2),$$

with other functions $\hat{P}$ and $\hat{P}_1$.

14.4.5 Reformulation of the Reverse Picard Iteration

This subsection concerns the reverse Picard iteration. On the basis of the two cases: $k=\pm\langle j\rangle$ and $k\neq\pm\langle j\rangle$, we split $c\zeta$ into two parts as follows:

$$\begin{cases} a\zeta_j^k=c\zeta_j^k & \text{if } k=\pm\langle j\rangle,\quad \text{and 0 else,}\\ b\zeta_j^k=c\zeta_j^k & \text{if (14.8) is satisfied,}\quad \text{and 0 else.}\end{cases} \tag{14.32}$$

It is noted that for $a\zeta=(a\zeta_j^k)\in\mathrm{H}^s$ and $b\zeta=(b\zeta_j^k)\in\mathrm{H}^s$, we have $a\zeta+b\zeta=c\zeta$ and $|||a\zeta|||_s^2+|||b\zeta|||_s^2=|||c\zeta|||_s^2$. Here, the same notation and property are used for $c\xi$.

We now rewrite the iterations (14.27) and (14.28) in an abstract form

$$\begin{cases} a\dot{\zeta}^{(n+1)} = \Omega^{-1}F(a\zeta^{(n)}, b\zeta^{(n)}) - Aa\zeta^{(n)}, \\ b\zeta^{(n+1)} = \Omega^{-1}\Psi G(a\zeta^{(n)}, b\zeta^{(n)}) - Bb\zeta^{(n)}, \end{cases} \tag{14.33}$$

where

$$(\Omega x)_j^k = (\omega_j + |k\cdot\boldsymbol{\omega}|)x_j^k, \quad (\Psi x)_j^k = 2\phi_2(h^2\omega_j^2)\cos\left(\frac{1}{2}h(k\cdot\boldsymbol{\omega})\right)x_j^k,$$

and the operators A, B are respectively given by

$$(Aa\zeta)_j^{\pm\langle j\rangle}(\tau) = \frac{1}{\pm 2\mathrm{i}s_{\langle j\rangle}h\varepsilon}\left(\frac{1}{2}\varepsilon^2h^2\sec\left(\frac{1}{2}h\omega_j\right)a\ddot{\zeta}_j^{\pm\langle j\rangle} + \cdots\right),$$

$$(Bb\zeta)_j^k(\tau) = \frac{c_k}{2s_{\langle j\rangle+k}s_{\langle j\rangle-k}}\left(\mathrm{i}\varepsilon h\frac{s_k(1+c_{\langle j\rangle+k}c_{\langle j\rangle-k})}{c_k^2}b\dot{\zeta}_j^k + \cdots\right)$$

for (j,k) subject to (14.8).

The functions $F = (F_j^k)$ and $G = (G_j^k)$ are defined respectively by

$$F_j^{\pm\langle j\rangle}(a\zeta, b\zeta) = \frac{1}{\mp\mathrm{i}\varepsilon}\frac{2\phi_2(h^2\omega_j^2)}{\mathrm{sinc}\left(\frac{1}{2}h\omega_j\right)}f_j^{\pm\langle j\rangle}(c\xi), \quad G_j^k(a\zeta, b\zeta) = -\frac{h^2(\omega_j + |k\cdot\boldsymbol{\omega}|)}{4s_{\langle j\rangle+k}s_{\langle j\rangle-k}}f_j^k(c\xi)$$

for (j,k) subject to (14.8).

Theorem 14.5 *The operators A and B are bounded by*

$$|||(Aa\zeta)(\tau)|||_s \leqslant C\sum_{l=2}^{N}h^{l-2}\varepsilon^{l-3/2}\left|\left|\left|\frac{\mathrm{d}^l}{\mathrm{d}\tau^l}(a\zeta)(\tau)\right|\right|\right|_s,$$

$$|||(Bb\zeta)(\tau)|||_s \leqslant C\varepsilon^{1/2}|||(b\dot{\zeta})(\tau)|||_s + C\sum_{l=2}^{N}h^{l-2}\varepsilon^{l-1/2}\left|\left|\left|\frac{\mathrm{d}^l}{\mathrm{d}\tau^l}(b\zeta)(\tau)\right|\right|\right|_s.$$

Moreover, we have

$$|||F|||_s \leqslant C\varepsilon^{1/2}, \quad ||G|||_s \leqslant C, \quad |||\Psi^{-1}\Omega^{-1}F|||_s \leqslant C.$$

Proof The bound of A follows from

$$\left|\frac{1}{\pm 2\mathrm{i}s_{\langle j\rangle}h\varepsilon}\frac{1}{2}\varepsilon^2h^2\sec\left(\frac{1}{2}h\omega_j\right)\right| = \left|\frac{\frac{1}{2}h\varepsilon}{\sin(h\omega_j)}\right| \leqslant \frac{1}{2}\varepsilon^{1/2}.$$

We compute

$$\left|\frac{c_k}{2s_{\langle j\rangle+k}s_{\langle j\rangle-k}}\mathrm{i}\varepsilon h\frac{s_k(1+c_{\langle j\rangle+k}c_{\langle j\rangle-k})}{c_k^2}\right| \leqslant \left|\frac{\varepsilon h}{\varepsilon^{1/2}h^2(\omega_j+|k\cdot\boldsymbol{\omega}|)}\frac{s_k(1+c_{\langle j\rangle+k}c_{\langle j\rangle-k})}{c_k}\right|$$

$$\leqslant \frac{\varepsilon^{1/2}}{h}\frac{\frac{h}{2}|k\cdot\boldsymbol{\omega}|}{\omega_j+|k\cdot\boldsymbol{\omega}|}\left|\frac{1+c_{\langle j\rangle+k}c_{\langle j\rangle-k}}{c_k}\right| \leqslant C\varepsilon^{1/2},$$

where $|s_k| \leqslant \frac{h}{2}|k\cdot\boldsymbol{\omega}|$ is used. Hence, we obtain the bound of B.

It follows from

$$\left|\frac{2\phi_2(h^2\omega_j^2)}{\mathrm{sinc}\left(\frac{1}{2}h\omega_j\right)}\right| = \left|\mathrm{sinc}\left(\frac{1}{2}h\omega_j\right)\right| \leqslant 1$$

and (14.30) that $|||F|||_s \leqslant C\varepsilon^{1/2}$. Then using (14.8) and (14.30) yields $|||G|||_s \leqslant C$. Furthermore, according to (14.11), we obtain

$$|||\Psi^{-1}\Omega^{-1}F|||_s^2 = \sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s}\left|(\Psi^{-1}\Omega^{-1}F)_j^k\right|^2 = \sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s}\left|\frac{h/2}{\varepsilon\sin(h\varepsilon)}\right|^2\left|f_j^{\pm\langle j\rangle}\right|^2$$

$$\leqslant C\sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s}\left|\frac{1}{\varepsilon^{3/2}}\right|^2\left|f_j^{\pm\langle j\rangle}\right|^2 = C\frac{1}{\varepsilon^3}|||f^{\pm\langle j\rangle}|||_s^2 \leqslant C.$$

This shows $|||\Psi^{-1}\Omega^{-1}F|||_s \leqslant C$. The proof is complete. □

With regard to the initial value condition (14.29), it can be rewritten as

$$a\zeta^{(n+1)}(0) = v + Pb\zeta^{(n)}(0) + Qb\zeta^{(n)}(0), \tag{14.34}$$

where $v_j^{\pm\langle j\rangle} = \frac{\omega_j}{\varepsilon}\left(\frac{1}{2}q_j(0) \mp \frac{\mathrm{i}}{2\omega_j}p_j(0)\right)$ and the operators P and Q are given by

$$(Pb\zeta)_j^{\pm\langle j\rangle}(0) = -\frac{1}{2}\frac{\omega_j}{\varepsilon}\sum_{k\neq\pm\langle j\rangle}\frac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{|k|}}b\zeta_j^k(0),$$

$$(Qb\zeta)_j^{\pm\langle j\rangle}(0) = \mp\frac{1}{2\omega_j}\frac{\omega_j}{\varepsilon}\sum_{k\neq\pm\langle j\rangle}\frac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{|k|}}b\eta_j^k(0).$$

