UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Article scientifique | Article | 2008

Accepted version | Open Access

This is an author manuscript post-peer-reviewing (accepted version) of the original publication. The layout of the published version may differ .

Conservation of energy, momentum and actions in numerical discretizations of non-linear wave equations

Cohen, David; Hairer, Ernst; Lubich, Christian

How to cite

COHEN, David, HAIRER, Ernst, LUBICH, Christian. Conservation of energy, momentum and actions in numerical discretizations of non-linear wave equations. In: Numerische Mathematik, 2008, vol. 110, p. 113–143. doi: 10.1007/s00211-008-0163-9

This publication URL: https://archive-ouverte.unige.ch//unige:5202
Publication DOI: 10.1007/s00211-008-0163-9

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

Numerische Mathematik manuscript No.
(will be inserted by the editor)

Conservation of energy, momentum and actions in numerical discretizations of nonlinear wave equations

David Cohen[1], Ernst Hairer[2], Christian Lubich[3]

[1] Mathematisches Institut, Univ. Basel, CH-4051 Basel, Switzerland,
e-mail: David.Cohen@unibas.ch
[2] Dept. de Mathématiques, Univ. de Genève, CH-1211 Genève 24, Switzerland,
e-mail: Ernst.Hairer@math.unige.ch
[3] Mathematisches Institut, Univ. Tübingen, D-72076 Tübingen, Germany,
e-mail: Lubich@na.uni-tuebingen.de

Received: April 2007 / Revised version: date

Summary For classes of symplectic and symmetric time-stepping methods — trigonometric integrators and the Störmer–Verlet or leap-frog method — applied to spectral semi-discretizations of semilinear wave equations in a weakly nonlinear setting, it is shown that energy, momentum, and all harmonic actions are approximately preserved over long times. For the case of interest where the CFL number is not a small parameter, such results are outside the reach of standard backward error analysis. Here, they are instead obtained via a modulated Fourier expansion in time.

1 Introduction

This paper is concerned with the long-time behaviour of symplectic integrators applied to Hamiltonian nonlinear partial differential equations, such as semilinear wave equations. For symplectic methods applied to Hamiltonian systems of *ordinary* differential equations, the numerically observed long-time near-conservation of the total energy, and of actions in near-integrable systems, can be rigorously proved with the help of backward error analysis. This interprets a step of a symplectic method as the exact flow of a modified Hamiltonian system, up to an error which in the case of an analytic Hamiltonian is exponentially small in $1/(h\omega)$, where h is the small step

size and ω represents the largest frequency in a local linearization of the system; see Benettin & Giorgilli [2], Hairer & Lubich [11], Reich [16], and Chapter IX in Hairer, Lubich & Wanner [14]. When the symplectic method is applied to a semi-discretization of a partial differential equation, however, then the product $h\omega$ corresponds to the CFL number, which in typical computations is not small but of size 1. In this situation, the "exponentially small" remainder terms become of magnitude $\mathcal{O}(1)$, and no conclusions on the long-time behaviour of the method can then be drawn from the familiar backward error analysis. Nevertheless, long-time conservation of energy, and of momentum and actions when appropriate, *is* observed in numerical computations with symplectic methods used with reasonable CFL numbers. The present paper presents a proof of such conservation properties in the case of semilinear wave equations in the weakly nonlinear regime, over time scales that go far beyond linear perturbation arguments. To our knowledge, the results of this paper are the first results that rigorously prove the remarkable long-time conservation properties of symplectic integrators on a class of nonlinear partial differential equations.

We consider the one-dimensional nonlinear wave equation

$$u_{tt} - u_{xx} + \rho u + g(u) = 0 \tag{1}$$

for $t > 0$ and $-\pi \le x \le \pi$ subject to periodic boundary conditions. We assume $\rho > 0$ and a nonlinearity g that is a smooth real function with $g(0) = g'(0) = 0$. We consider small initial data: in appropriate Sobolev norms, the initial values $u(\cdot, 0)$ and $u_t(\cdot, 0)$ are bounded by a small parameter ε. Notice that by rescaling u, this assumption could be rephrased as a $\mathcal{O}(1)$ initial datum but a small non-linearity.

In Section 2 we recall the exact conservation of energy and momentum and, less obvious, the near-conservation of actions over long times $t \le \varepsilon^{-N}$, where N only depends on a non-resonance condition on the frequencies, as shown by Bambusi [1] and Bourgain [3]. With the technique of modulated Fourier expansions that is central also to the present paper, the near-conservation of actions along solutions of (1) has been studied in our paper [6], and for spatial semi-discretizations of (1) by spectral methods in [13]. After discussing the semi-discretization in Section 3, we turn to the time discretization in Section 4.

We consider a class of symplectic and symmetric trigonometric integrators discussed in [14, Chap. XIII], and the familiar Störmer–Verlet or leapfrog method. In Section 4 we describe the trigonometric methods and present numerical experiments illustrating their

conservation properties, which appear particularly remarkable when confronted with the behaviour of a standard explicit Runge-Kutta method.

In Section 5 we state the main result of this paper, concerning the long-time near-conservation of energy, momentum and actions along numerical solutions in the full discretization. The result is proved in Sections 6 and 7, using the technique of *modulated Fourier expansions*. This approach was first used for studying long-time conservation properties of numerical methods for highly oscillatory ordinary differential equations with a single high frequency in [12], and later extended to several frequencies in [5]; see also [14, Chap. XIII] and further references given there. The extension of this technique to infinitely many frequencies, as occur in equation (1), was studied for the analytical problem in [6], and our treatment here essentially follows the lines of this previous work, with additional technical complications arising from the discretization.

In Section 8 we give similar long-time conservation results for the Störmer–Verlet/leapfrog method used with step sizes in the linear stability interval. These results follow from the previous ones by interpreting the leapfrog method as a trigonometric method with modified frequencies.

We are aware of two other papers that deal with long-time energy conservation of symplectic integrators for partial differential equations. Cano [4] also considers the nonlinear wave equation and aims at extending the classical backward error analysis to this situation. Long-time conservation properties are obtained under a list of unverified conditions formulated as conjectures. For symplectic splitting methods applied to the *linear* Schrödinger equation with a small potential, results on long-time energy conservation are given by Dujardin & Faou [8].

2 The nonlinear wave equation with small data

The semilinear wave equation (1) conserves several quantities along every solution $\big(u(x,t), v(x,t)\big)$, with $v = \partial_t u$. The *total energy* or Hamiltonian, defined for 2π-periodic functions u, v as

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

where the potential $U(u)$ is such that $U'(u) = g(u)$, and the *momentum*

$$K(u,v) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \partial_x u(x)\, v(x)\, dx = -\sum_{j=-\infty}^{\infty} \mathrm{i}\, j\, u_{-j}\, v_j \qquad (3)$$

are exactly conserved along every solution $\big(u(\cdot,t), v(\cdot,t)\big)$ of (1). Here, $u_j = \mathcal{F}_j u$ and $v_j = \mathcal{F}_j v$ are the Fourier coefficients in the series $u(x) = \sum_{j=-\infty}^{\infty} u_j \mathrm{e}^{\mathrm{i}jx}$ and correspondingly $v(x)$. Since we consider only real solutions, we note that $u_{-j} = \overline{u}_j$ and $v_{-j} = \overline{v}_j$. In terms of the Fourier coefficients, equation (1) reads

$$\partial_t^2 u_j + \omega_j^2 u_j + \mathcal{F}_j g(u) = 0, \qquad j \in \mathbb{Z}, \qquad (4)$$

with the frequencies

$$\omega_j = \sqrt{\rho + j^2}.$$

The *harmonic actions*

$$I_j(u,v) = \frac{\omega_j}{2}|u_j|^2 + \frac{1}{2\omega_j}|v_j|^2, \qquad (5)$$

for which we note $I_{-j} = I_j$, are conserved for the linear wave equation, that is, for $g(u) \equiv 0$. In the semilinear equation (1), they turn out to remain constant up to small deviations over long times for almost all values of $\rho > 0$, when the initial data are smooth and small [1, 3, 6]. We recall the precise statement of this result, because this will help to understand related assumptions for the numerical discretizations.

We work with the Sobolev space, for $s \geq 0$,

$$H^s = \{v \in L^2(\mathbb{T}) : \|v\|_s < \infty\}, \qquad \|v\|_s = \Big(\sum_{j=-\infty}^{\infty} \omega_j^{2s}\, |v_j|^2\Big)^{1/2},$$

where v_j denote the Fourier coefficients of a 2π-periodic function v. For the initial position and velocity we assume that for suitably large s and small ε,

$$\Big(\|u(\cdot,0)\|_{s+1}^2 + \|v(\cdot,0)\|_s^2\Big)^{1/2} \leq \varepsilon. \qquad (6)$$

Since the analysis of the near-conservation of actions encounters problems with small denominators, we prepare for the formulation of a non-resonance condition. Consider sequences of integers $\mathbf{k} = (k_\ell)_{\ell=0}^{\infty}$ with only finitely many $k_\ell \neq 0$. We denote $|\mathbf{k}| = (|k_\ell|)_{\ell=0}^{\infty}$ and let

$$\|\mathbf{k}\| = \sum_{\ell=0}^{\infty} |k_\ell|, \qquad \mathbf{k}\cdot\boldsymbol{\omega} = \sum_{\ell=0}^{\infty} k_\ell\, \omega_\ell, \qquad \boldsymbol{\omega}^{\sigma|\mathbf{k}|} = \prod_{\ell=0}^{\infty} \omega_\ell^{\sigma|k_\ell|} \qquad (7)$$

for real σ, where we use the notation $\boldsymbol{\omega} = (\omega_\ell)_{\ell=0}^\infty$. For $j \in \mathbb{Z}$, we write $\langle j \rangle = (0, \ldots, 0, 1, 0, \ldots)$ with the only entry at the $|j|$-th position.

For an arbitrary fixed integer $N \geq 1$ and for small $\varepsilon > 0$, we consider the set of near-resonant indices

$$\mathcal{R}_\varepsilon = \{(j, \mathbf{k}) : j \in \mathbb{Z}, \mathbf{k} \neq \pm\langle j \rangle, \|\mathbf{k}\| \leq 2N \text{ with } \big|\omega_j - |\mathbf{k} \cdot \boldsymbol{\omega}|\big| < \varepsilon^{1/2}\}. \tag{8}$$

We impose the following *non-resonance condition*: there are $\sigma > 0$ and a constant C_0 such that

$$\sup_{(j,\mathbf{k}) \in \mathcal{R}_\varepsilon} \frac{\omega_j^\sigma}{\boldsymbol{\omega}^{\sigma|\mathbf{k}|}} \, \varepsilon^{\|\mathbf{k}\|/2} \leq C_0 \, \varepsilon^N . \tag{9}$$

As is shown in [6], condition (9) is implied, for sufficiently large σ, by the non-resonance condition of Bambusi [1], which holds true for almost all (w.r.t. Lebesgue measure) ρ in any fixed interval of positive numbers.

Theorem 1 *[6, Theorem 1] Under the non-resonance condition (9) and assumption (6) on the initial data with $s \geq \sigma + 1$, the estimate*

$$\sum_{\ell=0}^{\infty} \omega_\ell^{2s+1} \, \frac{|I_\ell(t) - I_\ell(0)|}{\varepsilon^2} \leq C\varepsilon \quad \text{for} \quad 0 \leq t \leq \varepsilon^{-N+1}$$

with $I_\ell(t) = I_\ell\big(u(\cdot,t), v(\cdot,t)\big)$ holds with a constant C which depends on s, N, and C_0, but is independent of ε and t.

The smallness of the initial data, which implies that the nonlinearity is small compared to the linear terms, is essential for our analysis. Since we do not impose any further restrictions on the nonlinearity, such an assumption permits to avoid blow-up in finite time.

3 Spectral semi-discretization in space

For the numerical solution of (1) we first discretize in space (method of lines) and then in time (Section 4). Following [13], we consider pseudo-spectral semi-discretization in space with equidistant collocation points $x_k = k\pi/M$ (for $k = -M, \ldots, M-1$). This yields an approximation in form of real-valued trigonometric polynomials

$$u^M(x,t) = {\sum_{|j| \leq M}}' q_j(t) \mathrm{e}^{\mathrm{i}jx}, \qquad v^M(x,t) = {\sum_{|j| \leq M}}' p_j(t) \mathrm{e}^{\mathrm{i}jx} \tag{10}$$

where the prime indicates that the first and last terms in the sum are taken with the factor 1/2. We have $p_j(t) = \frac{d}{dt} q_j(t)$, and the $2M$-periodic coefficient vector $q(t) = (q_j(t))$ is a solution of the $2M$-dimensional system of ordinary differential equations

$$\frac{d^2 q}{dt^2} + \Omega^2 q = f(q) \qquad \text{with} \qquad f(q) = -\mathcal{F}_{2M} g(\mathcal{F}_{2M}^{-1} q). \tag{11}$$

The matrix Ω is diagonal with entries ω_j for $|j| \le M$, and $\mathcal{F}_{2M}$ denotes the discrete Fourier transform:

$$\big(\mathcal{F}_{2M} w\big)_j = \frac{1}{2M} \sum_{k=-M}^{M-1} w_k \, \mathrm{e}^{-\mathrm{i} j x_k}.$$

Since the components of the nonlinearity in (11) are of the form

$$f_j(q) = -\frac{\partial}{\partial q_{-j}} V(q) \quad \text{with} \quad V(q) = \frac{1}{2M} \sum_{k=-M}^{M-1} U\big((\mathcal{F}_{2M}^{-1} q)_k\big)\,,$$

we are concerned with a finite-dimensional complex Hamiltonian system with energy

$$H_M(q,p) = \frac{1}{2} {\sum_{|j|\le M}}' \Big(|p_j|^2 + \omega_j^2 \, |q_j|^2 \Big) + V(q), \tag{12}$$

which is exactly conserved along the solution $\big(q(t), p(t)\big)$ of (11) with $p(t) = dq(t)/dt$. We further consider the actions (for $|j| \le M$) and the momentum

$$I_j(q,p) = \frac{\omega_j}{2} |q_j|^2 + \frac{1}{2\omega_j} |p_j|^2, \qquad K(q,p) = -{\sum_{|j|\le M}}'' \mathrm{i}\, j \, q_{-j} p_j, \tag{13}$$

where the double prime indicates that the first and last terms in the sum are taken with the factor 1/4. The definition of these expressions is motivated by the fact that they agree with the corresponding quantities of Section 2 along the trigonometric polynomials u^M, v^M (with the exception of $I_{\pm M}$, where a factor 4 has been included to get a unified formula). Since we are concerned with real approximations (10), the Fourier coefficients satisfy $q_{-j} = \overline{q}_j$ and $p_{-j} = \overline{p}_j$, so that also $I_{-j} = I_j$.

