Plane wave stability of the split-step Fourier method for the nonlinear Schrödinger equation

Erwan Faou [1,2] Ludwig Gauckler [3] Christian Lubich [4]

Version of 2 Dezember 2013

Abstract

Plane wave solutions to the cubic nonlinear Schrödinger equation on a torus have recently been shown to behave orbitally stable. Under generic perturbations of the initial data that are small in a high-order Sobolev norm, plane waves are stable over long times that extend to arbitrary negative powers of the smallness parameter. The present paper studies the question as to whether numerical discretizations by the split-step Fourier method inherit such a generic long-time stability property. This can indeed be shown under a condition of linear stability and a non-resonance condition. They can both be verified if the time step-size is restricted by a CFL condition in the case of a constant plane wave. The proof first uses a Hamiltonian reduction and transformation and then modulated Fourier expansions in time. It provides detailed insight into the structure of the numerical solution.

Mathematics Subject Classification (2010): Primary 65P10, 65P40; secondary: 65M70.

1 Introduction

We consider the *cubic nonlinear Schrödinger* equation

$$\mathrm{i}\frac{\partial}{\partial t}u = -\Delta u + \lambda|u|^2 u, \qquad u = u(x,t) \tag{1}$$

in the defocusing ($\lambda = +1$) or focusing case ($\lambda = -1$). We impose periodic boundary conditions in arbitrary spatial dimension $d \geq 1$: the spatial variable x belongs to the d-dimensional torus $\mathbb{T}^d = \mathbb{R}^d/(2\pi\mathbb{Z})^d$.

This nonlinear Schrödinger equation has a class of simple solutions, the *plane wave solutions*

$$u(x,t) = \rho\, \mathrm{e}^{\mathrm{i}(\ell\cdot x - \omega t)} \tag{2}$$

for $\rho \geq 0$, $\ell \in \mathbb{Z}^d$ and $\omega = |\ell|^2 + \lambda\rho^2$, where $\ell \cdot x = \ell_1 x_1 + \cdots + \ell_d x_d$ and $|\ell|^2 = \ell \cdot \ell$. A natural question is whether these plane wave solutions (2) are stable under small

[1]INRIA and ENS Cachan Bretagne, Avenue Robert Schumann, F-35170 Bruz, France (Erwan.Faou@inria.fr).

[2]Département de mathématiques et applications, École normale supérieure, 45 rue d'Ulm, F-75230 Paris Cedex 05, France.

[3]Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, D-10623 Berlin, Germany (gauckler@math.tu-berlin.de).

[4]Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (lubich@na.uni-tuebingen.de).

perturbations of the initial value. In this context it is common knowledge that a *linear stability analysis*, where one examines the eigenvalues of the linearization of the nonlinear Schrödinger equation (1) around a plane wave, leads to the condition $1 + 2\lambda\rho^2 \geq 0$ for (linear) stability, see for instance [1, Sect. 5.1.1]. Since nonlinear effects are ignored, the validity of such a linear stability analysis is inherently restricted to a short time interval. Stability and instability on long time intervals of plane waves in the exact solution is discussed in the recent papers [7] and [18], respectively. Of particular importance for the present paper is [7], where orbital stability over long times is shown for perturbations in high-order Sobolev spaces. Orbital stability means that the solution stays close to the orbit (2).

From the viewpoint of numerical analysis, it is of interest whether (and if so why) a numerical method shares the stability or instability of the exact solution near plane waves. This is the topic of the present paper.

This problem can be traced back to the seminal paper [25] by Weideman & Herbst from 1986. In that paper, conditions on the discretization parameters for various numerical methods are derived that ensure that the numerical solution shares the *linear stability* of the exact solution. This is done by examining the eigenvalues of the linearization around a plane wave of a numerical method applied to (1). Such a linear stability analysis has recently been extended to different numerical methods [4, 6, 21, 22].

In the present paper, we take up this line of research. In contrast to previous work [4, 6, 21, 22, 25], however, we are interested in the *long-time behaviour* of a numerical method near plane waves, and hence a linear stability analysis is of limited use. We pursue the question as to whether the remarkably stable behaviour on long time intervals of the exact solution near plane waves [7] is shared by one of the most popular numerical methods for the nonlinear Schrödinger equation, the *split-step Fourier method* [19]. This method combines a Fourier collocation in space with a Strang splitting in time, see Subsect. 2.1. It integrates plane wave solutions (2) exactly. Our main result states that the long-time orbital stability of the exact solution near plane waves transfers to the numerical solution, see Subsect. 2.2 for a precise statement. In the case of a spatially constant plane wave ($\ell = 0$ in (2)), the case considered by Weideman & Herbst [25], it is further shown that the assumptions of this main result essentially hold under a CFL condition on the discretization parameters, see Subsect. 2.3.

The long-time stability result of the present paper deals with the *completely resonant* equation (1): the eigenvalues (frequencies) of the linear part of the equation are $|j|^2$, $j \in \mathbb{Z}^d$, whose integer linear combinations may vanish identically. This is in marked contrast to previous long-time stability results for numerical discretizations of nonlinear Hamiltonian partial differential equations that consider non-resonant situations. See [9, 10, 11, 14] for the split-step Fourier method applied to the nonlinear Schrödinger equation, where (1) is considered with an additional (generic but artificial) convolution term $V \star u$ in order to have non-resonant frequencies. Another feature of the result in the present paper is that it covers a much larger class of *initial values that are not small* than the aforementioned previous stability results that all deal with small initial values.

The proof of our stability result is given in Sects. 3–5. We first eliminate, in Sect. 3, the principal Fourier mode from the numerical scheme with a sequence of transformations and reductions. The resulting system of equations has small initial values. This enables us to use the technique of modulated Fourier expansions for its long-time analysis, see Sect. 4. It is likely that normal form techniques in the spirit of [9, 10] would lead to similar conclusions, but we have not worked out the details. In

order to obtain results that are valid on long time intervals, the frequencies have to satisfy a certain non-resonance condition. In fact, the completely resonant frequencies of the nonlinear Schrödinger equation are modified during the transformations of Sect. 3, and we are able to verify a non-resonance condition for the new frequencies in the final Sect. 5.

2 Numerical method and statement of the main results

2.1 The split-step Fourier method

We discretize the nonlinear Schrödinger equation (1) with the split-step Fourier method as introduced in [19, 24, 25]. In this method, the equation is discretized in space by a spectral collocation method and in time by a splitting integrator.

Discretization in space. For the discretization in space we make the ansatz

$$u_K(x,t) = \sum_{j\in\mathcal{K}} u_j(t)\mathrm{e}^{\mathrm{i}(j\cdot x)} \qquad \text{with} \qquad \mathcal{K} := \{-K,\dots,K-1\}^d$$

with the spatial discretization parameter K. For fixed t, $u_K(\cdot,t)$ is a *trigonometric polynomial* which is uniquely determined by its values in the collocation points $x_j = \pi j/K$, $j \in \mathcal{K}$. Requiring that the ansatz u_K fulfills the nonlinear Schrödinger equation (1) in the collocation points leads to the equation

$$\mathrm{i}\frac{\partial}{\partial t}u_K = -\Delta u_K + \lambda\mathcal{Q}\big(|u_K|^2 u_K\big), \qquad u_K(\cdot,0) = \mathcal{Q}(u(\cdot,0)), \tag{3}$$

where the trigonometric interpolation $\mathcal{Q}(u)$ (with respect to the spatial variable x) of a function $u(x) = \sum_{k\in\mathbb{Z}^d} u_k \mathrm{e}^{\mathrm{i}(k\cdot x)}$ is the uniquely determined trigonometric polynomial that interpolates u in the collocation points. This trigonometric interpolation is given by

$$\mathcal{Q}(u) = \sum_{j\in\mathcal{K}} \widetilde{u}_j \mathrm{e}^{\mathrm{i}(j\cdot x)} \qquad \text{with} \qquad \widetilde{u}_j = \sum_{k\in\mathbb{Z}^d : k\equiv j \bmod 2K} u_k,$$

where the congruence modulo $2K$ has to be understood entrywise.

Discretization in time. Equation (3) is then discretized in time by a splitting integrator with time step-size h. For this purpose we split (3) in its linear and its nonlinear part,

$$\mathrm{i}\frac{\partial}{\partial t}u_K = -\Delta u_K \qquad \text{and} \qquad \mathrm{i}\frac{\partial}{\partial t}u_K = \lambda\mathcal{Q}\big(|u_K|^2 u_K\big).$$

Denoting by Φ^h_{linear} and $\Phi^h_{\text{nonlinear}}$ the flows over a time h of these equations, we compute approximations u_K^n to $u_K(\cdot,t_n)$ at discrete times $t_n = nh$ by

$$u_K^{n+1} = \Phi^h_{\text{linear}} \circ \Phi^h_{\text{nonlinear}}(u_K^n). \tag{4a}$$

The initial value u_K^0 is chosen as

$$u_K^0 = u_K(\cdot,0) = \mathcal{Q}(u(\cdot,0)). \tag{4b}$$

Equations (4) provide a fully discrete scheme for the numerical solution of the nonlinear Schrödinger equation (1), the *split-step Fourier method*. Since its introduction in [19] it has become a widely used and well analysed method, see for example [1, 8, 11, 20, 24, 25] and references therein.

Computational aspects. In (4), both flows Φ^h_{linear} and $\Phi^h_{\text{nonlinear}}$ can be computed exactly in an efficient way. The flow of the linear equation is given in terms of the Fourier coefficients u_j of a trigonometric polynomial $u(x) = \sum_{j\in\mathcal{K}} u_j \mathrm{e}^{\mathrm{i}(j\cdot x)}$ by

$$\Phi^h_{\text{linear}}(u) = \sum_{j\in\mathcal{K}} \mathrm{e}^{-\mathrm{i}|j|^2 h} u_j \mathrm{e}^{\mathrm{i}(j\cdot x)}. \tag{5}$$

Thus, it can be computed easily in terms of these Fourier coefficients. On the other hand, the flow of the nonlinear equation is given by

$$\Phi^h_{\text{nonlinear}}(u) = \mathcal{Q}\Big(\mathrm{e}^{-\mathrm{i}\lambda|u|^2 h} u\Big), \tag{6}$$

i.e., $\Phi^h_{\text{nonlinear}}(u)(x_j) = \mathrm{e}^{-\mathrm{i}\lambda|u(x_j)|^2 h} u(x_j)$ for all $j \in \mathcal{K}$. This is easy to compute in terms of the function values in the collocation points. Note that the fast Fourier transform provides an efficient tool to switch from Fourier coefficients to function values in the collocation points and vice-versa. The computational cost per time step is thus of order $K^d \log K^d$.

Plane waves in the split-step Fourier method. The split-step Fourier method (4) has *plane wave solutions*

$$u^n_K(x) = \rho \mathrm{e}^{\mathrm{i}(\ell\cdot x - \omega t_n)} \qquad \text{for} \qquad u^0_K(x) = \rho \mathrm{e}^{\mathrm{i}(\ell\cdot x)} \tag{7}$$

with $\omega = |\ell|^2 + \lambda\rho^2$ and $\ell \in \mathcal{K}$. In other words, the plane wave solutions $\rho \mathrm{e}^{\mathrm{i}(\ell\cdot x - \omega t)}$ (2) of the nonlinear Schrödinger equation (1) are integrated exactly by the split-step Fourier method if $\ell \in \mathcal{K}$. It is the stability of these plane wave solutions (7) under perturbations of the initial value that we are interested in.

2.2 Long-time orbital stability

For the study of the stability of plane wave solutions (7), with fixed vector $\ell \in \mathcal{K}$, we impose the following assumptions (with constants that do not depend on the discretization parameters h and K).

Assumption 1. We assume that the time step-size h and the spatial discretization parameter K fulfill together with $\rho \geq 0$ (which will be chosen later as the L_2-norm of the initial value)

$$\big(\cos(n(j)h) - h\lambda\rho^2 \sin(n(j)h)\big)^2 \leq 1 - c_1 h^2 \qquad \text{for all} \qquad j \in \mathcal{Z} := \mathcal{K} \setminus \{0\} \tag{8}$$

with a positive constant c_1, where

$$n(j) = \tfrac{1}{2}|\ell + j \bmod 2K|^2 + \tfrac{1}{2}|\ell - j \bmod 2K|^2 - |\ell|^2. \tag{9}$$

Assumption 1 ensures that the *frequencies*

$$\omega_j = \tfrac{1}{2}|\ell + j \bmod 2K|^2 - \tfrac{1}{2}|\ell - j \bmod 2K|^2 + \frac{\arccos\big(\cos(n(j)h) - h\lambda\rho^2 \sin(n(j)h)\big)}{h \operatorname{sgn}\big(\sin(n(j)h) + h\lambda\rho^2 \cos(n(j)h)\big)} \tag{10}$$

are well defined for all $j \in \mathcal{Z}$. These frequencies show up after a linearization of the split-step Fourier method around a plane wave. This linearization has eigenvalues $\mathrm{e}^{-\mathrm{i}\omega_j h}$, see Sect. 3.

As a second assumption we need a *non-resonance condition*. Ideally we would like to impose this condition directly on the frequencies ω_j. For the verification of the non-resonance condition, however, it turns out to be appropriate to consider modifications of these frequencies.

Assumption 2. We assume that the time step-size h, the spatial discretization parameter K and $\rho \geq 0$ are chosen such that there exist *modified frequencies* ϖ_j, $j \in \mathcal{Z}$, with the following properties for some $N \geq 2$:

(a) The modified frequencies are close to the frequencies ω_j (10),

$$|\varpi_j - \omega_j| \leq \widehat{\varepsilon} \qquad \text{for all} \qquad j \in \mathcal{Z}$$

with a small parameter $\widehat{\varepsilon}$.

(b) There exist positive constants c_2, δ_2 and s_2 such that the following holds for all vectors $(k_j)_{j\in\mathcal{Z}} \in \mathbb{Z}^{\mathcal{Z}}$ of integers with $0 < \sum_{j\in\mathcal{Z}} |k_j| \leq N+1$ and with $k_j \neq 0$ only if $k_l = 0$ for all indices $l \neq j$ with $\varpi_l = \varpi_j$: if

$$\delta := \left| \frac{\mathrm{e}^{\mathrm{i}(\sum_{j\in\mathcal{Z}} k_j \varpi_j)h} - 1}{h} \right| \leq \delta_2,$$

then for all $l \in \mathcal{Z}$ satisfying $k_l \neq 0$,

$$\frac{|l|^4}{\prod_{j\in\mathcal{Z}} |j|^{2|k_j|}} \leq c_2 \delta^{N/s_2}.$$

(c) Complete resonances among the modified frequencies, i.e., $h \sum_{j\in\mathcal{Z}} k_j \varpi_j \in 2\pi\mathbb{Z}$ for a vector $(k_j)_{j\in\mathcal{Z}}$ of integers with $\sum_{j\in\mathcal{Z}} |k_j| \leq N+1$, can only occur if

$$\sum_{\substack{j\in\mathcal{Z} \\ n(j)=m}} k_j = 0 \qquad \text{for all} \qquad m \in \mathbb{Z}.$$

Under these assumptions we will prove the following main result. Here we denote, for a trigonometric polynomial $u(x) = \sum_{j\in\mathcal{K}} u_j \mathrm{e}^{\mathrm{i}(j\cdot x)}$, by

$$\|u\|_s^2 = |u_0|^2 + \sum_{j\in\mathcal{K}} |j|^{2s} |u_j|^2$$

its Sobolev H^s-norm. We further denote by

$$\mathcal{F}_{\neg\ell}(u) = \mathcal{Q}\big(\mathrm{e}^{-\mathrm{i}(\ell\cdot x)} u - u_\ell\big) = \sum_{0\neq j\in\mathcal{K}} u_{j+\ell \bmod 2K} \mathrm{e}^{\mathrm{i}(j\cdot x)}$$

the same function with the ℓth Fourier coefficient set to zero, followed by a shift of Fourier coefficients by ℓ and a trigonometric interpolation. Note that $\|\mathcal{F}_{\neg\ell}(u)\|_s$ measures the size of those Fourier coefficients whose subscript differs from ℓ modulo $2K$.

Theorem 2.1. *Fix an index $\ell \in \mathcal{K}$, an integer $N \geq 2$ and positive numbers c_1, c_2, s_2, δ_2 and ρ_1. There exist s_0 and C such that for every $s \geq s_0$ there exists $\varepsilon_0 > 0$ such that the following holds: If the time step-size h, the spatial discretization parameter K and $\rho \leq \rho_1$ fulfill Assumptions 1 and 2 with some $\widehat{\varepsilon} \leq \varepsilon_0$ (and with the prescribed constants c_1, c_2, s_2, δ_2), then for every initial value u_K^0 with*

$$\big\|u_K^0\big\|_0 = \rho \qquad \text{and} \qquad \big\|\mathcal{F}_{\neg\ell}(u_K^0)\big\|_s \leq \varepsilon \leq \varepsilon_0$$

we have the long-time stability estimate

$$\big\|\mathcal{F}_{\neg\ell}(u_K^n)\big\|_s \leq C\varepsilon \qquad \text{for} \qquad 0 \leq t_n = nh \leq \max(\varepsilon, \widehat{\varepsilon})^{-N/2}.$$

The proof of this theorem will be given in Sects. 3–4. Theorem 2.1 states that—under suitable assumptions—initial values that are close to a plane wave lead to numerical solutions that remain close to a plane wave for a long time, i.e., the numerical solution is concentrated in a single Fourier mode over long times. The closeness is measured by the Sobolev H^s-norm of $\mathcal{F}_{\neg\ell}(u)$. This implies long-time *orbital stability* in H^s, i.e., the numerical solution stays close to the orbit (7), see [7, Subsect. 3.4].

The bounds s_0, C and ε_0 are independent of the discretization parameters h and K subject to Assumptions 1 and 2 and of the small parameters ε and $\widehat{\varepsilon}$. In more detail, the proof of Theorem 2.1 shows that s_0 depends only on d and s_2; C depends only on c_1 and ρ_1; and ε_0 depends on c_1, c_2, d, ℓ, N, s, s_2, δ_2 and ρ_1.

