BIT Numer Math (2012) 52:877–903
DOI 10.1007/s10543-012-0385-1

One-stage exponential integrators for nonlinear Schrödinger equations over long times

David Cohen · Ludwig Gauckler

Received: 11 July 2011 / Accepted: 5 April 2012 / Published online: 9 May 2012

Abstract Near-conservation over long times of the actions, of the energy, of the mass and of the momentum along the numerical solution of the cubic Schrödinger equation with small initial data is shown. Spectral discretization in space and one-stage exponential integrators in time are used. The proofs use modulated Fourier expansions.

Keywords Nonlinear Schrödinger equation · Exponential integrators · Long-time behavior · Near-conservation of actions, energy, mass and momentum · Modulated Fourier expansion

Mathematics Subject Classification 65P10 · 65M70

1 Introduction

We consider the cubic Schrödinger equation with a potential of convolution type

$$\mathrm{i}\frac{\partial}{\partial t}u = -\Delta u + V * u + |u|^2 u, \tag{1.1}$$

where $u = u(x,t)$ with $x \in \mathbb{T}^d = \mathbb{R}^d/2\pi\mathbb{Z}^d$ and $t \geq 0$, in dimension $d \geq 1$ with periodic boundary conditions. We will consider small initial data: in appropriate Sobolev norms, the initial value $u(\cdot,0)$ is bounded by a small parameter ε. The potential

Communicated by Mechthild Thalhammer.

D. Cohen (✉)
Mathematisches Institut, Universität Basel, 4051 Basel, Switzerland
e-mail: David.Cohen@unibas.ch

L. Gauckler
Institut für Mathematik, TU Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: gauckler@math.tu-berlin.de

$V = V(x) \in L^2(\mathbb{T}^d)$ is assumed to be periodic with real Fourier coefficients. It acts by convolution on the function u. Such equations have been studied for example in [1, 2, 12].

It is known that this Hamiltonian partial differential equation possesses the following invariants, that follow from invariances of the equation under certain transformations, see for example the monograph [25, Sect. I.2.3]. For H^1-solutions, one has conservation of the total *energy* or Hamiltonian

$$H(u,\bar{u}) = \frac{1}{2(2\pi)^d} \int_{\mathbb{T}^d} \left(|\nabla u|^2 + (V * u)\bar{u} + \frac{1}{2}|u|^4 \right) \mathrm{d}x, \tag{1.2}$$

where $|\cdot|$ denotes the Euclidean norm. The L^2-norm, density, or *mass*

$$m(u,\bar{u}) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} |u|^2 \, \mathrm{d}x \tag{1.3}$$

is also a conserved quantity. Finally, the *momentum*

$$K(u,\bar{u}) = \mathrm{i}\frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} (u\nabla\bar{u} - \bar{u}\nabla u) \, \mathrm{d}x \tag{1.4}$$

is exactly conserved along the solution of our partial differential equation (1.1). But this is not all, this equation offers also another interesting geometric property which will turn out to be useful for our numerical analysis: It is reversible with respect to the complex conjugation ρ of the Fourier coefficients,

$$\mathrm{i}\frac{\partial}{\partial t}\rho(u) = -\big(-\Delta\rho(u) + V * \rho(u) + \big|\rho(u)\big|^2 \rho(u)\big)$$

for a solution $u = u(x,t)$ of (1.1), if $\rho(u) = \sum_{j\in\mathbb{Z}^d} \overline{u_j} \mathrm{e}^{\mathrm{i}(j\cdot x)}$ for $u = \sum_{j\in\mathbb{Z}^d} u_j \mathrm{e}^{\mathrm{i}(j\cdot x)}$. The Fourier coefficients of a function $u = u(x)$ are denoted throughout the paper by u_j, $j \in \mathbb{Z}^d$.

For the numerical solution of (1.1) we first discretize in space (method of lines) and then in time. In practice, in the periodic case, the use of a discrete Fourier transform is a favorable choice. We then discretize the resulting system of ordinary differential equations with an exponential integrator (Sect. 2). Exponential integrators are widely used and studied nowadays as witnessed by the recent review [22]. Here we use the exponential integrators for nonlinear Schrödinger equations introduced in [7]. An error analysis for exponential integrators of collocation type applied to Schrödinger equations was given in [10].

In this paper, we study the long-time behavior of the conserved quantities energy (1.2), mass (1.3) and momentum (1.4) along such a numerical solution of (1.1) by an exponential integrator. In recent years, there is a growing interest and an ongoing effort in explaining the long-time behavior of numerical schemes for Hamiltonian partial differential equations, see [3, 6, 8, 9, 11, 13–17, 19]. In the present article, we show for a class of one-stage exponential integrators that energy, mass and momentum of (1.1) are approximately conserved along the numerical solutions over long times, see Sect. 3 for a precise statement of the results and numerical experiments.

A property closely related to the *reversibility of the numerical scheme* turns out to be crucial to prove this result. We present numerical experiments that suggest that this property of the one-stage exponential integrator is not only sufficient but also necessary to have a good long-time behavior. This property requires a one-stage exponential integrator to be implicit.

Similar results have been shown in [16, 19] for splitting integrators, which are widely applied in the numerical integration of Schrödinger equations. This paper is neither aimed at comparing the (implicit) exponential integrators studied here with these splitting integrators nor at promoting their use at the expense of splitting integrators. The present paper contributes to the numerical analysis of exponential integrators, and, in a broader sense, tries to identify mechanisms that lead to a good long-time behavior of numerical integrators for *partial* differential equations (reversibility, for instance). Nonetheless we mention that, although implicit exponential integrators are slightly less efficient in comparison with splitting integrators, cf. [7, Introduction and Sect. 5.1], they are indeed used for the numerical integration of Schrödinger equations, see [4, 7, 10].

Similarly as in the aforementioned paper [19] on splitting integrators we start in Sect. 4 by showing, using a modulated Fourier expansion of the numerical solution, that the *actions*

$$I_l(u,\bar{u}) := \frac{1}{2}|u_l|^2 \quad \bigl(l \in \mathbb{Z}^d\bigr) \tag{1.5}$$

of the linear Schrödinger equation $\mathrm{i}\frac{\partial}{\partial t}u = -\Delta u + V * u$ are approximately conserved along the numerical solution of (1.1) over long times. Note that the actions (1.5) are also nearly conserved along the exact solution of the nonlinear equation (1.1) over long times, see [1, 18]. The long-time near-conservation of actions implies the regularity of the numerical solution over long times and is the key for the proof of the above mentioned conservation properties in Sect. 4.

2 Discretization of the nonlinear Schrödinger equation

In this section, we discretize (1.1) in space with a spectral collocation scheme and in time with a one-stage exponential integrator.

2.1 Spectral collocation method for the discretization in space

We start by denoting, for $j \in \mathbb{Z}^d$, the frequencies in the nonlinear Schrödinger (1.1) by

$$\omega_j := |j|^2 + V_j = j_1^2 + \cdots + j_d^2 + V_j \tag{2.1}$$

with the j-th Fourier coefficient $V_j \in \mathbb{R}$ of the potential V.

A spectral collocation discretization in space with collocation points

$$x_k := k\frac{\pi}{M} \quad \text{for } k \in \mathscr{M} = \{-M, \ldots, M-1\}^d$$

yields an approximation

$$u^M(x,t) := \sum_{k\in\mathscr{M}} q_k(t)\mathrm{e}^{\mathrm{i}(k\cdot x)}.$$

Requiring this ansatz to fulfill (1.1) at our collocation points, one obtains, using $u^M(x_k,t)_{k\in\mathscr{M}} = F_{2M}(q_k(t))_{k\in\mathscr{M}}$ with the d-dimensional discrete Fourier transform F_{2M}, the following system of ordinary differential equations

$$\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}t}u^M(x_k,t)_{k\in\mathscr{M}} = F_{2M}\Omega F_{2M}^{-1}u^M(x_k,t)_{k\in\mathscr{M}} + \left(\left|u^M(x_k,t)\right|^2 u^M(x_k,t)\right)_{k\in\mathscr{M}},$$

where $\Omega = \operatorname{diag}((\omega_k)_{k\in\mathscr{M}})$ is a diagonal matrix with frequencies ω_k, $k\in\mathscr{M}$. Or in terms of the approximation $u^M(x,t)$ one gets

$$\frac{\partial}{\partial t}u^M = Lu^M + f\left(u^M\right) \tag{2.2}$$

with

$$L = \mathrm{i}\Delta - \mathrm{i}V\, * \quad \text{and} \quad f(u) = -\mathrm{i}\mathscr{Q}\left(|u|^2u\right). \tag{2.3}$$

Here $\mathscr{Q}$ denotes the trigonometric interpolation

$$\mathscr{Q}\left(\sum_{j\in\mathbb{Z}^d} u_j\mathrm{e}^{\mathrm{i}(j\cdot x)}\right) := \sum_{j\in\mathscr{M}}\left(\sum_{\ell\in\mathbb{Z}^d:\ell\equiv j \bmod 2M} u_\ell\right)\mathrm{e}^{\mathrm{i}(j\cdot x)},$$

and this is defined in such a way that $\mathscr{Q}(u)(x_k) = u(x_k)$ for all $k\in\mathscr{M}$. The initial value is then given by

$$u^M(\cdot,0) = \mathscr{Q}\left(u(\cdot,0)\right).$$

We note that the above semi-discretized system is a finite dimensional complex Hamiltonian system with Hamiltonian

$$H_M\left(u^M,\overline{u^M}\right) = \frac{1}{2(2\pi)^d}\int_{\mathbb{T}^d}\left(\left|\nabla u^M\right|^2 + \left(V*u^M\right)\overline{u^M} + \frac{1}{2}\mathscr{Q}\left(\left|u^M\right|^4\right)\right)\mathrm{d}x.$$

We now discretize (2.2) with the above initial value in time.

2.2 Exponential integrators for the discretization in time

Exponential integrators, as their name suggests, use the exponential function of the Jacobian (or an approximation to it) inside the numerical scheme. They are particularly efficient for problems of the form, see (2.2)–(2.3),

$$\frac{d}{dt}u = Lu + f(u),$$

where $u = u(t)$, L is typically a linear unbounded differential operator, alternatively one can think of L as a matrix arising from a space discretization of such an operator and thus bounded for a fixed spatial resolution, but with a large norm. The map f is nonlinear but we assume that the size of $f(u)$ is small compared to L. For a survey of these methods see for instance [21, 24] and more recently the review [22] and references therein.

The schemes we consider here can all be cast in the form

$$\begin{aligned} F_r &= f\left(\mathrm{e}^{c_r hL}u^0 + h\sum_{j=1}^{s} a_{rj}(hL)F_j\right), \quad r = 1,\dots,s \\ u^1 &= \mathrm{e}^{hL}u^0 + h\sum_{r=1}^{s} b_r(hL)F_r. \end{aligned} \tag{2.4}$$

We use upper indices for denoting time steps with step-size h. The involved functions $a_{rj}(z)$ and $b_r(z)$ are complex functions that are used to define and to compute $a_{rj}(hL)$ and $b_r(hL)$ in terms of the spectral decomposition of the matrix describing L, i.e., $a_{rj}(hL) = F_{2M}a_{rj}(-ih\Omega)F_{2M}^{-1}$ and accordingly for b_r. They are often real entire or at least real analytic in a domain of the complex plane which includes the spectrum of hL for all h of interest. Such schemes have been applied to nonlinear Schrödinger equations in [7] and [5] for example. In applying them to this equation, it is of importance to choose functions $a_{rj}(z)$ and $b_r(z)$ which are bounded on the imaginary axis, a property which is rather common among popular exponential integrators.

