Zurich Lectures in Advanced Mathematics

Mathematics in Zurich has a long and distinguished tradition, in which the writing of lecture notes volumes and research monographs plays a prominent part. The *Zurich Lectures in Advanced Mathematics* series aims to make some of these publications better known to a wider audience. The series has three main constituents: lecture notes on advanced topics given by internationally renowned experts, graduate text books designed for the joint graduate program in Mathematics of the ETH and the University of Zurich, as well as contributions from researchers in residence at the mathematics research institute, FIM-ETH. Moderately priced, concise and lively in style, the volumes of this series will appeal to researchers and students alike, who seek an informed introduction to important areas of current research.

Previously published in this series:

Yakov B. Pesin, *Lectures on partial hyperbolicity and stable ergodicity*
Sun-Yung Alice Chang, *Non-linear Elliptic Equations in Conformal Geometry*
Sergei B. Kuksin, *Randomly forced nonlinear PDEs and statistical hydrodynamics in 2 space dimensions*
Pavel Etingof, *Calogero-Moser systems and representation theory*
Guus Balkema and Paul Embrechts, *High Risk Scenarios and Extremes – A geometric approach*
Demetrios Christodoulou, *Mathematical Problems of General Relativity I*
Camillo De Lellis, *Rectifiable Sets, Densities and Tangent Measures*
Paul Seidel, *Fukaya Categories and Picard–Lefschetz Theory*
Alexander H.W. *Schmitt, Geometric Invariant Theory and Decorated Principal Bundles*
Michael Farber, *Invitation to Topological Robotics*
Alexander Barvinok, *Integer Points in Polyhedra*
Christian Lubich, *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*
Shmuel Onn, *Nonlinear Discrete Optimization – An Algorithmic Theory*
Kenji Nakanishi and Wilhelm Schlag, *Invariant Manifolds and Dispersive Hamiltonian Evolution Equations*

Published with the support of the Huber-Kudlich-Stiftung, Zürich

Erwan Faou

Geometric Numerical Integration and Schrödinger Equations

European Mathematical Society

Author:

Erwan Faou
INRIA & ENS Cachan Bretagne
Département de mathématiques
Avenue Robert Schumann
Campus de Ker Lann
35170 Bruz
France

2010 Mathematics Subject Classification: 65P10, 37M15, 35Q41

Key words: Geometric numerical integration, symplectic integrators, backward error analysis, Schrödinger equations, long time behavior

ISBN 978-3-03719-100-2

The Swiss National Library lists this publication in The Swiss Book, the Swiss national bibliography, and the detailed bibliographic data are available on the Internet at http://www.helveticat.ch.

Contact address:

European Mathematical Society Publishing House
Seminar for Applied Mathematics
ETH-Zentrum FLI C4
CH-8092 Zürich
Switzerland

Phone: +41 (0)44 632 34 36
Email: info@ems-ph.org
Homepage: www.ems-ph.org

Printing and binding: Beltz Bad Langensalza GmbH, Bad Langensalza, Germany
∞ Printed on acid free paper
9 8 7 6 5 4 3 2 1

Preface

The goal of this book is to give some answers to the following general question: *How, and to which extent can we simulate numerically the long time behavior of Hamiltonian partial differential equations* typically arising in many application fields such as quantum mechanics or wave propagations phenomena. Starting from numerical examples, these notes try to provide a relatively complete analysis of the case of the Schrödinger equation in a simple setting (periodic boundary conditions, polynomial nonlinearities) approximated by splitting methods. The objective of this book is to analyze the possible stability and instability phenomena induced by space and time discretizations, and to provide rigorous mathematical explanations for them.

The results presented here originate from many collaborations done in the last 4 years. In particular, Chapter VI is largely inspired by joint works with Arnaud Debussche and Guillaume Dujardin. Chapter VII only exists because of several years of common work with Benoît Grébert. The final results of Chapter VII have been obtained with Rémi Carles. I am happy to warmly thank all of them for their contribution to the present analysis. Many parts of these notes have also taken benefit of many discussions and interactions with several mathematicians before and during my stay at ETH: Dario Bambusi, David Cohen, Ludwig Gauckler, Pierre Germain, Vasile Gradinaru, Ernst Hairer, Ralph Hiptmair, Thomas Kappeler, Peter Kauf, Christian Lubich, Eric Paturel, Katharina Schratz, Christoph Schwab and Julia Schweitzer. My sincere thanks go to all of them.

This book contains the notes of a graduate course (Nachdiplom Vorlesung) held at ETH Zürich in the spring semester 2010 and I would like to thank the FIM (Forschungsinstitut für Mathematik) for its hospitality during my stay at ETH, as well as all the members of the SAM (Seminar for applied mathematics) for their warm welcome. My grateful thanks also go to the participants to my lecture in Zürich.

Contents

I Introduction

The goal of this lecture is to give some results in the *geometric numerical integration* theory of linear and semi-linear Hamiltonian partial differential equations (PDEs). This means that we will study the ability of numerical schemes to reproduce *qualitative* properties of Hamiltonian PDEs over *long time periods*, properties such as preservation of the Hamiltonian, or energy exchanges between the eigenmodes of a solution. Rather than setting this study in a general abstract framework (as for instance in [15]), we will focus on linear and nonlinear Schrödinger equations, typically with polynomial nonlinearity. The results presented in these lecture notes follow the lines of [12] for the linear case, and [15] for the nonlinear case. The final Chapter VII gives a picture of the possible instabilities induced by numerical discretization – and ways to prevent them.

Before tackling the infinite dimensional case, we recall that many works exist in the finite dimensional case (ordinary differential equations): see [26] and [34]. We will discuss them in Chapter II. Relevant results concerning PDEs were obtained more recently, and using different techniques: see [9], [12], [13], [17], [18], [20], [21]. We will discuss these references throughout the text.

In this first chapter, we would like to show by numerical examples some nice or pathological behaviors observed in simulations obtained by using *splitting schemes* naturally induced by decomposition between the kinetic and potential parts. Such schemes are very easy to implement and for this reason, widely used in practical simulations (see for instance [3], [4], [30], [31] and the references therein). They also preserve the symplectic structure and the L^2 norm of the solution. For these reasons, we will restrict our analysis to such splitting methods, but consider many different situations: semi-discrete, implicit-explicit and fully discrete schemes.

1 Schrödinger equation

Let us consider the cubic nonlinear Schrödinger equation

$$i\,\partial_t u(t,x) = -\Delta u(t,x) + V(x)u(t,x) + \lambda |u(t,x)|^2 u(t,x), \quad u(0,x) = u^0(x) \tag{I.1}$$

where $u(t,x)$ is the wave function depending on the time $t \in \mathbb{R}$. We assume here periodic boundary conditions, which means that the space variable x belongs to the d-dimensional torus $\mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$. The function $V(x)$ is a real interaction potential function, and the operator $\Delta = \sum_{i=1}^d \partial_{x_i}^2$ is the Laplace operator. The constant λ is a real parameter. As initial condition, we impose that the function $u(t,x)$ at time $t = 0$ is equal to a given function u^0.

Such equations arise in many applications such as quantum dynamics and nonlinear optics. We refer to [36] for modeling aspects, and to [8] for the mathematical theory. The cubic nonlinearity arises in particular in the simulation of Bose–Einstein condensates (see for instance [3], [4]) while the case where $\lambda = 0$ constitutes the classical linear Schrödinger equation associated with a typical interaction potential $V(x)$.

Equation (I.1) is a Hamiltonian partial differential equation (PDE) possessing strong conservation properties. In quantum mechanics, the quantity $|u(t,x)|^2$ represents the probability density of finding the system in state x at time t, which is reflected by preservation of the L^2 norm: For any solution $u(t,x)$ we have

$$\|u(t,x)\|_{L^2}^2 = \frac{1}{(2\pi)^d}\int_{\mathbb{T}^d} |u(t,x)|^2 \mathrm{d}x = \|u(0,x)\|_{L^2}^2 .$$

Note that for concrete applications, many physical constants are present in equation (I.1) depending on the mass of the particle or the Planck constant. Here we consider a normalized version of the Schrödinger equation and address the question of its numerical approximation in relation with its Hamiltonian structure only. Specific algorithms for the semi-classical regime can be found for instance in [14] and [30]. Results concerning the case of the Gross–Pitaevskii associated with the harmonic oscillator, i.e. when $V(x) = x^2$ and $x \in \mathbb{R}$, can be also found in [3], [4], [19].

With the equation (I.1) is associated the *Hamiltonian energy* defined for any function u by the formula

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

where $|\nabla u|^2 = \sum_{i=1}^d |\partial_{x_i} u|^2$. This energy is preserved throughout the solution: for all times $t \in \mathbb{R}$ where the solution is defined and sufficiently smooth, we have

$$H(u(t),\bar{u}(t)) = H(u(0),\bar{u}(0)).$$

Note that this energy can be split into

$$H(u,\bar{u}) = T(u,\bar{u}) + P(u,\bar{u}), \tag{I.2}$$

where

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

is the kinetic energy of the system and

$$P(u,\bar{u}) = \frac{1}{(2\pi)^d}\int_{\mathbb{T}^d} V(x)|u(x)|^2 + \frac{\lambda}{2}|u(x)|^4 \mathrm{d}x$$

is the potential energy.

The goal of this lecture is to analyze the qualitative properties of numerical schemes applied to (I.1) and to discuss their long time behavior. In particular, we will try to show that in some situations, numerical method can or cannot reproduce physical properties of the Schrödinger equation, such as conservation of energy, stability of solitary waves, energy exchanges between modes, and preservation of regularity over long time periods.

2 Numerical schemes

One of the easiest ways to derive numerical schemes for (I.1) is to split the system according to the decomposition (I.2). For ease of presentation, we will mainly consider the case where $d = 1$.

2.1 Free Schrödinger equation. Let us consider the system

$$i\partial_t u(t,x) = -\Delta u(t,x), \quad u(0,x) = u^0(x), \tag{I.3}$$

set on the one-dimensional torus $\mathbb{T}$. To solve this system, we consider the Fourier transform $(\xi_a(t))_{a\in\mathbb{Z}}$ of $u(t,x)$ defined by

$$\widehat{(u(t,x))}_a = \xi_a(t) := \frac{1}{2\pi}\int_0^{2\pi} u(t,x)e^{-iax}\mathrm{d}x, \quad a \in \mathbb{Z},$$

and we plug the decomposition

$$u(t,x) = \sum_{a\in\mathbb{Z}} \xi_a(t)e^{iax}$$

into (I.3). Owing to the fact that $\widehat{(\partial_x u)}_a = ia\xi_a$, we see that (I.3) is equivalent to the collection of ordinary differential equations

$$\forall\, a \in \mathbb{Z}, \quad i\frac{\mathrm{d}}{\mathrm{d}t}\xi_a(t) = a^2\xi_a(t), \quad \xi_a(0) = \xi_a^0,$$

where ξ_a^0 are the Fourier coefficients of the initial function u^0. The solution of this equation can be written explicitly $\xi_a(t) = e^{-ita^2}\xi_a^0$. Hence in Fourier variables, the solution of the free Schrödinger equation can be computed exactly. Note that we have for all t, $|\xi_a(t)| = |\xi_a(0)|$. This means that the regularity of u^0, measured by the decay of the Fourier coefficients $\xi_a(t)$ with respect to $|a|$, is preserved by the flow of the kinetic part. We denote the solution of (I.3) by

$$u(t) = \varphi_T^t(u^0)$$

as the exact flow of the Hamiltonian PDE associated with the Hamiltonian T.

2.2 Potential part. Let us now consider the system

$$i\partial_t u(t,x) = V(x)u(t,x) + \lambda|u(t,x)|^2 u(t,x), \quad u(0,x) = u^0. \tag{I.4}$$

In this equation, we observe that x can be considered as a parameter (there is no derivative in x). Moreover, as V is real, the complex conjugate $\bar{u}(t,x)$ satisfies the equation

$$-i\partial_t \bar{u}(t,x) = V(x)\bar{u}(t,x) + \lambda|u(t,x)|^2 \bar{u}(t,x),$$

hence we see that for all t, we have for all x,

$$\begin{aligned}\partial_t |u(t,x)|^2 &= u(t,x)\partial_t \bar{u}(t,x) + \bar{u}(t,x)\partial_t u(t,x)\\ &= \left(V(x) + \lambda|u(t,x)|^2\right)\left(iu(t,x)\bar{u}(t,x) - i\bar{u}(t,x)u(t,x)\right)\\ &= 0,\end{aligned}$$

which means that the solution of (I.4) preserves the modulus of $u^0(x)$ for all fixed $x \in \mathbb{T}$: we have for all t, $|u(t,x)| = |u^0(x)|$. As an immediate consequence, the exact solution of (I.4) is given by

$$u(t,x) = \exp\left(-itV(x) - it\lambda|u^0(x)|^2\right)u^0(x).$$

We denote this solution by

$$u(t) = \varphi_P^t(u^0).$$

2.3 Splitting schemes. The previous paragraphs showed that we can solve exactly the Hamiltonian equations associated with the kinetic energy $T(u,\bar{u})$ and with the potential energy $P(u,\bar{u})$ appearing in the decomposition (I.2). Splitting schemes are based on this property: they consist in solving alternatively the free Schrödinger equation and the potential part. Denoting by φ_{T+P}^τ the exact flow defining the solution of the equation (I.1) (we will give a precise definition of this flow in Chapter III), then for a small time step $\tau > 0$, this leads to building the approximation

$$\varphi_{T+P}^\tau \simeq \varphi_T^\tau \circ \varphi_P^\tau, \tag{I.5}$$

known as the Lie splitting method. For a time $t = n\tau$, the solution is then approximated by

$$u(n\tau) \simeq u^n = \left(\varphi_T^\tau \circ \varphi_P^\tau\right)^n (u^0).$$

We will see later that this approximation is actually convergent in the following sense: if the solution $u(t,\cdot) = u(t)$ of (I.1) remains smooth in an interval $[0,T]$, then we have

$$\forall n\tau \in [0,T], \quad \|u(n\tau) - u^n\|_{L^2} \le C(T,u)\tau. \tag{I.6}$$

Here smooth means that the Fourier coefficients satisfy some decay properties uniformly in time and the constant $C(T,u)$ depends on the final time T and on *a priori* bounds on derivatives of the exact solution $u(t)$. Such a result is related to the Baker–Campbell–Hausdorff (BCH) formula which states that the error made in the approximation (I.5) is small and depends on the commutator between the two Hamiltonians T and P. In Chapter II, we will present a proof of this BCH formula, while convergence results are presented in Chapter IV.

Another approximation, known as the Strang splitting scheme, is given by

$$\varphi_{T+P}^{\tau} \simeq \varphi_{P}^{\tau/2} \circ \varphi_{T}^{\tau} \circ \varphi_{P}^{\tau/2}, \tag{I.7}$$

and it can be proved that this approximation is of order 2, which means that the error in (I.6) is $\mathcal{O}(\tau^2)$ provided the solution $u(t)$ remains smooth enough. More generally, high order splitting schemes can be constructed, but each time, their approximation properties rely on the *a priori* assumption that the solution remains smooth over the (finite) time interval considered (see for instance [27]).

Natural questions then arise: do these schemes preserve the energy over a long time? Do they preserve the regularity of the initial value over a long time? Are they stable? Do they correctly reproduce possible nonlinear exchanges between the modes $\xi_a(t)$? These questions constitute central questions of *geometric numerical integration* theory whose general aim is the study of the qualitative behavior of numerical schemes over a long time (see the classical references [26] and [34]). Note that since splitting schemes are built from exact solutions of Hamiltonian PDEs, they are naturally *symplectic*, something that is known to be fundamental to ensure the good behavior of numerical schemes applied to Hamiltonian ordinary differential equations.

Indeed, in the finite dimensional situation, a fundamental result known as *backward error analysis* shows that the numerical trajectory given by a symplectic integrator applied to a Hamiltonian ODE (almost) coincides with the exact solution of a *modified Hamiltonian system* over an extremely long time. This result implies in particular the existence of a modified energy preserved throughout the numerical solution, which turns out to be close to the original one. Before studying the case of Hamiltonian PDEs, we will consider extensively the finite dimensional situation in Chapter II, following the classical references in the field [5], [25], [26], [33], [34].

2.4 Practical implementation. To implement the previous splitting schemes, we define the grid $x_a = 2\pi a/K$ where K is an integer, and $a \in B^K$ belongs to a finite set $B^K \subset \mathbb{Z}$ depending on the parity of K:

$$B^K := \begin{cases} \{-P,\dots,P-1\} & \text{if} \quad K = 2P \quad \text{is even,} \\ \{-P,\dots,P\} & \text{if} \quad K = 2P+1 \quad \text{is odd.} \end{cases} \tag{I.8}$$

Note that in any case, $\sharp B^K = K$, and that the points x_a, $a \in B^K$ are made of K equidistant points in the interval $[-\pi,\pi]$. The discrete Fourier transform is defined as the mapping $\mathcal{F}_K : B^K \to B^K$ such that for all $v = (v_a) \in B^K$ with $a \in B^K$,

$$(\mathcal{F}_K v)_a = \frac{1}{K} \sum_{b \in B^K} e^{-2i\pi ab/K} v_b.$$

Its inverse is given by

$$\left(\mathcal{F}_K^{-1} v\right)_a = \sum_{b \in B^K} e^{2i\pi ab/K} v_b.$$

This Fourier transform entails many advantages. In particular, we can verify that $\sqrt{K}\mathcal{F}_K$ is a unitary transformation, and moreover, it can be easily computed using the Fast Fourier Transform algorithm, requiring a number of operations of order $\mathcal{O}(K \log K)$ instead of $\mathcal{O}(K^2)$ as a naive approach would indicate.

The practical implementation of the (abstract) splitting method

$$u(n\tau) \simeq u^n = \left(\varphi_T^\tau \circ \varphi_P^\tau\right)^n u^0$$

then consists in the approximation of the function $U^{K,n}(x)$ at each time step, evaluated at the grid points by the collection of numbers $v_b^{K,n}$, $b \in B^K$ such that

$$v_b^{K,n} \simeq u^n(x_b) \simeq u\left(n\tau, x_b\right).$$

Hence we see that K and τ represent the space and time discretization parameters respectively.

The algorithm to compute the numbers $v_b^{K,n+1}$ from the collection of numbers $v_b^{K,n}$ then reads:

1. Calculate the approximation

$$v_b^{K,n+1/2} = \exp\left(-i\tau V(x_b) - i\tau\lambda \left|v_b^{K,n}\right|^2\right) v_b^{K,n} \simeq (\varphi_P^\tau u^n)(x_b).$$

2. Take the Fourier transform

$$\xi_a^{K,n+1/2} = \left(\mathcal{F}_K v^{K,n+1/2}\right)_a, \quad a \in B^K.$$

3. Compute the solution of the free Schrödinger equation in Fourier variables

$$\xi_a^{K,n+1} = \exp\left(-i\tau a^2\right) \xi_a^{K,n+1/2}.$$

4. Take the inverse Fourier transform

$$v_b^{K,n+1} = \left(\mathcal{F}_K^{-1} \xi^{K,n+1}\right)_b \quad b \in B^K.$$

We can also interpret this algorithm as a splitting method for a finite dimensional system of the form

$$i\frac{\mathrm{d}}{\mathrm{d}t}\xi_a^K = a^2\xi_a^K + Q^K\left(\xi^K\right), \quad a \in B^K, \tag{I.9}$$

where $Q^K(\xi)$ is a nonlinear potential depending on K and on the Fourier coefficients ξ_a^K, $a \in B^K$, given roughly speaking by $Q^K = \mathcal{F}_K \circ \mathcal{P} \circ \mathcal{F}_K^{-1}$ where $\mathcal{P}$ is the potential part in (I.1) evaluated at the grid points. In terms of Fourier coefficients, Q^K can be viewed as a polynomial in the (large but) finite number of parameters ξ_a^K and $\bar{\xi}_a^K$, $a \in B^K$.

In Chapter IV, we will show that the previous scheme is convergent in the following sense: The trigonometric polynomial function $U^{K,n}(x) = \sum_{a\in B^K} \xi_a^{K,n} e^{iax}$ associated with the discrete Fourier coefficients $\xi_a^{K,n}$ defined above, constitutes an approximation of the exact solution $u(t,x)$ at time $t_n = n\tau \le T$, and we have the estimate

$$\forall\, t_n = n\tau \le T, \qquad \left\| U^{K,n}(x) - u(t_n,x) \right\|_{\ell^1} \le C(T,u)(\tau + K^{-s}), \tag{I.10}$$

where s is given by the *a priori* regularity of the exact solution $u(t,x)$ over the time interval $[0,T]$.

Note that in the previous formula, the error is measured in the ℓ^1 functional space associated with the norm

$$\|u\|_{\ell^1} = \sum_{a\in\mathbb{Z}} |\xi_a|, \quad \text{if} \quad u(x) = \sum_{a\in\mathbb{Z}} \xi_a\, e^{iax},$$

and called the Wiener algebra. This choice is driven by the simplicity of polynomial manipulations when acting on ℓ^1. In these notes, ℓ^1-based function spaces will constitute our main framework, though a similar analysis could be performed using standard Sobolev spaces H^s for s sufficiently large.

In the following, we will sometimes interpret the previous fully discretized algorithm as an (abstract) splitting method applied to a Hamiltonian PDE of the form

$$i\partial_t u = \frac{1}{\tau}\beta(-\tau\Delta)u + Q^K(u), \tag{I.11}$$

where β is a cut-off function such that $\beta(x) = x$ for $|x| \le \mathsf{cfl}$ and $\beta(x) = 0$ for $|x| > \mathsf{cfl}$ where the constant cfl corresponds to a Courant–Friedrich–Lewy (CFL) condition, see [10]. In the practical implementation described above, we have $\mathsf{cfl} = \tau K^2/4$ corresponding to the time step τ multiplied by the greatest eigenvalue of the discrete Laplace operator. In this situation, the potential Q^K will be assumed to satisfy bounds independent of K, and the analysis can then be made by only considering (I.11) with a given CFL number and a fixed polynomial potential $Q = Q^K$. This will be our abstract framework.

2.5 Semi-implicit schemes. As they are explicit schemes, splitting methods have the big advantage of their simplicity of implementation and their relatively low numerical cost. However, as we will see later, these schemes require often a strong CFL condition to be efficient. Even in the linear case ($\lambda = 0$ in (I.1)) they can lead to instabilities due to *numerical resonance* problems. The use of implicit or semi-implicit schemes often allows us to attenuate, if not avoid, these problems.

Let us consider a general semi-linear equation

$$i\partial_t u = -\Delta u + Q(u),$$

where Q is polynomial in u and $\bar{u}$. The midpoint approximation scheme is defined as the map $u^n \mapsto u^{n+1}$ such that

$$i\frac{u^{n+1}-u^n}{\tau} = -\Delta\left(\frac{u^{n+1}+u^n}{2}\right) + Q\left(\frac{u^{n+1}+u^n}{2}\right).$$

It turns out that this map is symplectic, but its practical computation requires solving a large nonlinear implicit problem at each time step.

An alternative consists in a combination of the splitting approach described above with an approximation of the solution of the free-linear Schrödinger by the midpoint method. Actually when $Q = 0$ in the previous equation, we can write down explicitly

$$u^{n+1} = R(i\tau\Delta)u^n := \left(\frac{1+i\tau\Delta/2}{1-i\tau\Delta/2}\right)u^n, \tag{I.12}$$

where this last expression is well defined in Fourier variables by the formula

$$\xi_a^{n+1} = \left(\frac{1-i\tau a^2/2}{1+i\tau a^2/2}\right)\xi_a^n, \quad a \in \mathbb{Z}, \tag{I.13}$$

where ξ_a^n are the Fourier coefficients of u^n on the torus $\mathbb{T}$. Note that this expression is explicit in the Fourier space. In a more general situation one has to rely on a linearly implicit equation to determine u^{n+1} in (I.14) at each step.

Instead of considering fully explicit splitting of the form (I.5), we can also consider semi-implicit schemes of the form

$$\varphi_{T+P}^\tau \simeq R(i\tau\Delta)\circ\varphi_P^\tau. \tag{I.14}$$

Such an algorithm can be viewed as the standard splitting scheme (I.5), where we replace the exact flow φ_T^τ by its approximation by the midpoint rule. Note that as the implicit midpoint is an order 2 scheme, such a numerical scheme will remain of order 1, which means that such an approximation will remain convergent for smooth solutions over finite time.

Before going on, let us mention that we can again interpret the previous implicit-explicit splitting method as a *classical* splitting method applied to a modified Hamiltonian PDE of the form (I.11). Indeed, for real number x, we have

$$\frac{1+ix}{1-ix} = \exp\left(2i\arctan(x)\right).$$

Hence the relation (I.13) can be written

$$\xi_a^{n+1} = \exp\left(-2i \text{ arctan}\left(\tau a^2/2\right)\right) \xi_a^n.$$

In an equivalent formulation, we can write $R(i\tau\Delta) = \exp(-2i \text{ arctan}(\tau\Delta/2))$, which means that the midpoint rule applied to the free Schrödinger equation is equivalent to the exact solution at time τ of the equation

$$i\partial_t u(t,x) = \frac{2}{\tau} \text{arctan}\left(-\frac{\tau\Delta}{2}\right) u(t,x). \tag{I.15}$$

We thus see that an implicit-explicit scheme can be again viewed as a standard splitting method applied to a modified equation of the form (I.11) where $\beta(x) = \frac{2}{\tau} \text{arctan}(x/2)$. Note the striking fact that the arctan function acts here as a regularized CFL condition: the high frequencies in equation (I.15) are smoothed, and the linear operator is now a (large but) bounded operator.

3 Examples

We now give various numerical examples of qualitative behavior of the previous schemes applied to (I.1).

3.1 Solitary waves. Let us consider the equation

$$i\partial_t u(t,x) = -\partial_{xx} u(t,x) - |u(t,x)|^2 u(t,x), \quad u(t,x) = u^0,$$

set on the real line, $x \in \mathbb{R}$, and for which there exists the particular family of solutions

$$u(t,x) = \rho(x - ct - x_0) \exp\left(i\left(\frac{1}{2}c(x - ct - x_0) + \theta_0\right)\right) \exp\left(i\left(\alpha + \frac{1}{4}c^2\right)t\right),$$

where α, c, x_0 and θ_0 are real parameters, and where

$$\rho(x) = \frac{\sqrt{2\alpha}}{\cosh(\sqrt{\alpha}x)}.$$

These solutions are called solitons or solitary waves, and they are stable in the sense that if the initial data is close to such a solution, it will remain close to this family of solutions over arbitrary long time periods. This is called the orbital stability (we refer to [8] and the reference therein).

Here, we aim at approximating the very particular solution corresponding to $\alpha = 1$, $c = 0$, $x_0 = 0$ and $\theta_0 = 0$, i.e. the solution

$$u(t,x) = \frac{\sqrt{2}e^{it}}{\cosh(x)}.$$

We first consider the standard Strang splitting method (I.7). As space discretization, we introduce a large window $[-\pi/L, \pi/L]$ where L is a small parameter, and use the spectral discretization method described in the previous section. This is justified because the solution we aim at simulating is exponentially decreasing with respect to $|x|$ and the approximation on the large windows will be correct for a small number L. In this scaled situation the CFL number is given by

$$\mathsf{cfl} = \tau L^2 \left(\frac{K}{2}\right)^2. \tag{I.16}$$

We take $K = 256$, $L = 0.11$ and $\tau = 0.1$ ($\mathsf{cfl} = 19.8$), $\tau = 0.05$ ($\mathsf{cfl} = 9.9$) and $\tau = 0.01$ ($\mathsf{cfl} = 1.9$).

In Figure I.1, we plot the evolution of the discrete approximation of the energy

$$H(u, \bar{u}) = \int_{\mathbb{R}} |\partial_x u(x)|^2 - \frac{1}{2}|u(x)|^4 \mathrm{d}x$$

throughout the numerical solution, with respect to time. We see that in the two cases $\mathsf{cfl} = 19.8$ and $\mathsf{cfl} = 9.9$, there is a serious drift, while in the case $\mathsf{cfl} = 1.9$, we observe a good preservation of energy.

In Figure I.2, we plot the absolute value of the numerical solution $|u^n(x)|$. In the case where $\mathsf{cfl} = 19.8$ we observe a deterioration at time $t = 300$ where the regularity of the initial solution seems to be lost. The bottom figure is obtained with a CFL number $\mathsf{cfl} = 1.9$ and we observe that the numerical solution is particularly stable. The profile of solution is almost the same as for the initial solution. This picture is drawn at time $t = 10000$.

To have a better understanding of the phenomenon, we plot the evolution of the *actions* associated with the numerical solution, i.e. the Fourier coefficients $|\xi_a(t)|^2$ for $a \in \mathbb{Z}$. In Figure I.3, we plot the evolution of these actions in logarithmic scale in the case where $\mathsf{cfl} = 19.8$. Since the function is regular, there is an exponential

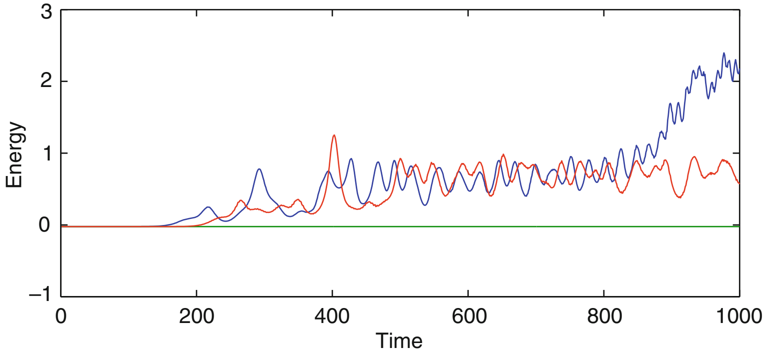


Figure I.1. Evolution of energy for the Strang splitting with cfl = 19.8, 9.9 and 1.9.

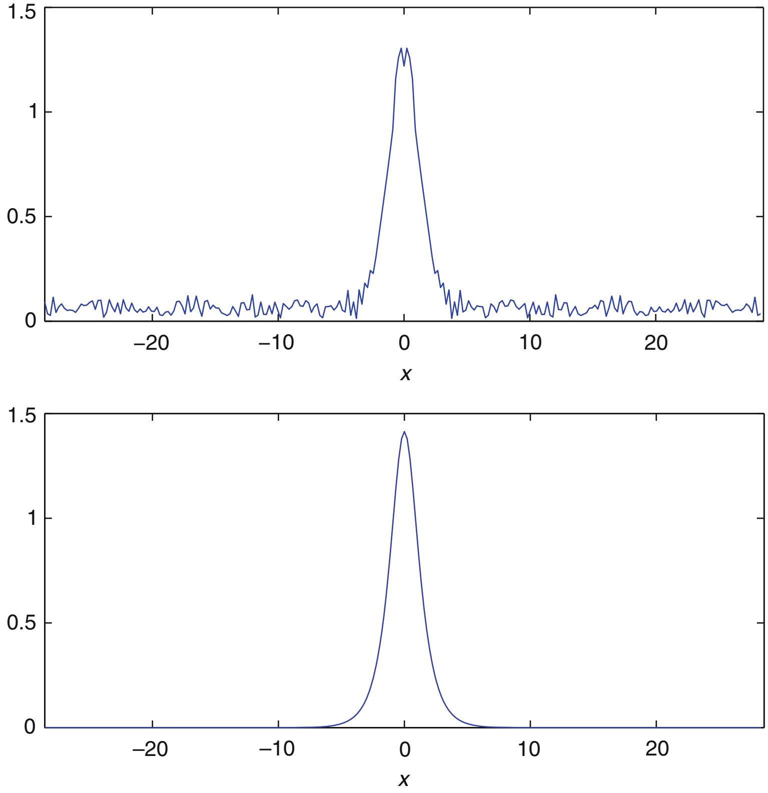


Figure I.2. $|u^n(x)|$ for the Lie splitting with cfl = 19.8 at time $t = 300$ (top) and cfl = 1.9 at time $t = 10000$ (bottom).

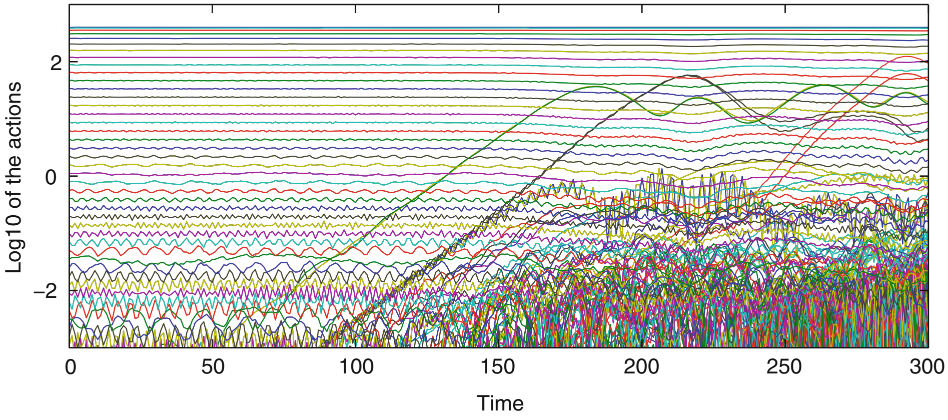


Figure I.3. Evolution of the actions for the Lie splitting with cfl = 19.8.

decay of the actions with respect to k, and the high modes are plotted at the bottom of the figure while the low modes are up. We observe that there are unexpected energy exchanges with the high modes: there is an energy leak from the low modes to the high modes producing a loss in the regularity of the solution.

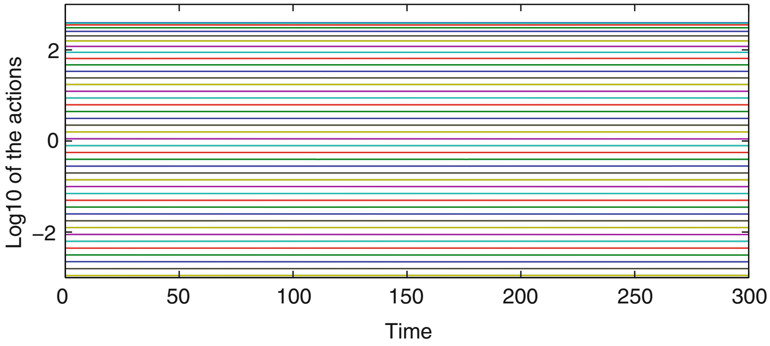


Figure I.4. Evolution of the actions for the Lie splitting with cfl = 1.9.

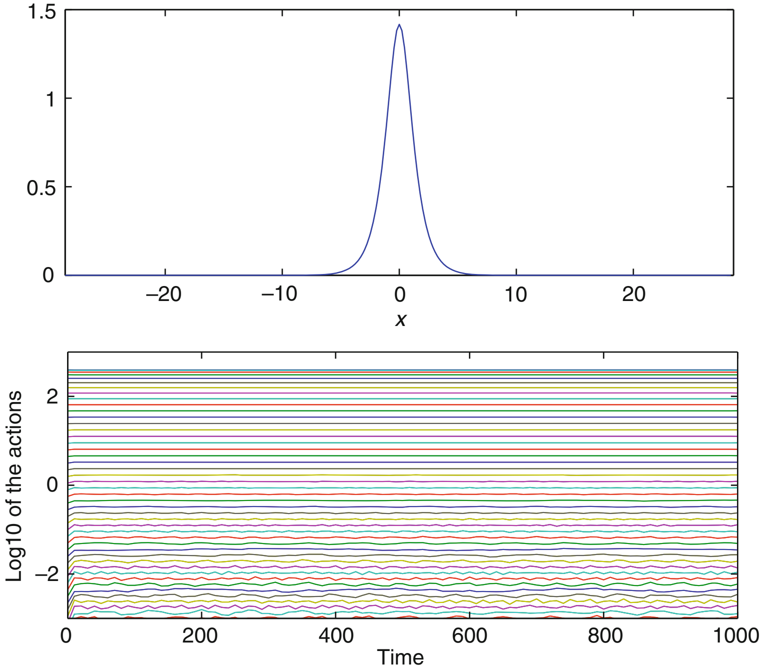


Figure I.5. Implicit-explicit integrator with cfl $=$ 19.8. Profile at $t = 1000$ (top) and evolution of the actions (bottom).

This phenomenon does not appear in the case where $\mathsf{cfl} = 1.9$, as shown in Figure I.4: the regularity of the solution expressed by the arithmetic decay of the actions in logarithmic scale is preserved over a very long time.

Now we repeat the same computations but with the implicit-explicit integrator (I.14). In Figure I.5 we plot both the evolution of the actions and the absolute value of the numerical solution at time $t = 1000$ by using a CFL condition of order $\mathsf{cfl} = 19.8$. Note that the results obtained are comparable to the classical splitting with $\mathsf{cfl} = 1.9$. In particular, we observe no deterioration of the regularity of the solution, and no energy drift.

3.2 Linear equations. The previous section showed that preservation of energy and long time behavior of the numerical solution are linked with the CFL number used in the simulation. To understand this phenomenon, we now consider the linear equation

$$i\partial_t u(t,x) = -\partial_{xx} u(t,x) - V(x)u(t,x), \quad u(t,x) = u^0,$$

with periodic boundary conditions ($x \in \mathbb{T}$) and where $V(x)$ and the initial solution are analytic. More precisely, we take

$$V(x) = \cos(x) + \cos(6x) \quad \text{and} \quad u^0 = \frac{2}{2 - \cos(x)}.$$

In Figure I.6, we plot the maximal deviation of the energy

$$H(u,\bar{u}) = \frac{1}{2\pi}\int_{\mathbb{T}} |\partial_x u(x)|^2 - V(x)|u(x)|^2 \mathrm{d}x,$$

between $t = 0$ and $t = 30$. For a fixed time step τ, we define a numerical solution u_τ^n from $t = 0$ to $t = 30$ (and hence $n\tau \leq T = 30$). With this discrete solution in hand, we compute the maximal energy deviation

$$E(\tau) := \max_{n,\, n\tau \in (0,30)} |H(u_\tau^n) - H(u^0)|.$$

We repeat this computation for time steps τ running from 0.01 to 0.1. We take $K = 128$ in this situation, so that the CFL condition runs from $\mathsf{cfl} = 40$ to 400. Note that the final time $t = 30$ cannot be considered as a very long time (it is of order τ^{-1}), however we are interested here in the behavior of the mapping $\tau \mapsto E(\tau)$ to have a better understanding on the possible existence of a modified energy for the numerical scheme, particularly for large CFL numbers.

In Figure I.6, we plot the function $\tau \mapsto E(\tau)$ for the explicit splitting (I.5) (top) and the same result for the implicit-explicit integrator (bottom).