Remark 2.2. The conclusion of Theorem 2.1 equally holds if the (Lie-Trotter) splitting (4a) is replaced by its symmetric version, the *Strang splitting*

$$u_K^{n+1} = \Phi_{\text{linear}}^{h/2} \circ \Phi_{\text{nonlinear}}^{h} \circ \Phi_{\text{linear}}^{h/2}(u_K^n).$$

In fact, both numerical schemes differ only by half a time step with the linear flow at the beginning and at the end of the interval of integration. This does not affect the long-time stability. The same remark applies to the other version of the Strang splitting,

$$u_K^{n+1} = \Phi_{\text{nonlinear}}^{h/2} \circ \Phi_{\text{linear}}^{h} \circ \Phi_{\text{nonlinear}}^{h/2}(u_K^n).$$

2.3 Discussion of the assumptions

Assumptions 1 and 2 for Theorem 2.1 exclude two different types of (potential) instabilities that show up on different time scales. Assumption 1, which is derived in [25, Sect. 5] and [22] with a slightly different meaning of ρ, ensures that (numerical) plane wave solutions (7) are *linearly stable*. This means that all eigenvalues of the linearization of the numerical scheme (4) around a plane wave (7) are of modulus one. Eigenvalues of modulus larger than one would lead to an instability right from the start. In contrast, the non-resonance condition of Assumption 2 on the frequencies is crucial for the proof of our *long-time* result. Indeed, the longer the time interval under consideration is, the more the nonlinear interaction becomes relevant, possibly leading to resonance phenomena if the frequencies are resonant or close to resonant.

A non-resonance condition as stated in Assumption 2 is typically required in a long-time analysis of Hamiltonian partial differential equations and their numerical discretizations, see for example [5, 9, 10, 11, 14] for uses and discussions of similar conditions. Note, however, that we do not (and cannot) impose this non-resonance condition on the completely resonant frequencies of the nonlinear Schrödinger equation (1) and its discretization by the split-step Fourier method, but only on the frequencies ω_j of (10) for the linearization around a plane wave.

In Sect. 5 we will prove the following theorem on a sufficient (though not necessary) condition under which Assumptions 1 and 2 hold in the case of a constant plane wave ($\ell = 0$) for many values of $\rho = \|u_K^0\|_0$ and h.

Theorem 2.3. *Let $\ell = 0$, and fix $\rho_0 > 0$ with*

$$1 + 2\lambda\rho_0^2 > 0, \tag{11}$$

$h_0 > 0$ and $N \geq 2$. Then we have the following result.

For every $\gamma > 0$ there exists a subset $\mathcal{P}(\gamma)$ of $[0, \rho_0] \times [0, h_0]$ of Lebesgue measure $|\mathcal{P}(\gamma)| \geq \rho_0 h_0 - \gamma$ such that Assumptions 1 and 2 hold for all $(\rho, h) \in \mathcal{P}(\gamma)$ and all K that satisfy the restriction

$$dhK^2 + 2h\rho_0^2 \leq \frac{\pi}{N+1} \tag{12}$$

with small parameter $\widehat{\varepsilon} = C_2 h^2$ *and constants* $c_1 = c_1(\rho_0)$, $C_2 = C_2(\rho_0)$, $c_2 = c_2(h_0, N, \gamma, \rho_0)$, $\delta_2 = 1$ *and* $s_2 = 5^4 N^5$.

Theorem 2.1 together with Theorem 2.3 is the discrete counterpart of [7, Theorem 1.1]: for $\ell = 0$, the long-time orbital stability of plane waves proven there for the exact solution transfers to the numerical discretization provided that the step-size restriction (12) is fulfilled and that $\rho = \|u_K^0\|_0$ and h belong to a large set. In comparison with the result [7, Theorem 1.1] for the exact solution, our discrete counterpart is valid on a time interval of length $\max(\varepsilon, h^2)^{-N/2}$ instead of $\varepsilon^{-N/2}$, and the value of N is restricted by the step-size restriction (12). These changes are due to the non-resonance condition that is much more involved for the numerical frequencies than for the analytical frequencies.

Let us finally comment on condition (11) in Theorem 2.3. This condition ensures linear stability of (analytical) plane wave solutions (2) to the nonlinear Schrödinger equation (1), i.e., that all eigenvalues of the linearization of the nonlinear Schrödinger equation around a plane wave (2) are real valued [25, 1, 7]. Theorem 2.3 states in particular that this implies linear stability of (numerical) plane wave solutions ((8) in Assumption 1) under the step-size restriction (12). Actually, weaker step-size restrictions that yield linear stability are discussed in detail in [25, Sect. 5], but (12) is sufficient for our long-time result because it allows us to verify Assumption 2. On the other hand, (8) reduces to (11) (with $\rho_0 = \rho$) in the limit $h \to 0$ for fixed K.

For nonzero but small ℓ, the condition of linear stability in Assumption 1 can still be expected to hold under a step-size restriction similar to (12). In this case, the frequencies ω_j differ from those for $\ell = 0$ only for large j (we have $n(j) = |j|^2$ and $|\ell + j \bmod 2K| = |\ell - j \bmod 2K|$ for all j that are not large, and we have $c|j|^2 \le n(j) \le C|j|^2$ for large j). This property can also be used to argue that the non-resonance condition of Assumption 2 can be expected to hold for nonzero but small ℓ, see Remark 5.6 in Sect. 5. In one dimension ($d = 1$), the linear stability of the split-step Fourier method for $\ell \neq 0$ has been recently analysed in detail by Lakoba [22].

2.4 Numerical experiments

We present numerical experiments which illustrate Theorem 2.1 in situations that are not covered by Theorem 2.3. They show in particular that the conditions of Theorem 2.3 are not necessary for the assumptions and conclusions of Theorem 2.1 to hold.

Throughout, we let $\lambda = -1$ and $\rho^2 = 0.4$ such that $1 + 2\lambda\rho^2 > 0$, and hence we have linear stability of plane waves in the exact solution. We consider the nonlinear Schrödinger equation in dimension one ($d = 1$) with an initial value that is chosen randomly such that for $\ell = 0$, $s = 5$ and $\varepsilon = 0.01$

$$\|u_K^0\|_0 = \rho \qquad \text{and} \qquad \|\mathcal{F}_{\neg\ell}(u_K^0)\|_s = \varepsilon,$$

i.e., the initial value is, in the H^5 norm, up to 0.01 close to the constant plane wave ρ.

For the numerical discretization with the split-step Fourier method we use 2^5 points for the Fourier collocation in space ($K = 2^4$), and we consider three different step-sizes for the discretization in time: the step-size $h = 0.04$ that does not fulfill the step-size restriction (12) of Theorem 2.3 and the slightly larger step-sizes $h = 0.042$ and $h = 0.044$.

We have checked numerically for $h = 0.04$ and $h = 0.044$ that Assumptions 1 and Assumption 2 of Theorem 2.1 are fulfilled for $N \le 5$ with $c_1 = 0.2$, $\widehat{\varepsilon} = 0$, $c_2 = 8$,

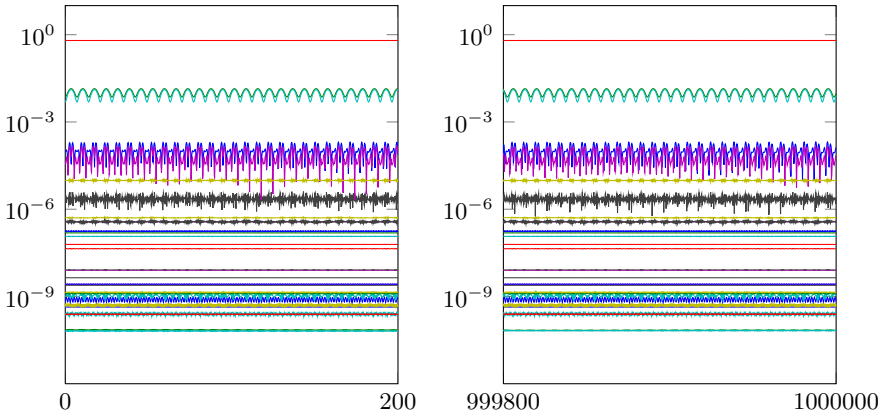

Figure 1: Evolution of the absolute values of the Fourier coefficients for $h = 0.04$.

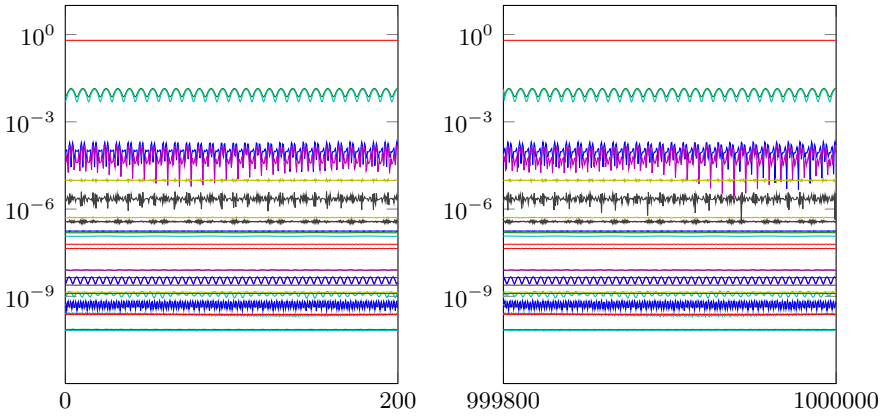

Figure 2: Evolution of the absolute values of the Fourier coefficients for $h = 0.044$.

$\delta_2 = 0.1$ and $s_2 = 5N$ for $h = 0.04$ and $s_2 = 8N/5$ for $h = 0.044$. Note, however, that the step-size restriction (12) of Theorem 2.3 is not fulfilled. For Figure 1 we compute the numerical solution with step-size $h = 0.04$ on a long time interval $t \leq 10^6$ and plot the absolute values of the Fourier coefficients on two subintervals of length 200. The same is done in Figure 2 with the step-size $h = 0.044$. As stated in Theorem 2.1 we observe in both cases that the solution stays concentrated in the ℓth Fourier mode over long times.

For the intermediate step size $h = 0.042$, however, Assumption 1 of Theorem 2.1 is not fulfilled. In Figure 3 we again plot the absolute values of the Fourier coefficients of the numerical solution and clearly observe an instability.

3 Reductions and transformations

From now on we omit the index K of the numerical solution u_K^n, $n = 0, 1, 2, \ldots$. Instead, we denote by u_j^n the jth Fourier coefficient of u^n: $u^n(x) = \sum_{j \in \mathcal{K}} u_j^n \mathrm{e}^{\mathrm{i}(j \cdot x)}$. We work with the numerical scheme (4a) in terms of these Fourier coefficients, which

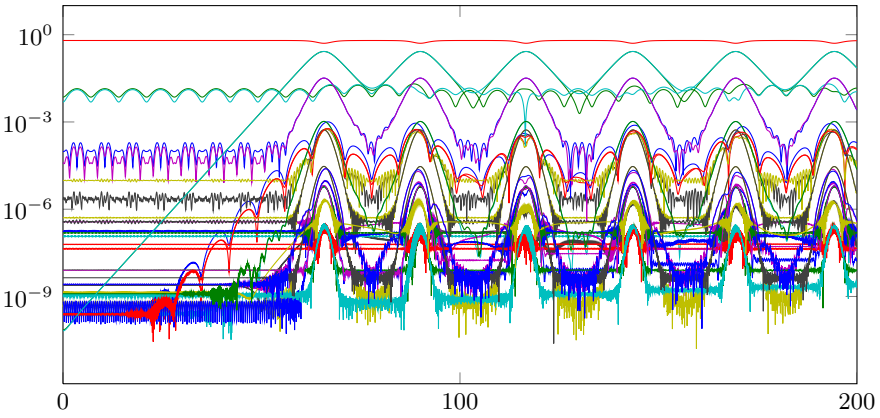

Figure 3: Evolution of the absolute values of the Fourier coefficients for $h = 0.042$.

takes the form (see (5) and (6))

$$u_j^{n+1} = \mathrm{e}^{-\mathrm{i}|j|^2 h} \sum_{m=0}^{\infty} \frac{(-\mathrm{i}h\lambda)^m}{m!} \sum_{\substack{k^1+\cdots+k^{m+1} \\ -l^1-\cdots-l^m \equiv j \bmod 2K}} u_{k^1}^n \cdots u_{k^{m+1}}^n \overline{u}_{l^1}^n \cdots \overline{u}_{l^m}^n. \tag{13}$$

The goal of this section is to eliminate the ℓth Fourier mode, which is not small, from u^n. To this end we apply similar reductions and transformations to those for the exact solution in [7, Sect. 2], which can be summarised as follows.

- Transformation $u \leftrightarrow v$ with $u = (u_j)_{j\in\mathcal{K}}$, $v = (v_j)_{j\in\mathcal{K}}$: shift to the case $\ell = 0$, see Subsect. 3.1.
- Transformation $v \leftrightarrow (a, \theta, w)$ with $a, \theta \in \mathbb{R}$ and $w = (w_j)_{j\in\mathcal{K}\setminus\{0\}}$: introduction of polar coordinates (a, θ) for v_0 and rotations w_j of v_j, see Subsect. 3.2.
- Reduction $(a, \theta, w) \leftrightarrow w$: elimination of a and θ using conservation of mass and gauge invariance, see Subsect. 3.2.
- Transformation $w \leftrightarrow \xi$ with $\xi = (\xi_j)_{j\in\mathcal{K}\setminus\{0\}}$: diagonalization of the linear part, see Subsects. 3.3–3.4.

These transformations and reductions are applied directly to the numerical scheme in the form (13). In Subsect. 3.5, we consider them from a different perspective, namely from the perspective of the differential equations that form the two steps of the splitting integrator (4a). Both perspectives will be important in the following Sect. 4.

3.1 Shift to the case $\ell = 0$

We introduce new variables

$$v_j = u_{\ell+j \bmod 2K} \qquad \text{for} \qquad j \in \mathcal{K}.$$

The numerical scheme (13) in the new variables v^n becomes

$$v_j^{n+1} = \mathrm{e}^{-\mathrm{i}|\ell+j \bmod 2K|^2 h} \sum_{m=0}^{\infty} \frac{(-\mathrm{i}h\lambda)^m}{m!} \sum_{\substack{k^1+\cdots+k^{m+1} \\ -l^1-\cdots-l^m \equiv j \bmod 2K}} v_{k^1}^n \cdots v_{k^{m+1}}^n \overline{v}_{l^1}^n \cdots \overline{v}_{l^m}^n. \tag{14}$$

3.2 Elimination of the zero mode

We introduce polar coordinates (a, θ) for v_0,

$$v_0 = a\mathrm{e}^{\mathrm{i}\theta} \qquad \text{with} \qquad a = |v_0|,$$

and new variables w_j, $0 \neq j \in \mathcal{K}$, by

$$v_j = w_j \mathrm{e}^{\mathrm{i}\theta} \qquad \text{for} \qquad j \in \mathcal{Z} = \mathcal{K} \setminus \{0\}. \tag{15}$$

In these new variables (a, θ, w) with $w = (w_j)_{j\in\mathcal{Z}}$ the numerical scheme (14) becomes

$$\begin{aligned} w_j^{n+1} = \mathrm{e}^{-\mathrm{i}|\ell+j \bmod 2K|^2 h} \mathrm{e}^{\mathrm{i}(\theta^n - \theta^{n+1})} \sum_{m=0}^{\infty} \frac{(-\mathrm{i}h\lambda)^m}{m!} \\ \sum_{\substack{k^1+\cdots+k^{m+1} \\ -l^1-\cdots-l^m \equiv j \bmod 2K}} w_{k^1}^n \cdots w_{k^{m+1}}^n \overline{w}_{l^1}^n \cdots \overline{w}_{l^m}^n, \end{aligned} \tag{16}$$

where we use the convention $w_0^n = \overline{w}_0^n = a^n$.

Now, we eliminate a and θ from (16). For the elimination of a we observe that the split-step Fourier method (13) conserves mass,

$$\left\|u^{n+1}\right\|_0 = \sum_{j\in\mathcal{K}} |u_j^{n+1}|^2 = \sum_{j\in\mathcal{K}} |u_j^n|^2 = \left\|u^n\right\|_0,$$

a fact that can be easily derived from the representations (5) and (6) of the flows composing the numerical scheme and the discrete Parseval identity $\sum_{j\in\mathcal{K}} |u_j^n|^2 = (2K)^{-d} \sum_{k\in\mathcal{K}} |u^n(x_k)|^2$. The conservation of mass allows us to express a^n in terms of w_j^n, $j \in \mathcal{Z}$, and $\rho = \|u^0\|_0$,

$$a^n = \Big(\rho^2 - \sum_{j\in\mathcal{Z}} |w_j^n|^2\Big)^{1/2}. \tag{17}$$

Also the factor $\mathrm{e}^{\mathrm{i}(\theta^n - \theta^{n+1})}$ in (16) can be expressed in terms of w_j using (16) for $j = 0$:

$$\mathrm{e}^{\mathrm{i}(\theta^n - \theta^{n+1})} = \frac{\mathrm{e}^{\mathrm{i}|\ell|^2 h}}{w_0^{n+1}} \sum_{m=0}^{\infty} \frac{(\mathrm{i}h\lambda)^m}{m!} \sum_{\substack{k^1+\cdots+k^{m+1} \\ -l^1-\cdots-l^m \equiv 0 \bmod 2K}} \overline{w}_{k^1}^n \cdots \overline{w}_{k^{m+1}}^n w_{l^1}^n \cdots w_{l^m}^n. \tag{18}$$

Hence, $a = w_0 = \overline{w}_0$ and $\mathrm{e}^{\mathrm{i}\theta}$ are determined by w_j, $j \in \mathcal{Z}$. The numerical scheme (16) is therefore completely described by the reduced set of variables $w = (w_j)_{j\in\mathcal{Z}}$.