It can be verified from (14.7) that v is bounded in H^s. For the bounds of the operators P and Q, we have

$$\begin{aligned}
|||Pb\zeta(0)|||_s^2 &= \sum_{k\in\mathscr{K}} \sum_{|j|\leqslant M}'' \omega_j^{2s} \left| \frac{1}{2}\frac{\omega_j}{\varepsilon} \sum_{k\neq\pm\langle j\rangle} \frac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{|k|}} b\zeta_j^k(0) \right|^2 \\
&\leqslant \frac{1}{4\varepsilon^2} \sum_{k\in\mathscr{K}} \sum_{|j|\leqslant M}'' \omega_j^{2s+2} \left(\sum_{k\neq\pm\langle j\rangle} \frac{\varepsilon^{2[[k]]}}{\boldsymbol{\omega}^{2|k|}} \right) \left(\sum_{k\neq\pm\langle j\rangle} b\zeta_j^k(0)^2 \right) \\
&\leqslant \frac{1}{4} \sum_{k\in\mathscr{K}} \sum_{|j|\leqslant M}'' \omega_j^{2s+2} \left(\sum_{k\neq\pm\langle j\rangle} \boldsymbol{\omega}^{-2|k|} \right) \left(\sum_{k\neq\pm\langle j\rangle} b\zeta_j^k(0)^2 \right) \\
&\leqslant C|||\Omega b\zeta(0)|||_s^2 \leqslant C|||b\zeta(0)|||_{s+1}^2.
\end{aligned}$$

Likewise, we can obtain

$$|||(Qb\zeta)(0)|||_s^2 \leqslant C|||b\eta(0)|||_s^2.$$

Therefore, the bounds $|||(Pb\zeta)(0)|||_s \leqslant C$ and $|||(Qb\zeta)(0)|||_s \leqslant C$ are confirmed. Finally, we remark that the starting iterates of (14.34) are chosen as $a\zeta^{(0)}(\tau) = v$ and $b\zeta^{(0)}(\tau) = 0$, respectively.

14.4.6 Bounds of the Coefficient Functions

Theorem 14.6 (Bounds of the Modulation Functions) *The modulation functions ζ^k of (14.15) are bounded by*

$$\sum_{\|k\|\leqslant 2N} \left(\frac{\boldsymbol{\omega}^{|k|}}{\varepsilon^{[[k]]}} \left\| \zeta^k(\varepsilon t) \right\|_s \right)^2 \leqslant C \tag{14.35}$$

and the same bound holds for any fixed number of derivatives of ζ^k with respect to the slow time $\tau = \varepsilon t$.

Proof According to the analysis stated above and by induction, we can prove that the iterates $a\zeta^{(n)}$, $b\zeta^{(n)}$ and their derivatives with respect to τ are bounded in H^s for $0 \leqslant \tau \leqslant 1$ and $n \leqslant 4N$. These bounds show that $c\zeta^{(n)} = a\zeta^{(n)} + b\zeta^{(n)}$ is bounded in H^s, and then the bound (14.35) follows. □

Theorem 14.7 (Bounds of the Expansion) *The expansion (14.15) is bounded by*

$$\|\tilde{q}(t)\|_{s+1} + \|\tilde{p}(t)\|_s \leqslant C\varepsilon \quad for \quad 0 \leqslant t \leqslant \varepsilon^{-1}. \tag{14.36}$$

For $|j| \leqslant M$, it further holds that

$$\tilde{q}_j(t) = \zeta_j^{\langle j\rangle}(\varepsilon t)\mathrm{e}^{\mathrm{i}\omega_j t} + \zeta_j^{-\langle j\rangle}(\varepsilon t)\mathrm{e}^{-\mathrm{i}\omega_j t} + r_j, \quad \textit{where} \quad \|r\|_{s+1} \leqslant C\varepsilon^2. \tag{14.37}$$

If the condition (14.12) is not satisfied, then the bound becomes $\|r\|_{s+1} \leqslant C\varepsilon^{3/2}$.

Proof The following bounds for the $(4N)$-th iterates can be obtained

$$\begin{aligned} |||a\zeta(0)|||_s \leqslant C,\ \ |||\Omega a\dot{\zeta}(\tau)|||_s \leqslant C\varepsilon^{1/2}, \\ |||\Psi^{-1}a\dot{\zeta}(\tau)|||_s \leqslant C,\ \ |||\Psi^{-1}\Omega b\zeta(\tau)|||_s \leqslant C, \end{aligned} \tag{14.38}$$

where C depends on N, but not on ε, h, M. It then follows from (14.38) that

$$|||a\dot{\zeta}|||_{s+1} = |||\Omega a\dot{\zeta}|||_s \leqslant C\varepsilon^{1/2},$$

$$\begin{aligned} |||b\zeta|||_{s+1}^2 &= \sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s+2}\left|b\zeta_j\right|^2 = \sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s}\frac{\omega_j^2}{(\omega_j+|k\cdot\boldsymbol{\omega}|)^2}\left|(\omega_j+|k\cdot\boldsymbol{\omega}|)b\zeta_j\right|^2 \\ &\leqslant |||\Omega b\zeta(\tau)|||_s^2 \leqslant C. \end{aligned}$$

We thus obtain

$$|||c\zeta(\tau) - a\zeta(0)|||_{s+1} = |||a\zeta(\tau) + b\zeta(\tau) - a\zeta(0)|||_{s+1} \leqslant |||a\dot{\zeta}|||_{s+1} + |||b\zeta|||_{s+1} \leqslant C.$$

On noticing the fact that $\zeta_j^k = \dfrac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{|k|}}(c\zeta_j^k - a\zeta_j^k(0) + a\zeta_j^k(0))$, we have

$$\begin{aligned} |||\tilde{q}|||_{s+1}^2 &= \sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s+2}\left|\sum_{\|k\|\leqslant 2N}\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t}\zeta_j^k\right|^2 \\ &\leqslant \sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s+2}\left[\frac{\varepsilon}{\omega_j}\left(\left|a\zeta_j^{\langle j\rangle}(0)\right| + \left|a\zeta_j^{-\langle j\rangle}(0)\right|\right) + \sum_{\|k\|\leqslant 2N}\frac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{|k|}}\left|c\zeta_j^k - a\zeta_j^k(0)\right|\right]^2 \\ &\leqslant 2\varepsilon^2\sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s}\left(\left|a\zeta_j^{\langle j\rangle}(0)\right| + \left|a\zeta_j^{-\langle j\rangle}(0)\right|\right)^2 \\ &\quad + 2\sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s+2}\left(\sum_{\|k\|\leqslant 2N}\frac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{|k|}}\left|c\zeta_j^k - a\zeta_j^k(0)\right|\right)^2 \\ &\leqslant 4\varepsilon^2|||a\zeta(0)|||_s^2 + 2\sum_{k\in\mathscr{K}}{\sum_{|j|\leqslant M}}'' \omega_j^{2s+2}\left(\sum_{\|k\|\leqslant 2N}\frac{\varepsilon^{2[[k]]}}{\boldsymbol{\omega}^{2|k|}}\right)\left(\sum_{\|k\|\leqslant 2N}\left|c\zeta_j^k - a\zeta_j^k(0)\right|^2\right) \\ &\leqslant 4\varepsilon^2|||a\zeta(0)|||_s^2 + 2C_{K,1}\varepsilon^2|||c\zeta - a\zeta(0)|||_{s+1}^2 \leqslant C\varepsilon^2. \end{aligned}$$

According to (14.26), with a similar analysis, it can be proved that $|||\tilde{p}|||_s \leqslant C\varepsilon$. Hence, the bound (14.36) holds.