What we observe is that the function $E(\tau)$ is not regular in τ in the case of the Lie splitting while it seems to be smoother for the implicit-explicit integrator. More precisely, in the case of the Strang splitting, for some specific values of the step size,

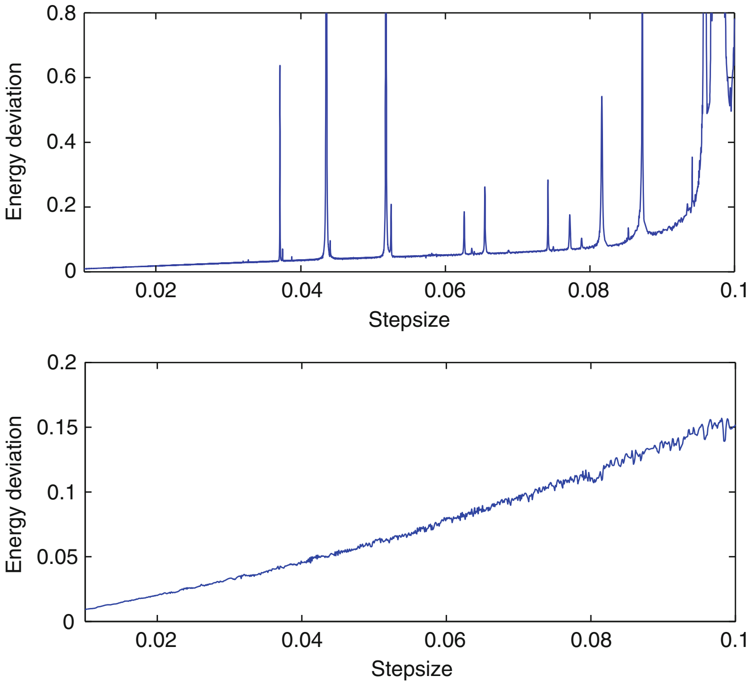


Figure I.6. Energy deviation as function of a time step for a Lie splitting (top) and the implicit-explicit scheme (bottom).

there is a drift in energy, while outside these pathological situations, the energy seems to be better preserved. Such particular time steps are called *resonant* step sizes.

To have a better view of the effect of these resonant step sizes, let us again plot the evolution of the actions in the case where the potential is small:

$$V(x) = 0.01 \frac{3}{5 - 4\sin(x)} \quad \text{and} \quad u^0(x) = \frac{2}{2 - \cos(x)}.$$

This smallness assumption on the potential attenuates the effect of the non diagonal (in Fourier variables) operator V: We thus expect for the exact solution a long time preservation of the smoothness of the initial data.

In Figure I.7, we plot the evolution of the actions $|\xi_a(t)|^2$ in logarithmic scale. We use step sizes:

$$\tau = \frac{2\pi}{6^2 - 2^2} \simeq 0.1963\ldots \text{ (top)} \quad \text{and} \quad \tau = 0.2 \quad \text{(bottom).} \qquad \text{(I.17)}$$

What we observe is that in the case of a resonant step size, the regularity of the initial solution is lost, while it is preserved for a non resonant step size. Note that the non resonant step size is very close to the resonant one. Later, we will explain that all step

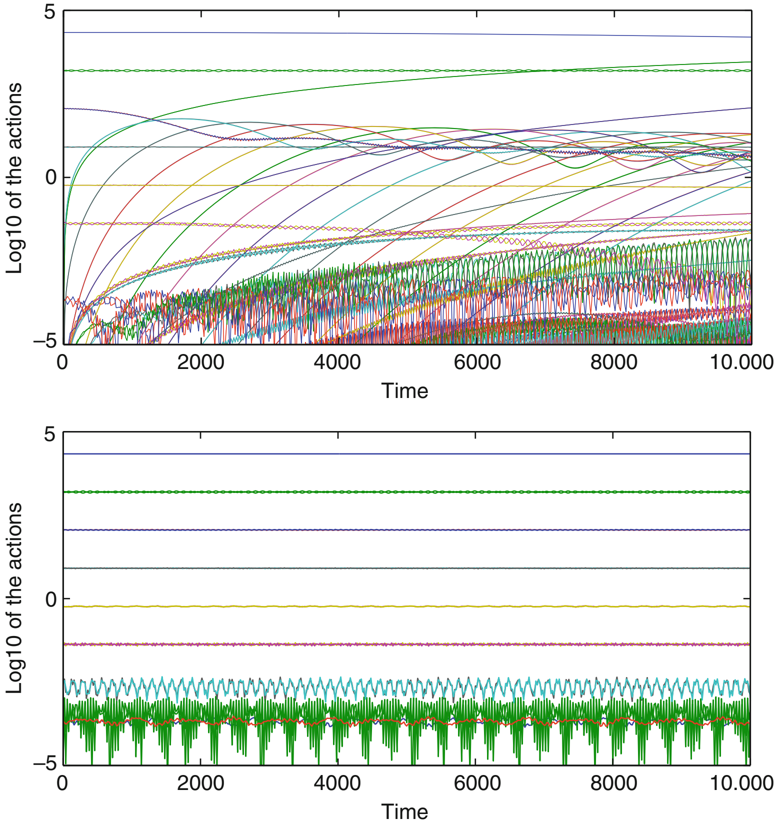


Figure I.7. Evolution of the actions (linear case) for a Lie splitting with resonant step size (top) and non resonant step size (bottom).

sizes of the form $2\pi/(a^2 - b^2)$ for two integers a and b are resonant. Moreover, when the time step is non resonant, we can actually show preservation of the regularity of the solution over a very long time, which in turn ensures preservation of energy even if the CFL number is large. We will however not prove this rigorously here, and refer to [13].

For explicit schemes with CFL condition, or implicit explicit integrators, such resonance effects do not appear. Let us explain this quickly: resonant step sizes can be shown to be such that there exist integers a, b with $a \neq \pm b$ and $\ell \neq 0$ such that

$$\tau(a^2 - b^2) \simeq 2\pi\ell.$$

We easily see that if a CFL condition is imposed with $\mathsf{cfl} < 2\pi$, then we will have $|\tau(a^2 - b^2)| \leq \mathsf{cfl} < 2\pi$ and the previous relation can never be satisfied. In the situation above, the CFL condition is large, so that resonant step sizes are indeed present. However the set of resonant step sizes can be proved to be very small, which explains the top figure I.6.

Now in the case of implicit-explicit integrators, the resonance condition reads (see (I.15))

$$2\arctan(\tau a^2/2) - 2\arctan(\tau b^2/2) \simeq 2\pi\ell$$

and as the arctan function is bounded by $\pi/2$, such a relation can never be satisfied for any step size τ! As we will see in Chapter V, this property ensures the existence of a *modified energy* associated with the implicit-explicit integrator, which is preserved along the numerical flow. This explains the regularity of the function $\tau \mapsto E(\tau)$ observed on the bottom in Figure I.6.

3.3 NLS in dimension 1: resonances and aliasing. We now consider the Schrödinger equation with a cubic nonlinearity and without potential (i.e. $V = 0$ in (I.4)). To measure the balanced effects between the linear and nonlinear parts, we introduce a scaling factor, and consider initial data to (I.4) that are *small*, i.e. of order δ where $\delta \to 0$ is a small parameter.

After a scaling of the solution, it is equivalent to study the family of nonlinear Schrödinger equations

$$i\partial_t u(t,x) = -\partial_{xx}u(t,x) + \varepsilon|u(t,x)|^2 u(t,x), \quad u(t,x) = u^0 \simeq 1 \tag{I.18}$$

where $\varepsilon = \delta^2 > 0$ is a small parameter, and $x \in \mathbb{T}$ the one-dimensional torus.

In dimension 1, this equation has the very nice property of being *integrable*, see [37], which implies in particular that it possesses an infinite number of invariants preserved throughout the exact solution. In particular, it can be shown that the actions $|\xi_a(t)|^2$ of $u(t,x)$ satisfy the preservation property

$$\forall\, a \in \mathbb{Z}, \quad \big||\xi_a(t)|^2 - |\xi_a(0)|^2\big| \le C\varepsilon, \tag{I.19}$$

for all time $t \ge 0$. In Chapter VII, and without considering the integrable nature of the equation, we will show this result for a long time of order $t \le \varepsilon^{-1}$ using a simple *averaging* argument.

A natural question in geometric integration theory is this: Does the discrete numerical approximation $\xi^{K,n}$ defined above satisfy the same preservation property? As we will see now, there are two sources of possible instabilities: one coming from the choice of the step size, and the other coming from the number K of grid points.

In a first simulation, we first consider the initial data

$$u^0(x) = \frac{1}{2-\cos(x)}$$

and take $\varepsilon = 0.01$ in (I.18), and $K = 512$ grid points.

We make two simulations with this initial data, and the same number of grid points: one with the step size $\tau = 0.09$, and the other with the step size

$$\tau = \frac{1}{12^2 - 5^2 - 7^2} \simeq 0.0898\dots. \tag{I.20}$$

In Figure I.8, we plot the evolution of the fully discrete actions $|\xi_a^{K,n}(t)|^2$ in logarithmic scale, as in the previous section. We observe that for $\tau = 0.09$, there is preservation of the actions over a long time, as expected from (I.19). But this preservation property is broken by the use of the resonant step size (I.20). As we will see in Chapter VI, such a step size impedes the existence of a *modified energy* preserved by the fully discrete solution. We will however show that if the CFL number (I.16) is sufficiently but reasonably small (of order $\simeq 1$), such a situation cannot occur, avoiding the possible use of a resonant step size (as in the linear case described above).

Let us now consider instabilities coming from the number of grid points K. In the next example, we perform a simulation with $\varepsilon = 0.01$, a step size $\tau = 10^{-3}$, and the

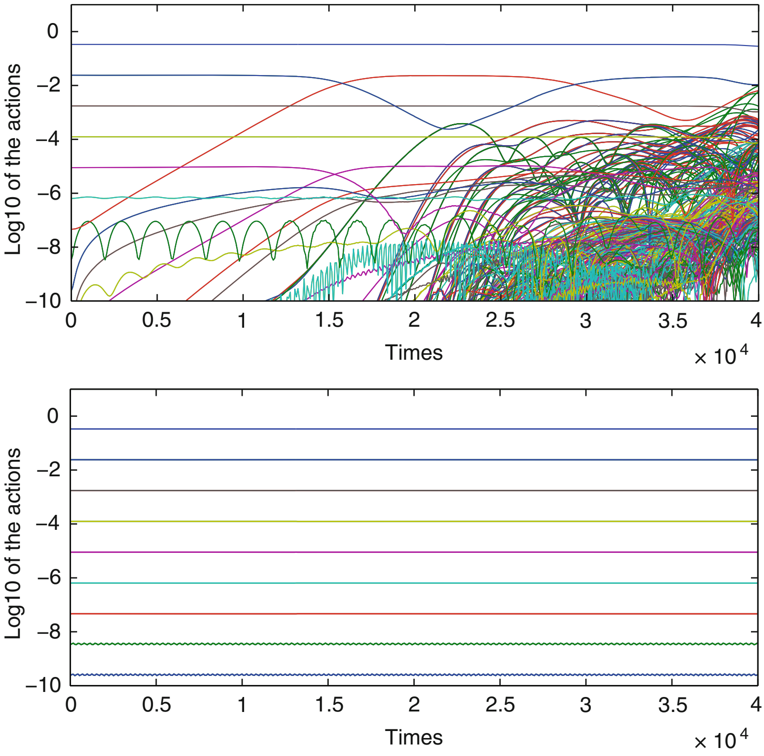


Figure I.8. Evolution of the actions in dimension 1 for resonant and non resonant step sizes.

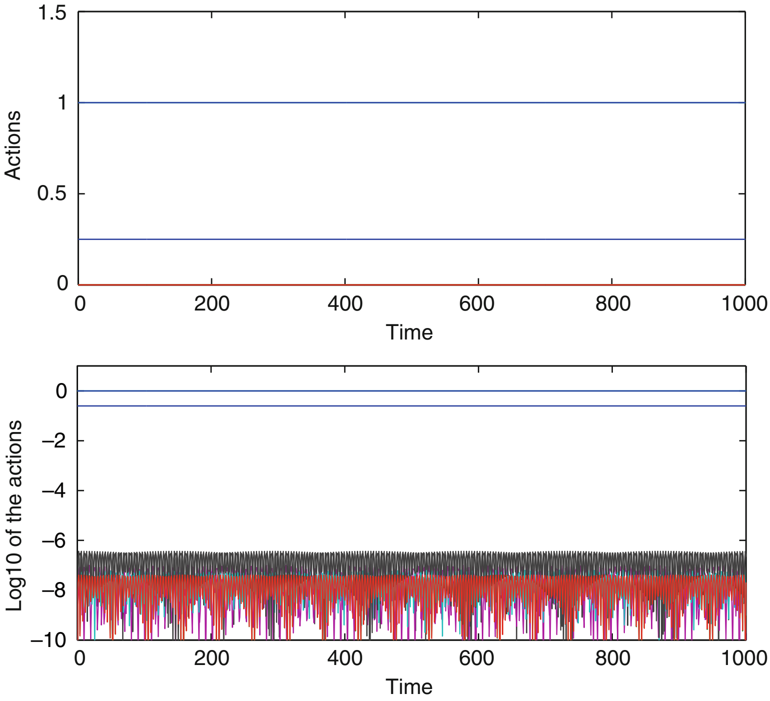


Figure I.9. Evolution of the actions in dimension 1 for $K = 31$ grid points.

initial value

$$u^0(x) = 2\sin(10x) - 0.5\, e^{i7x}.$$

Note that this initial value involves only the frequencies ±10 and 7. We make two simulations: one with $K = 31$ grid points, and the other one with $K = 34$ points. In Figure I.9, we plot the evolution of the actions $|\xi_a^{K,n}|^2$ both in standard and logarithmic scale for $K = 31$. We observe a very good preservation of the actions, as expected from (I.19). In Figure I.10, we use $K = 34$ and we observe exchanges between the actions. However, in this specific situation, a more careful analysis of the evolution of the actions show that there are only exchanges between symmetric frequencies, i.e., $|\xi_a(t)|^2$ and $|\xi_{-a}(t)|^2$ for $a \in B^K$, and the *super actions* $|\xi_a^{K,n}|^2 + |\xi_{-a}^{K,n}|^2$ are in fact preserved.

As we will see in Chapter VII, the persistence of (I.19) after space discretization holds only if K is a *prime number* (note that $K = 31$ is prime). In the situation where $K/2$ is a prime number, we can only show the long time preservation of the super actions defined above (this corresponds to Figure I.10 with $K = 34 = 2 \times 17$).

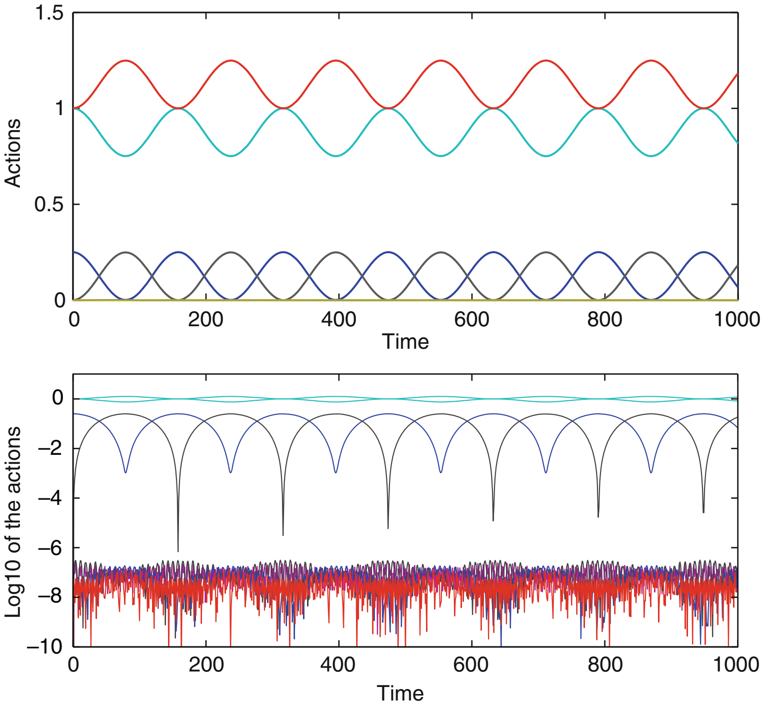


Figure I.10. Evolution of the actions in dimension 1 for $K = 34$ grid points.

In all other cases, nonlinear exchanges can always be observed. For example we perform another simulation with $\tau = 0.001$, $K = 30 = 2 \times 3 \times 5$, $\varepsilon = 0.05^2$, and

$$u^0(x) = 0.9\cos(-5x) + \sin(14x) + 1.1\exp(-10ix) + 1.2\cos(-11x). \qquad \text{(I.21)}$$

We plot the evolution of the actions in logarithmic norm in Figure I.11 both for $K = 30$ (top) and the prime number $K = 31$ (bottom). We observe that for $K = 30$ the dynamics of the actions is very complicated, while the preservation of the actions holds for $K = 31$ and the same step size and initial data.

In Chapter VII, we will show that the quadruplet of frequencies $(-5, 14, -10, -11)$ are non trivial frequencies belonging to the *numerical resonance modulus* associated with the modified energy of the numerical scheme. Note that in this situation, the step size is small enough to ensure the existence of the modified energy (the CFL number is of order 0.3), but the instability comes from the internal dynamics of this modified system and in particular the problem of aliasing.

3.4 Energy cascades in dimension 2. As a final example, we consider the same equation as before, but in dimension 2:

$$i\partial_t u = -\Delta u + \varepsilon |u|^2 u, \quad x \in \mathbb{T}^2, \qquad \text{(I.22)}$$

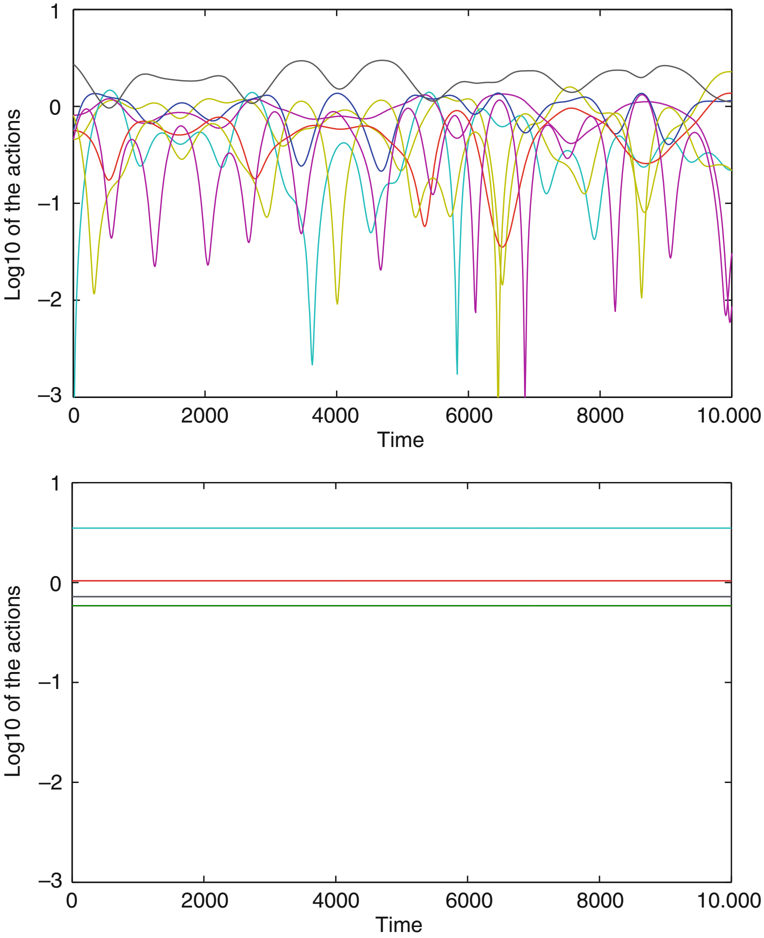


Figure I.11. Evolution of the actions for $K = 30$ (top) and $K = 31$ (bottom).

and we take as initial data

$$u(0, x) = 1 + 2\cos(x_1) + 2\cos(x_2). \tag{I.23}$$

As we will see in Chapter VII, the particular geometric configuration of the five modes associated with the initial data (I.23) makes possible energy exchanges between the Fourier modes of the exact solution. Following the methods used in [7] we will actually give some rigorous and explicit lower bounds for high modes, showing that some energy is actually transferred from low to high modes, in a time depending on the size of the high mode. Such a phenomenon is called an *energy cascade* and constitutes an interesting nonlinear test case for numerical schemes applied to (I.22).

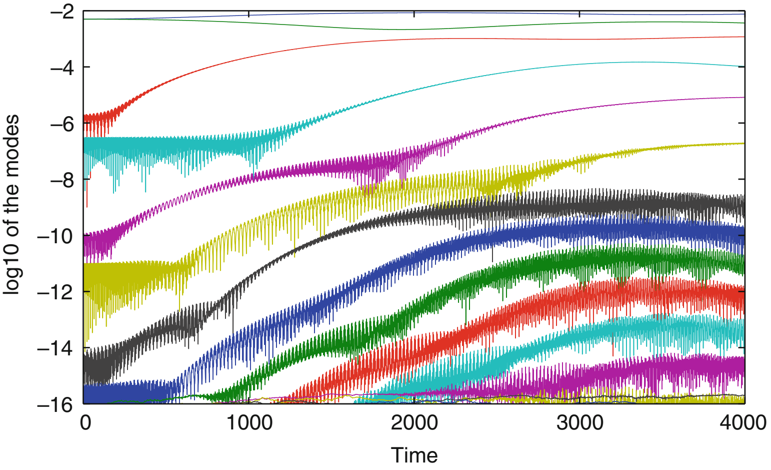


Figure I.12. Energy cascade.

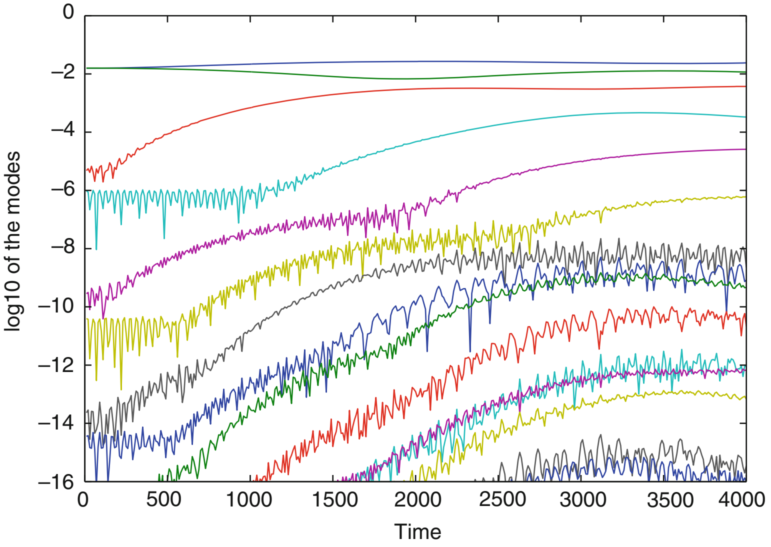


Figure I.13. Explicit scheme, $\tau = 0.1$, grid 128×128.

Such a phenomenon is linked with analysis of the (nonlinear) resonance relation $|a|^2 + |b|^2 - |c|^2 - |d|^2 = 0$ appearing for some quadruplet $(a, b, c, d) \in \mathbb{T}^2$ satisfying $a + b - c - d = 0$. Actually we will prove that such a relation is satisfied when (a, b, c, d) forms an affine rectangle in $\mathbb{Z}^2$, allowing energy exchanges between modes in such a configuration.

The reproduction of these energy exchanges by numerical simulation is not guaranteed in general. We give in Figure I.12 a numerical example with $\varepsilon = 0.0158$.

This simulation is made using an explicit splitting scheme with step size $\tau = 0.001$ and a 128×128 grid. We plot the evolution of the logarithms of the Fourier modes $\log|\xi_a(t)|$ for $a = (0, n)$, with $n = 0, \dots, 15$. We observe the energy exchanges between the modes.

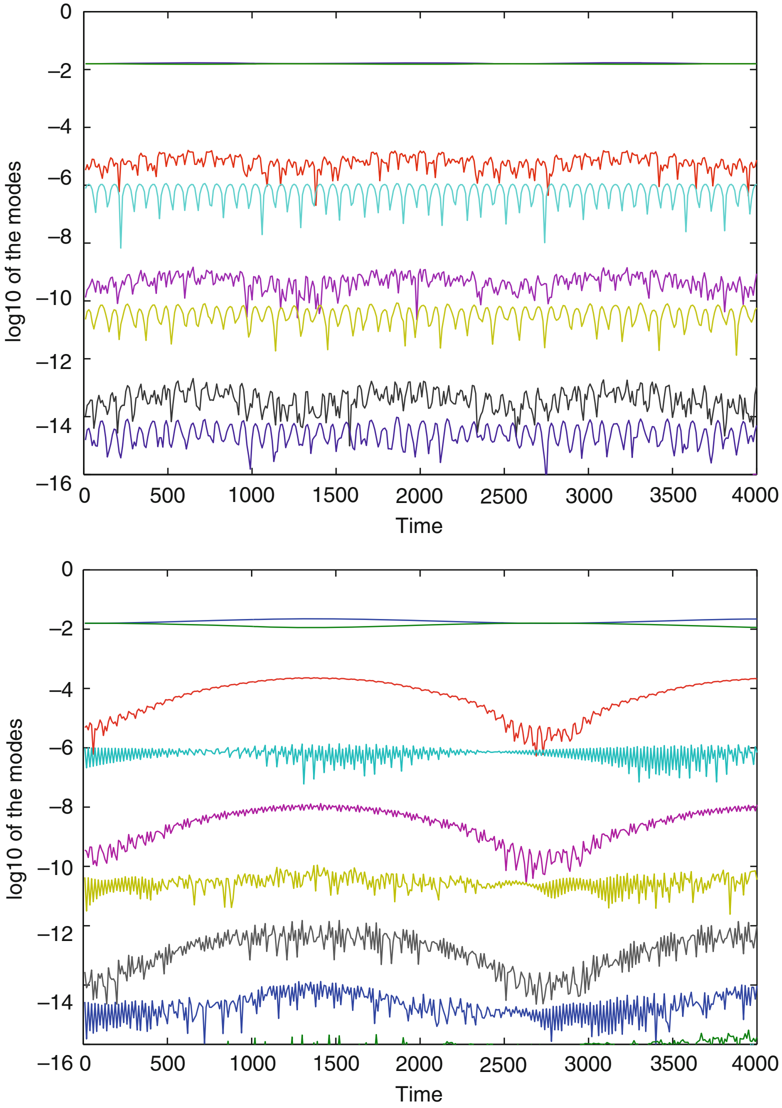


Figure I.14. Implicit-explicit integrator, $\tau = 0.1$ and $\tau = 0.05$.

Repeating the same experiment but with $\tau = 0.1$ and the explicit splitting scheme defined above, we observe that the energy exchanges are correctly reproduced (see Figure I.13).

Now we do the same simulation, but with the implicit-explicit integrator defined above, where the linear part is integrated using the midpoint rule. We observe in Figure I.14 that the energy exchanges are not correctly reproduced, even for a smaller time step $\tau = 0.05$.

The reason is again that the frequencies of the underlying operator associated with the implicit-explicit splitting scheme are slightly changed (see (I.12)), making the resonance relations $|a|^2 + |b|^2 - |c|^2 - |d|^2 = 0$, appearing for some a, b, c and d in $\mathbb{Z}^d$, destroyed by the numerical scheme. As these relations determine the energy transfers, the implicit-explicit cannot reproduce the energy cascade unless a very small time step is used.

4 Objectives

The main goal of this work is to give precise mathematical formulations of the numerical phenomena observed in the previous sections. In particular we will prove the existence of a modified energy for splitting schemes applied to very general linear and nonlinear situations, under some restrictions on the CFL number used. Using this modified energy, we will be able to make a resonance analysis in some specific situations.

We will first analyze in detail the finite dimensional situation. In this case, the results given by *backward error analysis* show that the numerical solution obtained by a symplectic integrator applied to a Hamiltonian system (almost) coincides with the exact solution of a modified Hamiltonian system, over an extremely long time. As we will only consider splitting methods, we will prove this result in Chapter II in this specific framework. This will be the occasion to introduce several tools that will be used later in the infinite dimensional case, such as the Baker–Campbell–Hausdorff formula and some Hamiltonian formalism.

We will then focus on Hamiltonian PDEs, first by defining symplectic flows in infinite dimension (Chapter III) and by considering semi-discrete flows after space discretization. We will also recall some global existence results for the nonlinear Schrödinger equation with defocusing nonlinearity, or for small initial data.

In Chapter IV, we will consider the approximation properties of splitting methods over finite time. This will lead us to state and prove convergence results in the case of semi-discrete and fully discrete numerical flows. In other words, we prove (I.10) for approximations of smooth solutions over finite time.

In Chapter V and VI, we will then give some backward error analysis results in the case of linear and cubic nonlinear Schrödinger equations. More precisely, we will show that under some CFL condition, the numerical methods almost coincides at each time step with the exact solution of a modified Hamiltonian PDE of the form (I.11). We show that there exists a modified Hamiltonian H_τ such that the following holds:

$$\left\| \varphi_P^\tau \circ \varphi_T^\tau(u) - \varphi_{H_\tau}^\tau(u) \right\|_{\ell^1} \leq C_N \tau^{N+1}, \tag{I.24}$$

where the error is estimated in the Wiener algebra ℓ^1, and where C_N depends on the size of the function u in ℓ^1. This result is valid for the explicit Lie splitting, as well as for the implicit-explicit splitting scheme, and can be also derived for fully discrete algorithms. The exponent N in the small error term $\mathcal{O}(\tau^{N+1})$ made at each step depends in general on the CFL condition.

It is important to note that the error in (I.24) is measured in the same Banach space used to bound the solution *a priori*. Using this result and a bootstrap argument, we prove the almost global existence of the numerical solution in H^1 for small fully discrete initial data of the nonlinear Schrödinger equation in one dimension of space. This is due to the fact that in dimension 1, the ℓ^1 norm in estimate (I.24) can be replaced by the Sobolev norm H^1, and that the modified Hamiltonian H_τ controls the H^1 norm of (small) fully discrete solutions.

With this modified energy H_τ in hand, we will then give in Chapter VII an introduction to long time analysis, and compare the one and two-dimensional cases. We will analyze the resonances of the nonlinear equation, their consequences on the long time behavior of the solution (preservation of the actions, energy cascade), and discuss the persistence of these qualitative properties in numerical discretizations.

II Finite dimensional backward error analysis

In this chapter, we consider Hamiltonian ordinary differential equations and show that splitting schemes can be expressed as exact flows of *modified* Hamiltonian systems, up to very small errors. We refer to [26], [34] for more general and extensive results in the same direction.

The proof relies on the Baker–Campbell–Hausdorff formula involving commutators of the two Hamiltonian vector fields associated with the splitting method. Such a tool will be fundamental for later applications to Hamiltonian PDEs.

In the following presentation, we give a *polynomial* version of backward error analysis: this means that the small error made above is of order $C_N \tau^N$ where τ is the small time discretization step size. Here, N is arbitrary, but the constant C_N depends on N (see for instance [24]). In other words, we consider the result in the sense of asymptotic expansions in powers of τ. In the case where the Hamiltonian function is *analytic*, a careful estimation of the constant C_N can be performed and the small error term can be optimized to obtain an exponentially small term of the form $\exp(-1/(c\tau))$ for some positive constant c, where the optimal N is taken of order $1/\tau$. We refer to [5], [25], [26, Chapter IX] and [33] for the highly technical proof of these exponential estimates.

1 Hamiltonian ODEs

1.1 Definitions and basic properties. We consider Hamiltonian ordinary differential systems of the form

$$\dot{y} = X_H(y) := J^{-1}\nabla H(y), \tag{II.1}$$

where $y = (p,q) \in \mathbb{R}^{2d}$, $H : \mathbb{R}^{2d} \to \mathbb{R}$ is the Hamiltonian of the system, and where

$$J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix} \tag{II.2}$$

is the canonical symplectic matrix satisfying $J^T = -J = J^{-1}$. The operator ∇ represents the derivative with respect to $y = (p_1,\dots,p_d,,q_1,\dots,q_d)$, and thus the system (II.1) can also be written

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}(p,q), \quad \dot{q}_i = \frac{\partial H}{\partial p_i}(p,q), \quad i = 1,\dots,d.$$

By analogy with the PDE case, we will assume that the Hamiltonian H can be split into $H = T + P$, and we will only consider the long time behavior of splitting methods built from this decomposition. Similar studies of more a general class of schemes can be found in the classical references [26], [34].

Many physical problems in classical mechanics possess a Hamiltonian structure. In particular, this is the case for all systems satisfying Newton's equations

$$m_i \ddot{q}_i = -\frac{\partial U}{\partial q_i}(q), \quad i = 1, \dots, d,$$

where $U(q)$ is the potential function and m_i the mass of the particle associated with the coordinates q_i. Such a system can be written

$$\dot{q}_i = \frac{p_i}{m_i}, \quad \dot{p}_i = -\frac{\partial U}{\partial q_i}(q),$$

which is a Hamiltonian system associated with the energy

$$H(p,q) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i} + U(q) =: T(p) + P(q), \tag{II.3}$$

where we denote by T the kinetic energy of the system and P the potential energy. Note that in this situation, as we will see below, we can calculate the exact solutions of the Hamiltonian system associated with the Hamiltonian $T(p)$ and $P(q)$ respectively, which makes splitting methods very easy and cheap to implement.

We now give some basic properties of Hamiltonian systems of the form (II.1). In the following, we denote by φ_H^t the flow of the system (II.1), i.e. the mapping $\varphi_H^t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ such that $\varphi_H^0(y) = y$ and for all t,

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_H(y) = X_H\left(\varphi_H^t(y)\right) = J^{-1}\nabla H\left(\varphi_H^t(y)\right). \tag{II.4}$$

We will always assume that φ_H^t is defined for all $t > 0$, all $y \in \mathbb{R}^{2d}$, and is smooth. This is guaranteed for example when H is smooth and when the solution remains bounded.

With the matrix J defined in (II.2) is associated the *symplectic form* $\omega : \mathbb{R}^{2d} \times \mathbb{R}^{2d} \to \mathbb{R}$ such that

$$\omega(\xi, \eta) = \xi^T J \eta.$$

Definition II.1. *A matrix A of size $2d$ is symplectic if it satisfies*

$$A^T J A = J.$$

A differentiable nonlinear mapping $\varphi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is said to be symplectic if, for all $y \in \mathbb{R}^{2d}$, the Jacobian matrix $\partial_y \varphi(y)$ is symplectic.

An easy consequence of the chain rule yields the following:

Proposition II.2. *Assume that φ and ψ are smooth symplectic mappings from $\mathbb{R}^{2d}$ to itself, then $\varphi \circ \psi$ is again symplectic.*

One of the major properties of the flow of a Hamiltonian system is given by the following result:

Proposition II.3. *Let φ_H^t be the flow associated with the Hamiltonian system (II.1). Then for all t, the mapping $y \mapsto \varphi_H^t(y)$ is symplectic and preserves energy in the sense that for all $t > 0$ and $y \in \mathbb{R}^{2d}$, we have*

$$H\left(\varphi_H^t(y)\right) = H(y). \tag{II.5}$$

Proof. Taking the derivative of (II.4), we see that

$$\frac{\mathrm{d}}{\mathrm{d}t}\partial_y \varphi_H^t(y) = J^{-1}\nabla^2 H\left(\varphi_H^t(y)\right)\cdot \partial_y \varphi_H^t(y)$$

where $\nabla^2 H$ is the Hessian (symmetric) matrix of H. This shows that (forgetting the dependence in y)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\left(\partial_y\varphi_H^t\right)^T J\,\partial_y\varphi_H^t\right) &= \left(\partial_y\varphi_H^t\right)^T\left(\nabla^2 H\left(\varphi_H^t\right)^T J^{-T}J + J^{-1}J\nabla^2 H\left(\varphi_H^t\right)\right)\partial_y\varphi_H^t\\ &= \left(\partial_y\varphi_H^t\right)^T\left(-\nabla^2 H\left(\varphi_H^t\right)^T + \nabla^2 H\left(\varphi_H^t\right)\right)\partial_y\varphi_H^t = 0\end{aligned}$$

as we have that $J^{-T}J = -I$, and because the Hessian matrix is symmetric. As for $t = 0$, $\varphi_H^0(y) = y$, we conclude that for all time t, we have

$$\left(\partial_y\varphi_H^t\right)^T J\,\partial_y\varphi_H^t = J \tag{II.6}$$

which means that the flow is symplectic.
The energy preservation (II.5) is an easy consequence of the fact that the matrix J is skew-symmetric. ■

Note that a consequence of the symplecticity of the flow is volume preservation: taking the determinant of (II.6) yields

$$\forall t > 0, \quad \forall y \in \mathbb{R}^{2d}, \quad |\det \partial_y \varphi_H^t(y)| = 1$$

which means that for all integrable functions $f : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$, we have

$$\forall t > 0 \quad \int_{\mathbb{R}^{2d}} f\left(\varphi_H^t(y)\right)\mathrm{d}y = \int_{\mathbb{R}^{2d}} f(y)\,\mathrm{d}y.$$

1.2 Expansion of the flow. Let $H, K : \mathbb{R}^{2d} \to \mathbb{R}$ be two smooth functions. We define the *Poisson bracket* of H and K by the formula

$$\{H, K\} = \nabla H^T J \nabla K = \sum_{i=1}^{d} \left(\frac{\partial H}{\partial p_i} \frac{\partial K}{\partial q_i} - \frac{\partial H}{\partial q_i} \frac{\partial K}{\partial p_i} \right). \tag{II.7}$$

We easily see that we have $\{H, K\} = -\{K, H\}$ and for three functions H, K and G, the Jacobi identity

$$\{H, \{G, K\}\} + \{G, \{K, H\}\} + \{K, \{H, G\}\} = 0. \tag{II.8}$$

Now let N be an integer, and $G : \mathbb{R}^{2d} \to \mathbb{R}^N$ be a smooth function with components G_j for $j = 1, \ldots, N$. We will mainly consider the cases $N = 1$ (Hamiltonian functions) or $N = 2d$ (vector fields).

We define the Lie derivative $\mathcal{L}_H$ associated with H by the following formula: for any function $G : \mathbb{R}^{2d} \to \mathbb{R}^N$, $\mathcal{L}_H[G]$ is a function from $\mathbb{R}^{2d}$ to $\mathbb{R}^N$ with components

$$(\mathcal{L}_H[G])_j = \nabla H^T J^{-T} \nabla G_j = \{H, G_j\}.$$

This Lie operator expresses the derivative in the direction given by X_H in the sense that we have

$$\frac{\mathrm{d}}{\mathrm{d}t} G\left(\varphi_H^t\right) = \mathcal{L}_H[G]\left(\varphi_H^t\right). \tag{II.9}$$

Note that taking $G = \mathrm{Id} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ the identity function yields

$$\mathcal{L}_H[\mathrm{Id}] = X_H.$$

Moreover, for two functions H and K, we calculate using (II.8) that we have

$$[\mathcal{L}_H, \mathcal{L}_K] = \mathcal{L}_H \circ \mathcal{L}_K - \mathcal{L}_K \circ \mathcal{L}_H = \mathcal{L}_{\{H,K\}}. \tag{II.10}$$

Now let us consider the flow φ_H^t. Formally for a fixed $y \in \mathbb{R}^{2d}$, we can write down the Taylor expansion around $t = 0$,

$$\varphi_H^t(y) = \sum_{k \geq 0} \frac{t^k}{k!} \left. \frac{\mathrm{d}^k \varphi_H^t(y)}{\mathrm{d}t^k} \right|_{t=0}.$$

But the successive derivatives of the flow with respect to t are easily expressed in terms of the Lie derivative: we have $\varphi_H^0 = \mathrm{Id} = \mathcal{L}_H^0[\mathrm{Id}]$,

$$\frac{\mathrm{d}\varphi_H^t}{\mathrm{d}t} = X_H\left(\varphi_H^t\right) = \mathcal{L}_H[\mathrm{Id}]\left(\varphi_H^t\right),$$

and by induction for all $k \geq 1$,

$$\frac{\mathrm{d}^k \varphi_H^t}{\mathrm{d}t^k} = \mathcal{L}_H^k[\mathrm{Id}]\left(\varphi_H^t\right).$$

Hence we can write at least formally

$$\varphi_H^t = \sum_{k \geq 0} \frac{t^k}{k!} \mathcal{L}_H^k[\mathrm{Id}] = \exp(t\mathcal{L}_H)[\mathrm{Id}].$$

Example II.4. In the case of a quadratic Hamiltonian of the form $H(y) = y^T Ay$ where $A^T = A$ is a symmetric matrix, then we have for all $k \geq 0$ that $\mathcal{L}_H^k[\mathrm{Id}] = (J^{-1}A)^k$. In this case, we have $\varphi_H^t(y) = \exp(tJ^{-1}A)y$ where the exp function is the classical matrix exponential defined as a convergent series.