In our numerical analysis, we will only consider one-stage exponential integrators ($s = 1$) for (2.2)–(2.3)

$$\begin{aligned} U &= \mathrm{e}^{chL}u^0 + ha(hL)f(U), \\ u^1 &= \mathrm{e}^{hL}u^0 + hb(hL)f(U). \end{aligned} \tag{2.5}$$

We will focus on two important geometric properties: *symmetry* and *reversibility* which we recall now.

Definition 2.1 (Symmetry, [20, Chap. V]) A numerical one-step method $y^1 = \Phi_h(y^0)$ is called *symmetric* if it satisfies

$$\Phi_h \circ \Phi_{-h} = id \quad \text{or equivalently} \quad \Phi_h = \Phi_{-h}^{-1}.$$

Definition 2.2 (Reversibility, [20, Chap. V]) Let ρ be an invertible linear transformation in the phase space of $\frac{d}{dt}y = g(y)$. This differential equation is called *ρ-reversible* if $\rho \circ g = -g \circ \rho$, implying $\rho \circ \varphi_t = \varphi_t^{-1} \circ \rho$ for the exact flow φ_t. A numerical one-step method $y^1 = \Phi_h(y^0)$ is called *ρ-reversible* if

$$\rho \circ \Phi_h = \Phi_h^{-1} \circ \rho.$$

Our nonlinear Schrödinger equation (1.1) and also its semi-discretization in space (2.2) are ρ-reversible for the complex conjugation of Fourier coefficients, $\rho(u) = \sum_j \overline{u_j}\mathrm{e}^{\mathrm{i}jx}$ for $u = \sum_j u_j\mathrm{e}^{\mathrm{i}jx}$ (note that (1.1) is in general not reversible for the complex conjugation of a function itself because of the convolution with V). In the following we will always study reversibility with respect to this complex conjugation.

For the one-stage numerical schemes (2.5) considered here, we obtain the following results.

Lemma 2.1 (Symmetry of exponential integrators, [7]) *A consistent one-stage exponential integrator* (2.5) *is symmetric if and only if*

$$c = a(0) = \frac{1}{2} \quad \textit{and} \quad b(z) = \mathrm{e}^{z/2}\big(a(z) + a(-z)\big) \quad \textit{for all } z \in \mathrm{i}\mathbb{R}. \tag{2.6}$$

This shows in particular that symmetric exponential integrators as considered here are implicit.

Lemma 2.2 (Reversibility of exponential integrators) *A consistent one-stage exponential integrator* (2.5) *is reversible if and only if*

$$c = \operatorname{Re}\big(a(0)\big) = \frac{1}{2} \quad \textit{and} \quad b(z) = 2\mathrm{e}^{z/2}\operatorname{Re}\big(a(z)\big) \quad \textit{for all } z \in \mathrm{i}\mathbb{R}. \tag{2.7}$$

Proof Let us first compute $v = \rho \circ \Phi_h(u^0)$. From the definition of the exponential integrator, we have

$$\begin{aligned} v &= \mathrm{e}^{-hL}\rho\big(u^0\big) - hb(hL)^* f\big(\rho(U)\big),\\ U &= \mathrm{e}^{chL}u^0 + ha(hL)f(U) \end{aligned}$$

with the adjoint matrix $b(hL)^*$. For the second term $w = \Phi_h^{-1} \circ \rho(u^0)$ in the definition of reversibility, we get $\rho(u^0) = \mathrm{e}^{hL}w + hb(hL)f(U)$ with $U = \mathrm{e}^{chL}w + ha(hL)f(U)$. We thus obtain

$$\begin{aligned} w &= \mathrm{e}^{-hL}\big(\rho\big(u^0\big) - hb(hL)f(U)\big),\\ U &= \mathrm{e}^{(c-1)hL}\big(\rho\big(u^0\big) - hb(hL)f(U)\big) + ha(hL)f(U). \end{aligned}$$

Comparing the two equations for v and w, the result follows. □

We also note, that the conditions of symmetry (2.6) and reversibility (2.7) are equivalent if $a(\overline{z}) = \overline{a(z)}$ for all $z \in \mathrm{i}\mathbb{R}$. Let us now illustrate these properties with some examples.

Example 2.1 (Symmetric Lawson method, [7, 23]) The symmetric one-stage *Lawson method* is an exponential integrator (2.5) with coefficients

$$a(z) = \frac{1}{2}, \qquad b(z) = \mathrm{e}^{z/2}, \qquad c = \frac{1}{2}.$$

It is symmetric and reversible.

Example 2.2 The method with coefficients

$$a(z) = \frac{1}{2}b(z/2), \qquad b(z) = \frac{\mathrm{e}^z - 1}{z}, \qquad c = \frac{1}{2}$$

is also symmetric and reversible.

Example 2.3 The exponential method (2.5) with coefficients

$$a(z) = \frac{1}{2}, \qquad b(z) = \frac{\mathrm{e}^z - 1}{z}, \qquad c = \frac{1}{2}$$

is neither symmetric nor reversible.

However, if $a(\overline{z}) \neq \overline{a(z)}$ for some $z \in \mathrm{i}\mathbb{R}$, then reversible methods are not symmetric and symmetric methods are not reversible.

Example 2.4 The method with coefficients

$$a(z) = \frac{1}{2} + i\,\frac{\mathrm{e}^{z/2} - 1 - \frac{z}{2}}{z}, \qquad b(z) = 2\mathrm{e}^{z/2}\,\mathrm{Re}\big(a(z)\big), \qquad c = \frac{1}{2}$$

is reversible but not symmetric.

Example 2.5 The method with coefficients

$$a(z) = \frac{1}{2} + i\,\frac{\mathrm{e}^{z/2} - 1 - \frac{z}{2}}{z}, \qquad b(z) = \mathrm{e}^{z/2}\big(a(z) + a(-z)\big), \qquad c = \frac{1}{2}$$

is symmetric but not reversible.

3 Main result and numerical experiments

We now formulate our main result on the long-time behavior of exponential integrators.

3.1 Assumptions on the exponential integrator (2.5)

We start this section by collecting the assumptions on the exponential integrator (2.5) we need to prove long-time near-conservation properties of the numerical solution. Our main assumption on the exponential integrator is that its coefficient functions $a(z)$ and $b(z)$ are linked in the following way:

$$b(z) = 2\mathrm{e}^{(1-c)z}\,\mathrm{Re}\big(a(z)\big) \quad \text{for all } z \in \mathrm{i}\mathbb{R}. \tag{3.1a}$$

If in addition $c = \mathrm{Re}(a(0)) = \frac{1}{2}$, this is equivalent to the condition of reversibility (2.7). In particular, all reversible methods (2.7) satisfy this condition. Also many symmetric methods (2.6) satisfy (3.1a), namely those that are reversible, i.e., those with $a(\overline{z}) = \overline{a(z)}$ for all $z \in \mathrm{i}\mathbb{R}$. But there are methods that satisfy (3.1a) which are neither symmetric nor reversible.

Example 3.1 The exponential method (2.5) with coefficients

$$a(z)=\frac{1}{2},\qquad b(z)=\mathrm{e}^{z/3},\qquad c=\frac{2}{3}$$

satisfies (3.1a) but is neither symmetric nor reversible.

Besides condition (3.1a) we need, as mentioned in Sect. 2, that the function a is bounded on the imaginary axis,

$$\bigl|a(z)\bigr|\le C_1\quad\text{for } z\in \mathrm{i}\mathbb{R}. \tag{3.1b}$$

Moreover, we assume that

$$b(hL)\quad\text{is invertible.} \tag{3.1c}$$

We do not hesitate to impose this assumption because it is typically less restrictive than the non-resonance condition on the frequencies ω_j that we will introduce in the following section. For the methods of Examples 2.1 and 3.1 the condition (3.1c) is trivially satisfied, and for the methods of Example 2.4 and 2.5 it is not difficult to verify this condition for positive frequencies ω_j, the eigenvalues of L. For the methods of Examples 2.2 and 2.3 condition (3.1c) amounts to a restriction on the time step-size h: One has to avoid time step-sizes that are integer multiples of $2\pi/\omega_j$ for some frequency ω_j. For comparison the non-resonance condition, that we will impose, requires that the time step-size h is not close to an integer multiple of 2π divided by many linear combinations of frequencies. We mention, however, that the method of Example 2.2 behaves in numerical experiments very well also for resonant time step-sizes that are not small and for time step-sizes that do not satisfy (3.1c). This may be due to the fact that the functions a and b decay for large frequencies and therefore act as filter functions.

The invertibility condition (3.1c) can be used to rewrite the exponential integrator (2.5). We solve the first equation of (2.5) for $f(U)$ and then plug it in the second one to obtain

$$U=\mathrm{e}^{chL}u^0+a(hL)b(hL)^{-1}\bigl(u^1-\mathrm{e}^{hL}u^0\bigr).$$

This yields

$$u^1=\mathrm{e}^{hL}u^0+hb(hL)f\bigl(\mathrm{e}^{chL}u^0+a(hL)b(hL)^{-1}\bigl(u^1-\mathrm{e}^{hL}u^0\bigr)\bigr). \tag{3.2}$$

In the following we will work with one-stage exponential integrators in this compact form.

3.2 Long-time near-conservation of actions, energy, mass and momentum

Let $N\ge 1$ be an arbitrary fixed integer. Our main result states near-conservation properties over long times $0\le t\le \varepsilon^{-N}$ of the numerical solutions of the cubic Schrödinger equation (1.1) with small initial data:

$$\bigl\|u^0\bigr\|_s\le\varepsilon\ll 1 \tag{3.3}$$

with the Sobolev norm

$$\|u\|_s^2 := \sum_{j\in\mathbb{Z}^d} |\omega_j|^s |u_j|^2,$$

where we recall that u_j denotes the j-th Fourier coefficient of a function u on $\mathbb{T}^d$ and that ω_j denotes the j-th frequency (2.1). In this definition, a zero frequency is replaced by 1. This is also tacitly assumed in the following whenever the absolute value of a frequency appears.

Moreover, we need a non-resonance condition on the frequencies that we introduce now. Recall that $\mathscr{M} = \{-M, \dots, M-1\}^d$ denotes the set of indices whose corresponding Fourier coefficients are used for the discretization in space, see Sect. 2.1. We denote by $\mathbf{k} = (k_l)_{l\in\mathscr{M}}$ a finite sequence of integers, by $\boldsymbol{\omega} = (\omega_l)_{l\in\mathscr{M}}$ the finite sequence of frequencies and by mod $2M$ the entry-wise reduction modulo $2M$ with representative chosen in $\mathscr{M}$. The non-resonance condition controls near-resonances among the frequencies, where the difference of a linear combination of frequencies

$$\mathbf{k}\cdot\boldsymbol{\omega} := \sum_{l\in\mathscr{M}} k_l\omega_l \quad \text{with small } \|\mathbf{k}\| := \sum_{l\in\mathscr{M}} |k_l|$$

and the frequency

$$\omega_{j(\mathbf{k})} \quad \text{with } j(\mathbf{k}) := \sum_{l\in\mathscr{M}} k_l l \bmod 2M \in \mathscr{M}, \tag{3.4}$$

is close to an integer multiple of $2\pi/h$. We recall that h is the step-size of our numerical integrator. More precisely, we require for near-resonant indices $(j,\mathbf{k})$ in the set

$$\mathscr{R}_{\varepsilon,M,h} = \left\{(j,\mathbf{k}) : j = j(\mathbf{k}), \mathbf{k} \neq \langle j\rangle, \left|\mathrm{e}^{\mathrm{i}(\omega_j - \mathbf{k}\cdot\boldsymbol{\omega})h} - 1\right| < \varepsilon^{\frac{1}{2}} h, \|\mathbf{k}\| \leq 2N+2\right\},$$

where $\langle j\rangle = (\delta_{jl})_{l\in\mathbb{Z}^d}$ with Kronecker's delta, that

$$\sup_{(j,\mathbf{k})\in\mathscr{R}_{\varepsilon,M,h}} \frac{|\omega_j|^{s-\frac{d+1}{2}}}{|\boldsymbol{\omega}^{(s-\frac{d+1}{2})|\mathbf{k}|}|}\varepsilon^{\|\mathbf{k}\|+1} \leq C_0\varepsilon^{2N+4} \tag{3.5}$$

with a constant C_0 independent of ε. Here,

$$\boldsymbol{\omega}^{\sigma|\mathbf{k}|} = \prod_{l\in\mathscr{M}} |\omega_l|^{\sigma|k_l|} \quad \text{for } \sigma\in\mathbb{R}.$$

This non-resonance condition is very similar to the one used for splitting integrators [19, Sect. 4]. As discussed in [19, Appendix] it reduces in the limit $h \to 0$ to a condition that is satisfied for almost all choices of the potential V and a time step-size restriction allows to exclude numerical resonances. Moreover, it is fulfilled for all step-sizes in a dense set under a restriction on the parameter M of the discretization in space in terms of ε, see [19, Appendix].