Now, we can replace $w_0^n = a^n$ in (16) and (18) by (17). Furthermore, we can use (17) with $n+1$ instead of n and $|w_j^{n+1}|^2$ replaced by (16) to replace w_0^{n+1} in (18). For sufficiently small w this leads, after a Taylor expansion of $(\rho^2 - \dots)^{\pm 1/2}$, to an equation for w^{n+1} of the following form (with right-hand side depending only on w^n):

$$\begin{aligned} w_j^{n+1} = \mathrm{e}^{-\mathrm{i}(|\ell+j \bmod 2K|^2 - |\ell|^2)h} \Big((1 - \mathrm{i}h\lambda\rho^2) w_j^n - \mathrm{i}h\lambda\rho^2 \overline{w}_{-j}^n \\ + \sum_{m+m'=2}^{\infty} \sum_{\substack{k^1+\cdots+k^m \\ -l^1-\cdots-l^{m'} \equiv j \bmod 2K}} h\widetilde{Q}_{j,k,l} w_{k^1}^n \cdots w_{k^m}^n \overline{w}_{l^1}^n \cdots \overline{w}_{l^{m'}}^n \Big). \end{aligned} \tag{19}$$

The subscripts here and in the following all belong to the reduced set $\mathcal{Z}$.

3.3 Linear stability and numerical frequencies

The linear part in equation (19) couples w_j to $\overline{w}_{-j}$. This leads us to consider the equation for w_j together with the one for $\overline{w}_{-j}$,

$$\begin{pmatrix} w_j^{n+1} \\ \overline{w}_{-j}^{n+1} \end{pmatrix} = \mathrm{e}^{-\mathrm{i}(|\ell+j \bmod 2K|^2/2-|\ell-j \bmod 2K|^2/2)h} A_j \begin{pmatrix} w_j^{n} \\ \overline{w}_{-j}^{n} \end{pmatrix} + \text{ higher order terms}$$

with the matrix

$$A_j = \begin{pmatrix} \alpha_j & \beta_j \\ \overline{\beta_j} & \overline{\alpha_j} \end{pmatrix},$$

where

$$\begin{aligned} \alpha_j &= (1 - \mathrm{i}h\lambda\rho^2)\mathrm{e}^{-\mathrm{i}(|\ell+j \bmod 2K|^2/2+|\ell-j \bmod 2K|^2/2-|\ell|^2)h}, \\ \beta_j &= -\mathrm{i}h\lambda\rho^2\mathrm{e}^{-\mathrm{i}(|\ell+j \bmod 2K|^2/2+|\ell-j \bmod 2K|^2/2-|\ell|^2)h}. \end{aligned}$$

This matrix has $|\alpha_j|^2 - |\beta_j|^2 = 1$ and its eigenvalues are

$$\lambda_j^{\pm} = \mathrm{Re}(\alpha_j) \pm \mathrm{i}\,\mathrm{sgn}(\mathrm{Im}(\alpha_j))\sqrt{1 - \mathrm{Re}(\alpha_j)^2}.$$

The reason for including $\mathrm{sgn}(\mathrm{Im}(\alpha_j))$ in the definition of the eigenvalues $\lambda_j^{\pm}$ will become clear in the following Subsect. 3.4.

Assumption 1 ensures that $\mathrm{Re}(\alpha_j)^2 \leq 1$, and hence the eigenvalues $\lambda_j^{\pm}$ of A are of modulus one: We have

$$\mathrm{e}^{-\mathrm{i}(|\ell+j \bmod 2K|^2/2-|\ell-j \bmod 2K|^2/2)h}\lambda_j^{\pm} = \mathrm{e}^{\mp\mathrm{i}\omega_j h}$$

with the *numerical frequencies* ω_j from (10),

$$\omega_j = \tfrac{1}{2}|\ell + j \bmod 2K|^2 - \tfrac{1}{2}|\ell - j \bmod 2K|^2 + \frac{\arccos(\mathrm{Re}(\alpha_j))}{-h\,\mathrm{sgn}(\mathrm{Im}(\alpha_j))},$$

where the branch of arccos with values in $[0, \pi]$ is used. Note that eigenvalues of A_j of modulus greater than one would lead to a growth of the corresponding modes in the linearization of (19). Assumption 1 excludes this scenario and thus ensures linear stability of the split-step Fourier method.

3.4 Diagonalization of the linear part

We introduce new variables ξ_j that diagonalize the linear part of (19):

$$\begin{pmatrix} \xi_j \\ \overline{\xi}_{-j} \end{pmatrix} = S_j \begin{pmatrix} w_j \\ \overline{w}_{-j} \end{pmatrix},$$

where, with the notation of the previous subsection,

$$S_j^{-1} = \frac{1}{\sqrt{|\beta_j|^2 - |\lambda_j^+ - \alpha_j|^2}} \begin{pmatrix} \beta_j & \lambda_j^- - \overline{\alpha_j} \\ \lambda_j^+ - \alpha_j & \overline{\beta_j} \end{pmatrix} \tag{20}$$

such that

$$\mathrm{e}^{-\mathrm{i}(|\ell+j \bmod 2K|^2/2-|\ell-j \bmod 2K|^2/2)h} S_j A_j S_j^{-1} = \begin{pmatrix} \mathrm{e}^{-\mathrm{i}\omega_j h} & 0 \\ 0 & \mathrm{e}^{\mathrm{i}\omega_j h} \end{pmatrix}.$$

Note that

$$|\beta_j|^2 - |\lambda_j^+ - \alpha_j|^2 = 2\sqrt{1-\mathrm{Re}(\alpha_j)^2}\Big(|\mathrm{Im}(\alpha_j)| - \sqrt{1-\mathrm{Re}(\alpha_j)^2}\Big) > 0, \tag{21}$$

and hence this change of variables, which defines ξ_j and $\overline{\xi}_j$, is well defined because of the structure of S_j (this is the reason for including the sign of $\mathrm{Im}(\alpha_j)$ in the definition of the numerical frequencies). Moreover, it is symplectic since $\det(S_j) = 1$. With this change of variables, (19) is transformed to

$$\xi_j^{n+1} = \mathrm{e}^{-\mathrm{i}\omega_j h}\xi_j^n + \sum_{m+m'=2}^{\infty} \sum_{\substack{k^1+\cdots+k^m \\ -l^1-\cdots-l^{m'} \equiv j \bmod 2K}} hQ_{j,k,l}\xi_{k^1}^n \cdots \xi_{k^m}^n \overline{\xi}_{l^1}^n \cdots \overline{\xi}_{l^{m'}}^n. \tag{22}$$

3.5 The splitting structure of the numerical scheme in the new variables

Recall that in the original variables u

$$u^{n+1} = \Phi^h_{\mathrm{linear}} \circ \Phi^h_{\mathrm{nonlinear}}(u^n),$$

see equation (4a). Here, $\Phi^h_{\mathrm{linear}} = \Phi^h_{\breve{H}_0}$ is the flow at time h of the Hamiltonian differential equation with Hamiltonian function $\breve{H}_0$,

$$\mathrm{i}\dot{u}_j = |j|^2 u_j = \frac{\partial \breve{H}_0}{\partial \overline{u}_j}(u,\overline{u}) \qquad \text{with} \qquad \breve{H}_0(u,\overline{u}) = \sum_j |j|^2 u_j \overline{u}_j.$$

Correspondingly, $\Phi^h_{\mathrm{nonlinear}} = \Phi^h_{\breve{P}}$ is the flow at time h of the Hamiltonian differential equation with Hamiltonian function $\breve{P}$,

$$\mathrm{i}\dot{u}_j = \frac{\partial \breve{P}}{\partial \overline{u}_j}(u,\overline{u}) \qquad \text{with} \qquad \breve{P}(u,\overline{u}) = \frac{\lambda}{2} \sum_{j^1+j^2-j^3-j^4 \equiv 0 \bmod 2K} u_{j^1} u_{j^2} \overline{u}_{j^3} \overline{u}_{j^4}.$$

Now, we consider the transformations $u \leftrightarrow v \leftrightarrow (a,\theta,w) \leftrightarrow w \leftrightarrow \xi$ from the previous subsections on the level of these differential equations (instead of their flows, as we have done in the previous subsections).

Shift $u \leftrightarrow v$. After the change of variables $u \leftrightarrow v$ described in Subsect. 3.1 (leading to the numerical scheme (14)) the splitting scheme becomes

$$v^{n+1} = \Phi^h_{\breve{H}_0} \circ \Phi^h_{\breve{P}}(v^n)$$

with the Hamiltonian functions

$$\breve{H}_0(v,\overline{v}) = \sum_j |\ell + j \bmod 2K|^2 v_j \overline{v}_j$$

and

$$\breve{P}(v,\overline{v}) = \frac{\lambda}{2} \sum_{j^1+j^2-j^3-j^4 \equiv 0 \bmod 2K} v_{j^1} v_{j^2} \overline{v}_{j^3} \overline{v}_{j^4}.$$

Transformation $v \leftrightarrow (a,\theta,w)$. In the variables (a,θ,w) introduced at the beginning of Subsect. 3.2 (leading to the numerical scheme (16)), the flow of the Hamiltonian differential equation with Hamiltonian function $\breve{H}_0$ has to be replaced by the flow of

$$\mathrm{i}\dot{w}_j = \dot{\theta} w_j + |\ell + j \bmod 2K|^2 w_j. \tag{23}$$

The corresponding equation for $a = w_0$ becomes, after taking the real part,

$$0 = \dot{\theta} a + |\ell|^2 a. \tag{24}$$

Correspondingly, the flow of the Hamiltonian differential equation with Hamiltonian function $\check{P}$ has to be replaced by the flow of

$$\mathrm{i}\dot{w}_j = \dot{\theta} w_j + \frac{\partial \widehat{P}}{\partial \overline{w}_j}(a, \theta, w, \overline{w}) \qquad \text{with} \qquad \widehat{P}(a, \theta, w, \overline{w}) = \check{P}(v, \overline{v}). \tag{25}$$

Note that the function $\widehat{P}$ is actually independent of θ (gauge invariance). The equation for $a = w_0$ becomes, after taking the real part,

$$0 = \dot{\theta} a + \frac{1}{2} \frac{\partial \widehat{P}}{\partial a}(a, \theta, w, \overline{w}). \tag{26}$$

Reduction $(a, \theta, w) \leftrightarrow w$. Solving (24) for $\dot{\theta}$ and inserting this into the equations (23) for $j \neq 0$ shows that (23) becomes, in the reduced set of variables w from Subsect. 3.2 (with the numerical scheme (19)),

$$\mathrm{i}\dot{w}_j = \frac{\partial \widetilde{H}_0}{\partial \overline{w}_j}(w, \overline{w}) \qquad \text{with} \qquad \widetilde{H}_0(w, \overline{w}) = \sum_j \big(|\ell + j \bmod 2K|^2 - |\ell|^2\big) w_j \overline{w}_j.$$

Solving (26) for $\dot{\theta}$ and inserting this into the equations (25) for $j \neq 0$ yields

$$\mathrm{i}\dot{w}_j = \frac{\partial \widehat{P}}{\partial \overline{w}_j}(a, \theta, w, \overline{w}) + \frac{-w_j}{2a} \frac{\partial \widehat{P}}{\partial a}(a, \theta, w, \overline{w}).$$

Using

$$\frac{\partial a}{\partial \overline{w}_j}(w, \overline{w}) = \frac{-w_j}{2a}$$

with a given by (17), we see that the equation (25) becomes, in the reduced set of variables w from Subsect. 3.2 (with the numerical scheme (19)),

$$\mathrm{i}\dot{w}_j = \frac{\partial \widetilde{P}}{\partial \overline{w}_j}(w, \overline{w}) \qquad \text{with} \qquad \widetilde{P}(w, \overline{w}) = \widehat{P}(a, \theta, w, \overline{w}),$$

which surprisingly is again of Hamiltonian form. We hence have

$$w^{n+1} = \Phi^h_{\widetilde{H}_0} \circ \Phi^h_{\widetilde{P}}(w^n).$$

The splitting integrator in the reduced set of variables w is still a Hamiltonian splitting, a splitting into two Hamiltonian equations.

Transformation $w \leftrightarrow \xi$. Concerning the final change of variables $w \leftrightarrow \xi$ of Subsect. 3.4 (leading to the numerical scheme (22)) we note first that the matrix S_j was chosen in such a way that it is symplectic. We therefore end up with

$$\xi^{n+1} = \Phi^h_{H_0} \circ \Phi^h_P(\xi^n) \tag{27a}$$

with

$$H_0(\xi, \overline{\xi}) = \widetilde{H}_0(w, \overline{w}) \qquad \text{and} \qquad P(\xi, \overline{\xi}) = \widetilde{P}(w, \overline{w}). \tag{27b}$$

While it is an obvious observation that the numerical scheme in the new variables ξ is still a splitting scheme, it is highly remarkable that the split equations retain their Hamiltonian structure.

By virtue of the expansion (22), we have a concrete expression for the flow $\Phi_P^h(\xi^n) = \big(\Phi_{H_0}^h\big)^{-1}(\xi^{n+1})$,

$$\Phi_P^h(\xi^n) = S_j \begin{pmatrix} \mathrm{e}^{\mathrm{i}(|\ell+j \bmod 2K|^2-|\ell|^2)h} & 0 \\ 0 & \mathrm{e}^{-\mathrm{i}(|\ell+j \bmod 2K|^2-|\ell|^2)h} \end{pmatrix} S_j^{-1} \begin{pmatrix} \xi_j^{n+1} \\ \overline{\xi}_{-j}^{n+1} \end{pmatrix}. \tag{28}$$

For later purposes we also introduce an expansion

$$P(\xi,\overline{\xi}) = \sum_{m+m'=2}^{\infty} \sum_{k\in\mathcal{Z}^m,\, l\in\mathcal{Z}^{m'}} P_{k,l}\xi_{k^1}\cdots\xi_{k^m}\overline{\xi}_{l^1}\cdots\overline{\xi}_{l^{m'}} \tag{29}$$

of the Hamiltonian function P.

3.6 Estimates for the transformation and for the transformed equation

We derive some bounds for the change of variables $u \leftrightarrow \xi$ described in Subsects. 3.1–3.4. We assume throughout that Assumption 1 is fulfilled.

Note that $|w_j| = |u_{\ell+j \bmod 2K}|$ for $j \in \mathcal{Z}$, and hence we first consider the last transformation $w \leftrightarrow \xi$ of Subsect. 3.4 described by the matrices S_j (20).

Lemma 3.1. *The absolute values of the entries of the matrices S_j and S_j^{-1} are bounded by $\sqrt{1+\rho^2/(2\sqrt{c_1})}$, independently of j and h.*

Proof. With the notations of Subsect. 3.3 we have by (21)

$$\left|\frac{\beta_j}{\sqrt{|\beta_j|^2-|\lambda_j^+-\alpha_j|^2}}\right|^2 = \frac{|\mathrm{Im}(\alpha_j)|+\sqrt{1-\mathrm{Re}(\alpha_j)^2}}{2\sqrt{1-\mathrm{Re}(\alpha_j)^2}},$$
$$\left|\frac{\lambda_j^+-\alpha_j}{\sqrt{|\beta_j|^2-|\lambda_j^+-\alpha_j|^2}}\right|^2 = \frac{|\mathrm{Im}(\alpha_j)|-\sqrt{1-\mathrm{Re}(\alpha_j)^2}}{2\sqrt{1-\mathrm{Re}(\alpha_j)^2}}.$$

This proves the statement of the lemma since $|\mathrm{Im}(\alpha_j)| \le \sqrt{1-\mathrm{Re}(\alpha_j)^2}+h\rho^2$ and since $\sqrt{1-\mathrm{Re}(\alpha_j)^2} \ge \sqrt{c_1}h$ by Assumption 1. □

Now we consider the norm

$$\|\xi\|_s = \Big(\sum_{j\in\mathcal{Z}} |j|^{2s}|\xi_j|^2\Big)^{1/2},$$

i.e., $\|\xi\|_s$ is the Sobolev H^s norm of the function $\sum_{j\in\mathcal{Z}}\xi_j\mathrm{e}^{\mathrm{i}(j\cdot x)}$ as introduced in Subsect. 2.2. The previous Lemma 3.1 implies the following result.

Lemma 3.2. *For the change of variables $u \leftrightarrow \xi$ there exist positive constants $\widehat{c}$ and $\widehat{C}$ depending only on c_1 and an upper bound of ρ such that*

$$\widehat{c}\|\xi\|_s \le \|\mathcal{F}_{\neg\ell}(u)\|_s \le \widehat{C}\|\xi\|_s.$$ □

In particular, the previous lemma shows that the condition $\|\mathcal{F}_{\neg\ell}(u^0)\|_s \le \varepsilon$ of Theorem 2.1 becomes in the new variables ξ

$$\|\xi^0\|_s \le \widehat{c}^{-1}\varepsilon. \tag{30}$$

We finally collect some estimates for the nonlinearity in the numerical scheme written in the new variables ξ as given by (22).

Lemma 3.3. *The nonlinearity given by the coefficients $Q_{j,k,l}$ in (22) satisfies for $s > d/2$*

$$\Big(\sum_{j\in\mathcal{Z}}|j|^{2s}\Big(\sum_{\substack{k^1+\cdots+k^m\\ -l^1-\cdots-l^{m'}\equiv j \bmod 2K}}\big|Q_{j,k,l}\eta^1_{k^1}\cdots\eta^m_{k^m}\eta^{m+1}_{l^1}\cdots\eta^{m+m'}_{l^{m'}}\big|\Big)^2\Big)^{1/2}$$
$$\le C_{m,m',s}\|\eta^1\|_s\cdots\|\eta^{m+m'}\|_s$$

for vectors $\eta^1,\dots,\eta^{m+m'}\in\mathbb{C}^{\mathcal{Z}}$. The constants $C_{m,m',s}$ depend only on m, m', s, c_1 and ρ and satisfy

$$\sum_{m+m'=2}^{\infty} C_{m,m',s}r^{m+m'} \le C$$

for some positive constants r and C depending only on c_1, s and ρ.