On the space of $2M$-periodic sequences $q = (q_j)$ we consider the weighted norm

$$\|q\|_s = \Big({\sum_{|j|\le M}}'' \omega_j^{2s} |q_j|^2 \Big)^{1/2}, \tag{14}$$

which is defined such that it equals the H^s norm of the trigonometric polynomial with coefficients q_j. We assume that the initial data $q(0)$ and $p(0)$ satisfy a condition corresponding to (6):

$$\Big(\|q(0)\|_{s+1}^2 + \|p(0)\|_s^2\Big)^{1/2} \le \varepsilon. \tag{15}$$

Theorem 2 *[13, Theorems 3.1 and 3.2] Under the non-resonance condition (9) with exponent σ and the assumption (15) of small initial data with $s \ge \sigma + 1$, the near-conservation estimates*

$$\begin{aligned} \sum_{\ell=0}^{M} \omega_\ell^{2s+1} \frac{|I_\ell(t) - I_\ell(0)|}{\varepsilon^2} &\le C\varepsilon \\ \frac{|K(t) - K(0)|}{\varepsilon^2} &\le C\, t\, \varepsilon M^{-s-1} \end{aligned} \qquad \text{for} \quad 0 \le t \le \varepsilon^{-N+1}$$

for actions $I_\ell(t) = I_\ell\big(q(t), p(t)\big)$ and momentum $K(t) = K\big(q(t), p(t)\big)$ hold with a constant C that depends on s, N, and C_0, but is independent of ε, M, and t.

Since the expression $\sum_{\ell=0}^{M} \omega_\ell^{2s+1} I_\ell(t)$ is essentially (up to the factors in the boundary terms) equal to the squared $H^{s+1} \times H^s$ norm of the solution $\big(q(t), p(t)\big)$, Theorem 2 implies long-time spatial regularity:

$$\Big(\|q(t)\|_{s+1}^2 + \|p(t)\|_s^2\Big)^{1/2} \le \varepsilon(1 + C\varepsilon) \quad \text{for} \quad t \le \varepsilon^{-N+1}. \tag{16}$$

Theorems 1 and 2 have been included as a motivation of our results. They will not be used in the following.

4 Full discretization and numerical phenomena

We consider the class of time discretization methods studied in [14, Chapter XIII], which gives the exact solution for linear problems (11) with $f(q) = 0$, and reduces to the Störmer–Verlet/leapfrog method for (11) with $\Omega = 0$:

$$q^{n+1} - 2\cos(h\Omega)\, q^n + q^{n-1} = h^2 \Psi\, f(\Phi q^n), \tag{17}$$

where $\Psi = \psi(h\Omega)$ and $\Phi = \phi(h\Omega)$ with filter functions ψ and ϕ that are real-valued, bounded, even, and satisfy $\psi(0) = \phi(0) = 1$. A velocity approximation p^n is obtained from

$$2h\,\mathrm{sinc}\,(h\Omega)\, p^n = q^{n+1} - q^{n-1} \tag{18}$$

provided that $\operatorname{sinc}(h\Omega)$ is invertible. Here we use the notation $\operatorname{sinc}\xi = \sin\xi/\xi$.

For an implementation it is more convenient to work with an equivalent one-step mapping $(q^n, p^n) \mapsto (q^{n+1}, p^{n+1})$, which is obtained from adding and subtracting the formulas (17) and (18) and which reads

$$q^{n+1} = \cos(h\Omega)q^n + h\operatorname{sinc}(h\Omega)p^n + \tfrac{1}{2}h^2\,\Psi f(\Phi q^n) \tag{19}$$

$$p^{n+1} = -\Omega\sin(h\Omega)q^n + \cos(h\Omega)p^n + \tfrac{1}{2}h\,\Big(\Psi_0 f(\Phi q^n) + \Psi_1 f(\Phi q^{n+1})\Big).$$

Here, $\Psi_0 = \psi_0(h\Omega)$ and $\Psi_1 = \psi_1(h\Omega)$, where the functions $\psi_i(\xi)$ are defined by the relations $\psi(\xi) = \operatorname{sinc}(\xi)\psi_1(\xi)$ and $\psi_0(\xi) = \cos(\xi)\psi_1(\xi)$. These methods are symmetric for all choices of ψ and ϕ; they are symplectic if

$$\psi(\xi) = \operatorname{sinc}(\xi)\,\phi(\xi) \qquad \text{for all real } \xi. \tag{20}$$

The methods (19) with this property are precisely the mollified impulse methods introduced in [9]. Interpreted as a splitting method, they can be extended for fully non-linear problems.

Condition (20) will be assumed in the following. We note, however, that for non-symplectic methods, the transformation of variables

$$\widehat{q}^n = \chi(h\Omega)q^n, \qquad \widehat{p}^n = \chi(h\Omega)p^n, \tag{21}$$

turns the method (19) into a symplectic method if χ can be chosen as a positive solution of $\chi(\xi)^2 = \phi(\xi)\operatorname{sinc}(\xi)/\psi(\xi)$.

In our numerical experiments we consider the nonlinear wave equation (1) with the following data: $\rho = 0.5$, $g(u) = -u^2$, and initial data

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

for $-\pi \le x \le \pi$. The spatial discretization is (11) with dimension $2M = 2^7$. Considered as 2π-periodic functions, the initial data $u(\cdot,0)$ and $\partial_t u(\cdot,0)$ have a jump discontinuity in the second and first derivative, respectively. The assumption (15) is therefore satisfied for $s < 1.5$.

We first apply a standard explicit Runge–Kutta method in the variable stepsize implementation DOPRI5 of [15], with local error tolerances $Atol = 10^{-5}$ and $Rtol = 10^{-4}$. The program chose 32 735 accepted steps for the integration over the interval $0 \le t \le 550$, which corresponds to an average stepsize $\overline{h} = 0.0168$ and average CFL number $\overline{h}\omega_M = 1.075$. In both pictures of Figure 1 we plot the actions

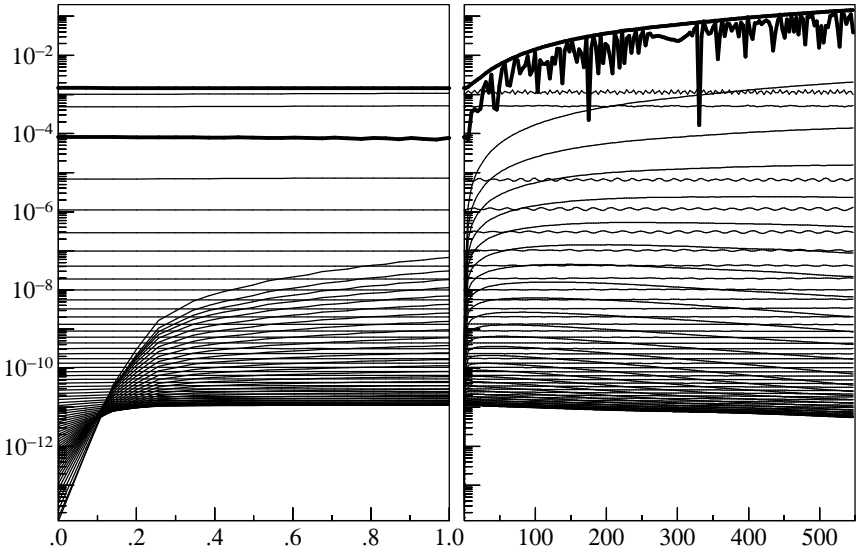


Fig. 1. Actions, total energy (upper bold line), and momentum (lower bold line) along the numerical solution of DOPRI5, average CFL number 1.075.

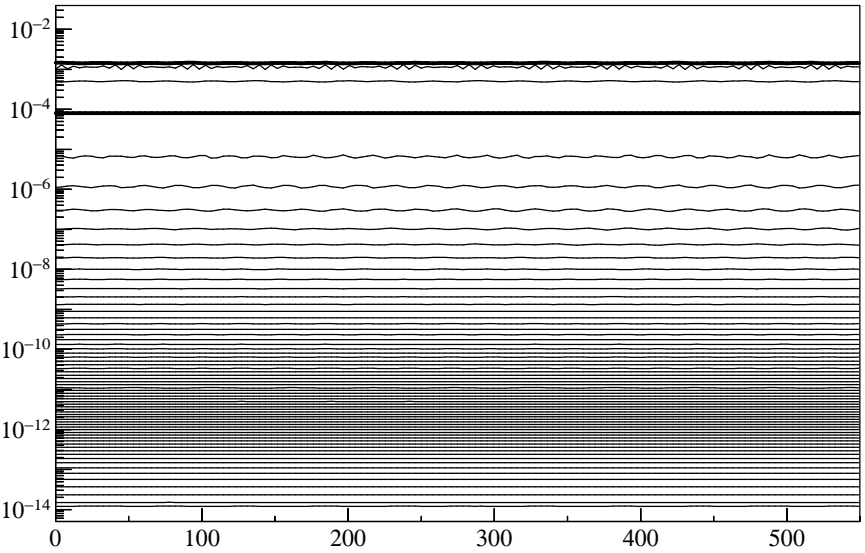


Fig. 2. Actions, total energy (upper bold line), and momentum (lower bold line) along the numerical solution of the trigonometric integrator (19) with $\psi = \mathrm{sinc}$ and $\phi = 1$ for the CFL number $h\omega_M \approx 6.4$.

I_j of (5), the total energy H_M of (12), and the momentum K of (13) along the numerical solution. The left-hand picture illustrates that even on the short interval $0 \leq t \leq 1$, the actions with values below

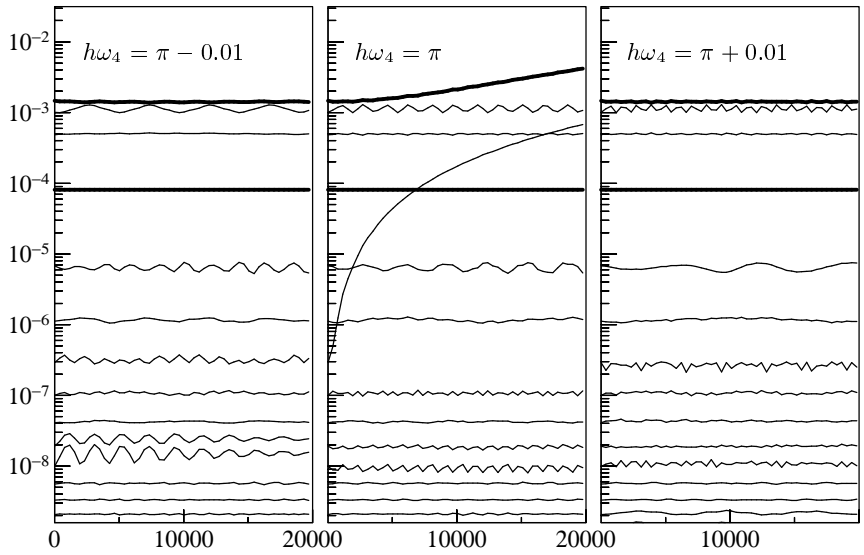


Fig. 3. Illustration of numerical resonance; method (19) with $\psi = \text{sinc}$ and $\phi = 1$

the tolerance are not at all conserved. The right-hand picture shows substantial drifts in all the quantities over a longer time interval.

We now consider method (19) with $\psi = \text{sinc}$ and $\phi = 1$, which was originally proposed in [7]. The method can also be viewed as a special case of the impulse method used in molecular dynamics [10, 17]. We apply the method with stepsize $h = 0.1$ to the above problem. The CFL number then is $h\omega_M \approx 6.4$. Figure 2 illustrates that energy, momentum and actions are very well conserved. Since the regularity of our initial data is not very high ($s < 1{,}5$), this shows that the regularity assumption $s \geq \sigma + 1$, where $\sigma = 2^9$ already for $N = 2$ (c.f. [6]) can be relaxed in concrete examples when a fixed ρ is considered.

In a further experiment with the same problem, we choose stepsizes such that $h\omega_4$ is close to π (Figure 3). In this situation of a numerical resonance, the action I_4 is no longer preserved, which on longer time scales also affects the conservation of energy. The resonance behaviour depends strongly on the choice of the filter functions, cf. [14, Section XIII.2]. For example, with $\phi = \text{sinc}$ and $\psi = \text{sinc}^2$, a method proposed in [9], no numerical resonance is visible.

5 Main results

To get rigorous statements on the good long-time behaviour illustrated in Section 4, we combine the techniques of [6], where the long-time preservation of the harmonic actions along exact solutions of

the semilinear wave equation (1) is shown, with those of [5], where the long-time behaviour of the numerical method (17) is studied for oscillatory Hamiltonian systems with a fixed number of large frequencies. As for spectral semi-discretizations (cf. [13]), we are interested in results that are valid uniformly in M, where $2M$ is the dimension of the spatially discretized system (11).

The analytical tool for understanding the long-time behaviour of the numerical solution of (11) is given by a modulated Fourier expansion in time (see [14, Chapter XIII] and [6]),

$$\widetilde{q}(t) = \sum_{\|\mathbf{k}\| \le 2N} z^{\mathbf{k}}(\varepsilon t) \mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}, \tag{22}$$

approximating the numerical solution q^n at $t = nh$. We use the notation introduced in (7), where now $k_\ell = 0$ for $\ell > M$, since only the frequencies ω_ℓ for $0 \le \ell \le M$ appear in the spatial discretization (11).