We are now able to state the main result of this paper, whose proof will be given in Sect. 4.

Theorem 3.1 *For given $N \geq 1$ and $s \geq d+1$ there exists $\varepsilon_0 > 0$ such that the following holds: Under the conditions (3.1a)–(3.1c) on the exponential integrator, the condition of small initial data (3.3) with $\varepsilon \leq \varepsilon_0$ and the non-resonance condition (3.5), the estimates*

$$\sum_{l\in\mathscr{M}} |\omega_l|^s \frac{|I_l(u^n,\overline{u^n}) - I_l(u^0,\overline{u^0})|}{\varepsilon^2} \leq C\varepsilon^{\frac{3}{2}},$$

$$\frac{|H(u^n,\overline{u^n}) - H(u^0,\overline{u^0})|}{\varepsilon^2} \leq C\varepsilon^{\frac{3}{2}}, \qquad \frac{|H_M(u^n,\overline{u^n}) - H_M(u^0,\overline{u^0})|}{\varepsilon^2} \leq C\varepsilon^{\frac{3}{2}},$$

$$\frac{|m(u^n,\overline{u^n}) - m(u^0,\overline{u^0})|}{\varepsilon^2} \leq C\varepsilon^{\frac{3}{2}}, \qquad \sum_{r=1}^{d} \frac{|K_r(u^n,\overline{u^n}) - K_r(u^0,\overline{u^0})|}{\varepsilon^2} \leq C\varepsilon^{\frac{3}{2}}$$

hold for the numerical solution u^n described in Sect. 2 with time step-size $h \leq 1$ over long times

$$0 \leq t_n = nh \leq \varepsilon^{-N}$$

with a constant C which depends on C_0 from the non-resonance condition (3.5), C_1 from assumption (3.1b), the dimension d, N, s and the norm of the potential V but is independent of n, the size of the initial value ε and the discretization parameters M and h.

3.3 Numerical experiments

We conclude this section with some numerical experiments in order to illustrate Theorem 3.1. We use data as in the experiments of [19]. The initial value $u(\cdot,0)$ is chosen as ($d=1$)

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

and the potential V is chosen such that

$$\omega_j = \sqrt{|j|^4 + r_j},$$

where $r_j = 0.5$ for $j \geq 0$ and $r_j = 0.8$ for $j < 0$. We use $2M = 2^8$ collocation points for the discretization in space, and we use a time step $h = 0.1$ with different exponential integrators. For the solution of the nonlinear equations defining U in (2.5) we apply the standard fixed point iteration to the nonlinear equation as described in [7, Sect. 5.1]. To be on the safe side, we use 15 iterations although the convergence is much faster, in particular due to the small nonlinearity (cf. the analysis of the nonlinear equation in Sect. 4.7). For the role of rounding errors in a long-time integration and a possible way to reduce them we refer to [20, VIII.5].

In Fig. 1 we plot some of the actions, the discrete energy H_M, the mass m and the momentum K for the symmetric and reversible Lawson method from Example 2.1.

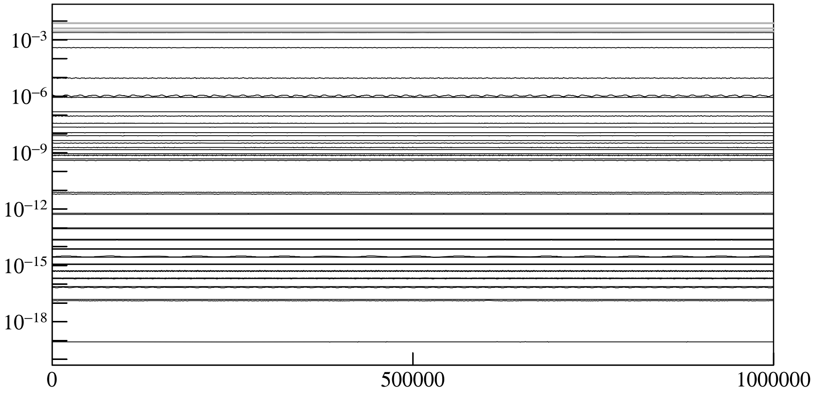

Fig. 1 Actions (*black lines*), discrete energy (*middle bold grey line*), mass (*upper bold grey line*) and momentum (*lower bold grey line*) for the method from Example 2.1. The methods from Example 2.2, Example 2.4 and Example 3.1 show the same behavior

This method satisfies the main assumption (3.1a) on the exponential integrator. As explained by Theorem 3.1, the plotted quantities are nearly conserved on a long time interval of length 10^6. We observe the same behavior for the other methods that satisfy the main assumption (3.1a), the symmetric and reversible method from Example 2.2, the reversible but non-symmetric method from Example 2.4 and the non-reversible and non-symmetric method from Example 3.1.

We repeat the experiment in Fig. 2 using *methods that do not satisfy the main assumption* (3.1a): The non-symmetric and non-reversible method from Example 2.3 and the symmetric but non-reversible method from Example 2.5. For these methods, the actions are no longer nearly conserved on long time intervals.

4 Modulated Fourier expansions and proof of the main result

In this section we prove Theorem 3.1 on the long-time near-conservation of actions (1.5), energy (1.2), mass (1.3) and momentum (1.4) along the numerical solution. Throughout this section we work under the assumptions of this theorem.

The proof relies on a careful study of a *modulated Fourier expansion* in time of the numerical solution (3.2),

$$\tilde{u}(x,t) = \sum_{\|\mathbf{k}\|\le K} z^{\mathbf{k}}(x,\varepsilon t)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t} = \sum_{\|\mathbf{k}\|\le K}\sum_{j\in\mathcal{M}} z_j^{\mathbf{k}}(\varepsilon t)\mathrm{e}^{\mathrm{i}(j\cdot x)}\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}. \tag{4.1a}$$

We require this modulated Fourier expansion to describe at time $t_n = nh$ the numerical solution u^n after n time steps. The *modulation functions* $z^{\mathbf{k}}$ evolve on a slow time-scale $\tau = \varepsilon t$. It turns out that we can assume these functions to be single spatial waves,

$$z^{\mathbf{k}}(x,\varepsilon t) = z^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t)\mathrm{e}^{\mathrm{i}(j(\mathbf{k})\cdot x)}, \tag{4.1b}$$

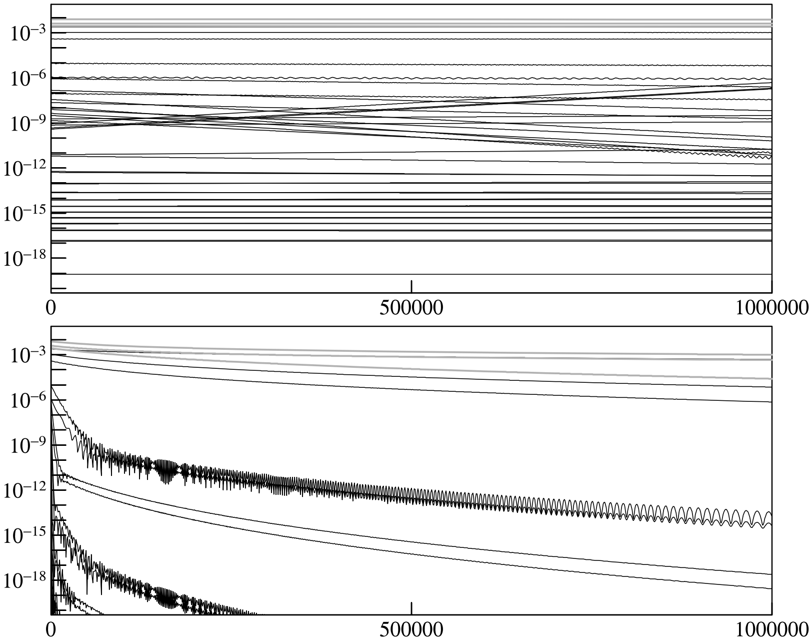

Fig. 2 Actions (*black lines*), discrete energy (*middle bold grey line*), mass (*upper bold grey line*) and momentum (*lower bold grey line*) for the methods from Example 2.3 (*first subfigure*) and Example 2.5 (*second subfigure*)

i.e., their Fourier coefficients $z_j^{\mathbf{k}}$ vanish for $j \neq j(\mathbf{k})$ with $j(\mathbf{k})$ as introduced as in (3.4). The outline of the proof is as follows.

– In Sect. 4.1 we derive a system of equations for the modulation functions.
– In Sect. 4.2 we show the existence of invariants for this system of equations.
– Then we construct an approximate solution of this system in Sect. 4.3.
– We study the size of the constructed modulation functions in Sect. 4.5 using a rescaling of these functions introduced in Sect. 4.4,
– and we study the defect for the approximate solution in the modulation system in Sect. 4.6.
– Then we control the size of the numerical solution and the difference of the numerical solution and its modulated Fourier expansion in Sects. 4.7 and 4.8.
– We study the invariants of the modulation system along the approximate solution of this system and establish their relationship with the actions in Sect. 4.9.
– Finally, we extend the previous results, that are valid only on a short time interval of length ε^{-1}, to a long time interval in Sects. 4.10 and 4.11.

This is the standard approach to study the long-time behavior of numerical solutions of Hamiltonian partial differential equations using modulated Fourier expansions [8, 17]. The proof of Theorem 3.1 presented here closely follows the one given in [19] for splitting integrators applied to (1.1). We refer the reader to that article, whenever arguments are very similar or identical, but try to present however the main line of arguments. A major difference compared to the corresponding proof for splitting integrators [19] is that the modulation system for the exponential integrators studied here directly provides invariants. In contrast to that, one has to consider an auxiliary modulation system in the case of splitting integrators. Differences in the analysis of the modulation system further arise due to the implicitness of the considered exponential integrators.