Proof. (a) By carefully going through the construction of the coefficients $Q_{j,k,l}$ in Subsects. 3.2 and 3.4 one shows for the coefficients $Q_{j,k,l}$ that there exists a constant C depending on c_1 and ρ such that

$$|Q_{j,k,l}| \le C^{m+m'} \tag{31}$$

for all $j\in\mathcal{Z}$, all $k\in\mathcal{Z}^m$ and all $l\in\mathcal{Z}^{m'}$.

(b) The first estimate of the lemma follows by applying (31) and the Cauchy-Schwarz inequality and by using

$$\sum_{\substack{k^1+\cdots+k^m\\ -l^1-\cdots-l^{m'}\equiv j \bmod 2K}}\left(\frac{|j|}{|k^1|\cdots|k^m||l^1|\cdots|l^{m'}|}\right)^{2s} \le c^{m+m'}$$

with a constant c depending only on $s > d/2$. The second estimate of the lemma then follows also from (31). □

4 Modulated Fourier expansions

In this section we will prove Theorem 2.1 using *modulated Fourier expansions* originally introduced in [15], see also [17]. Throughout we will work with the numerical scheme in the new variables ξ introduced in Sect. 3, see (22).

There are two main steps:

- Construction of a short-time approximation of ξ^n from (22) by a modulated Fourier expansion in Subsects. 4.1–4.5.
- Almost-invariants of the modulated Fourier expansion that allow us to prove a result on a long time interval in Subsects. 4.6–4.9.

For the first main step it is convenient to work with the numerical scheme as given by the composition of flows (22), whereas for the derivation of the almost-invariants it is necessary to switch to the level of the differential equations whose flows compose the numerical scheme (27). We ultimately show that $\|\xi^n\|_s$ stays of order ε for initial values ξ^0 of order ε. This preservation of smallness and regularity of ξ^n is the main ingredient for the final proof of Theorem 2.1 in Subsect. 4.10.

The proof via modulated Fourier expansions given here uses and combines ideas from several previous proofs using such expansions: The aforementioned idea of switching between the flows and the differential equations is loosely based on [11, 14], the construction of the modulated Fourier expansion with an asymptotic expansion is based on [13, 16], the idea of using modified frequencies ϖ_j instead of the original (numerical) frequencies ω_j of (10) for the modulated Fourier expansion is also used in [12], the non-resonance condition in Assumption 2 is used in a similar way to [5], and the use of almost-invariants of the modulated Fourier expansion to prove long-time almost-conservation properties can be traced back to [15].

In the following analysis, the (generic) constants C, s_0 and δ_0 are all independent of the small parameters ε from (30) and $\widehat{\varepsilon}$ from Assumption 2. The constants C and δ_0 will depend on the constants c_1, c_2, s_2, δ_2 and N of Assumptions 1 and 2, on s from (30), on an upper bound of $\rho = \|u^0\|_0$, on the index $\ell \in \mathcal{K}$ from (7) and on the dimension d. The constant s_0 will depend only on d and s_2.

4.1 Resonant modulated Fourier expansion

In order to motivate the modulated Fourier expansion we consider here, let us first have a look at (22) in the linear case (all $Q_{j,k,l} = 0$). In this case, the evolution of the jth mode is given by the multiplication with $\mathrm{e}^{-\mathrm{i}\omega_j t}$. In the presence of the nonlinearity, we seek for an expansion, the modulated Fourier expansion, in terms of products of these exponentials that are multiplied (modulated) by slowly varying coefficients.

There are two pitfalls in the present situation that have to be handled with care. First, it turns out that the frequencies ω_j of (10) are inconvenient when it comes to resonance issues. Therefore we use the modified frequencies ϖ_j of Assumption 2 instead and consider products of the exponentials $\mathrm{e}^{-\mathrm{i}\varpi_j t}$:

$$\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t} \qquad \text{with} \qquad \mathbf{k}\cdot\boldsymbol{\varpi} = \sum_{j\in\mathcal{Z}} k_j \varpi_j$$

for vectors of integers $\mathbf{k} = (k_j)_{j\in\mathcal{Z}} \in \mathbb{Z}^{\mathcal{Z}}$ and the vector $\boldsymbol{\varpi} = (\varpi_j)_{j\in\mathcal{Z}}$ of modified frequencies.

Second, the modified frequencies ϖ_j of Assumption 2 are by definition exactly resonant, for instance $\varpi_j = \varpi_l$ for $|j| = |l|$ in the case $\ell = 0$. Hence, we cannot distinguish all products $\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t}$, and we therefore introduce the *resonance module*

$$\mathcal{M} = \big\{\, \mathbf{k} \in \mathbb{Z}^{\mathcal{Z}} : \mathbf{k}\cdot\boldsymbol{\varpi} = 0,\ j(\mathbf{k}) = 0 \,\big\},$$

where

$$j(\mathbf{k}) = \sum_{l\in\mathcal{Z}} k_l l \bmod 2K.$$

The restriction $j(\mathbf{k}) = 0$ in the definition of the resonance module comes from the fact that the products $\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t}$ are attached to some specific mode ξ_j, namely $j = j(\mathbf{k})$, as we will see in the following.

With these preliminaries, we introduce the *resonant modulated Fourier expansion*

$$\xi_j(t) = \sum_{[\mathbf{k}]} z_j^{[\mathbf{k}]}(\delta t) \mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t} \tag{32}$$

Here,

$$\delta = \max(\varepsilon, \widehat{\varepsilon})^{1/2} \tag{33}$$

is a small parameter and the sum is over all residue classes $[\mathbf{k}] \in \mathbb{Z}^{\mathcal{Z}}/\mathcal{M}$. The coefficients of the modulated Fourier expansion, the *modulation functions* $z_j^{[\mathbf{k}]}$, are required to be polynomials on a slow time scale $\tau = \delta t$ with δ from (33) that have all derivatives bounded independently of the small parameters. By a slight abuse of notation we write in the following $z_j^{\mathbf{k}}$ instead of $z_j^{[\mathbf{k}]}$ and $\sum_{\mathbf{k}}$ instead of $\sum_{[\mathbf{k}]}$.

4.2 Modulation equations

Requiring $\xi_j(t_n) = \xi_j^n$ for $n \geq 1$ with ξ_j^n given by (22) yields, after a comparison of the coefficients of $\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t}$, *modulation equations* for the modulation functions $z_j^{\mathbf{k}}$:

$$\begin{aligned} z_j^{\mathbf{k}}(\tau+\delta h)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})h} = \mathrm{e}^{-\mathrm{i}\omega_j h} z_j^{\mathbf{k}}(\tau) + \sum_{m+m'=2}^{\infty} \sum_{\substack{\mathbf{k}^1+\cdots+\mathbf{k}^m \\ -\mathbf{l}^1-\cdots-\mathbf{l}^{m'}\in[\mathbf{k}]}} \\ \sum_{k\in\mathcal{Z}^m,\, l\in\mathcal{Z}^{m'}} h Q_{j,k,l} z_{k^1}^{\mathbf{k}^1}(\tau)\cdots z_{k^m}^{\mathbf{k}^m}(\tau) \overline{z}_{l^1}^{\mathbf{l}^1}(\tau)\cdots\overline{z}_{l^{m'}}^{\mathbf{l}^{m'}}(\tau). \end{aligned} \tag{34a}$$

The condition $\xi_j(0) = \xi_j^0$ yields

$$\sum_{\mathbf{k}} z_j^{\mathbf{k}}(0) = \xi_j^0. \tag{34b}$$

For the approximate solution of the modulation equations (34) it is useful to expand the modulation functions in powers of ε and δ,

$$z_j^{\mathbf{k}}(\tau) = \sum_{p=0}^{\infty} \varepsilon\delta^p z_{j,p}^{\mathbf{k}}(\tau) \tag{35}$$

with polynomials $z_{j,p}^{\mathbf{k}}$ in $\tau = \delta t$. We call the functions $z_{j,p}^{\mathbf{k}}$ *modulation coefficient functions* and set $z_{j,p}^{\mathbf{k}} = 0$ for $p < 0$. After dividing by $\delta h \mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})h}$, expanding $z_j^{\mathbf{k}}(\tau+\delta h)$ around τ and (formally) comparing the coefficients of $\varepsilon\delta^p$, the modulation equations (34a) become

$$\begin{aligned} \frac{1-\mathrm{e}^{-\mathrm{i}(\omega_j-\mathbf{k}\cdot\boldsymbol{\varpi})h}}{\delta h} z_{j,p}^{\mathbf{k}} + \dot{z}_{j,p}^{\mathbf{k}} = -\sum_{r=2}^{\infty} \frac{h^{r-1}}{r!} \frac{\mathrm{d}^r}{\mathrm{d}\tau^r} z_{j,p+1-r}^{\mathbf{k}} \\ + \sum_{m+m'=2}^{\infty} \sum_{\substack{p_1+\cdots+p_m \\ +q_1+\cdots+q_{m'}=p+3-2(m+m')}} \frac{\varepsilon^{m+m'-1}}{\delta^{2m+2m'-2}} \sum_{\substack{\mathbf{k}^1+\cdots+\mathbf{k}^m \\ -\mathbf{l}^1-\cdots-\mathbf{l}^{m'}\in[\mathbf{k}]}} \\ \sum_{k\in\mathcal{Z}^m,\, l\in\mathcal{Z}^{m'}} \mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})h} Q_{j,k,l} z_{k^1,p_1}^{\mathbf{k}^1}\cdots z_{k^m,p_m}^{\mathbf{k}^m} \overline{z}_{l^1,q_1}^{\mathbf{l}^1}\cdots\overline{z}_{l^{m'},q_{m'}}^{\mathbf{l}^{m'}}. \end{aligned} \tag{36a}$$

Condition (34b) yields

$$z_{j,p}^{\langle j\rangle}(0) = -\sum_{\mathbf{k}\neq\langle j\rangle} z_{j,p}^{\mathbf{k}}(0) + \begin{cases} \varepsilon^{-1}\xi_j^0, & p=0, \\ 0, & p>0, \end{cases} \tag{36b}$$

where $\langle j\rangle$ denotes the jth unit vector in $\mathbb{Z}^{\mathcal{Z}}$.

4.3 Construction of modulation functions

We construct modulation functions $z_j^{\mathbf{k}}$ that solve the modulation equations (34) up to a small defect. We work with the asymptotic expansion (35) and consider the equations (36). The crucial observation is that the right-hand side of (36a) depends only on modulation coefficient functions $z_{k,q}^{\mathbf{l}}$ with $q<p$. This allows us to solve the equations (36) up to a small defect by the following simple recursion.

Fix $p\geq 0$ and assume that we have computed all modulation coefficient functions $z_{j,q}^{\mathbf{k}}$ with $q<p$ (this is true for $p=0$). Equation (36a) is then of the form

$$\alpha z_{j,p}^{\mathbf{k}} + \dot{z}_{j,p}^{\mathbf{k}} = P$$

with a polynomial P. The unique polynomial solution of this equation is given for $\alpha\neq 0$ by

$$z_{j,p}^{\mathbf{k}}(\tau) = \sum_{m=0}^{\deg(P)} (-1)^m \alpha^{-m-1} \frac{\mathrm{d}^m}{\mathrm{d}\tau^m} P(\tau). \tag{37}$$

We therefore compute $z_{j,p}^{\mathbf{k}}$ for all j and all $\mathbf{k}$ as follows.

(i) For indices $(j,\mathbf{k})$ with $j\neq j(\mathbf{k})$ or $\|\widetilde{\mathbf{k}}\|>p$ for all $\widetilde{\mathbf{k}}\in[\mathbf{k}]$ we set

$$z_{j,p}^{\mathbf{k}} = 0. \tag{38a}$$

This is consistent with (36a) since the right-hand side of this equation vanishes for these indices by induction (recall that $Q_{j,k,l}=0$ if $j\not\equiv k^1+\cdots+k^m-l^1-\cdots-l^{m'} \bmod 2K$).

(ii) For indices $(j,\mathbf{k})$ with $|1-\mathrm{e}^{-\mathrm{i}(\omega_j-\mathbf{k}\cdot\boldsymbol{\varpi})h}|\geq \delta h/2$ that are not covered by (i) we

$$\text{compute } z_{j,p}^{\mathbf{k}} \text{ from (36a) and (37).} \tag{38b}$$

Indeed, the factor in front of $z_{j,p}^{\mathbf{k}}$ in (36a) is bounded for these indices away from zero, and the comparison of coefficients used to derive (36a) thus makes sense.

(iii) For indices $(j,\mathbf{k})\neq(j,\langle j\rangle)$ that are neither covered by (i) nor by (ii) we set

$$z_{j,p}^{\mathbf{k}} = 0. \tag{38c}$$

Of course, this introduces a defect which, however, can be controlled using the non-resonance condition of Assumption 2 as we shall see in Subsect. 4.5.
For the considered indices $(j,\mathbf{k})$ we have $|1-\mathrm{e}^{-\mathrm{i}(\omega_j-\mathbf{k}\cdot\boldsymbol{\varpi})h}|<\delta h/2$, and they are in this sense close to a resonance. We therefore call them *near-resonant* in the following.

(iv) Having computed $z^{\mathbf{k}}_{j,p}$ for all j and all $\mathbf{k} \neq \langle j \rangle$ in (i)–(iii) we can

$$\text{compute } z^{\langle j \rangle}_{j,p}(0) \text{ from (36b).} \tag{38d}$$

Moreover, since the factor in front of $z^{\mathbf{k}}_{j,p}$ in (36a) vanishes for $\mathbf{k} = \langle j \rangle$, we can

$$\text{compute } \dot{z}^{\langle j \rangle}_{j,p} \text{ from (36a).} \tag{38e}$$

This allows us to compute the diagonal modulation coefficient functions $z^{\langle j \rangle}_{j,p}$.

We stop the above construction (38) of modulation coefficient function $z^{\mathbf{k}}_{j,p}$ after $p = N$,

$$z^{\mathbf{k}}_{j,p} = 0 \qquad \text{for} \qquad p > N. \tag{38f}$$

It is clear that the construction leads to modulation coefficient functions $z^{\mathbf{k}}_{j,p}$ that are polynomials in τ, of degree bounded by p. Moreover, we have

$$z^{\mathbf{k}}_{j,0} = 0 \qquad \text{for} \qquad \mathbf{k} \neq \langle j \rangle \tag{39}$$

because the right-hand side of (36a) vanishes for $p = 0$.

4.4 Size of the modulation functions

We estimate the modulation coefficient functions constructed in (38). For fixed index p we collect them in the vectors

$$\mathbf{z}_p = (z^{\mathbf{k}}_{j,p})_{j \in \mathcal{Z}, \mathbf{k} \in \mathbb{Z}^{\mathcal{Z}}}.$$

We also consider their rescalings

$$(\mathbf{\Gamma}^{s-\widehat{s}} \mathbf{z}_p)^{\mathbf{k}}_j := \big(\Gamma^{\mathbf{k}}\big)^{s-\widehat{s}} \cdot z^{\mathbf{k}}_{j,p} \qquad \text{with} \qquad \Gamma^{\mathbf{k}} := \min_{\widetilde{\mathbf{k}} \in [\mathbf{k}]} \Big(2^{\|\widetilde{\mathbf{k}}\|} \prod_{l \in \mathcal{Z}} |l|^{|\widetilde{k}_l|} \Big) \tag{40}$$

and with $s \geq \widehat{s} := (d+1)/2$ such that Lemma 3.3 is applicable for s and $\widehat{s}$.

For vectors $\mathbf{v} = (v^{\mathbf{k}}_j)_{j \in \mathcal{Z}, \mathbf{k} \in \mathbb{Z}^{\mathcal{Z}}}$ of polynomials $v^{\mathbf{k}}_j = v^{\mathbf{k}}_j(\tau)$ in τ we use the norm

$$|||\mathbf{v}|||_{s,\tau} = \Big\| \Big(\sum_{\mathbf{k}} |v^{\mathbf{k}}_j|_\tau \Big)_{j \in \mathcal{Z}} \Big\|_s = \Big(\sum_{j \in \mathcal{Z}} |j|^{2s} \Big(\sum_{\mathbf{k}} |v^{\mathbf{k}}_j|_\tau \Big)^2 \Big)^{1/2},$$

where

$$|v|_\tau = \sum_{m=0}^{\infty} \frac{1}{m!} \Big| \frac{\mathrm{d}^m}{\mathrm{d}\tau^m} v(\tau) \Big|.$$

Lemma 4.1. *The modulation coefficient functions (38) satisfy on $0 \leq \tau \leq 1$ for $\delta \leq \delta_0$ and $s \geq \widehat{s}$*

$$|||\mathbf{z}_p|||_{s,\tau} \leq C \qquad \textit{and} \qquad |||\mathbf{\Gamma}^{s-\widehat{s}} \mathbf{z}_p|||_{\widehat{s},\tau} \leq C$$

for all p with constants C and δ_0.

Proof. This follows from the recursive construction (38): The property

$$|vw|_\tau \leq |v|_\tau |w|_\tau$$

together with Lemma 3.3 yields inductively an estimate of the nonlinearity on the right-hand side of (36a) in the norm $|||\cdot|||_{s,\tau}$ (note that terms in the nonlinearity with $2(m+m') > p+3$ vanish, and hence the sum over m and m' is finite). Then, the property $|\dot{v}|_\tau \leq \deg(v)|v|_\tau$ allows us to estimate the norm $|||\cdot|||_{s,\tau}$ for the vector consisting of modulation coefficient functions constructed with (38b). For the remaining nonzero modulation coefficient functions constructed with (38d)–(38e), the estimate in the norm $|||\cdot|||_{s,\tau}$ then follows using the smallness of the initial value (30) and the property $|v|_\tau \leq |v(0)| + \sup_{0\leq\tilde{\tau}\leq\tau}|\dot{v}|_{\tilde{\tau}}$.