It then follows from (14.30) and (14.33) that $\Big(\sum_{\|k\|=1} \big\|(\Psi^{-1}\Omega b\zeta)^k\big\|_s^2\Big)^{1/2} \leqslant C\varepsilon$ for $b\zeta = (b\zeta)^{(4N)}$. Furthermore, using (14.12), we obtain that

$$\sum_{|j|\leqslant M} \sum_{j_1+j_2=j} \sum_{k=\pm\langle j_1\rangle\pm\langle j_2\rangle} \omega_j^{2(s+1)} |b\zeta_j^k|^2 \leqslant C\varepsilon.$$

These bounds as well as (14.38) lead to (14.37). The proof is complete. □

Concerning the alternative scaling (14.31), we can obtain the same bounds

$$|||\hat{a}\zeta(0)|||_1 \leqslant C, \quad |||\Omega\hat{a}\dot{\zeta}(\tau)|||_1 \leqslant C\varepsilon^{1/2}, \quad |||\Psi^{-1}\Omega\hat{b}\zeta(\tau)|||_1 \leqslant C. \tag{14.39}$$

Moreover, the following bound is also true for this scaling:

$$\Big(\sum_{\|k\|=1} \Big\|(\Psi^{-1}\Omega\hat{b}\zeta)^k\Big\|_1^2\Big)^{1/2} \leqslant C\varepsilon. \tag{14.40}$$

14.4.7 Defects

In this subsection, we pay attention to the so-called defect. It follows from (14.5) that the defect can be put in another form

$$\begin{aligned}\delta_j(t) = {} & \frac{\tilde{q}_j(t+h) - 2\cos(h\omega_j)\tilde{q}_j(t) + \tilde{q}_j(t-h)}{h^2\phi_2(h^2\omega_j^2)} \\ & - \Big[\int_0^1 f_j((1-\sigma)\tilde{q}_h(t) + \sigma\tilde{q}_h(t+h))\mathrm{d}\sigma + \int_0^1 f_j((1-\sigma)\tilde{q}_h(t-h) + \sigma\tilde{q}_h(t))\mathrm{d}\sigma\Big],\end{aligned} \tag{14.41}$$

where $\tilde{q}_j$ is determined in (14.15) with $\zeta_j^k = (\zeta_j^k)^{(4N)}$ obtained after $4N$ iterations of the procedure in Sect. 14.4.3. Here, $\delta_j(t)$ can also be rewritten as

$$\delta_j(t) = \sum_{\|k\|\leqslant NK} \mathrm{d}^k(\varepsilon t)\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t} + R(t),$$

where

$$d_j^k = \frac{1}{h^2\phi_2(h^2\omega_j^2)}\tilde{L}_j^k\zeta_j^k + \sum_{m=2}^{N}\frac{g^{(m)}(0)}{m!}\sum_{k^1+\cdots+k^m=k}\ \sum_{j_1+\cdots+j_m\equiv j \bmod 2M}{}' \cdot \int_0^1 \left[\left(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\right)(t\varepsilon,\sigma)\right]\mathrm{d}\sigma. \tag{14.42}$$

It is remarked that we consider $\|k\| \leqslant NK$ for d_j^k, and assume that $\zeta_j^k = \eta_j^k = 0$ for $\|k\| > K := 2N$. We denote by $\tilde{L}_j^k$ the truncation of the operator L_j^k after the ε^N term. The remainder terms of the Taylor expansion of f after N terms are absorbed in $R(t)$. Then it can be confirmed by the bound (14.36) and the estimates (14.38) that

$$\|R(t)\|_{s+1} \leqslant C\varepsilon^{N+1}.$$

Furthermore, using the Cauchy–Schwarz inequality and Lemma 14.1 results in

$$\begin{aligned}
&\left\|\sum_{\|k\|\leqslant NK} d^k(\varepsilon t)\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t}\right\|_s^2 = \sum_{|j|\leqslant M}{}'' \omega_j^{2s}\left|\sum_{\|k\|\leqslant NK} d_j^k\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t}\right|^2 \\
&= \sum_{|j|\leqslant M}{}'' \omega_j^{2s}\left|\sum_{\|k\|\leqslant NK}\boldsymbol{\omega}^{-|k|}(\boldsymbol{\omega}^{|k|}d_j^k\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\omega})t})\right|^2 \\
&\leqslant \sum_{|j|\leqslant M}{}'' \omega_j^{2s}\left(\sum_{\|k\|\leqslant NK}\boldsymbol{\omega}^{-2|k|}\right)\left(\sum_{\|k\|\leqslant NK}(\boldsymbol{\omega}^{|k|}d_j^k)^2\right) \\
&\leqslant C_{NK,1}\sum_{\|k\|\leqslant NK}\left\|\boldsymbol{\omega}^{|k|}d^k(\varepsilon t)\right\|_s^2.
\end{aligned}$$

This result leads to bounds on the defects. In fact, the right-hand side of this inequality can be estimated as follows.

Theorem 14.8 (Bounds of the Defects) *It can be deduced that* $\sum_{\|k\|\leqslant NK}\left\|\boldsymbol{\omega}^{|k|}d^k(\varepsilon t)\right\|_s^2 \leqslant C\varepsilon^{2(N+1)}$, *and then the defect* (14.41) *implies the bound* $\|\delta(t)\|_s \leqslant C\varepsilon^{N+1}$.

Proof To prove this result we will consider three different cases: truncated, near-resonant and non-resonant modes.

- **Truncated and near-resonant modes.** The result for these two cases can be obtained by using the similar analysis given in Sect. 6.8 of [1].

- **Non-resonant mode.** For the non-resonant mode ($\|k\| > K$ and (j,k) satisfies (14.8)), we first reformulate the defect in the scaled variables of Sect. 14.4.4 as

$$\boldsymbol{\omega}^{|k|} d_j^k = \varepsilon^{[[k]]} \left(\frac{1}{h^2 \phi_2(h^2\omega_j^2)} \tilde{L}_j^k c\zeta_j^k + f_j^k(c\xi) \right).$$

Then splitting them into $k = \pm\langle j \rangle$ and $k \neq \pm\langle j \rangle$ yields

$$\begin{aligned}
\omega_j d_j^{\pm\langle j\rangle} &= \varepsilon \left(\pm \mathrm{i}\varepsilon\omega_j \frac{\operatorname{sinc}(h\omega_j/2)}{\phi_2(h^2\omega_j^2)} \big(a\dot{\zeta}_j^{\pm\langle j\rangle} + (Aa\zeta)_j^{\pm\langle j\rangle} \big) + f_j^{\pm\langle j\rangle}(c\xi) \right), \\
\boldsymbol{\omega}^{|k|} d_j^k &= \varepsilon^{[[k]]} \left(\frac{2 s_{\langle j\rangle + k} s_{\langle j\rangle - k}}{h^2 c_k \phi_2(h^2\omega_j^2)} \big(b\zeta_j^k + (Bb\zeta)_j^k \big) + f_j^k(c\xi) \right).
\end{aligned}$$

We remark that the functions here are actually the $4N$-th iterates of the iteration in Sect. 14.4.3. Expressing $f_j^{\pm\langle j\rangle}$ and f_j^k in terms of F, G and inserting them from (14.33) into this defect, we obtain

$$\begin{aligned}
&\omega_j d_j^{\pm\langle j\rangle} = 2\omega_j \alpha_j^{\pm\langle j\rangle} \big([a\dot{\zeta}_j^{\pm\langle j\rangle}]^{(4N)} - [a\dot{\zeta}_j^{\pm\langle j\rangle}]^{(4N+1)} \big), \quad \alpha_j^{\pm\langle j\rangle} = \pm \mathrm{i}\varepsilon^2 \frac{\operatorname{sinc}(h\omega_j/2)}{2\phi_2(h^2\omega_j^2)}, \\
&\boldsymbol{\omega}^{|k|} d_j^k = \beta_j^k \big([b\zeta_j^k]^{(4N)} - [b\zeta_j^k]^{(4N+1)} \big), \qquad \beta_j^k = \varepsilon^{[[k]]} \frac{2 s_{\langle j\rangle + k} s_{\langle j\rangle - k}}{h^2 c_k \phi_2(h^2\omega_j^2)}.
\end{aligned}$$