In our analysis, we must deal with small denominators (see Section 6). To control these terms, we will use non-resonance conditions. As soon as, for a given step size h, the inequality

$$\Big| \sin\Big(\frac{h}{2}(\omega_j - \mathbf{k}\cdot\boldsymbol{\omega})\Big) \cdot \sin\Big(\frac{h}{2}(\omega_j + \mathbf{k}\cdot\boldsymbol{\omega})\Big) \Big| \ge \varepsilon^{1/2} h^2 \big(\omega_j + |\mathbf{k}\cdot\boldsymbol{\omega}|\big) \tag{23}$$

is violated, we have to make an assumption on the pair of indices $(j, \mathbf{k})$. For a fixed integer $N \ge 1$, subsequently used in the truncation of the expansion (22), the set of near-resonant indices becomes, instead of (8),

$$\mathcal{R}_{\varepsilon,h} = \big\{(j,\mathbf{k}) : \ |j| \le M, \ \|\mathbf{k}\| \le 2N, \ \mathbf{k} \ne \pm\langle j\rangle, \text{ not satisfying (23)}\big\}.$$

Similar to (9), we require the following non-resonance condition: there are $\sigma > 0$ and a constant C_0 such that

$$\sup_{(j,\mathbf{k})\in\mathcal{R}_{\varepsilon,h}} \frac{\omega_j^\sigma}{\boldsymbol{\omega}^{\sigma|\mathbf{k}|}} \, \varepsilon^{\|\mathbf{k}\|/2} \le C_0 \, \varepsilon^N . \tag{24}$$

Notice that, in the limit $h \to 0$, condition (23) becomes equivalent (up to a non-zero constant factor) to $\big|\omega_j^2 - (\mathbf{k}\cdot\boldsymbol{\omega})^2\big| \ge \varepsilon^{1/2} \cdot \big|\omega_j + |\mathbf{k}\cdot\boldsymbol{\omega}|\big|$, so that (24) corresponds precisely to the non-resonance condition (9) for the semilinear wave equation.

We assume the further numerical non-resonance condition

$$|\sin(h\omega_j)| \ge h\varepsilon^{1/2} \quad \text{for} \quad |j| \le M. \tag{25}$$

Yet another non-resonance condition, which leads to improved conservation estimates, reads as follows:

$$\Big|\sin\Big(\frac{h}{2}(\omega_j - \mathbf{k}\cdot\boldsymbol{\omega})\Big)\cdot\sin\Big(\frac{h}{2}(\omega_j + \mathbf{k}\cdot\boldsymbol{\omega})\Big)\Big| \geq c\,h^2\,|\psi(h\omega_j)| \qquad (26)$$
$$\text{for } (j,\mathbf{k}) \text{ of the form } j = j_1 + j_2 \text{ and } \mathbf{k} = \pm\langle j_1\rangle \pm \langle j_2\rangle,$$

with a positive constant $c > 0$. In the limit $h \to 0$, this inequality becomes $\big|\omega_j^2 - (\mathbf{k}\cdot\boldsymbol{\omega})^2\big| \geq 4c$ which is automatically fulfilled for the considered pairs $(j,\mathbf{k})$.

We are now in the position to state the main result of this paper.

Theorem 3 *Under the symplecticity condition (20), under the non-resonance conditions (24) with exponent σ and (25)-(26), and under the assumption (15) of small initial data with $s \geq \sigma+1$ for $(q^0,p^0) = \big(q(0),p(0)\big)$, the near-conservation estimates*

$$\frac{|H_M(q^n,p^n) - H_M(q^0,p^0)|}{\varepsilon^2} \leq C\varepsilon$$
$$\frac{|K(q^n,p^n) - K(q^0,p^0)|}{\varepsilon^2} \leq C\big(\varepsilon + M^{-s} + \varepsilon t M^{-s+1}\big)$$
$$\sum_{\ell=0}^{M} \omega_\ell^{2s+1}\,\frac{|I_\ell(q^n,p^n) - I_\ell(q^0,p^0)|}{\varepsilon^2} \leq C\varepsilon$$

for energy, momentum and actions hold for long times

$$0 \leq t = nh \leq \varepsilon^{-N+1}$$

with a constant C which depends on s, N, and C_0, but is independent of the small parameter ε, the dimension $2M$ of the spatial discretization, the time stepsize h, and the time $t = nh$. If condition (26) fails to be satisfied, then $C\varepsilon$ is weakened to $C\varepsilon^{1/2}$ in the above bounds.

In addition we obtain, by the argument of Section 6.2 in [13], that the original Hamiltonian H of (2) along the trigonometric interpolation polynomials $\big(u^n(x),v^n(x)\big)$ with Fourier coefficients (q_j^n,p_j^n) satisfies the long-time near-conservation estimate

$$\frac{|H(u^n,v^n) - H(u^0,v^0)|}{\varepsilon^2} \leq C\varepsilon \quad \text{for} \quad 0 \leq nh \leq \varepsilon^{-N+1}\,.$$

For a non-symplectic symmetric method (19) the result remains valid in the transformed variables (21).

The proof of Theorem 3 is given in the subsequent Sections 6 and 7. It is based on the idea of interpolating the numerical solution

by a function where different time scales are well separated. This is done by the ansatz (22) which is a truncated series of products of $e^{i\omega_j t}$ (oscillations with respect to the fast time t) with coefficient functions that are smooth in the slow time $\tau = \varepsilon t$. The proof then proceeds as follows:

- Proving existence of smooth functions $z^{\mathbf{k}}(\tau)$ with derivatives bounded independently of ε (on intervals of length ε^{-1}). This is the technically difficult part and elaborated in Section 6. It requires non-resonance conditions and a careful truncation of the series.
- Establishing a Hamiltonian structure and the existence of formal invariants in the differential and algebraic equations for the functions $z^{\mathbf{k}}(\tau)$ (Sections 7.1 – 7.3).
- Proving closeness (on intervals of length ε^{-1}) of the formal invariants to actions I_ℓ, to the total energy H, and to the momentum K (Section 7.4).
- Stretching from short to long intervals of length ε^{-N+1} by patching together previous results along an invariant.

6 Modulated Fourier expansion

Our principal tool for the long-time analysis of the nonlinearly perturbed wave equation is a short-time modulation expansion constructed in this section. To construct this expansion, we combine the tools and techniques developed in [5], [6], and [13].

6.1 Statement of the result

In this section we consider, instead of the symplecticity condition (20), the weaker condition

$$|\psi(h\omega_j)| \le C\,|\operatorname{sinc}(h\omega_j)| \quad \text{for} \quad |j| \le M. \tag{27}$$

In the following result we use the abbreviations (7) and set

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

Theorem 4 *Under the assumptions of Theorem 3 (with the symplecticity assumption (20) relaxed to (27)), there exist truncated asymptotic expansions (with N from (24))*

$$\begin{aligned} \widetilde{q}(t) &= \sum_{\|\mathbf{k}\|\le 2N} z^{\mathbf{k}}(\varepsilon t)\, \mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}, \\ \widetilde{p}(t) &= \operatorname{sinc}(h\Omega)^{-1}\, \frac{\widetilde{q}(t+h)-\widetilde{q}(t-h)}{2h}, \end{aligned} \tag{28}$$

such that the numerical solution q^n, p^n given by method (19), satisfies

$$\|q^n - \widetilde{q}(t)\|_{s+1} + \|p^n - \widetilde{p}(t)\|_s \le C\varepsilon^N \quad \text{for} \quad 0 \le t = nh \le \varepsilon^{-1}. \tag{29}$$

The truncated modulated Fourier expansion is bounded by

$$\|\widetilde{q}(t)\|_{s+1} + \|\widetilde{p}(t)\|_s \le C\varepsilon \quad \text{for} \quad 0 \le t \le \varepsilon^{-1}. \tag{30}$$

On this time interval, we further have, for $|j| \le M$,

$$\widetilde{q}_j(t) = z_j^{\langle j\rangle}(\varepsilon t)\, \mathrm{e}^{\mathrm{i}\omega_j t} + z_j^{-\langle j\rangle}(\varepsilon t)\, \mathrm{e}^{-\mathrm{i}\omega_j t} + r_j\,, \quad \text{with} \quad \|r\|_{s+1} \le C\varepsilon^2. \tag{31}$$

(If condition (26) fails to be satisfied, then the bound is $\|r\|_{s+1} \le C\varepsilon^{3/2}$.) The modulation functions $z^{\mathbf{k}}$ are bounded by

$$\sum_{\|\mathbf{k}\|\le 2N} \left(\frac{\boldsymbol{\omega}^{|\mathbf{k}|}}{\varepsilon^{[\![\mathbf{k}]\!]}} \|z^{\mathbf{k}}(\varepsilon t)\|_s \right)^2 \le C\,. \tag{32}$$

Bounds of the same type hold for any fixed number of derivatives of $z^{\mathbf{k}}$ with respect to the slow time $\tau = \varepsilon t$. Moreover, the modulation functions satisfy $z_{-j}^{-\mathbf{k}} = \overline{z_j^{\mathbf{k}}}$. The constants C are independent of ε, M, h, and of $t \le \varepsilon^{-1}$.

The proof of this result will cover the remainder of this section. It is organized in the same way as the proof of the analogous result for the analytical solution in [6].

6.2 Formal modulation equations

We are looking for a truncated series (28) such that, up to a small defect,

$$\widetilde{q}(t+h) - 2\cos(h\Omega)\,\widetilde{q}(t) + \widetilde{q}(t-h) = h^2 \Psi\, f(\Phi\widetilde{q}(t))$$

with $\widetilde{q}(0) = q^0$, $\widetilde{p}(0) = p^0$, see (17) and (28). We insert the ansatz (28) into this equation, expand the right-hand side into a Taylor series around zero and compare the coefficients of $\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}$. We then get

$$L_j^{\mathbf{k}} z_j^{\mathbf{k}} = -h^2\psi(h\omega_j) \sum_{m\geq 2} \frac{g^{(m)}(0)}{m!} \times \sum_{\mathbf{k}^1+\ldots+\mathbf{k}^m=\mathbf{k}} {\sum_{j_1+\ldots+j_m\equiv j \bmod 2M}}' \phi(h\omega_{j_1}) z_{j_1}^{\mathbf{k}^1} \cdot \ldots \cdot \phi(h\omega_{j_m}) z_{j_m}^{\mathbf{k}^m}, \tag{33}$$

where the right-hand side is obtained as in [13]. The prime on the sum over $j_1, \ldots, j_m$ indicates that with every appearance of $z_{j_i}^{\mathbf{k}^i}$ with $j_i = \pm M$ a factor $1/2$ is included. The operator $L_j^{\mathbf{k}}$ is given as

$$\begin{aligned}\big(L_j^{\mathbf{k}} z_j^{\mathbf{k}}\big)(\tau) &= \mathrm{e}^{\mathrm{i}h(\mathbf{k}\cdot\boldsymbol{\omega})} z_j^{\mathbf{k}}(\tau+\varepsilon h) - 2\cos(h\omega_j) z_j^{\mathbf{k}}(\tau) + \mathrm{e}^{-\mathrm{i}h(\mathbf{k}\cdot\boldsymbol{\omega})} z_j^{\mathbf{k}}(\tau-\varepsilon h) \\ &= 4 s_{\langle j\rangle+\mathbf{k}} s_{\langle j\rangle-\mathbf{k}} z_j^{\mathbf{k}}(\tau) + 2\mathrm{i} s_{2\mathbf{k}} h\varepsilon \dot{z}_j^{\mathbf{k}}(\tau) + c_{2\mathbf{k}} h^2\varepsilon^2 \ddot{z}_j^{\mathbf{k}}(\tau) + \ldots\ .\end{aligned} \tag{34}$$

Here, $s_{\mathbf{k}} = \sin(\frac{h}{2}\,\mathbf{k}\cdot\boldsymbol{\omega})$ and $c_{\mathbf{k}} = \cos(\frac{h}{2}\,\mathbf{k}\cdot\boldsymbol{\omega})$, and the dots on $z_j^{\mathbf{k}}$ represent derivatives with respect to the slow time $\tau = \varepsilon t$. The higher order terms are linear combinations of the rth derivative of $z_j^{\mathbf{k}}$ (for $r \geq 3$) multiplied by $h^r\varepsilon^r$ and containing one of the factors $s_{2\mathbf{k}}$ or $c_{2\mathbf{k}}$.

The first term in (34) vanishes for $\mathbf{k} = \pm\langle j\rangle$, so that in this case the dominating term becomes $\pm 2\mathrm{i}h\sin(h\omega_j)\varepsilon\dot{z}_j^{\pm\langle j\rangle}$ due to condition (25). For $\mathbf{k} \neq \pm\langle j\rangle$ the first term becomes dominant, if the inequality (23) holds. Else, it is not clear which term is dominant, but then the non-resonance condition (24) will ensure that the defect in simply setting $z_j^{\mathbf{k}} \equiv 0$ is of size $\mathcal{O}(\varepsilon^{N+1})$ in an appropriate Sobolev-type norm.

In addition, the initial conditions $\widetilde{q}(0) = q^0$ and $\widetilde{p}(0) = p^0$ need to be taken care of. The condition $\widetilde{q}(0) = q^0$ reads

$$\sum_{\|\mathbf{k}\|\leq 2N} z_j^{\mathbf{k}}(0) = q_j^0, \tag{35}$$

and for $\widetilde{p}(0) = p^0$, we obtain from (28)

$$\frac{1}{2h\ \mathrm{sinc}\,(h\omega_j)} \sum_{\|\mathbf{k}\|\leq 2N} \Big(z_j^{\mathbf{k}}(\varepsilon h)\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} - z_j^{\mathbf{k}}(-\varepsilon h)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} \Big) = p_j^0. \tag{36}$$

6.3 Reverse Picard iteration

We now turn to an iterative construction of the functions $z_j^{\mathbf{k}}$ such that after $4N$ iteration steps, the defect in equations (33), (35), and (36) is of size $\mathcal{O}(\varepsilon^{N+1})$ in the H^s norm. The iteration procedure we employ can be viewed as a Picard iteration on (33) to (36), where we keep only the dominant terms on the left-hand side. We call it reverse Picard iteration, because the highest appearing derivatives do not carry the new iteration number $n+1$.