For the estimate of the rescaling $\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{z}_p$ in the norm $|||\cdot|||_{\widehat{s},\tau}$ we can use essentially the same argument: We just have to take into account that

$$\Gamma^{\mathbf{k}^1+\mathbf{k}^2} \leq \Gamma^{\mathbf{k}^1}\Gamma^{\mathbf{k}^2} \tag{41}$$

and that $\Gamma^{\langle j\rangle} = 2|j|$. The latter follows from

$$|j| = |j(\widetilde{\mathbf{k}})| \leq \sum_{l\in\mathcal{Z}}|\widetilde{k}_l||l| \leq \|\widetilde{\mathbf{k}}\|\prod_{l\in\mathcal{Z}}|l|^{|\widetilde{k}_l|}$$

for $\widetilde{\mathbf{k}} \in [\langle j\rangle]$ and $j \in \mathcal{Z}$. □

4.5 Defect and error

The modulation functions constructed in (38) via their modulation coefficient functions (35) are supposed to fulfill the modulation system (34). However, there are two sources of error in their construction: First, we stopped the construction of modulation coefficient functions $z^{\mathbf{k}}_{j,p}$ after $p = N$ (38f). Second, the modulation functions for near-resonant indices $(j,\mathbf{k})$ were set to zero (38c). In other words, the constructed modulation functions satisfy the equations of motion (34a) of the modulation system only up to a defect,

$$\begin{aligned} d^{\mathbf{k}}_j + e^{\mathbf{k}}_j = -z^{\mathbf{k}}_j(\cdot+\delta h)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})h} + \mathrm{e}^{-\mathrm{i}\omega_j h}z^{\mathbf{k}}_j + \sum_{m+m'=2}^{\infty}\sum_{\substack{\mathbf{k}^1+\cdots+\mathbf{k}^m\\ -\mathbf{l}^1-\cdots-\mathbf{l}^{m'}=[\mathbf{k}]}} \\ \sum_{k\in\mathcal{Z}^m,\,l\in\mathcal{Z}^{m'}} hQ_{j,k,l}z^{\mathbf{k}^1}_{k^1}\cdots z^{\mathbf{k}^m}_{k^m}\overline{z}^{\mathbf{l}^1}_{l^1}\cdots\overline{z}^{\mathbf{l}^{m'}}_{l^{m'}}, \end{aligned} \tag{42}$$

whereas the initial condition (34b) is met exactly. Here, $\mathbf{d} = (d^{\mathbf{k}}_j)_{j\in\mathcal{Z},\mathbf{k}\in\mathbb{Z}^z}$ denotes the defect from the cut-off (38f), i.e.,

$$d^{\mathbf{k}}_j = \sum_{p=N+1}^{\infty}\varepsilon\delta^p h F^{\mathbf{k}}_{j,p}\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})h}, \tag{43}$$

where $F^{\mathbf{k}}_{j,p}$ is the right-hand side of (36a). The defect in near-resonant indices that is not yet covered by $\mathbf{d}$ is denoted by $\mathbf{e} = (e^{\mathbf{k}}_j)_{j\in\mathcal{Z},\mathbf{k}\in\mathbb{Z}^z}$, i.e., $e^{\mathbf{k}}_j$ is different from zero only for near-resonant indices and in this case

$$e^{\mathbf{k}}_j = \sum_{p=1}^{N}\varepsilon\delta^p h F^{\mathbf{k}}_{j,p}\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})h}. \tag{44}$$

(Recall that $F^{\mathbf{k}}_{j,p} = 0$ for $p = 0$.) Both defects are estimated in the following lemma.

Lemma 4.2. *The defects (42)–(44) satisfy on $0 \le \tau \le 1$ for $\delta \le \delta_0$ and $s \ge s_0$*

$$|||\mathbf{d}|||_{s,\tau} \le C\varepsilon\delta^{N+1}h, \qquad |||\mathbf{e}|||_{s,\tau} \le C\varepsilon\delta^{N+1}h,$$
$$|||\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{d}|||_{\widehat{s},\tau} \le C\varepsilon\delta^{N+1}h, \qquad |||\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{e}|||_{\widehat{s},\tau} \le C\varepsilon\delta h$$

with constants C, s_0 and δ_0.

Proof. (a) For the bound of $\mathbf{d}$, we note that by Lemmas 3.3 and 4.1

$$|||\mathbf{d}|||_{s,\tau} \le C\varepsilon\delta^{N+1}h + \varepsilon h \sum_{m+m'=2}^{\infty} C_{m,m',s}C^{m+m'} \sum_{p=\max(N+1,2(m+m')-3)}^{\infty} \delta^p,$$

where $C/(N+1)$ denotes the constant of Lemma 4.1. Splitting the sum over m and m' in a part with $m+m' \le N+3$ and another part with $m+m' \ge N+4$ proves the claimed estimate of $\mathbf{d}$ using the second part of Lemma 3.3 for the sum with $m+m' \ge N+4$ and sufficiently small δ. The estimates of $\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{d}$ and $\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{e}$ follow similarly.

(b) Concerning the defect $\mathbf{e}$ in near-resonant indices, we note that for those indices $(j,\mathbf{k})$

$$\begin{aligned}\left|\mathrm{e}^{-\mathrm{i}(\varpi_j - \mathbf{k}\cdot\boldsymbol{\varpi})h} - 1\right| &\le \left|\mathrm{e}^{-\mathrm{i}\varpi_j h} - \mathrm{e}^{-\mathrm{i}\omega_j h}\right| + \left|\mathrm{e}^{-\mathrm{i}(\omega_j - \mathbf{k}\cdot\boldsymbol{\varpi})h} - 1\right| \\ &< 2\left|\sin((\varpi_j - \omega_j)h/2)\right| + \frac{\delta h}{2} \le \widehat{\varepsilon}h + \frac{\delta h}{2} \le \delta h\end{aligned}$$

by Assumption 2 and for $2\delta \le 1$. The non-resonance condition of Assumption 2 (used with $\mathbf{k} - \langle j \rangle$ in place of $\mathbf{k}$) thus implies

$$|j|^{s-\widehat{s}} \le c_2^{(s-\widehat{s})/2}\delta^N \left(\Gamma^{\mathbf{k}}\right)^{s-\widehat{s}}$$

for $s - \widehat{s} \ge 2s_2$. This shows that

$$|||\mathbf{e}|||_{s,\tau} \le c_2^{(s-\widehat{s})/2}\delta^N |||\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{e}|||_{\widehat{s},\tau},$$

and the claimed estimate of $\mathbf{e}$ follows from Lemma 4.1. □

Now, we study the difference $\xi^n - \xi(t_n)$ of the numerical solution ξ^n of (22) and its modulated Fourier expansion $\xi(t)$ of (32). In this modulated Fourier expansion $\xi(t)$ of (32) we use the modulation functions constructed in (38) at discrete times $t_n = nh$.

Proposition 4.3. *We have for $\delta \le \delta_0$ and $s \ge s_0$*

$$\|\xi^n - \xi(t_n)\|_s \le C\varepsilon\delta^N \qquad \text{for} \qquad 0 \le t_n = nh \le \delta^{-1}$$

with constants C, s_0 and δ_0.

Proof. (a) From Lemma 4.1 we know for the modulated Fourier expansion the estimate

$$\|\xi(t_n)\|_s \le C\varepsilon \qquad \text{for} \qquad 0 \le t_n \le \delta^{-1}.$$

(b) A corresponding estimate holds, for sufficiently small ε, also for the numerical solution:

$$\|\xi^n\|_s \le 2\widehat{c}^{-1}\varepsilon \qquad \text{for} \qquad 0 \le t_n \le \varepsilon^{-1/2}$$

with $\widehat{c}$ from (30), since by Lemma 3.3 and induction

$$\|\xi^{n+1}\|_s \le \|\xi^0\|_s + h\sum_{n'=0}^{n}\sum_{m+m'=2}^{\infty} C_{m,m',s}\|\xi^{n'}\|_s^{m+m'} \le \widehat{c}^{-1}\varepsilon + \frac{4C}{\widehat{c}^2 r^2}\, nh\varepsilon^2$$

for $2\varepsilon \le r\widehat{c}$ with C and r from Lemma 3.3. We may assume without loss of generality that the constant C from (a) is larger than $2/\widehat{c}$.

(c) The modulated Fourier expansion satisfies by (42)

$$\xi_j(t_{n+1}) = \mathrm{e}^{-\mathrm{i}\omega_j h}\xi_j(t_n) - \sum_{\mathbf{k}}\big(d_j^{\mathbf{k}}(\delta t_n) + e_j^{\mathbf{k}}(\delta t_n)\big)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t} + \sum_{m+m'=2}^{\infty} \sum_{k\in\mathcal{Z}^m,\, l\in\mathcal{Z}^{m'}} hQ_{j,k,l}\xi_{k^1}(t_n)\cdots\xi_{k^m}(t_n)\overline{\xi}_{l^1}(t_n)\cdots\overline{\xi}_{l^{m'}}(t_n).$$

Subtracting the numerical solution ξ^{n+1} of (22) and using (a), (b), Lemma 3.3 and Lemma 4.2 shows that for $0 \le t_n \le \delta^{-1}$ and for sufficiently small ε

$$\|\xi^{n+1} - \xi(t_{n+1})\|_s \le \|\xi^n - \xi(t_n)\|_s + C\varepsilon\delta^{N+1}h + C\varepsilon h\|\xi^n - \xi(t_n)\|_s.$$

The claimed estimate follows inductively. □

4.6 The splitting structure of the modulated Fourier expansion

In the previous subsections, a modulated Fourier expansion was constructed and analysed based on the representation (22) of the numerical scheme, i.e., based on flows of differential equations. Recall that we have derived in Subsect. 3.5 differential equations (Hamiltonian functions) that underly the flows that compose the numerical scheme. In this subsection, we will derive corresponding differential equations for the modulated Fourier expansion.

Motivated by (27), we denote by $\Phi^h_{\mathbf{H}_0}$ the flow at time h of the Hamiltonian differential equation

$$\mathrm{i}\dot{z}_j^{\mathbf{k}} = \frac{\partial \mathbf{H}_0}{\partial \overline{z}_j^{\mathbf{k}}}(\mathbf{z},\overline{\mathbf{z}})$$

with Hamiltonian function

$$\mathbf{H}_0(\mathbf{z},\overline{\mathbf{z}}) = \sum_{\mathbf{k}}\sum_{j\in\mathcal{Z}}\big(|\ell + j \bmod 2K|^2 - |\ell|^2\big)|w_j^{\mathbf{k}}|^2, \qquad \begin{pmatrix} w_j^{\mathbf{k}} \\ \overline{w}_{-j}^{-\mathbf{k}} \end{pmatrix} = S_j^{-1}\begin{pmatrix} z_j^{\mathbf{k}} \\ \overline{z}_{-j}^{-\mathbf{k}} \end{pmatrix}.$$

Correspondingly, we denote by $\Phi^h_{\mathbf{P}}$ the flow at time h of the Hamiltonian differential equation with Hamiltonian function

$$\mathbf{P}(\mathbf{z},\overline{\mathbf{z}}) = \sum_{m+m'=0}^{\infty} \sum_{\substack{\mathbf{k}^1+\cdots+\mathbf{k}^m \\ -\mathbf{l}^1-\cdots-\mathbf{l}^{m'}\in\mathcal{M}}} \sum_{k\in\mathcal{Z}^m,\, l\in\mathcal{Z}^{m'}} P_{k,l} z_{k^1}^{\mathbf{k}^1}\cdots z_{k^m}^{\mathbf{k}^m}\overline{z}_{l^1}^{\mathbf{l}^1}\cdots\overline{z}_{l^{m'}}^{\mathbf{l}^{m'}},$$

compare (29).

The splitting structure of the modulation system for the modulation functions $\mathbf{z}$ is revealed in the following lemma: advancing the modulation functions by δh corresponds, up to a small defect, to solving Hamiltonian differential equations with Hamiltonian functions $\mathbf{H}_0$ and $\mathbf{P}$ one after another.

Lemma 4.4. *We have*

$$\Phi^h_{\mathbf{H}_0} \circ \Phi^h_{\mathbf{P}}(\mathbf{z}(\delta t_n)) = \widetilde{\mathbf{z}}(\delta t_{n+1}) + \mathbf{d}(\delta t_n) + \mathbf{e}(\delta t_n)$$

with the defects $\mathbf{d}$ *and* $\mathbf{e}$ *of* (42)–(44) *and where* $\widetilde{z}^{\mathbf{k}}_j(\delta t_{n+1}) = z^{\mathbf{k}}_j(\delta t_{n+1})\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})h}$.

Proof. Let

$$\big(\Phi^h_P(\xi)\big)_j = \sum_{m+m'=0}^{\infty} \sum_{k\in\mathcal{Z}^m,\, l\in\mathcal{Z}^{m'}} P_{j,k,l}\,\xi_{k^1}\cdots\xi_{k^m}\overline{\xi}_{l^1}\cdots\overline{\xi}_{l^{m'}}$$

be the expansion of the flow Φ^h_P given by (28). Then one verifies that the flow $\Phi^h_{\mathbf{P}}$ is given by the same coefficients $P_{j,k,l}$,

$$\big(\Phi^h_{\mathbf{P}}(\mathbf{z})\big)^{\mathbf{k}}_j = \sum_{m+m'=0}^{\infty} \sum_{k\in\mathcal{Z}^m,\, l\in\mathcal{Z}^{m'}} \sum_{\substack{\mathbf{k}^1+\cdots+\mathbf{k}^m \\ -\mathbf{l}^1-\cdots-\mathbf{l}^{m'}=[\mathbf{k}]}} P_{j,k,l}\, z^{\mathbf{k}^1}_{k^1}\cdots z^{\mathbf{k}^m}_{k^m}\overline{z}^{\mathbf{l}^1}_{l^1}\cdots\overline{z}^{\mathbf{l}^{m'}}_{l^{m'}}.$$

This implies that also the coefficients of the expansions of $\Phi^h_{H_0}\circ\Phi^h_P(\xi)$ and $\Phi^h_{\mathbf{H}_0}\circ\Phi^h_{\mathbf{P}}(\mathbf{z})$ coincide. The coefficients in the expansion of $\Phi^h_{H_0}\circ\Phi^h_P(\xi)$ are given by (22), and they also appear in the expansion (42) of $\widetilde{\mathbf{z}}(\delta t_{n+1}) + \mathbf{d}(\delta t_n) + \mathbf{e}(\delta t_n)$. The statement of the lemma follows. □

4.7 Discrete almost-invariants

An essential property of modulated Fourier expansions is the existence of formal invariants. These invariants will finally allow us to consider long time intervals by patching together many of the short time intervals considered so far. They take the form

$$\mathcal{I}_m(\mathbf{z}) = \sum_{\mathbf{k}} \sum_{j\in\mathcal{Z}} \sum_{\substack{l\in\mathcal{Z} \\ n(l)=m}} k_l |z^{\mathbf{k}}_j|^2 \qquad \text{for} \qquad m\in\mathcal{N} := \{\, n(j) : j\in\mathcal{Z}\,\}. \tag{45}$$

This is well defined (recall that the $\sum_{\mathbf{k}}$ stands for the sum over the equivalence classes $[\mathbf{k}]\in\mathbb{Z}^{\mathcal{Z}}/\mathcal{M}$) since $\sum_{l:n(l)=m} k_l = 0$ for $\mathbf{k}\in\mathcal{M}$ by part (c) of Assumption 2.

Lemma 4.5. *We have*

$$\mathcal{I}_m\big(\Phi^h_{\mathbf{H}_0}\circ\Phi^h_{\mathbf{P}}(\mathbf{z}(\delta t_n))\big) = \mathcal{I}_m\big(\mathbf{z}(\delta t_n)\big) \qquad \textit{for} \qquad m\in\mathcal{N}.$$

Proof. Let $\mathbf{S}(\theta)$ be defined by

$$\big(\mathbf{S}(\theta)\mathbf{z}\big)^{\mathbf{k}}_j = \mathrm{e}^{\mathrm{i}\theta\sum_{l:n(l)=m} k_l} z^{\mathbf{k}}_j$$

for $m\in\mathcal{N}$. The Hamiltonian function $\mathbf{P}$ from the previous subsection is invariant under the transformations $\mathbf{S}(\theta)$, and this leads to conserved quantities by Noether's theorem: We have along a solution $\mathbf{z} = \Phi^t_{\mathbf{P}}\mathbf{z}^0$ of the Hamiltonian differential equation with Hamiltonian function $\mathbf{P}$

$$\begin{aligned} 0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\Big|_{\theta=0} \mathbf{P}\big(\mathbf{S}(\theta)\mathbf{z}, \overline{\mathbf{S}(\theta)\mathbf{z}}\big) &= -\mathrm{i}\sum_{\mathbf{k}}\sum_{j\in\mathcal{Z}}\sum_{\substack{l\in\mathcal{Z} \\ n(l)=m}} k_l\left(\overline{z}^{\mathbf{k}}_j\frac{\partial\mathbf{P}}{\partial\overline{z}^{\mathbf{k}}_j}(\mathbf{z},\overline{\mathbf{z}}) - z^{\mathbf{k}}_j\frac{\partial\mathbf{P}}{\partial z^{\mathbf{k}}_j}(\mathbf{z},\overline{\mathbf{z}})\right) \\ &= \sum_{\mathbf{k}}\sum_{j\in\mathcal{Z}}\sum_{\substack{l\in\mathcal{Z} \\ n(l)=m}} k_l\frac{\mathrm{d}}{\mathrm{d}t}|z^{\mathbf{k}}_j|^2 = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}_m(\mathbf{z}). \end{aligned}$$

This implies conservation of $\mathcal{I}_m$ along the flow of $\mathbf{P}$,

$$\mathcal{I}_m\big(\Phi^h_{\mathbf{P}}(\mathbf{z}(\delta t_n))\big) = \mathcal{I}_m\big(\mathbf{z}(\delta t_n)\big).$$

In the same way, one shows conservation of $\mathcal{I}_m$ along the flow of $\mathbf{H}_0$, and the statement of the lemma follows. □

In the end, we are interested more in $\mathbf{z}(\delta t_{n+1})$ than in $\Phi^h_{\mathbf{H}_0} \circ \Phi^h_{\mathbf{P}}(\mathbf{z}(\delta t_n))$. The following lemma shows that $\mathcal{I}_m$ is an almost-invariant along the modulation functions $\mathbf{z}$.