Looking closer at these expressions, we introduce new variables as follows:

$$\tilde{a}\zeta_j^{\pm\langle j\rangle} = \alpha_j^{\pm\langle j\rangle} a\zeta_j^{\pm\langle j\rangle}, \quad \tilde{b}\zeta_j^k = \beta_j^k b\zeta_j^k$$

and then rewrite the iteration (14.33) in these variables as

$$\begin{aligned}
\tilde{a}\dot{\zeta}^{(n+1)} &= \Omega^{-1} \tilde{F}(\tilde{a}\zeta^{(n)}, \tilde{b}\zeta^{(n)}) - A\tilde{a}\zeta^{(n)}, \\
\tilde{b}\zeta^{(n+1)} &= \tilde{G}(\tilde{a}\zeta^{(n)}, \tilde{b}\zeta^{(n)}) - B\tilde{b}\zeta^{(n)}.
\end{aligned}$$

In such a way, the transformed functions are determined by

$$\begin{aligned}
\tilde{F}_j^{\pm\langle j\rangle}(\tilde{a}\zeta, \tilde{b}\zeta) &= \alpha_j^{\pm\langle j\rangle} F_j^{\pm\langle j\rangle}(\alpha^{-1}\tilde{a}\zeta, \beta^{-1}\tilde{b}\zeta) = -\varepsilon f_j^{\pm\langle j\rangle}(\alpha^{-1}\tilde{a}\zeta + \beta^{-1}\tilde{b}\zeta), \\
\tilde{G}_j^k(\tilde{a}\zeta, \tilde{b}\zeta) &= \beta_j^k (\Psi\Omega^{-1}G)_j^k(\alpha^{-1}\tilde{a}\zeta, \beta^{-1}\tilde{b}\zeta) = -\varepsilon^{[[k]]} f_j^k(\alpha^{-1}\tilde{a}\zeta + \beta^{-1}\tilde{b}\zeta).
\end{aligned}$$

As for the initial values of the iteration, we have

$$\tilde{a}\zeta^{(n+1)}(0) = \alpha v + \tilde{P}\tilde{b}\zeta^{(n)}(0) + \tilde{Q}\tilde{b}\zeta^{(n)}(0),$$

where $\tilde{P} = \alpha P \beta^{-1}$, $\tilde{Q} = \alpha Q \beta^{-1}$. For the bound of $\tilde{P}$, we obtain

$$
\begin{aligned}
&|||\tilde{P}\tilde{b}\zeta(0)|||_s^2 \\
&= \sum_{k\in\mathscr{K}} {\sum_{|j|\leqslant M}}'' \omega_j^{2s} \left| \mathrm{i}\varepsilon^2 \frac{\mathrm{sinc}(h\omega_j/2)}{2\phi_2(h^2\omega_j^2)} \frac{1}{2}\frac{\omega_j}{\varepsilon} \sum_{k\neq\pm\langle j\rangle} \frac{h^2 c_k \phi_2(h^2\omega_j^2)}{\varepsilon^{[[k]]} 2 s_{\langle j\rangle+k} s_{\langle j\rangle-k}} \frac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{|k|}} \tilde{b}\zeta_j^k(0) \right|^2 \\
&\leqslant \frac{\varepsilon^2 h^4}{64} \sum_{k\in\mathscr{K}} {\sum_{|j|\leqslant M}}'' \omega_j^{2s} \left(\sum_{k\neq\pm\langle j\rangle} \frac{\omega_j}{\left|s_{\langle j\rangle+k} s_{\langle j\rangle-k}\right|} \boldsymbol{\omega}^{-|k|} \tilde{b}\zeta_j^k(0) \right)^2 \\
&\leqslant \frac{\varepsilon^2 h^4}{64} \sum_{k\in\mathscr{K}} {\sum_{|j|\leqslant M}}'' \omega_j^{2s} \left(\sum_{k\neq\pm\langle j\rangle} \frac{1}{\varepsilon^{1/2} h^2} \boldsymbol{\omega}^{-|k|} \tilde{b}\zeta_j^k(0) \right)^2 \\
&\leqslant \frac{\varepsilon}{64} \sum_{k\in\mathscr{K}} {\sum_{|j|\leqslant M}}'' \omega_j^{2s} \left(\sum_{k\neq\pm\langle j\rangle} \boldsymbol{\omega}^{-2|k|} \sum_{k\neq\pm\langle j\rangle} (\tilde{b}\zeta_j^k(0))^2 \right) \leqslant C\varepsilon |||\tilde{b}\zeta(0)|||_s^2.
\end{aligned}
$$

In a similar way, the following result can be achieved:

$$
|||\tilde{Q}\tilde{b}\zeta(0)|||_s^2 \leqslant C\varepsilon |||\tilde{b}\zeta(0)|||_s^2.
$$

Clearly, it can be verified that in an H^s-neighbourhood of 0 where the bounds (14.38) hold, the partial derivatives of $\tilde{F}$ with respect to $\tilde{a}\zeta$ and $\tilde{b}\zeta$ are bounded by $\mathscr{O}(\varepsilon^{1/2})$. Moreover, the partial derivative of $\tilde{G}$ with respect to $\tilde{b}\zeta$ is bounded by $\mathscr{O}(\varepsilon^{1/2})$ but that of $\tilde{G}$ with respect to $\tilde{a}\zeta$ is only $\mathscr{O}(1)$. In fact, these results are the same as those described in Sect. 6.9 of [1]. Similarly, we can obtain

$$
\begin{aligned}
&|||\Omega(\tilde{a}\dot{\zeta}^{(4N+1)} - \tilde{a}\dot{\zeta}^{(4N)})|||_s \leqslant C\varepsilon^{N+2}, \\
&|||\tilde{b}\zeta^{(4N+1)} - \tilde{b}\zeta^{(4N)})|||_s \leqslant C\varepsilon^{N+2}, \\
&|||\tilde{a}\zeta(0)^{(4N+1)} - \tilde{a}\zeta(0)^{(4N)})|||_s \leqslant C\varepsilon^{N+2}.
\end{aligned}
$$

Hence, for $\tau \leqslant 1$ and $(j,k) \in \mathscr{R}_{\varepsilon,h}$, these results yield the bound

$$
\left(\sum_{\|k\|\leqslant K} \left\| \boldsymbol{\omega}^{|k|} d^k(\tau) \right\|_s^2 \right)^{1/2} \leqslant C\varepsilon^{N+1}. \tag{14.43}
$$

It then follows from (14.43) that the defect (14.41) has the bound $\|\delta(t)\|_s \leqslant C\varepsilon^{N+1}$ for $t \leqslant \varepsilon^{-1}$. Concerning the defect in the initial conditions (14.24) and (14.25), it is true that

$$
\|q(0) - \tilde{q}(0)\|_{s+1} + \|p(0) - \tilde{p}(0)\|_s \leqslant C\varepsilon^{N+1}.
$$

Finally, we turn to the alternative scaling (14.31). For this case, we can obtain

$$\Big(\sum_{\|k\|\leqslant K}\Big\|\boldsymbol{\omega}^{s|k|}d^k(\tau)\Big\|_1^2\Big)^{1/2}\leqslant C\varepsilon^{N+1}. \tag{14.44}$$

The proof is complete. □

14.4.8 Remainders

In this subsection, we are concerned with the difference between the numerical solution and its modulated Fourier expansion.