Indicating by $[\cdot]^n$ the nth iterate of all appearing variables $z_j^{\mathbf{k}}$ taken within the bracket, we set for $\mathbf{k} = \pm\langle j\rangle$

$$\pm 2\mathrm{i}h\varepsilon s_{2j}\Big[\dot z_j^{\pm\langle j\rangle}\Big]^{n+1} = \Big[-h^2\psi(h\omega_j)\sum_{m\ge 2}\frac{g^{(m)}(0)}{m!}\ \times \tag{37}$$
$$\sum_{\mathbf{k}^1+\ldots+\mathbf{k}^m=\mathbf{k}}\ \sum_{j_1+\ldots+j_m\equiv j \bmod 2M}{}' \phi(h\omega_{j_1})z_{j_1}^{\mathbf{k}^1}\cdot\ldots\cdot\phi(h\omega_{j_m})z_{j_m}^{\mathbf{k}^m}$$
$$-\Big(c_{2j}h^2\varepsilon^2\ddot z_j^{\pm\langle j\rangle}+\ldots\Big)\Big]^n$$

with the sines and cosines s_{2j} and c_{2j} defined after formula (34). For $\mathbf{k}\neq\pm\langle j\rangle$ and j that are non-resonant with (23), we set

$$4s_{\langle j\rangle+\mathbf{k}}s_{\langle j\rangle-\mathbf{k}}\Big[z_j^{\mathbf{k}}\Big]^{n+1} = \Big[-h^2\psi(h\omega_j)\sum_{m\ge 2}\frac{g^{(m)}(0)}{m!}\ \times \tag{38}$$
$$\sum_{\mathbf{k}^1+\ldots+\mathbf{k}^m=\mathbf{k}}\ \sum_{j_1+\ldots+j_m\equiv j \bmod 2M}{}' \phi(h\omega_{j_1})z_{j_1}^{\mathbf{k}^1}\cdot\ldots\cdot\phi(h\omega_{j_m})z_{j_m}^{\mathbf{k}^m}$$
$$-\Big(2\mathrm{i}s_{2\mathbf{k}}h\varepsilon\dot z_j^{\mathbf{k}}+c_{2\mathbf{k}}h^2\varepsilon^2\ddot z_j^{\mathbf{k}}+\ldots\Big)\Big]^n,$$

whereas we let $z_j^{\mathbf{k}}=0$ for $\mathbf{k}\neq\pm\langle j\rangle$ in the near-resonant set $\mathcal{R}_{\varepsilon,h}$. The dots indicate the remainder in (34), truncated after the ε^N term.

On the initial conditions we iterate by

$$\Big[z_j^{\langle j\rangle}(0)+z_j^{-\langle j\rangle}(0)\Big]^{n+1} = \Big[q_j^0-\sum_{\mathbf{k}\neq\pm\langle j\rangle}z_j^{\mathbf{k}}(0)\Big]^n \tag{39}$$

and on (36) by

$$\mathrm{i}\omega_j\Big[z_j^{\langle j\rangle}(0)-z_j^{-\langle j\rangle}(0)\Big]^{n+1} = p_j^0$$
$$-\frac{1}{2h\ \mathrm{sinc}\,(h\omega_j)}\Big[\sum_{\mathbf{k}\neq\pm\langle j\rangle}z_j^{\mathbf{k}}(0)\Big(\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h}-\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h}\Big) \tag{40}$$
$$-\sum_{\|\mathbf{k}\|\le K}\Big(\big(z_j^{\mathbf{k}}(\varepsilon h)-z_j^{\mathbf{k}}(0)\big)\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h}-\big(z_j^{\mathbf{k}}(-\varepsilon h)-z_j^{\mathbf{k}}(0)\big)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h}\Big)\Big]^n.$$

In all the above formulas, it is tacitly assumed that $\|\mathbf{k}\| \le K := 2N$ and $\|\mathbf{k}^i\| \le K$ for $i = 1, \dots, m$. In each iteration step, we thus have an initial value problem of first-order differential equations for $z_j^{\pm\langle j\rangle}$ (for $|j| \le M$) and algebraic equations for $z_j^{\mathbf{k}}$ with $\mathbf{k} \ne \pm\langle j\rangle$.

The starting iterates ($n = 0$) are chosen as $z_j^{\mathbf{k}}(\tau) = 0$ for $\mathbf{k} \ne \pm\langle j\rangle$, and $z_j^{\pm\langle j\rangle}(\tau) = z_j^{\pm\langle j\rangle}(0)$ with $z_j^{\pm\langle j\rangle}(0)$ determined from the above formula.

For real initial data we have $q_{-j}^0 = \overline{q_j^0}$ and $p_{-j}^0 = \overline{p_j^0}$, and we observe that the above iteration yields $\big[z_{-j}^{-\mathbf{k}}\big]^n = \overline{\big[z_j^{\mathbf{k}}\big]^n}$ for all iterates n and all $j, \mathbf{k}$ and hence gives real approximations (28).

6.4 Rescaling and estimation of the nonlinear terms

As in [6], we will work with the more convenient rescaling

$$c_j^{\mathbf{k}} = \frac{\boldsymbol{\omega}^{|\mathbf{k}|}}{\varepsilon^{[\![\mathbf{k}]\!]}} z_j^{\mathbf{k}}, \qquad c^{\mathbf{k}} = \big(c_j^{\mathbf{k}}\big)_{|j|\le M} = \frac{\boldsymbol{\omega}^{|\mathbf{k}|}}{\varepsilon^{[\![\mathbf{k}]\!]}} z^{\mathbf{k}}$$

considered in the space $\mathbf{H}^s = (H^s)^{\mathcal{K}} = \{\mathbf{c} = (c^{\mathbf{k}})_{\mathbf{k}\in\mathcal{K}} : c^{\mathbf{k}} \in H^s\}$ with norm $|||\mathbf{c}|||_s^2 = \sum_{\mathbf{k}\in\mathcal{K}} \|c^{\mathbf{k}}\|_s^2$ and where the superscripts $\mathbf{k}$ are in the set

$$\mathcal{K} = \{\mathbf{k} = (k_\ell)_{\ell=0}^M \text{ with integers } k_\ell \,:\, \|\mathbf{k}\| \le K\}$$

with $K = 2N$. The nonlinear function $\mathbf{f} = (f_j^{\mathbf{k}})$ defined as

$$f_j^{\mathbf{k}}(\mathbf{c}) = \frac{\boldsymbol{\omega}^{|\mathbf{k}|}}{\varepsilon^{[\![\mathbf{k}]\!]}} \sum_{m=2}^{N} \frac{g^{(m)}(0)}{m!} \sum_{\mathbf{k}^1+\ldots+\mathbf{k}^m=\mathbf{k}} \frac{\varepsilon^{[\![\mathbf{k}^1]\!]+\cdots+[\![\mathbf{k}^m]\!]}}{\boldsymbol{\omega}^{|\mathbf{k}^1|+\cdots+|\mathbf{k}^m|}} \times \sum_{j_1+\ldots+j_m\equiv j \bmod 2M}{}' \phi(h\omega_{j_1}) c_{j_1}^{\mathbf{k}^1} \cdot \ldots \cdot \phi(h\omega_{j_m}) c_{j_m}^{\mathbf{k}^m}$$

expresses the nonlinearity in (33) in the rescaled variables. With the fact that H^s is a normed algebra, the following bounds are obtained as in [6, Section 3.5] by exploiting the connection between the $2M$-periodic sequence $c_j^{\mathbf{k}}$ and the corresponding trigonometric polynomial (cf. [13]):

$$\sum_{\|\mathbf{k}\|\le K} \|f^{\mathbf{k}}(\mathbf{c})\|_s^2 \le \varepsilon P(|||\mathbf{c}|||_s^2) \tag{41}$$

$$\sum_{|j|\le M} \|f^{\pm\langle j\rangle}(\mathbf{c})\|_s^2 \le \varepsilon^3 P_1(|||\mathbf{c}|||_s^2), \tag{42}$$

where P and P_1 are polynomials with coefficients bounded independently of ε, h, and M. Notice that the function ϕ is bounded.

With the different rescaling

$$\widehat{c}_j^{\mathbf{k}} = \frac{\boldsymbol{\omega}^{s|\mathbf{k}|}}{\varepsilon^{[\![\mathbf{k}]\!]}}\, z_j^{\mathbf{k}}\,, \qquad \widehat{c}^{\mathbf{k}} = \big(\widehat{c}_j^{\mathbf{k}}\big)_{|j|\le M} = \frac{\boldsymbol{\omega}^{s|\mathbf{k}|}}{\varepsilon^{[\![\mathbf{k}]\!]}}\, z^{\mathbf{k}} \tag{43}$$

considered in the space $\mathbf{H}^1 = (H^1)^{\mathcal{K}}$ with norm $|||\widehat{\mathbf{c}}|||_1^2 = \sum_{\|\mathbf{k}\|\le K} \|\widehat{c}^{\mathbf{k}}\|_1^2$, for $\widehat{f}_j^{\mathbf{k}}$ defined as $f_j^{\mathbf{k}}$ but with $\boldsymbol{\omega}^{|\mathbf{k}|}$ replaced by $\boldsymbol{\omega}^{s|\mathbf{k}|}$, we have similar bounds

$$\begin{aligned} \sum_{\|\mathbf{k}\|\le K} \|\widehat{f}^{\mathbf{k}}(\widehat{\mathbf{c}})\|_1^2 &\le \varepsilon \widehat{P}(|||\widehat{\mathbf{c}}|||_1^2) \\ \sum_{|j|\le M} \|\widehat{f}^{\pm\langle j\rangle}(\widehat{\mathbf{c}})\|_1^2 &\le \varepsilon^3\, \widehat{P}_1(|||\widehat{\mathbf{c}}|||_1^2) \end{aligned} \tag{44}$$

with other polynomials $\widehat{P}$ and $\widehat{P}_1$.

6.5 Abstract reformulation of the iteration

For $\mathbf{c} = (c_j^{\mathbf{k}}) \in \mathbf{H}^s$ with $c_j^{\mathbf{k}} = 0$ for all $\mathbf{k} \neq \pm\langle j\rangle$ with $(j,\mathbf{k}) \in \mathcal{R}_{\varepsilon,h}$, we split the components of $\mathbf{c}$ corresponding to $\mathbf{k} = \pm\langle j\rangle$ and $\mathbf{k} \neq \pm\langle j\rangle$ and collect them in $\mathbf{a} = (a_j^{\mathbf{k}}) \in \mathbf{H}^s$ and $\mathbf{b} = (b_j^{\mathbf{k}}) \in \mathbf{H}^s$, respectively:

$$\begin{aligned} a_j^{\mathbf{k}} &= c_j^{\mathbf{k}} \quad \text{if } \mathbf{k} = \pm\langle j\rangle, \quad \text{and 0 else} \\ b_j^{\mathbf{k}} &= c_j^{\mathbf{k}} \quad \text{if (23) is satisfied,} \quad \text{and 0 else.} \end{aligned} \tag{45}$$

We then have $\mathbf{a} + \mathbf{b} = \mathbf{c}$ and $|||\mathbf{a}|||_s^2 + |||\mathbf{b}|||_s^2 = |||\mathbf{c}|||_s^2$. We now introduce differential operators A, B acting on functions $\mathbf{a}(\tau)$ and $\mathbf{b}(\tau)$, respectively:

$$\begin{aligned} (A\mathbf{a})_j^{\pm\langle j\rangle}(\tau) &= \frac{1}{\pm 2\mathrm{i} h\varepsilon s_{2j}} \Big(c_{2j} h^2 \varepsilon^2 \ddot{a}_j^{\pm\langle j\rangle}(\tau) + \dots \Big) \\ (B\mathbf{b})_j^{\mathbf{k}}(\tau) &= \frac{1}{4 s_{\langle j\rangle+\mathbf{k}} s_{\langle j\rangle-\mathbf{k}}} \Big(2\mathrm{i} s_{2\mathbf{k}} h\varepsilon \dot{b}_j^{\mathbf{k}}(\tau) + c_{2\mathbf{k}} h^2 \varepsilon^2 \ddot{b}_j^{\mathbf{k}}(\tau) + \dots \Big) \end{aligned}$$

for $(j,\mathbf{k})$ satisfying (23). These definitions are motivated by formulas (37) and (38), and as in these formulas, the dots represent a truncation after the ε^N terms. In terms of the nonlinear function $\mathbf{f}$ of the preceding subsection, we introduce the functions $\mathbf{F} = (F_j^{\mathbf{k}})$ and $\mathbf{G} = (G_j^{\mathbf{k}})$ with non-vanishing entries

$$\begin{aligned} F_j^{\pm\langle j\rangle}(\mathbf{a},\mathbf{b}) &= -\frac{1}{\pm\mathrm{i}\varepsilon} \frac{\psi(h\omega_j)}{\operatorname{sinc}(h\omega_j)}\, f_j^{\pm\langle j\rangle}(\mathbf{a}+\mathbf{b})\,, \\ G_j^{\mathbf{k}}(\mathbf{a},\mathbf{b}) &= -\frac{h^2(\omega_j + |\mathbf{k}\cdot\boldsymbol{\omega}|)}{4 s_{\langle j\rangle+\mathbf{k}} s_{\langle j\rangle-\mathbf{k}}}\, f_j^{\mathbf{k}}(\mathbf{a}+\mathbf{b}) \end{aligned}$$

for $(j,\mathbf{k})$ satisfying (23). Further we write

$$\big(\boldsymbol{\Omega}\mathbf{c}\big)_j^{\mathbf{k}} = (\omega_j + |\mathbf{k}\cdot\boldsymbol{\omega}|)\, c_j^{\mathbf{k}}, \qquad \big(\boldsymbol{\Psi}\mathbf{c}\big)_j^{\mathbf{k}} = \psi(h\omega_j)\, c_j^{\mathbf{k}}.$$

In terms of $\mathbf{a}$ and $\mathbf{b}$, the iterations (37) and (38) then become of the form

$$\begin{aligned} \dot{\mathbf{a}}^{(n+1)} &= \boldsymbol{\Omega}^{-1}\mathbf{F}(\mathbf{a}^{(n)},\mathbf{b}^{(n)}) - A\mathbf{a}^{(n)} \\ \mathbf{b}^{(n+1)} &= \boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}\,\mathbf{G}(\mathbf{a}^{(n)},\mathbf{b}^{(n)}) - B\mathbf{b}^{(n)}. \end{aligned} \tag{46}$$

By (42), condition (27) gives the bound $|||\mathbf{F}|||_s \le C\varepsilon^{1/2}$, whereas condition (25) yields $|||\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}^{-1}\mathbf{F}|||_s \le C$. By (41) and (23), we have the bound $|||\mathbf{G}|||_s \le C$. These bounds hold uniformly in ε, h, M on bounded subsets of $\mathbf{H}^s$. Analogous bounds are obtained for the derivatives of $\mathbf{F}$ and $\mathbf{G}$. The operators A and B are estimated as

$$\begin{aligned} |||(A\mathbf{a})(\tau)|||_s &\le C\sum_{l=2}^{N} h^{l-2}\varepsilon^{l-3/2}\Big|\Big|\Big|\frac{d^l}{d\tau^l}\mathbf{a}(\tau)\Big|\Big|\Big|_s \\ |||(B\mathbf{b})(\tau)|||_s &\le C\varepsilon^{1/2}|||\dot{\mathbf{b}}(\tau)|||_s + C\sum_{l=2}^{N} h^{l-2}\varepsilon^{l-1/2}\Big|\Big|\Big|\frac{d^l}{d\tau^l}\mathbf{b}(\tau)\Big|\Big|\Big|_s. \end{aligned}$$

The bound for A is obtained with (25), that for B uses (23) and the trivial estimate $|s_{2\mathbf{k}}| = |\sin(h\,\mathbf{k}\cdot\boldsymbol{\omega})| \le h|\mathbf{k}\cdot\boldsymbol{\omega}|$.