More generally, the convergence of this series holds for an analytic function H. Here we will consider that the previous equality holds in the sense of asymptotic expansions:

Proposition II.5. *Assume that H is $\mathcal{C}^\infty$ over $\mathbb{R}^{2d}$ and let M be fixed. Then there exists t_0 such that for all N, there exists a constant C_N satisfying the following: for all y with $\|y\| \leq M$ and all $t \leq t_0$, we have*

$$\left\| \varphi_H^t(y) - \sum_{k \geq 0}^{N} \frac{t^k}{k!} \mathcal{L}_H^k[\mathrm{Id}](y) \right\| \leq C_N t^{N+1}.$$

Proof. As H is smooth, it is clear that for all y, the mapping $t \mapsto \varphi_H^t(y)$ is smooth, and we have by using a Taylor expansion

$$\varphi_H^t(y) - \sum_{k \geq 0}^{N} \frac{t^k}{k!} \mathcal{L}_H^k[\mathrm{Id}](y) = \int_0^t \frac{(t-s)^N}{N!} \mathcal{L}_H^{N+1}[\mathrm{Id}]\left(\varphi_H^s(y)\right) \mathrm{d}s.$$

Now by standard arguments there exists t_0 such that, for all y with $\|y\| \leq M$, and for all $s \leq t_0$, we have $\left\|\varphi_H^s(y)\right\| \leq 2M$. As $\mathcal{L}_H^{N+1}$ is made of multiple derivatives of H, we see that $(s, y) \mapsto \mathcal{L}_H^{N+1}[\mathrm{Id}](\varphi_H^s(y))$ is bounded for $s \in (0, t_0)$ and $\|y\| \leq M$. This shows the result. ■

Remark II.6. The norm $\|\cdot\|$ considered above can be any norm on $\mathbb{R}^{2d}$. As all the norms are equivalent in finite dimension, a particular choice does not affect the final result. The situation is of course totally different in infinite dimension, where the norms will be defined on infinite dimensional functional spaces.

2 Numerical integrators

We give here a rough but convenient definition of a numerical integrator and its corresponding order of approximation. A numerical flow is defined as a mapping

$\Phi^\tau : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ such that for small τ, we have $\Phi^\tau \simeq \varphi_H^\tau$ up to a small error $\mathcal{O}(\tau^{p+1})$ where p is the order of the integrator. More precisely:

Definition II.7. *Assume that $H \in \mathcal{C}^\infty(\mathbb{R}^{2d}, \mathbb{R})$. A numerical integrator Φ^τ of order $p \in \mathbb{N}$ associated with the Hamiltonian system (II.1) is a mapping from $\mathbb{R}^{2d}$ to itself such that for all $M > 0$, there exist constants $\tau_0, L, C > 0$ such that for all $\tau \le \tau_0$, $x, y \in \mathbb{R}^{2d}$ with $\|y\| \le M$ and $\|x\| \le M$ we have*

$$\left\|\varphi_H^\tau(x) - \Phi^\tau(y)\right\| \le (1 + L\tau)\,\|x - y\| + C\tau^{p+1}. \tag{II.11}$$

The numerical integrator Φ^τ is said to be symplectic if for all τ, the mapping $\Phi^\tau : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is symplectic.

According to Proposition II.5, setting $x = y$ yields that the asymptotic expansion around $\tau = 0$ of an integrator $\Phi^\tau(x)$ of given order p should coincide with the Taylor expansion of the exact flow $\varphi_H^\tau(x)$ up to the order p included. Now if this is the case, we have for $\|y\| \le M$,

$$\varphi_H^\tau(y) - \Phi^\tau(x) = \varphi_H^\tau(y) - \varphi_H^\tau(x) + \varphi_H^\tau(x) - \Phi^\tau(x),$$

which yields (II.11) where L is the Lipschitz constant of X_H over the compact set $\{y \mid \|y\| \le 2M\}$.

Let us now consider splitting methods applied to a Hamiltonian system with $H = T + P$, and assume that we can compute the exact solution of the Hamiltonian systems associated with the Hamiltonian functions T and P. Using Proposition II.5, we can write in the sense of asymptotic expansions that

$$\varphi_H^\tau = \mathrm{Id} + \tau X_H + \frac{\tau^2}{2}\mathcal{L}_H^2[\mathrm{Id}] + \mathcal{O}(\tau^3).$$

But this implies

$$\begin{aligned}\varphi_T^\tau \circ \varphi_P^\tau &= \mathrm{Id} + \tau X_T + \tau X_P + \mathcal{O}(\tau^2)\\ &= \mathrm{Id} + \tau X_H + \mathcal{O}(\tau^2),\end{aligned}$$

and we easily conclude that the Lie splitting method

$$\Phi_H^\tau := \varphi_T^\tau \circ \varphi_P^\tau$$

is a numerical method of order 1. By similar calculations we can show that

$$\Phi_H^\tau := \varphi_P^{\tau/2} \circ \varphi_T^\tau \circ \varphi_P^{\tau/2}$$

is a numerical method of order 2.

The symplecticity of a method is not straightforward, and we refer to [26] for extensive analysis of the conditions on the coefficients of general numerical methods to construct symplectic integrators. For example we can show that the midpoint rule defined as the mapping $y^{n+1} = \Phi^\tau(y^n)$ such that

$$y^{n+1} = y^n + \tau X_H \left(\frac{y^n + y^{n+1}}{2} \right)$$

is a symplectic mapping. Note that the mapping $y^{n+1} = \Phi^\tau(y^n)$ is well defined on bounded domains if τ is small enough.

In the following, we will only consider splitting methods applied to a Hamiltonian functions that can be decomposed into $H = T + P$. In the case where the flows φ_T^τ and φ_P^τ can be calculated explicitly, then the splitting schemes will always be symplectic as compositions of symplectic maps. If this is not the case, we can apply a symplectic method to each part (for example the midpoint rule) and hence still obtain symplectic maps. For this reason the implicit-explicit integrators discussed in Chapter I are symplectic numerical integrators.

With this definition of *local* order, we obtain the following classical *global* result:

Proposition II.8. *Let Φ^τ be a numerical integrator of order p applied to the smooth Hamiltonian system (II.1). Let $y^0 \in \mathbb{R}^{2d}$, and let $t_* > 0$. Assume that for all $t \in (0, t_*)$, $\varphi_H^t(y^0)$ is well defined and let $B = \sup\{ \|\varphi_H^t(y^0)\| \mid t \in (0, t_*)\}$. Then there exist constants τ_0 and C such that, if $\tau \le \tau_0$ and if y^n is the sequence defined by induction by setting $y^{n+1} = \Phi^\tau(y^n)$, $n \ge 0$, then we have*

$$\forall t = n\tau \le t_*, \quad \left\|\varphi_H^t\left(y^0\right) - y^n\right\| \le C\tau^p.$$

Proof. We apply (II.11) with $M = 2B$. We set $y(t) = \varphi_H^t(y^0)$ and $t_n = n\tau$. Of course we have $\|y^0\| \le M$. Now for $n \ge 1$, assume that $\|y^{n-1}\| \le M$. Then we can write using $(1 + L\tau) \le e^{L\tau}$:

$$\|y(t_n) - y^n\| = \left\|\varphi_H^\tau\left(y\left(t_{n-1}\right)\right) - \Phi^\tau\left(y^{n-1}\right)\right\| \le C\tau^{p+1} + e^{Lt}\left\|y\left(t_{n-1}\right) - y^{n-1}\right\|.$$

Using the fact that $y(t_0) = y^0$, then as long as $\|y^n\| \le M$ and $n\tau \le t_*$, we have

$$\|y(t_n) - y^n\| \le Cne^{Ln\tau}\tau^{p+1} \le \left(Ct_*e^{Lt_*}\right)\tau^p.$$

Now this relation implies that if τ_0 satisfies $(Ct_*e^{Lt_*})\tau_0^p \le B$, we have $\|y^n\| \le 2B = M$ for $n\tau \le t_*$. This finishes the proof. ■

3 Backward analysis for splitting methods

3.1 Setting of the problem. We now assume that $H = T + P$ can be split into two parts for which we are able to compute the corresponding exact flows. In the case where the Hamiltonian can be split into $H(p, q) = T(p) + P(q)$ as in (II.3), this is actually the case: the solution of the Hamiltonian system $y(t) = \varphi_T^t(y^0)$ associated with T starting in $y^0 = (p^0, q^0)$ satisfies $y(t) = (p(t), q(t))$ with $p(t) = p^0$ and $q(t) = q^0 + t\partial_p T(p^0)$ while the solution associated with P is given by $q(t) = q^0$ and $p(t) = p^0 - t\partial_q P(q^0)$.

Let us for example consider the Lie splitting method $\Phi^\tau := \varphi_P^\tau \circ \varphi_T^\tau$. The backward error analysis problem consists in searching for a Hamiltonian $Z(\tau)$ such that

$$\varphi_{Z(\tau)}^1 = \varphi_P^\tau \circ \varphi_T^\tau \tag{II.12}$$

at least in the sense of asymptotic expansion in powers of τ. If such a function $Z(\tau)$ can be constructed, the numerical trajectory can then be interpreted as the exact solution of the modified Hamiltonian $\frac{1}{\tau}Z(\tau)$ evaluated at the discrete times $n\tau$: we have for all n,

$$\left(\varphi_P^\tau \circ \varphi_P^\tau\right)^n = \varphi_{Z(\tau)}^n = \varphi_{Z(\tau)/\tau}^{n\tau}.$$

In particular such a relation implies that $Z(\tau)$ is a conserved quantity along the numerical trajectory. Of course, as (II.12) holds in the sense of asymptotic expansions, a small error is made at each step, but this error is of order $C_N\tau^N$ for any given N as long as the numerical trajectory remains bounded, and hence the conservation of the *modified energy* $Z(\tau)$ holds over a very long time.

The equation (II.12) can be written (note the reverse order of T and P)

$$\exp\left(\mathcal{L}_{Z(\tau)}\right) = \exp\left(\tau\mathcal{L}_T\right) \circ \exp\left(\tau\mathcal{L}_P\right) \tag{II.13}$$

and we see that the construction of $Z(\tau)$ relies on the Baker–Campbell–Hausdorff formula (see [2], [28]). As such formal calculations are central in our analysis of splitting methods in finite and infinite dimension, we give here a complete proof of this formula.

3.2 Baker–Campbell–Hausdorff formula. We consider here only the case of matrices, but in essence the result holds in the sense of formal series. This formula thus turns out to be valid for more general linear operators.

Lemma II.9. *Let A and Z be two square matrices, then we have*

$$\exp(Z)A\exp(-Z) = \exp(\mathrm{ad}_Z)A = \sum_{k\geq 0}\frac{1}{k!}\mathrm{ad}_Z^k A \tag{II.14}$$

where for two square matrices A and B we have

$$\mathrm{ad}_A B = [A, B] = AB - BA.$$

Proof. Let us consider the mapping $s \mapsto U(s) = \exp(sZ)A\exp(-sZ)$. We easily see by induction that for all $k \geq 1$,

$$\frac{\mathrm{d}^k U}{\mathrm{d}s^k}(s) = \exp(sZ)\mathrm{ad}_Z^k A\exp(-sZ)$$

and therefore (II.14) corresponds to the Taylor series of $U(s)$ around $s = 0$ evaluated at $s = 1$. The convergence of the series is clear considering the relation

$$\|\mathrm{ad}_A B\| \leq 2\,\|A\|\,\|B\|$$

for any subordinate matrix norm. ■

The second result gives an expression of the derivative of the exponential:

Lemma II.10. *Let $t \mapsto Z(t)$ be a differentiable mapping from $\mathbb{R}$ to the space of square matrices. Then we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\exp(Z(t)) = \left[\left(\frac{\exp\left(\mathrm{ad}_{Z(t)}\right) - 1}{\mathrm{ad}_{Z(t)}}\right)\frac{\mathrm{d}Z(t)}{\mathrm{d}t}\right]\exp\left(Z(t)\right), \tag{II.15}$$

where

$$\frac{\exp(\mathrm{ad}_Z) - 1}{\mathrm{ad}_Z} := \sum_{k\geq 0}\frac{1}{(k+1)!}\mathrm{ad}_Z^k.$$

Proof. Let us consider

$$U(s,t) = \left(\frac{\mathrm{d}}{\mathrm{d}t}\exp\left(sZ(t)\right)\right)\exp\left(-sZ(t)\right).$$

We calculate directly that

$$\begin{aligned}\frac{\partial}{\partial s}U(s,t) &= \left(\frac{\mathrm{d}}{\mathrm{d}t}Z(t)\exp\left(sZ(t)\right)\right)\exp\left(-sZ(t)\right)\\ &\quad - \left(\frac{\mathrm{d}}{\mathrm{d}t}\exp\left(sZ(t)\right)\right)Z(t)\exp\left(-sZ(t)\right)\\ &= \frac{\mathrm{d}Z(t)}{\mathrm{d}t} + [Z(t), U(s,t)] = \frac{\mathrm{d}Z(t)}{\mathrm{d}t} + \mathrm{ad}_{Z(t)}U(s,t).\end{aligned}$$

Using the Duhamel formula, this shows that

$$U(s,t) = \exp\left(s\,\mathrm{ad}_{Z(t)}\right)U(0,t) + \int_0^s \exp\left((s-\sigma)\mathrm{ad}_{Z(t)}\right)Z'(t)\mathrm{d}\sigma.$$

As $U(0,t) = 0$, we have (after doing a change of variable $\sigma \mapsto 1-\sigma$ in the formula)

$$U(1,t) = \int_0^1 \exp\left(\sigma\,\mathrm{ad}_{Z(t)}\right)Z'(t)\mathrm{d}\sigma,$$

which gives the result. ■

The final lemma gives the inverse formula for the derivative of the exponential:

Lemma II.11. *Let B_k be the Bernoulli numbers defined by the formula*

$$\frac{z}{e^z - 1} = \sum_{k \geq 0} \frac{B_k}{k!} z^k \tag{II.16}$$

for any complex number z such that $|z| < 2\pi$. Let Z be a square matrix of norm $\|Z\| < 2\pi$. Then we have

$$\left(\frac{\exp(\mathrm{ad}_Z) - 1}{\mathrm{ad}_Z} \right)^{-1} = \sum_{k \geq 0} \frac{B_k}{k!} \mathrm{ad}_Z^k. \tag{II.17}$$

The proof of this lemma is clear. With these preparations, we can prove the Baker–Campbell–Hausdorff (BCH) formula:

Theorem II.12. *Let A and B two square matrices. Then there exists t_0 sufficiently small and a smooth mapping $t \mapsto Z(t)$ for $|t| \leq t_0$, and such that for all $t \in (0, t_0)$ we have*

$$\exp(Z(t)) = \exp(tA) \exp(tB). \tag{II.18}$$

Moreover, $Z(t)$ satisfies the differential equation

$$Z'(t) = A + B + [Z(t), B] + \sum_{k \geq 1} \frac{B_k}{k!} \mathrm{ad}_{Z(t)}^k (A + B), \quad Z(0) = 0. \tag{II.19}$$

Proof. Taking the derivative of (II.18) with respect to t yields, using the previous Lemmas

$$\left[\left(\frac{\exp\left(\mathrm{ad}_{Z(t)}\right) - 1}{\mathrm{ad}_{Z(t)}} \right) Z'(t) \right] \exp(Z(t)) = A \exp(tA) \exp(tB) + \exp(tA) \exp(tB) B,$$

whence

$$\begin{aligned} \left(\frac{\exp\left(\mathrm{ad}_{Z(t)}\right) - 1}{\mathrm{ad}_{Z(t)}} \right) Z'(t) &= A + \exp(Z(t)) B \exp(-Z(t)) \\ &= A + \exp(\mathrm{ad}_Z(t)) B. \end{aligned}$$

Now we have

$$\begin{aligned} \exp(\mathrm{ad}_Z) B &= B + \sum_{k \geq 0} \frac{1}{(k+1)!} \mathrm{ad}_Z^k (\mathrm{ad}_Z B) \\ &= B + \frac{\exp(\mathrm{ad}_Z) - 1}{\mathrm{ad}_Z} (\mathrm{ad}_Z B). \end{aligned}$$

Hence we have

$$\left[\left(\frac{\exp\left(\mathrm{ad}_{Z(t)}\right)-1}{\mathrm{ad}_{Z(t)}}\right)\left(Z'(t)-\mathrm{ad}_Z B\right)\right]=A+B$$

which yields (II.19). The existence of Z then follows from standard ODE arguments. Note that the convergence of the series in the right-hand side is guaranteed as long as $\|Z(t)\|<2\pi$ which, as $Z(0)=0$, holds for small $t>0$. ■

3.3 Recursive equations. Let us now go back to equation (II.13) and the decomposition $H=T+P$. By applying the previous calculations, we find that the operator $\mathcal{L}_{Z(t)}$ has to satisfy (at least formally) for $t\in(0,\tau)$ the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{Z(t)}=\mathcal{L}_T+\mathcal{L}_P+\left[\mathcal{L}_{Z(t)},\mathcal{L}_P\right]+\sum_{k\geq 1}\frac{B_k}{k!}\mathrm{ad}^k_{\mathcal{L}_{Z(t)}}\left(\mathcal{L}_T+\mathcal{L}_P\right).$$

But using the identification (II.10), we see that this formula is equivalent[1] to a similar equation at the level of the Hamiltonian functions:

$$\frac{\mathrm{d}}{\mathrm{d}t}Z(t)=T+P+\{Z(t),P\}+\sum_{k\geq 1}\frac{B_k}{k!}\mathrm{ad}^k_{Z(t)}(T+P), \qquad \text{(II.20)}$$

where this time

$$\mathrm{ad}_H K=\{H,K\}$$

with $\{\cdot,\cdot\}$ the Poisson brackets of two functions defined in (II.7). Here, $Z=Z(t,y)$ is a function depending on the space variable $y=(p,q)$ and the time t. Hence equation (II.20) can be seen as a nonlinear transport equation.

Now the big difference with the linear matrix case is that there is *a priori* no hope for this differential equation to have a solution $Z(t)$ on a small interval $(0,\tau)$. Indeed, we cannot find a norm on nonlinear Hamiltonian functions satisfying $\|\{H,K\}\|\leq C\,\|H\|\,\|K\|$ for a uniform constant C independent on H and K. This is due to the presence of derivatives in the Poisson brackets. Hence the infinite series in (II.20) is in general divergent.

Note however that this is possible in the class of quadratic Hamiltonians of the form $H=y^TAy$ for some symmetric matrix A. In this situation the Poisson bracket of two quadratic Hamiltonians remains quadratic and hence can always be identified with a symmetric matrix. This corresponds to the linear case.

Going back to the general case of nonlinear Hamiltonian functions, equation (II.20) can be solved in the sense of *formal series* in powers of t and hence as T and P are assumed to be smooth, in the sense of asymptotic expansions in powers of the small parameter t. Let us make the formal Ansatz

$$Z(t)=\sum_{\ell\geq 0}t^\ell Z_\ell, \quad Z_0=0. \qquad \text{(II.21)}$$

[1] The equivalence between $\mathcal{L}_H$ and H holds up to an integration constant that we fix to 0.

We first observe that

$$\mathrm{ad}_{Z(t)} = \sum_{\ell \geq 0} t^\ell \mathrm{ad}_{Z_\ell}.$$

Hence we have that

$$\mathrm{ad}^k_{Z(t)} = \sum_{n \geq 0} t^n \sum_{\ell_1 + \cdots + \ell_k = n} \mathrm{ad}_{Z_{\ell_1}} \cdots \mathrm{ad}_{Z_{\ell_k}}$$

and by identifying the powers of t, the collection of Hamiltonian functions Z_n, $n \geq 1$ has to satisfy the induction formula: for all $n \geq 0$,

$$\begin{aligned}(n+1)Z_{n+1} = \delta_n^0(T+P) + \{Z_n, P\} \\ + \sum_{k \geq 1} \frac{B_k}{k!} \sum_{\ell_1 + \cdots + \ell_k = n} \mathrm{ad}_{Z_{\ell_1}} \cdots \mathrm{ad}_{Z_{\ell_k}}(T+P),\end{aligned}$$

where δ_n^0 is the Kronecker symbol. As $Z_0 = 0$, in the previous sum, the indices ℓ_j in the sum are all greater than 1. Hence k cannot be greater than n, and moreover we have $\ell_j \leq n - k + 1$ for all j. The previous formula can thus be written:

$$(n+1)Z_{n+1} = \delta_n^0(T+P) + \{Z_n, P\} + \sum_{k=0}^{n} \frac{B_k}{k!} \sum_{\substack{\ell_1 + \cdots + \ell_k = n \\ 1 \leq \ell_j \leq n-k+1}} \mathrm{ad}_{Z_{\ell_1}} \cdots \mathrm{ad}_{Z_{\ell_k}}(T+P). \tag{II.22}$$

For $n = 0$, this formula yields

$$Z_1 = T + P. \tag{II.23}$$

For $n \geq 1$, if we assume that $Z_1, \ldots, Z_n$ are given, we note that the right-hand side of (II.22) is a linear combination of Poisson brackets of the Hamiltonians Z_ℓ, $\ell \leq n$ containing a finite number of terms. By induction this shows the existence of functions Z_ℓ such that the formal series (II.21) satisfies (II.20). Note moreover that all the Z_ℓ are made of multiple derivatives of the Hamiltonian functions T and P.

Roughly speaking, this means that for a fixed N, we can construct the Hamiltonian $Z^N(\tau) = \tau(T + P) + \sum_{\ell=2}^N \tau^\ell Z_\ell$, which will satisfy (II.20) up to a small error of order $\mathcal{O}(\tau^N)$ with a constant depending on N. This result constitutes a "polynomial" version of backward error analysis:

Theorem II.13. *Assume that T and P are smooth Hamiltonian functions and $H = T + P$. Let $N \in \mathbb{N}$ and $M > 0$ and τ_0 be fixed. Then there exist constants C depending on M, N and τ_0 such that for all $\tau \leq \tau_0$, there exists a smooth modified Hamiltonian H_τ^N such that for all $y \in \mathbb{R}^{2d}$ with $\|y\| \leq M$ we have*

$$\left| H(y) - H_\tau^N(y) \right| \leq C\tau, \tag{II.24}$$

and

$$\left\| \varphi^\tau_{H_\tau^N}(y) - \varphi_P^\tau \circ \varphi_T^\tau(y) \right\| \leq C\tau^{N+1}.$$

Proof. For all $\tau > 0$ and $N \in \mathbb{N}$, we define the function

$$Z^N(\tau) = \sum_{\ell=1}^{N} \tau^\ell Z_\ell$$

where the Z_ℓ are given by the recursive formula (II.22), and we set $H_\tau^N = \frac{1}{\tau} Z^N(\tau)$. The expression (II.23) shows that for $\|y\| \leq M$ and $\tau \leq \tau_0$,

$$\left| H_\tau^N(y) - H(y) \right| \leq \sum_{k=2}^{N} \tau^{\ell-1} Z_\ell(y)| \leq \tau \left(\sum_{\ell=2} \tau_0^{\ell-2} \left(\sup_{\|y\| \leq M} |Z_\ell(y)| \right) \right) \leq C\tau,$$

and this shows (II.24).
Let us now consider the expression

$$R^N(t, y) := \exp\left(\mathcal{L}_{Z^N(t)}\right) [\mathrm{Id}](y) - \exp(t\mathcal{L}_T) \circ \exp(t\mathcal{L}_P)[\mathrm{Id}](\mathrm{y})$$

in the sense of formal series in powers of t.
By construction of the functions Z_ℓ, the relations (II.22) are satisfied for $n = 0, \dots, N - 1$ (the term Z_{N+1} is not present). This means that the derivative of $R^N(t, y)$ is a formal series with vanishing coefficients up to the order t^N included. After integration, we thus obtain $R^N(t, y) = \mathcal{O}(t^{N+1})$ in the sense of formal series. Using now Proposition II.5, we thus see that for a fixed y, the coefficients of the Taylor expansion of the *function*

$$\tau \mapsto r^N(\tau, y) = \varphi_{H_\tau}^\tau(y) - \varphi_P^\tau \circ \varphi_T^\tau(y)$$

vanish up to the order $N + 1$. Using a Taylor formula applied to $r^N(t, y)$ for $\tau \in (0, \tau_0)$ then yields the result. ■

The following corollary shows the preservation of energy over a long time:

Corollary II.14. *Let $y^0 \in \mathbb{R}^{2d}$, M and $N > 0$ be fixed. Then there exists τ_0 such that the following holds: for all $\tau \leq \tau_0$, let H_τ^N be the Hamiltonian defined in the previous theorem, and let y^n be the sequence defined by $y^{n+1} = \varphi_P^\tau \circ \varphi_T^\tau(y^n)$, $n \geq 0$. Assume that for all $n \geq 0$, we have $\|y^n\| \leq M$. Then we have*

$$\left| H_\tau^N(y^n) - H_\tau^N\left(y^0\right) \right| \leq Cn\tau^{N+1}, \tag{II.25}$$

where C depends on M and N. In particular, we have that

$$\left| H(y^n) - H\left(y^0\right) \right| \leq c\tau, \quad \text{for} \quad n \leq \tau^{-N} \tag{II.26}$$

for some constant c depending on N and M.

Proof. For $n \in \mathbb{N}$, we have

$$\begin{aligned} H_\tau^N\left(y^{n+1}\right) - H_\tau^N\left(y^n\right) &= H_\tau^N\left(\varphi_P^\tau \circ \varphi_T^\tau\left(y^n\right)\right) - H_\tau^N\left(y^n\right) \\ &= H_\tau^N\left(\varphi_P^\tau \circ \varphi_T^\tau\left(y^n\right)\right) - H_\tau^N\left(\varphi_{H_\tau^N}^\tau\left(y^n\right)\right) \end{aligned}$$

where we used the preservation of the Hamiltonian H_τ^N by the exact flow $\varphi_{H_\tau^N}^\tau$. Now as for all n we have $\|y^n\| \leq M$, we can always assume that τ_0 is such that $\left\|\varphi_{H_\tau^N}^\tau(y^n)\right\| \leq 2M$. We deduce that

$$\begin{aligned} \left|H_\tau^N\left(y^{n+1}\right) - H_\tau^N\left(y^n\right)\right| &\leq \left(\sup_{\substack{\|y\| \leq 3M \\ \tau \leq \tau_0}} \left\|\nabla H_\tau^N(y)\right\|\right) \left\|\varphi_{H_\tau^N}^\tau\left(y^n\right) - \varphi_P^\tau \circ \varphi_T^\tau\left(y^n\right)\right\| \\ &\leq C\tau^{N+1} \end{aligned}$$

for some constant C depending on M and N (as τ_0 does). This shows (II.25) by induction. The second equation is a consequence of (II.24) which implies that for all n we have

$$\begin{aligned} \left|H\left(y^n\right) - H\left(y^0\right)\right| &\leq \left|H\left(y^n\right) - H_\tau^N\left(y^n\right)\right| + \left|H_\tau^N\left(y^n\right) - H_\tau^N\left(y^0\right)\right| \\ &\quad + \left|H_\tau^N\left(y^0\right) - H\left(y^0\right)\right| \\ &\leq C\left(\tau + n\tau^{N+1}\right) \leq 2C\tau \quad \text{for} \quad n \leq \tau^{-N}, \end{aligned}$$

for some constant C. ■

Remark II.15. Note that a similar analysis can be performed for the Strang splitting $\varphi_P^{\tau/2} \circ \varphi_T^\tau \circ \varphi_P^{\tau/2}$. The only difference is that (II.24) can be replaced by

$$\left|H_\tau^N(y) - H(y)\right| \leq C\tau^2$$

for another modified energy H_τ^N. The reason is that the Strang splitting is an integrator of order 2 which means that the modified Hamiltonian determined by the BCH formula coincides with H up to the order (τ^3) (or equivalently that $Z_2 = 0$ in the previous construction). Hence the preservation of energy as expressed by (II.26) can be improved to $\mathcal{O}(\tau^2)$.

So far we have proved that in the finite dimensional case, splitting methods do preserve energy over a very long time, up to some small constant decaying with the step size τ, see (II.26). This important result has been used in [26, Chap. X] to prove the stability of symplectic numerical methods applied to integrable systems, using perturbation theory. We will not give more details here, and refer to [26], [34] for more general results on symplectic integrators.

III Infinite dimensional and semi-discrete Hamiltonian flow

In this chapter, we define the solutions of a class of nonlinear Schrödinger equations with polynomial nonlinearities. We consider the case where the equation is set on the torus $\mathbb{T}^d$ with $d \geq 1$. Let us consider a nonlinear Schrödinger equation (abbreviated as NLS in the following) of the form

$$i\partial_t u = -\Delta u + Q(u, \bar{u}). \tag{III.1}$$

The first problem we face in studying such a partial differential equation is that the right-hand side does not act on L^2 or any Sobolev space H^s. If for example $u \in H^s$ with $s > 0$, then $\Delta u \in H^{s-2}$ and we cannot apply the standard fixed point argument to define a solution. However, as mentioned in Chapter I, we can always define the flow $\varphi_T^t = e^{it\Delta}$ of the free Schrödinger equation $i\partial_t u = -\Delta u$ in a Fourier space: in dimension d, if $\xi_a(t)$ are the Fourier coefficients of the solution $u(t) = e^{it\Delta}u(0)$, then we have $\xi_a(t) = e^{-it|a|^2}\xi_a(0)$ (here $a = (a^1, \dots, a^d) \in \mathbb{Z}^d$ and $|a|^2 = (a^1)^2 + \cdots + (a^d)^2$). Hence we see that $e^{it\Delta}$ is an *isometry* of the Sobolev spaces H^s. The solution of (III.1) is then defined using Duhamel's formula

$$u(t) = e^{it\Delta}u^0 + \int_0^t e^{i(t-s)\Delta} \circ Q(u(s), \bar{u}(s))\mathrm{d}s,$$

to which we can now apply a fixed point procedure. Such a solution is called a *mild solution* of (III.1) and we will only consider such solutions in the following.

Another problem coming into play is the presence of the nonlinearity Q. To deal with these terms, we will not use the classical Sobolev spaces, but Banach spaces based on the space of functions with integrable Fourier coefficients, called a Wiener algebra. In such spaces, we will show that the right-hand sides will always be locally Lipschitz, so that we can easily derive some existence results.

We then discuss some global existence results in dimension $d = 1$, in the case where the nonlinearity is defocusing (which means that the polynomial Hamiltonian associated with Q is non negative) and in the case of small initial data.

All the present analysis will be made by discussing the Hamiltonian structure of the equation (III.1).

To conclude this chapter, we consider the space discretization of nonlinear Schrödinger equations with polynomial nonlinearity using pseudo-spectral collocation methods, as mentioned in the introduction. Because of the presence of aliasing problems, this makes the bounds for the existence time *a priori* depend on the discretization parameter (the number of points K on the torus). However, we show that we can obtain explicit bounds that will help later to prove results for fully discrete schemes under CFL condition.

1 NLS in Fourier space

Let us consider the cubic Schrödinger equation,

$$i\partial_t u(t,x) = -\Delta u(t,x) + |u(t,x)|^2 u(t,x) \tag{III.2}$$

set on the d-dimensional torus $\mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$. We can decompose $u(t,x)$ at least formally in Fourier series

$$u(t,x) = \sum_{a\in\mathbb{Z}^d} \xi_a(t)e^{ia\cdot x}, \tag{III.3}$$

where for $a = (a^1,\dots,a^d) \in \mathbb{Z}^d$ and $x = (x_1,\dots,x_d) \in \mathbb{T}^d$, we set $a\cdot x = a^1x_1 + \cdots + a^d x_d$. Plugging this decomposition into (III.2) yields

$$\sum_{a\in\mathbb{Z}^d} i\dot{\xi}_a(t)e^{ia\cdot x} = \sum_{a\in\mathbb{Z}^d} |a|^2\xi_a(t)e^{ia\cdot x} + \sum_{a_1,a_2,a_3\in\mathbb{Z}^d} e^{i(a_1-a_2+a_3)\cdot x}\xi_{a_1}(t)\bar{\xi}_{a_2}(t)\xi_{a_3}(t),$$

where $\dot{\xi}_a(t)$ denote the derivative with respect to the time t, and where $|a|^2 = (a^1)^2 + \cdots + (a^d)^2$. By identifying the components in the Fourier basis, we thus see that (III.2) is formally equivalent to the collection of coupled ordinary differential equations

$$\forall a\in\mathbb{Z}^d, \quad i\dot{\xi}_a(t) = |a|^2\xi_a(t) + \sum_{a=a_1-a_2+a_3} \xi_{a_1}(t)\bar{\xi}_{a_2}(t)\xi_{a_3}(t), \tag{III.4}$$

where the last sum holds for all triplets $(a_1,a_2,a_3)\in(\mathbb{Z}^d)^3$ such that $a = a_1-a_2+a_3$.

Let us now consider the (normalized) Hamiltonian associated with (III.2):

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

Plugging the decomposition (III.3) into this expression, this energy can be written in terms of $\xi = (\xi_a)_{a\in\mathbb{Z}^d} \in \mathbb{C}^{\mathbb{Z}^d}$,

$$\begin{aligned} H(\xi,\bar{\xi}) &= \sum_{a,b\in\mathbb{Z}^d} \frac{1}{(2\pi)^d}\int_{\mathbb{T}^d}(ia)(-ib)e^{i(a-b)\cdot x}\xi_a\bar{\xi}_b\mathrm{d}x \\ &\quad + \frac{1}{2}\sum_{a_1,a_2,b_1,b_2\in\mathbb{Z}^d}\frac{1}{(2\pi)^d}\int_{\mathbb{T}^d} e^{i(a_1+a_2-b_1-b_2)\cdot x}\xi_{a_1}\xi_{a_2}\bar{\xi}_{b_1}\bar{\xi}_{b_2}\mathrm{d}x \\ &= \sum_{a\in\mathbb{Z}^d}|a|^2|\xi_a|^2 + \frac{1}{2}\sum_{a_1+a_2-b_1-b_2=0}\xi_{a_1}\xi_{a_2}\bar{\xi}_{b_1}\bar{\xi}_{b_2}. \end{aligned} \tag{III.5}$$

Considering ξ_a and $\eta_a := \bar{\xi}_a$ as independent variables, we thus see that the system (II.1) can be written

$$\forall a\in\mathbb{Z}^d, \quad \dot{\xi}_a(t) = -i\frac{\partial H}{\partial\eta_a}(\xi,\eta),$$

where we identify the function u with the collection $(\xi_a)_{a\in\mathbb{Z}^d}$.

We see that at least formally, the cubic nonlinear Schrödinger equation can be embedded into a class of Hamiltonian equations of the form

$$\forall a \in \mathbb{Z}^d, \quad \dot{\xi}_a(t) = -i\frac{\partial H}{\partial \eta_a}(\xi, \eta) \quad \text{and} \quad \dot{\eta}_a(t) = i\frac{\partial H}{\partial \xi_a}(\xi, \eta), \tag{III.6}$$

where $H(\xi, \eta)$ is a polynomial in the collection $\xi = (\xi_a)_{a\in\mathbb{Z}^d} \in \mathbb{C}^{\mathbb{Z}^d}$ and $\eta = (\eta_a)_{a\in\mathbb{Z}^d} \in \mathbb{C}^{\mathbb{Z}^d}$.

Let us now consider an initial value $(\xi^0, \eta^0) \in \mathbb{C}^{\mathbb{Z}^d} \times \mathbb{C}^{\mathbb{Z}^d}$, and assume that for all $a \in \mathbb{Z}^d$, $\xi_a^0 = \bar{\eta}_a^0$. Then using the expression (III.5) of the Hamiltonian associated with the NLS equation, we see that for all a, we have $\xi_a(t) = \bar{\eta}_a(t)$ throughout the solution of the previous system. In other words, the collection (ξ, η) actually corresponds to a function u through the identification

$$u(x) = \sum_{a\in\mathbb{Z}^d} \xi_a e^{ia\cdot x}, \quad \text{and} \quad \bar{u}(x) = \sum_{a\in\mathbb{Z}^d} \eta_a e^{-ia\cdot x}. \tag{III.7}$$

In the following, we will only consider Hamiltonian systems satisfying this property, and we say such Hamiltonians are *real*, and we will give some examples below.

Remark III.1. The system (III.6) is a complex Hamiltonian system. The connexion with the finite dimensional case can be made as follows: Along a solution satisfying $\xi = \bar{\eta}$, we can also define the real variables p_a and q_a given by

$$\xi_a = \frac{1}{\sqrt{2}}(p_a + iq_a) \quad \text{and} \quad \bar{\xi}_a = \frac{1}{\sqrt{2}}(p_a - iq_a).$$

Then the system (III.6) is equivalent to the system

$$\begin{cases} \dot{p}_a = -\dfrac{\partial H}{\partial q_a}(q, p) & a \in \mathbb{Z}^d, \\ \dot{q}_a = \dfrac{\partial H}{\partial p_a}(q, p), & a \in \mathbb{Z}^d, \end{cases}$$

which is an infinite (real) Hamiltonian system in the sense of the previous chapter.

Now consider a more general case where the Hamiltonian function is given by

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

where P is a polynomial in u and $\bar{u}$ such that for all u, $P(u, \bar{u}) \in \mathbb{R}$. The corresponding PDE is written

$$i\partial_t u = -\Delta u + Q(u, \bar{u}), \tag{III.8}$$

where

$$Q(u,\bar{u}) = \partial_2 P(u,\bar{u})$$

is a polynomial in u and $\bar{u}$. Typical examples are obtained by considering nonlinearities of the form $P(u,\bar{u}) = \frac{\lambda}{\sigma+1}|u|^{2\sigma+2}$ for $\lambda \in \mathbb{R}$ and $\sigma \in \mathbb{N}$, for which we have $Q = \lambda |u|^{2\sigma} u$. We can verify that such a Hamiltonian is *real* in the sense defined above.

Now consider the decomposition $P(u,\bar{u}) = \sum_{k=1}^{r} P_k(u,\bar{u})$ where each P_k is a homogeneous polynomial in $(u,\bar{u})$ of degree k. We can write

$$P_k(u,\bar{u}) = \sum_{p+q=k} a_{pq} u^p \bar{u}^q,$$

where a_{pq} are complex coefficients satisfying the condition

$$\forall p,q, \quad \bar{a}_{pq} = a_{qp} \tag{III.9}$$

ensuring the fact that P_k are *real*.

Now plugging again the decomposition (III.7) into the Hamiltonian associated with P_k, we obtain that

$$\begin{aligned} &\frac{1}{(2\pi)^d}\int_{\mathbb{T}^d} P_k(u(x),\bar{u}(x))\,\mathrm{d}x \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{T}^d} \sum_{p+q=k} a_{pq} \Big(\sum_{a\in\mathbb{Z}^d} \xi_a e^{ia\cdot x}\Big)^p \Big(\sum_{b\in\mathbb{Z}^d} \eta_b e^{ib\cdot x}\Big)^q \\ &= \sum_{p+q=k} a_{pq} \sum_{a_1+\cdots+a_p-b_1-\cdots-b_q=0} \xi_{a_1}\cdots\xi_{a_p}\eta_{b_1}\cdots\eta_{b_q}. \end{aligned}$$

Here the summation in the right-hand side is made over the set of indices

$$(a_1,\ldots,a_p,b_1,\ldots,b_q) \in (\mathbb{Z}^d)^{p+q}$$

satisfying the *zero momentum* condition $a_1+\cdots+a_p-b_1-\cdots-b_q=0$.