Theorem 14.9 (Remainders) *The bound on the difference between the numerical solution and its modulated Fourier expansion satisfies*

$$\|q_n-\tilde{q}(t)\|_{s+1}+\|p_n-\tilde{p}(t)\|_s\leqslant C\varepsilon^N\quad for\quad 0\leqslant t=nh\leqslant\varepsilon^{-1}. \tag{14.45}$$

Proof Let $\Delta q_n=\tilde{q}(t_n)-q_n,\ \Delta p_n=\tilde{p}(t_n)-p_n$. We have

$$\begin{pmatrix}\Delta q_{n+1}\\ \Omega^{-1}\Delta p_{n+1}\end{pmatrix}=\begin{pmatrix}\cos(h\Omega) & \sin(h\Omega)\\ -\sin(h\Omega) & \cos(h\Omega)\end{pmatrix}\begin{pmatrix}\Delta q_n\\ \Omega^{-1}\Delta p_n\end{pmatrix}+h\begin{pmatrix}h\Omega\phi_2(V)\Omega^{-1}(\Delta f+\delta)\\ \phi_1(V)\Omega^{-1}(\Delta f+\delta)\end{pmatrix},$$

where

$$\Delta f=\int_0^1\big(f((1-\sigma)q_n+\sigma q_{n+1})-f((1-\sigma)\tilde{q}(t_n)+\sigma\tilde{q}(t_n+h))\big)\mathrm{d}\sigma.$$

According to the Lipschitz bound given in Sect. 4.2 of [3] and Sect. 6.10 of [1], it is clear that

$$\big\|\Omega^{-1}\Delta f\big\|_{s+1}=\|\Delta f\|_s\leqslant\varepsilon(\|\Delta q_n\|_s+\|\Delta q_{n+1}\|_s).$$

Moreover, we have $\big\|\Omega^{-1}\delta(t)\big\|_{s+1}=\|\delta(t)\|_s\leqslant C\varepsilon^{N+1}$. We then obtain

$$\left\|\begin{pmatrix}\Delta q_{n+1}\\ \Omega^{-1}\Delta p_{n+1}\end{pmatrix}\right\|_{s+1}\leqslant\left\|\begin{pmatrix}\Delta q_n\\ \Omega^{-1}\Delta p_n\end{pmatrix}\right\|_{s+1}+h\Big(C\varepsilon\,\|\Delta q_n\|_s+C\varepsilon\,\|\Delta q_{n+1}\|_s+C\varepsilon^{N+1}\Big).$$

This leads to $\|\Delta q_n\|_{s+1}+\big\|\Omega^{-1}\Delta p_n\big\|_{s+1}\leqslant C(1+t_n)\varepsilon^{N+1}$ for $t_n\leqslant\varepsilon^{-1}$. This proves (14.45). □

14.4.9 Almost Invariants

This subsection concerns almost-invariants of the modulated Fourier expansions.

According to the analysis presented in Sect. 14.4.7, we can rewrite the defect formula (14.42) as

$$\frac{1}{h^2\phi_2(h^2\omega_j^2)}\tilde{L}_j^k\zeta_j^k + \nabla_{-j}^{-k}\mathscr{U}(\xi(t)) = d_j^k, \tag{14.46}$$

where $\nabla_{-j}^{-k}\mathscr{U}(y)$ is the partial derivative with respect to y_{-j}^{-k} of the extended potential (see, e.g. [1, 3])

$$\mathscr{U}(\xi(t,\sigma)) = \sum_{l=-N}^{N}\mathscr{U}_l(\xi(t,\sigma)),$$

$$\mathscr{U}_l(\xi(t,\sigma)) = \sum_{m=2}^{N}\frac{U^{(m+1)}(0)}{(m+1)!}\sum_{k^1+\cdots+k^{m+1}=0}\ \sum_{j_1+\cdots+j_{m+1}=2Ml}{}'\int_0^1(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_{m+1}}^{k^{m+1}})(t,\sigma)\mathrm{d}\sigma.$$

We define (see [1])

$$S_{\boldsymbol{\mu}}(\theta)y = \left(\mathrm{e}^{\mathrm{i}(k\cdot\boldsymbol{\mu})\theta}y_j^k\right)_{|j|\leqslant M,\|k\|\leqslant K}$$

and

$$T(\theta)y = \left(\mathrm{e}^{\mathrm{i}j\theta}y_j^k\right)_{|j|\leqslant M,\|k\|\leqslant K},$$

where $\boldsymbol{\mu} = (\mu_l)_{l\geqslant 0}$ is an arbitrary real sequence for $\theta \in R$. Using the results given in [1], we obtain $\mathscr{U}(S_{\boldsymbol{\mu}}(\theta)y) = \mathscr{U}(y)$ and $\mathscr{U}_0(T(\theta)y) = \mathscr{U}_0(y)$ for $\theta \in \mathbb{R}$. Hence,

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\,|_{\theta=0}\,\mathscr{U}(S_{\boldsymbol{\mu}}(\theta)\xi(t,\sigma)),\quad 0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\,|_{\theta=0}\,\mathscr{U}_0(T(\theta)\xi(t,\sigma)). \tag{14.47}$$

Theorem 14.10 (Two Almost-Invariants) *There exist two functions* $\mathscr{J}_l[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)$ *and* $\mathscr{K}[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)$ *such that*

$$\sum_{l=1}^{M}\omega_l^{2s+1}\left|\frac{\mathrm{d}}{\mathrm{d}\tau}\mathscr{J}_l[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)\right| \leqslant C\varepsilon^{N+1},$$

$$\left|\frac{\mathrm{d}}{\mathrm{d}\tau}\mathscr{K}[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)\right| \leqslant C(\varepsilon^{N+1} + \varepsilon^2 M^{-s+1}) \tag{14.48}$$

for $\tau \leqslant 1$. *Moreover, it is true that*

$$\begin{aligned}\mathscr{J}_l[\boldsymbol{\zeta},\boldsymbol{\eta}](\varepsilon t_n) &= \hat{J}_l(p_n,q_n) + \gamma_l(t_n)\varepsilon^3,\\ \mathscr{K}[\boldsymbol{\zeta},\boldsymbol{\eta}](\varepsilon t_n) &= \hat{K}(p_n,q_n) + \mathscr{O}(\varepsilon^3) + \mathscr{O}(\varepsilon^2 M^{-s}),\end{aligned} \tag{14.49}$$

where

$$\hat{J}_l = \hat{I}_l + \hat{I}_{-l} = 2\hat{I}_l \quad \textit{for} \quad 0 < l < M, \quad \hat{J}_0 = \hat{I}_0, \quad \hat{J}_M = \hat{I}_M.$$

Here, all the constants in (14.48) *and* (14.49) *are independent of* ε, M, h, *and* n, *and* $\sum\limits_{l=0}^{M} \omega_l^{2s+1}\gamma_l(t_n) \leqslant C$ *for* $t_n \leqslant \varepsilon^{-1}$.

Proof

- **Proof of** (14.48)**.**

  It follows from the first equality of (14.47) that

$$\begin{aligned}0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\,|_{\theta=0}\,\mathscr{U}(S_{\boldsymbol{\mu}}(\theta)\xi(t,\sigma)) &= \sum_{\|k\|\leqslant K}\sum_{|j|\leqslant M}{}' \mathrm{i}(k\cdot\boldsymbol{\mu})\xi_{-j}^{-k}(t,\sigma)\nabla_{-j}^{-k}\mathscr{U}(\xi(t,\sigma))\\ &= \sum_{\|k\|\leqslant K}\sum_{|j|\leqslant M}{}' \mathrm{i}(k\cdot\boldsymbol{\mu})L_4^{-k}(\sigma)\zeta_{-j}^{-k}\\ &\quad\times\left(\frac{1}{h^2\phi_2(h^2\omega_j^2)}\tilde{L}_j^k\zeta_j^k - d_j^k\right).\end{aligned} \tag{14.50}$$