The initial value conditions (39) and (40) translate into an equation for $\mathbf{a}^{(n+1)}$ of the form

$$\mathbf{a}^{(n+1)}(0) = \mathbf{v} + P\mathbf{b}^{(n)}(0) + Q(\mathbf{a}+\mathbf{b})^{(n)}(\varepsilon h) \tag{47}$$

where $\mathbf{v}$ has the components

$$v_j^{\pm\langle j\rangle} = \frac{\omega_j}{\varepsilon}\Big(\frac{1}{2}q_j^0 \mp \frac{\mathrm{i}}{2\omega_j}p_j^0\Big).$$

By assumption (15), $\mathbf{v}$ is bounded in $\mathbf{H}^s$. The operators P and Q are given by

$$\begin{aligned} (P\mathbf{b})_j^{\pm\langle j\rangle}(0) &= -\frac{\omega_j}{2\varepsilon s_{2j}}\sum_{\mathbf{k}\neq\pm\langle j\rangle}\Big(\sin(\omega_j h) \pm \sin\big((\mathbf{k}\cdot\boldsymbol{\omega})h\big)\Big)\frac{\varepsilon^{[\![\mathbf{k}]\!]}}{\boldsymbol{\omega}^{|\mathbf{k}|}}\, b_j^{\mathbf{k}}(0) \\ (Q\mathbf{c})_j^{\pm\langle j\rangle}(\tau) &= \mp\frac{\omega_j}{4\mathrm{i}\varepsilon s_{2j}}\sum_{\|\mathbf{k}\|\le K}\Big(e^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h}\,\frac{\varepsilon^{[\![\mathbf{k}]\!]}}{\boldsymbol{\omega}^{|\mathbf{k}|}}\Big(c_j^{\mathbf{k}}(\tau) - c_j^{\mathbf{k}}(0)\Big) \\ &\qquad - e^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h}\,\frac{\varepsilon^{[\![\mathbf{k}]\!]}}{\boldsymbol{\omega}^{|\mathbf{k}|}}\Big(c_j^{\mathbf{k}}(-\tau) - c_j^{\mathbf{k}}(0)\Big)\Big). \end{aligned}$$

For these expressions we have the bounds

$$\begin{aligned}
|||(P\mathbf{b})(0)|||_s &\le C\, |||\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}\, \mathbf{b}(0)|||_s \\
|||(Q\mathbf{c})(\varepsilon h)|||_s &\le C\,\varepsilon \sup_{-\varepsilon h<\tau<\varepsilon h} |||\boldsymbol{\Psi}^{-1}\dot{\mathbf{c}}(\tau)|||_s
\end{aligned}$$

with a constant C that is independent of ε, h, and M, but depends on $K = 2N$. For the first estimate we use $|\sin(\omega_j h) \pm \sin((\mathbf{k}\cdot\boldsymbol{\omega})h)| \le h(\omega_j + |\mathbf{k}\cdot\boldsymbol{\omega}|)$, condition (27), and the Cauchy–Schwarz inequality together with the bound (cf. Lemma 2 of [6])

$$\sum_{\|\mathbf{k}\|\le K} \boldsymbol{\omega}^{-2|\mathbf{k}|} \le C < \infty\,. \tag{48}$$

Similarly, applying the mean value theorem to $\mathbf{c}(\tau)$ yields the second estimate.

The starting iterates are $\mathbf{a}^{(0)}(\tau) = \mathbf{v}$ and $\mathbf{b}^{(0)}(\tau) = 0$.

6.6 Bounds of the modulation functions

In view of the non-resonance conditions (23) and (25), and using the assumption on the filter function (27), we can show by induction that the iterates $\mathbf{a}^{(n)}$ and $\mathbf{b}^{(n)}$ and their derivatives with respect to the slow time $\tau = \varepsilon t$ are bounded in $\mathbf{H}^s$ for $0 \le \tau \le 1$ and $n \le 4N$: more precisely, the $(4N)$-th iterates $\mathbf{a} = \mathbf{a}^{(4N)}$ and $\mathbf{b} = \mathbf{b}^{(4N)}$ satisfy

$$\begin{aligned}
&|||\mathbf{a}(0)|||_s \le C\,, \quad |||\boldsymbol{\Omega}\dot{\mathbf{a}}(\tau)|||_s \le C\varepsilon^{1/2}\,, \quad |||\boldsymbol{\Psi}^{-1}\dot{\mathbf{a}}(\tau)|||_s \le C\,, \\
&|||\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}\mathbf{b}(\tau)|||_s \le C\,,
\end{aligned} \tag{49}$$

with a constant C independent of ε, h, M, but dependent on N. We also obtain analogous bounds for higher derivatives of $\mathbf{a}$ and $\mathbf{b}$ with respect to $\tau = \varepsilon t$. For $z_j^{\mathbf{k}} = \varepsilon^{[\![\mathbf{k}]\!]}\boldsymbol{\omega}^{-|\mathbf{k}|}\, c_j^{\mathbf{k}}$ with $(c_j^{\mathbf{k}}) = \mathbf{c}^{(4N)} = \mathbf{a}^{(4N)} + \mathbf{b}^{(4N)}$, the bounds for $\mathbf{a}$ and $\mathbf{b}$ together yield the bound (32).

These bounds imply $|||\mathbf{c}(\tau) - \mathbf{a}(0)|||_{s+1} \le C$ and as in [6, Section 3.7] give, using (48), the bound (30) for $\widetilde{q}(t)$. For the function $\widetilde{p}(t)$, defined in (28), we use $z_j^{\mathbf{k}}(\varepsilon t + \varepsilon h)\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} - z_j^{\mathbf{k}}(\varepsilon t - \varepsilon h)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} = z_j^{\mathbf{k}}(\varepsilon t)2\mathrm{i}\sin\big((\mathbf{k}\cdot\boldsymbol{\omega})h\big) + r_j^{\mathbf{k}}$, where by the mean value theorem $|r_j^{\mathbf{k}}| \le 2\varepsilon h \max_{-\varepsilon h<\tau<\varepsilon h} |\dot{z}_j^{\mathbf{k}}(\tau)|$. Using the condition (27), the bounds (49) yield in a similar way also the statement (30) for the function $\widetilde{p}(t)$.

Using (42) and (46) we also obtain the bound, for $\mathbf{b} = \mathbf{b}^{(4N)}$,

$$\Big(\sum_{\|\mathbf{k}\|=1} \|(\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}\mathbf{b})^{\mathbf{k}}\|_s^2\Big)^{1/2} \le C\varepsilon\,.$$

Moreover, condition (26) ensures that

$$\sum_{|j|\le M}\sum_{j_1+j_2=j}\sum_{\mathbf{k}=\pm\langle j_1\rangle\pm\langle j_2\rangle}\omega_j^{2(s+1)}\,|b_j^{\mathbf{k}}|^2\le C\varepsilon.$$

These bounds together with (49) yield (31).

With the alternative scaling (43) we obtain the same bounds (for $\tau=\varepsilon t\le 1$),

$$\|\!|\!|\widehat{\mathbf{a}}(0)|\!|\!|_1\le C\,,\quad \|\!|\!|\boldsymbol{\Omega}\dot{\widehat{\mathbf{a}}}(\tau)|\!|\!|_1\le C\varepsilon^{1/2}\,,\quad \|\!|\!|\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}\widehat{\mathbf{b}}(\tau)|\!|\!|_1\le C\,. \tag{50}$$

and again

$$\Big(\sum_{\|\mathbf{k}\|=1}\|(\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}\widehat{\mathbf{b}})^{\mathbf{k}}\|_1^2\Big)^{1/2}\le C\varepsilon\,. \tag{51}$$

For the function $\widehat{\mathbf{a}}(\tau)$ these statements follow at once from the fact that $\|\widehat{a}^{\mathbf{k}}\|_1=\|a^{\mathbf{k}}\|_s$. For the function $\widehat{\mathbf{b}}(\tau)$ one has to repeat the argumentation from before, but one needs no longer take care of initial values.

In addition to these bounds, we also obtain that the map

$$B_\varepsilon\subset H^{s+1}\times H^s\to\mathbf{H}^1:\ \big(u(0),v(0)\big)\mapsto\widehat{\mathbf{c}}(0)$$

(with B_ε the ball of radius ε centered at 0) is Lipschitz continuous with a Lipschitz constant proportional to ε^{-1}: at $t=0$,

$$\|\!|\!|\widehat{\mathbf{a}}_2-\widehat{\mathbf{a}}_1|\!|\!|_1^2+\|\!|\!|\boldsymbol{\Omega}(\widehat{\mathbf{b}}_2-\widehat{\mathbf{b}}_1)|\!|\!|_1^2\le\frac{C}{\varepsilon^2}\Big(\|u_2-u_1\|_{s+1}^2+\|v_2-v_1\|_s^2\Big). \tag{52}$$

6.7 Defects

We consider the defect $\delta(t)=\big(\delta_j(t)\big)_{|j|\le M}$ in (17) divided by $h^2\psi(h\omega_j)$:

$$\delta_j(t)=\frac{\widetilde{q}_j(t+h)-2\cos(h\omega_j)\widetilde{q}_j(t)+\widetilde{q}_j(t-h)}{h^2\psi(h\omega_j)}-f_j(\Phi\widetilde{q}(t))$$

where $f=(f_j)$ is given in (11) and the approximation $\widetilde{q}(t)=\big(q_j(t)\big)$ is given by (28) with $z_j^{\mathbf{k}}=(z_j^{\mathbf{k}})^{(4N)}$ obtained after $4N$ iterations of the procedure in Section 6.3. We write this defect as

$$\delta(t)=\sum_{\|\mathbf{k}\|\le NK}d^{\mathbf{k}}(\varepsilon t)\,\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}+R_{N+1}(\widetilde{q})(t).$$

Here we have set

$$d_j^{\mathbf{k}} = \frac{1}{h^2\psi(h\omega_j)}\,\widetilde{L}_j^{\mathbf{k}} z_j^{\mathbf{k}} + \sum_{m=2}^{N} \frac{g^{(m)}(0)}{m!} \times \qquad (53)$$
$$\sum_{\mathbf{k}^1+\ldots+\mathbf{k}^m=\mathbf{k}} {\sum_{j_1+\ldots+j_m\equiv j \bmod 2M}}' \phi(h\omega_{j_1}) z_{j_1}^{\mathbf{k}^1} \cdot \ldots \cdot \phi(h\omega_{j_m}) z_{j_m}^{\mathbf{k}^m},$$

which is to be considered for $\|\mathbf{k}\| \le NK$, and where we set $z_j^{\mathbf{k}} = 0$ for $\|\mathbf{k}\| > K = 2N$. The operator $\widetilde{L}_j^{\mathbf{k}}$ denotes the truncation of the expansion (34) after the ε^N term. The function R_{N+1} collects the remainder term of the Taylor expansion of f after N terms, and that due to the truncation of the series in (34) after the ε^N term. Using the bound (30) for the remainder in the Taylor expansion of f and the estimates (49) for the $(N+1)$-th derivative for $z_j^{\mathbf{k}}(\tau)$, we have $\|R_{N+1}(\widetilde{q})\|_{s+1} \le C\varepsilon^{N+1}$.

We now use the bound of [6, Section 3.8] to obtain

$$\Big\| \sum_{\|\mathbf{k}\|\le NK} d^{\mathbf{k}}(\varepsilon t)\, \mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t} \Big\|_s^2 \le C \sum_{\|\mathbf{k}\|\le NK} \Big\| \boldsymbol{\omega}^{|\mathbf{k}|}\, d^{\mathbf{k}}(\varepsilon t) \Big\|_s^2. \qquad (54)$$

In he following two subsections we estimate the right-hand side of (54) by $C\varepsilon^{2(N+1)}$.