Hence in Fourier variables, the Hamiltonian P can be viewed as a polynomial in the variables ξ and η. Note that the condition (III.9) ensures the fact that P is real, i.e. that the Hamiltonian system (III.6) associated with P degenerates into two copies of the same system when the initial value satisfies the condition $\bar{\xi} = \eta$.

2 Function spaces

As explained in the previous section, we work in the complex Fourier variables $z = (\xi,\eta) \in \mathbb{C}^{\mathbb{Z}^d} \times \mathbb{C}^{\mathbb{Z}^d}$. More precisely, we introduce the set $\mathcal{Z} = \mathbb{Z}^d \times \{\pm 1\}$ and the variables $z \in \mathbb{C}^{\mathcal{Z}}$ such that

$$\forall j = (a, \delta) \in \mathbb{Z}^d \times \{\pm 1\}, \quad z_j = \begin{cases} \xi_a & \text{if} \quad \delta = 1, \\ \eta_a & \text{if} \quad \delta = -1. \end{cases}$$

Moreover, we define the absolute value $|z_j|$ for $j = (a, \delta)$ by $|z_j| = |\xi_a|$ if $\delta = 1$ and $|z_j| = |\eta_a|$ if $\delta = -1$. Similarly, for $j = (a, \delta) \in \mathbb{Z}^d \times \{\pm 1\}$, we set $|j| = \max(1, |a|)$.

Finally, for $s \geq 0$, we define the following norm:

$$\|z\|_{\ell^1_s} = \sum_{j \in \mathcal{Z}} |j|^s |z_j|. \tag{III.10}$$

Note that in terms of ξ and η, this norm can be written

$$\|z\|_{\ell^1_s} = \sum_{a \in \mathbb{Z}^d} \max(1, |a|)^s \left(|\xi_a| + |\eta_a|\right),$$

and in the case where $\eta = \bar{\xi}$, this norm is $2 \sum_{a \in \mathbb{Z}^d} \max(1, |a|)^s |\xi_a|$.

We define the Banach space

$$\ell^1_s := \left\{ z \in \mathbb{C}^{\mathcal{Z}} \mid \|z\|_{\ell^1_s} < +\infty \right\}.$$

Now let $z = (\xi, \eta) \in \ell^1_s$ with $\xi = \bar{\eta}$, and u the associated complex function defined by (III.7). Then we say that, by abuse of notation, $u \in \ell^1_s$. In the case where $s = 0$, this space is called Wiener algebra, and we denote it simply by $\ell^1 := \ell^1_0$. With this identification, $u \in \ell^1_s$ with $s \in \mathbb{N}$ if and only $\partial^k u \in \ell^1$ for $|k| = 0, \ldots, s$.

We will also consider the classical Sobolev norms, for $s \geq 0$,

$$\|z\|_{\ell^2_s} = \left(\sum_{j \in \mathcal{Z}} |j|^{2s} |z_j|^2 \right)^{1/2},$$

and the space

$$\ell^2_s := \left\{ z \in \mathbb{C}^{\mathcal{Z}} \mid \|z\|_{\ell^2_s} < +\infty \right\} = H^s\left(\mathbb{T}^d\right),$$

where

$$H^s\left(\mathbb{T}^d\right) = \left\{ u(x) \mid \partial^k u \in L^2\left(\mathbb{T}^d\right), \ |k| = 0, \ldots, s \right\}.$$

The identification between the Fourier coefficients and the function space is made by using the Fourier transform which is an isometry. The big difference is that ℓ^2_s are now Hilbert spaces. However, to deal with polynomial nonlinearities, the spaces ℓ^1_s will be much more convenient. These spaces are imbricated via the following relation:

Proposition III.2. *Let s and s' be such that $s' - s > d/2$. Then we have*

$$\ell^2_{s'} \subset \ell^1_s \subset \ell^2_s, \tag{III.11}$$

and there exists a constant C such that for all z,

$$\|z\|_{\ell^2_s} \le \|z\|_{\ell^1_s} \le C\, \|z\|_{\ell^2_{s'}}\,. \tag{III.12}$$

Proof. Let $z \in \ell^2_{s'}$. We have, using the Cauchy–Schwartz inequality,

$$\begin{aligned}\|z\|_{\ell^1_s} &= \sum_{j\in\mathcal{Z}} |j|^s |z_j| = \sum_{j\in\mathcal{Z}} |j|^{s-s'} |z_j| |j|^{s'} \\ &\le \Big(\sum_{j\in\mathcal{Z}} |j|^{2s'} |z_j|^2\Big)^{1/2} \Big(\sum_{j\in\mathcal{Z}} |j|^{2(s-s')}\Big)^{1/2} \le C\, \|z\|_{\ell^2_{s'}}\,,\end{aligned}$$

owing to the fact that the series in the right-hand side is convergent for $s-s' < -d/2$. Now assume that $z \in \ell^1_s$, then we have in particular that for all $j \in \mathcal{Z}$, $|j|^s|z_j| \le \|z\|_{\ell^1_s}$. Hence we have

$$\|z\|^2_{\ell^2_s} = \sum_{j\in\mathcal{Z}} |j|^{2s}|z_j|^2 \le \|z\|_{\ell^1_s} \sum_{j\in\mathcal{Z}} |j|^s |z_j| = \|z\|^2_{\ell^1_s}\,. \qquad \blacksquare$$

3 Polynomials and vector fields

Let $r \in \mathbb{N}$ be given. For a collection of indices $\boldsymbol{j} = (j_1, \dots, j_r) \in \mathcal{Z}^r$, we define the *momentum* $\mathcal{M}(\boldsymbol{j})$ by the following formula: if for all $i = 1, \dots, r$ we have $j_i = (a_i, \delta_i) \in \mathcal{Z}$ we set

$$\mathcal{M}(\boldsymbol{j}) := \sum_{i=1}^{r} a_i \delta_i\,. \tag{III.13}$$

Moreover, for such a multi-index, we set

$$z_{\boldsymbol{j}} = z_{j_1} \dots z_{j_r}\,.$$

Note that such a term mixes the ξ_a and the η_a depending on the signs of the δ_i. For example in the nonlinear term of the Hamiltonian (III.5), the sum is made over indices $j_1 = (a_1, 1)$, $j_2 = (a_2, 1)$, $j_3 = (b_1, -1)$ and $j_4 = (b_2, -1)$ and involves the monomials $\xi_{a_1}\xi_{a_2}\eta_{b_1}\eta_{b_2}$. The relation $a_1 + a_2 - b_1 - b_2 = 0$ is then equivalent to $\mathcal{M}(j_1, j_2, j_3, j_4) = 0$.

For $r \in \mathbb{N}$, we define the following set of indices:

$$\mathcal{I}_r := \{\boldsymbol{j} \in \mathcal{Z}^r \mid \mathcal{M}(\boldsymbol{j}) = 0\}\,. \tag{III.14}$$

Moreover, for $j = (a, \delta) \in \mathcal{Z}$, we set $\bar{j} = (a, -\delta)$, and for a multi-index $\boldsymbol{j} = (j_1, \ldots, j_r)$, we set $\bar{\boldsymbol{j}} = (\bar{j}_1, \ldots, \bar{j}_r)$.

We now give a definition of the polynomial nonlinearities that we consider:

Definition III.3. *We say that a polynomial Hamiltonian $P \in \mathcal{P}_k$ if P is of degree k, has a zero of order at least 2 in $z = 0$, and if*

- *P contains only monomials $a_{\boldsymbol{j}} z_{\boldsymbol{j}}$ having zero momentum, i.e. such that $\mathcal{M}(\boldsymbol{j}) = 0$ when $a_{\boldsymbol{j}} \neq 0$ and thus P formally reads*

$$P(z) = \sum_{\ell=2}^{k} \sum_{\boldsymbol{j} \in \mathcal{I}_\ell} a_{\boldsymbol{j}} z_{\boldsymbol{j}} \tag{III.15}$$

  with the relation $a_{\bar{\boldsymbol{j}}} = \bar{a}_{\boldsymbol{j}}$ ensuring the fact that P is real.

- *The coefficients $a_{\boldsymbol{j}}$ are bounded, i.e. satisfy*

$$\forall \ell = 2, \ldots, k, \quad \forall \boldsymbol{j} = (j_1, \ldots, j_\ell) \in \mathcal{I}_\ell, \quad |a_{\boldsymbol{j}}| \leq C.$$

  In the following, we set

$$\|P\| = \sum_{\ell=2}^{k} \sup_{\boldsymbol{j} \in \mathcal{I}_\ell} |a_{\boldsymbol{j}}|. \tag{III.16}$$

Definition III.4. *We say that $P \in \mathcal{SP}_k$ if $P \in \mathcal{P}_k$ has coefficients $a_{\boldsymbol{j}}$ such that $a_{\boldsymbol{j}} \neq 0$ implies that $\boldsymbol{j}$ contains the same numbers of positive and negative indices:*

$$\sharp \{i \mid j_i = (a_i, +1)\} = \sharp \{i \mid j_i = (a_i, -1)\}. \tag{III.17}$$

In other words, P contains only monomials with the same numbers of ξ_i and η_i. Note that this implies that k is even.

Example III.5. In the cubic nonlinearity associated with the Hamiltonian

$$P(u, \bar{u}) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} |u(x)|^4 \mathrm{d}x,$$

the corresponding polynomial in Fourier variables is given by

$$\begin{aligned} P(z) = P(\xi, \eta) &= \sum_{k_1 + k_2 - \ell_1 - \ell_2 = 0} \xi_{k_1} \xi_{k_2} \eta_{\ell_1} \eta_{\ell_2} \\ &= \sum_{\mathcal{M}(j_1, j_2, j_3, j_4) = 0} a_{j_1 j_2 j_3 j_4} z_{j_1} z_{j_2} z_{j_3} z_{j_4} \end{aligned}$$

with the relation

$$a_{j_1 j_2 j_3 j_4} = \begin{cases} 1 & \text{if} \quad \delta_1 = \delta_2 = 1 \quad \text{and} \quad \delta_3 = \delta_4 = -1, \\ 0 & \text{otherwise} \end{cases}$$

Of course we have in this case $\|P\| = 1$ and $P \in \mathcal{SP}_4$.

With such a polynomial we associate its gradient

$$\nabla P(z) = \left(\frac{\partial P}{\partial z_j}\right)_{j\in\mathcal{Z}}.$$

We then define the Hamiltonian vector fields $X_P(z)$ by the formula

$$\forall j \in \mathcal{Z}, \quad (X_P(z))_j = (J\nabla P(z))_j = \begin{cases} -i\dfrac{\partial P}{\partial \eta_a} & \text{if} \quad \delta = 1, \\ i\dfrac{\partial P}{\partial \xi_a} & \text{if} \quad \delta = -1. \end{cases}$$

Note that here J is an infinite dimensional symplectic operator similar to the one studied in the previous chapter, but with a multiplication by the complex number i. Actually, for two functions F and G, the Poisson Bracket is (formally) defined as

$$\{F,G\} = \nabla F^T J \nabla G = i\sum_{a\in\mathbb{Z}^d} \frac{\partial F}{\partial \xi_a}\frac{\partial G}{\partial \eta_a} - \frac{\partial F}{\partial \eta_a}\frac{\partial G}{\partial \xi_a}. \tag{III.18}$$

Hence with a given Hamiltonian polynomial $P \in \mathcal{P}_k$, we associate the Hamiltonian system

$$\dot{z} = X_P(z)$$

which can be written in the form (III.6).

All the previous calculations were formal. The following proposition will show that for polynomials with bounded norm, they actually make sense on the spaces ℓ^1_s. This is the most technical result of this chapter, but it is essential for the construction of the modified Hamiltonian in Chapter VI.

Proposition III.6. *Let $k \geq 2$ and $s \geq 0$ and let $P \in \mathcal{P}_k$. Then we have $P \in \mathcal{C}^\infty(\ell^1_s, \mathbb{C})$ and $X_P \in \mathcal{C}^\infty(\ell^1_s, \ell^1_s)$. Moreover we have the estimates*

$$|P(z)| \leq \|P\| \left(\max_{n=2,\dots,k} \|z\|^n_{\ell^1_s}\right) \tag{III.19}$$

and

$$\forall z \in \ell^1_s, \quad \|X_P(z)\|_{\ell^1_s} \leq 2k(k-1)^s \|P\| \, \|z\|_{\ell^1_s} \left(\max_{n=1,\dots,k-2} \|z\|^n_{\ell^1_s}\right). \tag{III.20}$$

Moreover, for z and y in ℓ^1_s, we have

$$\|X_P(z) - X_P(y)\|_{\ell^1_s} \leq 4k(k-1)^s \|P\| \left(\max_{n=1,\dots,k-2} \left(\|y\|^n_{\ell^1_s}, \|z\|^n_{\ell^1_s}\right)\right) \|z-y\|_{\ell^1_s}. \tag{III.21}$$

Eventually, for $P \in \mathcal{P}_k$ and $Q \in \mathcal{P}_\ell$, then $\{P, Q\} \in \mathcal{P}_{k+\ell-2}$ and we have the estimate

$$\|\{P, Q\}\| \leq 2k\ell \, \|P\| \, \|Q\| \, . \tag{III.22}$$

If now $P \in \mathcal{S}\mathcal{P}_k$ and $Q \in \mathcal{S}\mathcal{P}_k$, then $\{P, Q\} \in \mathcal{S}\mathcal{P}_{k+\ell-2}$.

Proof. Assume that P is given by (III.15), and denote by P_i the homogeneous component of degree i of P, i.e.,

$$P_i(z) = \sum_{\boldsymbol{j} \in \mathcal{J}_i} a_{\boldsymbol{j}} z_{\boldsymbol{j}}, \quad i = 2, \ldots, k.$$

We have for all z and using the definition of $\|P_i\| = \sup_{\boldsymbol{j} \in \mathcal{J}_i} |a_{\boldsymbol{j}}|$,

$$|P_i(z)| \leq \|P_i\| \, \|z\|_{\ell^1}^i \leq \|P_i\| \, \|z\|_{\ell^1_s}^i \, .$$

The first inequality (III.19) is then a consequence of the fact that

$$\|P\| = \sum_{i=2}^{k} \|P_i\| \, . \tag{III.23}$$

Now let $j = (a, \epsilon) \in \mathcal{Z}$ be fixed. The derivative of a given monomial $z_{\boldsymbol{j}} = z_{j_1} \cdots z_{j_i}$ with respect to z_j vanishes except if $j \subset \boldsymbol{j}$. Assume for instance that $j = j_i$. Then the zero momentum condition implies that $\mathcal{M}(j_1, \ldots, j_{i-1}) = -\epsilon a$ and we can write

$$|j|^s \left| \frac{\partial P_i}{\partial z_j} \right| \leq i \, \|P_i\| \sum_{\boldsymbol{j} \in \mathcal{Z}^{i-1}, \, \mathcal{M}(\boldsymbol{j}) = -\epsilon a} |j|^s |z_{j_1} \cdots z_{j_{i-1}}|. \tag{III.24}$$

Now in this formula, for a fixed multi-index $\boldsymbol{j}$, the zero momentum condition implies that

$$|j|^s \leq (|j_1| + \cdots + |j_{i-1}|)^s \leq (i-1)^s \max_{n = 1, \ldots, i-1} |j_n|^s. \tag{III.25}$$

Therefore, after summing in a and ϵ we get

$$\begin{aligned} \left\| X_{P_i}(z) \right\|_{\ell^1_s} &\leq 2i(i-1)^s \, \|P_i\| \sum_{\boldsymbol{j} \in \mathcal{Z}^{i-1}} \max_{n = 1, \ldots, i-1} |j_n|^s |z_{j_1}| \cdots |z_{j_{i-1}}| \\ &\leq 2i(i-1)^s \, \|P_i\| \, \|z\|_{\ell^1_s} \, \|z\|_{\ell^1}^{i-2} \end{aligned} \tag{III.26}$$

which yields (III.20) after summing in $i = 2, \ldots, k$.
Note that this implies that

$$\|\nabla P(z)\|_{\ell^\infty} := \sup_{j \in \mathcal{Z}} \left| \frac{\partial P(z)}{\partial z_j} \right| \leq \|X_P(z)\|_{\ell^1_0} < \infty$$

which shows that $\nabla P \in \mathcal{C}(\ell_s^1, \mathbb{C})$ and hence P is in $\mathcal{C}^1(\ell_s^1, \mathbb{C})$.
Now for z and y in ℓ_s^1 we have, with the previous notation,

$$|j|^s \left| \frac{\partial P_i}{\partial z_j}(z) - \frac{\partial P_i}{\partial z_j}(y) \right| \leq \sum_{q \in \mathcal{Z}} |j|^s \left| \int_0^1 \frac{\partial P_i}{\partial z_j \partial z_q}(ty + (1-t)z)\, \mathrm{d}t \right| |z_q - y_q|.$$

But we have, for fixed $j = (\epsilon, a)$ and $q = (b, \delta)$ in $\mathcal{Z}$, and for all $u \in \ell_s^1$,

$$|j|^s \left| \frac{\partial P_i}{\partial z_j \partial z_q}(u) \right| \leq i \,\|P_i\| \sum_{\boldsymbol{j} \in \mathcal{Z}^{i-2},\, \mathcal{M}(\boldsymbol{j}) = -\epsilon a - \delta b} |j|^s |u_{j_1} \cdots u_{j_{i-2}}|.$$

In the previous sum, we necessarily have that $\mathcal{M}(\boldsymbol{j}, j, q) = 0$, and hence

$$|j|^s \leq (|j_1| + \cdots + |j_{i-2}| + |q|)^s \leq (i-1)^s |q|^s \prod_{n=1}^{i-2} |j_n|^s.$$

Let $u(t) = ty + (1-t)z$; we have for all $t \in [0, 1]$, with the previous estimates,

$$|j|^s \left| \int_0^1 \frac{\partial P_i}{\partial z_j \partial z_q}(u(t)) \mathrm{d}t \right| \leq i(i-1)^s |q|^s \,\|P_i\| \int_0^1 \sum_{\boldsymbol{j} \in \mathcal{Z}^{i-2},\, \mathcal{M}(\boldsymbol{j}) = -\epsilon a - \delta b} \times |j_1|^s |u_{j_1}(t)| \cdots |j_{i-2}|^s |u_{j_{i-2}}(t)| \mathrm{d}t.$$

Multiplying by $(z_q - y_q)$ and summing in k and j, we obtain

$$\left\| X_{P_i}(z) - X_{P_i}(y) \right\|_{\ell_s^1} \leq 4i(i-1)^s \,\|P_i\| \left(\int_0^1 \|u(t)\|_{\ell_s^1}^{i-2} \mathrm{d}t \right) \|z - y\|_{\ell_s^1}.$$

Hence we obtain the result after summing in i, using the fact that

$$\|ty + (1-t)z\|_{\ell_s^1} \leq \max(\|y\|_{\ell_s^1}, \|z\|_{\ell_s^1}).$$

Note that the previous calculations show that for all $z \in \ell_s^1$,

$$\nabla X_P(z) \in \mathcal{C}(\ell_s^1, \ell_s^1)$$

and this implies that X_P is in $\mathcal{C}^1(\ell_s^1, \ell_s^1)$. The fact that the Hamiltonian P and the vector field X_P are $\mathcal{C}^\infty$ can be verified using similar calculations.
Assume now that P and Q are homogeneous polynomials of degrees k and ℓ respectively and with coefficients $a_{\boldsymbol{k}}$, $\boldsymbol{k} \in \mathcal{J}_k$ and $b_{\boldsymbol{\ell}}$, $\boldsymbol{\ell} \in \mathcal{J}_\ell$. It is clear that $\{P, Q\}$ is a monomial of degree $k + \ell - 2$ satisfying the zero momentum condition. Furthermore writing

$$\{P, Q\}(z) = \sum_{\boldsymbol{j} \in \mathcal{J}_{k+\ell-2}} c_{\boldsymbol{j}} z_{\boldsymbol{j}},$$

$c_{\boldsymbol{j}}$ is expressed as a sum of coefficients $a_{\boldsymbol{k}}b_{\boldsymbol{\ell}}$ for which there exists an $a \in \mathbb{Z}$ and $\epsilon \in \{\pm 1\}$ such that

$$(a,\epsilon) \subset \boldsymbol{k} \in \mathcal{I}_k \quad \text{and} \quad (a,-\epsilon) \subset \boldsymbol{\ell} \in \mathcal{I}_\ell,$$

and such that if for instance $(a,\epsilon) = k_1$ and $(a,-\epsilon) = \ell_1$, we necessarily have

$$(k_2,\ldots,k_k,\ell_2,\ldots,\ell_\ell) = \boldsymbol{j}\,.$$

Hence for a given $\boldsymbol{j}$, the zero momentum condition on $\boldsymbol{k}$ and on $\boldsymbol{\ell}$ determines the value of ϵa which in turn determines two possible values of (ϵ, a) (as $a \in \mathbb{Z}^d$).
This proves (III.22) for monomials. If

$$P = \sum_{i=2}^{k} P_i \quad \text{and} \quad Q = \sum_{j=2}^{\ell} Q_j$$

where P_i and Q_j are homogeneous polynomials of degree i and j respectively, then we have

$$P = \sum_{n=2}^{k+\ell-2} \sum_{i+j-2=n} \{P_i, Q_j\}.$$

Hence by definition of $\|P\|$ (see (III.16)) and the fact that all the polynomials $\{P_i, Q_j\}$ in the sum are homogeneous of degree $i+j-2$, we have by the previous calculations

$$\|P\| = \sum_{n=2}^{k+\ell-2} \left\| \sum_{i+j-2=n} \{P_i, Q_j\} \right\| \leq 2 \sum_{n=2}^{k+\ell-2} \sum_{i+j-2=n} ij\, \|P_i\|\, \|Q_j\|$$

$$\leq 2k\ell \left(\sum_{i=2}^{k} \|P_i\| \right) \left(\sum_{j=2}^{\ell} \|Q_j\| \right) = 2k\ell\, \|P\|\, \|Q\|\,,$$

where we used (III.23) for the last equality.
The last assertion, as well as the fact that the Poisson bracket of two real Hamiltonians is real, follow immediately from the definition of the Poisson bracket. ■

4 Local existence of the flow

We are now prepared to define the flow of a Hamiltonian PDE of the form (III.8) with a polynomial nonlinearity $Q(u,\bar{u}) = \partial_2 P(u,\bar{u})$. So far we have shown how this

PDE can be interpreted as an infinite dimensional Hamiltonian system involving the coefficients $z = (\xi, \eta)$ and the Hamiltonian function

$$H(z) = T(z) + P(z),$$

where $P \in \mathscr{P}_r$ and where T is the Hamiltonian associated with the Laplace operator. This one is defined by

$$T(z) = T(\xi, \eta) := \sum_{a \in \mathbb{Z}^d} |a|^2 \xi_a \eta_a, \tag{III.27}$$

compare (III.5): When $z \in \ell_1^2 = H^1(\mathbb{T}^d)$ or $z \in \ell_1^1 \subset \ell_1^2$, and when z is *real* which means $z = (\xi, \eta)$ with $\xi = \bar{\eta}$, we have

$$T(z) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} |\nabla u(x)|^2 \mathrm{d}x,$$

where $u(x) = \sum_{a \in \mathbb{Z}^d} \xi_a \, e^{i a \cdot x}$.

The Hamiltonian system (III.6) associated with the function $H(z)$ then reads

$$\begin{cases} \dot{\xi}_a = -i|a|^2 \xi_a - i \dfrac{\partial P}{\partial \eta_a}(\xi, \eta), & a \in \mathbb{Z}^d, \\ \dot{\eta}_a = i|a|^2 \eta_a + i \dfrac{\partial P}{\partial \xi_a}(\xi, \eta), & a \in \mathbb{Z}^d, \end{cases} \tag{III.28}$$

and corresponds to the nonlinear Schrödinger equation

$$i \partial_t u = -\Delta u + \partial_2 P(u, \bar{u}) \tag{III.29}$$

where $\partial_2 P(u, \bar{u}) = i X_P(z)$ after the identification between u and z.

Note that the vector field $X_T(z)$ acts from ℓ_s^1 to ℓ_{s-2}^1 and hence the flow of the previous Hamiltonian equation cannot be defined by a direct use of the Cauchy Lipschitz Theorem in a Banach space. However, as mentioned in the introduction, the flow of T, $\varphi_T^t(z)$ can always be defined: it is given by the formula

$$z(t) = (\xi(t), \eta(t)) = \varphi_T^t\left(\xi^0, \eta^0\right), \quad \text{where} \quad \begin{cases} \xi_a(t) = \exp\left(-it|a|^2\right) \xi_a^0, & a \in \mathbb{Z}^d, \\ \eta_a(t) = \exp\left(it|a|^2\right) \eta_a^0, & a \in \mathbb{Z}^d. \end{cases}$$

In particular, we observe that the flow of T acts as a rotation in the Fourier coefficients. Hence we have for all time $t > 0$ and all $z \in \ell_s^1$,

$$\left\| \varphi_T^t(z) \right\|_{\ell_s^1} = \|z\|_{\ell_s^1}. \tag{III.30}$$

Instead of considering the equation

$$\dot{z} = X_H(z) = X_T(z) + X_P(z), \quad \text{with} \quad z(0) = z^0,$$

we reformulate it using Duhamel's formula as

$$\forall t > 0 \quad z(t) = \varphi_T^t(z^0) + \int_0^t \varphi_T^{t-s} \circ X_P(z(s))\, \mathrm{d}s, \tag{III.31}$$

in which all the terms are now well defined in ℓ_s^1. A $\mathcal{C}^1$ function $z(t)$ solution of the previous system is called a *mild solution* of the Hamiltonian system (III.28). Note that such a formulation expressed in terms of the function $u(t) = u(t,x)$ solution of (III.29) can be written

$$u(t) = e^{it\Delta}u^0 + \int_0^t e^{i(t-s)\Delta}\partial_2 P\,(u(s), \bar{u}(s))\, \mathrm{d}s.$$

In the following, for a given number M, we define the open ball

$$B_M^s = \left\{ z \in \ell_s^1 \mid \|z\|_{\ell_s^1} < M \right\}. \tag{III.32}$$

Theorem III.7. *Let $P \in \mathcal{P}_k$ for some given $k \in \mathbb{N}$, $M > 0$ and $s \geq 0$. Then there exists t_* and for all $|t| \leq t_*$ a mapping $\varphi_H^t : B_M^s \to \ell_s^1$ of class $\mathcal{C}^1$ and such that for all $z^0 \in B_M^s$, $z(t) = \varphi_H^t(z^0)$ is the unique mild solution in ℓ_s^1 of the Hamiltonian system* (III.28)*. If moreover $z^0 = (\xi^0, \bar{\xi}^0)$ is real, then $\varphi_H^t(z^0)$ is real for $|t| \leq t_*$.*

Proof. Let us fix $z^0 \in B_M^s$ and $\bar{t} > 0$, and let us consider the Banach space $\mathcal{E} = \mathcal{C}^0([-\bar{t}, \bar{t}], \ell_s^1)$ equipped with the norm

$$\|\psi(\cdot)\|_{\mathcal{E}} = \sup_{\sigma \in [-\bar{t}, \bar{t}]} \|\psi(\sigma)\|_{\ell_s^1}.$$

The mapping

$$(\mathcal{T}\psi)(t) = \varphi_T^t(z^0) + \int_0^t \varphi_T^{t-\sigma} \circ X_P(\psi(\sigma))\, \mathrm{d}\sigma,$$

defines an mapping $\mathcal{T} : \mathcal{E} \mapsto \mathcal{E}$. This is a consequence of (III.30) and Proposition III.6. Now consider the function

$$[-\bar{t}, \bar{t}] \ni t \mapsto \psi^0(t) = \varphi_T^t(z^0).$$

Note that as φ_T^t is an isometry, we have that for all $t \in [-\bar{t}, \bar{t}]$, $\psi^0(t) \in B_M^s$ and $\psi^0 \in \mathcal{E}$.

Let $\eta > 0$, and consider the ball in $\mathcal{E}$, centered in ψ^0 and with radius η:

$$\mathcal{B}_\eta\left(\psi^0\right) := \left\{ \psi(\cdot) \in \mathcal{E} \mid \left\|\psi - \psi^0\right\|_{\mathcal{E}} \leq \eta \right\}.$$

Note that if $\psi \in \mathcal{B}_\eta(\psi^0)$, we have for all $t \in [-\bar{t}, \bar{t}]$, $\|\psi(t)\|_{\ell_s^1} \leq M + \eta \leq 2M$ if we assume that $\eta < M$. Then for such ψ, using estimate (III.20), we see that for all $t \in [-\bar{t}, \bar{t}]$,

$$\left\|(\mathcal{T}\psi)(t) - \psi^0(t)\right\|_{\ell_s^1} \leq \int_0^{|t|} 2k(k-1)^s \|P\| \, \|\psi(s)\|_{\ell_s^1} \left(\max_{n=1,\dots,k-2} \|\psi(s)\|_{\ell_s^1}^n \right) \mathrm{d}s$$
$$\leq 2|t|k(k-1)^s \|P\| \, 2M \max\left(1, (2M)^{k-2}\right) \leq C\bar{t},$$

where C depends on k, M, $\|P\|$ and s. A similar calculation for $t \in [-\bar{t}, 0]$ shows that

$$\left\|\mathcal{T}\psi - \psi^0\right\|_{\mathcal{E}} \leq C\bar{t}.$$

Hence for $\bar{t} \leq \eta/C$, the mapping $\mathcal{T}$ maps $\mathcal{B}_\eta(\psi^0)$ into itself. Now consider ψ^1 and ψ^2 in $\mathcal{B}_\eta(\psi^0)$. Using (III.21) we see that there exists a constant L depending on M, $\|P\|$, s and k such that

$$\left\|\mathcal{T}\psi^1 - \mathcal{T}\psi^2\right\|_{\mathcal{E}} \leq \bar{t}L \left\|\psi^1 - \psi^2\right\|_{\mathcal{E}}.$$

Hence for $\bar{t} \leq 1/(2L)$, the mapping $\mathcal{T}$ is a contraction mapping from $\mathcal{B}_\eta(\psi^0)$ to itself. Taking $t_* = \min(\eta/C, 1/(2L))$, the fixed point theorem then ensures the existence and uniqueness of a solution $z(t)$, $t \in [-t_*, t_*]$, satisfying $\mathcal{T}z = z$ which means that z is a mild solution of (III.28). The properties of $z(t) =: \varphi_H^t(z^0)$ are then easily verified by using Proposition III.6. ■

Remark III.8. Note that a mild solution in ℓ_s^1 is also a mild solution in $\ell_s^2 = H^s(\mathbb{T}^d)$, owing to (III.12). However the converse is not true: a mild solution in ℓ_s^2 is a mild solution in $\ell_{s'}^1$ only if $s - s' > d/2$.

With this definition of the flow, we get the following:

Corollary III.9. *Assume that $z^0 \in \ell_1^1$ is real, and that $\varphi_H^t(z^0)$ is well defined for $t \in [0, t_*]$. Then we have*

$$H\left(\varphi_H^t\left(z^0\right)\right) = H\left(z^0\right), \quad \textit{for} \quad t \in [0, t_*]. \tag{III.33}$$

Proof. Using Theorem III.7, the flow $\varphi_H^t(z_0)$ is well defined for sufficiently small t in the space ℓ_1^1. Let $K \in \mathbb{N}$. We define the projection operator $\Pi_K : \ell_1^1 \mapsto \ell_1^1$ such that

$$\forall j \in \mathcal{Z}, \quad (\Pi_K z)_j = \begin{cases} z_j & \text{if} \quad |j| \leq K, \\ 0 & \text{if} \quad |j| > K. \end{cases}$$

It is clear that for all $z \in \ell_1^1$, we have $\|\Pi_K z\|_{\ell_1^1} \leq \|z\|_{\ell_1^1}$ and $\|\Pi_K z - z\|_{\ell_1^1} \to 0$ when $K \to +\infty$. Let us consider the Hamiltonian $H^K := T + P \circ \Pi_K$, and $z^{(K)} = \Pi_K z^0$. It is clear that for all K, we can define a solution in ℓ_1^1 of (III.28) associated with the Hamiltonian H^K and with initial value $z^{(K)}$. As $\|P \circ \Pi_K\| \leq \|P\|$, we can always assume that this solution is well defined for $t \in [0, t_K]$ with $t_K > t_*$. Moreover, as T is diagonal in Fourier, we easily see that for all t, $\Pi_K \circ \varphi^t_{H^K}(z^{(K)}) = \varphi^t_{H^K}(z^{(K)})$. Hence the flow $\varphi^t_{H^K}$ is finite dimensional. We deduce that for all $t \in [0, t_K]$ we have

$$H^K\left(\varphi^t_{H^K}\left(z^{(K)}\right)\right) = H^K\left(z^{(K)}\right).$$

Now this relation holds for all K and all $t \in [0, t_*]$. Hence by letting $K \to +\infty$, and as $\varphi^t_H(z^0) \in \ell_1^1 \subset \ell_1^2$, we can prove that the relation (III.33) holds true. ■

5 Cases of global existence

In the rest of this chapter, we will consider the case where $d = 1$. The global existence results given below use in a crucial manner the preservation of energy. Hence we need to consider solutions in $\ell_1^2 = H^1(\mathbb{T})$. In the previous section, we have seen that if the initial data z^0 is in ℓ_1^1, there exists a local solution in ℓ_1^1 and hence in ℓ_1^2. However in Corollary III.9 we used the fact that the flow $\varphi^t_{H^K}$ was converging towards φ^t_H in ℓ_1^1, which is a consequence of Proposition III.6. Hence to derive the energy preservation of a solution in ℓ_1^2, we need to make the assumption that the nonlinearity X_P acts on ℓ_1^2, which will be the case for standard polynomials in dimension 1.

The first result concerns the case where the nonlinearity has a positive sign. In this situation the solutions are global in $\ell_1^2 = H^1(\mathbb{T})$.

Proposition III.10. *Let us consider the Hamiltonian system* (III.28) *with a nonlinearity satisfying $X_P \in \mathscr{C}^1(\ell_1^2, \ell_1^2)$. Assume moreover that for all real $z \in \ell_1^2$, we have*

$$|P(z)| \geq 0.$$

Let $z^0 = (\xi^0, \eta^0) \in \ell_1^2$ be real. Then the flow $\varphi^t_H(z^0)$ exists in ℓ_1^2 for all time $t \in \mathbb{R}$.

Proof. Let us first note that with the assumption on the nonlinearity X_P, we can define a mild solution $\varphi^t_H(z^0)$ of (III.28) in $\ell_1^2 = H^1(\mathbb{T})$. Moreover, in this situation, we can show that (III.33) holds by using the same kind of proof. Now as $P \geq 0$ and using the fact that $H(\varphi^t_H(z^0)) \in \mathbb{R}$ because z^0 is real, we obtain

$$\left\|\varphi^t_H\left(z^0\right)\right\|^2_{\ell_1^2} = 2T\left(\varphi^t_H\left(z^0\right)\right) \leq 2H\left(\varphi^t_H\left(z^0\right)\right) = 2H\left(z^0\right) < +\infty.$$

By standard arguments, this shows the solution is global in ℓ_1^2. ■

Remark III.11. If the initial data is in ℓ^1_1 and the nonlinearity acts on ℓ^2_1, then the previous result shows that the solution is global in ℓ^2_1 and hence in $\ell^1 := \ell^1_0$ is the one-dimensional case.

Note that in dimension 1, the condition that $X_P \in \mathcal{C}^1(\ell^2_1, \ell^2_1)$ will be satisfied for polynomials in $(u, \bar{u})$. This is a consequence of the following:

Lemma III.12. *There exists a constant C such that for all u and $v \in H^1(\mathbb{T})$ we have*

$$\|uv\|_{H^1(\mathbb{T})} \leq C \|u\|_{H^1(\mathbb{T})} \|v\|_{H^1(\mathbb{T})} .$$

Proof. First, we note that if $u(x) = \sum_{a \in \mathbb{Z}} \xi_a \, e^{iax}$ we have that $u(x) \in \mathcal{C}(\mathbb{T}, \mathbb{R})$ with the estimate

$$\|u\|_{L^\infty} \leq \|z\|_{\ell^1} \leq c \|z\|_{\ell^2_1} = c \|u\|_{H^1} ,$$

where $z = (\xi, \eta)$ and for some constant c given by the inclusions (III.11). Hence we immediately obtain that $\|uv\|_{L^2} \leq 2c \|u\|_{H^1} \|v\|_{H^1}$.
Moreover, we have

$$\partial_x (uv) = v \partial_x u + u \partial_x v$$

and hence by integration, $\|\partial_x (uv)\|_{L^2} \leq \|v\|_{L^\infty} \|u\|_{H^1} + \|u\|_{L^\infty} \|v\|_{H^1}$ which yields the result. ■

Example III.13. On the one-dimensional torus, the defocusing cubic nonlinear Schrödinger equation

$$i \partial_t u = -\Delta u + \lambda |u|^2 u$$

with $\lambda > 0$ has global solutions in $H^1(\mathbb{T})$. The nonlinearity is here

$$P(u, \bar{u}) = \frac{1}{2\pi} \int_{\mathbb{T}} \frac{\lambda}{2} |u(x)|^4 \, \mathrm{d}x \geq 0, \quad \forall u \in \ell^2_1.$$

We conclude this section by giving another case of global existence: In dimension 1 and when the initial data is small enough in $H^1(\mathbb{T}) = \ell^2_1$. In the next statement, we say that $P \in \mathcal{P}_k$ has a zero of order m at the origin $z = 0$, with $m \geq 1$, if P involves only monomials $z_{\boldsymbol{j}}$ with $\boldsymbol{j} \in \mathcal{I}_r$ with $r \geq m$. In other words, the components of order $0, 1, \ldots, m-1$ in the decomposition of P in homogeneous polynomials vanish.