4.1 The modulation system

We insert the modulated Fourier expansion (4.1a), (4.1b) in the numerical scheme (3.2) and require that the numerical solution u^n defined by (3.2) is described by the modulated Fourier expansion $\tilde{u}(x,t_n)$ at time $t_n = nh$. Comparing the coefficients of $\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}$, this yields the following system of equations for the modulation functions $z^{\mathbf{k}}_{j(\mathbf{k})}$:

$$\begin{aligned}\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} z^{\mathbf{k}}_j\big(\varepsilon(t+h)\big) &= \mathrm{e}^{-\mathrm{i}\omega_j h} z^{\mathbf{k}}_j(\varepsilon t) - \mathrm{i}hb(-\mathrm{i}\omega_j h)\\ &\quad\times \sum_{\mathbf{k}^1+\mathbf{k}^2-\mathbf{k}^3=\mathbf{k}} w^{\mathbf{k}^1}_{j(\mathbf{k}^1)}(\varepsilon t) w^{\mathbf{k}^2}_{j(\mathbf{k}^2)}(\varepsilon t) \overline{w^{\mathbf{k}^3}_{j(\mathbf{k}^3)}(\varepsilon t)}\end{aligned} \tag{4.2a}$$

for $j = j(\mathbf{k})$ (note that $j(\mathbf{k}) = j(\mathbf{k}^1) + j(\mathbf{k}^2) - j(\mathbf{k}^3) \bmod 2M$ if $\mathbf{k} = \mathbf{k}^1 + \mathbf{k}^2 - \mathbf{k}^3$). Here and in the following, we assume that $\|k\| \le K := 2N + 2$ unless stated otherwise. The functions $w^{\mathbf{k}}_j$ in the nonlinearity take the form

$$w^{\mathbf{k}}_j(\varepsilon t) = \mathrm{e}^{-\mathrm{i}\omega_j ch} z^{\mathbf{k}}_j(\varepsilon t) + \frac{a(-\mathrm{i}\omega_j h)}{b(-\mathrm{i}\omega_j h)}\big(\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} z^{\mathbf{k}}_j\big(\varepsilon(t+h)\big) - \mathrm{e}^{-\mathrm{i}\omega_j h} z^{\mathbf{k}}_j(\varepsilon t)\big) \tag{4.2b}$$

for $j = j(\mathbf{k})$. The initial condition further yields

$$u^0_j = \sum_{\mathbf{k}} z^{\mathbf{k}}_j(0). \tag{4.2c}$$

The system of equations (4.2a)–(4.2c) for the coefficients of the modulated Fourier expansion is called the *modulation system.*

4.2 Invariants of the modulation system

A remarkable property of the modulation system (4.2a)–(4.2c) is the presence of many conserved quantities or invariants provided that the exponential integrator satisfies condition (3.1a). These invariants, that we derive next, form the cornerstone for the study of long time intervals.

Let

$$\mathscr{U}(\mathbf{w}) = \sum_{\mathbf{k}^1+\mathbf{k}^2-\mathbf{k}^3-\mathbf{k}^4=\mathbf{0}} w^{\mathbf{k}^1}_{j(\mathbf{k}^1)} w^{\mathbf{k}^2}_{j(\mathbf{k}^2)} \overline{w^{\mathbf{k}^3}_{j(\mathbf{k}^3)} w^{\mathbf{k}^4}_{j(\mathbf{k}^4)}}, \tag{4.3}$$

with $\mathbf{w} = (w^{\mathbf{k}})_{\mathbf{k}}$, be the extended potential. We have for real sequences $\boldsymbol{\mu}$

$$
\begin{aligned}
0 &= h\frac{d}{d\theta}\bigg|_{\theta=0} \mathscr{U}\big((\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\mu})\theta} w^{\mathbf{k}})_{\mathbf{k}}\big) \\
&= -4h\,\mathrm{Re}\Bigg(\sum_{\mathbf{k}} i(\mathbf{k}\cdot\boldsymbol{\mu})\overline{w^{\mathbf{k}}_{j(\mathbf{k})}} \sum_{\mathbf{k}^1+\mathbf{k}^2-\mathbf{k}^3=\mathbf{k}} w^{\mathbf{k}^1}_{j(\mathbf{k}^1)} w^{\mathbf{k}^2}_{j(\mathbf{k}^2)} \overline{w^{\mathbf{k}^3}_{j(\mathbf{k}^3)}}\Bigg).
\end{aligned}
$$

Using the modulation system (3.1a) we get for $\mathbf{w} = \mathbf{w}(\varepsilon t)$ as defined in (4.2b)

$$
0 = \mathrm{Re}\Bigg(\sum_{\mathbf{k}} \frac{4(\mathbf{k}\cdot\boldsymbol{\mu})}{b(-\mathrm{i}\omega_{j(\mathbf{k})}h)} \overline{w^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t)}\big(\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} z^{\mathbf{k}}_{j(\mathbf{k})}\big(\varepsilon(t+h)\big) - \mathrm{e}^{-\mathrm{i}\omega_{j(\mathbf{k})}h} z^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t)\big)\Bigg). \tag{4.4}
$$

Under the main condition (3.1a) on the exponential integrator we have

$$
w^{\mathbf{k}}_j(\varepsilon t) = \frac{1}{b(-\mathrm{i}\omega_j h)}\big(a(-\mathrm{i}\omega_j h)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} z^{\mathbf{k}}_j\big(\varepsilon(t+h)\big) + \overline{a(-\mathrm{i}\omega_j h)}\mathrm{e}^{-\mathrm{i}\omega_j h} z^{\mathbf{k}}_j(\varepsilon t)\big)
$$

for $j = j(\mathbf{k})$, and (4.4) simplifies to

$$
0 = \sum_{\mathbf{k}} \frac{(\mathbf{k}\cdot\boldsymbol{\mu})}{\mathrm{Re}(a(-\mathrm{i}\omega_{j(\mathbf{k})}h))}\big(\big|z^{\mathbf{k}}_{j(\mathbf{k})}\big(\varepsilon(t+h)\big)\big|^2 - \big|z^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t)\big|^2\big).
$$

Choosing $\boldsymbol{\mu} = \frac{1}{2}\mathrm{Re}(a(-\mathrm{i}\omega_l h))\langle l\rangle$ for $l \in \mathbb{Z}$, this shows that

$$
\mathscr{I}_l(\mathbf{z}) = \frac{1}{2}\sum_{\mathbf{k}} k_l \frac{\mathrm{Re}(a(-\mathrm{i}\omega_l h))}{\mathrm{Re}(a(-\mathrm{i}\omega_{j(\mathbf{k})}h))}\big|z^{\mathbf{k}}_{j(\mathbf{k})}\big|^2 \tag{4.5}
$$

is conserved along a solution $\mathbf{z}$ of the modulation system (4.2a)–(4.2c) from one time step to another. Recalling the conditions (3.1a) and (3.1b) on the coefficients of the numerical scheme, we will see in Lemma 4.1 that these quantities are well defined. These invariants are the same as for splitting integrators derived in [19, Sect. 6.1] except for the fraction $\mathrm{Re}(a(-\mathrm{i}\omega_l h))/\mathrm{Re}(a(-\mathrm{i}\omega_{j(\mathbf{k})}h))$. Note, however, that our invariants (4.5) are invariants of the modulation system itself and not of an auxiliary modulation system as in [19, Sect. 6].

4.3 Iterative solution of the modulation system

In this subsection we introduce an iterative procedure that we use to compute an approximate solution of the modulation system (4.2a)–(4.2c) in the same way as in [19, Sect. 5.3]. The modulation system here takes the form

$$
\big(1 - \mathrm{e}^{-\mathrm{i}(\omega_j - \mathbf{k}\cdot\boldsymbol{\omega})h}\big) z^{\mathbf{k}}_j(\varepsilon t) + \varepsilon h \dot{z}^{\mathbf{k}}_j(\varepsilon t) = \mathbf{A}\big(\mathbf{z}(\varepsilon t)\big)^{\mathbf{k}}_{j(\mathbf{k})} + \mathbf{N}\big(\mathbf{w}(\varepsilon t)\big)^{\mathbf{k}}_{j(\mathbf{k})}, \tag{4.6a}
$$

$$
\begin{aligned}
w^{\mathbf{k}}_j(\varepsilon t) = \mathrm{e}^{-\mathrm{i}\omega_j ch} z^{\mathbf{k}}_j(\varepsilon t) + \frac{a(-\mathrm{i}\omega_j h)}{b(-\mathrm{i}\omega_j h)}\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} \\
\times \big(\big(1 - \mathrm{e}^{-\mathrm{i}(\omega_j - \mathbf{k}\cdot\boldsymbol{\omega})h}\big) z^{\mathbf{k}}_j(\varepsilon t) - \mathbf{B}\big(\mathbf{z}(\varepsilon t)\big)^{\mathbf{k}}_{j(\mathbf{k})}\big)
\end{aligned} \tag{4.6b}
$$

for $j = j(\mathbf{k})$, where the dot on $z^{\mathbf{k}}_j$ stands for the derivative with respect to the slow time $\tau = \varepsilon t$, and where we use the differential operators

$$\mathbf{A}(\mathbf{z})_{j(\mathbf{k})}^{\mathbf{k}} = -\sum_{l=2}^{\infty} \frac{\varepsilon^l h^l}{l!} \frac{d^l}{d\tau^l} z_{j(\mathbf{k})}^{\mathbf{k}} \quad \text{and} \quad \mathbf{B}(\mathbf{z})_{j(\mathbf{k})}^{\mathbf{k}} = -\sum_{l=1}^{\infty} \frac{\varepsilon^l h^l}{l!} \frac{d^l}{d\tau^l} z_{j(\mathbf{k})}^{\mathbf{k}}$$

and the nonlinearity

$$\mathbf{N}(\mathbf{w})_{j(\mathbf{k})}^{\mathbf{k}} = -\mathrm{i} h b(-\mathrm{i}\omega_{j(\mathbf{k})} h) \mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} \sum_{\mathbf{k}^1+\mathbf{k}^2-\mathbf{k}^3=\mathbf{k}} w_{j(\mathbf{k}^1)}^{\mathbf{k}^1} w_{j(\mathbf{k}^2)}^{\mathbf{k}^2} \overline{w_{j(\mathbf{k}^3)}^{\mathbf{k}^3}}.$$

For the iterative solution of the modulation system we distinguish modulation functions corresponding to near-resonant indices $(j,\mathbf{k}) \in \mathscr{R}_{\varepsilon,M,h}$ or large indices $\|\mathbf{k}\| > K = 2N+2$, "diagonal" modulation functions with indices $(j, \langle j\rangle)$, and the remaining modulation functions with indices in the set

$$\mathscr{S}_{\varepsilon,M,h} = \big\{(j,\mathbf{k}) : j = j(\mathbf{k}), \mathbf{k} \neq \langle j\rangle, (j,\mathbf{k}) \notin \mathscr{R}_{\varepsilon,M,h}, \|\mathbf{k}\| \le K\big\}.$$

We start by setting

$$\big[z_j^{\langle j\rangle}(\tau)\big]^0 = u_j^0 \quad \text{and} \quad \big[z_{j(\mathbf{k})}^{\mathbf{k}}(\tau)\big]^0 = 0 \quad \text{for } \mathbf{k} \neq \big\langle j(\mathbf{k})\big\rangle$$

for $0 \le \varepsilon t = \tau \le 1$. We iterate, motivated by (4.6a), by

$$\big[z_j^{\mathbf{k}}(\tau)\big]^{n+1} = \frac{1}{1-\mathrm{e}^{-\mathrm{i}(\omega_j - \mathbf{k}\cdot\boldsymbol{\omega})h}} \big[\mathbf{B}\big(\mathbf{z}(\tau)\big)_j^{\mathbf{k}} + \mathbf{N}\big(\mathbf{w}(\tau)\big)_j^{\mathbf{k}}\big]^n$$

for $(j,\mathbf{k}) \in \mathscr{S}_{\varepsilon,M,h}$ and $0 \le \varepsilon t = \tau \le 1$ with $\mathbf{w}$ defined as in (4.6b). The notation $[\cdot]^n$ means that the n-th iterates of the modulation functions within the brackets are taken. For $\mathbf{k} = \langle j\rangle$ the first term in (4.6a) cancels, and we define $z_j^{\langle j\rangle}$ as solution of the differential equation

$$\big[\dot{z}_j^{\langle j\rangle}(\tau)\big]^{n+1} = \varepsilon^{-1} h^{-1} \big[\mathbf{A}\big(\mathbf{z}(\tau)\big)_j^{\langle j\rangle} + \mathbf{N}\big(\mathbf{w}(\tau)\big)_j^{\langle j\rangle}\big]^n$$

with initial value

$$\big[z_j^{\langle j\rangle}(0)\big]^{n+1} = u_j^0 - \Bigg[\sum_{\mathbf{k}\neq\langle j\rangle} z_j^{\mathbf{k}}(0)\Bigg]^n$$

by (4.2c). For near-resonant indices $(j,\mathbf{k}) \in \mathscr{R}_{\varepsilon,M,h}$ or for large $\mathbf{k}$ with $\|\mathbf{k}\| > K$ we set for $0 \le \varepsilon t = \tau \le 1$

$$\big[z_j^{\mathbf{k}}(\tau)\big]^{n+1} = 0.$$

With this iterative construction, the iterated modulation functions $[z_j^{\mathbf{k}}]^n$ are polynomials in τ of degree bounded in terms of the number of iterations n.