6.8 Defect in the truncated and near-resonant modes

For $\|\mathbf{k}\| > K = 2N$ (truncated modes) and for $(j,\mathbf{k})$ in the set $\mathcal{R}_{\varepsilon,h}$ of near-resonances we have by definition $z_j^{\mathbf{k}} = 0$. In both situations the defect reads

$$d_j^{\mathbf{k}} = \sum_{m=2}^{N} \frac{g^{(m)}(0)}{m!} \times$$
$$\sum_{\mathbf{k}^1+\ldots+\mathbf{k}^m=\mathbf{k}} {\sum_{j_1+\ldots+j_m\equiv j \bmod 2M}}' \phi(h\omega_{j_1}) z_{j_1}^{\mathbf{k}^1} \cdot \ldots \cdot \phi(h\omega_{j_m}) z_{j_m}^{\mathbf{k}^m}.$$

For truncated modes we write the defect as $d_j^{\mathbf{k}} = \varepsilon^{[\![\mathbf{k}]\!]}\, \boldsymbol{\omega}^{-|\mathbf{k}|}\, f_j^{\mathbf{k}}$, and we notice that by (49) and (41), used with NK in place of K, the bound $|||\mathbf{f}|||_s^2 \le C\varepsilon$ holds. We thus have

$$\sum_{\|\mathbf{k}\|>K} {\sum_{|j|\le M}}' \omega_j^{2s} \big|\boldsymbol{\omega}^{|\mathbf{k}|}\, d_j^{\mathbf{k}}\big|^2 = \sum_{\|\mathbf{k}\|>K} {\sum_{|j|\le M}}' \omega_j^{2s} \big|f_j^{\mathbf{k}}\big|^2\, \varepsilon^{2[\![\mathbf{k}]\!]}$$

and hence, since $2[\![\mathbf{k}]\!] = \|\mathbf{k}\| + 1 \geq K + 2 = 2(N+1)$,

$$\sum_{\|\mathbf{k}\|>K} {\sum_{|j|\leq M}}' \omega_j^{2s} \left|\boldsymbol{\omega}^{|\mathbf{k}|}\, d_j^{\mathbf{k}}\right|^2 \leq C\varepsilon^{2(N+1)}.$$

For the near-resonant modes we consider the rescaling (43), so that $d_j^{\mathbf{k}} = \varepsilon^{[\![\mathbf{k}]\!]}\, \boldsymbol{\omega}^{-s|\mathbf{k}|}\, \widehat{f}_j^{\mathbf{k}}$. We have $|||\widehat{\mathbf{f}}|||_1^2 \leq C\varepsilon$ by (50) and (44), so that

$$\begin{aligned}\sum_{(j,\mathbf{k})\in\mathcal{R}_{\varepsilon,h}} \omega_j^{2s} \left|\boldsymbol{\omega}^{|\mathbf{k}|}\, d_j^{\mathbf{k}}\right|^2 &= \sum_{(j,\mathbf{k})\in\mathcal{R}_{\varepsilon,h}} \frac{\omega_j^{2(s-1)}}{\boldsymbol{\omega}^{2(s-1)|\mathbf{k}|}}\, \varepsilon^{2[\![\mathbf{k}]\!]}\, \omega_j^2 |\widehat{f}_j^{\mathbf{k}}|^2 \\ &\leq C \sup_{(j,\mathbf{k})\in\mathcal{R}_{\varepsilon,h}} \frac{\omega_j^{2(s-1)}\, \varepsilon^{2[\![\mathbf{k}]\!]+1}}{\boldsymbol{\omega}^{2(s-1)|\mathbf{k}|}}\,.\end{aligned}$$

The non-resonance condition (24) is formulated such that the supremum is bounded by $C_0^2\, \varepsilon^{2(N+1)}$, and hence

$$\sum_{(j,\mathbf{k})\in\mathcal{R}_{\varepsilon,h}} \omega_j^{2s} \left|\boldsymbol{\omega}^{|\mathbf{k}|}\, d_j^{\mathbf{k}}\right|^2 \leq C\, \varepsilon^{2(N+1)}. \tag{55}$$

6.9 Defect in the non-resonant modes

We now assume that $\|\mathbf{k}\| \leq K$ and that $(j,\mathbf{k})$ satisfies the non-resonance condition (23), so that in the scaled variables $c_j^{\mathbf{k}}$ of Section 6.4 the defect satisfies

$$\boldsymbol{\omega}^{|\mathbf{k}|} d_j^{\mathbf{k}} = \varepsilon^{[\![\mathbf{k}]\!]} \Big(\frac{1}{h^2 \psi(h\omega_j)}\, \widetilde{L}_j^{\mathbf{k}} c_j^{\mathbf{k}} + f_j^{\mathbf{k}}(\mathbf{c}) \Big).$$

Written in terms of the components $\mathbf{a}$ and $\mathbf{b}$ of (45) we have

$$\begin{aligned}\omega_j d_j^{\pm\langle j\rangle} &= \varepsilon \Big(\pm 2\mathrm{i}\varepsilon\omega_j\, \frac{\operatorname{sinc}(h\omega_j)}{\psi(h\omega_j)} \left(\dot{a}_j^{\pm\langle j\rangle} + (A\mathbf{a})_j^{\pm\langle j\rangle}\right) + f_j^{\pm\langle j\rangle}(\mathbf{a}+\mathbf{b}) \Big) \\ \boldsymbol{\omega}^{|\mathbf{k}|} d_j^{\mathbf{k}} &= \varepsilon^{[\![\mathbf{k}]\!]} \Big(\frac{4\, s_{\langle j\rangle+\mathbf{k}} s_{\langle j\rangle-\mathbf{k}}}{h^2 \psi(h\omega_j)} \left(b_j^{\mathbf{k}} + (B\mathbf{b})_j^{\mathbf{k}}\right) + f_j^{\mathbf{k}}(\mathbf{a}+\mathbf{b}) \Big).\end{aligned}$$

It should be noted that the functions in this defect are actually the $4N$-th iterates $\mathbf{a}^{(4N)}$ and $\mathbf{b}^{(4N)}$ of the iteration in Section 6.3. Expressing $f_j^{\pm\langle j\rangle}(\mathbf{a}+\mathbf{b})$ and $f_j^{\mathbf{k}}(\mathbf{a}+\mathbf{b})$ in terms of $\mathbf{F}(\mathbf{a},\mathbf{b})$ and $\mathbf{G}(\mathbf{a},\mathbf{b})$

and inserting $\mathbf{F}$ and $\mathbf{G}$ from (46) into this defect, relates it to the increment of the iteration in the following way:

$$\begin{aligned}
\omega_j d_j^{\pm\langle j\rangle} &= 2\omega_j \alpha_j^{\pm\langle j\rangle} \Big(\big[\dot a_j^{\pm\langle j\rangle}\big]^{(4N)} - \big[\dot a_j^{\pm\langle j\rangle}\big]^{(4N+1)} \Big), \\
\alpha_j^{\pm\langle j\rangle} &:= \pm \mathrm{i}\varepsilon^2 \, \frac{\operatorname{sinc}(h\omega_j)}{\psi(h\omega_j)}, \\
\boldsymbol{\omega}^{|\mathbf{k}|} d_j^{\mathbf{k}} &= \beta_j^{\mathbf{k}} \Big(\big[b_j^{\mathbf{k}}\big]^{(4N)} - \big[b_j^{\mathbf{k}}\big]^{(4N+1)} \Big), \\
\beta_j^{\mathbf{k}} &:= \varepsilon^{[\![\mathbf{k}]\!]} \, \frac{4\, s_{\langle j\rangle+\mathbf{k}} s_{\langle j\rangle-\mathbf{k}}}{h^2 \psi(h\omega_j)}.
\end{aligned}$$

Motivated by these relations we introduce new variables

$$\widetilde a_j^{\pm\langle j\rangle} := \alpha_j^{\pm\langle j\rangle} a_j^{\pm\langle j\rangle}, \qquad \widetilde b_j^{\mathbf{k}} := \beta_j^{\mathbf{k}} b_j^{\mathbf{k}}. \tag{56}$$

Collecting these variables into vectors and using the transformed functions

$$\begin{aligned}
\widetilde F_j^{\pm\langle j\rangle}(\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}) &:= \alpha_j^{\pm\langle j\rangle} \, F_j^{\pm\langle j\rangle}(\boldsymbol{\alpha}^{-1}\widetilde{\mathbf{a}}, \boldsymbol{\beta}^{-1}\widetilde{\mathbf{b}}) \\
&= -\varepsilon \, f_j^{\pm\langle j\rangle}(\boldsymbol{\alpha}^{-1}\widetilde{\mathbf{a}} + \boldsymbol{\beta}^{-1}\widetilde{\mathbf{b}}) \\
\widetilde G_j^{\mathbf{k}}(\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}) &:= \frac{\beta_j^{\mathbf{k}} \psi(h\omega_j)}{\omega_j + |\mathbf{k}\cdot\boldsymbol{\omega}|} \, G_j^{\mathbf{k}}(\boldsymbol{\alpha}^{-1}\widetilde{\mathbf{a}}, \boldsymbol{\beta}^{-1}\widetilde{\mathbf{b}}) \\
&= -\varepsilon^{[\![\mathbf{k}]\!]} \, f_j^{\mathbf{k}}(\boldsymbol{\alpha}^{-1}\widetilde{\mathbf{a}} + \boldsymbol{\beta}^{-1}\widetilde{\mathbf{b}})
\end{aligned}$$

the iteration (46)-(47) becomes

$$\begin{aligned}
\dot{\widetilde{\mathbf{a}}}^{(n+1)} &= \boldsymbol{\Omega}^{-1} \widetilde{\mathbf{F}}(\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}) - A\widetilde{\mathbf{a}}^{(n)} \\
\widetilde{\mathbf{b}}^{(n+1)} &= \widetilde{\mathbf{G}}(\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}) - B\widetilde{\mathbf{b}}^{(n)} \\
\widetilde{\mathbf{a}}^{(n+1)}(0) &= \boldsymbol{\alpha}\mathbf{v} + \widetilde P \widetilde{\mathbf{b}}^{(n)}(0) + Q\widetilde{\mathbf{a}}^{(n)}(\varepsilon h) + \widetilde Q \widetilde{\mathbf{b}}^{(n)}(\varepsilon h).
\end{aligned} \tag{57}$$

In the iteration for the initial values we abbreviate $\widetilde P = \boldsymbol{\alpha} P \boldsymbol{\beta}^{-1}$, $\widetilde Q = \boldsymbol{\alpha} Q \boldsymbol{\beta}^{-1}$, which are bounded by

$$\begin{aligned}
\|\!|(\widetilde P \widetilde{\mathbf{b}})(0)|\!\|_s &\le C\, \varepsilon^{1/2} \, \|\!|\widetilde{\mathbf{b}}(0)|\!\|_s \\
\|\!|(Q\widetilde{\mathbf{a}})(\varepsilon h)|\!\|_s &\le C\, \varepsilon^{1/2} \sup_{-\varepsilon h < \tau < \varepsilon h} \|\!|\dot{\widetilde{\mathbf{a}}}(\tau)|\!\|_s \\
\|\!|(\widetilde Q \widetilde{\mathbf{b}})(\varepsilon h)|\!\|_s &\le C\, \varepsilon^{3/2} \sup_{-\varepsilon h < \tau < \varepsilon h} \|\!|\boldsymbol{\Omega}^{-1}\dot{\widetilde{\mathbf{b}}}(\tau)|\!\|_s .
\end{aligned}$$

In an $\mathbf{H}^s$ neighbourhood of 0 where the bounds (49) hold, the partial derivatives of $\widetilde{\mathbf{F}}$ with respect to $\widetilde{\mathbf{a}}$ and $\widetilde{\mathbf{b}}$ and those of $\widetilde{\mathbf{G}}$ with respect to

$\widetilde{\mathbf{b}}$ are bounded by $\mathcal{O}(\varepsilon^{1/2})$, whereas the derivatives of $\widetilde{\mathbf{G}}$ with respect to $\widetilde{\mathbf{a}}$ is only $\mathcal{O}(1)$. This is the same situation as we had for the exact solution in [6]. As in that paper one proves

$$\begin{aligned} \|\!|\boldsymbol{\Omega}(\dot{\widetilde{\mathbf{a}}}^{(4N+1)}(\tau) - \dot{\widetilde{\mathbf{a}}}^{(4N)}(\tau))|\!\|_s &\le C\,\varepsilon^{N+2} \\ \|\!|\widetilde{\mathbf{b}}^{(4N+1)}(\tau) - \widetilde{\mathbf{b}}^{(4N)}(\tau)|\!\|_s &\le C\,\varepsilon^{N+2} \\ \|\!|\widetilde{\mathbf{a}}^{(4N+1)}(0) - \widetilde{\mathbf{a}}^{(4N)}(0)|\!\|_s &\le C\,\varepsilon^{N+2}. \end{aligned}$$

These estimates yield the desired bound of the defect in the non-resonant modes $(j,\mathbf{k}) \notin \mathcal{R}_{\varepsilon,h}$. Combined with the corresponding estimates of Subsection 6.8 we obtain

$$\Big(\sum_{\|\mathbf{k}\|\le K} \|\boldsymbol{\omega}^{|\mathbf{k}|} d^{\mathbf{k}}(\tau)\|_s^2 \Big)^{1/2} \le C\varepsilon^{N+1} \quad \text{for} \quad \tau \le 1. \tag{58}$$

Consequently, the defect $\delta(t)$ (see Subsection 6.7) satisfies

$$\|\Omega^{-1}\delta(t)\|_{s+1} = \|\delta(t)\|_s \le C\varepsilon^{N+1} \quad \text{for} \quad t \le \varepsilon^{-1}. \tag{59}$$

For the defect in the initial conditions (35) and (36) we obtain

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

For the alternative scaling $\widehat{c}_j^{\mathbf{k}} = \boldsymbol{\omega}^{s|\mathbf{k}|} z_j^{\mathbf{k}}$, we obtain

$$\Big(\sum_{\|\mathbf{k}\|\le K} \|\boldsymbol{\omega}^{s|\mathbf{k}|} d^{\mathbf{k}}(\tau)\|_1^2 \Big)^{1/2} \le C\varepsilon^{N+1} \quad \text{for} \quad \tau \le 1. \tag{60}$$