Proposition III.14. *Let us consider the Hamiltonian system (III.28) with a polynomial $P \in \mathcal{P}_k$ having a zero of order at least 3 at the origin $z = 0$. Then there exists*

$\varepsilon > 0$ *such that for real* $z^0 \in \ell_1^2 = H^1(\mathbb{T})$ *with* $\left\| z^0 \right\|_{\ell_1^2} \leq \varepsilon$, *the flow* $\varphi_H^t(z^0)$ *exists in* $\ell_1^2 = H^1(\mathbb{T})$ *for all time* $t \in \mathbb{R}$ *and satisfies*

$$\forall t \in \mathbb{R}, \quad \left\| \varphi_H^t\left(z^0\right) \right\|_{\ell_1^2} \leq 2\varepsilon. \tag{III.34}$$

Proof. The hypothesis on P and equation (III.19) imply that for all $z \in \ell_1^2$ with $\|z\|_{\ell_1^2} \leq 1$,

$$|P(z)| \leq \|P\| \max_{n=3,\ldots,k} \left(\|z\|_{\ell^1}^n \right) \leq C \|z\|_{\ell_1^2}^3 \leq C T(z)^{3/2},$$

by definition of $T(z) = \frac{1}{2} \|z\|_{\ell_1^2}^2$, and for some constant C. Hence we have for all $z \in \ell_1^2$ with $\|z\|_{\ell_1^2} \leq 1$,

$$T(z)\left(1 - CT(z)^{1/2}\right) \leq H(z) \leq T(z)\left(1 + CT(z)^{1/2}\right).$$

Let $z(t) = \varphi_H^t(z^0)$. For all time t where the flow is well defined in ℓ_1^2 and remains of norm smaller than 1, we can write

$$T\left(z(t)\right)\left(1 - CT\left(z(t)\right)^{1/2}\right) \leq H\left(z(t)\right) = H\left(z^0\right) \leq T\left(z^0\right)\left(1 + CT\left(z^0\right)^{1/2}\right).$$

Assume that $\left\| z^0 \right\|_{\ell_1^2} \leq \varepsilon$. Then we have $T(z^0) \leq \frac{1}{2}\varepsilon^2$, and we have

$$H\left(z^0\right) \leq T\left(z^0\right)\left(1 + CT\left(z^0\right)^{1/2}\right) \leq \varepsilon^2$$

provided $\frac{C}{\sqrt{2}}\varepsilon < 1$. This shows that for all time t such that $\|z(t)\|_{\ell_1^2} \leq 1$,

$$T\left(z(t)\right)\left(1 - CT\left(z(t)\right)^{1/2}\right) \leq \varepsilon^2.$$

Now assume that $T(z(t)) \leq 2\varepsilon^2 \leq 1$, we can write

$$T\left(z(t)\right) \leq \varepsilon^2 + CT\left(z(t)\right)^{3/2} \leq 2\varepsilon^2,$$

provided $C2^{3/2}\varepsilon < 1$. By classical arguments, this shows that for all t, we have $T(z(t)) \leq 2\varepsilon^2$ and $z(t)$ is well defined in ℓ_1^2 for all time $t \in \mathbb{R}$ and satisfies (III.34). ■

Example III.15. The previous theorem holds true for polynomial nonlinearities of the form

$$P\left(u, \bar{u}\right) = \frac{1}{2\pi} \int_{\mathbb{T}} \frac{\lambda}{\sigma + 1} |u|^{2\sigma+2} \, \mathrm{d}x,$$

for any $\lambda \in \mathbb{R}$, and in particular for the cubic nonlinear Schrödinger equation. Note that the ε given by the previous proof is not very small in general (of order $1/\lambda$ in the case where $\sigma = 1$).

6 Semi-discrete flow

Following the example in the introductory chapter, we consider now a space discretization of the previous class of semi-linear Schrödinger equations. However, we will restrict the presentation to the cubic case. Similarly, we will only consider the case where the dimension $d = 1$. Note however that the results below can be easily extended to higher dimensions and other polynomial nonlinearities.

Let K be an integer. We define the set (see (I.8))

$$B^K := \begin{cases} \{-P, \dots, P-1\} & \text{if} \quad K = 2P \quad \text{is even,} \\ \{-P, \dots, P\} & \text{if} \quad K = 2P+1 \quad \text{is odd.} \end{cases} \tag{III.35}$$

With this set is associated the grid $x_a = 2\pi a/K$ with $a \in B^K$ made of K equidistant points in the interval $[-\pi, \pi]$. Recall that the discrete Fourier transform is defined as the mapping $\mathcal{F}_K : \mathbb{C}^K \to \mathbb{C}^K$ such that for all $a \in B^K$,

$$(\mathcal{F}_K v)_a = \frac{1}{K} \sum_{b \in B^K} e^{-2i\pi ab/K} v_b \quad \text{and} \quad \left(\mathcal{F}_K^{-1} v\right)_a = \sum_{b \in B^K} e^{2i\pi ab/K} v_b. \tag{III.36}$$

Let us now consider the cubic nonlinear Schrödinger equation

$$i\partial_t u(t,x) = -\Delta u(t,x) + \lambda |u(t,x)|^2 u(t,x), \quad u(0,x) = u^0(x),$$

where $\lambda \in \mathbb{R}$ and $x \in \mathbb{T}$. We consider the pseudo-spectral collation method defined as follows: Find a trigonometric polynomial

$$U^K(t,x) = \sum_{a \in B^K} e^{ixa} \xi_a^K(t)$$

such that for all $b \in B^K$, the equation

$$\begin{aligned} & i\partial_t U^K(t,x_b) = -\Delta U^K(t,x_b) + \lambda \left|U^K(t,x_b)\right|^2 U^K(t,x_b), \\ & U^K(0,x_b) = u^0(x_b), \end{aligned} \tag{III.37}$$

is satisfied for all time t. For a fixed $b \in B^K$, we calculate that

$$\left|U^K(t,x_b)\right|^2 U^K(t,x_b) = \sum_{a_1,a_2,a_3 \in B^K} \xi_{a_1}^K(t) \eta_{a_2}^K(t) \xi_{a_3}^K(t) e^{ix_b(a_1-a_2+a_3)},$$

where $\eta_a^K = \bar{\xi}_a^K$. In particular, we get for $a \in B^K$,

$$\begin{aligned} & \left(\mathcal{F}_K\left(\left|U^K(t,x_b)\right|^2 U^K(t,x_b)\right)\right)_a \\ & = \frac{1}{K} \sum_{b \in B^K} \sum_{a_1,a_2,a_3 \in B^K} \xi_{a_1}^K(t) \eta_{a_2}^K(t) \xi_{a_3}^K(t) e^{ix_b(a_1-a_2+a_3-a)}. \end{aligned}$$

But for a fixed $d \in B^K$, we calculate that for even and odd K,

$$\frac{1}{K}\sum_{b\in B^K} e^{idx_b} = \frac{1}{K}\sum_{b\in B^K} \exp\left(\frac{2\pi i b d}{K}\right) = \begin{cases} 1 & \text{if} \quad d = mK, \quad m \in \mathbb{Z}, \\ 0 & \text{if} \quad d \neq mK, \quad m \in \mathbb{Z}. \end{cases} \tag{III.38}$$

Hence we have

$$\left(\mathcal{F}_K\left(\left|U^K(t,x_b)\right|^2 U^K(t,x_b)\right)\right)_a = \sum_{m\in\mathbb{Z}}\ \sum_{a_1-a_2+a_3-a=mK} \xi^K_{a_1}(t)\eta^K_{a_2}(t)\xi^K_{a_3}(t).$$

Using this formula, and taking the discrete Fourier transform of the expression (III.37), we get the following equation for the Fourier coefficients $\xi^K_a(t)$:

$$\forall a \in B^K, \quad i\dot{\xi}^K_a = a^2\xi^K_a + \lambda \sum_{m\in\mathbb{Z}}\ \sum_{a=a_1-a_2+a_3+mK} \xi^K_{a_1}\eta^K_{a_2}\xi^K_{a_3}. \tag{III.39}$$

Note that in the last sum, we have $(a_1, a_2, a_3) \in (B^K)^3$ and hence as $a \in B^K$, we verify that we cannot have $|m| \geq 2$. We recognize here a Hamiltonian PDE of the form studied above, with a Hamiltonian of the form

$$H^K(\xi,\eta) = T^K + P^K := \sum_{a\in B^K} a^2\xi_a\eta_a + \frac{\lambda}{2} \sum_{\substack{a_1+a_2-a_3-a_4=mK \\ a_i\in B^K,\,|m|\leq 1}} \xi_{a_1}\xi_{a_2}\eta_{a_3}\eta_{a_4}. \tag{III.40}$$

We thus see that the only difference with the continuous case, is that the *zero momentum* condition is not satisfied, but is replaced by a zero momentum modulo K. This is a typical problem of *aliasing*. Fortunately in the case where P is a polynomial, the possible values for m in the equation above will always be bounded. This defines the class of discretized polynomials below.

Of course, the local existence of the solution of equation (III.39) is guaranteed by the finite dimensional Cauchy–Lipschitz Theorem, but we can also view this equation as posed on a finite dimensional subspace (of dimension K) of the Banach spaces ℓ^1_s or ℓ^2_s. The following results extend Proposition III.6 to the case of a polynomial Hamiltonian of the form above. It will be used later to prove the existence of a modified energy for fully discrete splitting schemes applied to NLS.

We define the set of signed indices

$$\mathcal{B}^K = \left\{ j = (a,\delta) \in B^K \times \{\pm 1\}\right\} \subset \mathcal{Z}. \tag{III.41}$$

For $r \in \mathbb{N}$ and $m \in \mathbb{Z}$ we define the following set of indices (compare (III.14))

$$\mathcal{I}^K_{r,m} := \left\{ \boldsymbol{j} \in \left(\mathcal{B}^K\right)^r \mid \mathcal{M}(\boldsymbol{j}) = mK\right\}.$$

We extend now the notion of polynomial given by Definition III.3:

Definition III.16. *Let $K \in \mathbb{N}$, $K \geq 1$. We say that a discrete polynomial Hamiltonian $P^K \in \mathcal{P}^K_{k,p}$ if P^K is of degree k, has a zero of order at least 3 in $z = 0$, and if*

- *P^K is written*
$$P^K(z) = \sum_{\ell=2}^{k} \sum_{|m| \leq p} \sum_{\boldsymbol{j} \in \mathcal{I}^K_{\ell,m}} a^m_{\boldsymbol{j}} z_{\boldsymbol{j}} \tag{III.42}$$
with the relation $a^m_{\bar{\boldsymbol{j}}} = \bar{a}^{-m}_{\boldsymbol{j}}$.
- *The coefficients $a^m_{\boldsymbol{j}}$ are bounded, i.e. satisfy*
$$\forall \ell = 2, \ldots, k, \quad \forall |m| \leq p, \quad \forall \boldsymbol{j} = (j_1, \ldots, j_\ell) \in \mathcal{I}^K_{\ell,m}, \quad |a^m_{\boldsymbol{j}}| \leq C.$$
The norm $\left\| P^K \right\|$ is defined as
$$\left\| P^K \right\| = \sum_{\ell=2}^{k} \sum_{m=-p}^{p} \sup_{\boldsymbol{j} \in \mathcal{I}^K_{\ell,m}} |a^m_{\boldsymbol{j}}|. \tag{III.43}$$

Echoing Definition III.4, we define:

Definition III.17. *We say that $P \in \mathcal{SP}^K_{k,p}$ if $P \in \mathcal{P}^K_{k,p}$ has coefficients $a_{\boldsymbol{j}}$ such that $a_{\boldsymbol{j}} \neq 0$ implies that $\boldsymbol{j}$ contains the same numbers of positive and negative indices, i.e. satisfies* (III.17).

The next proposition corresponds to the extension of Proposition III.6 for $s = 0$:

Proposition III.18. *Let $k \geq 2$, $p \in \mathbb{N}$ and $K \in \mathbb{N}$, $K \geq 1$ and let $P^K \in \mathcal{P}^K_{k,p}$. Then we have $P^K \in \mathcal{C}^\infty(\ell^1, \mathbb{C})$ and $X_{P^K} \in \mathcal{C}^\infty(\ell^1, \ell^1)$. Moreover we have the estimates*
$$|P^K(z)| \leq \left\| P^K \right\| \left(\max_{n=2,\ldots,k} \|z\|^n_{\ell^1} \right) \tag{III.44}$$
and
$$\forall z \in \ell^1, \quad \|X_{P^K}(z)\|_{\ell^1} \leq 2k \left\| P^K \right\| \|z\|_{\ell^1} \left(\max_{n=1,\ldots,k-2} \|z\|^n_{\ell^1} \right). \tag{III.45}$$
Moreover, for z and y in ℓ^1, we have
$$\|X_{P^K}(z) - X_{P^K}(y)\|_{\ell^1} \leq 4k \left\| P^K \right\| \left(\max_{n=1,\ldots,k-2} \left(\|y\|^n_{\ell^1}, \|z\|^n_{\ell^1} \right) \right) \|z - y\|_{\ell^1}. \tag{III.46}$$
Eventually, for $P^K \in \mathcal{P}^K_{k,p}$ and $Q^K \in \mathcal{P}^K_{\ell,q}$, then $\{P^K, Q^K\} \in \mathcal{P}^K_{k+\ell-2,p+q}$ and we have the estimate
$$\left\| \{P^K, Q^K\} \right\| \leq 2(p+q)k\ell \left\| P^K \right\| \left\| Q^K \right\|. \tag{III.47}$$

Proof. As in the proof of Proposition III.6, we denote by $P^K_{i,m}$ the homogeneous component of degree i of P^K, and involving only coefficients of momentum mK, i.e.,

$$P^K_{i,m}(z) = \sum_{\boldsymbol{j} \in \mathcal{I}^K_{i,m}} a^m_{\boldsymbol{j}} z_{\boldsymbol{j}}, \quad i = 2, \dots, k.$$

The first inequality (III.19) is then a consequence of the definition of the norm of $\left\| P^K \right\|$ by using similar arguments as in the proof of Proposition III.6.

Now let $j = (a, \epsilon) \in B^K \times \{\pm 1\} \subset \mathcal{Z}$ be fixed. As for (III.24), we have

$$\left| \frac{\partial P^K_{i,m}}{\partial z_j} \right| \leq i \left\| P^K_{i,m} \right\| \sum_{\substack{\boldsymbol{j} \in \mathcal{Z}^{i-1} \\ \mathcal{M}(\boldsymbol{j}) = -\epsilon a + mK}} \left| z_{j_1} \cdots z_{j_{i-1}} \right|. \tag{III.48}$$

Therefore, after summing in a and ϵ we get

$$\left\| X_{P^K_{i,m}}(z) \right\|_{\ell^1} \leq 2i \left\| P^K_{i,m} \right\| \sum_{\boldsymbol{j} \in \mathcal{Z}^{i-1}} \left| z_{j_1} \right| \cdots \left| z_{j_{i-1}} \right| \leq 2i \left\| P^K_{i,m} \right\| \|z\|^{i-1}_{\ell^1} \tag{III.49}$$

which yields (III.45) after summing in $i = 2, \dots, k$ and m. Note that a similar estimate in the Banach space ℓ^1_s would involve terms of order K^s: the equation (III.25) is true only when the zero momentum condition is fulfilled.

The equations (III.46) and (III.47) can be proved similarly and are left to the reader.

■

Hence the collocation space discretization of a nonlinear Schrödinger equation with polynomial nonlinearity leads to consider discrete Hamiltonian functions of the form

$$H^K = T^K + P^K \tag{III.50}$$

where $T^K = \sum_{a \in B^K} a^2 \xi_a \eta_a$ and $P^K \in \mathcal{P}^K_{k,p}$ for some constants k and p, and hence satisfying the bounds independent on K given by the above proposition. Note the the discretization of the cubic nonlinear Schrödinger equation described above can be written in the previous form, with $k = 4$ and $p = 1$. Note moreover that such a Hamiltonian leaves the space

$$\mathcal{A}^K := \left\{ z_a = (\xi_a, \eta_a) \,\middle|\, \xi_a = \eta_a = 0 \quad \text{if} \quad a \notin B^K \right\} \simeq \mathbb{C}^K \times \mathbb{C}^K \tag{III.51}$$

invariant by the flow, and that T^K is the restriction of T on the subspace $\mathcal{A}^K$.

Using the previous proposition, we thus see that if the norms of the polynomials P^K are uniformly bounded, we can prove the existence of a mild solution to the system (III.50) viewed as a finite dynamical system embedded in ℓ^1, for times independent of K. In the next chapter, we will use this fact to prove the convergence of the semi-discrete flow towards the exact flow, over finite time intervals and for smooth solutions.

IV Convergence results

In this chapter, we still consider semi-linear Schrödinger equations of the form

$$i\partial_t u = -\Delta u + Q(u, \bar{u}) \tag{IV.1}$$

with polynomial Hamiltonian nonlinearity Q. We prove that under the hypothesis that the exact solution remains smooth on a finite interval, then the splitting methods are convergent. Such a result can be found in [31] for the cubic NLS. We then extend this result to more general splitting methods where the linear operator is smoothed in high frequencies either with the help of an implicit integrator, or directly using more general filter functions as in (I.11).

In Section 4, we consider the case where the equation (IV.1) is discretized in space by a pseudo-spectral collocation method as described in the end of the previous chapter. We conclude in Section 5 with the case of fully discrete systems discretized both in space and time, and show the convergence of the fully discrete splitting method over finite time, under the assumption that the exact solution is smooth. Though relatively standard from the point of view of numerical analysis, such results for fully discrete schemes are difficult to find in the existing literature (see however in [19] in the case of the Gross–Pitaevskii equation, and [29] in the linear case).

1 Splitting methods and Lie derivatives

As in the previous chapter, we associate with (IV.1) an infinite dimensional Hamiltonian system in the variable $z = (\xi, \eta)$ made of the Fourier coefficients of $u = \sum_{a\in\mathbb{Z}^d} \xi_a e^{ia\cdot x}$ and $\bar{u} = \sum_{a\in\mathbb{Z}^d} \eta_a e^{-ia\cdot x}$. With the notation of the previous chapter, we consider a Hamiltonian system associated with a Hamiltonian function of the form

$$H(z) = T(z) + P(z),$$

where $T(z) = \sum_{a\in\mathbb{Z}^d} |a|^2 \xi_a \eta_a$ is the Hamiltonian associated with the Laplace operator, and $P \in \mathscr{P}_k$, $k \geq 3$ is a polynomial Hamiltonian. Note that we could also consider a quadratic Hamiltonian of the form $T(z) = \sum_{a\in\mathbb{Z}^d} \omega_a \xi_a \eta_a$ with frequencies ω_a satisfying the bound $\omega_a \leq C|a|^2$. This would allow us to consider a more general splitting scheme based on a decomposition between the linear and nonlinear parts of the Hamiltonian PDE.

The splitting methods we consider are based on the following approximation, for a small time step τ:

$$\varphi_H^\tau \simeq \varphi_T^\tau \circ \varphi_P^\tau \tag{IV.2}$$

known as the Lie splitting method. The higher order symmetric approximation

$$\varphi_H^\tau \simeq \varphi_T^{\tau/2} \circ \varphi_P^\tau \circ \varphi_T^{\tau/2} \tag{IV.3}$$

is known as the Strang splitting approximation. Note that we can also consider the same methods where we exchange the role of T and P. This does not affect the results of this chapter.

As in the finite dimensional case, we define the Lie derivative as follows (for the notations, see the previous chapter):

Definition IV.1. *Let $g \in \mathcal{C}^\infty(\ell_s^1, \mathbb{C})$ and H a Hamiltonian; we define the Lie derivative $\mathcal{L}_H[g]$ by the formula*

$$\mathcal{L}_H[g] = \sum_{j \in \mathcal{Z}} (X_H)_j \frac{\partial g}{\partial z_j} = i \sum_{a \in \mathbb{Z}^d} \frac{\partial H}{\partial \xi_a}\frac{\partial g}{\partial \eta_a} - \frac{\partial H}{\partial \eta_a}\frac{\partial g}{\partial \xi_a} = \{H, g\}.$$

Let $Y \in \mathcal{C}^\infty(\ell_s^1, \ell_s^1)$ with $Y = (Y_j)_{j \in \mathcal{Z}}$, then we set

$$(\mathcal{L}_H[Y])_j = \left\{H, Y_j\right\}, \quad j \in \mathcal{Z}.$$

As in the finite dimensional case we have, for two Hamiltonians functions H and G

$$[\mathcal{L}_H, \mathcal{L}_G] := \mathcal{L}_H \circ \mathcal{L}_G - \mathcal{L}_G \circ \mathcal{L}_H = \mathcal{L}_{\{H,G\}}.$$

The following result will be used to define the asymptotic expansion of the solution of (IV.1), provided it fulfills some regularity assumptions.

Proposition IV.2. *Let $s, s' \geq 0$ with $s' \geq s$. Assume that $Y \in \mathcal{C}^\infty(\ell_{s'}^1, \ell_s^1)$. Then*

$$\mathcal{L}_T[Y] \in \mathcal{C}^\infty\left(\ell_{s'+2}^1, \ell_s^1\right) \quad \text{and} \quad \mathcal{L}_P[Y] \in \mathcal{C}^\infty\left(\ell_{s'}^1, \ell_s^1\right).$$

Proof. Recall that $P \in \mathcal{P}_k$ and assume that $Y = (Y_j)_{j \in \mathcal{Z}} \in \mathcal{C}^1(\ell_{s'}^1, \ell_s^1)$. For all $z \in \ell_{s'}^1$, we have $\nabla Y(z) \in \mathcal{L}(\ell_{s'}^1, \ell_s^1)$, and the Lie derivative can be written

$$\mathcal{L}_P[Y](z) = \nabla Y(z) \cdot X_P(z).$$

Hence the fact that $X_P(z) \in \ell_{s'}^1$ for $z \in \ell_{s'}^1$ (see (III.20)) shows that $\mathcal{L}_P[Y](z) \in \ell_s^1$. Now for the Hamiltonian T, to obtain that $X_T(z) \in \ell_{s'}^1$, we need that $z \in \ell_{s'+2}^1$ as T acts as the multiplication by $|a|^2$ in each component.
The result follows by a similar argument repeated on the successive derivatives of $\mathcal{L}_P[Y]$ and $\mathcal{L}_T[Y]$. ■

Let us consider the flow $\varphi_H^t(z)$ which is defined as the solution in ℓ_s^1 of the equation

$$z(t) = \varphi_T^t(z) + \int_0^t \varphi_T^{t-\sigma} \circ X_P(z(\sigma))\, \mathrm{d}\sigma. \tag{IV.4}$$

It is easy to verify that if $z(t) \in \ell^1_{s+2}$ for $t \in [0, t_*]$, then $z(t)$ satisfies the equation

$$z'(t) = X_T(z(t)) + X_P\,(z(t)) \quad \text{in} \quad \ell^1_s,$$

for $t \in [0, t_*]$. Now if $z(t) \in \ell^1_s$ for all $s \geq 0$, we can consider the Taylor expansion around $t = 0$ as in the finite dimensional situation:

$$\varphi^t_H(z) = \sum_{k \geq 0} \frac{t^k}{k!} \left. \frac{\mathrm{d}^k \varphi^t_H(z)}{\mathrm{d}t^k} \right|_{t=0},$$

and we can write at least formally

$$\varphi^t_H = \sum_{k \geq 0} \frac{t^k}{k!} \mathcal{L}^k_H[\mathrm{Id}] = \exp(t\mathcal{L}_H)\,[\mathrm{Id}].$$

With these calculations and the previous proposition, we get the following (compare Proposition II.5).

Proposition IV.3. *Let M be fixed, and let $z(t) = \varphi^t_H(z)$ be a mild solution of* (IV.4) *in ℓ^1_s. Assume that $z \in B^{s+2N+2}_M$. Then there exists t_0 such that for all $N \in \mathbb{N}$, there exists a constant C_N such that for all $t \in [0, t_0]$, we have*

$$\left\| \varphi^t_H(z) - \sum_{k \geq 0}^{N} \frac{t^k}{k!} \mathcal{L}^k_H[\mathrm{Id}](z) \right\|_{\ell^1_s} \leq C_N t^{N+1}.$$

Proof. Recall that B^s_M is defined in (III.32) as the ball of radius M in ℓ^1_s. By assumption, and using the results of the previous chapter, there exists t_0 such that for all $t \in [0, t_0]$, $\varphi^t_H(z) \in B^{s+2N+2}_{2M}$. Now using Proposition IV.2, we have by induction that

$$\forall k \geq 0, \quad \mathcal{L}^k_H[\mathrm{Id}] \in \mathcal{C}^\infty\left(\ell^1_{s+2k}, \ell^1_s\right).$$

The result is then obtained as in the finite dimensional case, using a Taylor expansion. ■

The previous proposition shows that if the solution is smooth enough, the representation of the flow as an exponential makes sense. If such an assumption is relevant over a small time interval for smooth initial value, this is in general not fair over long time intervals.

2 Convergence of the Lie splitting methods

Let us begin with the following

Lemma IV.4. *Let $s \geq 0$, and M be given. Then there exist constants L and τ_0 such that for all $\tau \leq \tau_0$, and all z and y in B^s_M, we have*

$$\left\| \varphi^\tau_T \circ \varphi^\tau_P(z) - \varphi^\tau_T \circ \varphi^\tau_P(y) \right\|_{\ell^1_s} \leq e^{L\tau} \left\| z - y \right\|_{\ell^1_s}. \tag{IV.5}$$

Proof. For $y \in B^s_M$ and $t \in (0, \tau)$, we have

$$\varphi^t_P(y) = y + \int_0^t X_P\left(\varphi^\sigma_P(y)\right) \mathrm{d}\sigma.$$

As P is a polynomial of degree k, equation (III.20) of Proposition III.6 shows that for $y \in B^s_M$,

$$\left\|\varphi^t_P(y)\right\|_{\ell^1_s} \leq M + 2k(k-1)^s \|P\| \int_0^t \left\|\varphi^\sigma_P(y)\right\|_{\ell^1_s} \left(\max_{n=0,\dots,k-2} \left\|\varphi^\sigma_P(y)\right\|^n_{\ell^1_s}\right) \mathrm{d}\sigma.$$

Hence as long as $\left\|\varphi^\sigma_P(y)\right\|^1_{\ell_s} \leq 2M$ for $\sigma \in (0, t)$ we have the estimate

$$\left\|\varphi^t_P(y)\right\|_{\ell^1_s} \leq M + 2tMk(k-1)^s \|P\| \max\left(1, (2M)^{k-2}\right). \tag{IV.6}$$

This shows that for $\tau \leq \tau_0$ where τ_0 is small enough (depending on M, k and s), we have $\varphi^\tau_P(y) \in B^s_{2M}$.
As φ^τ_T is a linear isometry, we have

$$\left\|\varphi^\tau_T \circ \varphi^\tau_P(z) - \varphi^\tau_T \circ \varphi^\tau_P(y)\right\|_{\ell^1_s} = \left\|\varphi^\tau_P(z) - \varphi^\tau_P(y)\right\|_{\ell^1_s}.$$

Hence using Proposition III.6 and the fact that $\varphi^\sigma_P(z)$ and $\varphi^\sigma_P(y)$ are bounded by $2M$ in ℓ^1_s, we obtain

$$\left\|\varphi^t_P(y) - \varphi^t_P(z)\right\|_{\ell^1_s} \leq \|y - z\|_{\ell^1_s} + L \int_0^t \left\|\varphi^\sigma_P(y) - \varphi^\sigma_P(z)\right\|_{\ell^1_s} \mathrm{d}\sigma,$$

where L is the Lipschitz constant of P over B^s_{2M} given by Proposition III.6, see equation (III.21). We conclude by using the Gronwall Lemma. ∎

We now give the following local error result:

Proposition IV.5. *Let $s \geq 0$, and assume that $z \in B^{s+2}_M$ for some $M > 0$. Then there exist τ_0 and a constant C such that for all $\tau < \tau_0$, we have*

$$\left\|\varphi^\tau_H(z) - \varphi^\tau_T \circ \varphi^\tau_P(z)\right\|_{\ell^1_s} \leq C\tau^2. \tag{IV.7}$$

Before proving this result, let us show how the argument used in the finite dimensional case studied in Chapter II can easily be adapted to the present situation, provided the *a priori* regularity of z is ℓ^1_{s+4} and not ℓ^1_{s+2} as stated in the result above.

Indeed, under the hypothesis $z \in \ell^1_{s+4}$, then for all $\sigma \leq \tau \leq \tau_0$ where τ_0 is sufficiently small, we can assume that $\varphi^\sigma_H(z)$ and $\varphi^\sigma_P(z)$ are in B^{s+4}_{2M}. Using Proposition IV.3, we can write

$$\begin{aligned}\varphi^\tau_H(z) &= z + \tau\mathcal{L}_H[\mathrm{Id}](z) + \int_0^\tau (\tau - \sigma)\mathcal{L}^2_H[\mathrm{Id}]\left(\varphi^\sigma_H(z)\right) \mathrm{d}\sigma \\ &= z + \tau\left(\mathcal{L}_T[\mathrm{Id}] + \mathcal{L}_P[\mathrm{Id}]\right)(z) + \mathcal{O}_s\left(\tau^2\right),\end{aligned}$$

where the rest is bounded by $C\tau^2$ in ℓ_s^1 because $\varphi_H^\sigma(z) \in \ell_{s+4}$ for all σ. Similarly, we have

$$\varphi_T^\tau(z) = z + \tau \mathcal{L}_T[\mathrm{Id}](z) + \int_0^\tau (\tau - \sigma)\mathcal{L}_T^2[\mathrm{Id}]\left(\varphi_T^\sigma(z)\right) \mathrm{d}\sigma,$$

and hence

$$\varphi_T^\tau \circ \varphi_P^\tau(z) = \varphi_P^\tau(z) + \tau \mathcal{L}_T[\mathrm{Id}]\left(\varphi_P^\tau(z)\right) + \mathcal{O}_s\left(\tau^2\right).$$

Now we have

$$\varphi_P^\tau(z) = z + \tau \mathcal{L}_P[\mathrm{Id}](z) + \mathcal{O}_s\left(\tau^2\right),$$

and by definition of the Lie derivative

$$\mathcal{L}_T[\mathrm{Id}]\left(\varphi_P^\tau(z)\right) = \mathcal{L}_T[\mathrm{Id}](z) + \int_0^\tau \mathcal{L}_P\mathcal{L}_T[\mathrm{Id}]\left(\varphi_P^\sigma(z)\right) \mathrm{d}\sigma.$$

Hence

$$\varphi_T^\tau \circ \varphi_P^\tau(z) = z + \tau\left(\mathcal{L}_T[\mathrm{Id}](z) + \mathcal{L}_P[\mathrm{Id}](z)\right) + \mathcal{O}_s\left(\tau^2\right),$$

which proves the result, but under the assumption that $z \in \ell_{s+4}^1$. The goal is now to show that the result still holds when $z \in \ell_{s+2}^1$ only.

Proof of Proposition IV.5. As $z \in B_M^{s+2}$, and as φ_P^τ is well defined on ℓ_{s+2}^1, the same argument as before based on the estimate (IV.6) shows that there exists τ_0 such that for all $z \in B_M^{s+2}$ and all $\sigma \leq \tau_0$, we have $\varphi_P^\sigma(z) \in B_{2M}^{s+2}$. As φ_T^σ is an isometry, the same holds for $\varphi_T^\sigma \circ \varphi_P^\tau(z)$ with $\tau \leq \tau_0$.
Let us start with the formula defining the mild solution $\varphi_H^\tau(z)$:

$$\varphi_H^\tau(z) = \varphi_T^\tau(z) + \int_0^\tau \varphi_T^{\tau-t} X_P\left(\varphi_H^t(z)\right) \mathrm{d}t.$$

By definition of the flow $\varphi_P^\tau(z)$, we have

$$\varphi_P^\tau(z) = z + \int_0^\tau X_P\left(\varphi_P^t(z)\right) \mathrm{d}t.$$

As φ_T^τ is linear, we thus have

$$\varphi_T^\tau \circ \varphi_P^\tau(z) = \varphi_T^\tau(z) + \int_0^\tau \varphi_T^\tau X_P\left(\varphi_P^t(z)\right) \mathrm{d}t.$$

We define

$$d_\tau(z) = \varphi_H^\tau(z) - \varphi_T^\tau \circ \varphi_P^\tau(z).$$

As φ_T^τ is inversible, the previous calculations show that

$$\begin{aligned}\varphi_T^{-\tau} d_\tau(z) &= \int_0^\tau \varphi_T^{-t} X_P\left(\varphi_H^t(z)\right) - X_P\left(\varphi_P^t(z)\right) \mathrm{d}t \\ &= \int_0^\tau \left(\varphi_T^{-t} \circ X_P - X_P \circ \varphi_T^{-t}\right)\left(\varphi_H^t(z)\right) \mathrm{d}t \\ &\quad + \int_0^\tau X_P\left(\varphi_T^{-t}\varphi_H^t(z)\right) - X_P\left(\varphi_P^t(z)\right) \mathrm{d}t \\ &=: r_\tau^1(z) + r_\tau^2(z).\end{aligned}$$

As $\varphi_H^t(z)$ and $\varphi_P^t(z)$ remain in the ball B_{2M}^s, we have using Proposition III.6 of the previous chapter,

$$\left\| X_P\left(\varphi_T^{-t}\varphi_H^t(z)\right) - X_P\left(\varphi_P^t(z)\right)\right\|_{\ell_s^1} \leq C \left\|\varphi_T^{-t}\varphi_H^t(z) - \varphi_P^t(z)\right\|_{\ell_s^1},$$

for some constant C depending on M. But we have

$$\varphi_T^{-t}\varphi_H^t(z) - \varphi_P^t(z) = \varphi_T^{-t} \circ \left(\varphi_H^t - \varphi_T^t\varphi_P^t\right)(z) = \varphi_T^{-t} d_t(z).$$

As φ_T^t is an isometry, we get

$$\left\|r_\tau^2(z)\right\|_{\ell_s^1} \leq C \int_0^\tau \|d_t(z)\|_{\ell_s^1} \mathrm{d}t.$$

Let us consider now $y \in \ell_{s+2}^1$, and define

$$f(t, y) = \varphi_T^{-t} \circ X_P(y) - X_P \circ \varphi_T^{-t}(y).$$

We have $f(0, y) = 0$, and

$$\frac{\partial f}{\partial t}(t, y) = -\varphi_T^{-t} \circ X_T \circ X_P(y) + \mathcal{L}_T[X_P]\left(\varphi_T^{-t}(y)\right). \tag{IV.8}$$

Note that we calculate

$$\begin{aligned}\frac{\partial f}{\partial t}(0, y) &= -X_T \circ X_P(y) + \mathcal{L}_T\mathcal{L}_P[\mathrm{Id}](y) \\ &= -\mathcal{L}_P\mathcal{L}_T[\mathrm{Id}](y) + \mathcal{L}_T\mathcal{L}_P[\mathrm{Id}](y) \\ &= \mathcal{L}_{\{T,P\}}[\mathrm{Id}](y),\end{aligned}$$

which means that the error term is driven by the commutator between H and P. Now it is clear that we have

$$\|X_T(y)\|_{\ell_s^1} \leq \|y\|_{\ell_{s+2}^1}.$$

Hence using (III.20) we get that for all $y \in B_{2M}^{s+2}$ and all $t \in [0, \tau_0]$, we have

$$\left\| \frac{\partial f}{\partial t}(t, y) \right\|_{\ell_s^1} \leq C \left(\|y\|_{\ell_{s+2}^1} \right),$$

for some constant C depending on M. But we have

$$r_\tau^1(z) = \int_0^\tau f\left(t, \varphi_H^t(z)\right) \mathrm{d}t = \int_0^\tau \int_0^t \frac{\partial f}{\partial \sigma}\left(\sigma, \varphi_H^t(z)\right) \mathrm{d}\sigma \mathrm{d}t.$$

Hence we get for all $z \in B_M^{s+2}$,

$$\|r_\tau^1\|_{\ell_s^1} \leq c\tau^2,$$

for some constant c depending on M. Gathering the estimates on $r_\tau^1(z)$ and $r_\tau^2(z)$, we thus get for all $\tau \leq \tau_0$,

$$\|d_\tau(z)\|_{\ell_s^1} \leq c\tau^2 + C \int_0^\tau \|d_t(z)\|_{\ell_s^1} \mathrm{d}t,$$

and the Gronwall Lemma then yields the result. ■

Proposition IV.6. *Let $z^0 \in \ell_{s+2}^1$, $M > 0$ and $t_* > 0$. Assume that for all $t \in (0, t_*)$, $\varphi_H^t(z^0)$ is well defined in ℓ_{s+2}^1 and remains in the ball B_M^{s+2}. Then there exist constants C and τ_0 such that for $0 \leq \tau \leq \tau_0$, if z^n is the sequence defined by induction:*

$$z^{n+1} = \varphi_T^\tau \circ \varphi_P^\tau(z^n), \quad n \geq 0,$$

then we have

$$\forall t = n\tau \leq t_*, \quad \left\|\varphi_H^t(z^0) - z^n\right\|_{\ell_s^1} \leq C\tau.$$

Proof. Setting $z(t) = \varphi_H^t(z^0)$ and $t_n = n\tau$, we have for $n \geq 0$,

$$\begin{aligned}\left\|z\left(t_{n+1}\right) - z^{n+1}\right\|_{\ell_s^1} \leq& \left\|\varphi_H^\tau\left(z(t_n)\right) - \varphi_T^\tau \circ \varphi_P^\tau\left(z(t_n)\right)\right\|_{\ell_s^1} \\ &+ \left\|\varphi_T^\tau \circ \varphi_P^\tau\left(z(t_n)\right) - \varphi_T^\tau \circ \varphi_P^\tau\left(z^n\right)\right\|_{\ell_s^1}.\end{aligned}$$

By assumption, for all n such that $n\tau \leq t_*$, $z(t_n)$ is in a ball B_M^{s+2}. Hence Proposition IV.5 shows that

$$\left\|\varphi_H^\tau\left(z\left(t_n\right)\right) - \varphi_T^\tau \circ \varphi_P^\tau\left(z\left(t_n\right)\right)\right\|_{\ell_s^1} \leq C\tau^2$$

for some constant C depending only on M and s. Using (IV.5), we see that there exists a constant L such that

$$\left\|\varphi_T^\tau \circ \varphi_P^\tau\left(z\left(t_n\right)\right)-\varphi_T^\tau \circ \varphi_P^\tau\left(z^n\right)\right\|_{\ell_s^1} \leq e^{L \tau}\left\|z\left(t_n\right)-z^n\right\|_{\ell_s^1}$$

as long as $z^n \in B_{2M}^s$ (while $z(t_n) \in B_M^{s+2} \subset B_{2M}^s$).
Using the fact that $z(t_0) = z^0$, then as long as $n\tau \leq t_*$ and $z^n \in B_{2M}^s$, we have

$$\left\|z\left(t_n\right)-z^n\right\|_{\ell_s^1} \leq C n e^{L n \tau} \tau^2 \leq \left(C t_* e^{L t_*}\right) \tau .$$

This shows that for τ_0 sufficiently small, we have $z^n \in B_{2M}^s$ – and hence the previous estimate – for $n\tau \leq t_*$. This concludes the proof. ■

Remark IV.7. A similar result holds for the Strang splitting method (IV.3). Indeed we can show the local error estimate

$$\left\|\varphi_H^\tau(z)-\varphi_T^{\tau/2} \circ \varphi_P^\tau \circ \varphi_T^{\tau/2}(z)\right\|_{\ell_s^1} \leq C\tau^3$$

when z remains bounded in B_M^{s+4}. This shows that the Strang splitting is of (global) order 2 for smooth functions. We do not give the details here.