4.4 Rescaling the modulation functions

In order to take into account the powers of ε that accumulate in the modulation functions, we now rescale and split these functions as in [19, Sect. 5.4]. Let

$$[[\mathbf{k}]] = \begin{cases} \max(\frac{1}{2}(\|\mathbf{k}\|+1), 2), & \mathbf{k} \neq \langle j\rangle, \\ \frac{1}{2}(\|\mathbf{k}\|+1) = 1, & \mathbf{k} = \langle j\rangle \end{cases}$$

and

$$z_j^{\mathbf{k}} = \varepsilon^{[[\mathbf{k}]]} a_j^{\mathbf{k}} + \varepsilon^{[[\mathbf{k}]]} b_j^{\mathbf{k}}$$

with diagonal entries $a_j^{\mathbf{k}}$ and off-diagonal entries $b_j^{\mathbf{k}}$, i.e., $a_j^{\mathbf{k}} \neq 0$ only for $\mathbf{k} = \langle j \rangle$ and $b_j^{\mathbf{k}} \neq 0$ only for $\mathbf{k} \neq \langle j \rangle$. We write

$$\mathbf{a} = \big(a^{\mathbf{k}}\big)_{\mathbf{k}} = \big(a_{j(\mathbf{k})}^{\mathbf{k}} \mathrm{e}^{\mathrm{i}(j(\mathbf{k})\cdot x)}\big)_{\mathbf{k}} \quad \text{and} \quad \mathbf{b} = \big(b^{\mathbf{k}}\big)_{\mathbf{k}} = \big(b_{j(\mathbf{k})}^{\mathbf{k}} \mathrm{e}^{\mathrm{i}(j(\mathbf{k})\cdot x)}\big)_{\mathbf{k}}$$

and set additionally $\mathbf{u} = (u^{\mathbf{k}})_{\mathbf{k}}$ with

$$u^{\mathbf{k}} = \varepsilon^{-[[\mathbf{k}]]} w^{\mathbf{k}} = \varepsilon^{-[[\mathbf{k}]]} w_{j(\mathbf{k})}^{\mathbf{k}} \mathrm{e}^{\mathrm{i} j(\mathbf{k}) x}.$$

We further define $\mathbf{F}(\mathbf{u})_j^{\mathbf{k}} = \varepsilon^{-\max([[\mathbf{k}]],2)} \mathbf{N}(\mathbf{w})$ and

$$(\boldsymbol{\Omega}\mathbf{c})_j^{\mathbf{k}} = \begin{cases} (1 - \mathrm{e}^{-\mathrm{i}(\omega_j - \mathbf{k}\cdot\boldsymbol{\omega})h}) c_j^{\mathbf{k}}, & (j, \mathbf{k}) \in \mathscr{S}_{\varepsilon,M,h}, \\ \varepsilon^{\frac{1}{2}} h c_j^{\mathbf{k}}, & \text{else.} \end{cases}$$

In the rescaled variables the iteration from the previous Sect. 4.3 becomes

$$\begin{aligned} \big[b_j^{\mathbf{k}}\big]^{n+1} &= \big[\big(\boldsymbol{\Omega}^{-1}\mathbf{B}(\mathbf{b})\big)_j^{\mathbf{k}}\big]^n + \big[\big(\boldsymbol{\Omega}^{-1}\mathbf{F}(\mathbf{u})\big)_j^{\mathbf{k}}\big]^n \quad \text{for } (j, \mathbf{k}) \in \mathscr{S}_{\varepsilon,M,h}, \\ \big[\dot{a}_j^{\langle j \rangle}\big]^{n+1} &= \varepsilon^{-1} h^{-1} \big[\mathbf{A}(\mathbf{a})_j^{\langle j \rangle}\big]^n + h^{-1} \big[\mathbf{F}(\mathbf{u})_j^{\langle j \rangle}\big]^n, \\ \big[a_j^{\langle j \rangle}(0)\big]^{n+1} &= \varepsilon^{-1} u_j^0 - \Bigg[\sum_{\mathbf{k} \neq \langle j \rangle} \varepsilon^{[[\mathbf{k}]]-1} b_j^{\mathbf{k}}(0)\Bigg]^n \end{aligned}$$

with $[u_j^{\mathbf{k}}]^n = \varepsilon^{-[[\mathbf{k}]]} [w_j^{\mathbf{k}}]^n$ defined by (4.6b).

We also use a second rescaling of the variables,

$$\hat{a}_j^{\mathbf{k}} = \big|\boldsymbol{\omega}^{\frac{2s-d-1}{4}|\mathbf{k}|}\big| a_j^{\mathbf{k}}, \qquad \hat{b}_j^{\mathbf{k}} = \big|\boldsymbol{\omega}^{\frac{2s-d-1}{4}|\mathbf{k}|}\big| b_j^{\mathbf{k}} \quad \text{and} \quad \hat{u}_j^{\mathbf{k}} = \big|\boldsymbol{\omega}^{\frac{2s-d-1}{4}|\mathbf{k}|}\big| u_j^{\mathbf{k}}.$$

With $\hat{\mathbf{F}}(\hat{\mathbf{u}})_j^{\mathbf{k}} = |\boldsymbol{\omega}^{\frac{2s-d-1}{4}|\mathbf{k}|}| \cdot \mathbf{F}(\mathbf{u})_j^{\mathbf{k}}$ the iteration for $\hat{\mathbf{b}}$ becomes

$$\big[\hat{b}_j^{\mathbf{k}}\big]^{n+1} = \big[\big(\boldsymbol{\Omega}^{-1}\mathbf{B}(\hat{\mathbf{b}})\big)_j^{\mathbf{k}}\big]^n + \big[\big(\boldsymbol{\Omega}^{-1}\hat{\mathbf{F}}(\hat{\mathbf{u}})\big)_j^{\mathbf{k}}\big]^n \quad \text{for } (j, \mathbf{k}) \in \mathscr{S}_{\varepsilon,M,h}.$$

4.5 Size of the iterated modulation functions

In order to control the size of the iterated modulation functions we use the norm

$$|||\mathbf{z}|||_s = \bigg(\sum_j |\omega_j|^s \bigg(\sum_{\mathbf{k}} \big|z_j^{\mathbf{k}}\big|\bigg)^2\bigg)^{\frac{1}{2}}.$$

Note that we do not only need to control the modulation functions themselves but also products of the modulation functions with $a(-\mathrm{i}\omega_j h)/b(-\mathrm{i}\omega_j h)$, see the definition of $w_j^{\mathbf{k}}$ in (4.6b). Fortunately, this is not needed for the diagonal modulation functions collected in $\mathbf{a}$ but only for their derivatives and the off-diagonal modulation functions collected in $\mathbf{b}$ including all their derivatives. Since $a(-\mathrm{i}\omega_j h)$ is bounded by (3.1b), we therefore set

$$\gamma_j = \max\left(1, \frac{1}{|b(-\mathrm{i}\omega_j h)|}\right) \quad \text{and} \quad (\boldsymbol{\Gamma}\mathbf{c})_j^{\mathbf{k}} = \gamma_j c_j^{\mathbf{k}},$$

and we study $\boldsymbol{\Gamma}\mathbf{b}$ and $\boldsymbol{\Gamma}\dot{\mathbf{a}}$ instead of $\mathbf{b}$ and $\dot{\mathbf{a}}$.

Lemma 4.1 *We have for* $0 \le \tau = \varepsilon t \le 1$

$$\begin{aligned} &\left|\!\left|\!\left| \left[\mathbf{a}(\tau)\right]^n \right|\!\right|\!\right|_s \le C, \qquad \left|\!\left|\!\left| \left[\boldsymbol{\Gamma}\mathbf{a}^{(\ell)}(\tau)\right]^n \right|\!\right|\!\right|_s \le C\varepsilon \quad \text{for } \ell \ge 1, \\ &\left|\!\left|\!\left| \left[\boldsymbol{\Gamma}\mathbf{b}^{(\ell)}(\tau)\right]^n \right|\!\right|\!\right|_s \le C\varepsilon^{\frac{1}{2}} \quad \text{for } \ell \ge 0, \end{aligned}$$

for all n with a constant C depending only on C_0, C_1, d, n, s and the norm of V. The same estimates hold for $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ instead of $\mathbf{a}$ and $\mathbf{b}$ if we replace $|||\cdot|||_s$ by $|||\cdot|||_{\frac{d+1}{2}}$.

In particular, it follows that the modulated Fourier expansion of the numerical scheme $\tilde{u}$ is small

$$\left\|\tilde{u}(\cdot, t)\right\|_s \le C\varepsilon$$

and that its coefficients z are also small

$$\sum_{j\in\mathscr{M}} |\omega_j|^s \left|z_j^{\langle j\rangle}\right|^2 \le C\varepsilon^2 \quad \text{and} \quad \sum_{j\in\mathscr{M}} |\omega_j|^s \left(\sum_{\mathbf{k}\neq\langle j\rangle} \left|z_j^{\mathbf{k}}\right|\right)^2 \le C\varepsilon^5.$$

Proof Initially, we have for $0 \le \tau \le 1$

$$\begin{aligned} &\left|\!\left|\!\left| \left[\mathbf{a}(\tau)\right]^0 \right|\!\right|\!\right|_s \le 1, \qquad \left|\!\left|\!\left| \left[\boldsymbol{\Gamma}\mathbf{a}^{(\ell)}(\tau)\right]^0 \right|\!\right|\!\right|_s = 0 \quad \text{for } \ell \ge 1, \\ &\left|\!\left|\!\left| \left[\boldsymbol{\Gamma}\mathbf{b}^{(\ell)}(\tau)\right]^0 \right|\!\right|\!\right|_s = 0 \quad \text{for } \ell \ge 0, \end{aligned}$$

and the same estimates hold for $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ if we replace $|||\cdot|||_s$ by $|||\cdot|||_{\frac{d+1}{2}}$.

The bounds for the iterated modulation functions are obtained as in [19, Sect. 5.6] by analyzing the iteration using

- the non-resonance condition (3.5) to control $\boldsymbol{\Omega}^{-1}$,
- the fact that $\mathbf{A}(\mathbf{a})$ contains only derivatives of $\mathbf{a}$ to estimate $\boldsymbol{\Gamma}\mathbf{A}(\mathbf{a})$ in the same way as $\boldsymbol{\Gamma}\mathbf{B}(\mathbf{b})$ inductively,
- the factor $b(-\mathrm{i}\omega_j h)$ in front of the nonlinearity to estimate $\boldsymbol{\Gamma}\mathbf{F}(\mathbf{u})$ in $|||\cdot|||_s$ by $C\varepsilon h|||\mathbf{u}|||_s^3$ using [18, Lemma 2] and the bound (3.1b) together with the condition (3.1a),
- the bound $|||\mathbf{u}|||_s \le |||\mathbf{a}+\mathbf{b}|||_s + C\max_{\ell\ge1}|||\boldsymbol{\Gamma}\mathbf{a}^{(\ell)}|||_s + C\max_{\ell\ge0}|||\boldsymbol{\Gamma}\mathbf{b}^{(\ell)}|||_s$
- and the fact that the modulation functions are polynomials in τ of degree bounded in terms of the number of iterations n.