It is noted that the right-hand side is independent of σ. We thus choose $\sigma = 1/2$ in the following analysis. In this case, (14.50) gives

$$\begin{aligned}&\sum_{\|k\|\leqslant K}\sum_{|j|\leqslant M}{}' \mathrm{i}(k\cdot\boldsymbol{\mu})L_4^{-k}\left(\frac{1}{2}\right)\zeta_{-j}^{-k}\frac{1}{h^2\phi_2(h^2\omega_j^2)}\tilde{L}_j^k\zeta_j^k\\ =&\sum_{\|k\|\leqslant K}\sum_{|j|\leqslant M}{}' \mathrm{i}(k\cdot\boldsymbol{\mu})L_4^{-k}\left(\frac{1}{2}\right)\zeta_{-j}^{-k}d_j^k.\end{aligned} \tag{14.51}$$

It then follows from the expansions of $L_4^{-k}\left(\frac{1}{2}\right)$ and $\tilde{L}_j^k$ and the "magic formulas" on p. 508 of [20] that the left-hand side of (14.51) is a total derivative of function $\varepsilon\mathscr{J}_{\boldsymbol{\mu}}[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)$ which depends on $\boldsymbol{\zeta}(\tau), \boldsymbol{\eta}(\tau)$ and their up to $(N-1)$th order

derivatives. This implies that (14.51) is identical to the following equation

$$-\varepsilon\frac{\mathrm{d}}{\mathrm{d}\tau}\mathscr{J}_{\boldsymbol{\mu}}[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)=\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant M}}'\mathrm{i}(k\cdot\boldsymbol{\mu})L_4^{-k}\left(\frac{1}{2}\right)\zeta_{-j}^{-k}d_j^k.$$

In what follows, we consider the special case where $\boldsymbol{\mu}=\langle l\rangle$. Let $z_j^k=L_4^k(1/2)\zeta_j^k$. It follows from the property of $L_4^k(1/2)$ that the bounds on z_j^k and ζ_j^k are of the same magnitude. Splitting $d=ad+bd$ into two parts: the diagonal ($k=\pm\langle j\rangle$) and nondiagonal ($k\neq\pm\langle j\rangle$), gives

$$|||ad|||_s^2+\sum_{\|k\|\leqslant K}|||\boldsymbol{\omega}^{s|k|}bd|||_0^2=\sum_{\|k\|\leqslant K}\left\|\boldsymbol{\omega}^{s|k|}d^k\right\|_0^2\leqslant C\varepsilon^{2N+2},$$

where (14.44) is used. Using Lemma 3 of [2] and the facts that

- $z_j^k=\dfrac{\varepsilon}{\omega_j^s}\hat{a}z_j^k+\dfrac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{s|k|}}\hat{a}z_j^k$,
- $|||\hat{a}z^k|||_1\leqslant C$,
- $|||\Omega\hat{b}z^k|||_1\leqslant$ from (14.39),

we obtain

$$\begin{aligned}&\sum_{l=1}^M\omega_l^{2s+1}\left|\frac{\mathrm{d}}{\mathrm{d}\tau}\mathscr{J}_l[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)\right|=\frac{1}{\varepsilon}\sum_{l=1}^M\omega_l^{2s+1}\left|\sum_{\|k\|\leqslant K}k_l\sum_{j=-\infty}^{\infty}\zeta_j^kd_j^k\right|\\&\leqslant\frac{1}{\varepsilon}\Bigg[|||\frac{\varepsilon}{\omega_j^s}\hat{a}\zeta_j^k|||_{s+1}|||ad|||_s+\Bigg(\sum_{\|k\|\leqslant K}\left\|\boldsymbol{\omega}^{s|k|}(1+|k\cdot\boldsymbol{\omega}|)\frac{\varepsilon^{[[k]]}}{\boldsymbol{\omega}^{s|k|}}\hat{a}\zeta_j^k\right\|_0^2\Bigg)^{1/2}\cdot\\&\Bigg(\sum_{\|k\|\leqslant K}\left\|\boldsymbol{\omega}^{s|k|}bd^k\right\|_0^2\Bigg)^{1/2}\Bigg]\\&\leqslant C\varepsilon^{N+1}.\end{aligned}$$

The first statement of (14.48) is proved.

In a similar way, using the second equality of (14.47), we obtain

$$\begin{aligned}&\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant M}}'\mathrm{i}jL_4^{-k}\left(\frac{1}{2}\right)\zeta_{-j}^{-k}\frac{1}{h^2\phi_2(h^2\omega_j^2)}\tilde{L}_j^k\zeta_j^k\\&=\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant M}}'\mathrm{i}jL_4^{-k}\left(\frac{1}{2}\right)\zeta_{-j}^{-k}\left(d_j^k-\sum_{l\neq0}\nabla_{-j}^k\left(\mathscr{U}_l(\xi(t,\sigma))\right)\right).\end{aligned}\tag{14.52}$$

A careful analysis shows that the left-hand side of (14.52) can be written as a total derivative of function $\varepsilon\mathscr{K}[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)$, which yields

$$-\varepsilon\frac{\mathrm{d}}{\mathrm{d}\tau}\mathscr{K}[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau)=\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant M}}'\mathrm{i}jL_4^{-k}\big(\frac{1}{2}\big)\zeta_{-j}^{-k}\Big(d_j^k-\sum_{l\neq 0}\nabla_{-j}^{-k}\big(\mathscr{U}_l(\xi(t,\sigma))\big)\Big). \tag{14.53}$$

It follows from the Cauchy–Schwarz inequality and the bound $|j|\leqslant\omega_j$ that

$$\begin{aligned}\left|\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant K}}'\mathrm{i}jz_{-j}^{-k}d_j^k\right|&\leqslant\Big(\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant K}}'\omega_j^2\left|z_j^k\right|^2\Big)^{1/2}\Big(\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant K}}'\left|d_j^k\right|^2\Big)^{1/2}\\&\leqslant C\varepsilon\Big(\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant K}}'\frac{\omega_j^2}{\boldsymbol{\omega}^{|k|}}\frac{\varepsilon^{[[k]]}}{\varepsilon^2}\frac{\boldsymbol{\omega}^{|k|}}{\varepsilon^{[[k]]}}\left|z_j^k\right|^2\Big)^{1/2}\Big(\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant K}}'\left|d_j^k\right|^2\Big)^{1/2}\leqslant C\varepsilon^{N+2}.\end{aligned}$$

Furthermore, we note that

$$\begin{aligned}&\sum_{\|k\|\leqslant K}{\sum_{|j|\leqslant M}}'\mathrm{i}jz_{-j}^{-k}\nabla_{-j}^{-k}\mathscr{U}_l(\xi(t,\sigma))\\&=\sum_{m=2}^{N}\frac{U^{(m+1)}(0)}{m!}\sum_{k^1+\cdots+k^{m+1}=k}{\sum_{j_1+\cdots+j_{m+1}=2Ml}}'z_{j_1}^{k^1}\cdots z_{j_m}^{k^m}\cdot\mathrm{i}j_{m+1}z_{j_{m+1}}^{k^{m+1}},\end{aligned}$$

is the $2Ml$-th Fourier coefficient of the function (see [3])

$$w(x):=\sum_{m=2}^{N}\frac{U^{(m+1)}(0)}{m!}\sum_{k^1+\cdots+k^{m+1}=k}\mathscr{P}z^{k^1}(x)\cdots\mathscr{P}z^{k^m}(x)\cdot\frac{\mathrm{d}}{\mathrm{d}x}\mathscr{P}z^{k^{m+1}}(x).$$

We then can deduce that $\|w\|_{s-1}\leqslant C\varepsilon^3$, and the $2Ml$-th Fourier coefficient of w is bounded by $C\varepsilon^3\omega_{2Ml}^{-s+1}\leqslant C\varepsilon^3(2Ml)^{-s+1}$, as shown in the proof of Theorem 5.2 of [3]. In such a way, the second statement of (14.48) is confirmed by (14.53).

- **Proof of** (14.49)**.**

We will prove only the second statement of (14.49) since the first one can be dealt with in a similar way.