Proposition 4.6. *We have for $\delta \le \delta_0$ and $s \ge s_0$*

$$\sum_{m\in\mathcal{N}} \max(1,m)^s \big|\mathcal{I}_m(\mathbf{z}(\delta t_n)) - \mathcal{I}_m(\mathbf{z}(0))\big| \le C\varepsilon^2\delta^N \qquad \text{for} \qquad 0 \le t_n = nh \le \delta^{-1}$$

with constants C, s_0 and δ_0.

Proof. Throughout the proof, we work with the representative $\mathbf{k}$ of $[\mathbf{k}]$ for which the minimum in the definition (41) of $\Gamma^{\mathbf{k}}$ is attained.

(a) By Lemma 4.4 and Lemma 4.5 we have

$$\big|\mathcal{I}_m(\mathbf{z}(\delta t_{n+1})) - \mathcal{I}_m(\mathbf{z}(\delta t_n))\big| \le 2\sum_{\mathbf{k}}\sum_{j\in\mathcal{Z}}\sum_{\substack{l\in\mathcal{Z}\\ n(l)=m}} |k_l|\big(|z^{\mathbf{k}}_j||d^{\mathbf{k}}_j| + |d^{\mathbf{k}}_j|^2 + |e^{\mathbf{k}}_j|^2\big)$$

since $z^{\mathbf{k}}_j = 0$ if $e^{\mathbf{k}}_j \ne 0$. Here, the modulation functions $z^{\mathbf{k}}_j$ on the right-hand side are evaluated at time $\tau = \delta t_{n+1}$ and the defects $d^{\mathbf{k}}_j$ and $e^{\mathbf{k}}_j$ at time $\tau = \delta t_n$.

(b) Let $\mathbf{k} \ne \mathbf{0}$ and $j = j(\mathbf{k}) = \sum_{l\in\mathcal{Z}} k_l l \bmod 2M \in \mathcal{Z}$, and let $\bar{l} \in \mathcal{Z}$ be the index of largest norm $|\cdot|$ with $k_{\bar{l}} \ne 0$. Then we have

$$|\bar{l}|^2 \le |j| \cdot \Gamma^{\mathbf{k}}$$

Indeed, if $|\bar{l}|^2 > \Gamma^{\mathbf{k}}$, then necessarily $|k_{\bar{l}}| = 1$ and $\|\mathbf{k}\| \cdot |l| \le |\bar{l}|$ for all $l \ne \bar{l}$ with $k_l \ne 0$, and hence

$$|\bar{l}| \le \Big|j - \sum_{\bar{l}\ne l\in\mathcal{Z}} k_l l\Big| \le |j| + \frac{\|\mathbf{k}\|-1}{\|\mathbf{k}\|}\,|\bar{l}|,$$

i.e., $|\bar{l}| \le \|\mathbf{k}\| \cdot |j|$. This implies for $s \ge 2\widehat{s}$ that

$$\sum_{l\in\mathcal{Z}} |k_l||l|^{2s} \le \|\mathbf{k}\| \cdot |\bar{l}|^{2s} \le |j|^{2\widehat{s}}\big(\Gamma^{\mathbf{k}}\big)^{2(s-\widehat{s})}.$$

The last estimate improves for near-resonant indices (for which $e^{\mathbf{k}}_j \ne 0$) by the non-resonance condition in Assumption 2 to

$$\sum_{l\in\mathcal{Z}} |k_l||l|^{2s} \le c_2^{s-2\widehat{s}}|j|^{2\widehat{s}}\big(\Gamma^{\mathbf{k}}\big)^{2(s-\widehat{s})}\delta^N$$

if $s - 2\widehat{s} \ge s_2$.

(c) By (a), (b), the Cauchy-Schwarz inequality and Lemma 4.7 below we have

$$\sum_{m\in\mathcal{N}} \max(1,m)^s \big|\mathcal{I}_m(\mathbf{z}(\delta t_{n+1})) - \mathcal{I}_m(\mathbf{z}(\delta t_n))\big|$$
$$\le C|||\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{z}|||_{\widehat{s}}|||\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{d}|||_{\widehat{s}} + C|||\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{d}|||^2_{\widehat{s}} + Cc_2^{s-2\widehat{s}}\delta^N|||\mathbf{\Gamma}^{s-\widehat{s}}\mathbf{e}|||^2_{\widehat{s}}.$$

The statement of the proposition thus follows from Lemma 4.1 and Lemma 4.2 by summing up. □

Lemma 4.7. *We have*

$$\tfrac{1}{C}|j|^2 \le \max\big(1,|n(j)|\big) \le C|j|^2 \qquad \text{for all} \qquad j \in \mathcal{Z}$$

with a positive constant C depending only on ℓ.

Proof. We have $-|\ell|^2 \le n(j) \le |j|^2$ since $|\ell \pm j \bmod 2K| \le |\ell \pm j|$, and hence $|n(j)| \le C|j|^2$. To get a lower bound for $|n(j)|$ we note that $|\ell_1 \pm j_1 \bmod 2K| \ge \min(|\ell_1 + j_1|, |\ell_1 - j_1|)$, and hence

$$\tfrac{1}{2}|\ell_1 + j_1 \bmod 2K|^2 + \tfrac{1}{2}|\ell_1 - j_1 \bmod 2K|^2 - \ell_1^2 \ge j_1^2 - 2|j_1| \cdot |\ell_1|.$$

This holds not only for the first component, and we get by summing up all components

$$n(j) \ge |j|^2 - 2|\ell| \cdot |j|.$$

We therefore get $n(j) \ge \frac{1}{2}|j|^2$ for $4|\ell| < |j|$. For $4|\ell| \ge |j|$ we have $\max(1,|n(j)|) \ge 1 \ge \frac{1}{C}|j|^2$. □

Next we show that the almost-invariants $\mathcal{I}_m$ of (45) are close to the *super-actions*

$$I_m(\xi) = \sum_{\substack{l \in \mathcal{Z} \\ n(l)=m}} |\xi_l|^2, \qquad m \in \mathcal{N}, \tag{46}$$

that collect those *actions* $|\xi_l|^2$ with the same value $n(l)$.

Proposition 4.8. *We have for $\delta \le \delta_0$ and $s \ge s_0$*

$$\sum_{m \in \mathcal{N}} \max(1,m)^s \big|\mathcal{I}_m(\mathbf{z}(\delta t_n)) - I_m(\xi^n)\big| \le C\varepsilon^2\delta \qquad \text{for} \qquad 0 \le t_n = nh \le \delta^{-1}$$

with constants C, s_0 and δ_0.

Proof. We omit the argument δt_n of the modulation functions. We have by (38a)

$$\mathcal{I}_m(\mathbf{z}) - \sum_{\substack{l \in \mathcal{Z}: \\ n(l)=m}} |z_l^{\langle l \rangle}|^2 = \sum_{\substack{l \in \mathcal{Z}: \\ n(l)=m}} \sum_{\mathbf{k} \ne \langle l \rangle} k_l |z_{j(\mathbf{k})}^{\mathbf{k}}|^2,$$

and part (b) of the proof of Proposition 4.6 together with Lemma 4.1, Lemma 4.7 and (39) implies

$$\sum_{m \in \mathcal{N}} \max(1,m)^s \Big|\mathcal{I}_m(\mathbf{z}) - \sum_{\substack{l \in \mathcal{Z}: \\ n(l)=m}} |z_l^{\langle l \rangle}|^2\Big| \le C\varepsilon^2\delta^2.$$

On the other hand, we have for the modulated Fourier expansions $\xi(t)$ (32)

$$\Big|I_m(\xi(t_n)) - \sum_{\substack{l \in \mathcal{Z}: \\ n(l)=m}} |z_l^{\langle l \rangle}|^2\Big| \le 2 \sum_{\substack{l \in \mathcal{Z}: \\ n(l)=m}} \Big(\sum_{\mathbf{k} \ne \langle l \rangle} |z_l^{\mathbf{k}}|\Big)\Big(\sum_{\mathbf{k}} |z_l^{\mathbf{k}}|\Big),$$

and hence by the Cauchy-Schwarz inequality together with Lemma 4.1, Lemma 4.7 and (39)

$$\sum_{m \in \mathcal{N}} \max(1,m)^s \Big|I_m(\xi(t_n)) - \sum_{\substack{l \in \mathcal{Z}: \\ n(l)=m}} |z_l^{\langle l \rangle}|^2\Big| \le C\varepsilon^2\delta.$$

Finally, we have by Proposition 4.3 and Lemma 4.7 for the numerical solution ξ^n

$$\sum_{m \in \mathcal{N}} \max(1,m)^s \big|I_m(\xi(t_n)) - I_m(\xi^n)\big| \le C\varepsilon^2\delta^N,$$

where we have used that $||\xi_j(t_n)|^2 - |\xi_j^n|^2| \le |\xi_j(t_n) - \xi_j^n|(|\xi_j(t_n)| + |\xi_j^n|)$. Putting all this together proves the statement of the proposition. □

4.8 Modulated Fourier expansion on another time interval

All estimates of the previous subsections are valid on the time interval $0 \le t_n = nh \le \delta^{-1}$. We assume without loss of generality that

$$\delta^{-1} = \widetilde{n}h \tag{47}$$

for some $\widetilde{n} \in \mathbb{N}$. In this subsection, we consider consecutive short time intervals

$$\nu\delta^{-1} \le t_n = nh \le (\nu+1)\delta^{-1} \qquad \text{for} \qquad \nu = 0, 1, 2, \dots .$$

In principle, we can repeat the construction of a modulated Fourier expansion described in Subsects. 4.2–4.3 on these time intervals, taking $\xi^{\nu\widetilde{n}}$ as initial value instead of ξ^0. This gives us modulation functions $\mathbf{z}^\nu$ on the νth time interval constructed in such a way that

$$\xi_j^n \approx \xi_j^\nu(t_n) = \sum_{\mathbf{k}} z_j^{\mathbf{k},\nu}(\delta t_n) \mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t_n} \qquad \text{for} \qquad \nu\delta^{-1} \le t_n \le (\nu+1)\delta^{-1}.$$

The estimates of Subsects. 4.4–4.7 remain valid provided that $\xi^{\nu\widetilde{n}}$ satisfies a smallness condition as ξ^0 (30). In the following lemma, we bound the difference of the modulated Fourier expansions $\mathbf{z}^\nu$ and $\mathbf{z}^{\nu-1}$ at the interface $\delta t_{\nu\widetilde{n}}$ of their intervals of validity.

Lemma 4.9. *Assume that for $\nu \ge 1$*

$$\|\xi^{(\nu-1)\widetilde{n}}\|_s \le 2\widehat{c}^{-1}\varepsilon \qquad \textit{and} \qquad \|\xi^{\nu\widetilde{n}}\|_s \le 2\widehat{c}^{-1}\varepsilon.$$

Then we have for $\delta \le \delta_0$ and $s \ge s_0$

$$|||\boldsymbol{\Gamma}^{s-\widehat{s}}\mathbf{z}^\nu(\delta t_{\nu\widetilde{n}}) - \boldsymbol{\Gamma}^{s-\widehat{s}}\mathbf{z}^{\nu-1}(\delta t_{\nu\widetilde{n}})|||_{\widehat{s}} \le C\varepsilon\delta^N$$

with constants C, s_0 and δ_0.

Proof. Let $n = \nu\widetilde{n}$.

(a) We first show by induction on $q = 0, \dots, N$ that

$$M_q^\nu(\mathbf{z}) := \Big|\Big|\Big| \sum_{p=0}^{q} \varepsilon\delta^p \big(\mathbf{z}_p^\nu(\delta t_n) - \mathbf{z}_p^{\nu-1}(\delta t_n)\big) \Big|\Big|\Big|_s \le C\varepsilon\delta^q, \tag{48}$$

where $\mathbf{z}_p^\nu = (z_{j,p}^{\mathbf{k},\nu})_{j\in\mathcal{Z},\mathbf{k}\in\mathbb{Z}^\mathcal{Z}}$. For this purpose we split the modulation functions, $\mathbf{z}_p = \mathbf{a}_p + \mathbf{b}_p$ with $a_{j,p}^{\langle j\rangle} = z_{j,p}^{\langle j\rangle}$ and $b_{j,p}^{\mathbf{k}} = z_{j,p}^{\mathbf{k}}$ for $\mathbf{k} \ne \langle j\rangle$.

We consider the nonlinearities $F_{j,p}^{\mathbf{k},\nu}$ and $F_{j,p}^{\mathbf{k},\nu-1}$ on the right-hand sides of (36a). By Lemma 3.3 and Lemma 4.1 we have

$$M_q^\nu(\mathbf{F}) \le C\varepsilon\delta^{-1} M_{q-1}^\nu(\mathbf{z}),$$

and hence by construction (38b) and (38e)

$$M_q^\nu(\mathbf{b}) \le C\delta M_{q-1}^\nu(\mathbf{z}) \qquad \text{and} \qquad M_q^\nu(\dot{\mathbf{a}}) \le C\delta M_{q-1}^\nu(\mathbf{z}). \tag{49}$$

In order to complete the inductive proof of (48), we need a similar estimate also for $M_q^\nu(\mathbf{a})$. Note that by construction (38d) of $\mathbf{z}^\nu$

$$\sum_{\mathbf{k}}\sum_{p=0}^{q} \varepsilon\delta^p z_{j,p}^{\mathbf{k},\nu}(\delta t_n)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t_n} = \sum_{\mathbf{k}}\sum_{p=0}^{q} \varepsilon\delta^p z_{j,p}^{\mathbf{k},\nu-1}(\delta t_n)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\varpi})t_n} + r_j^n$$

with $\|r^n\|_s \le C\varepsilon\delta^q$ by Proposition 4.3 (applied on the $(\nu-1)$th interval with modulation functions $z_j^{\mathbf{k},\nu-1}$ truncated after $p=q$ instead of $p=N$ as in (38f)). This shows that

$$\Big(\sum_{j\in\mathcal{Z}}|j|^{2s}\Big|\sum_{p=0}^{q}\varepsilon\delta^p\big(a_{j,p}^{\langle j\rangle,\nu}(\delta t_n)-a_{j,p}^{\langle j\rangle,\nu-1}(\delta t_n)\big)\Big|^2\Big)^{1/2}\le M_q^\nu(\mathbf{b})+C\varepsilon\delta^q, \tag{50}$$

and hence by (49)

$$M_q^\nu(\mathbf{a})\le C\delta M_{q-1}^\nu(\mathbf{z})+C\varepsilon\delta^q.$$

This completes the proof of (48).

(b) In order to prove the statement of the lemma, we have to consider

$$\widehat{M}_q^\nu(\mathbf{z}):=\Big|\Big|\Big|\sum_{p=0}^{q}\varepsilon\delta^p\mathbf{\Gamma}^{s-\widehat{s}}\big(\mathbf{z}_p^\nu(\delta t_n)-\mathbf{z}_p^{\nu-1}(\delta t_n)\big)\Big|\Big|\Big|_{\widehat{s}}$$

instead of $M_q^\nu(\mathbf{z})$. Note that $\widehat{M}_q^\nu(\mathbf{a})=M_q^\nu(\mathbf{a})$ and that the estimates (49) transfer to $\widehat{M}_q^\nu$ by (41). This finishes the proof of the lemma. □

The following proposition bounds, in the situation of the above lemma, the difference of the almost-invariants $\mathcal{I}_m$ (45) at the interface of two time intervals.

Proposition 4.10. *Assume that for $\nu\ge 1$*

$$\|\xi^{(\nu-1)\widetilde{n}}\|_s\le 2\widehat{c}^{-1}\varepsilon\qquad\text{and}\qquad\|\xi^{\nu\widetilde{n}}\|_s\le 2\widehat{c}^{-1}\varepsilon.$$

Then we have for $\delta\le\delta_0$ and $s\ge s_0$

$$\sum_{m\in\mathcal{N}}\max(1,m)^s\big|\mathcal{I}_m(\mathbf{z}^\nu(\delta t_{\nu\widetilde{n}}))-\mathcal{I}_m(\mathbf{z}^{\nu-1}(\delta t_{\nu\widetilde{n}}))\big|\le C\varepsilon^2\delta^N$$

with constants C, s_0 and δ_0.

Proof. This follows using part (b) of the proof of Proposition 4.6, the Cauchy-Schwarz inequality and the estimates of Lemma 4.1, Lemma 4.7 and Lemma 4.9. □

4.9 Long-time near-conservation of super-actions

We put the results of all previous subsections together to show near-conservation of the super-actions (46) on *long* time intervals of length δ^{-N}.

Theorem 4.11. *The super-actions are nearly conserved for $\delta\le\delta_0$ and $s\ge s_0$:*

$$\sum_{m\in\mathcal{N}}\max(1,m)^s\big|I_m(\xi^n)-I_m(\xi^0)\big|\le C\varepsilon^2\delta\qquad\text{for}\qquad 0\le t_n=nh\le\delta^{-N}$$

with constants C, s_0 and δ_0.

Proof. Let $C/5$ be the maximum of the constant of Proposition 4.10 and the constants that appear in Propositions 4.6 and 4.8 if $\widehat{c}^{-1}$ in (30) is replaced by $2\widehat{c}^{-1}$.