The same arguments also yield estimates for $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ in the norm $|||\cdot|||_{\frac{d+1}{2}}$. □

4.6 Defect of the iterated modulation functions

The defect in the modulation system (4.6a), (4.2c) after n iterations is

$$\begin{aligned}\big[d_j^{\mathbf{k}}\big]^n &= \big[\big(1-\mathrm{e}^{-\mathrm{i}(\omega_j-\mathbf{k}\cdot\boldsymbol{\omega})h}\big)z_j^{\mathbf{k}} + \varepsilon h \dot{z}_j^{\mathbf{k}} - \mathbf{A}(\mathbf{z})_j^{\mathbf{k}} - \mathbf{N}(\mathbf{w})_j^{\mathbf{k}}\big]^n,\\ \big[\tilde{d}_j^{\langle j\rangle}(0)\big]^n &= u_j^0 - \Big[\sum_{\mathbf{k}} z_j^{\mathbf{k}}(0)\Big]^n.\end{aligned} \tag{4.7}$$

In contrast to [19, Sect. 5.7], we have here no defect resulting from a truncation of a Taylor expansion. We decompose the defect as

$$\big[d_j^{\mathbf{k}}\big]^n = \big[e_j^{\mathbf{k}} + f_j^{\mathbf{k}} + g_j^{\mathbf{k}} + \dot{h}_j^{\mathbf{k}}\big]^n$$

with $[e_j^{\mathbf{k}}]^n = 0$ for $(j,\mathbf{k}) \notin \mathscr{S}_{\varepsilon,M,h}$, $[\dot{h}_j^{\mathbf{k}}]^n = 0$ for $\mathbf{k} \neq \langle j\rangle$, $[f_j^{\mathbf{k}}]^n = 0$ for non-near-resonant indices $(j,\mathbf{k}) \notin \mathscr{R}_{\varepsilon,M,h}$ and $[g_j^{\mathbf{k}}]^n = 0$ for $\|\mathbf{k}\| \leq K$. The defect can be estimated as follows.

Lemma 4.2 *We have for* $0 \leq \tau \leq 1$

$$\begin{aligned}&\big|\!\big|\!\big|\big[\boldsymbol{\Gamma}\mathbf{f}(\tau)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{N+3}h, \qquad \big|\!\big|\!\big|\big[\boldsymbol{\Gamma}\mathbf{g}(\tau)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{N+3}h,\\ &\big|\!\big|\!\big|\big[\boldsymbol{\Gamma}\mathbf{e}(\tau)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{\frac{n+4}{2}}h, \qquad \big|\!\big|\!\big|\big[\boldsymbol{\Gamma}\dot{\mathbf{h}}(\tau)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{\frac{n+4}{2}}h, \qquad \big|\!\big|\!\big|\big[\tilde{\mathbf{d}}(0)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{\frac{n+2}{2}}\end{aligned}$$

for all n with a constant C depending only on C_0, C_1, d, N, n, s and the norm of the potential V. The same estimates hold for $\hat{\mathbf{e}}$ and $\hat{\mathbf{h}}$ instead of $\mathbf{e}$ and $\mathbf{h}$ if we replace $|\!|\!|\cdot|\!|\!|_s$ by $|\!|\!|\cdot|\!|\!|_{\frac{d+1}{2}}$.

Proof The estimate of the defect $\mathbf{f}$ in the near-resonant indices is obtained as in [18, Sect. 3.7] and [19, Sect. 5.7] using the non-resonance condition (3.5) and in addition the bound (3.1b). Also the defect $\mathbf{g}$ can be estimated as there using that $\|\mathbf{k}\| > K$ implies $[[\mathbf{k}]] \geq \frac{1}{2}(K+2) = N+2$.

The diagonal part $\dot{\mathbf{h}}$ and the off-diagonal part $\mathbf{e}$ of the defect take the form

$$\begin{aligned}\big[e_j^{\mathbf{k}}\big]^n &= \varepsilon^{[[k]]}\big(\big[(\boldsymbol{\Omega}\mathbf{b})_j^{\mathbf{k}}\big]^n - \big[(\boldsymbol{\Omega}\mathbf{b})_j^{\mathbf{k}}\big]^{n+1}\big),\\ \big[h_j^{\mathbf{k}}\big]^n &= \varepsilon^{\frac{3}{2}}\big(\big[(\boldsymbol{\Omega}\mathbf{a})_j^{\mathbf{k}}\big]^n - \big[(\boldsymbol{\Omega}\mathbf{a})_j^{\mathbf{k}}\big]^{n+1}\big).\end{aligned}$$

Using a Lipschitz estimate [18, Lemma 2] for the nonlinearity in the modulation system, we get as in [19, Sect. 5.7] by an analysis of the iteration

$$\begin{aligned}&\big|\!\big|\!\big|\big[\mathbf{h}(\tau)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{\frac{n+4}{2}}h, \qquad \big|\!\big|\!\big|\big[\boldsymbol{\Gamma}\mathbf{h}^{(\ell)}(\tau)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{\frac{n+4}{2}}h \quad \text{for } \ell \geq 1,\\ &\big|\!\big|\!\big|\big[\boldsymbol{\Gamma}\mathbf{e}^{(\ell)}(\tau)\big]^n\big|\!\big|\!\big|_s \leq C\varepsilon^{\frac{n+4}{2}}h \quad \text{for } \ell \geq 0,\end{aligned}$$

for $0 \leq \tau \leq 1$, and the same estimates also for $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ if we replace $|\!|\!|\cdot|\!|\!|_s$ by $|\!|\!|\cdot|\!|\!|_{\frac{d+1}{2}}$.

Also the defect for the initial condition $\tilde{\mathbf{d}}$ can be estimated as in [19, Sect. 5.7]:

$$\big\|\big[\tilde{\mathbf{d}}(0)\big]^n\big\|_s \le \big\|\boldsymbol{\Omega}^{-1}\big[\mathbf{e}(0)\big]^{n-1}\big\|_s \le C\varepsilon^{-\frac{1}{2}}\varepsilon^{\frac{n+3}{2}}.$$

This conclues the proof of the lemma. □

4.7 The numerical solution on short time intervals

We study the size of the numerical solution u^n on a short time interval of length ε^{-1}. Its control uses fixed point arguments since the considered exponential integrators (3.2) are implicit schemes.

Lemma 4.3 *We have for* $0 \le t_n = nh \le \varepsilon^{-1}$

$$\left\|u^n\right\|_s \le 2\varepsilon$$

for ε *sufficiently small compared to* C_1, d, s *and the norm of the potential* V.

Proof We show by induction on n that

$$\left\|u^n\right\|_s \le \varepsilon + 27Cnh\varepsilon^3 \quad \text{for } 0 \le nh \le \varepsilon^{-1}, \tag{4.8}$$

and we let ε be sufficiently small compared to C such that (4.8) implies $\|u^n\|_s \le 2\varepsilon$. Here, C is a constant depending only on C_1, d, s and the norm of V such that

$$\left\|a(hL)\mathscr{Q}(UVW)\right\|_s + \left\|b(hL)\mathscr{Q}(UVW)\right\|_s \le C\|U\|_s\|V\|_s\|W\|_s, \tag{4.9}$$

which exists by [18, Lemmas 1 and 4]. Note that $f(U) = -\mathrm{i}\mathscr{Q}(|U|^2U)$ is the nonlinearity defined in (2.3).

For $n = 0$ the estimate (4.8) is trivial. For $n > 0$ we have by (4.9) and the definition of the integrator

$$\left\|u^n\right\|_s \le \left\|u^{n-1}\right\|_s + h\left\|b(hL)f\big(U^{n-1}\big)\right\|_s \le \left\|u^{n-1}\right\|_s + Ch\left\|U^{n-1}\right\|_s^3 \tag{4.10}$$

with a fixed point U^{n-1} of

$$g : U \mapsto \mathrm{e}^{chL}u^{n-1} + ha(hL)f(U).$$

For $0 \le nh \le \varepsilon^{-1}$ this function g maps by (4.9) the ball $\{U : \|U\|_s \le 3\varepsilon\}$ to itself since $\|u^{n-1}\|_s \le 2\varepsilon$ by induction. Moreover, using the fact that

$$|U|^2U - |\tilde{U}|^2\tilde{U} = |U|^2(U - \tilde{U}) + U\tilde{U}(\overline{U} - \overline{\tilde{U}}) + |\tilde{U}|^2(U - \tilde{U})$$

and the form of our nonlinearity, see (2.3), we obtain from (4.9) that

$$\left\|a(hL)\big(f(U) - f(\tilde{U})\big)\right\|_s \le 3C\max\big(\|U\|_s, \|\tilde{U}\|_s\big)^2\|U - \tilde{U}\|_s. \tag{4.11}$$

This shows that the map g has for sufficiently small ε in the norm $\|\cdot\|_s$ on the ball $\{U : \|U\|_s \le 3\varepsilon\}$ a Lipschitz constant smaller than one. The Banach fixed point theorem then ensures

$$\left\|U^{n-1}\right\|_s \le 3\varepsilon$$

for the fixed point U^{n-1} of g, and finally the induction hypothesis applied to (4.10) yields (4.8). □

4.8 The modulated Fourier expansion and the numerical solution

In this subsection we study the error $u^n - \tilde{u}(\cdot, t_n)$ of the modulated Fourier expansion

$$\tilde{u}(x,t) = \sum_{\mathbf{k}} \bigl[z^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t)\bigr]^L \mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t} \mathrm{e}^{\mathrm{i}(j(\mathbf{k})\cdot x)},$$

where the iterated modulation functions $z^{\mathbf{k}}_j = [z^{\mathbf{k}}_j]^L$ after $L := 2N+2$ iterations replace the exact solution of the modulation system that is not available. By a slight abuse of notation, we omit the index L in the following, keeping in mind that the modulation system is then satisfied only up to a small defect. We show that the modulated Fourier expansion $\tilde{u}(\cdot, t_n)$ describes the numerical solution u^n up to a very small error on a short time interval of length ε^{-1}. Again, as in the previous subsection, we employ fixed point arguments in contrast to the direct arguments used in [19, Sect. 5.8].

Proposition 4.1 *We have for* $0 \le t_n = nh \le \varepsilon^{-1}$

$$\left\|u^n - \tilde{u}(\cdot, t_n)\right\|_s \le C\varepsilon^{N+2}$$

for ε *sufficiently small compared to* C_0, C_1, d, N, s *and the norm of* V *with a constant* C *depending only on* C_0, C_1, d, N, s *and the norm of* V.

Proof Let

$$\tilde{U}(x,t) = \sum_{\mathbf{k}} w^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t) \mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t} \mathrm{e}^{\mathrm{i}(j(\mathbf{k})\cdot x)}.$$

Then, by definition of the modulation system (4.2a)–(4.2c) and with (4.7), we obtain

$$\tilde{u}(\cdot, t_n) = \mathrm{e}^{hL}\tilde{u}(\cdot, t_{n-1}) + hb(hL) f\bigl(\tilde{U}(\cdot, t_{n-1})\bigr) + \delta(\cdot, t_{n-1})$$

with the defect

$$\delta(x,t) = \sum_{\mathbf{k}} d^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t) \mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})(t+h)} \mathrm{e}^{\mathrm{i}(j(\mathbf{k})\cdot x)}.$$

Note that for $0 \le t \le \varepsilon^{-1}$ by Lemma 4.2

$$\left\|\delta(\cdot, t)\right\|_s \le C\varepsilon^{N+3} h$$

with a constant C depending only on C_0, C_1, d, N, s and the norm of the potential V.