3 Filtered splitting schemes

The standard Lie–Trotter splitting methods for PDEs associated with the Hamiltonian $T + P$ consists in replacing the flow generated by $H = T + P$ during the time τ (the small time step) by the composition of the flows generated by T and P during the same time, namely

$$\varphi_T^\tau \circ \varphi_P^\tau = \exp\left(\tau \mathcal{L}_P\right) \circ \exp\left(\tau \mathcal{L}_T\right)[\mathrm{Id}].$$

As explained in the introduction, it turns out that it is convenient to consider more general splitting methods that induce smoothing effects to the high frequencies of the linear part. Thus we replace the linear operator $\tau\mathcal{L}_T$ by a more general Hamiltonian operator associated with a Hamiltonian A_0.
More precisely let $\beta(x)$ be a real function, possibly depending on the step size τ and satisfying $\beta(0) = 0$ and $\beta(x) \simeq x$ for small x. We define the diagonal operator X_{A_0} by the relation

$$\forall j=(a, \delta) \in \mathcal{Z}, \quad \left(X_{A_0}(z)\right)_j=\delta \beta\left(\tau|a|^2\right) z_j . \tag{IV.9}$$

In other words, the system $\dot{z} = X_{A_0}(z)$ can be written for $a \in \mathbb{Z}^d$ (compare (III.28))

$$\dot{\xi}_a=-i \beta\left(\tau|a|^2\right) \xi_a, \quad \text{and} \quad \dot{\eta}_a=i \beta\left(\tau|a|^2\right) \eta_a,$$

and for real $z = (\xi, \bar{\xi})$ associated with a function $u(x) = \sum_{a \in \mathbb{Z}^d} \xi_a \, e^{ia \cdot x}$, we can rewrite this equation in shorter form,

$$i \partial_t u = \beta(-\tau \Delta) u.$$

For $a \in \mathbb{Z}^d$, we set $\lambda_a = \beta(\tau |a|^2)$. The Hamiltonian associated with X_{A_0} is given by

$$A_0(z) = A_0(\xi, \eta) = \sum_{a \in \mathbb{Z}^d} \lambda_a \xi_a \eta_a. \tag{IV.10}$$

In the following, we consider the splitting methods

$$\varphi_P^\tau \circ \varphi_{A_0}^1 \quad \text{and} \quad \varphi_{A_0}^1 \circ \varphi_P^\tau, \tag{IV.11}$$

where φ_P^τ is the exact flow associated with the Hamiltonian P, and where $\varphi_{A_0}^1$ is defined by the relation

$$\forall j = (a, \delta) \in \mathcal{Z}, \quad \left(\varphi_{A_0}^1(z)\right)_j = \exp\left(-i \delta \lambda_a\right) z_j$$

which is the flow of the Hamiltonian A_0 given by (IV.10) at time 1 (recall that the step size τ is included in the definition of the eigenvalues λ_a of A_0). Note that $\varphi_{A_0}^1$ can also be viewed as the time τ flow of the equation

$$i \partial_t u = \frac{1}{\tau} \beta(-\tau \Delta) u, \tag{IV.12}$$

that can be viewed as a regularization of the linear free Schrödinger equation $i \partial_t u = -\Delta u$. Hence $\varphi_{A_0}^1$ is a regularization of the exact flow φ_T^τ.

In these notes, we will mainly consider two cases:

(i) The case where $\beta(x) = x$ which corresponds to the case $A_0 = \tau T$, i.e. $\varphi_{A_0}^1 = \varphi_T^\tau$ and the spitting method (IV.11) coincides with (IV.2).

(ii) The case where $\beta(x) = 2 \arctan(x/2)$ corresponding to the implicit-explicit integrator introduced in the introduction (see also [1], [35]).

The second case corresponds to the approximation of the system

$$\dot{\xi}_a = -i |a|^2 \xi_a, \quad \text{and} \quad \dot{\eta}_a = i |a|^2 \eta_a, \quad a \in \mathbb{Z}^d,$$

by the midpoint rule. Starting from a given point (ξ_a^0, η_a^0), the midpoint rule applied to the first equation of the previous system is defined by the implicit relation

$$\xi_a^1 = \xi_a^0 - i \tau |a|^2 \left(\frac{\xi_a^1 + \xi_a^0}{2} \right).$$

Owing to the classical relation

$$\forall x \in \mathbb{R}, \quad \frac{1 + ix}{1 - ix} = \exp\left(2i \arctan(x)\right),$$

we can write

$$\xi_a^1 = \left(\frac{1 - i\tau|a|^2/2}{1 + i\tau|a|^2/2}\right)\xi_a^0 = \exp\left(-2i \arctan(\tau|a|^2/2)\right)\xi_a^0.$$

We easily see that a similar relation holds for η_a^1, and we eventually observe that we can interpret the numerical approximation (ξ_a^1, η_a^1) as the *exact* flow at time $t = 1$, of the Hamiltonian A_0 defined by (see formula (IV.10))

$$A_0(z) = A_0(\xi, \eta) := \sum_{a \in \mathbb{Z}^d} 2 \arctan(\tau|a|^2/2)\, \xi_a \eta_a. \tag{IV.13}$$

Remarkably, the previous calculations show the following: We can do backward error analysis for the midpoint rule applied to the free-linear Schrödinger equation, which can be interpreted at the flow at time τ of the *modified* system (IV.12) with $\beta(x) = 2\arctan(x/2)$.

Finally, we allow the possibility of making a cut-off in high frequencies, that is to consider

$$\beta(x) = x\mathbb{1}_{x \le c_0}(x), \quad \text{or} \quad \beta(x) = 2\arctan(x/2)\mathbb{1}_{x \le c_0}(x)$$

where c_0 is a given number (the CFL number). In the case of a fully discrete system, this number will be naturally determined by the highest mode in the space discretized system, but we will also consider such a high frequency cut-off in the abstract formulation. This will make possible the construction of the modified energy of Chapter VI in an abstract framework.

To analyze the convergence of the *filtered* splitting methods (IV.11), we only have to evaluate the difference between φ_T^τ and $\varphi_{A_0}^1$, and combine it with the estimate of the previous section. For the implicit-explicit integrator based on the midpoint rule, we have the following result:

Proposition IV.8. *Let $s \ge 0$, and assume that $z \in B_M^{s+4}$ for some $M > 0$. Let A_0 be defined by (IV.13) the quadratic Hamiltonian associated with the filter function $\beta(x) = 2\arctan(x/2)$. Then there exist τ_0 and a constant C such that for all $\tau \le \tau_0$, we have*

$$\left\|\varphi_H^\tau(z) - \varphi_{A_0}^1 \circ \varphi_P^\tau(z)\right\|_{\ell_s^1} \le C\tau^2. \tag{IV.14}$$

Proof. For all $x \in \mathbb{R}$, we have

$$\arctan(x) - x = -\int_0^x \frac{y^2}{1+y^2}\mathrm{d}y.$$

For $a \in \mathbb{Z}^d$, this yields

$$2\arctan\left(\frac{\tau|a|^2}{2}\right) - \tau|a|^2 = -2\int_0^{\tau|a|^2/2} \frac{y^2}{1+y^2}\mathrm{d}y.$$

Let $\gamma \in [0, 2]$; it is clear that for all $y \in \mathbb{R}$,

$$\frac{y^2}{1+y^2} \le y^\gamma.$$

Hence we have for all $a \in \mathbb{Z}^d$,

$$\left| 2 \arctan\left(\frac{\tau|a|^2}{2}\right) - \tau|a|^2 \right| \le 2 \int_0^{\tau|a|^2/2} y^\gamma \mathrm{d}y \le C\tau^{\gamma+1}|a|^{2\gamma+2}, \tag{IV.15}$$

for some constant C independent of a. Hence, owing to the fact that $|e^{ix} - e^{iy}| \le |x-y|$ for real x and y,

$$\left|\exp\left(-i\tau|a|^2\right) - \exp\left(-2i \arctan(\tau|a|^2/2)\right)\right| \le C\tau^{\gamma+1}|a|^{2\gamma+2}.$$

Hence we get for all z,

$$\left\|\varphi_T^\tau(z) - \varphi_{A_0}^1(z)\right\|_{\ell_s^1} \le C\tau^{\gamma+1} \|z\|_{\ell^1_{s+2\gamma+2}}. \tag{IV.16}$$

Combining the results of the previous Section, we get: There exists τ_0 such that for $\tau \le \tau_0$, $\varphi_P^\tau(z) \in B_{2M}^{s+4}$. Using thus (IV.16) with $\gamma = 1$, we obtain

$$\left\|\varphi_T^\tau \circ \varphi_P^\tau(z) - \varphi_{A_0}^1 \circ \varphi_P^\tau(z)\right\|_{\ell_s^1} \le C\tau^2.$$

The equation (IV.7) then yields the result. ■

In [15], different other choices for the filter function $\beta(x)$ are studied. In particular the cases where

$$\beta(x) = \tau^\nu \arctan\left(\tau^{-\nu}x\right), \quad \text{and} \quad \beta(x) = \frac{x + x^2/\tau^\nu}{1 + x/\tau^\nu + x^2/\tau^{2\nu}}$$

for $1 > \nu \ge 0$. These functions induce a slightly stronger smoothing in the high frequencies which helps the construction of the modified energy made in Chapter VI, but requires more regularity of the initial solution to obtain convergence results over finite time. We refer to [15] for an extensive discussion of these generalized cases. In the rest of this book, we will only focus on the two cases (i) and (ii) described above.

Remark IV.9. The previous proposition, in combination with the proof of Proposition IV.6, show the convergence of the implicit-explicit splitting scheme associated with the filter function $\beta(x) = 2\arctan(x/2)$. Note that the smoothness required for the exact solution is higher than for the exact splitting scheme: $s+4$ for the implicit-explicit integrator instead of $s+2$ for the exact splitting, to obtain the convergence in ℓ_s^1.

4 Space approximation

With the same kind of technics as the ones used in Section 2, we would like to prove now the convergence of the semi-discrete flow – as defined in the last section of Chapter III – towards the exact solution $\varphi_H^t(z)$. As for the splitting methods studied above, we need some smoothness assumption for the exact solution to obtain the convergence estimates. As in the end of Chapter III, we only consider the case where the dimension $d = 1$.

Let $P \in \mathcal{P}_k$ for some $k \geq 3$, and let P^K a family of discrete Hamiltonian in the space $\mathcal{P}_{k,p}^K$ (see Definition III.16). Here p is a fixed integer. We recall that for a fixed K, P^K acts on the finite dimensional space $\mathcal{A}^K$ defined in (III.51) and made of sequences z_j with $j \in \mathcal{B}^K$ (see (III.41)) the finite set of indices $(a, \delta) \in B^K \times \{\pm 1\}$ where B^K depends on the parity of K, and is defined in (III.35). Of course we have $\mathcal{A}^K \subset \ell_s^1$ for all s. We make the following assumptions:

Hypothesis IV.10. *There exists a constant C_0 such that*

$$\forall K \in \mathbb{N}, \qquad \left\| P^K \right\| \leq C_0 \left\| P \right\| \tag{IV.17}$$

where the first norm is defined in (III.43) and the second is the norm (III.16). Moreover, if we denote by

$$P(z) = \sum_{\ell=2}^{k} \sum_{\substack{\boldsymbol{j} \in \mathcal{Z}^\ell \\ \mathcal{M}(\boldsymbol{j})=0}} a_{\boldsymbol{j}} z_{\boldsymbol{j}}, \quad \text{and} \quad P^K(z) = \sum_{\ell=2}^{k} \sum_{|m| \leq p} \sum_{\substack{\boldsymbol{j} \in (\mathcal{B}^K)^\ell \\ \mathcal{M}(\boldsymbol{j})=mK}} a_{\boldsymbol{j}}^m z_{\boldsymbol{j}} \tag{IV.18}$$

the expressions of $P(z)$ and $P^K(z)$ in terms of their coefficients (see (III.42)), then we have

$$\forall \ell = 2, \dots, k, \quad \boldsymbol{j} \in \left(\mathcal{B}^K\right)^\ell \Longrightarrow a_{\boldsymbol{j}}^0 = a_{\boldsymbol{j}}. \tag{IV.19}$$

In other words, the coefficients of P^K corresponding to the indices with zero momentum coincide with the coefficients of P.

Example IV.11. In the case of the semi-discrete equation (III.39) obtained after space discretization of the cubic nonlinear Schrödinger equation, and with the definition of the norm, the previous estimate (IV.17) holds with the constant $C_0 = 3$ in dimension 1. We also easily see that the second condition (IV.19) is satisfied.

Lemma IV.12. *Assume that the polynomial P and the family P^K, $K \in \mathbb{N}$ satisfy the hypothesis IV.10, and let $s > 0$. Assume that $\|z\|_{\ell_s^1} \leq M$, then we have*

$$\|X_P(z) - X_{P^K}(z)\|_{\ell^1} \leq CK^{-s}, \tag{IV.20}$$

where the constant C only depends on M, k, p and s.

Proof. Let $P^K_{\ell,m}$ denote the component of degree ℓ associated with the index $m \neq 0$ in the decomposition (IV.18). We thus can write

$$\begin{aligned} P(z) - P^K(z) &= Q^K_1(z) - Q^K_2(z) \\ &= \sum_{\ell=2}^{k} \sum_{\substack{\boldsymbol{j} \in \mathcal{Z}^\ell \setminus (\mathcal{B}^K)^\ell \\ \mathcal{M}(\boldsymbol{j})=0}} a_{\boldsymbol{j}} z_{\boldsymbol{j}} - \sum_{\ell=2}^{k} \sum_{\substack{|m| \leq p \\ m \neq 0}} P^K_{\ell,m}(z). \end{aligned} \qquad \text{(IV.21)}$$

Using the same calculation as for equation (III.48) we have for $j = (a, \epsilon) \in \mathcal{B}^K$,

$$\left| \frac{\partial P^K_{\ell,m}}{\partial z_j} \right| \leq \ell \left\| P^K_{\ell,m} \right\| \sum_{\substack{\boldsymbol{j} \in (\mathcal{B}^K)^{\ell-1} \\ \mathcal{M}(\boldsymbol{j}) = -\epsilon a + mK}} \left| z_{j_1} \cdots z_{j_{\ell-1}} \right|,$$

but now in the decomposition, we have $m \neq 0$ and hence by definition of B^K, we have $|-a\epsilon + mK| \geq K/2$. Thus we easily see that there is always an index j_i such that $|j_i| \geq \frac{K}{4\ell}$. The previous equation thus yields

$$\begin{aligned} \left| \frac{\partial P^K_{\ell,m}}{\partial z_j} \right| &\leq \ell \left\| P^K_{\ell,m} \right\| \sum_{\substack{\boldsymbol{j} \in (\mathcal{B}^K)^{\ell-1} \\ \mathcal{M}(\boldsymbol{j}) = -\epsilon a + mK}} \frac{1}{|j_1|^s \cdots |j_{\ell-1}|^s} |j_1|^s |z_{j_1}| \cdots |j_{\ell-1}|^s |z_{j_{\ell-1}}| \\ &\leq 4^s K^{-s} \ell^{s+1} \left\| P^K_{\ell,m} \right\| \sum_{\substack{\boldsymbol{j} \in (\mathcal{B}^K)^{\ell-1} \\ \mathcal{M}(\boldsymbol{j}) = -\epsilon a + mK}} |j_1|^s |z_{j_1}| \cdots |j_{\ell-1}|^s |z_{j_{\ell-1}}|. \end{aligned}$$

Therefore, after summing in a and ϵ we get (compare (III.49))

$$\left\| X_{P^K_{\ell,m}}(z) \right\|_{\ell^1} \leq (4\ell)^{s+1} K^{-s} \left\| P^K_{\ell,m} \right\| \|z\|^{\ell-1}_{\ell^1_s},$$

and this shows with the notation (IV.21) that $\left\| X_{Q^K_2}(z) \right\|_{\ell^1} \leq CK^{-s}$ under the assumption $z \in B^s_M$, and with a constant C depending on ℓ, p, s and M.
Considering now $Q^K_1(z)$, we observe that for a multi-index $\boldsymbol{j} \in \mathcal{Z}^\ell \setminus (\mathcal{B}^K)^\ell$, there exists at least one index j_i such that $|j_i| \geq K/4$. We conclude as before that $\left\| X_{Q^K_1}(z) \right\|_{\ell^1} \leq CK^{-s}$, which finishes the proof. ■

Let us now consider an initial data function $u^0(x) = \sum_{a \in \mathbb{Z}} \xi^0_a e^{iax}$. We denote by $\xi^{K,0} = (\xi^K_a)_{a \in B^K} \in \mathbb{C}^K$ the complex number defined by

$$\xi^{K,0} = \mathcal{F}_K^{-1} \circ \operatorname{diag}\left(u^0(x_b)\right),$$

which represent the initial data after discrete Fourier transform. Using the aliasing formula (III.38), we have

$$\begin{aligned}\xi_a^{K,0} &= \frac{1}{K}\sum_{b\in B^K} e^{-iax_b}u^0(x_b)\\ &= \frac{1}{K}\sum_{b\in B^K}\sum_{c\in\mathbb{Z}} e^{i(c-a)x_b}\xi_c^0 = \sum_{m\in\mathbb{Z}}\xi_{a+mK}^0.\end{aligned} \tag{IV.22}$$

Denoting by $z^{K,0}$ the vector $z^{K,0} = (\xi^{K,0},\bar{\xi}^{K,0}) \in \mathbb{C}^K\times\mathbb{C}^K$, the equation above shows that $z^{K,0}\in\ell^1$ satisfies $\left\|z^{K,0}\right\|_{\ell^1} \le \left\|z^0\right\|_{\ell^1}$ where $z^0 = (\xi^0,\bar{\xi}^0)\in\mathbb{C}^{\mathbb{Z}}\times\mathbb{C}^{\mathbb{Z}}$ is associated with the function u^0.

Lemma IV.13. *With the previous notation, then if $z^0\in\ell^1_s$ we have*

$$\left\|z^0 - z^{K,0}\right\|_{\ell^1} \le CK^{-s}\left\|z^0\right\|_{\ell^1_s}. \tag{IV.23}$$

Proof. Recall that $\xi_a^{K,0}$ is a finite dimensional vector with indices in B^K. We have

$$\left\|z^0 - z^{0,K}\right\|_{\ell^1} \le 2\sum_{a\notin B^K}|\xi_a^0| + 2\sum_{a\in B^K}\sum_{\substack{m\in\mathbb{Z}\\ m\neq 0}}|\xi_{a+mK}^0|.$$

In the first term of the right-hand side of the previous equation, we have for the first $|a|\ge|K/2-1|$ and hence there exists a constant C such that

$$\sum_{a\notin B^K}|\xi_a^0| \le \frac{C}{K^{-s}}\sum_{a\notin B^K}|a|^s|\xi_a^0| \le CK^{-s}\left\|z^0\right\|_{\ell^1_s}.$$

For the second term, we observe that the indices $a+mK$ with $|m|\ge 1$ and $a\in B^K$ satisfy $|a+mK|\ge K/2$ and we conclude with a similar estimate. ■

We are now ready to prove the following

Proposition IV.14. *Let s and M be fixed, and $z^0\in\ell^1_s$. Assume that $z(t) = \varphi_H^t(z^0)$ is well defined for $t\in(0,t_*)$ and remains in a ball B_M^s. Let $z^{K,0}$ the discrete initial data defined by (IV.22), and let $z^K(t) := \varphi_{H^K}^t(z^{K,0})$ be the solution of the (finite dimensional) Hamiltonian system associated with the discrete Hamiltonian $H^K = T^K + P^K$ where $T^K = \sum_{a\in B^K}|a|^2\xi_a\eta_a$ and where the family P^K, $K\in\mathbb{N}$ satisfies the Hypothesis IV.10. Then there exist constants C and K_0 depending on M, s and t_* such that for $K\ge K_0$,*

$$\forall t\in(0,t_*),\quad \left\|z(t) - z^K(t)\right\|_{\ell^1} \le CK^{-s}. \tag{IV.24}$$

Proof. By definition, the exact flow $z(t) = \varphi_H^t(z)$ satisfies, for $t \in (0, t_*)$,

$$z(t) = \varphi_T^t\left(z^0\right) + \int_0^t \varphi_T^{t-\sigma} \circ X_P\left(z(\sigma)\right) \mathrm{d}\sigma.$$

Using the fact that for $z \in \mathcal{A}^K$ (see (III.51)) we have $T^K(z) = T(z)$, the flow $\varphi_{H^K}^t$ satisfies

$$z^K(t) = \varphi_T^t\left(z^{K,0}\right) + \int_0^t \varphi_T^{t-\sigma} \circ X_{P^K}\left(z^K(\sigma)\right) \mathrm{d}\sigma.$$

Hence we obtain, using the fact that φ_T^t is a linear isometry of ℓ^1,

$$\begin{aligned}\left\|z(t) - z^K(t)\right\|_{\ell^1} \leq \left\|z^0 - z^{K,0}\right\|_{\ell_1} &+ \int_0^t \|X_P(z(\sigma)) - X_{P^K}(z(\sigma))\|_{\ell^1} \mathrm{d}\sigma \\ &+ \int_0^t \left\|X_{P^K}(z(\sigma)) - X_{P^K}(z^K(\sigma))\right\|_{\ell^1} \mathrm{d}\sigma.\end{aligned}$$

Now using (III.46), we see that as long as $z^K(t) \in B_{2M}^0$ the ball of radius $2M$ in ℓ^1, then we can write with (IV.20) and the previous lemma

$$\forall t \in (0, t_*), \quad \left\|z(t) - z^K(t)\right\|_{\ell^1} \leq C_1 K^{-s} + \int_0^t C_2 \left\|z(\sigma) - z^K(\sigma)\right\|_{\ell^1} \mathrm{d}\sigma$$

where C_1 depends on t_* and M and C_2 on $\left\|P^K\right\|$ which is bounded independently of K. Using Gronwall's lemma, we get

$$\forall t \in (0, t_*), \quad \left\|z(t) - z^K(t)\right\|_{\ell^1} \leq C_1 K^{-s} e^{C_2 t_*}$$

and we conclude using the same bootstrap argument as before: As long as $z^K(t) \in B_{2M}^0$ the previous estimate holds, and hence if $K \geq K_0$ is sufficiently large, we have $\left\|z^K(t)\right\|_{\ell^1} \leq 2M$ for $t \in (0, t_*)$. This shows (IV.24). ■

5 Fully discrete splitting method

We consider now a full discretization of a Hamiltonian PDE associated with a Hamiltonian of the form $H(z) = T(z) + P(z)$ as studied before. The numerical solution is obtained using a time discretization of the semi-discrete flow $\varphi_{H^K}^t$ by a splitting method, where $H^K = T^K + P^K$ is a discrete approximation of H, as in the previous section.

We prove the convergence in ℓ^1 of the fully discrete numerical solution towards the exact solution, provided the exact solution remains bounded in ℓ_s^1 with $s \geq 2$ (i.e. at least 2 derivatives in the Wiener algebra). The goal will be here to obtain explicit bounds in term of the discretization parameter τ and K.

Lemma IV.15. *Let $z^0 \in \ell^1_{s+2}$, $M > 0$ and $t_* > 0$. Assume that for all $t \in (0, t_*)$, $\varphi^t_H(z^0)$ is well defined in ℓ^1_{s+2} and remains in the ball B^{s+2}_M. Then there exist constants C and τ_0 such that for $\tau \le \tau_0$, if z^n is the sequence defined by*

$$z^{n+1} = \varphi^\tau_T \circ \varphi^\tau_P(z^n), \quad n = 0, \dots, t_*/\tau, \tag{IV.25}$$

then we have

$$\forall n\tau \le t_*, \quad \|z^n\|_{\ell^1_s} \le 2M.$$

Proof. This is a straightforward Corollary of Proposition IV.6. ■

Lemma IV.16. *Let $k \in \mathbb{N}$, $s \ge 0$, $P \in \mathcal{P}_k$ and P^K, $K > 0$ a collection of Hamiltonians satisfying Hypothesis IV.10. Let $M > 0$ and $z \in \ell^1_s$ such that $z \in B^s_M$. Then there exists constants C and τ_0 such that for $\tau \le \tau_0$ we have*

$$\left\|\varphi^\tau_P(z) - \varphi^\tau_{P^K}(z)\right\|_{\ell^1} \le C\tau K^{-s}.$$

Proof. For $t \in (0, \tau)$, we set $z(t) = \varphi^t_P(z)$ and $z^K(t) = \varphi^t_{P^K}(z)$. We have by definition

$$z(t) = z + \int_0^t X_P(z(\sigma))\, \mathrm{d}\sigma$$

and a similar formula for $z^K(t)$. Hence we can write

$$\begin{aligned}\left\|z(t) - z^K(t)\right\|_{\ell^1} &\le \int_0^t \left\|X_P(z(\sigma)) - X_{P^K}(z^K(\sigma))\right\|_{\ell^1} \mathrm{d}\sigma \\ &\le \int_0^t \|X_P(z(\sigma)) - X_{P^K}(z(\sigma))\|_{\ell^1} \mathrm{d}\sigma \\ &\quad + \int_0^t \left\|X_{P^K}(z(\sigma)) - X_{P^K}(z^K(\sigma))\right\|_{\ell^1} \mathrm{d}\sigma. \end{aligned} \tag{IV.26}$$

Now we can assume that τ_0 is small enough to have $z(\sigma) \in B^s_{2M}$ for $\sigma \in (0, \tau_0)$. Using (IV.20), this shows that

$$\|X_P(z(\sigma)) - X_{P^K}(z(\sigma))\|_{\ell^1} \le CK^{-s}$$

with a constant C uniform in $\sigma \in (0, \tau_0)$.
To deal with the term (IV.26), we observe that the relation (III.46) ensures that X_{P^K} is Lipschitz on bounded sets of ℓ^1 with a constant independent of K. This implies that we can assume that $z^K(\sigma) \in B^0_M$ for $\sigma \in (0, \tau_0)$, and that there exists a constant L such that for all $t \in (0, \tau)$,

$$\left\|z(t) - z^K(t)\right\|_{\ell^1} \le C\tau K^{-s} + \int_0^t L \left\|z(\sigma) - z^K(\sigma)\right\|_{\ell^1} \mathrm{d}\sigma.$$

This shows the result. ■

Theorem IV.17. *Let $k \in \mathbb{N}$, $s \geq 0$, $P \in \mathcal{P}_k$ and P^K, $K > 0$ a collection of Hamiltonians satisfying Hypothesis IV.10.*

Let $z^0 \in \ell^1_{s+2}$, $M > 0$ and $t_* > 0$. Assume that for all $t \in (0, t_*)$, $\varphi^t_H(z^0)$ is well defined in ℓ^1_{s+2} and remains in the ball B^{s+2}_M. Then there exist constants C, K_0 and τ_0 such that for all $\tau \leq \tau_0$ and $K \geq K_0$, the sequence $z^{K,n}$ defined by induction:

$$z^{K,n+1} = \varphi^\tau_{T^K} \circ \varphi^\tau_{P^K}\left(z^{K,n}\right), \quad n = 0, \ldots, t_*/\tau,$$

with initial value $z^{K,0}$ defined as in (IV.22) satisfies

$$\forall t = n\tau \leq t_*, \quad \left\| z(t) - z^{K,n} \right\|_{\ell^1} \leq C\left(\tau + K^{-s}\right).$$

Proof. Let z^n be the sequence defined by (IV.25). Using Lemma IV.15, we have $z^n \in B^s_{2M}$ for all $n \leq t_*/\tau$.

As T leaves invariant the space $\mathcal{A}^K$, we have that for all n, $z^{K,n+1} = \varphi^\tau_T \circ \varphi^\tau_{P^K}(z^{K,n})$, while $z^{n+1} = \varphi^\tau_T \circ \varphi^\tau_P(z^n)$. Hence we calculate that

$$\begin{aligned}\left\| z^{n+1} - z^{K,n+1} \right\|_{\ell^1} &= \left\| \varphi^\tau_P(z^n) - \varphi^\tau_{P^K}(z^{K,n}) \right\|_{\ell^1} \\ &\leq \left\| \varphi^\tau_P(z^n) - \varphi^\tau_{P^K}(z^n) \right\|_{\ell^1} + \left\| \varphi^\tau_{P^K}(z^n) - \varphi^\tau_{P^K}(z^{K,n}) \right\|_{\ell^1}.\end{aligned}$$

Using Lemma IV.16 the first term in this equation is bounded by $C\tau K^{-s}$.

To deal with the second term, we observe that as long as $\left\| z^{K,n} \right\|_{\ell^1} \leq 2M$, we have

$$\left\| \varphi^\tau_{P^K}(z^n) - \varphi^\tau_{P^K}(z^{K,n}) \right\|_{\ell^1} \leq e^{L\tau} \left\| z^n - z^{K,n} \right\|_{\ell^1}$$

where L is the Lipschitz constant of P^K over the ball of radius $2M$ in ℓ^1. Note that L is actually independent of K using (III.46).

Hence as long as $\left\| z^{K,n} \right\|_{\ell^1} \leq 2M$ we can write

$$\left\| z^{n+1} - z^{K,n+1} \right\|_{\ell^1} \leq C\tau K^{-s} + e^{L\tau} \left\| z^n - z^{K,n} \right\|_{\ell^1}.$$

By induction we obtain

$$\left\| z^n - z^{K,n} \right\|_{\ell^1} \leq C t_* e^{Lt_*}(K^{-s} + \left\| z^0 - z^{K,0} \right\|_{\ell^1}) \leq C_* K^{-s}$$

where C_* depends on t_*, C and L. For $K \geq K_0$ sufficiently large, this shows that the previous relation holds for $n\tau \leq t_*$. We conclude by gathering this estimate with the result of Proposition IV.6. ■

V Modified energy in the linear case

In this chapter, we consider the *linear* Schrödinger equation

$$i\partial_t u(t,x) = -\Delta u(t,x) + V(x)u(t,x), \quad u(0,x) = u^0(x), \tag{V.1}$$

set on the d-dimensional torus $\mathbb{T}^d$, with initial condition u^0 and smooth potential function $V(x) \in \mathbb{R}$. We consider splitting methods induced by the natural decomposition between the kinetic energy represented by the Laplace operator $-\Delta$ and the potential energy associated with $V(x)$. Such schemes are convergent in the sense of the previous chapter: they yield convergent approximations over finite time intervals if the exact solution is smooth. We will not give the details here, as the analysis is similar to the one performed in Chapter IV, but we refer to [29] for a complete analysis.

In this chapter, we prove backward error analysis results in the sense of Chapter II: we construct a modified energy for the numerical scheme and prove that the numerical flow can be interpreted as the exact flow of this modified energy. Such a result holds true without any further assumption in the case of implicit-explicit integrators where the solution of the free Schrödinger equation is approximated by the midpoint rule. For the classical splitting scheme, the existence of a modified energy relies on the use of a CFL condition. The presentation here roughly follows the lines of [12].

We then consider the case of fully discrete splitting schemes and show the existence of a modified energy under a CFL condition. Using the preservation of this modified energy, we then give some long time control of the regularity of the numerical solution.

1 Operators, flow and splitting methods

In this section, we would like to introduce a specific framework to perform the analysis of the linear case. In contrast to the nonlinear case, the Hamiltonian functions associated with linear equations are always quadratic in $(u, \bar{u})$. But as the potential depends on x, the *zero momentum* condition is not satisfied. To deal with these quadratic Hamiltonians and the associated operators, we introduce here some operator spaces in which we will be able to construct the modified energy.

1.1 Operators. In the linear situation, the Hamiltonian function associated with the previous equation is quadratic in $(u, \bar{u})$: The Hamiltonian function $H = T + P$ is written

$$H(u,\bar{u}) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} |\nabla u(x)|^2 + V(x)|u(x)|^2 \, \mathrm{d}x.$$

As in the previous chapters, we decompose u and $\bar{u}$ in Fourier series with the notation

$$u(x) = \sum_{a\in\mathbb{Z}^d} \xi_a\, e^{ia\cdot x} \quad \text{and} \quad \bar{u}(x) = \sum_{a\in\mathbb{Z}^d} \eta_a\, e^{-ia\cdot x}.$$

Then if $V(x) = \sum_{a\in\mathbb{Z}^d} V_a\, e^{ia\cdot x}$, we have

$$\begin{aligned} H(u,\bar{u}) &= \sum_{a\in\mathbb{Z}^d} |a|^2\xi_a\eta_a + \sum_{a_1-a_2+a_3=0} V_{a_3}\xi_{a_1}\eta_{a_2} \\ &= \sum_{a\in\mathbb{Z}^d} |a|^2\xi_a\eta_a + \sum_{a,b\in\mathbb{Z}^d} V_{a-b}\xi_b\eta_a, \end{aligned}$$

where we recall that $|a|^2 = (a^1)^2 + \cdots + (a^d)^2$ if $a = (a^1,\ldots,a^d) \in \mathbb{Z}^d$. In Fourier variables, the linear equation (V.1) can be written using the formalism of the previous chapters

$$i\dot{\xi}_a = |a|^2\xi_a + \sum_{b\in\mathbb{Z}^d} V_{a-b}\xi_b = \frac{\partial H}{\partial\eta_a}(\xi,\eta).$$

We observe here that in the equation above, the term depending on the potential does not satisfy the *zero momentum* condition as in the nonlinear case (see (III.14)). Hence in our framework, we cannot see the linear situation as a particular case of the nonlinear PDEs studied above, where the nonlinearity does not depend on x. However, if V is smooth, the term V_{a-b} decays with respect to the momentum $a-b$. A similar property holds for nonlinear Hamiltonians depending smoothly on x. Note that we could have considered a very general situation encompassing both the linear and nonlinear cases, but this would have led to many more technical difficulties (see Proposition III.6 where only the zero momentum case is studied).

To measure the decay of the operators with respect to the diagonal level $|b-a|$, we introduce the following operator norm:

Definition V.1. *An operator A is an element $A = (A_{ab})_{a,b\in\mathbb{Z}^d}$ acting as a linear map in the Fourier space $\mathbb{C}^{\mathbb{Z}^d}$. For $\alpha > 1$ we define the norm*

$$\|A\|_\alpha = \sup_{a,b} |A_{ab}|\,(1+|a-b|^\alpha)\,.$$

We say that A is symmetric if $\overline{A_{ab}} = A_{ba}$, and we write

$$\mathcal{L}_\alpha = \{A = (A_{ab})_{a,b\in\mathbb{Z}^d} \quad \textit{symmetric} \quad |\, \|A\|_\alpha < \infty\}.$$

With a real function $W(x)$ we associate the operator $W = (W_{ab})_{a,b\in\mathbb{Z}^d}$ with components $W_{ab} = W_{a-b}$ where W_a denotes the Fourier coefficient of W associated with $a \in \mathbb{Z}^d$. Thus the operator $(W_{ab})_{a,b\in\mathbb{Z}^d}$ acting in the Fourier space corresponds to multiplication by W and we see that if the function W belongs to the Sobolev space H^s for some $s \geq 0$, then the operator $W \in \mathcal{L}_s$. Note moreover that with this identification, $\|W\|_\alpha < \infty$ with $\alpha > d$ implies that $\|W\|_{L^\infty} < \infty$.

Let $A \in \mathcal{L}_\alpha$ be a (symmetric) operator. Then we set

$$\langle u|A|u\rangle = \sum_{a,b\in\mathbb{Z}^d} \bar{\xi}_a A_{ab}\xi_b \in \mathbb{R}.$$

We thus see that the Hamiltonian energy $H(u,\bar{u})$ can be written

$$H(u,\bar{u}) = \langle u| - \Delta - V|u\rangle.$$

Finally, for two operators A and B, we set

$$\mathrm{ad}_A(B) = AB - BA,$$

where the product of two operators is defined by the formula

$$\forall a,b \in \mathbb{Z}^d, \quad (AB)_{ab} = \sum_{c\in\mathbb{Z}^d} A_{ac}B_{cb}.$$

Lemma V.2. *Assume that $\alpha > d$. There exists a constant C_α such that for all operators A and B,*

$$\|AB\|_\alpha \leq C_\alpha \|A\|_\alpha \|B\|_\alpha .$$

Proof. We have for $a,b \in \mathbb{Z}^d$,

$$\begin{aligned}|(AB)_{ab}|(1+|a-b|^\alpha) &\leq (1+|a-b|^\alpha)\sum_{c\in\mathbb{Z}^d} |A_{ac}||B_{cb}|\\ &\leq \|A\|_\alpha \|B\|_\alpha \sum_{c\in\mathbb{Z}^d} \frac{1+|a-b|^\alpha}{(1+|a-c|^\alpha)(1+|c-b|^\alpha)}.\end{aligned}$$

As the function $x \to x^\alpha$ is convex for $x > 0$, we have

$$1+|a-b|^\alpha \leq 1+(|a-c|+|c-b|)^\alpha \leq 2^{\alpha-1}\left(1+|a-c|^\alpha+1+|c-b|^\alpha\right).$$

Hence we have

$$|(AB)_{ab}|(1+|a-b|^\alpha) \leq 2^{\alpha-1}\|A\|_\alpha \|B\|_\alpha \sum_{c\in\mathbb{Z}^d}\left(\frac{1}{1+|c-b|^\alpha}+\frac{1}{1+|a-c|^\alpha}\right)$$

and this shows the result, the condition $\alpha > d$ ensuring the convergence of the series. ■

Lemma V.3. *Let $\alpha > d$. There exists a constant M_α such that for all symmetric operators B and for all $u \in L^2$, we have*

$$|\langle u|B|u\rangle| \leq M_\alpha \|B\|_\alpha \|u\|_{L^2}^2.$$

Proof. We have

$$\begin{aligned} |\langle u|B|\,u\rangle| &\leq \sum_{a,b} |B_{ab}||\xi_a||\xi_b|, \\ &\leq \|B\|_\alpha \sum_{a,b} \frac{1}{1+|a-b|^\alpha}|\xi_a||\xi_b|, \\ &\leq \|B\|_\alpha \sum_{a,b} \frac{1}{1+|a-b|^\alpha}|\xi_a|^2, \end{aligned}$$

after using the formula $|\xi_a||\xi_b| \leq \frac{1}{2}(|\xi_a|^2 + |\xi_b|^2)$. This yields the result. ■

1.2 Linear flow. When $V = 0$, we can define the solution of the free Schrödinger equation as the flow φ_T^t of the previous chapter. Here we denote this flow as $\exp(it\Delta)$ defined in Fourier series by the formula

$$\forall a \in \mathbb{Z}^d, \quad \xi_a(t) = \exp(-it|a|^2)\xi_a(0)$$

if $u(t) = \sum_{a\in\mathbb{Z}^d} \xi_a(t)e^{ia\cdot x} = \exp(it\Delta)u(0)$. Equipped with the previous lemmas, we can prove the existence and uniqueness of global mild solutions to the linear equation (V.1).

Theorem V.4. *Assume that $V \in \mathcal{L}_\alpha$ with $\alpha > d$, and assume that $u^0 \in L^2$. Then there exists a unique solution $u(t,x)$ in L^2 satisfying for all $t \in \mathbb{R}$,*

$$u(t,x) = e^{it\Delta}u^0(x) + \int_0^t e^{i(t-\sigma)\Delta}V(x)u(\sigma,x)\,d\sigma.$$

Proof. The argument is the same as in the proof of Theorem III.7. The key here is that the mapping

$$u(x) \mapsto V(x)u(x)$$

is globally Lipschitz from L^2 to itself, which is a consequence of the fact that for $\alpha > d$, $V \in L^\infty$. ■

In the following, we denote this solution by

$$u(t,x) = \exp(it(\Delta - V))u^0(x).$$

1.3 Splitting methods. Note that as V is real, the solution of the potential equation

$$i\partial_t u(t,x) = V(x)u(t,x), \quad u(0,x) = u^0(x),$$

is directly given by the formula

$$\forall x \in \mathbb{T}^d, \quad u(t,x) = \exp(-itV(x))u^0(x).$$

The splitting methods studied in the previous chapter can be written, for a small time step $\tau > 0$,

$$\exp(i\tau(\Delta - V)) \simeq \exp(-i\tau V)\exp(-iA_0) = \varphi_V^\tau \circ \varphi_{A_0}^1, \tag{V.2}$$

where $A_0 = \beta(-\tau\Delta)$ is the operator associated with a filter function β (see (IV.9) and (IV.11)). In terms of Fourier coefficients, we have for all $a \in \mathbb{Z}^d$, $\xi_a^1 = \exp(-i\beta(\tau|a|^2))\xi_a^0$. As in the previous chapter we will mainly consider the cases where

$$\beta(x) = 2\arctan(x/2)\mathbb{1}_{x\leq c_0} \quad \text{and} \quad \beta(x) = x\mathbb{1}_{x\leq c_0}$$

where c_0 is a given CFL number (possibly infinite).

Note that we have $A_0 \in \mathcal{L}_\alpha$ for all $\alpha > 0$ without restriction on c_0 in the case of the implicit-explicit integrator, and as soon as $c_0 < \infty$ in the case of the classical splitting.