With this constant C, we prove the theorem for $\nu\delta^{-1}\le t_n\le(\nu+1)\delta^{-1}$ and $\nu=0,1,2,\dots$ by induction on ν. The main observation is that for $\nu\delta^{-1}\le t_n\le(\nu+1)\delta^{-1}$

and with $\widetilde{n} = 1/(\delta h)$ as in (47)

$$\begin{aligned}\big|I_m(\xi^n) - I_m(\xi^0)\big| \le{}& \big|I_m(\xi^n) - \mathcal{I}_m(\mathbf{z}^\nu(\delta t_n))\big| + \big|\mathcal{I}_m(\mathbf{z}^\nu(\delta t_n)) - \mathcal{I}_m(\mathbf{z}^\nu(\delta t_{\nu\widetilde{n}}))\big| \\ &+ \sum_{\widetilde{\nu}=1}^{\nu} \big|\mathcal{I}_m(\mathbf{z}^{\widetilde{\nu}}(\delta t_{\widetilde{\nu}\widetilde{n}})) - \mathcal{I}_m(\mathbf{z}^{\widetilde{\nu}-1}(\delta t_{\widetilde{\nu}\widetilde{n}}))\big| \\ &+ \sum_{\widetilde{\nu}=1}^{\nu} \big|\mathcal{I}_m(\mathbf{z}^{\widetilde{\nu}-1}(\delta t_{\widetilde{\nu}\widetilde{n}})) - \mathcal{I}_m(\mathbf{z}^{\widetilde{\nu}-1}(\delta t_{(\widetilde{\nu}-1)\widetilde{n}}))\big| \\ &+ \big|\mathcal{I}_m(\mathbf{z}^0(0)) - I_m(\xi^0)\big|.\end{aligned} \tag{51}$$

After multiplying by $\max(1,m)^s$ and summing over $m \in \mathcal{N}$ we can apply Propositions 4.6, 4.8 and 4.10 to the different terms in (51), since, in case $\nu \ge 1$, the induction hypothesis implies for $0 \le n \le \nu\widetilde{n}$

$$\|\xi^n\|_s^2 \le \|\xi^0\|_s^2 + C_1^s \sum_{m\in\mathcal{N}} \max(1,m)^s \big|I_m(\xi^n) - I_m(\xi^0)\big| \le \widehat{c}^{\,-2}\varepsilon^2 + C_1^s C\varepsilon^2\delta \le 2\widehat{c}^{\,-2}\varepsilon^2$$

provided that $\delta \le 1/(C_1^s C\widehat{c}^{\,2})$ with the constant C_1 of Lemma 4.7. This gives

$$\sum_{m\in\mathcal{N}} \max(1,m)^s \big|I_m(\xi^n) - I_m(\xi^0)\big| \le \tfrac{2}{5}C\varepsilon^2\delta + \tfrac{1}{5}C(2\nu+1)\varepsilon^2\delta^N.$$

The statement of the theorem follows for $\nu \le \delta^{-N+1}$, i.e., $t_n \le \delta^{-N}$. □

4.10 Proof of Theorem 2.1

In order to complete the proof of Theorem 2.1, we go in a final step back from the new variables ξ introduced in Sect. 3 to the original variables u, in which the split-step Fourier method (4) is formulated. Under the conditions and with the constant C of Theorem 4.11, we have for $0 \le t_n = nh \le \delta^{-N}$ and $\delta \le 1/(C_1^s C\widehat{c}^{\,2})$

$$\|\xi^n\|_s^2 \le \|\xi^0\|_s^2 + C_1^s \sum_{m\in\mathcal{N}} \max(1,m)^s \big|I_m(\xi^n) - I_m(\xi^0)\big| \le 2\widehat{c}^{\,-2}\varepsilon^2$$

with the constant C_1 of Lemma 4.7. By Lemma 3.2, this transfers to a statement in the original variables,

$$\|\mathcal{F}_{\neg\ell}(u^m)\|_s \le \widehat{C}\|\xi^n\|_s \le \sqrt{2}\widehat{C}\widehat{c}^{\,-1}\varepsilon \qquad \text{for} \qquad 0 \le t_n = nh \le \delta^{-N},$$

as claimed in Theorem 2.1.

5 On the non-resonance condition

In this section, we give the proof of Theorem 2.3 on a sufficient condition under which Assumptions 1 and 2 in Theorem 2.1 hold. The first subsection deals with the (numerical) linear stability of Assumption 1, while the remaining main part of this section is devoted to the non-resonance condition of Assumption 2.

From now on, we let $\ell = 0$. In this case, we have $n(j) = |j|^2$, and the frequencies (10) become

$$\omega_j = \frac{\arccos\big(\cos(|j|^2 h) - h\lambda\rho^2 \sin(|j|^2 h)\big)}{h\,\mathrm{sgn}\big(\sin(|j|^2 h) + h\lambda\rho^2\cos(|j|^2 h)\big)} \tag{52}$$

for $j \in \mathcal{Z}$. We introduce the set of possible values of $n(j)$:

$$\mathcal{N} = \{\, n(j) : j \in \mathcal{Z}\,\} = \{\, |j|^2 : j \in \mathcal{Z}\,\}.$$

5.1 Linear stability

We show that Assumption 1 is fulfilled, for $\ell = 0$, under the conditions (11) and (12) of Theorem 2.3.

Lemma 5.1. *Under the step-size restriction (12) the condition (11) of analytical linear stability implies the condition (8) of numerical linear stability for $0 \le \rho \le \rho_0$ with $c_1 = c_1(\rho_0)$. Moreover,*

$$1 - h^2\rho^4 + \frac{2\lambda\rho^2}{\mu_n} \ge \min\big(\tfrac{1}{2}, 1 + 2\lambda\rho_0^2\big) > 0 \qquad \textit{for all} \qquad n \in \mathcal{N} \tag{53}$$

with

$$\mu_n = \frac{\sin(nh)}{h\cos(nh)} = \frac{\tan(nh)}{h}. \tag{54}$$

Proof. We note that for all $n \in \mathcal{N}$

$$0 \le \frac{1}{\tan(nh)} \le \frac{1}{\tan(h)} \le \frac{1}{h} - \frac{h}{3}$$

by the step-size restriction (12). This yields for $\lambda = +1$

$$1 - h^2\rho^4 + \frac{2\lambda\rho^2}{\mu_n} \ge 1 - h^2\rho^4$$

and for $\lambda = -1$

$$1 - h^2\rho^4 + \frac{2\lambda\rho^2}{\mu_n} \ge 1 - h^2\rho^4 - 2h\rho^2\Big(\frac{1}{h} - \frac{h}{3}\Big) \ge (1 - 2\rho^2)\Big(1 + \frac{h^2\rho^2}{2}\Big).$$

We hence get (53) for $0 \le \rho \le \rho_0$ from (11) and (12).

The estimate (53) together with (12) implies that there exists $c_1 = c_1(\rho_0)$ such that

$$c_1 h^2 \le \sin(nh)^2\Big(1 - h^2\rho^4 + \frac{2\lambda\rho^2}{\mu_n}\Big) = 1 - \big(\cos(nh) - h\lambda\rho^2\sin(nh)\big)^2.$$

for all $n \in \mathcal{N}$, and hence (8) holds. □

5.2 Modified frequencies

Now, we turn to Assumption 2, again for $\ell = 0$. We begin by constructing the modified frequencies. The reason, why we use modified frequencies in the theory developed in the present paper, is that it seems to be very hard to verify the non-resonance condition in part (b) of Assumption 2 directly for the frequencies ω_j of (52). For the frequencies that show up after the linearization of the nonlinear Schrödinger equation itself around a plane wave, however, a suitable non-resonance condition can be established, see [7, Lemma 2.2]. These frequencies are $\sqrt{|j|^4 + 2\lambda\rho^2|j|^2}$, and we therefore seek modified frequencies of a similar form.

We fix $\rho_0 > 0$ with (11) and h and K with (12) for some $N \ge 2$ as in Theorem 2.3. The frequencies ω_j of (52) are considered henceforth as functions of $\sigma = \rho^2$ with $0 \le \sigma \le \sigma_0 := \rho_0^2$:

$$\omega_j = \omega_j(\sigma) = \frac{\arccos\big(\cos(|j|^2h) - h\lambda\sigma\sin(|j|^2h)\big)}{h}. \tag{55}$$

(Note that the step-size restriction (12) together with (53) ensures that the sign of ω_j in (52) is positive.)

The derivative of the frequencies ω_j with respect to σ is given by

$$\frac{\mathrm{d}\omega_j(\sigma)}{\mathrm{d}\sigma} = \frac{\lambda}{\sqrt{1 - h^2\sigma^2 + \frac{2\lambda\sigma}{\mu_{|j|^2}}}}$$

with $\mu_{|j|^2}$ from (54) which is positive for $j \in \mathcal{Z}$ by (12). This motivates the definition

$$\varpi_j = \varpi_j(\sigma) = |j|^2 - \mu_{|j|^2} + \sqrt{\mu_{|j|^2}^2 + 2\lambda\sigma\mu_{|j|^2}}, \qquad j \in \mathcal{Z}, \tag{56}$$

of the *modified frequencies* since we then have

$$\frac{\mathrm{d}\varpi_j(\sigma)}{\mathrm{d}\sigma} = \frac{\lambda}{\sqrt{1 + \frac{2\lambda\sigma}{\mu_{|j|^2}}}} \qquad \text{and} \qquad \varpi_j(0) = \omega_j(0).$$

This implies

$$\frac{\mathrm{d}(\omega_j - \varpi_j)(\sigma)}{\mathrm{d}\sigma} = \frac{\lambda h^2\sigma^2}{2(1 - \xi + \frac{2\lambda\sigma}{\mu_{|j|^2}})^{3/2}}$$

for all $j \in \mathcal{Z}$ and some $0 \le \xi = \xi_j \le h^2\sigma^2$, and hence

$$|\omega_j - \varpi_j| \le C_2 h^2 \qquad \text{for all} \qquad j \in \mathcal{K} \tag{57}$$

with $C_2 = C_2(\sigma_0)$ by (53). The modified frequencies ϖ_j are hence close to the original frequencies ω_j as required in part (a) of Assumption 2.

5.3 Bambusi's non-resonance condition for the modified frequencies

We study resonances among the modified frequencies ϖ_j of (56) derived in the previous subsection. As $\varpi_j = \varpi_l$ for $|j|^2 = |l|^2$, we introduce

$$\Omega_n = \Omega_n(\sigma) = \varpi_j(\sigma) \qquad \text{for} \qquad n \in \mathcal{N},\ j \in \mathcal{Z} \qquad \text{with} \qquad n = |j|^2. \tag{58}$$

We verify for these modified frequencies a non-resonance condition that has been introduced by Bambusi and is widely used in the long-time analysis of infinite dimensional Hamiltonian systems. The verification is an adaptation of [3, Sect. 5.1] to the present situation along with some simplifications.

We proceed roughly as follows. The aim is to show that there are a lot of "good" values of σ for which linear combinations of frequencies do not become small. More precisely, a value of σ is considered as "good" if for all vectors $\mathbf{k} \in \mathbb{Z}^{\mathcal{N}}$ and $\mathbf{l} \in \mathbb{Z}^{\mathcal{N}}$ with $\|\mathbf{k}\| \le N$ and $\|\mathbf{l}\| \le 2$ the linear combinations

$$\mathbf{k} \cdot \mathbf{\Omega} + \mathbf{l} \cdot \mathbf{\Omega} = \sum_{n \in \mathcal{N}} k_n \Omega_n + \sum_{n \in \mathcal{N}} l_n \Omega_n$$

are bounded away from zero by a negative power of $\mathrm{argmax}(\mathbf{k})$, where we denote by $\mathrm{argmax}(\mathbf{k})$ the largest index $n \in \mathcal{N}$ with $k_n \neq 0$ (and set $\mathrm{argmax}(\mathbf{k}) = 1$ for $\mathbf{k} = \mathbf{0}$). The first step is to observe that it suffices to consider

$$\mathbf{k} \cdot \mathbf{\Omega} + m$$

with integers m instead of $\mathbf{k}\cdot\boldsymbol{\Omega}+\mathbf{l}\cdot\boldsymbol{\Omega}$, the reason being that $\mathbf{l}\cdot\boldsymbol{\Omega}=\pm\Omega_n\pm\Omega_{n'}$ is either close to an integer by the asymptotic behaviour $\Omega_n\sim n+\lambda\sigma$ of the frequencies (see Lemma 5.2 below) or may be absorbed into $\mathbf{k}\cdot\boldsymbol{\Omega}$. This is done in Proposition 5.5 below. Furthermore, if one excludes some values of σ, a bound of $\mathbf{k}\cdot\boldsymbol{\Omega}+m$ can be obtained from a bound of some derivative

$$\frac{\mathrm{d}^k(\mathbf{k}\cdot\boldsymbol{\Omega}+m)}{\mathrm{d}\sigma^k}$$

of $\mathbf{k}\cdot\boldsymbol{\Omega}+m$ (Lemma 5.4). Therefore we study in Lemma 5.3 a matrix made up of derivatives of the frequencies Ω_n. This matrix is such that its inverse multiplied with the vector containing the first derivatives of $\mathbf{k}\cdot\boldsymbol{\Omega}$ is just the vector containing the nonzero entries of $\mathbf{k}$. Bounding its inverse (see Lemma 5.3) thus helps to study the derivatives of $\mathbf{k}\cdot\boldsymbol{\Omega}+m$.

As in the previous subsection we fix $\rho_0>0$ with (11) and h and K with (12) for some $N\geq 2$. Let us emphasize, however, that again all constants will be independent of the discretization parameters h and K. We will make extensive use of the asymptotic behaviour of μ_n from (54) and of the modified frequencies described in the following lemma.

Lemma 5.2. *We have for* $0\leq\rho\leq\rho_0$

$$n\leq\mu_n\leq Cn\qquad\text{and}\qquad -\frac{C}{n}\leq\Omega_n-n-\lambda\sigma\leq 0\qquad\text{for}\qquad n\in\mathcal{N}$$

with a constant $C=C(\sigma_0)$.

Proof. The estimates of μ_n follow from $nh\leq\tan(nh)\leq Cnh$ by (12). For the estimates of Ω_n we note that

$$\Omega_n-n-\lambda\sigma=\sqrt{\big(\mu_n+\lambda\sigma\big)^2-\sigma^2}-\big(\mu_n+\lambda\sigma\big).$$

This shows

$$0\geq\Omega_n-n-\lambda\sigma\geq\frac{-\sigma_0^2}{2\mu_n\sqrt{1+\frac{2\lambda\sigma}{\mu_n}}}.$$

The estimate (53) of Lemma 5.1 and $\mu_n\geq n$ thus lead to the claimed lower bound of $\Omega_n-n-\lambda\sigma$. □

Now we begin with the investigation of (integer) linear combinations $\mathbf{k}\cdot\boldsymbol{\Omega}=\sum_{n\in\mathcal{N}}k_n\Omega_n$ of modified frequencies (58). The following lemma will help us to control derivatives of these linear combinations with respect to σ, which in turn will allow us to control the linear combinations themselves.

Lemma 5.3. *Let* $1\leq n_1<n_2<\cdots<n_M$, *and let* $A=(a_{kl})_{k,l=1}^M$ *be the matrix with entries*

$$a_{kl}=\frac{\mathrm{d}^k\Omega_{n_l}(\sigma)}{\mathrm{d}\sigma^k}.$$

Then for all $k,l=1,\dots,M$ *and all* $0\leq\sigma\leq\sigma_0$

$$|a_{kl}|\leq Cn_l^{-k+1}\qquad\text{and}\qquad\|A^{-1}\|_\infty\leq Cn_M^{2M}$$

with a constant $C=C(M,\sigma_0)$.

Proof. We have

$$a_{kl} = \frac{\mathrm{d}^k \Omega_{n_l}(\sigma)}{\mathrm{d}\sigma^k} = d_k e_l x_l^{k-1} \tag{59}$$

with $d_1 = \lambda$ and $d_{k+1} = -\lambda(2k-1)d_k$ for $k \geq 1$, $e_l = 1/\sqrt{1 + 2\lambda\sigma/\mu_{n_l}}$ and $x_l = e_l^2/\mu_{n_l}$. Note that for all $k, l = 1, \ldots, M$

$$c' \leq |d_k| \leq C', \quad c' \leq e_l \leq C' \qquad \text{and} \qquad \frac{c'}{n_l} \leq x_l \leq \frac{C'}{n_l} \tag{60}$$

with positive constants $c' = c'(\sigma_0)$ and $C' = C'(M, \sigma_0)$ by Lemmas 5.1 and 5.2. Hence, the bound on the entry a_{kl} as stated in the lemma follows from the representation (59).

Moreover, this representation shows that

$$A = DVE$$

with the diagonal matrices $D = \operatorname{diag}(d_k)_{k=1}^M$ and $E = \operatorname{diag}(e_l)_{l=1}^M$ and the Vandermonde matrix $V = (x_l^{k-1})_{k,l=1}^M$. In order to examine the inverse of A, we first invert V. Its inverse is given by

$$V^{-1} = \left(\frac{v_{ij}}{w_i}\right)_{i,j=1}^M$$

with

$$v_{ij} = \sum_{\substack{1 \leq l_1 < \cdots < l_{M-j} \leq M \\ l_1, \ldots, l_{M-j} \neq i}} (-1)^j x_{l_1} \cdots x_{l_{M-j}} \qquad \text{and} \qquad w_i = \prod_{\substack{1 \leq l \leq M \\ l \neq i}} (x_i - x_l),$$

see for example [23, Sect. 2.8.1]. Since

$$|x_i - x_l| = x_i x_l |\mu_{n_l} - \mu_{n_i}| = x_i x_l \frac{|n_l - n_i|}{\cos^2(\xi h)}$$

with $\min(n_i, n_l) \leq \xi \leq \max(n_i, n_l)$, the bounds (60) and the step-size restriction (12) imply

$$\|V^{-1}\|_\infty \leq C n_M^{2M}, \quad \|D^{-1}\|_\infty \leq C \qquad \text{and} \qquad \|E^{-1}\|_\infty \leq C$$

with $C = C(M, \sigma_0)$. The estimate of $\|A^{-1}\|_\infty$ stated in the lemma follows. □

Now we consider sets of values of σ for which linear combinations of modified frequencies are small. We define for vectors $\mathbf{k} \in \mathbb{Z}^\mathcal{N}$ and $\mathbf{l} \in \mathbb{Z}^\mathcal{N}$ and integers m the sets

$$\mathcal{Q}_{\mathbf{k},\mathbf{l},m}(\gamma, \alpha) = \left\{ \sigma \in [0, \sigma_0] : \left|(\mathbf{k} + \mathbf{l}) \cdot \boldsymbol{\Omega}(\sigma) + m\right| < \frac{\gamma}{\operatorname{argmax}(\mathbf{k})^\alpha} \right\}. \tag{61}$$

We first estimate the Lebesgue measure $|\cdot|$ of these sets in the case $\mathbf{l} = \mathbf{0}$.