It follows from the AAVF formula that

$$2h\mathrm{sinc}(h\Omega)\tilde{p}(t)=\tilde{q}(t+h)-\tilde{q}(t-h)+\mathscr{O}(h^2).$$

This shows that

$$\tilde{p}_j(t) = \mathrm{i}\omega_j\left(\eta_j^{\langle j\rangle}(\varepsilon t)\mathrm{e}^{\mathrm{i}\omega_j t} - \eta_j^{-\langle j\rangle}(\varepsilon t)\mathrm{e}^{-\mathrm{i}\omega_j t}\right) + \mathscr{O}(h\varepsilon^2) + \mathscr{O}(h^3\varepsilon^2).$$

We then have

$$\zeta_j^{\langle j\rangle} = \frac{1}{2}\left(\tilde{q}_j + \frac{1}{\mathrm{i}\omega_j}\tilde{p}_j\right) + \mathscr{O}(\varepsilon^2)$$

and

$$\zeta_j^{-\langle j\rangle} = \frac{1}{2}\left(\tilde{q}_j - \frac{1}{\mathrm{i}\omega_j}\tilde{p}_j\right) + \mathscr{O}(\varepsilon^2).$$

On the basis of these results, an analysis of $\mathscr{K}$ is presented below:

$$\mathscr{K}[\boldsymbol{\zeta},\boldsymbol{\eta}](\tau) = {\sum_{|j|\leqslant M}}' j\frac{1}{2}\frac{4\varepsilon h\sin\left(\frac{1}{2}h\omega_j\right)\cos\left(\frac{1}{2}h\omega_j\right)}{2h^2\phi_2(h^2\omega_j^2)}\left(|\zeta_j^{\langle j\rangle}|^2 - |\zeta_j^{-\langle j\rangle}|^2\right) + \mathscr{O}(\varepsilon^3)$$

$$= {\sum_{|j|\leqslant M}}' j\omega_j\frac{\cos\left(\frac{1}{2}h\omega_j\right)}{\operatorname{sinc}\left(\frac{1}{2}h\omega_j\right)}\left(|\zeta_j^{\langle j\rangle}|^2 - |\zeta_j^{-\langle j\rangle}|^2\right) + \mathscr{O}(\varepsilon^3)$$

$$= {\sum_{|j|\leqslant M}}' \frac{j\omega_j}{4}\frac{\cos\left(\frac{1}{2}h\omega_j\right)}{\operatorname{sinc}\left(\frac{1}{2}h\omega_j\right)}\left(|\tilde{q}_j + \frac{1}{\mathrm{i}\omega_j}\tilde{p}_j|^2 - |\tilde{q}_j - \frac{1}{\mathrm{i}\omega_j}\tilde{p}_j|^2\right) + \mathscr{O}(\varepsilon^3)$$

$$= {\sum_{|j|\leqslant M}}' \frac{\cos\left(\frac{1}{2}h\omega_j\right)}{\operatorname{sinc}\left(\frac{1}{2}h\omega_j\right)}\frac{j\omega_j}{4}4\frac{1}{\mathrm{i}\omega_j}\tilde{q}_{-j}\tilde{p}_j + \mathscr{O}(\varepsilon^3)$$

$$= \hat{K}(\tilde{p},\tilde{q}) + \mathscr{O}(\varepsilon^3) + \mathscr{O}(\varepsilon^2 M^{-s}) = \hat{K}(p_n,q_n) + \mathscr{O}(\varepsilon^3) + \mathscr{O}(\varepsilon^2 M^{-s}),$$

where the results (14.37) and (14.45) are used. □

14.4.10 From Short to Long-Time Intervals

According to the analysis stated above in this chapter, the statement of Theorem 14.3 can be confirmed by patching together many intervals of length ε^{-1} in the same way as that used in [1, 2].

14.5 Analysis for the AAVF Method with a Quadrature Rule

The previous analysis was made for the AAVF method with the integral appearing in (14.5), which usually cannot be solved exactly. Normally a quadrature rule is required. For this reason, we will show that the main result for the AAVF method with the integral is still true for the AAVF method with a quadrature approximation instead of the integral.

As an example, we consider the following AAVF method with the midpoint rule

$$\begin{cases} q_{n+1} = \phi_0(V)q_n + h\phi_1(V)p_n + h^2\phi_2(V)f((q_n+q_{n+1})/2), \\ p_{n+1} = -h\Omega^2\phi_1(V)q_n + \phi_0(V)p_n + h\phi_1(V)f((q_n+q_{n+1})/2). \end{cases} \tag{14.54}$$

The main result presented in Theorem 14.3 can be adapted for this method with the following modifications for the operator and the nonlinearity. We next present only the main differences and omit the details for brevity.

- Modifications for Sect. 14.4.2.
  Since the term $\int_0^1 f((1-\sigma)q_n+\sigma q_{n+1})\mathrm{d}\sigma$ is replaced by $f((q_n+q_{n+1})/2)$, the function $\xi^k\left(\varepsilon\left(t+\frac{h}{2}\right),\sigma\right)$ should be changed to $\xi^k\left(\varepsilon\left(t+\frac{h}{2}\right),1/2\right)$ and the operator $L_4^k(\sigma)$ is replaced by $L_4^k(1/2)$. Then all the analyses and results in Sect. 14.4.2 still hold for (14.54).
- Modifications for Sect. 14.4.3.
  For this part, we only need to change $\int_0^1\left[\left(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\right)(t\varepsilon,\sigma)\right]\mathrm{d}\sigma$ to $\left(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\right)(t\varepsilon,1/2)$.
- Modifications for Sect. 14.4.4.
  One part of the function $f_j^k\left(c\xi(\tau)\right)$ here is $\left(c\xi_{j_1}^{k^1}\cdot\cdots\cdot c\xi_{j_m}^{k^m}\right)(\tau,1/2)$ instead of $\int_0^1\left(c\xi_{j_1}^{k^1}\cdot\cdots\cdot c\xi_{j_m}^{k^m}\right)(\tau,\sigma)\mathrm{d}\sigma$ and then the property of $f_j^k\left(c\xi(\tau)\right)$ stated in Proposition 14.2 is still true.
- Modifications for Sect. 14.4.7.
  Since the defect expressed by (14.41) needs to be modified according to the scheme (14.54), the term $\int_0^1\left[\left(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\right)(t\varepsilon,\sigma)\right]\mathrm{d}\sigma$ appearing in (14.42) should be replaced by $\left(\xi_{j_1}^{k^1}\cdot\cdots\cdot\xi_{j_m}^{k^m}\right)(t\varepsilon,1/2)$. In this situation, we still obtain the same bounds of the defects as those stated previously.
- Modifications for Sect. 14.4.8.
  Here only the expression of Δf should be modified in the light of (14.54).
- Modifications for Sect. 14.4.9.

A new function

$$\mathscr{U}_l(\xi) = \sum_{m=2}^{N} \frac{U^{(m+1)}(0)}{(m+1)!} \sum_{k^1+\cdots+k^{m+1}=0} {\sum_{j_1+\cdots+j_{m+1}=2Ml}}' \left(\xi_{j_1}^{k^1} \cdot \cdots \cdot \xi_{j_{m+1}}^{k^{m+1}}\right)(t, 1/2)$$

will be used here instead of the previous one.

At the end of this section, we remark that since the AAVF method with the integral is of only order two, the long-time momentum and actions behaviour does not change for (14.54). For the AAVF method with other higher-order quadrature rules, the main result can also be obtained by following the same approach.

14.6 Conclusions and Discussions

It is known that the preservation of geometric or physical properties of the numerical flow can assist in long-time integration and produce improved qualitative behaviour in comparison with a general-purpose numerical method. In this chapter, we have investigated in detail the long-time behaviour of the AAVF method when applied to semilinear wave equations via spatial spectral semidiscretisations. With the semidiscretisation, the AAVF method exactly preserves the energy and nearly conserves modified actions and modified momentum over long times. The main result has been presented by developing a modulated Fourier expansion of the AAVF method and showing two almost-invariants of the modulated system.