(a) We first examine the difference $U^{n-1}-\tilde{U}(\cdot,t_{n-1})$ with the solution U^{n-1} of the nonlinear equation in the numerical method for computing u^n (see the proof of Lemma 4.3). Note that $\tilde{U}(\cdot,t_{n-1})$ is by (4.2a)–(4.2c) and (4.7) a fixed point of

$$\tilde{g}:\tilde{U}\mapsto \mathrm{e}^{chL}\tilde{u}(\cdot,t_{n-1})+ha(hL)f(\tilde{U})+a(hL)b(hL)^{-1}\delta(\cdot,t_{n-1})$$

and by Lemma 4.2

$$\big\|a(hL)b(hL)^{-1}\delta(\cdot,t)\big\|_s\le C\varepsilon^{N+3}h.$$

Recall from the proof of Lemma 4.3 that the fixed point iteration $[U]^l=g([U]^{l-1})$, $[U]^0=\mathrm{e}^{chL}u^{n-1}$ converges in the norm $\|\cdot\|_s$ to U^{n-1} and is bounded in this norm by 3ε. We study $[U]^l-\tilde{U}$ with $\tilde{U}=\tilde{U}(\cdot,t_{n-1})$. Since $\|\tilde{U}\|_s\le C\varepsilon$ by Lemma 4.1, we get with (4.9) and the estimate of the defect

$$\big\|[U]^0-\tilde{U}\big\|_s=\big\|\mathrm{e}^{chL}u^{n-1}-\tilde{g}(\tilde{U})\big\|_s\le\big\|u^{n-1}-\tilde{u}(\cdot,t_{n-1})\big\|_s+C\varepsilon^3h+C\varepsilon^{N+3}h.$$

For $l>0$ we use (4.11) to obtain

$$\begin{aligned}\big\|[U]^l-\tilde{U}\big\|_s&=\big\|g\big([U]^{l-1}\big)-\tilde{g}(\tilde{U})\big\|_s\\&\le\big\|u^{n-1}-\tilde{u}(\cdot,t_{n-1})\big\|_s+C\varepsilon^2h\big\|[U]^{l-1}-\tilde{U}\big\|_s+C\varepsilon^{N+3}h\end{aligned}$$

with a constant C independent of l. A recursion on l and the above result for $l=0$ now yields

$$\big\|[U]^l-\tilde{U}\big\|_s\le\big(\big\|u^{n-1}-\tilde{u}(\cdot,t_{n-1})\big\|_s+C\varepsilon^{N+3}h\big)\sum_{j=0}^{l}\big(C\varepsilon^2h\big)^j+C\varepsilon^3h\big(C\varepsilon^2h\big)^l.$$

For $l\to\infty$ and $C\varepsilon^2h\le\frac{1}{2}$ one then obtains

$$\big\|U^{n-1}-\tilde{U}(\cdot,t_{n-1})\big\|_s\le 2\big\|u^{n-1}-\tilde{u}(\cdot,t_{n-1})\big\|_s+2C\varepsilon^{N+3}h.\tag{4.12}$$

(b) Finally, we consider $u^n-\tilde{u}(\cdot,t_n)$. For $n>0$ we have using (4.11) with $b(hL)$ instead of $a(hL)$

$$\big\|u^n-\tilde{u}(\cdot,t_n)\big\|_s\le\big\|u^{n-1}-\tilde{u}(\cdot,t_{n-1})\big\|_s+C\varepsilon^2h\big\|U^{n-1}-\tilde{U}(\cdot,t_{n-1})\big\|_s+C\varepsilon^{N+3}h.$$

Together with (4.12) we get by induction on n

$$\big\|u^n-\tilde{u}(\cdot,t_n)\big\|_s\le\big(1+2C\varepsilon^2h\big)^n\big(C\varepsilon^{N+3}nh+\big\|u^0-\tilde{u}(\cdot,0)\big\|_s\big).$$

This yields the desired result if ε is sufficiently small since

$$\big\|u^0-\tilde{u}(\cdot,0)\big\|_s\le\big\|\big[\tilde{\mathbf{d}}(0)\big]^n\big\|_s\le C\varepsilon^{N+2}$$

by Lemma 4.2 with the defect $\tilde{\mathbf{d}}$ in the initial condition. □

4.9 Almost invariants close to the actions

Let $z_j^{\mathbf{k}} = [z_j^{\mathbf{k}}]^L$ be the iterated modulation functions after $L = 2N+2$ iterations as in the previous subsection. These modulation functions satisfy the modulation system (4.2a)–(4.2c) only up to a defect $d_j^{\mathbf{k}} = [d_j^{\mathbf{k}}]^L$ studied in Lemma 4.2. Of course, the formal invariants $\mathscr{I}_l(\mathbf{z})$ of the modulation system introduced in (4.5) are then no longer exact invariants, but they turn out to be almost invariants.

Proposition 4.2 *We have for* $0 \le t_n = nh \le \varepsilon^{-1}$

$$\sum_{l\in\mathscr{M}} |\omega_l|^s \big|\mathscr{I}_l\big(\mathbf{z}(\varepsilon t_n)\big) - \mathscr{I}_l\big(\mathbf{z}(0)\big)\big| \le C\varepsilon^{N+3}$$

for ε *sufficiently small compared to* C_0, C_1, d, N, s *and the norm of the potential* V *with a constant* C *depending only on* C_0, C_1, d, N, s *and the norm of* V.

Proof Repeating the calculation for the derivation of the invariants of the modulation system in Sect. 4.2 we get

$$\mathscr{I}_l\big(\mathbf{z}\big(\varepsilon(t+h)\big)\big) = \mathscr{I}_l\big(\mathbf{z}(\varepsilon t)\big) + \mathrm{Re}\bigg(\sum_{\mathbf{k}} 2k_l \mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})h} \frac{\mathrm{Re}(a(-\mathrm{i}\omega_l h))}{b(-\mathrm{i}\omega_{j(\mathbf{k})}h)} \overline{w_{j(\mathbf{k})}^{\mathbf{k}}(\varepsilon t) d_j^{\mathbf{k}}(\varepsilon t)}\bigg).$$

Lemma 3 from [18] (with the adaption to the spatially discrete setting in [18, Sect. 6.2]) together with the bound (3.1b) for $a(z)$ then tells us that

$$\sum_{l\in\mathscr{M}} |\omega_l|^s \big|\mathscr{I}_l\big(\mathbf{z}(\varepsilon t_{n+1})\big) - \mathscr{I}_l\big(\mathbf{z}(\varepsilon t_n)\big)\big| \le C\|\hat{\mathbf{w}}\|_{\frac{d+1}{2}} \|\boldsymbol{\Gamma}\hat{\mathbf{e}} + \boldsymbol{\Gamma}\dot{\hat{\mathbf{h}}}\|_{\frac{d+1}{2}}$$

with a constant C depending only on d, N, s and the norm of V. Using the estimates from Lemma 4.1 on the size of the iterated modulation functions and from Lemma 4.2 on the defect of these functions, we get the statement of the proposition by summing up. □

We can show as in [18, Proposition 6] using in addition the bound (3.1b) for $a(z)$ that the almost invariants $\mathscr{I}_l$ are close to the actions I_l.

Proposition 4.3 *We have for* $0 \le t_n = nh \le \varepsilon^{-1}$

$$\sum_{l\in\mathscr{M}} |\omega_l|^s \big|\mathscr{I}_l\big(\mathbf{z}(\varepsilon t_n)\big) - I_l\big(u^n, \overline{u^n}\big)\big| \le C\varepsilon^{\frac{7}{2}}$$

for ε *sufficiently small compared to* C_0, C_1, d, N, s *and the norm of* V *with a constant* C *depending only on* C_0, C_1, d, N, s *and the norm of* V.

4.10 Interface between modulated Fourier expansions

So far, we only considered a short time interval of length ε^{-1}. In order to get longer time intervals as announced in Theorem 3.1 we have to patch many of these short time intervals together. In this subsection we consider a second short time interval $\varepsilon^{-1} \le t \le 2\varepsilon^{-1}$ (if ε^{-1} is not a multiple of the time step-size h we consider instead the time interval $n_\varepsilon h \le t \le 2n_\varepsilon h$, where n_ε denotes the largest integer with $n_\varepsilon h \le \varepsilon^{-1}$). On this second time interval we consider again a modulated Fourier expansion

$$\sum_{\mathbf{k}} \tilde{z}^{\mathbf{k}}_{j(\mathbf{k})}(\varepsilon t)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})t}$$

of the numerical solution, starting with the numerical solution u^{n_ε} at the end of the first time interval (after n_ε time steps) as initial value. The initial condition (4.2c) of the modulation system then becomes

$$u_j^{n_\varepsilon} = \sum_{\mathbf{k}} \tilde{z}^{\mathbf{k}}_j(\varepsilon n_\varepsilon h)\mathrm{e}^{-\mathrm{i}(\mathbf{k}\cdot\boldsymbol{\omega})n_\varepsilon h},$$

whereas the remaining part of the modulation system (4.2a)–(4.2c) remains unchanged. This modulation system is again solved approximately with the iterative procedure described in Sect. 4.3, and we denote, again by an abuse of notation, by $\tilde{z}^{\mathbf{k}}_j$ (and also $z^{\mathbf{k}}_j$) the iterated modulation functions after $L = 2N + 2$ iterations. Since $\|u^{n_\varepsilon}\|_s \le 2\varepsilon$ by Lemma 4.1, all the results on the modulation functions $\mathbf{z}$ proven so far are also valid for $\tilde{\mathbf{z}}$ with constants depending on the same parameters.

It is possible to control the difference of the almost invariants $\mathscr{I}_l(\mathbf{z})$ and $\mathscr{I}_l(\tilde{\mathbf{z}})$ at the interface $n_\varepsilon h \approx \varepsilon^{-1}$ between the modulated Fourier expansions on the first two time intervals.

Proposition 4.4 *We have*

$$\sum_{l\in\mathscr{M}} |\omega_l|^s \left|\mathscr{I}_l\big(\mathbf{z}(\varepsilon n_\varepsilon h)\big) - \mathscr{I}_l\big(\tilde{\mathbf{z}}(\varepsilon n_\varepsilon h)\big)\right| \le C\varepsilon^{N+3}$$

for ε sufficiently small compared to C_0, C_1, d, N, s and the norm of the potential V with a constant C depending only on C_0, C_1, d, N, s and the norm V.

Proof We first show that

$$\left|\!\left|\!\left| \hat{\mathbf{z}}(\varepsilon n_\varepsilon h) - \hat{\tilde{\mathbf{z}}}(\varepsilon n_\varepsilon h) \right|\!\right|\!\right|_{\frac{d+1}{2}} \le C\varepsilon^{N+2} \tag{4.13}$$

for the rescaled modulation functions defined in Sect. 4.4. Together with [18, Lemma 3] and Lemma 4.1 this yields the stated result. For the proof of (4.13) we have to study once more the iterative procedure, this time the iteration for $\tilde{\mathbf{z}}$. This is done in the same way as in the proof of [18, Proposition 4], considering again $\boldsymbol{\Gamma}\mathbf{a}^{(\ell)}$ for $\ell \ge 1$ and $\boldsymbol{\Gamma}\mathbf{b}^{(\ell)}$ for $\ell \ge 0$ instead of $\mathbf{a}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$ (but not $\boldsymbol{\Gamma}\mathbf{a}$). □

4.11 From short to long time intervals

We are now in the position to prove Theorem 3.1. We start with the long-time near-conservation of actions, that we can control so far only on a short time interval of length ε^{-1}. The almost invariants $\mathscr{J}_l$ permit to patch many of these short time intervals together to a long time interval of length ε^{-N} exactly as in [19, Sect. 6.3]: We consider modulated Fourier expansions on short time intervals of length $\approx \varepsilon^{-1}$ starting on the numerical solution as described in Sect. 4.10. The almost invariants (Propositions 4.2 and 4.4) close to the actions (Proposition 4.3) ensure that the numerical solution satisfies a smallness condition $\|u^n\|_s \le 2\varepsilon$ over long times $0 \le nh \le \varepsilon^{-N}$ and imply the near-conservation of actions on these time intervals.