Lemma 5.4. *There exists a constant $C = C(N, \alpha, \sigma_0)$ such that for all $0 < \gamma \leq 1$, all $\|\mathbf{k}\| \leq N$ and all $m \in \mathbb{Z}$ with $\|k\| + |m| \neq 0$*

$$|\mathcal{Q}_{\mathbf{k},\mathbf{0},m}(\gamma, \alpha)| \leq \frac{C\gamma^{1/M}}{\operatorname{argmax}(\mathbf{k})^{\alpha/M - 4M}},$$

where M denotes the number of nonzero entries of $\mathbf{k}$.

Proof. We fix a vector $\mathbf{k}$ with $\|\mathbf{k}\| \le N$. We may assume $\mathbf{k} \neq \mathbf{0}$ because the statement is trivial for $\mathbf{k} = \mathbf{0}$ since $\gamma \le 1$.

(a) Lemma 5.3 shows that there exists a constant $C = C(N, \sigma_0)$ such that for any $0 \le \sigma \le \sigma_0$ there exists $1 \le k \le M$ with

$$\left| \frac{\mathrm{d}^k (\mathbf{k} \cdot \boldsymbol{\Omega})(\sigma)}{\mathrm{d}\sigma^k} \right| \ge C \operatorname{argmax}(\mathbf{k})^{-2M}. \tag{62}$$

(b) The function $g : [0, \sigma_0] \to \mathbb{R}$, $\sigma \mapsto \mathbf{k} \cdot \Omega(\sigma) + m$ is infinitely differentiable and its first $M+1$ derivatives are uniformly bounded on $[0, \sigma_0]$ by a constant depending only on σ_0 and N (Lemma 5.3). The property (62) then enables us to apply [2, Lemma 8.4], which yields the statement of the lemma. □

Setting

$$\mathcal{Q}(\gamma, \alpha) = \bigcup_{\substack{\mathbf{k}:\|\mathbf{k}\| \le N \\ \mathbf{l}:\|\mathbf{l}\| \le 2 \\ \mathbf{k}+\mathbf{l} \neq \mathbf{0}}} \mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha) \tag{63}$$

with $\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha)$ from (61) we can now prove the following non-resonance result in the spirit of Bambusi's non-resonance condition, see [3, Lemma 5.7].

Proposition 5.5. *Let $\alpha \ge (5N)^4$. Then there exists a constant $C = C(N, \alpha, \sigma_0)$ such that for all $0 < \gamma \le 1$*

$$|\mathcal{Q}(\gamma, \alpha)| \le C \gamma^{1/(2\sqrt{\alpha}(N+2))}.$$

Proof. We consider the sets $\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha)$ for vectors $\mathbf{l}$ with $\|\mathbf{l}\| \le 2$. Throughout this discussion we fix $0 < \gamma \le 1$ and $\mathbf{k}$ with $\|\mathbf{k}\| \le N$.

(a) For the vector $\mathbf{l} = \mathbf{0}$ the measure of the set $\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha)$ can be estimated with Lemma 5.4.

(b) For $\mathbf{l} = \pm\langle n \rangle$ let $c' \ge 1$ be a constant such that $|\pm\Omega_n + \mathbf{k} \cdot \boldsymbol{\Omega}| \ge 1$ if $n > c' \operatorname{argmax}(\mathbf{k})$. This constant exists by Lemma 5.2 and depends on σ_0 and N. Then for $n \le c' \operatorname{argmax}(\mathbf{k})$

$$\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha) \subseteq \mathcal{Q}_{\mathbf{k}+\mathbf{l},\mathbf{0},0}\big((c')^\alpha \gamma, \alpha\big),$$

whereas $\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha) = \emptyset$ for $n > c' \operatorname{argmax}(\mathbf{k})$.

(c) For $\mathbf{l} = \pm(\langle n \rangle + \langle n' \rangle)$ let similarly be $c'' = c''(N, \sigma_0) \ge 1$ be a constant such that $|\pm(\Omega_n + \Omega_{n'}) + \mathbf{k} \cdot \boldsymbol{\Omega}| \ge 1$ if $n + n' > c'' \operatorname{argmax}(\mathbf{k})$. Then for $n + n' \le c'' \operatorname{argmax}(\mathbf{k})$

$$\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha) \subseteq \mathcal{Q}_{\mathbf{k}+\mathbf{l},\mathbf{0},0}\big((c'')^\alpha \gamma, \alpha\big),$$

whereas $\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha) = \emptyset$ for $n + n' > c'' \operatorname{argmax}(\mathbf{k})$.

(d) For $\mathbf{l} = \pm(\langle n \rangle - \langle n' \rangle)$, where without loss of generality $n < n'$, note that with the constant C of Lemma 5.2

$$|\mathbf{l} \cdot \boldsymbol{\Omega} - m| \le \frac{2C}{n} \qquad \text{for} \qquad m = \pm(n \pm n').$$

Then for $n \ge C \operatorname{argmax}(\mathbf{k})^{\sqrt{\alpha}} / \gamma^{1/(2\sqrt{\alpha})}$

$$\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha) \subseteq \mathcal{Q}_{\mathbf{k},\mathbf{0},m}\big(3\gamma^{1/(2\sqrt{\alpha})}, \sqrt{\alpha}\big),$$

and this set is empty for $|m| \ge c''' \operatorname{argmax}(\mathbf{k})$ with a constant $c''' = c'''(N, \sigma_0)$ by Lemma 5.2. On the other hand, we have for $n < C \operatorname{argmax}(\mathbf{k})^{\sqrt{\alpha}} / \gamma^{1/(2\sqrt{\alpha})}$

$$\mathcal{Q}_{\mathbf{k},\mathbf{l},0}(\gamma, \alpha) \subseteq \mathcal{Q}_{\mathbf{k} \pm \langle n \rangle, \mp \langle n' \rangle, 0}(C^{\sqrt{\alpha}} \sqrt{\gamma}, \sqrt{\alpha}),$$

a situation that is covered by (b).

The results (a)-(d) show that there exists a constant $c = c(N, \alpha, \sigma_0)$ such that

$$\mathcal{Q}(\gamma,\alpha) \subseteq \bigcup_{\substack{\mathbf{k}:\|\mathbf{k}\|\le N+2\\ m\in\mathbb{Z}:|m|<c'''\operatorname{argmax}(\mathbf{k})\\ \|\mathbf{k}\|+|m|\neq 0}} \mathcal{Q}_{\mathbf{k},\mathbf{0},m}\big(c\gamma^{1/(2\sqrt{\alpha})},\sqrt{\alpha}\big).$$

Since the number of vectors $\mathbf{k}$ with $\|\mathbf{k}\| \le N+2$ and $\operatorname{argmax}(\mathbf{k}) = L$ is at most $(N+3)L^{N+2}$ we have by Lemma 5.4

$$|\mathcal{Q}(\gamma,\alpha)| \le C\gamma^{1/(2\sqrt{\alpha}(N+2))} \sum_{L=1}^{\infty} L^{5N+11-\sqrt{\alpha}/(N+2)}$$

with a constant $C = C(N, \alpha, \sigma_0)$. The choice of α ensures that $\sqrt{\alpha} \ge (N+2)(5N+13)$, and hence the latter sum converges and the proposition is proven. □

Remark 5.6 (case $\ell \neq 0$)**.** For $\ell \neq 0$ but small, the frequencies ω_j from (10) are different from those for $\ell = 0$ only for large j. For these large j, we have to deal with two differences.

First, the frequencies of (10) (and also the modified frequencies) contain an additional summand $\frac{1}{2}|\ell+j \bmod 2K|^2 - \frac{1}{2}|\ell-j \bmod 2K|^2$. This is an integer and does not affect the proof of Lemmas 5.3 and 5.4, where also the integer summand $|j|^2 - \mu_{|j|^2}$ in the modified frequencies (56) does not pose a problem. It does neither pose a problem in the proof of Proposition 5.5 since it is of order one for small ℓ (for part (b) and (c) of the proof) and is an integer (for part (d) of the proof).

Second, the quantity $n(j)$ appearing in the frequencies of (10) can be different from $|j|^2$. But for small ℓ, these two quantities are of the same order, see Lemma 4.7. We therefore expect the statements of Lemmas 5.3 and 5.4 and of Proposition 5.5 to transfer to this situation with constants depending on ℓ.

5.4 Proof of Theorem 2.3

We have already verified in Lemma 5.1 that Assumption 1 is satisfied under the conditions (11) and (12) of Theorem 2.3. We have also verified in (57) that the modified frequencies (56) are close to the original frequencies as required in part (a) of Assumption 2. We will now prove that they satisfy the non-resonance condition in part (b) of Assumption 2 for many values of h and ρ. Note that, in the considered case $\ell = 0$, part (c) of Assumption 2 follows from part (b) since $\varpi_j = \varpi_l$ for all $j, l \in \mathcal{Z}$ with $n(j) = |j|^2 = |l|^2 = n(l)$.

Fix $\rho_0 > 0$ with (11), $h_0 > 0$ and N. In contrast to the previous subsection, we do not fix the time step-size h anymore. We consider for all $0 < h \le h_0$ the corresponding sets (63) for $\sigma_0 = \rho_0^2$ which we denote now by $\mathcal{Q}_h(\gamma, \alpha)$ to emphasize the dependence (of the modified frequencies, and hence the sets) on h. We set for $0 < \gamma \le 1$

$$\mathcal{P}(\gamma) = \big\{\, (h,\rho) \in [0,h_0] \times [0,\rho_0] : \rho^2 \notin \mathcal{Q}_h(\gamma,\alpha) \,\big\}$$

with $\alpha = (5N)^4$. As mentioned above, all $(h, \rho) \in \mathcal{P}(\gamma)$ satisfy Assumption 1 with constant $c_1 = c_1(\rho_0)$ and part (a) of Assumption 2 with $\widehat{\varepsilon} = C_2 h^2$ and constant $C_2 = C_2(\rho_0)$ provided that K satisfies (12).

We still have to show that for all $(h, \rho) \in \mathcal{P}(\gamma)$ the modified frequencies $\Omega_n = \Omega_n(\rho^2)$ satisfy the non-resonance condition in part (b) of Assumption 2 provided

that (12) holds. For this purpose let $\mathbf{k} \in \mathbb{Z}^{\mathcal{N}}$ with $\|\mathbf{k}\| \le N+1$. Then we have

$$\frac{2}{\pi} |\mathbf{k} \cdot \boldsymbol{\Omega}| \le \left| \frac{\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\Omega})h} - 1}{h} \right| =: \delta \tag{64}$$

since the (strong[1]) step-size restriction (12) ensures together with Lemma 5.2 that $|\mathbf{k} \cdot \boldsymbol{\Omega}| h \le \pi$. Now we write

$$\mathbf{k} \cdot \boldsymbol{\Omega} = \pm \Omega_{n_M} \pm \Omega_{n_{M-1}} \pm \cdots \pm \Omega_{n_1}$$

with $n_M \ge n_{M-1} \ge \cdots \ge n_1$ in such a way that there is no pairwise cancellation ($M = \|\mathbf{k}\|$). For $\delta \le 1$ we have by (64) and by Lemma 5.2 that $n_M \le c n_{M-1}$ with $c = c(N\rho_0)$. Moreover, the choice of the set $\mathcal{Q}_h(\gamma, \alpha)$ yields

$$|\mathbf{k} \cdot \boldsymbol{\Omega}| \ge \gamma \left(\frac{n_M^2}{c\, n_M n_{M-1} \cdots n_1} \right)^{\alpha}.$$

Combining this with (64) we get

$$\left(\frac{n_M^2}{\prod_{n \in \mathcal{N}} n^{|k_n|}} \right)^{\alpha} \le \frac{c^\alpha \pi}{2\gamma}\, \delta.$$

The non-resonance condition of Assumption 2 thus holds for $c_2 = c_2(N, \gamma, \rho_0)$, $\delta_2 = 1$ and $s_2 = \alpha N$.

We finally have to estimate the Lebesgue measure of $\mathcal{P}(\gamma)$. By Fubini's theorem and Proposition 5.5 we have

$$|\mathcal{P}(\gamma)| = \rho_0 h_0 - \int_0^{h_0} \mathcal{Q}_h(\gamma, \alpha)\, dh \ge \rho_0 h_0 - C h_0 \gamma^{1/(2\sqrt{\alpha}(N+2))}$$

because the constant in this proposition is independent of h. The proof of Theorem 2.3 is thus complete if we redefine γ.

References

[1] G. P. Agrawal, *Nonlinear fiber optics*, fifth ed., Academic Press, 2013.

[2] D. Bambusi, *On long time stability in Hamiltonian perturbations of non-resonant linear PDEs*, Nonlinearity **12** (1999), 823–850.

[3] D. Bambusi and B. Grébert, *Birkhoff normal form for partial differential equations with tame modulus*, Duke Math. J. **135** (2006), 507–567.

[4] B. Cano and A. González-Pachón, *Plane waves numerical stability of some explicit exponential methods for cubic Schrödinger equation*, Preprint, 2013.

[5] D. Cohen, E. Hairer and Ch. Lubich, *Long-time analysis of nonlinearly perturbed wave equations via modulated Fourier expansions*, Arch. Ration. Mech. Anal. **187** (2008), 341–368.

[6] M. Dahlby and B. Owren, *Plane wave stability of some conservative schemes for the cubic Schrödinger equation*, M2AN Math. Model. Numer. Anal. **43** (2009), 677–687.

[7] E. Faou, L. Gauckler and Ch. Lubich, *Sobolev stability of plane wave solutions to the cubic nonlinear Schrödinger equations on a torus*, Comm. Partial Differential Equations **38** (2013), 1123–1140.

[1] This is the first and only place, where we need that the right-hand side of (12) is $\pi/(N+1)$ and not only $\pi/3$, say.

[8] E. Faou, *Geometric numerical integration and Schrödinger equations*, Zurich Lectures in Advanced Mathematics, European Mathematical Society (EMS), Zürich, 2012.

[9] E. Faou, B. Grébert and E. Paturel, *Birkhoff normal form for splitting methods applied to semilinear Hamiltonian PDEs. I. Finite-dimensional discretization*, Numer. Math. **114** (2010), 429–458.

[10] E. Faou, B. Grébert and E. Paturel, *Birkhoff normal form for splitting methods applied to semilinear Hamiltonian PDEs. II. Abstract splitting*, Numer. Math. **114** (2010), 459–490.

[11] L. Gauckler, *Long-time analysis of Hamiltonian partial differential equations and their discretizations*, Dissertation (doctoral thesis), Universität Tübingen, 2010, http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-47540.

[12] L. Gauckler, E. Hairer and Ch. Lubich, *Energy separation in oscillatory Hamiltonian systems without any non-resonance condition*, Comm. Math. Phys. **321** (2013), 803–815.

[13] L. Gauckler, E. Hairer, Ch. Lubich and D. Weiss, *Metastable energy strata in weakly nonlinear wave equations*, Comm. Partial Differential Equations **37** (2012), 1391–1413.

[14] L. Gauckler and Ch. Lubich, *Splitting integrators for nonlinear Schrödinger equations over long times*, Found. Comput. Math. **10** (2010), 275–302.

[15] E. Hairer and Ch. Lubich, *Long-time energy conservation of numerical methods for oscillatory differential equations*, SIAM J. Numer. Anal. **38** (2000), 414–441 (electronic).

[16] E. Hairer and Ch. Lubich, *On the energy distribution in Fermi-Pasta-Ulam lattices*, Arch. Ration. Mech. Anal. **205** (2012), 993–1029.

[17] E. Hairer, Ch. Lubich and G. Wanner, *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*, second ed., Springer Series in Computational Mathematics, vol. 31, Springer-Verlag, Berlin, 2006.

[18] Z. Hani, *Long-time instability and unbounded Sobolev orbits for some periodic nonlinear Schrödinger equations*, Arch. Ration. Mech. Anal. (to appear), doi:10.1007/s00205-013-0689-6.

[19] R. H. Hardin and F. D. Tappert, *Applications of the split-step Fourier method to the numerical solution of nonlinear and variable coefficient wave equations*, SIAM Rev. **15** (1973), 423.

[20] S. Jin, P. Markowich and Ch. Sparber, *Mathematical and computational methods for semiclassical Schrödinger equations*, Acta Numer. **20** (2011), 121–209.

[21] M. Khanamiryan, O. Nevanlinna and T. Vesanen, *Long-term behavior of the numerical solution of the cubic non-linear Schrödinger equation using Strang splitting method*, Preprint, 2012, http://www.damtp.cam.ac.uk/user/na/people/Marianna/papers/NLS.pdf.

[22] T. I. Lakoba, *Instability of the split-step method for a signal with nonzero central frequency*, J. Opt. Soc. Am. B **30** (2013), 3260–3271.

[23] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical recipes. The art of scientific computing*, third ed., Cambridge University Press, Cambridge, 2007.

[24] T. R. Taha and M. J. Ablowitz, *Analytical and numerical aspects of certain nonlinear evolution equations. II. Numerical, nonlinear Schrödinger equation*, J. Comput. Phys. **55** (1984), 203–230.

[25] J. A. C. Weideman and B. M. Herbst, *Split-step methods for the solution of the nonlinear Schrödinger equation*, SIAM J. Numer. Anal. **23** (1986), 485–507.