6.10 Remainder term of the modulated Fourier expansion

We write the method (19) in the form

$$\begin{aligned} \begin{pmatrix} q^{n+1} \\ \Omega^{-1}p^{n+1} \end{pmatrix} &= \begin{pmatrix} \cos(h\Omega) & \sin(h\Omega) \\ -\sin(h\Omega) & \cos(h\Omega) \end{pmatrix} \begin{pmatrix} q^n \\ \Omega^{-1}p^n \end{pmatrix} \\ &\quad + \frac{h}{2}\,\Psi_1 \begin{pmatrix} \sin(h\Omega) f^n \\ \cos(h\Omega) f^n + f^{n+1} \end{pmatrix} \end{aligned}$$

where $f^n = \Omega^{-1} f(\Phi q^n)$, and we notice that Ψ_1 is a matrix, bounded independently of h and the dimension M. The differences $\Delta q^n := \widetilde{q}(t_n) - q^n$ and $\Delta p^n := \widetilde{p}(t_n) - p^n$, where $t_n := nh$, satisfy the same relation with f^n replaced by $\Omega^{-1}\big(f(\Phi\widetilde{q}(t_n)) - f(\Phi q^n)\big) + \delta(t_n)$. Using

the Lipschitz bound (cf. Section 4.2 in [13] on the relation between $f(q)$ and $g(u)$ of (1))

$$\begin{aligned}\big\|\Omega^{-1}\big(f(q_1)-f(q_2)\big)\big\|_{s+1} &= \|f(q_1)-f(q_2)\|_s \\ &\le C\varepsilon\|q_1-q_2\|_s \le C\varepsilon\|q_1-q_2\|_{s+1}\end{aligned}$$

for $q_1, q_2 \in H^s$ satisfying $\|q_i\|_s \le M\varepsilon$, and the estimate (59) for the defect yields

$$\begin{aligned}\left\| \begin{pmatrix} \Delta q^{n+1} \\ \Omega^{-1}\Delta p^{n+1}\end{pmatrix} \right\|_{s+1} \le \left\| \begin{pmatrix} \Delta q^{n} \\ \Omega^{-1}\Delta p^{n}\end{pmatrix} \right\|_{s+1} \\ + \frac{h}{2}\Big(C\varepsilon\|\Delta q^n\|_{s+1} + C\varepsilon\|\Delta q^{n+1}\|_{s+1} + C\varepsilon^{N+1}\Big).\end{aligned}$$

Solving this inequality gives the estimate

$$\|\Delta q^n\|_{s+1} + \|\Omega^{-1}\Delta p^n\|_{s+1} \le C(1+t_n)\varepsilon^{N+1} \qquad \text{for} \quad t_n \le \varepsilon^{-1}$$

and thus completes the proof of Theorem 4.

7 Conservation properties

We now show that the system of equations determining the modulation functions has almost-invariants close to the actions, the momentum and the total energy along numerical solutions given by the full discretization (17)–(18). The proof takes up arguments of [6] for the conservation of actions, of [13] for the conservation of momentum and aspects of the space discretization, and of [14, Ch. XIII] for the conservation of energy and for the aspects arising from the time discretization.

7.1 The extended potential

The defect formula (53) can be rewritten as

$$\frac{1}{h^2\psi(h\omega_j)}\,\widetilde{L}_j^{\mathbf{k}} z_j^{\mathbf{k}} + \nabla_{-j}^{-\mathbf{k}}\mathcal{U}(\Phi\mathbf{z}) = d_j^{\mathbf{k}}, \tag{61}$$

where $\nabla_{-j}^{-\mathbf{k}}\mathcal{U}(\mathbf{y})$ is the partial derivative with respect to $y_{-j}^{-\mathbf{k}}$ of the *extended potential* (see [13])

$$\mathcal{U}(\mathbf{y}) = \sum_{l=-N}^{N} \mathcal{U}_l(\mathbf{y}) \tag{62}$$

$$\mathcal{U}_l(\mathbf{y}) = \sum_{m=2}^{N} \frac{U^{(m+1)}(0)}{(m+1)!} \sum_{\mathbf{k}^1+\cdots+\mathbf{k}^{m+1}=\mathbf{0}} {\sum_{j_1+\cdots+j_{m+1}=2Ml}}' y_{j_1}^{\mathbf{k}^1}\dots y_{j_{m+1}}^{\mathbf{k}^{m+1}},$$

where again $\|\mathbf{k}^i\| \le 2N$ and $|j_i| \le M$, and $U(u)$ is the potential in (2).

7.2 Invariance under group actions

The existence of almost-invariants for the system (61) turns out to be a consequence, in the spirit of Noether's theorem, of the invariance of the extended potential under continuous group actions: for an arbitrary real sequence $\boldsymbol{\mu} = (\mu_\ell)_{\ell\ge 0}$ and for $\theta \in \mathbb{R}$, let

$$\begin{aligned} S_{\boldsymbol{\mu}}(\theta)\mathbf{y} &= \Big(\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\mu})\theta} y_j^{\mathbf{k}}\Big)_{|j|\le M, \|\mathbf{k}\|\le K}, \\ T(\theta)\mathbf{y} &= \Big(\mathrm{e}^{\mathrm{i}j\theta} y_j^{\mathbf{k}}\Big)_{|j|\le M, \|\mathbf{k}\|\le K}. \end{aligned} \tag{63}$$

Since the sum in the definition of $\mathcal{U}$ is over $\mathbf{k}^1 + \cdots + \mathbf{k}^{m+1} = \mathbf{0}$ and that in $\mathcal{U}_0$ over $j_1 + \cdots + j_{m+1} = 0$, we have

$$\mathcal{U}(S_{\boldsymbol{\mu}}(\theta)\mathbf{y}) = \mathcal{U}(\mathbf{y}), \qquad \mathcal{U}_0(T(\theta)\mathbf{y}) = \mathcal{U}_0(\mathbf{y}) \quad \text{for} \quad \theta \in \mathbb{R}.$$

Differentiating these relations with respect to θ yields

$$\begin{aligned} 0 &= \frac{d}{d\theta}\Big|_{\theta=0} \mathcal{U}(S_{\boldsymbol{\mu}}(\theta)\mathbf{y}) = \sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' \mathrm{i}(\mathbf{k}\cdot\boldsymbol{\mu})\, y_j^{\mathbf{k}} \nabla_j^{\mathbf{k}} \mathcal{U}(\mathbf{y}) \\ 0 &= \frac{d}{d\theta}\Big|_{\theta=0} \mathcal{U}_0(T(\theta)\mathbf{y}) = \sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' \mathrm{i}j\, y_j^{\mathbf{k}} \nabla_j^{\mathbf{k}} \mathcal{U}_0(\mathbf{y}). \end{aligned} \tag{64}$$

7.3 Almost-invariants of the modulation system

We now multiply (61) once with $\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\mu})\phi(h\omega_j) z_{-j}^{-\mathbf{k}}$ and once with $\mathrm{i}j\phi(h\omega_j) z_{-j}^{-\mathbf{k}}$, and sum over j and $\mathbf{k}$ with $|j| \le M$ and $\|\mathbf{k}\| \le K$. Thanks to (64), we obtain

$$\begin{aligned} &\sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' \mathrm{i}(\mathbf{k}\cdot\boldsymbol{\mu}) \frac{\phi(h\omega_j)}{h^2\psi(h\omega_j)}\, z_{-j}^{-\mathbf{k}} \widetilde{L}_j^{\mathbf{k}} z_j^{\mathbf{k}} \\ &\qquad = \sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' \mathrm{i}(\mathbf{k}\cdot\boldsymbol{\mu})\phi(h\omega_j)\, z_{-j}^{-\mathbf{k}} d_j^{\mathbf{k}}, \end{aligned} \tag{65}$$

$$\begin{aligned} &\sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' \mathrm{i}j \frac{\phi(h\omega_j)}{h^2\psi(h\omega_j)}\, z_{-j}^{-\mathbf{k}} \widetilde{L}_j^{\mathbf{k}} z_j^{\mathbf{k}} \\ &\qquad = \sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' \mathrm{i}j\phi(h\omega_j)\, z_{-j}^{-\mathbf{k}} \Big(d_j^{\mathbf{k}} - \sum_{l\ne 0} \nabla_{-j}^{-\mathbf{k}} \mathcal{U}_l(\Phi\mathbf{z}) \Big). \end{aligned} \tag{66}$$

By the expansion (34) of the operator $\widetilde{L}_j^{\mathbf{k}}$, only expressions of the following type appear for $z(\tau) = z_j^{\mathbf{k}}(\tau)$ and $\overline{z}(\tau) = z_{-j}^{-\mathbf{k}}(\tau)$ on the left-hand side of the above equations:

$$\begin{aligned} \mathrm{Re}\,\overline{z}z^{(2l+1)} &= \mathrm{Re}\,\frac{d}{d\tau}\Big(\overline{z}z^{(2l)} - \ldots \pm \overline{z}^{(l-1)}z^{(l+1)} \mp \frac{1}{2}\,\overline{z}^{(l)}z^{(l)}\Big) \\ \mathrm{Im}\,\overline{z}z^{(2l+2)} &= \mathrm{Im}\,\frac{d}{d\tau}\Big(\overline{z}z^{(2l+1)} - \dot{\overline{z}}z^{(2l)} + \ldots \pm \overline{z}^{(l)}z^{(l+1)}\Big). \end{aligned} \tag{67}$$

Therefore, the left-hand sides can be written as total derivatives of functions $\varepsilon\mathcal{J}_{\boldsymbol{\mu}}[\mathbf{z}](\tau)$ and $\varepsilon\mathcal{K}[\mathbf{z}](\tau)$ which depend on $\mathbf{z}(\tau)$ and its derivatives $\varepsilon^\ell \mathbf{z}^{(\ell)}(\tau)$ for $\ell = 1, \ldots, N-1$. In this way, (65) and (66) become

$$-\varepsilon\,\frac{d}{d\tau}\mathcal{J}_{\boldsymbol{\mu}}[\mathbf{z}] = \sum_{\|\mathbf{k}\|\le K}\sum_{|j|\le M}{}' \mathrm{i}(\mathbf{k}\cdot\boldsymbol{\mu})\phi(h\omega_j)\, z_{-j}^{-\mathbf{k}} d_j^{\mathbf{k}} \tag{68}$$

$$-\varepsilon\,\frac{d}{d\tau}\mathcal{K}[\mathbf{z}] = \sum_{\|\mathbf{k}\|\le K}\sum_{|j|\le M}{}' \mathrm{i}j\phi(h\omega_j)\, z_{-j}^{-\mathbf{k}}\Big(d_j^{\mathbf{k}} - \sum_{l\ne 0}\nabla_{-j}^{-\mathbf{k}}\mathcal{U}_l(\Phi\mathbf{z})\Big). \tag{69}$$

In the following we consider the special case of $\boldsymbol{\mu} = \langle\ell\rangle$. From the smallness of the right-hand sides in (68) and (69) we infer the following.

Theorem 5 *Under the conditions of Theorem 4 we have, for $\tau \le 1$,*

$$\sum_{\ell=0}^{M}\omega_\ell^{2s+1}\,\Big|\frac{d}{d\tau}\mathcal{J}_{\langle\ell\rangle}[\mathbf{z}](\tau)\Big| \le C\,\varepsilon^{N+1},$$

$$\Big|\frac{d}{d\tau}\mathcal{K}[\mathbf{z}](\tau)\Big| \le C\,\big(\varepsilon^{N+1} + \varepsilon^2 M^{-s+1}\big).$$

Proof The result is obtained from (68) and (69) with the arguments of [6,13] as follows. With the bounds (50) and (60), the estimate for the functions $\mathcal{J}_{\langle\ell\rangle}[\mathbf{z}]$ follows with the proof of Theorem 3 in [6]. With the bound (32) and with the bounds $|||\mathbf{z}|||_1 \le C\varepsilon$ and $|||\mathbf{d}|||_0 \le C\varepsilon^{N+1}$, which follow from (32) and (58), the estimate for $\mathcal{K}[\mathbf{z}]$ is obtained as in Theorem 5.2 of [13]. □

A further almost-invariant is obtained by multiplying (61) with the expression $\phi(h\omega_j)\big(\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})z_{-j}^{-\mathbf{k}} + \varepsilon\dot{z}_{-j}^{-\mathbf{k}}\big)$, summing over j and $\mathbf{k}$, and using (64):

$$\begin{aligned} &\sum_{\|\mathbf{k}\|\le K}\sum_{|j|\le M}{}'\frac{\phi(h\omega_j)}{h^2\psi(h\omega_j)}\,\big(\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})z_{-j}^{-\mathbf{k}} + \varepsilon\dot{z}_{-j}^{-\mathbf{k}}\big)\widetilde{L}_j^{\mathbf{k}} z_j^{\mathbf{k}} \\ &\qquad + \varepsilon\,\frac{d}{d\tau}\mathcal{U}(\Phi\mathbf{z}) = \sum_{\|\mathbf{k}\|\le K}\sum_{|j|\le M}{}'\phi(h\omega_j)\,\big(\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})z_{-j}^{-\mathbf{k}} + \varepsilon\dot{z}_{-j}^{-\mathbf{k}}\big)d_j^{\mathbf{k}}. \end{aligned} \tag{70}$$

In addition to the identities (67) we also use

$$\begin{aligned}\operatorname{Re}\dot{\overline{z}}z^{(2l)} &= \operatorname{Re}\frac{d}{dt}\Big(\dot{\overline{z}}z^{(2l-1)} - \ldots \mp \overline{z}^{(l-1)}z^{(l+1)} \pm \frac{1}{2}\overline{z}^{(l)}z^{(l)}\Big)\\ \operatorname{Im}\dot{\overline{z}}z^{(2l+1)} &= \operatorname{Im}\frac{d}{dt}\Big(\dot{\overline{z}}z^{(2l)} - \ddot{\overline{z}}z^{(2l-1)} + \ldots \mp \overline{z}^{(l)}z^{(l+1)}\Big).\end{aligned}$$

Therefore, the left-hand side of (70) can be written as the total derivative of a function $\varepsilon\mathcal{H}[\mathbf{z}](\tau)$, so that (70) becomes

$$\varepsilon\,\frac{d}{d\tau}\mathcal{H}[\mathbf{z}] = \sum_{\|\mathbf{k}\|\le K}{\sum_{|j|\le M}}' \phi(h\omega_j)\,\big(\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})z_{-j}^{-\mathbf{k}} + \varepsilon\dot{z}_{-j}^{-\mathbf{k}}\big)d_j^{\mathbf{k}}. \qquad (71)$$

As in Theorem 5, the Cauchy-Schwarz inequality and the estimates for $z_j^{\mathbf{k}}$ and $d_j^{\mathbf{k}}$ then yield the following estimate.