2 Formal series

Following the principle of backward error analysis, we try to find an operator $Z(\tau)$ in some $\mathcal{L}_\alpha$ space such that

$$\exp(-i\tau V)\exp(-iA_0) = \exp(-iZ(\tau)).$$

To do this, the standard method inspired by the finite dimensional case consists in expanding this expression in powers of τ and determining $Z(\tau)$ by solving a differential equation obtained by the use of BCH-like formulas (see (II.19)). However the successive derivatives of $A_0 = -\beta(\tau\Delta)$ with respect to τ yield unbounded operators, which makes such an equation ill-posed on $\mathcal{L}_\alpha$.

To remedy this difficulty, the strategy is the following: we consider the operator A_0 as fixed, and we search for a function $t \to Z(t)$ taking values in some $\mathcal{L}_\alpha$ space, such that $Z(0) = A_0$ and

$$\forall t \in [0,\tau], \quad \exp(-itV)\exp(-iA_0) = \exp(-iZ(t)). \tag{V.3}$$

If such an operator can be found, then setting $t = \tau$ in the previous equation will yield the result.

Taking the derivative of the expression (V.3) with respect to t, we obtain using (II.15),

$$-iV\exp(-itV)\exp(-iA_0) = -i\left[\left(\frac{\exp\left(\mathrm{ad}_{-iZ(t)}\right)-1}{\mathrm{ad}_{-iZ(t)}}\right)\frac{\mathrm{d}Z(t)}{\mathrm{d}t}\right]\exp(-iZ(t)).$$

Hence using (II.17), $Z(t)$ has to satisfy the differential equation

$$Z'(t) = \sum_{k\geq 0}\frac{B_k}{k!}(-1)^k\mathrm{ad}^k_{iZ(t)}(V). \tag{V.4}$$

Recall that here, the B_k are the Bernoulli numbers defined by the relation (II.16).

We define the formal series

$$Z(t) = \sum_{\ell\geq 0} t^\ell Z_\ell,$$

where

$$Z_0 := A_0 = -\beta(\tau\Delta)$$

is the diagonal operator with coefficients

$$\lambda_a = (A_0)_{aa} = \beta(\tau|a|^2), \tag{V.5}$$

and where Z_ℓ, $\ell \geq 1$, are unknown operators.

Plugging this expression into (V.4) we find

$$\begin{aligned}\sum_{\ell\geq 1}\ell t^{\ell-1}Z_\ell &= \sum_{k\geq 0}\frac{B_k}{k!}\left(-i\sum_{\ell\geq 0}t^\ell\mathrm{ad}_{Z_\ell}\right)^k(V)\\ &= \sum_{\ell\geq 0}t^\ell\sum_{k\geq 0}\frac{B_k}{k!}(-i)^k\sum_{\ell_1+\cdots+\ell_k=\ell}\mathrm{ad}_{Z_{\ell_1}}\cdots\mathrm{ad}_{Z_{\ell_k}}(V).\end{aligned}$$

Identifying the coefficients in the formal series in powers of t, we find the induction formula:

$$\forall\ell\geq 1,\quad (\ell+1)Z_{\ell+1} = \sum_{k\geq 0}\frac{B_k}{k!}(-i)^k\sum_{\ell_1+\cdots+\ell_k=\ell}\mathrm{ad}_{Z_{\ell_1}}\cdots\mathrm{ad}_{Z_{\ell_k}}(V). \tag{V.6}$$

Note that we easily show by induction that if they are defined, then for all ℓ, Z_ℓ is symmetric. For $\ell = 1$, this equation yields

$$Z_1 = \sum_{k\geq 0}\frac{B_k}{k!}(-i)^k\mathrm{ad}^k_{A_0}(V). \tag{V.7}$$

Note that the main difference with the finite dimensional situation is that the “first” term in the expansion is given by an infinite series and that it depends on the small parameter τ through the operator $Z_0 = A_0$. The key to controlling this term is to estimate the norm of the operator ad_{A_0}.

3 Analytic estimates

Lemma V.5. *Let A_0 be the diagonal operator with eigenvalues $\lambda_a = \beta(\tau|a|^2)$, and assume that*

$$\forall a \in \mathbb{Z}^d, \quad 0 \le \lambda_a \le \pi. \tag{V.8}$$

Let $W = (W_{ab})_{a,b\in\mathbb{Z}^d}$ be an operator in $\mathcal{L}_\alpha$ for some $\alpha > 1$. Then we have

$$\|\mathrm{ad}_{A_0} W\|_\alpha \le \pi \, \|W\|_\alpha \,. \tag{V.9}$$

Proof. For $a, b \in \mathbb{Z}^d$ we have, as A_0 is diagonal,

$$(\mathrm{ad}_{A_0} W)_{ab} = (\lambda_a - \lambda_b) W_{ab}.$$

Hence we have for all $a, b \in \mathbb{Z}^d$,

$$\left|(\mathrm{ad}_{A_0} W)_{ab}\right| \le \pi |W_{ab}|$$

and this shows the result. ■

Remark V.6. The condition (V.8) will be fulfilled as soon as

$$\forall x > 0, \quad 0 \le \beta(x) \le \pi.$$

In the case where A_0 is associated with the filter function $\beta(x) = 2\arctan(x/2)$, this condition is automatically satisfied. In the case where β is of the form $\beta(x) = x\mathbb{1}_{x\le c_0}$ with a CFL number c_0, then this condition will be fulfilled as soon as $c_0 \le \pi$.

We are now ready to prove the main result of this chapter:

Theorem V.7. *Let $\alpha > d$, and assume that $\|V\|_\alpha < \infty$. Assume that the eigenvalues λ_a of the operator A_0 satisfy the hypothesis (V.8). Then there exist $\tau_0 > 0$ and a constant C such that for all $\tau \in (0, \tau_0)$, there exists a symmetric operator $S(\tau)$ such that*

$$\exp(i\tau V)\exp(-iA_0) = \exp(-i\tau S(\tau)).$$

Moreover we have

$$S(\tau) = -\frac{1}{\tau}\beta(\tau\Delta) + V(\tau) + \tau W(\tau)$$

where $V(\tau)$ and $W(\tau)$ satisfy,

$$\|V(\tau)\|_\alpha \le C\,\|V\|_\alpha \quad \text{and} \quad \|W(\tau)\|_\alpha \le C\,\|V\|_\alpha^2\,, \tag{V.10}$$

and where moreover $V(\tau)$ is given by the convergent series in $\mathcal{L}_\alpha$

$$V(\tau) = V + \sum_{k\geq 1} \frac{B_k}{k!}(-i)^k \mathrm{ad}_{A_0}^k(V), \tag{V.11}$$

where the B_k are the Bernoulli numbers.

Proof. Recall that the power series (II.16) defining the Bernoulli numbers has a radius of convergence equal to 2π.
Let us consider the equation (V.7). Using (V.9), we see that

$$\|Z_1\|_\alpha \leq \|V\|_\alpha \sum_{k\geq 0} \frac{|B_k|}{k!}\pi^k \leq C\,\|V\|_\alpha \tag{V.12}$$

is bounded. In terms of the components of the operator Z_1, we calculate using the expression of ad_{A_0} that

$$(Z_1)_{ab} = V_{ab}\frac{i(\lambda_a - \lambda_b)}{\exp(i(\lambda_a - \lambda_b)) - 1}. \tag{V.13}$$

Note that for any bounded operator A and B, we always have

$$\|\mathrm{ad}_A(B)\|_\alpha \leq 2C_\alpha\,\|A\|_\alpha\,\|B\|_\alpha$$

where C_α is given by Lemma V.2. We define now the following numbers:

$$\zeta_0 = \pi \quad\text{and}\quad \zeta_\ell = 2C_\alpha\,\|Z_\ell\|_\alpha\,, \quad\text{for}\quad \ell \geq 1.$$

Using (V.6) and Lemma V.5, we see that we have the estimates

$$\forall \ell \geq 1, \quad \frac{1}{2C_\alpha}(\ell+1)\zeta_{\ell+1} \leq \|V\|_\alpha \sum_{k\geq 0}\frac{|B_k|}{k!} \sum_{\ell_1+\cdots+\ell_k=\ell} \zeta_{\ell_1}\cdots\zeta_{\ell_k}.$$

Now for any ρ such that $\pi < \rho < 2\pi$, using Cauchy estimates, there exists a constant M such that for all k, $|B_k| \leq k!M\rho^{-k}$. Hence we can write

$$\forall \ell \geq 1, \quad \frac{1}{2C_\alpha}(\ell+1)\zeta_{\ell+1} \leq M\,\|V\|_\alpha \sum_{k\geq 0}\rho^{-k} \sum_{\ell_1+\cdots+\ell_k=\ell} \zeta_{\ell_1}\cdots\zeta_{\ell_k}.$$

Let $\zeta(t)$ be the formal series $\zeta(t) = \sum_{\ell\geq 0} t^\ell\zeta_\ell$. Multiplying the previous equation by t^ℓ and summing over $\ell \geq 0$, we find

$$\frac{1}{2C_\alpha}\zeta'(t) \leq M\,\|V\|_\alpha \sum_{k\geq 0}\rho^{-k}\zeta(t)^k = M\,\|V\|_\alpha \frac{1}{1-\zeta(t)/\rho}.$$

Let $\eta(t)$ be the solution of the differential equation:

$$\eta'(t) = 2MC_\alpha \|V\|_\alpha \frac{1}{1-\eta(t)/\rho}, \quad \eta(0) = \pi.$$

Taking $\rho = 3\pi/2$, we see that for $t \leq \frac{\pi}{48MC_\alpha\|V\|_\alpha}$, the solution can be written

$$\eta(t) = \frac{3\pi}{2}\left(1 - \sqrt{\frac{1}{9} - \frac{8}{3\pi}MC_\alpha \|V\|_\alpha t}\right),$$

and defines an analytic function of t. Expanding $\eta(t) = \sum_{\ell \geq 0} t^\ell \eta_\ell$, we see that the coefficients satisfy the relations $\eta_0 = \pi$ and

$$\forall \ell \geq 1, \quad \frac{1}{2C_\alpha}(\ell+1)\eta_{\ell+1} = M\|V\|_\alpha \sum_{k\geq 0} \rho^{-k} \sum_{\ell_1+\cdots+\ell_k=\ell} \eta_{\ell_1}\cdots\eta_{\ell_k}$$

with $\rho = \frac{3\pi}{2}$. By induction, this shows that $\zeta_\ell \leq \eta_\ell$. Moreover, for all $z \in \mathbb{C}$ with $|z| \leq \frac{\pi}{48MC_\alpha\|V\|_\alpha}$, we have that the coefficients ζ_ℓ are positive,

$$|\zeta(z)| = \left|\sum_{\ell=0}^{\infty} \zeta_\ell z^\ell\right| \leq \sum_{\ell=0}^{\infty} \zeta_\ell |z|^\ell = \zeta(|z|) \leq \eta(|z|) \leq \frac{3\pi}{2}.$$

Using Cauchy estimates, we see that

$$\forall \ell \geq 1, \quad \|Z_\ell\| = \frac{1}{2C_\alpha}\zeta_\ell = \frac{1}{2C_\alpha}\frac{\zeta^{(\ell)}(0)}{\ell!} \leq \frac{3\pi}{4C_\alpha}\left(\frac{48MC_\alpha\|V\|_\alpha}{\pi}\right)^\ell. \tag{V.14}$$

The theorem is now proved by setting

$$V(\tau) = Z_1, \quad \text{and} \quad W(\tau) = -\sum_{\ell \geq 2} \tau^{\ell-2} Z_\ell$$

which defines a convergent power series for $|\tau| < \tau_0 = \frac{\pi}{48MC_\alpha\|V\|_\alpha}$. The estimate (V.10) on $V(\tau)$ is then an easy consequence of (V.12). The estimate (V.10) on $W(\tau)$ is obtained from (V.14). ■

4 Properties of the modified equation

The following result shows that $S(\tau)$ given by the previous result defines a "modified" energy when applied to smooth functions.

Proposition V.8. *Let $\nu \in [0,1]$, and suppose that A_0 is associated with the filter function $\beta(x) = 2\arctan(x/2)$. Assume that $u \in H^{1+\nu}(\mathbb{T}^d)$, then we have for $\tau \in (0, \tau_0)$,*

$$|\langle u|S(\tau)|u\rangle - \langle u| -\Delta + V|u\rangle| \leq C\tau^\nu \|u\|^2_{H^{1+\nu}} \tag{V.15}$$

where C depends on ν and V.

Proof. Using (IV.15), we have for all $a \in \mathbb{Z}^d$,

$$\left|\frac{2}{\tau}\arctan\left(\frac{\tau|a|^2}{2}\right) - |a|^2\right| \leq \frac{2}{\tau}\int_0^{\tau|a|^2/2} y^\gamma \mathrm{d}y \leq C\tau^\gamma |a|^{2\gamma+2}.$$

This shows that for all v,

$$\left|\langle v| - \frac{2}{\tau}\arctan\left(\frac{\tau\Delta}{2}\right)|v\rangle - \langle v| - \Delta|v\rangle\right| \leq C\tau^\gamma \|v\|^2_{H^{1+\gamma}}. \tag{V.16}$$

Now we have

$$\langle v|V(\tau)|v\rangle - \langle v|V|v\rangle = \sum_{k\geq 1}\frac{B_k}{k!}\langle v|(-i)^k \mathrm{ad}^k_{A_0}(V)|v\rangle.$$

Recall that $A_0 = -2\arctan\left(\frac{\tau\Delta}{2}\right)$ is a positive operator. As $\nu \in [0,1]$, the operators A_0^ν and $A_0^{1-\nu}$ are hence well defined, and for an operator W we have in terms of the coefficients of the operators

$$\left(A_0^{1-\nu}W\right)_{ab} = \left(2\arctan\left(\frac{\tau|a|^2}{2}\right)\right)^{1-\nu} W_{ab}.$$

Hence we have for all $\alpha > 1$,

$$\left\|A_0^{1-\nu}W\right\|_\alpha \leq \pi^{1-\nu}\|W\|_\alpha \quad \text{and} \quad \left\|WA_0^{1-\nu}\right\|_\alpha \leq \pi^{1-\nu}\|W\|_\alpha.$$

Now using Lemma V.3 and the fact that A_0 is symmetric, we have for all v and all operators W,

$$\begin{aligned}|\langle v|\mathrm{ad}_{A_0}(W)|v\rangle| &\leq \left(\left\|A_0^{1-\nu}W\right\|_\alpha + \left\|WA_0^{1-\nu}\right\|_\alpha\right)\left\|A_0^\nu v\right\|_{L^2}\|v\|_{L^2}\\ &\leq 2\pi^{1-\nu}\|W\|_\alpha\left\|A_0^\nu v\right\|_{L^2}\|v\|_{L^2}.\end{aligned}$$

Hence we have

$$\begin{aligned}|\langle v|V(\tau)|v\rangle - \langle v|V|v\rangle| &\leq 2\sum_{k\geq 1}\frac{|B_k|}{k!}\pi^{k-\nu}\|V\|_\alpha\left\|A_0^\nu v\right\|_{L^2}\|v\|_{L^2}\\ &\leq C\|V\|_\alpha\left\|A_0^\nu v\right\|_{L^2}\|v\|_{L^2}.\end{aligned}$$

As for all $y \geq 0$ the relation $\arctan(y) \leq y$ holds, we have $\left\|A_0^\nu v\right\|_{L^2} \leq 2^\nu\tau^\nu\|v\|_{H^{2\nu}}$ and hence

$$|\langle v|V(\tau)|v\rangle - \langle v|V|v\rangle| \leq C\|V\|_\alpha\tau^\nu\|v\|_{H^{2\nu}}\|v\|_{L^2}.$$

Finally, we have using (V.10) that

$$|\langle v|W(\tau)|v\rangle| \leq C\|V\|^2_\alpha\,\tau\,\|v\|^2_{L^2}.$$

Summing the previous inequalities with $\gamma = \nu$ in (V.16) we have that

$$|\langle\, v|S(\tau)|v\rangle - \langle\, v| - \Delta + V|v\rangle| \leq C\tau^{\nu}\left(\|v\|^{2}_{H^{1+\nu}} + \|v\|_{H^{2\nu}}\|v\|_{L^2}\right)$$

for a constant C depending on V and ν. As $\|v\|_{H^{2\nu}} \leq \|v\|_{H^{1+\nu}}$ for $\nu \in [0,1]$ this yields the result. ■

The next result shows the conservation of the modified energy $S(\tau)$ along the (semi-discrete) numerical solution associated with the splitting propagator. As a consequence, we give a regularity bound for the numerical solution over arbitrary long time periods.

Corollary V.9. *Assume that $u^0 \in L^2(\mathbb{T}^d)$ and $\tau \in (0,\tau_0)$ given in Theorem V.7. For all $n \geq 1$, we define*

$$u^n = (\exp(-i\tau V)\exp(-iA_0))^n\, u^0.$$

Then for all $n \geq 0$ we have the preservation of the modified energy:

$$\langle u^n|S(\tau)|u^n\rangle = \langle u^0|S(\tau)|u^0\rangle. \tag{V.17}$$

If moreover $u^0 \in H^1$ and A_0 is associated with $\beta(x) = 2\arctan(x/2)$, then there exists a constant C_0 depending on V and α such that for all $n \in \mathbb{N}$,

$$\sum_{|a|\leq 1/\sqrt{\tau}} |a|^2|\xi_a^n|^2 + \frac{1}{\tau}\sum_{|a|>1/\sqrt{\tau}} |\xi_a^n|^2 \leq C_0\left\|u^0\right\|^2_{H^1}, \tag{V.18}$$

where $\xi_a^n, a \in \mathbb{Z}^d$ are the Fourier coefficients of the function u^n.

Proof. Let us first note that as $S(\tau)$ commutes with $\exp(-i\tau S(\tau))$ we have for all v,

$$\begin{aligned}\langle\exp(-i\tau S(\tau))v|S(\tau)|\exp(-i\tau S(\tau))v\rangle &= \langle v|\exp(i\tau S(\tau))S(\tau)\exp(-i\tau S(\tau))|v\rangle\\ &= \langle v|S(\tau)|v\rangle,\end{aligned}$$

and this shows (V.17) by induction.
Using the fact that V is symmetric, we have for all n, $\|u^n\|_{L^2} = \left\|u^0\right\|_{L^2}$. Hence, using Lemma V.3, we can write for all $v \in L^2$,

$$\langle v|S(\tau)|v\rangle = \frac{1}{\tau}\langle v| - 2\arctan\left(\frac{\tau\Delta}{2}\right)|v\rangle + \langle v|V(\tau) + \tau W(\tau)|v\rangle,$$

whence using (V.10), Lemma V.3 and the fact that A_0 is a positive operator,

$$|\langle v|S(\tau)|v\rangle| \geq \frac{1}{\tau}\langle v| - 2\arctan\left(\frac{\tau\Delta}{2}\right)|v\rangle - C\,\|V\|_{\alpha}\,\|v\|^2_{L^2}\,.$$

Hence using (V.17) we have that for all n,

$$\begin{aligned}\frac{1}{\tau}\langle u^n| - 2\arctan\left(\frac{\tau\Delta}{2}\right)|u^n\rangle &\leq \langle u^n|S(\tau)|u^n\rangle + C\,\|V\|_\alpha\,\|u^n\|_{L^2}^2\\ &\leq \langle u^0|S(\tau)|u^0\rangle + C\,\|V\|_\alpha\,\left\|u^0\right\|_{L^2}^2.\end{aligned}$$

Using (V.15) with $\nu = 0$, we find that there exists a constant such that for all n,

$$\frac{1}{\tau}\langle u^n| - 2\arctan\left(\frac{\tau\Delta}{2}\right)|u^n\rangle \leq C_0\left\|u^0\right\|_{H^1}^2. \tag{V.19}$$

Now we have for all $x > 0$,

$$x > \frac{1}{2} \Longrightarrow \arctan x > \arctan\left(\frac{1}{2}\right) \quad \text{and} \quad x \leq \frac{1}{2} \Longrightarrow \arctan x > \frac{2x}{3}. \tag{V.20}$$

Applying this inequality to (V.19) by considering the set of frequencies $\tau|a|^2 \leq 1$ and $\tau|a|^2 > 1$ then yields the result. ■

This last result shows that H^1 estimates are preserved over arbitrary long time periods only for “low” modes $|a| < 1/\sqrt{\tau}$ whereas the remaining high frequencies part is small in L^2.

Remark V.10. The previous results extend to the splitting scheme

$$\exp(-i\tau A_0)\exp(-i\tau V)$$

and to the Strang splitting

$$\exp(-i\tau V/2)\exp(-iA_0)\exp(-i\tau V/2). \tag{V.21}$$

Note that in this last situation, the fact that the method is of order 2 allows us to take $\nu \in [0,2]$ in (V.15).

The previous result shows long time bounds for the regularity of the solution, measured in the norm (V.18), and in the case of the implicit-explicit integrator. In the case where the high frequencies are cut by the use of a CFL restriction with number c_0, the L^2 norm of the high modes cannot be controlled as the operator A_0 is not positive anymore in the high modes. However, in the case of a fully discrete system, where c_0 is naturally defined by the number of modes in the discretization, no high modes are present, and we do not need to control them. In this case, we can prove the preservation of the H^1 norm of the fully discrete solution. This is the goal of the next section.

5 Fully discrete splitting method

We now consider a full discretization of the previous splitting schemes, using the pseudo-spectral method described in the previous chapters. For simplicity of notation, we consider only the case where the dimension $d = 1$.

Let K be an integer. As in Section 6 of Chapter III, we define the grid $x_a = 2\pi a/K$ made of K equidistant points in the interval $[-\pi, \pi]$, with $a \in B^K$ defined in (III.35). With this grid is associated the discrete Fourier transform (III.36).

As in the nonlinear case (see Section 6 of Chapter III), we search for a trigonometric polynomial

$$U^K(t,x) = \sum_{a \in B^K} e^{iax} \xi_a^K(t)$$

such that for all $b \in B^K$, the equation

$$i\partial_t U^K(t,x_b) = -\Delta U^K(t,x_b) + V(x_b)U^K(t,x_b), \quad U^K(0,x_b) = u^0(x_b),$$

is satisfied for all time t (compare (III.37) for the nonlinear case).

In terms of the vector $\xi^K(t) := (\xi_a^K(t))_{a \in B^K}$ constructed with the coefficients of the polynomial U^K, we see that we can write the previous equation as

$$i\mathcal{F}_K^{-1}\dot{\xi}^K(t) = \mathcal{F}_K^{-1}D^K\xi^K(t) + V^K\mathcal{F}_K^{-1}\xi^K(t),$$

where D^K and V^K are the K-dimensional diagonal matrices given by:

$$D^K = \operatorname{diag}(a^2), \quad \text{and} \quad V^K = \operatorname{diag}(V(x_a)), \quad a \in B^K.$$

Hence after taking the inverse of the Fourier transform, we see that the vector $\xi^K(t)$ satisfies the linear system of differential equation (of dimension K)

$$i\dot{\xi}^K(t) = D^K\xi^K(t) + W^K\xi^K(t), \tag{V.22}$$

where

$$W^K = \mathcal{F}_K V^K \mathcal{F}_K^{-1},$$

and with initial condition $\xi^K(0) = \mathcal{F}_K \circ \operatorname{diag}(u^0(x_b))$. Note that as $\sqrt{K}\mathcal{F}_K$ is unitary, the matrix W^K is symmetric.

Let us consider a standard splitting method applied to this equation. It can be written as the numerical scheme

$$\begin{aligned}\xi^{K,n+1} &= \exp(-i\tau D^K) \circ \exp(-i\tau W^K)\xi^{K,n} \\ &= \exp(-i\tau D^K) \circ \mathcal{F}_K \circ \exp(-i\tau V^K) \circ \mathcal{F}_K^{-1}\xi^{K,n}\end{aligned}$$

acting on $\mathbb{C}^K$, which corresponds to the numerical scheme defined in Chapter I.

Note that the computational cost of this method is relatively low: As the matrices D^K and V^K are diagonal, the evaluation of the exponentials is cheap, while the evaluation of the Fourier transforms $\mathcal{F}_K$ and $\mathcal{F}_K^{-1}$ can be easily made using the Fast Fourier Transform (FFT) algorithm.

To obtain backward error analysis as in the previous section with bounds independent of the spectral parameter K, we consider D^K and W^K as operators acting on $\mathbb{C}^{\mathbb{Z}}$ and leaving invariant the space $\{\xi_a \in \mathbb{C}^Z | \xi_a = 0 \text{ if } a \notin B^K\}$. To apply the previous result, we need two ingredients:

- We can consider that $\tau D^K = A_0^K = -\beta(\tau\Delta)$ where $\beta(x) = x\mathbb{1}_{x\leq c_0}$ with the CFL number c_0 naturally defined as $c_0 = \tau K^2/4$. We can also consider the case of implicit-explicit integrator, i.e., $\beta(x) = 2\arctan(x/2)\mathbb{1}_{x\leq c_0}$.
- The operator W^K is a finite dimensional matrix operator satisfying
$$W_{ab}^K = 0, \quad \text{if} \quad a \notin B^K \quad \text{or} \quad b \notin B^K.$$
Hence it is clear that W^K belongs to all the space $\mathcal{L}_\alpha$, $\alpha > 0$, but with a norm depending *a priori* on K.

The key to applying Theorem V.7 is to estimate the norm of W^K.

Proposition V.11. *Assume that $V(x)$ defines an operator $V_{ab} \in \mathcal{L}_\alpha$ for $\alpha > 1$. For all $K \geq 1$, let $W^K = (W_{ab}^K)_{a,b\in\mathbb{Z}}$ be the operator defined by*

$$W^K = \mathcal{F}_K \circ (\mathrm{diag}(V(x_b))) \circ \mathcal{F}_K^{-1}.$$

Then for all K and for all $\nu > 0$, we have that $W^K \in \mathcal{L}_\nu$, and satisfies

$$\left\| W^K \right\|_\nu \leq c_\alpha K^\nu \left\| V \right\|_\alpha \tag{V.23}$$

for some constant depending on α only.

Proof. We calculate directly that for $a, b \in B^K$, we have

$$W_{ab}^K = \frac{1}{K} \sum_{j\in B^K} V(x_j) e^{-i(a-b)x_j}.$$

With this (finite dimensional) operator, we can naturally associate an operator acting on $\mathbb{C}^{\mathbb{Z}}$, by setting $W_{ab}^K = 0$ when $a \notin B^K$ or $b \notin B^K$. By decomposing the function V in the Fourier basis, we obtain for $a, b \in B^K$,

$$W_{ab}^K = \frac{1}{K} \sum_{j\in B^K} \sum_{c\in\mathbb{Z}} \hat{V}_c e^{-i(a-b-c)x_j} = \sum_{d\in\mathbb{Z}} \hat{V}_{d+a-b} \frac{1}{K} \sum_{j\in B^K} e^{idx_j}.$$

Using formula (III.38), we obtain

$$W_{ab}^{K} = \sum_{m \in \mathbb{Z}} \hat{V}_{a-b+mK}.$$

Assuming that $V \in \mathcal{L}_\alpha$, we thus have

$$|W_{ab}^{K}| \leq \|V\|_\alpha \sum_{m \in \mathbb{Z}} \frac{1}{1 + |a - b + mK|^\alpha}.$$

We thus obtain as $|a - b| \leq K - 1$,

$$\begin{aligned}(1 + |a-b|^\nu)\,|W_{ab}^{K}| &= \|V\|_\alpha \sum_{m \in \mathbb{Z}} \frac{1 + |a-b|^\nu}{1 + |a - b + mK|^\alpha} \\ &\leq \|V\|_\alpha \sum_{m \in \mathbb{Z}} \frac{1 + |K|^\nu}{1 + |a - b + mK|^\alpha} \\ &\leq \|V\|_\alpha \sum_{p \in \mathbb{Z}} \frac{1 + K^\nu}{1 + |p|^\alpha},\end{aligned}$$

which yields the result. ■

Remark V.12. The bound (V.23) is sharp in the sense that we cannot obtain a bound independent of K for $\nu > 0$. This is due to the *aliasing* problem. To see this, take K even, $m = -1$, $a = K/2 - 1$ and $b = -K/2$ in the previous sum.

With this result, we are now ready to prove the following result for the fully discrete splitting method applied to the linear Schrödinger equation.

Theorem V.13. *Consider the linear equation* (V.1) *on the one-dimensional torus* $\mathbb{T}^1$. *Let* $\alpha > 1$, *and assume that* $\|V\|_\alpha < \infty$. *Let* K *be a given number, and consider the approximation* (V.22) *of* (V.1) *by collocation method in the Fourier basis.*

Let A_0^K *be the diagonal operator with eigenvalues* $\lambda_a^K = \tau a^2$ *or* $\lambda_a^K = 2\arctan(\tau a^2/2)$, $a \in B^K$. *Assume that the two following conditions are satisfied:*

$$0 \leq \lambda_a^K \leq \pi \quad \textit{and} \quad \tau c_\alpha K^\nu \leq \tau_0 \tag{V.24}$$

for some $\nu > 1$, *where* τ_0, *given by Theorem* V.7, *depends only on* V, *and where* c_α *is the constant appearing in* (V.23).

Then there exists a symmetric matrix $S^K(\tau)$ *such that*

$$\exp - \left(i\tau W^K\right) \circ \exp\left(-iA_0^K\right) = \exp\left(-i\tau S^K(\tau)\right),$$

satisfying for all $\tau \in (0, \tau_0)$,

$$S^K(\tau) = A_0^K + V^K(\tau)$$

where $V^K(\tau)$ *satisfy,*

$$\left\| V^K(\tau) \right\|_\nu \le C$$

for some constant C *depending on* $\|V\|_\alpha$ *but not on* K.

Proof. Let A_0 be the diagonal operator associated with the eigenvalues λ_a defined by

$$\lambda_a = \begin{cases} \lambda_a^K & \text{if} \quad a \in B^K, \\ 0 & \text{if} \quad a \notin B^K. \end{cases}$$

The bound on the eigenvalues λ_a^K ensures that A_0 satisfies the condition of Theorem V.7.
As W^K satisfy the bound (V.23), the same arguments as in the proof of Theorem V.7 show that the coefficients of the formal series $\sum t^\ell Z_\ell^K$ constructed from the potential W^K satisfy the bounds (compare (V.14)),

$$\left\| Z_\ell^K \right\|_\nu \le \frac{3\pi}{4C_\alpha} \left(\frac{48 M C_\alpha c_\alpha K^\nu \|V\|_\alpha}{\pi} \right)^\ell .$$

This shows that the series $\sum t^\ell Z_\ell^K$ is convergent in $\mathcal{L}_\nu$ under the condition $\tau c_\alpha K^\nu \le \tau_0$ defined in the proof of Theorem V.7. The result is then obtained using the same arguments.
Note that by construction the operators Z_ℓ^K are matrix operators acting on the subspace spanned by the indices $a \in B^K$. This can be verified from the fact that the bracket of two operators acting on the same subspace defines an operator on this subspace. ■

Let us comment on the condition (V.24). In the case of the standard splitting method, it can be written (using the fact that all the frequencies in B^K are smaller than $K^2/4$),

$$\tau K^2 \le 4\pi \quad \text{and} \quad \tau K^\nu \le C$$

for some constant C depending on $\|V\|_\alpha$. As the condition on ν is only $\nu > 1$, we thus see that the second condition will be automatically satisfied for K sufficiently large.

In the case of the implicit-explicit midpoint rule, the first condition $\lambda_a \le \pi$ is always satisfied. Concerning the second condition, as ν can be arbitrarily close to 1 (of course with a possible deterioration of the constant) it can be viewed as a softer

CFL condition than the one expected from the asymptotic eigenvalues of the operator. Roughly speaking, we can say that the use of an implicit-explicit scheme thus allows us to replace a CFL condition of the form $\tau K^2 \leq c$ by a milder one $\tau K \leq c$ to obtain a modified energy.

Using the modified energy constructed in the previous theorem, we can prove a long time H^1 bound for the fully discrete solution:

Corollary V.14. *With the notation of the previous Theorem, let $u^0 \in \ell_1^1 \subset H^1(\mathbb{T})$, and let $(\xi_a^{K,0})_{a \in B^K} = \mathcal{F}_K \circ \mathrm{diag}(u^0(x_b))$ be the vector defined by the Formula (IV.22). Let $\xi^{K,n}(t) = (\xi_a^{K,n})_{a \in B^K}$ be the sequence in $\mathbb{C}^{B^K}$ defined by the formula*

$$\xi^{K,n+1} = \exp\left(-i\tau W^K\right) \circ \exp\left(-iA_0^K\right) \xi^{K,n}, \qquad n \geq 0.$$

Assume that τ and K satisfy the condition $\tau K^2 \leq 4\pi$. Then there exists a constant C independent of K such that

$$\forall n \geq 0, \qquad \sum_{a \in B^K} |a|^2 \left|\xi_a^{K,n}\right|^2 \leq C \left\|\xi^{K,0}\right\|_{H^1}^2 \leq C \left\|u^0\right\|_{\ell_1^1}^2. \tag{V.25}$$

Proof. The condition $\tau K^2 \leq 4\pi$ ensures that the condition (V.24) is always satisfied (both in the case of the standard and implicit-explicit scheme). Moreover, the preservation of the modified energy $S^K(\tau)$ given by the previous theorem combined with the proof of (V.18) easily show that

$$\forall n \geq 0, \qquad \sum_{a \in B^K} |a|^2 \left|\xi_a^{K,n}\right|^2 \leq C \left\|\xi^{K,0}\right\|_{H^1}^2$$

for some constant C independent on K and $\xi^{K,0}$: this is due to the fact that no high modes are present in the fully discrete version of (V.18) (as $\tau K^2 \leq 4\pi$).
To prove the last estimate, we use (IV.22) and obtain for all $a \in B^K$,

$$|a| \left|\xi_a^{K,0}\right| \leq \sum_{m \in \mathbb{Z}} |a| \left|\xi_{a+mK}^0\right| \leq \left\|u^0\right\|_{\ell_1^1} = \sum_{a \in \mathbb{Z}} |a| \left|\xi_a^0\right|.$$

Hence

$$\begin{aligned}\left\|\xi^{K,0}\right\|_{H^1}^2 := \sum_{a \in B^K} |a|^2 \left|\xi_a^{K,0}\right|^2 &\leq \left\|u^0\right\|_{\ell_1^1} \sum_{a \in B^K} |a| \left|\xi_a^{K,0}\right| \\ &\leq \left\|u^0\right\|_{\ell_1^1} \sum_{a \in B^K} \sum_{m \in \mathbb{Z}} |a| \left|\xi_{a+mK}^0\right| \\ &\leq \left\|u^0\right\|_{\ell_1^1} \sum_{b \in \mathbb{Z}} |b| \left|\xi_b^0\right| = \left\|u^0\right\|_{\ell_1^1}^2,\end{aligned}$$

and this concludes the proof. ∎

6 Resonance analysis

In the construction made in the above sections, the first term of the modified energy is given by (see (V.13))

$$(Z_1)_{ab} = V_{ab} \frac{i(\lambda_a - \lambda_b)}{\exp(i(\lambda_a - \lambda_b)) - 1},$$

where $a, b \in \mathbb{Z}^d$, and $\lambda_a = \beta(\tau|a|^2)$. The hypothesis $\lambda_a \le \pi$ implies that for all a and b in $\mathbb{Z}^d$, the term $\lambda_a - \lambda_b$ belongs to the interval $[-\pi, \pi]$ and hence avoids the poles $\pm 2\pi$.

In the case of the implicit-explicit integrator with $\beta(x) = 2\arctan(x/2)$, we always have $\lambda_a \le \pi$, and thus we can observe that the function $\tau \mapsto Z(\tau)$ is continuous in τ. This explains the absence of numerical resonances, as shown on the bottom of Figure I.6.

Now in the case of the standard splitting scheme with $\beta(x) = x$, we see that resonances appear for some values of the time step τ such that there exist a, b and k such that

$$\tau(|a|^2 - |b|^2) \simeq 2\pi k, \tag{V.26}$$

in which case the term Z_1 above is not well defined. This makes the function $\tau \mapsto Z(\tau)$ not well defined beyond the CFL regime $\tau|a|^2 < 2\pi$. However, as shown in Figure I.6, these singularities seem to appear for very specific values of τ. Such a phenomenon is called *resonance* effect, and a step-size τ satisfying (V.26) is called *resonant*. When the step-size is not resonant, we remark that the modified energy Z_1 is still well defined, and hence we expect that the numerical scheme is stable.

As we will see below, such resonance relation is not *generic*, in the sense that very few step-sizes τ are resonant, and satisfy (V.26). Said differently, a step-size τ chosen randomly in an interval $[0, \tau_0]$ has many chances to be non resonant, and hence to yield a stable long time integration of the equation.

The goal of this last section is to quantify this fact. To do this, let us consider the following non resonance condition:

$$\forall n \in \mathbb{Z}, \quad n \neq 0, \quad \left| \frac{1 - e^{i\tau n}}{\tau} \right| \ge \frac{\gamma}{|n|^\nu}. \tag{V.27}$$

If such a diophantine relation is satisfied, long time results can be obtained for classical splitting methods applied to the linear Schrödinger equation with small potential, see [13].

Here we will not give details about this result, but show that such a relation is *generic*. As the method is standard in resonance analysis, we give here a complete proof of the following proposition (see also [26]):

Proposition V.15. *Let $\gamma > 0$ and $\nu > 1$ be fixed, and let*

$$Z(\tau_0; \gamma, \nu) = \{\tau \in (0, \tau_0) | \tau \textit{ does not satisfy } (\text{V.27})\}.$$

Then we have

$$\operatorname{meas} Z(\tau_0) \leq C\gamma\tau_0^2$$

for some constant C independent of γ and ν. As a consequence for a fixed $\nu > 1$, the set of $\tau \in (0, \tau_0)$ for which there exists γ such that (V.27) is of full measure in $(0, \tau_0)$.

Proof. Assume that τ does not satisfy (V.27). Then there exists $k \in \mathbb{Z}\backslash\{0\}$ such that

$$\left|1 - e^{i\tau k}\right| \leq \frac{\gamma\tau}{|k|^\nu}.$$

Now for this k, there exists ℓ such that $|k\tau - 2\pi\ell| < \pi$, and hence as for $x \in [-\pi, \pi]$ we have $|1 - e^{ix}| \geq \frac{2}{\pi}|x|$, we get

$$\left|1 - e^{i\tau k}\right| \geq \frac{2}{\pi}|k\tau - 2\pi\ell| \geq \frac{2|k|}{\pi}\left|\tau - \frac{2\pi\ell}{k}\right|.$$

But as $|k\tau - 2\pi\ell| < \pi$, we have for this ℓ the bound

$$2\pi|\ell| \leq \pi + |k|\tau_0.$$

Hence τ is in the set $Z(\tau_0)$ if there exists $k \neq 0$ and ℓ such that

$$\frac{2|k|}{\pi}\left|\tau - \frac{2\pi\ell}{k}\right| \leq \frac{\gamma\tau}{|k|^\nu}$$

or

$$\left|\tau - \frac{2\pi\ell}{k}\right| \leq \frac{\pi\gamma\tau}{2|k|^{\nu+1}}.$$

Note that if $\ell = 0$ in the previous inequality, we must have

$$1 \leq \frac{\pi\gamma}{2|k|^{\nu+1}} \leq \frac{\pi\gamma}{2}$$

which contradicts $\gamma < 2/\pi$. Hence ℓ is submitted to the restriction

$$\ell \neq 0, \quad \text{and} \quad |\ell| \leq \frac{1}{2} + \frac{|k|\tau_0}{2\pi},$$

and we note that there are at most $\frac{|k|\tau_0}{\pi}$ such ℓ for a given k. Hence we have

$$\begin{aligned}\operatorname{meas} Z(\tau_0; \gamma, \nu) &\leq \sum_{k\neq 0} \sum_{\substack{|\ell| \leq \frac{1}{2} + \frac{|k|\tau_0}{2\pi} \\ \ell \neq 0}} \frac{\pi\gamma\tau_0}{2|k|^{\nu+1}} \\ &\leq \sum_{k\neq 0} \frac{|k|\tau_0}{\pi} \frac{\pi\gamma\tau_0}{2|k|^{\nu+1}} \\ &\leq \frac{\gamma}{2}\tau_0^2 \sum_{k\neq 0} \frac{1}{|k|^\nu} \leq C\gamma\tau_0^2.\end{aligned}$$

The last statement follows from the fact that

$$\operatorname{meas} \bigcap_{\gamma > 0} Z(\tau_0; \gamma, \nu) = 0. \qquad \blacksquare$$

This result partly explains the top Figure I.6: the set of resonant step-sizes is very small and generically, we can hope that the energy is well preserved. However, the complete analysis has only be performed for small potential, see [13].