The main result of this chapter explains rigorously the good long-time behaviour of EP methods for the numerical solution of semilinear wave equations. The analysis for multi-dimensional wave equations deserves further investigation. It is also noted that the long-term analysis of many different methods other than EP methods has been given recently for Schrödinger equations and the reader is referred to [18, 38–40]. The Schrödinger equation has become one of the most studied PDEs. It is hoped to obtain near-conservation of actions, momentum and density as well as exact-conservation of energy for some EP schemes when applied to the Schrödinger equation.

The material in this chapter is based on the work by Wang and Wu [41].

References

1. Cohen, D., Hairer, E., Lubich, C.: Conservation of energy, momentum and actions in numerical discretizations of nonlinear wave equations. Numer. Math. **110**, 113–143 (2008)
2. Cohen, D., Hairer, E., Lubich, C.: Long-time analysis of nonlinearly perturbed wave equations via modulated Fourier expansions. Arch. Ration. Mech. Anal. 187, 341–368 (2008)
3. Hairer, E., Lubich, C.: Spectral semi-discretisations of weakly nonlinear wave equations over long times. Found. Comput. Math. **8**, 319–334 (2008)

4. Bambusi, D.: Birkhoff normal form for some nonlinear PDEs. Commun. Math. Phys. **234**, 253–285 (2003)
5. Cano, B.: Conservation of invariants by symmetric multistep cosine methods for second-order partial differential equations. BIT Numer. Math. **53**, 29–56 (2013)
6. Cano, B.: Conserved quantities of some Hamiltonian wave equations after full discretization. Numer. Math. **103**, 197–223 (2006)
7. Cano, B., Moreta, M.J.: Multistep cosine methods for second-order partial differential systems. IMA J. Numer. Anal. **30**, 431–461 (2010)
8. Gauckler, L.: Error analysis of trigonometric integrators for semilinear wave equations. SIAM J. Numer. Anal. **53**, 1082–1106 (2015)
9. Gauckler, L., Weiss, D.: Metastable energy strata in numerical discretizations of weakly nonlinear wave equations. Disc. Contin. Dyn. Syst. **37**, 3721–3747 (2017)
10. Gauckler, L., Lu, J., Marzuola, J., et al.: Trigonometric integrators for quasilinear wave equations. Math. Comput. **88**, 717–749 (2019)
11. Grimm, V.: On the Use of the Gautschi-Type Exponential Integrator for Wave Equations Numerical Mathematics and Advanced Applications. Springer, Berlin (2006), pp. 557–563
12. Liu, C., Iserles, A., Wu, X.: Symmetric and arbitrarily high-order Birkhoff-Hermite time integrators and their long-time behaviour for solving nonlinear Klein-Gordon equations. J. Comput. Phys. **356**, 1–30 (2018)
13. Wang, B., Iserles, A., Wu, X.: Arbitrary-order trigonometric Fourier collocation methods for multi-frequency oscillatory systems. Found. Comput. Math. **16**, 151–181 (2016)
14. Wu, X., Wang, B.: Recent Developments in Structure-Preserving Algorithms for Oscillatory Differential Equations. Springer, Nature Singapore Pte Ltd., Singapore (2018)
15. Gauckler, L., Hairer, E., Lubich, C.: Long-term analysis of semilinear wave equations with slowly varying wave speed. Commun. Partial. Differ. Equ. **41**, 1934–1959 (2016)
16. Gauckler, L., Hairer, E., Lubich, C., et al.: Metastable energy strata in weakly nonlinear wave equations. Commun. Partial. Differ. Equ. **37**, 1391–1413 (2012)
17. Hairer, E., Lubich, C.: Long-time energy conservation of numerical methods for oscillatory differential equations. SIAM J. Numer. Anal. **38**, 414–441 (2000)
18. Cohen, D., Gauckler, L.: One-stage exponential integrators for nonlinear Schrödinger equations over long times. BIT Numer. Math. **52**, 877–903 (2012)
19. Hairer, E., Lubich, C.: Long-term analysis of the Störmer-Verlet method for Hamiltonian systems with a solution-dependent high frequency. Numer. Math. **134**, 119–138 (2016)
20. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer, Berlin (2006)
21. McLachlan, R.I., Stern, A.: Modified trigonometric integrators. SIAM J. Numer. Anal. **52**, 1378–1397 (2014)
22. Sanz-Serna, J.M.: Modulated Fourier expansions and heterogeneous multiscale methods. IMA J. Numer. Anal. **29**, 595–605 (2009)
23. Brugnano, L., Frasca Caccia, G., Iavernaro, F.: Energy conservation issues in the numerical solution of the semilinear wave equation. Appl. Math. Comput. **270**, 842–870 (2015)
24. Celledoni, E., Grimm, V., McLachlan, R.I., et al.: Preserving energy resp. dissipation in numerical PDEs using the "Average Vector Field" method. J. Comput. Phys. **231**, 6770–6789 (2012)
25. Li, Y.W., Wu, X.: General local energy-preserving integrators for solving multi-symplectic Hamiltonian PDEs. J. Comput. Phys. **301**, 141–166 (2015)
26. Liu, C., Wu, X.: An energy-preserving and symmetric scheme for nonlinear Hamiltonian wave equations. J. Math. Anal. Appl. **440**, 167–182 (2016)
27. Liu K, Wu X, Shi W. A linearly-fitted conservative (dissipative) scheme for efficiently solving conservative (dissipative) nonlinear wave PDEs. J. Comput. Math. **35**, 780–800 (2017)
28. Mei, L., Liu, C., Wu, X.: An essential extension of the finite-energy condition for extended Runge-Kutta-Nyström integrators when applied to nonlinear wave equations. Commun. Comput. Phys. **22**, 742–764 (2017)

29. Wang, B., Wu, X.: The formulation and analysis of energy-preserving schemes for solving high-dimensional nonlinear Klein-Gordon equations. IMA. J. Numer. Anal. **39**, 2016–2044 (2019)
30. Hairer, E.: Energy-preserving variant of collocation methods. J. Numer. Anal. Ind. Appl. Math. **5**, 73–84 (2010)
31. Li, Y.W., Wu, X.: Exponential integrators preserving first integrals or Lyapunov functions for conservative or dissipative systems. SIAM J. Sci. Comput. **38**, 1876–1895 (2016)
32. Quispel, G.R.W., McLaren, D.I.: A new class of energy-preserving numerical integration methods. J. Phys. A **41**, 045206 (2008)
33. Wang, B., Wu, X.: A new high precision energy preserving integrator for system of oscillatory second-order differential equations. Phys. Lett. A **376**, 1185–1190 (2012)
34. Wu, X., Wang, B., Shi, W.: Efficient energy preserving integrators for oscillatory Hamiltonian systems. J. Comput. Phys. **235**, 587–605 (2013)
35. Hairer, E., Lubich, C.: Long-term analysis of a variational integrator for charged-particle dynamics in a strong magnetic field. Numeri. Math. **144**, 699–728 (2020)
36. Wang, B., Wu, X.: Global error bounds of one-stage extended RKN integrators for semilinear wave equations. Numer. Algor. **81**, 1203–1218 (2019)
37. Feng, K., Qin, M.: The Symplectic Methods for the Computation of Hamiltonian Equations, Numerical Methods for Partial Differential Equations. Springer, Berlin (2006), pp. 1–37
38. Gauckler, L.: Numerical long-time energy conservation for the nonlinear Schrödinger equation. IMA J. Numer. Anal. **37**, 2067–2090 (2017)
39. Gauckler, L., Lubich, C.: Nonlinear Schrödinger equations and their spectral semi-discretizations over long times. Found. Comput. Math. **10**, 141–169 (2010)
40. Gauckler, L., Lubich, C.: Splitting integrators for nonlinear Schrödinger equations over long times. Found. Comput. Math. **10**, 275–302 (2010)
41. Wang, B., Wu, X.: Long-time momentum and actions behaviour of energy-preserving methods for semi-linear wave equations via spatial spectral semi-discretisations. Adv. Comput. Math. **45**, 2921–2952 (2019)