Theorem 6 *Under the conditions of Theorem 4 we have, for $\tau \le 1$,*

$$\Big|\frac{d}{d\tau}\mathcal{H}[\mathbf{z}](\tau)\Big| \le C\,\varepsilon^{N+1}.$$

7.4 Relationship with actions, momentum, and energy

We now show that the almost-invariant $\mathcal{J}_{\langle\ell\rangle}$ of the modulated Fourier expansion is close to the corresponding harmonic action (13) of the numerical solution,

$$J_\ell = I_\ell + I_{-\ell} = 2I_\ell \quad \text{for} \quad 0 < \ell < M, \quad J_0 = I_0, \quad J_M = I_M,$$

and that $\mathcal{H}$ and $\mathcal{K}$ are close to the Hamiltonian H_M and the momentum K of (12) and (13), respectively.

Theorem 7 *Under the conditions of Theorem 3, along the numerical solution (q^n, p^n) of (19) and the associated modulation sequence $\mathbf{z}(\varepsilon t)$, it holds that*

$$\begin{aligned}\mathcal{H}[\mathbf{z}](\varepsilon t_n) &= H_M\big(q^n, p^n\big) + \mathcal{O}(\varepsilon^3)\\ \mathcal{K}[\mathbf{z}](\varepsilon t_n) &= K\big(q^n, p^n\big) + \mathcal{O}(\varepsilon^3) + \mathcal{O}(\varepsilon^2 M^{-s})\\ \mathcal{J}_{\langle\ell\rangle}[\mathbf{z}](\varepsilon t_n) &= J_\ell\big(q^n, p^n\big) + \gamma_\ell(t_n)\,\varepsilon^3\end{aligned}$$

with $\sum_{\ell=0}^{M}\omega_\ell^{2s+1}\gamma_\ell(t_n) \le C$ for $t_n \le \varepsilon^{-1}$. All appearing constants are independent of ε, M, h, and n.

Proof With the identities (67) we obtain from (66) that

$$\mathcal{K}[\mathbf{z}] = \sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' j \frac{\phi(h\omega_j)}{\psi(h\omega_j)} \Big((\mathbf{k}\cdot\boldsymbol{\omega})\,\mathrm{sinc}\,(h\,\mathbf{k}\cdot\boldsymbol{\omega}) |z_j^{\mathbf{k}}|^2 + 2\varepsilon c_{2\mathbf{k}}\,\mathrm{Im}\,\big(z_{-j}^{-\mathbf{k}} \dot z_j^{\mathbf{k}}\big) + \dots \Big).$$

Separating the terms with $\mathbf{k} = \pm\langle j\rangle$ and using the symplecticity condition (20), and applying the bounds (50) and (51) to the remaining terms, we find

$$\mathcal{K}[\mathbf{z}] = {\sum_{|j|\le M}}' j\,\omega_j \left(|z_j^{\langle j\rangle}|^2 - |z_j^{-\langle j\rangle}|^2 \right) + \mathcal{O}(\varepsilon^3).$$

In terms of the Fourier coefficients of the modulated Fourier expansion $\widetilde q_j(t) = \sum_{\|\mathbf{k}\|\le K} z_j^{\mathbf{k}}(\varepsilon t)\,\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}$, we have at $t = t_n$

$$\begin{aligned}\mathcal{K}[\mathbf{z}] &= {\sum_{|j|\le M}}' j\,\frac{\omega_j}{4}\Big(\big|\widetilde q_j + (\mathrm{i}\omega_j)^{-1}\widetilde p_j\big|^2 - \big|\widetilde q_j - (\mathrm{i}\omega_j)^{-1}\widetilde p_j\big|^2 \Big) + \mathcal{O}(\varepsilon^3) \\ &= K(\widetilde q, \widetilde p) + \mathcal{O}(\varepsilon^3) + \mathcal{O}(\varepsilon^2 M^{-s}) \\ &= K(q^n, p^n) + \mathcal{O}(\varepsilon^3) + \mathcal{O}(\varepsilon^2 M^{-s}),\end{aligned}$$

where we have used (31). The $\mathcal{O}(\varepsilon^2 M^{-s})$ terms come from the boundary terms in the sum. The last equality is a consequence of the remainder bound of Theorem 4.

Similarly, we obtain from (70) that

$$\mathcal{H}[\mathbf{z}] = \sum_{\|\mathbf{k}\|\le K} {\sum_{|j|\le M}}' (\mathbf{k}\cdot\boldsymbol{\omega}) \frac{\phi(h\omega_j)}{\psi(h\omega_j)} \Big((\mathbf{k}\cdot\boldsymbol{\omega})\,\mathrm{sinc}\,(h\,\mathbf{k}\cdot\boldsymbol{\omega}) |z_j^{\mathbf{k}}|^2 + \dots \Big) + \mathcal{U}(\Phi\mathbf{z}),$$

which yields, using in addition $\mathcal{U}(\Phi\mathbf{z}) = \mathcal{O}(\varepsilon^3)$,

$$\mathcal{H}[\mathbf{z}] = {\sum_{|j|\le M}}' \omega_j^2 \left(|z_j^{\langle j\rangle}|^2 + |z_j^{-\langle j\rangle}|^2 \right) + \mathcal{O}(\varepsilon^3),$$

and shows that $\mathcal{H}[\mathbf{z}] = H_M(q^n, p^n) + \mathcal{O}(\varepsilon^3)$.

The result for J_ℓ is obtained in the same way, using in addition Lemma 3 of [6] to estimate the remainder terms. □

With an identical argument to that of [6, Section 4.5], Theorems 5 and 6 together with the estimates of Theorem 4 and the Lipschitz continuity (52) yield the statement of Theorem 3 by patching together many intervals of length ε^{-1}.

8 The Störmer–Verlet/leapfrog discretization

The leapfrog discretization of (11) reads, in the two-step formulation,

$$q^{n+1} - 2\,q^n + q^{n-1} = h^2\big(-\Omega^2 q^n + f(q^n)\big), \tag{72}$$

with the velocity approximation p^n given by

$$2h\,p^n = q^{n+1} - q^{n-1}. \tag{73}$$

The starting value is chosen as $q^1 = q^0 + hp^0 + \frac{h^2}{2}\,f(q^0)$. Conservation properties of this method will be obtained by reinterpreting it as a trigonometric method (17) with modified frequencies $\widehat{\omega}_j$ satisfying $1 - \frac{1}{2}h^2\omega_j^2 = \cos(h\widehat{\omega}_j)$, that is,

$$\sin\Big(\frac{1}{2}\,h\widehat{\omega}_j\Big) = \tfrac{1}{2}h\omega_j. \tag{74}$$

This is possible as long as $h\omega_j \le 2$.

Theorem 8 *Under the stepsize restriction $h\omega_M \le c < 2$, under the non-resonance conditions (24) and (26) for the modified frequencies $\widehat{\omega}_j$ of (74), and under the assumption (15) of small initial data with $s \ge \sigma + 1$ for $(q^0, p^0) = \big(q(0), p(0)\big)$, the near-conservation estimates*

$$\begin{aligned}
\frac{|H_M(q^n,p^n) - H_M(q^0,p^0)|}{\varepsilon^2} &\le C(\varepsilon + h^2) \\
\frac{|K(q^n,p^n) - K(q^0,p^0)|}{\varepsilon^2} &\le C\big(\varepsilon + h^2 + M^{-s} + \varepsilon t M^{-s+1}\big) \\
\sum_{\ell=0}^{M} \omega_\ell^{2s-1}\,\frac{|I_\ell(q^n,p^n) - I_\ell(q^0,p^0)|}{\varepsilon^2} &\le C(\varepsilon + h^2)
\end{aligned}$$

for energy, momentum and actions hold for long times

$$0 \le t = nh \le \varepsilon^{-N+1}$$

with a constant C which depends on s, N, C_0, and c, but is independent of ε, M, h, and t.

Proof Denoting by $\widehat{\Omega}$ the diagonal matrix with entries $\widehat{\omega}_j$, we introduce the transformed variables

$$\widehat{q}^n = \operatorname{sinc}\big(h\widehat{\Omega}\big)\,q^n, \qquad \widehat{p}^n = p^n,$$

which are solutions to the symplectic trigonometric method (17)-(18) with $\psi = \operatorname{sinc}$ and $\phi = 1$. Under the stepsize restriction $h\omega_M \le c < 2$

the non-resonance condition (25) is trivially satisfied for $\widehat{\omega}_j$, and we have

$$\omega_j \leq \widehat{\omega}_j \leq C\omega_j,$$

where C depends only on c. Hence, the assumption (15) of small initial data is satisfied with the same exponent s for the weighted norms defined with $\widehat{\omega}_j$ or ω_j. We can therefore apply Theorem 3 in the transformed variables $(\widehat{q}^n, \widehat{p}^n)$. With the estimate $|\operatorname{sinc}(h\widehat{\omega}_j) - 1| \leq \frac{1}{6}h^2\widehat{\omega}_j^2$, the result stated for the original variables (q^n, p^n) then follows. □

We apply the leapfrog method to the problem of Section 4 with stepsize $h = 0.3$, so that the CFL number $h\omega_M \approx 1.92$ is close to the linear stability limit. In Figure 4 we observe oscillations with large relative amplitude proportional to $h^2\omega_j^2$ for the actions I_j corresponding to high frequencies, but no drift in actions, energy, and momentum.

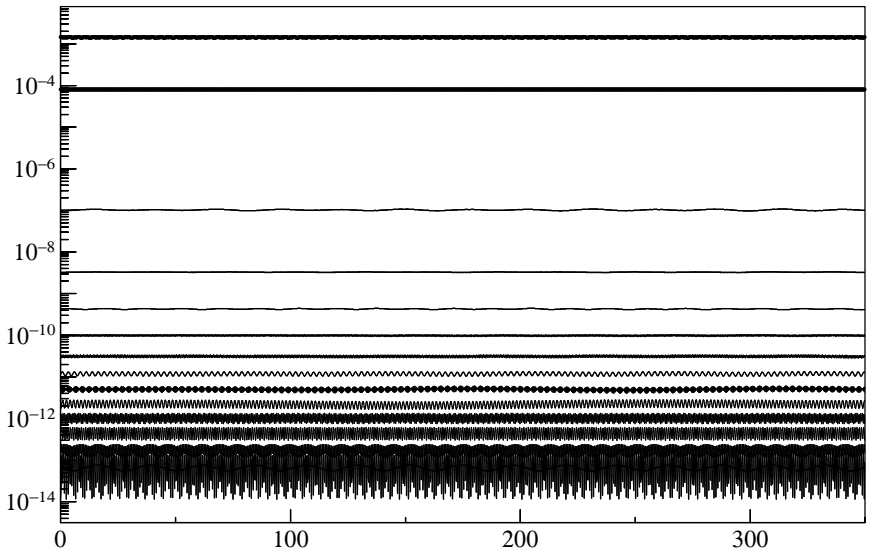


Fig. 4. Actions, energy, and momentum along the numerical solution of the leapfrog method, every 5th action is plotted.

Acknowledgement

This work was partially supported by the Fonds National Suisse, project No. 200020-113249/1, and by DFG, Project LU 532/4-1. A large part of this work was carried out when the authors visited the Isaac Newton Institute in Cambridge.

References

1. D. Bambusi, *Birkhoff normal form for some nonlinear PDEs*, Comm. Math. Phys. **234** (2003), 253–285.
2. G. Benettin and A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys. **74** (1994), 1117–1143.
3. J. Bourgain, *Construction of approximative and almost periodic solutions of perturbed linear Schrödinger and wave equations*, Geom. Funct. Anal. **6** (1996), 201–230.
4. B. Cano, *Conserved quantities of some Hamiltonian wave equations after full discretization*, Numer. Math. **103** (2006), 197–223.
5. D. Cohen, E. Hairer, and C. Lubich, *Numerical energy conservation for multi-frequency oscillatory differential equations*, BIT **45** (2005), 287–305.
6. ———, *Long-time analysis of nonlinearly perturbed wave equations via modulated Fourier expansions*, Arch. Ration. Mech. Anal. **187** (2008), 341–368.
7. P. Deuflhard, *A study of extrapolation methods based on multistep schemes without parasitic solutions*, Z. Angew. Math. Phys. **30** (1979), 177–189.
8. G. Dujardin and E. Faou, *Normal form and long time analysis of splitting schemes for the linear Schrödinger equation with small potential*, Numer. Math. **108** (2007), 223–262.
9. B. García-Archilla, J. M. Sanz-Serna, and R. D. Skeel, *Long-time-step methods for oscillatory differential equations*, SIAM J. Sci. Comput. **20** (1999), 930–963.
10. H. Grubmüller, H. Heller, A. Windemuth, and P. Tavan, *Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions*, Mol. Sim. **6** (1991), 121–142.
11. E. Hairer and C. Lubich, *The life-span of backward error analysis for numerical integrators*, Numer. Math. **76** (1997), 441–462, Erratum: http://www.unige.ch/math/folks/hairer/.
12. ———, *Long-time energy conservation of numerical methods for oscillatory differential equations*, SIAM J. Numer. Anal. **38** (2001), 414–441.
13. ———, *Spectral semi-discretisations of nonlinear wave equations over long times*, Foundations of Comput. Math. (2008).
14. E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*, 2nd ed., Springer Series in Computational Mathematics 31, Springer-Verlag, Berlin, 2006.
15. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations I. Nonstiff problems*, 2nd ed., Springer Series in Computational Mathematics 8, Springer, Berlin, 1993.
16. S. Reich, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal. **36** (1999), 1549–1570.
17. M. Tuckerman, B. J. Berne, and G. J. Martyna, *Reversible multiple time scale molecular dynamics*, J. Chem. Phys. **97** (1992), 1990–2001.