Finally, the near-conservation of energy H, discrete energy H_M, mass m and momentum K is shown as in [19, Sect. 6.4]. The main point is that all these quantities are sums of scaled actions plus, in the case of the energies, a higher order term of size ε^4. The long-time near-conservation of actions thus implies the long-time near-conservation of discrete and continuous energy, of mass and of momentum as stated in Theorem 3.1.

5 Conclusion and open problems

We have shown long time near-conservation of actions, energy, mass and momentum for the numerical solution of a cubic Schrödinger equation given by a one-stage exponential integrator. This has been done for a class of methods that contains all reversible methods. We have presented a numerical experiment with a symmetric exponential integrator, that does not belong to this class, and that does not show this good long-time behavior.

An extension of the results to nonlinear Schrödinger equation (1.1) with more general nonlinearities of the form $g(|u(x,t)|^2)u(x,t)$ is easy if g is real analytic in a neighborhood of zero and $g(0)=0$. On the contrary, it is an open problem to extend the theoretical results of the present paper to exponential integrators with more than one stage ($s>1$ in (2.4)). The technical difficulty seems to be the identification of invariants in the corresponding modulation system. In fact, in the modulation system for a method with more than one stage there are several nonlinear terms coming from an extended potential $\mathscr{U}$ (4.3) with different arguments. This prevents the derivation of the invariants in Sect. 4.2 from working. We expect, however, that reversible exponential integrators with two or more stages have a similar long-time behavior as the one-stage methods studied here. In order to support this conjecture, we present in Fig. 3 a numerical experiment with the same parameters as in the experiments of Sect. 3.3 but with the two stage exponential integrator

$$\begin{aligned} U^1 &= \mathrm{e}^{c_1 hL} u^0 + \frac{h}{2}\Big(\frac{1}{2} f\big(U^1\big) + c_1 \mathrm{e}^{(c_1-c_2)hL} f\big(U^2\big)\Big), \\ U^2 &= \mathrm{e}^{c_2 hL} u^0 + \frac{h}{2}\Big(c_2 \mathrm{e}^{(c_2-c_1)hL} f\big(U^1\big) + \frac{1}{2} f\big(U^2\big)\Big), \\ u^1 &= \mathrm{e}^{hL} u^0 + \frac{h}{2}\big(\mathrm{e}^{(1-c_1)hL} f\big(U^1\big) + \mathrm{e}^{(1-c_2)hL} f\big(U^2\big)\big) \end{aligned} \tag{5.1}$$

with $c_1 = \frac{1}{2} - \frac{\sqrt{3}}{6}$ and $c_2 = \frac{1}{2} + \frac{\sqrt{3}}{6}$. This is a reversible and symmetric Lawson method with the Hammer and Hollingsworth method as an underlying Runge–Kutta

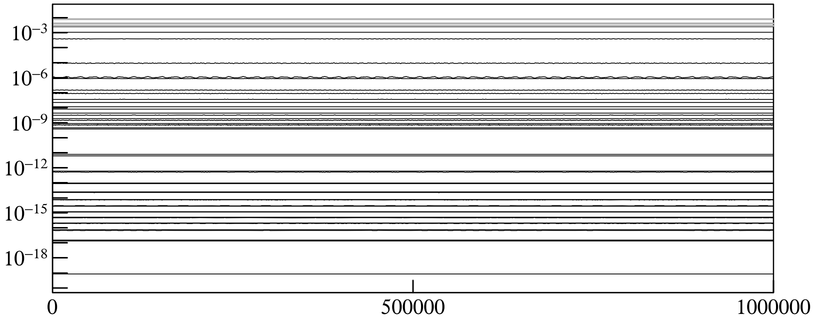

Fig. 3 Actions (*black lines*), discrete energy (*middle bold grey line*), mass (*upper bold grey line*) and momentum (*lower bold grey line*) for the two-stage method (5.1)

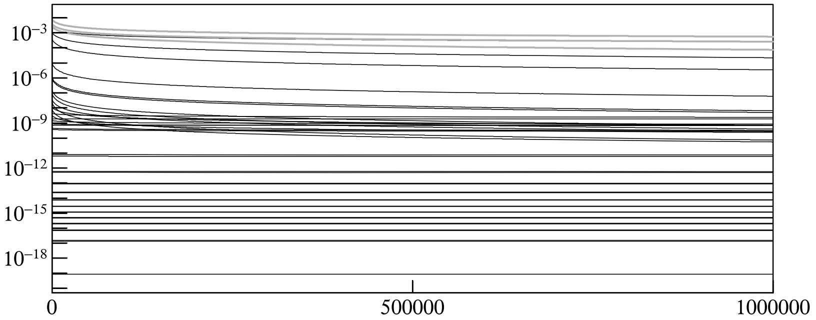

Fig. 4 Actions (*black lines*), discrete energy (*middle bold grey line*), mass (*upper bold grey line*) and momentum (*lower bold grey line*) for the pseudo steady state approximation (5.2)

method, see [7, Sect. 2]. We also perform the same experiment for the *pseudo steady-state approximation* [24, 26]

$$\begin{aligned} U^1 &= u^0, \\ U^2 &= \mathrm{e}^{hL}u^0 + h\frac{\mathrm{e}^{hL}-1}{hL}f\bigl(U^1\bigr), \\ u^1 &= \mathrm{e}^{hL}u^0 + \frac{h}{2}\frac{\mathrm{e}^{hL}-1}{hL}\bigl(f\bigl(U^1\bigr)+f\bigl(U^2\bigr)\bigr), \end{aligned} \tag{5.2}$$

a two-stage explicit exponential integrator (neither symmetric nor reversible) that is used in chemistry, and plot the result in Fig. 4. A loss of energy, mass, momentum and actions is observed for this method.

Our main result explains rigorously the good long-time behavior of certain exponential integrators in the situation where the initial values are small and the frequencies in the equation as well as the time-step size satisfy a non-resonance condition. For initial values that are not small we can not expect long-time near-conservation of actions (neither along the exact nor along a numerical solution). Concerning the behavior of energy and mass we refer to [7] for many numerical experiments when initial values are not small and frequencies are resonant. A rigorous numerical analysis of this situation on long time intervals is still missing.

We finally mention that mass can be exactly conserved by exponential integrators, whereas our main result only explains its near-conservation over long times. For example the Lawson method from Example 2.1 conserves mass exactly. For a characterization of exponential integrators, that preserve mass exactly, we refer the reader to [7, Sect. 3].

Acknowledgements We greatly appreciate the referees' comments on an earlier version. We would like to thank Christian Lubich for interesting discussions.

References

1. Bambusi, D., Grébert, B.: Birkhoff normal form for partial differential equations with tame modulus. Duke Math. J. **135**(3), 507–567 (2006)
2. Bourgain, J.: Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equations. Ann. Math. (2) **148**(2), 363–439 (1998)
3. Cano, B.: Conserved quantities of some Hamiltonian wave equations after full discretization. Numer. Math. **103**(2), 197–223 (2006)
4. Cano, B., González-Pachón, A.: Exponential time integration of solitary waves of cubic Schrödinger equations. Preprint (2011)
5. Cano, B., González-Pachón, A.: Exponential methods for the time integration of Schrödinger equation. AIP Conf. Proc. **1281**(1), 1821–1823 (2010)
6. Castella, F., Dujardin, G.: Propagation of Gevrey regularity over long times for the fully discrete Lie Trotter splitting scheme applied to the linear Schrödinger equation. M2AN Math. Model. Numer. Anal. **43**(4), 651–676 (2009)
7. Celledoni, E., Cohen, D., Owren, B.: Symmetric exponential integrators with an application to the cubic Schrödinger equation. Found. Comput. Math. **8**(3), 303–317 (2008)
8. Cohen, D., Hairer, E., Lubich, C.: Conservation of energy, momentum and actions in numerical discretizations of non-linear wave equations. Numer. Math. **110**(2), 113–143 (2008)
9. Debussche, A., Faou, E.: Modified energy for split-step methods applied to the linear Schrödinger equation. SIAM J. Numer. Anal. **47**(5), 3705–3719 (2009)
10. Dujardin, G.: Exponential Runge–Kutta methods for the Schrödinger equation. Appl. Numer. Math. **59**(8), 1839–1857 (2009)
11. Dujardin, G., Faou, E.: Normal form and long time analysis of splitting schemes for the linear Schrödinger equation with small potential. Numer. Math. **108**(2), 223–262 (2007)
12. Eliasson, L.H., Kuksin, S.B.: KAM for the nonlinear Schrödinger equation. Ann. Math. (2) **172**(1), 371–435 (2010)
13. Faou, E., Grébert, B.: Resonances in long time integration of semi linear Hamilonian PDEs. Preprint (2009). http://www.irisa.fr/ipso/perso/faou
14. Faou, E., Grébert, B.: Hamiltonian interpolation of splitting approximations for nonlinear PDEs. Found. Comput. Math. **11**(4), 381–415 (2011)
15. Faou, E., Grébert, B., Paturel, E.: Birkhoff normal form for splitting methods applied to semilinear Hamiltonian PDEs. I. Finite-dimensional discretization. Numer. Math. **114**(3), 429–458 (2010)
16. Faou, E., Grébert, B., Paturel, E.: Birkhoff normal form for splitting methods applied to semilinear Hamiltonian PDEs. II. Abstract splitting. Numer. Math. **114**(3), 459–490 (2010)
17. Gauckler, L.: 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
18. Gauckler, L., Lubich, C.: Nonlinear Schrödinger equations and their spectral semi-discretizations over long times. Found. Comput. Math. **10**(2), 141–169 (2010)
19. Gauckler, L., Lubich, C.: Splitting integrators for nonlinear Schrödinger equations over long times. Found. Comput. Math. **10**(3), 275–302 (2010)

20. Hairer, E., Lubich, C., Wanner, G.: Geometric numerical integration. In: Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer Series in Computational Mathematics, vol. 31. Springer, Berlin (2006)
21. Hochbruck, M., Lubich, C., Selhofer, H.: Exponential integrators for large systems of differential equations. SIAM J. Sci. Comput. **19**(5), 1552–1574 (1998) (electronic)
22. Hochbruck, M., Ostermann, A.: Exponential integrators. Acta Numer. **19**, 209–286 (2010)
23. Lawson, J.D.: Generalized Runge–Kutta processes for stable systems with large Lipschitz constants. SIAM J. Numer. Anal. **4**, 372–380 (1967)
24. Minchev, B., Wright, W.M.: A review of exponential integrators for semilinear problems. Tech. rep. 2/05, Department of Mathematical Sciences, NTNU, Norway (2005). http://www.math.ntnu.no/preprint/
25. Sulem, C., Sulem, P.L.: The Nonlinear Schrödinger Equation. Applied Mathematical Sciences, vol. 139. Springer, New York (1999)
26. Verwer, J.G., van Loon, M.: An evaluation of explicit pseudo-steady-state approximation schemes for stiff ODE systems from chemical kinetics. J. Comput. Phys. **113**(2), 347–352 (1994)