VI Modified energy in the semi-linear case

In this chapter, we go back to the nonlinear case. For simplicity, we will only consider the case of the cubic nonlinear Schrödinger equation

$$i\partial_t u = -\Delta u + \lambda |u|^2 u, \tag{VI.1}$$

set on the one-dimensional torus $\mathbb{T}^1$, but the results are valid in more general cases, where the nonlinearity is polynomial and the equation set is on a torus of arbitrary dimension d. We refer to [15] for a more general analysis.

Following the idea of Chapter V concerning the linear case, we consider splitting methods of the form

$$\varphi_{A_0}^1 \circ \varphi_P^\tau \simeq \varphi_H^\tau,$$

where A_0 is a filtered Laplace operator smoothed in the high frequencies, and where P is the nonlinear part of (VI.1). We have proved in Chapter IV that such schemes are convergent over finite time intervals, provided the exact solution is smooth. Here, we prove the existence of a modified Hamiltonian H_τ such that

$$\varphi_{A_0}^1 \circ \varphi_P^\tau = \varphi_{H_\tau}^\tau + \mathcal{O}(\tau^N), \tag{VI.2}$$

where the degree of precision N depends on a bound for the operator A_0. Note that in our situation, the operator A_0 is bounded but not small (as would be the case in the finite dimensional case): typically its norm behaves like $\beta(\tau K^2/4)$ where β is the filter function, and K the number of grid points in the underlying pseudo-spectral collocation method, as described in previous chapters. Hence we see that the precision level N can be viewed as depending on the CFL number.

Such a backward error analysis result is valid in the Wiener algebra ℓ^1: the relation (VI.2) is valid in ℓ^1, and the constant error depends on *a priori* bounds of the numerical solution in the same Banach algebra. Hence as long as the numerical solution remains in ℓ^1, the energy H_τ is preserved along the numerical solution. Using this property in the case of fully discretized numerical schemes, we prove by a bootstrap argument an almost global existence result for small discrete initial data, which constitutes a fully discrete version of the global existence result stated in Chapter III (Proposition III.14).

Such a result extends to more general situations where the Hamiltonian function is of the form (with the notation of the previous chapters)

$$H(z) = T(z) + P(z)$$

where $T(z) = \sum_{a \in \mathbb{Z}^d} \omega_a \xi_a \eta_a$ and $P \in \mathcal{P}_{r_0}$ with $r_0 \geq 3$, and where the frequencies satisfy the bound $\omega_a \leq C|a|^2$. We refer to [15] for detailed proofs.

1 Recursive equations

In this section we explain the mechanism of construction of the modified energy. As we will see, the formalism turns out to be the same as in the previous chapter, except that the bracket of two operators has to be replaced by the Poisson bracket between two polynomial Hamiltonians. For simplicity, we consider only the splitting method $\varphi_P^\tau \circ \varphi_{A_0}^1$. The splitting method $\varphi_{A_0}^1 \circ \varphi_P^\tau$ can be treated similarly. Note that the long time behavior of these two methods are the same, as we have for all $n \geq 1$,

$$\left(\varphi_{A_0}^1 \circ \varphi_P^\tau\right)^n = \varphi_{A_0}^1 \circ \left(\varphi_P^\tau \circ \varphi_{A_0}^1\right)^{n-1} \circ \varphi_P^\tau.$$

Here, recall that A_0 is a diagonal operator obtained by smoothing the frequencies with a filter function β: The operator A_0 is written (see (IV.9))

$$A_0 = \sum_{a \in \mathbb{Z}} \lambda_a \xi_a \eta_a, \quad \text{with} \quad \lambda_a = \beta(\tau a^2), \quad a \in \mathbb{Z}.$$

As before, we will mainly consider the cases where $\beta(x) = x$, $\beta(x) = 2\arctan(x/2)$, possibly in combination with the use of a CFL condition.

With the notation of the previous chapters, the Hamiltonian P is given by

$$P(\xi, \eta) = \frac{\lambda}{4\pi} \int_{\mathbb{T}} |u(x)|^4 \mathrm{d}x = \frac{\lambda}{2} \sum_{a+b-c-d=0} \xi_a \xi_b \eta_c \eta_d, \tag{VI.3}$$

with the usual notation $u(x) = \sum_{a \in \mathbb{Z}} \xi_a \, e^{iax}$ and $\eta_a = \bar{\xi}_a$. This is the polynomial Hamiltonian associated with the cubic nonlinearity (VI.1).

Following the strategy developed in the previous chapter in the linear case, we look for a real Hamiltonian *polynomial* function $Z(t) := Z(t; \xi, \eta)$ such that for all $t \leq \tau$ we have

$$\varphi_P^t \circ \varphi_{A_0}^1 = \varphi_{Z(t)}^1, \tag{VI.4}$$

and such that $Z(0) = A_0$. With the notation of Chapter III, we can write for a given Hamiltonian K,

$$\varphi_K^1 = \exp(\mathcal{L}_K)[\mathrm{Id}]. \tag{VI.5}$$

Differentiating the exponential map we calculate, as in Chapter II, that

$$\frac{\mathrm{d}}{\mathrm{d}t} \varphi_{Z(t)}^1 = X_{Q(t)} \circ \varphi_{Z(t)}^1,$$

where (at least formally) the differential operator associated with $Q(t)$ is given by

$$\mathcal{L}_{Q(t)} = \sum_{k \geq 0} \frac{1}{(k+1)!} \mathrm{Ad}_{\mathcal{L}_{Z(t)}}^k \left(\mathcal{L}_{Z(t)}\right),$$

with

$$\mathrm{Ad}_{\mathcal{L}_A}(\mathcal{L}_H) = [\mathcal{L}_A, \mathcal{L}_H]$$

the commutator of two vector fields. As the vector fields are Hamiltonian, we have for two Hamiltonian functions A and H,

$$[\mathcal{L}_A, \mathcal{L}_H] = \mathcal{L}_{\{A,H\}}.$$

Hence we obtain the formal series equation for Q:

$$Q(t) = \sum_{k\geq 0} \frac{1}{(k+1)!} \mathrm{ad}^k_{Z(t)} Z'(t), \tag{VI.6}$$

where $Z'(t)$ denotes the derivative with respect to t of the Hamiltonian function $Z(t)$, and where for two Hamiltonian functions K and G,

$$\mathrm{ad}_K(G) = \{K, G\}.$$

Therefore taking the derivative of (VI.4), we obtain

$$X_P \circ \varphi^t_P \circ \varphi^1_{A_0} = X_{Q(t)} \circ \varphi^1_{Z(t)}$$

and hence the equation to be satisfied by $Z(t)$ reads:

$$\sum_{k\geq 0} \frac{1}{(k+1)!} \mathrm{ad}^k_{Z(t)} Z'(t) = P. \tag{VI.7}$$

So formally, using the results in Chapter II, equation (VI.7) is equivalent to the formal series equation

$$Z'(t) = \sum_{k\geq 0} \frac{B_k}{k!} \mathrm{ad}^k_{Z(t)} P. \tag{VI.8}$$

Remark VI.1. Equation (VI.8) is formally the same as equation (V.4) in the previous chapter. Note that in the case of quadratic Hamiltonians, then the Poisson bracket of two Hamiltonian is again a quadratic Hamiltonian associated with an operator given by the bracket of two linear operators. In this situation, (VI.8) and (V.4) are equivalent.

Plugging an Ansatz expansion $Z(t) = \sum_{\ell\geq 0} t^\ell Z_\ell$ into this equation, we get $Z_0 = A_0$ and for $n \geq 0$

$$(n+1)Z_{n+1} = \sum_{k\geq 0} \frac{B_k}{k!} \sum_{\ell_1+\cdots+\ell_k=n} \mathrm{ad}_{Z_{\ell_1}} \cdots \mathrm{ad}_{Z_{\ell_k}} P. \tag{VI.9}$$

In the following, we will show that this formula allows us to construct the terms Z_n up to a level N depending in general on a CFL condition imposed on the system.

2 Construction of the modified energy

2.1 First terms. Let us write down the formula (VI.9) for $n = 0$. We obtain

$$Z_1 = \sum_{k \geq 0} \frac{B_k}{k!} \mathrm{ad}_{A_0}^k P. \tag{VI.10}$$

Recall that here we consider a class of polynomials of the form (III.15), where for a given ℓ, $\mathcal{J}_\ell$ is the set of multi-indices in $\mathcal{Z} = \mathbb{Z}^d \times \{\pm 1\}$ with zero momentum (see (III.14)). To calculate the first term Z_1 defined above (and further the other terms in the development), we use the following result:

Lemma VI.2. *Let $r \geq 2$ and assume that*

$$Q(z) = \sum_{\boldsymbol{j} \in \mathcal{J}_r} a_{\boldsymbol{j}} z_{\boldsymbol{j}},$$

is a homogeneous polynomial of degree r, and let $A_0(z) = \sum_{a \in \mathbb{Z}} \lambda_a \xi_a \eta_a$ for $z = (\xi, \eta)$, then we have

$$\mathrm{ad}_{A_0}(Q) = \sum_{\boldsymbol{j} \in \mathcal{J}_r} i\Lambda(\boldsymbol{j}) a_{\boldsymbol{j}} z_{\boldsymbol{j}}$$

where for a multi-index $\boldsymbol{j} = (j_1, \dots, j_r)$ with $j_i = (a_i, \delta_i) \in \mathbb{Z}^d \times \{\pm 1\}$, for $i = 1, \dots, r$, we set

$$\Lambda(\boldsymbol{j}) = \delta_1 \lambda_{a_1} + \cdots + \delta_r \lambda_{a_r}.$$

In particular, we have

$$\sum_{k \geq 0} \frac{B_k}{k!} \mathrm{ad}_{A_0}^k Q = \sum_{\boldsymbol{j} \in \mathcal{Z}^r} \frac{i\Lambda(\boldsymbol{j})}{\exp(i\Lambda(\boldsymbol{j})) - 1} a_{\boldsymbol{j}} z_{\boldsymbol{j}}.$$

Proof. This is just a calculation made from the expression of the Poisson bracket of two Hamiltonians. ∎

Remark VI.3. According to the definition III.4 of polynomials containing the same number of ξ's and η's, we see that if $Q \in \mathcal{SP}_r$, then we also have $\mathrm{ad}_{A_0}(Q) \in \mathcal{SP}_r$.

With the expression (VI.3) of the polynomial P associated with the NLS equation (VI.1), we obtain using (VI.10) that the first term of the modified equation is given by

$$Z_1 = \frac{\lambda}{2} \sum_{a+b-c-d=0} \frac{i\Lambda_{abcd}}{\exp(i\Lambda_{abcd}) - 1} \xi_a \xi_b \eta_c \eta_d, \tag{VI.11}$$
$$\text{with} \quad \Lambda_{abcd} = \lambda_a + \lambda_b - \lambda_c - \lambda_d.$$

We see that to be well defined, we need to avoid configurations such that $\Lambda_{abcd} \simeq 2\pi m$ for some $m \in \mathbb{Z}$, $m \neq 0$. Let us examine this condition in the two main cases of applications: First when $\beta(x) = x$, we have

$$\Lambda_{abcd} = \tau\left(a^2 + b^2 - c^2 - d^2\right).$$

Note that in the example of the introduction, we have precisely made a simulation with a resonant step-size τ such that $\Lambda_{abcd} = 2\pi$, see equation (I.20). More precisely, we have taken $a = 0$, $b = 12$, $c = 5$ and $d = 7$. Figure I.8 shows the instabilities observed for the numerical solution in this case. It corresponds to a singularity in the first term Z_1 defined above. To avoid such a situation, we have to impose the fact that Λ_{abcd} is never a multiple of 2π. This condition can be interpreted as a non resonance condition on the step size τ.

Let us assume that a CFL condition is imposed, i.e. that

$$\beta(x) = x\mathbb{1}_{x<c_0}(x),$$

where c_0 is the CFL number. In this situation, we have $\lambda_a = 0$ if $\tau a^2 \geq c_0$. Hence taking into account the positivity of the eigenvalues, we have for all a, b, c and d in $\mathbb{Z}$,

$$|\Lambda_{abcd}| \leq 2c_0.$$

This shows that the term Z_1 is well defined as soon as $c_0 < \pi$. Now if we consider the case where $\beta(x) = 2\arctan(x/2)$, we can perform a similar analysis, and the previous condition yields

$$|\Lambda_{abcd}| \leq 4\arctan\left(c_0^2/2\right) < 2\pi,$$

for all c_0. Hence we see that for all c_0, we can construct the first term of the modified energy. In other words, there is no restriction on the CFL number when using the implicit-explicit scheme (at least to construct the first term).

Let us now write down the formula (VI.9) for $n = 1$. The second term Z_2 satisfies

$$Z_2 = \frac{1}{2}\sum_{k\geq 0}\frac{B_k}{k!}\sum_{m=0}^{k-1}\mathrm{ad}_{A_0}^m\mathrm{ad}_{Z_1}\mathrm{ad}_{A_0}^{k-1-m}P.$$

Using Proposition III.6, the first thing to note is that Z_2 is a *homogeneous* polynomial of degree 6, and made of monomials of the form

$$\xi_{a_1}\xi_{a_2}\xi_{a_3}\eta_{b_1}\eta_{b_2}\eta_{b_3},$$

in other words, $Z_2 \in \mathcal{SP}_6$. In order to give a meaning to Z_2, we will see that we need now a condition on the form $|\Lambda(\boldsymbol{j})| < 2\pi$ for multi-indices $\boldsymbol{j}$ of degree 6 satisfying the symmetry condition (III.17). As expected, this requires a stronger CFL condition, even in the case of the implicit-explicit integrator.

2.2 Iterative construction. The following proposition gives a condition to construct the terms Z_n defined by (VI.9) up to some fixed level $n = N$.

Proposition VI.4. *Let $N > 0$ be fixed, and let P be the Hamiltonian* (VI.3). *Assume that the eigenvalues λ_a of the operator A_0 satisfy*

$$\lambda_a = \begin{cases} \beta(\tau a^2) & \text{if} \quad \tau a^2 \leq c_0, \\ 0 & \text{if} \quad \tau a^2 > c_0, \end{cases} \tag{VI.12}$$

where $\beta(x) = x$ or $\beta(x) = 2\arctan(x/2)$ and $c_0 > 0$.

Assume that c_0 satisfies the condition

$$c_0 < \beta^{-1}\left(\frac{2\pi}{N+1}\right). \tag{VI.13}$$

Then for $n \leq N$ we can define homogeneous symmetric polynomials $Z_n \in \mathcal{SP}_{2n+2}$ satisfying the equations (VI.9) *up to the order n, and such that $\|Z_n\| < +\infty$.*

Proof. Let $\mathcal{SI}_n$ be the set of multi-indices $\boldsymbol{j} = (j_1, \ldots, j_n)$ satisfying the symmetry condition (III.17). In other words, for all $\boldsymbol{j} \in \mathcal{SI}_n$, the number of ξ_a and η_a is the same in the monomial $z_{\boldsymbol{j}}$. Using the definition of $\Lambda(\boldsymbol{j})$ and the condition (VI.13), we see that there exists $\delta > 0$ such that

$$\forall n \leq 2N+2 \quad \forall \boldsymbol{j} \in \mathcal{SI}_n, \quad |\Lambda(\boldsymbol{j})| \leq 2\pi - \delta. \tag{VI.14}$$

Using Lemma VI.2, this condition implies that for any homogeneous polynomial $Q \in \mathcal{SP}_r$ with $r \leq 2N+2$, we have the estimate

$$\|\mathrm{ad}_{Z_0} Q\| \leq (2\pi - \delta)\,\|Q\|\,. \tag{VI.15}$$

Under this assumption, we easily see that Z_1 satisfies $\|Z_1\| \leq c_\delta \|P\|$ for some constant c_δ.

Assume now that the Z_k are constructed for $0 \leq k \leq n$, $n \geq 1$ and are such that Z_k is a homogeneous symmetric polynomial of degree $2k+2$. Formally Z_{n+1} is defined as a series

$$Z_{n+1} = \frac{1}{n+1} \sum_{k \geq 0} \frac{B_k}{k!} A_k,$$

where

$$A_k = \sum_{\ell_1 + \cdots + \ell_k = n} \mathrm{ad}_{Z_{\ell_1}} \cdots \mathrm{ad}_{Z_{\ell_k}} P.$$

Let us prove that this series converges absolutely. In the previous sum, we separate the number of indices j for which $\ell_j = 0$. For them, we can use (VI.15). Only for the other indices, we will use the estimates of Proposition III.6 by taking into account that the right-hand side is a sum of terms that are all real polynomials of degree $(\ell_1 + \cdots + \ell_k)2 + 4 = 2(n+1) + 2$ and hence the inequality of Proposition III.6 is only used with polynomials of order less than $4(n+1)$. Thus we write for $k \geq n$,

$$\begin{aligned}
\|A_k\| &:= \left\| \sum_{\ell_1 + \cdots + \ell_k = n} \mathrm{ad}_{Z_{\ell_1}} \cdots \mathrm{ad}_{Z_{\ell_k}} P \right\| \\
&\leq \sum_{i=1}^{n} \frac{k!\,(2\pi - \delta)^{k-i}}{(k-i)!\,i!} \sum_{\substack{\ell_1 + \cdots + \ell_i = n \\ \ell_j > 0}} (n+1)^{i-1} (32)^i \ell_1 \left\|Z_{\ell_1}\right\| \cdots \ell_i \left\|Z_{\ell_i}\right\| \|P\| \\
&\leq (2\pi - \delta)^{k-n} k^n \sum_{i=1}^{n} \sum_{\substack{\ell_1 + \cdots + \ell_i = n \\ \ell_j > 0}} (n+1)^{i-1} (32)^i \ell_1 \left\|Z_{\ell_1}\right\| \cdots \ell_i \left\|Z_{\ell_i}\right\| \|P\|,
\end{aligned}$$

and thus $\sum_{k \geq 0} \frac{B_k}{k!} A_k$ converges and Z_{n+1} is well defined up to $n + 1 \leq N$. This shows the result. ■

3 Backward error analysis result

For $s \geq 0$, we recall that

$$B_M^s = \left\{ z \in \ell_s^1 \,|\, \|z\|_{\ell_s^1} \leq M \right\}$$

and we will use the notation $B_M = B_M^0$.

Theorem VI.5. *Let $N \geq 1$, $s \geq 0$ and $M_0 \geq 1$ be fixed. Then there exist constants τ_0 and C_N depending on s, $\|P\|$, M_0 and N such that the following holds: For all $\tau \leq \tau_0$ such that the eigenvalues λ_a of the operator A_0 defined by (VI.12) with c_0 satisfying (VI.13), there exists a real Hamiltonian polynomial $H_\tau \in \mathcal{S}\mathcal{P}_{2N+2}$ such that for all $M < M_0$ and $z \in B_M^s$, we have*

$$\left\| \varphi_P^\tau \circ \varphi_{A_0}^1(z) - \varphi_{H_\tau}^\tau(z) \right\|_{\ell_s^1} \leq C_N M^{2N+1} \tau^{N+1}. \tag{VI.16}$$

Moreover, for $z \in B_M^s$ we have

$$\left| H_\tau(z) - H_\tau^{(1)}(z) \right| \leq C_N \tau M^6 \tag{VI.17}$$

where

$$H_\tau^{(1)}(z) = \sum_{a \in \mathbb{Z}} \frac{1}{\tau} \lambda_a \xi_a \eta_a + \frac{\lambda}{2} \sum_{a+b-c-d=0} \frac{i\Lambda_{abcd}}{\exp(i\Lambda_{abcd}) - 1} \xi_a \xi_b \eta_c \eta_d \tag{VI.18}$$

with the notation (VI.11).

Proof. We define the real Hamiltonian $H_\tau = \frac{Z_N(\tau)}{\tau}$, where

$$Z_N(t) = \sum_{j=0}^{N} t^j Z_j,$$

and where, for $j = 0, \cdots, N$, the polynomials Z_j are defined in Proposition VI.4. By definition, $Z_N(t)(z)$ is a polynomial of order $2N + 2$ and using Proposition VI.4 we get

$$\|Z_N(t)\| \leq \sum_{j=0}^{N} \|Z_j\| < \infty.$$

Thus $Z_N \in \mathcal{SP}_{2N+2}$.
Now, as for all j, Z_j is a homogeneous polynomial of order $2j + 2$, we have, using Proposition III.6 and Proposition VI.4 that for $z \in B^s_M$ with $M < M_0$ and $j \geq 1$,

$$\left\|X_{Z_j}(z)\right\|_{\ell^1_s} \leq C_j \left\|Z_j\right\| \left(\sup_{k=2,\ldots,2j+1} \|z\|^k_{\ell^1_s} \right), \leq D_j \left\|Z_j\right\| M M_0^{2j} \leq M(C_1)^j$$

where the constants D_j are given by Proposition III.6. Note that the constant C_1 depends on P, s, N and M_0. On the other hand we have using Lemma VI.2 and (VI.14),

$$\|X_{Z_0}(z)\|_{\ell^1_s} \leq 2\pi \|z\|_{\ell^1_s} \leq 2\pi M.$$

Hence, for $t \leq (2C_1)^{-1}$ we have

$$\left\|X_{Z_N(t)}(z)\right\|_{\ell^1_s} \leq 2\pi M + M \sum_{j=1}^{N} (tC_1)^j < (2\pi + 1)M < 8M. \tag{VI.19}$$

Therefore by a classical bootstrap argument, the time 1 flow $\Phi^1_{Z_N(t)}$ maps B^s_M into B^s_{9M} provided that $t \leq (2C_1)^{-1}$.
On the other hand, $\varphi^1_{A_0}$ is an isometry of ℓ^1_s and hence maps B^s_M into itself, while using again Proposition III.6, we see that φ^t_P maps B^s_M into B^s_{9M} as long as $t \leq C_2^{-1}$, where C_2 depends on $\|P\|$, M_0 and s. We then define

$$T := \min\left\{(2C_1)^{-1}, C_2^{-1}\right\} \tag{VI.20}$$

and we assume in the sequel that $0 \leq t \leq T$ in such a way that all the flows remain in the ball B_{9M}.
Let $u(t) = \varphi^t_P \circ \varphi^1_{A_0}(z) - \varphi^1_{Z_N(t)}(z)$ and denote by $Q_N(t)$ the Hamiltonian defined by

$$Q_N(t) = \sum_{k \geq 0} \frac{1}{(k+1)!} \mathrm{ad}^k_{Z_N(t)} Z'_N(t).$$

By construction (see (VI.6)), the following relation holds: For $t \leq T$ given in (VI.20), the Hamiltonian $Q_N(t) \in \mathcal{C}^\infty(\ell_s^1, \mathbb{C})$ satisfies for $z \in B_M^s$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi^1_{Z_N(t)}(z) = X_{Q_N(t)} \circ \varphi^1_{Z_N(t)}(z). \tag{VI.21}$$

Using this result, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = X_P \circ \varphi_P^t \circ \varphi_{A_0}^1(z) - X_{Q_N(t)} \circ \varphi^1_{Z_N(t)}(z).$$

As $u(0) = 0$, we get for $t \leq T$ given in (VI.20),

$$\|u(t)\|_{\ell_s^1} \leq \int_0^t \left\| X_P \circ \varphi_P^\sigma \circ \varphi_{A_0}^1(z) - X_{Q_N(\sigma)} \circ \varphi^1_{Z_N(\sigma)}(z) \right\|_{\ell_s^1} \mathrm{d}\sigma$$

and hence

$$\begin{aligned}\|u(t)\|_{\ell_s^1} \leq &\int_0^t \left\| X_P \circ \varphi^1_{Z_N(\sigma)}(z) - X_{Q_N(\sigma)} \circ \varphi^1_{Z_N(\sigma)}(z) \right\|_{\ell_s^1} \mathrm{d}\sigma \\ &+ \int_0^t \left\| X_P \circ \varphi_P^\sigma \circ \varphi_{A_0}^1(z) - X_P \circ \varphi^1_{Z_N(\sigma)}(z) \right\|_{\ell_s^1} \mathrm{d}\sigma.\end{aligned}$$

Therefore for $t \leq T$,

$$\|u(t)\|_{\ell_s^1} \leq \int_0^t \sup_{z \in B_{9M}} \left\| X_P(z) - X_{Q_N(\sigma)}(z) \right\|_{\ell_s^1} \mathrm{d}\sigma + L_P \int_0^t \|u(\sigma)\|_{\ell_s^1} \mathrm{d}\sigma \tag{VI.22}$$

where using equation (III.21) in Proposition III.6, we can take

$$L_P = 4^{s+2} \|P\| \, (9M_0)^2.$$

So it remains to estimate $\sup_{z \in B_{9M}} \left\| X_P(z) - X_{Q_N(t)}(z) \right\|_{\ell_s^1}$ for $z \in B_{9M}$ and $t \leq T$. Now by definition of $Q_N(t)$ we have

$$Z_N'(t) = \sum_{k=0}^\infty \frac{B_k}{k!} \mathrm{ad}^k_{Z_N(t)} Q_N(t),$$

where the right-hand side actually defines a convergent series by the argument used in the proof of Proposition VI.4. By construction (cf. Section 3), we have

$$\sum_{k=0}^\infty \frac{B_k}{k!} \mathrm{ad}^k_{Z_N(t)} (Q_N(t) - P) = \mathcal{O}(t^N)$$

in the sense of real Hamiltonians in the space $\mathcal{C}^\infty(\ell_s^1, \mathbb{C})$. Taking the inverse of the series, we see

$$Q_N(t) - P = \sum_{n \geq N} t^n K_n \tag{VI.23}$$

where we have the explicit expressions

$$K_n = \sum_{\substack{\ell+m=n\\ m<N}} (m+1) \sum_{k\geq 0} \frac{1}{(k+1)!} \sum_{\substack{\ell_1+\cdots+\ell_k=\ell\\ \ell_j\leq N}} \mathrm{ad}_{Z_{\ell_1}} \cdots \mathrm{ad}_{Z_{\ell_k}} Z_{m+1}. \quad \text{(VI.24)}$$

Estimates similar to the one in the proof of Proposition VI.4 lead to

$$\|K_n\| \leq \sum_{\substack{\ell+m=n\\ m<N}} (m+1) \sum_{i=0}^{\ell} (32)^i (n+1)^i \sum_{k\geq i} \frac{(2\pi-\delta)^{(k-i)}}{i!\,(k-i)!}$$
$$\times \sum_{\substack{\ell_1+\cdots+\ell_i=\ell\\ 0<\ell_j\leq N}} \ell_1 \left\|Z_{\ell_1}\right\| \cdots \ell_{i-1} \left\|Z_{\ell_{i-1}}\right\| \ell_i \left\|Z_{\ell_i}\right\| \left\|Z_{m+1}\right\|,$$

where δ is given by (VI.15). Hence, after summing in k,

$$\|K_n\| \leq C_1 \sum_{\substack{\ell+m=n\\ m<N}} (m+1) \sum_{i=0}^{\ell} \frac{(32)^i (n+1)^i}{i!}$$
$$\times \sum_{\substack{\ell_1+\cdots+\ell_i=\ell\\ 0<\ell_j\leq N}} \ell_1 \left\|Z_{\ell_1}\right\| \cdots \ell_{i-1} \left\|Z_{\ell_{i-1}}\right\| \ell_i \left\|Z_{\ell_i}\right\| \left\|Z_{m+1}\right\|,$$

for some constant C_1 depending on δ. Using the estimates in Proposition VI.4, we see that there exist a constant C_3 depending on N such that

$$\|K_n\| \leq C_3^{n+1}.$$

As K_n is a polynomial of order $2n+2 \leq 4n$, we deduce from the previous estimate and Proposition III.6 that, for $z \in B^s_{9M}$,

$$\|X_{K_n}(z)\|_{\ell^1_s} \leq 2(4n)^{s+1} C_3^{n+1} (9M)^{2n+1}.$$

Using (VI.23) and the previous bound, we get

$$\begin{aligned}\left\|X_{Q_N(t)}(z) - X_P(z)\right\|_{\ell^1_s} &\leq \sum_{n\geq N} t^n \left\|X_{K_n}(z)\right\|_{\ell^1_s}\\ &\leq \sum_{n\geq N} t^n 2(4n)^{s+1} C_3^{n+1} (9M)^{2n+1}\\ &\leq C_5 M^{2N+1} t^N,\end{aligned}$$

for $t \leq C_4$ and for some constant C_5, with C_4 and C_5 depending on $\|P\|$, s, δ, M_0 and N.

Let us set

$$\tau_0 = \tau_0(M_0, N, \delta, s, \|P\|) := \min\left\{(2C_1)^{-1}, C_2^{-1}(9M_0)^{-4}, C_4^{-1}\right\}.$$

For $t \leq \tau_0$, inserting the last estimate in (VI.22) we get

$$\|u(t)\|_{\ell_s^1} \leq t^{N+1} M^{2N+1} C_5 + L_P \int_0^t \|u(s)\|_{\ell_s^1} \, \mathrm{d}s$$

and this leads to

$$\|u(t)\|_{\ell_s^1} \leq t^{N+1} M^{2N+1} C_6$$

for some constant C_6 depending on δ, s $\|P\|$, M_0 and N. This implies (VI.16) defining $H_\tau = Z_\tau(\tau)/\tau$ for $\tau \leq \tau_0$.
The second assertion of the theorem is just a calculus defining

$$H_\tau^{(1)} = \frac{1}{\tau} Z_0 + Z_1.$$

Using the previous bounds and the first inequality in Proposition III.6, we then calculate that for $z \in B_M^s$,

$$|H_\tau(z) - H_\tau^{(1)}(z)| \leq \sum_{j=2}^{N} \tau^{j-1} \left\|Z_j(z)\right\|_{\ell_s^1}$$

which yields the result. ■

We conclude this section by giving explicitly the CFL condition (VI.13) required to obtain a given precision τ^{N+1} in the previous theorem. The numbers given in Table VI.1 are given by the function $\beta^{-1}(\frac{2\pi}{N+1})$.

τ^{N+1}	$\beta(x) = x$	$\beta(x) = 2\arctan(x/2)$
τ^2	3.14	∞
τ^3	2.10	3.46
τ^4	1.57	2.00
τ^5	1.27	1.45
τ^6	1.05	1.15
τ^7	0.90	0.96
τ^8	0.80	0.83
τ^9	0.70	0.73
τ^{10}	0.63	0.65

Table VI.1. CFL conditions for cubic NLS

Note that in the case of a general polynomial nonlinear term in (VI.1), the previous theorem remains true, but the CFL condition depends on the degree of the polynomial. We refer to [15] for a complete analysis.

The previous construction allows us to prove the preservation of the modified energy over a long time depending on N:

Corollary VI.6. *Under the hypothesis of the previous theorem, let $z^0 = (\xi^0, \bar{\xi}^0) \in \ell^1$ and the sequence z^n defined by*

$$z^{n+1} = \varphi_P^\tau \circ \varphi_{A_0}^1(z^n), \quad n \geq 0, \tag{VI.25}$$

for $\tau \leq \tau_0$. Assume that for all n, the numerical solution z^n remains in a ball B_M of ℓ^1 for a given $M < M_0$. Then there exists a constant c such that,

$$H_\tau(z^n) = H_\tau(z^0) + \mathcal{O}\left(M^{2N+1}\right), \quad \textit{for} \quad n\tau \leq c\tau^{-N}.$$

Proof. As all the Hamiltonian functions considered are real (and in fact homogeneous symmetric polynomials), we have for all n, $z^n = (\xi^n, \bar{\xi}^n)$, i.e. z^n is real. Hence for all n, $H_\tau(z^n) \in \mathbb{R}$.
We use the notation of the previous theorem and we notice that $H_\tau(z)$ is a conserved quantity by the flow generated by H_τ. Therefore we have

$$H_\tau\left(z^{n+1}\right) - H_\tau(z^n) = H_\tau\left(\varphi_P^\tau \circ \varphi_{A_0}^1(z^n)\right) - H_\tau\left(\varphi_{H_\tau}^\tau(z^n)\right)$$

and hence

$$\left|H_\tau\left(z^{n+1}\right) - H_\tau(z^n)\right| \leq \left(\sup_{z \in B_{2M_0}} \|\nabla H_\tau(z)\|_{\ell^\infty}\right) \left\|\varphi_P^\tau \circ \varphi_{A_0}^1(z^n) - \varphi_{H_\tau}^\tau(z^n)\right\|_{\ell^1}.$$

Now using (VI.19) and the fact that $z^n \in B_M$, we obtain for all n,

$$\left|H_\tau\left(z^{n+1}\right) - H_\tau(z^n)\right| \leq 4\pi C_N M^{2N+1}\tau^{N+1},$$

and hence

$$\left|H_\tau\left(z^n\right) - H_\tau\left(z^0\right)\right| \leq (n\tau)c^{-1}M^{2N+1}\tau^N, \tag{VI.26}$$

for some constant c. This implies the result. ■

4 Fully discrete scheme

In this section, we consider the case of fully discrete approximations of the solution $u(t,x)$ of the cubic NLS equation (VI.1), obtained by splitting methods. Let us recall that the discrete Hamiltonian associated with the cubic nonlinear Schrödinger equation is given by (see (III.40))

$$H^K(\xi,\eta) = T^K + P^K := \sum_{a\in B^K} a^2 \xi_a \eta_a + \frac{\lambda}{2} \sum_{\substack{a_1+a_2-a_3-a_4=mK\\ a_i\in B^K,\, |m|\le 1}} \xi_{a_1}\xi_{a_2}\eta_{a_3}\eta_{a_4}, \tag{VI.27}$$

where we recall that B^K is the finite set of indices defined in (III.35).

Note that with the notation of Chapter III, we have (see Definition III.16),

$$P^K \in \mathcal{P}^K_{4,1}, \quad \text{and} \quad \left\| P^K \right\| = 3.$$

Moreover, P^K is homogeneous of degree 4, and symmetric in (ξ,η): with the definition (III.17) we have $P^K \in \mathcal{S}\mathcal{P}^K_{4,1}$. In particular the condition (IV.17) is satisfied (with $C_0 = 3$).

Recall moreover that the exact flows of P^K, T^K and H^K preserve the space (see (III.51))

$$\mathcal{A}^K := \left\{ z_a = (\xi_a,\eta_a) \,|\, \xi_a = \eta_a = 0 \quad \text{if} \quad a \notin B^K \right\} \subset \ell^1.$$

In this section, we consider the fully discrete splitting method

$$\varphi^\tau_{P^K} \circ \varphi^1_{A_0^K} \simeq \varphi^\tau_{H^K}$$

where A_0 is a diagonal operator

$$A_0 = \sum_{a\in B^K} \lambda_a \xi_a \eta_a, \quad \text{with} \quad \lambda_a = \beta(\tau a^2), \quad a\in\mathbb{Z},$$

with $\beta(x) = x$ or $\beta(x) = 2\arctan(x/2)$. Note that here the natural CFL number is given by $c_0 = \tau K^2/4$, as all the frequencies of the linear operator are smaller than $K/2$.

We see that we can follow the same strategy as in Section 1: We seek a modified Hamiltonian $Z^K(\tau) = A_0^K + \tau Z_1^K + \tau^2 Z_2^K + \cdots$ such that for all $t < \tau$, we have

$$\varphi^t_{P^K} \circ \varphi^1_{A_0^K} = \varphi^1_{Z^K(t)}.$$

As before, the equations for the Hamiltonian functions Z_n^K are given by (compare (VI.28)).

$$(n+1) Z^K_{n+1} = \sum_{k\ge 0} \frac{B_k}{k!} \sum_{\ell_1+\cdots+\ell_k=n} \mathrm{ad}_{Z^K_{\ell_1}} \cdots \mathrm{ad}_{Z^K_{\ell_k}} P^K. \tag{VI.28}$$

The following proposition is the extension of Proposition VI.4 to this fully discrete case:

Proposition VI.7. *Let $N > 0$ be fixed, and let P^K be the Hamiltonian (VI.27). Assume that τ and K satisfy the condition*

$$\tau K^2 < 4\beta^{-1}\left(\frac{2\pi}{N+1}\right). \tag{VI.29}$$

Then for $n \leq N$ we can define homogeneous symmetric polynomials $Z_n^K \in \mathcal{SP}_{2n+2,3n}^L$ satisfying the equations (VI.9) up to the order n, and such that there exist constants $C_n < \infty$ independent of K, and such that $\left\|Z_n^K\right\| < C_n$.

Proof. The proof is very similar to the proof of Proposition VI.4. The condition (VI.29) corresponds to the CFL regime (VI.13). The fact that the norms of the polynomials Z_n^K are independent of K is a consequence of the estimate (III.47) of Proposition III.18. ∎

The following result is the fully discrete version of the backward error analysis Theorem VI.5. Recall that we have $\mathcal{A}^K \subset \ell^1$, and that B_M denotes the ball of radius M in ℓ^1.

Theorem VI.8. *Let $N \geq 1$ and $M_0 \geq 1$ be fixed, and P^K the family of Hamiltonian functions (VI.27) depending on K. Then there exist constants τ_0 and C_N depending on M_0 and N such that the following holds: For all $\tau \leq \tau_0$ and $K \geq 1$ satisfying the CFL condition (VI.29), there exists a real Hamiltonian polynomial $H_\tau^K \in \mathcal{SP}_{2N+2,3N}$ such that for all $M < M_0$ and $z^K \in \mathcal{A}^K \cap B_M$, we have*

$$\left\|\varphi_{P^K}^{\tau} \circ \varphi_{A_0^K}^{1}\left(z^K\right) - \varphi_{H_\tau^K}^{\tau}\left(z^K\right)\right\|_{\ell^1} \leq C_N M^{2N+1}\tau^{N+1}. \tag{VI.30}$$

Moreover, for $z^K \in \mathcal{A}^K \cap B_M$ we have

$$\left|H_\tau^K\left(z^K\right) - H_\tau^{K,(1)}\left(z^K\right)\right| \leq C_N \tau M^6 \tag{VI.31}$$

where

$$H_\tau^{K,(1)}(z) = \sum_{a\in B^K}\frac{1}{\tau}\beta\left(\tau a^2\right)\xi_a\eta_a + \frac{\lambda}{2}\sum_{\substack{a+b-c-d=mK\\(a,b,c,d)\in\left(B^K\right)^4,\,|m|\leq 1}}\frac{i\Lambda_{abcd}}{\exp\left(i\Lambda_{abcd}\right)-1}\xi_a\xi_b\eta_c\eta_d, \tag{VI.32}$$

where we recall that $\Lambda_{abcd} = \lambda_a + \lambda_b - \lambda_c - \lambda_d$ with $\lambda_a = \beta(\tau a^2)$, $a \in B^K$.

Proof. The proof of this theorem is straightforward using the proof developed in the previous Section. The key argument is that the constants appearing in the control of the Hamiltonian polynomials Z_j, $j = 1, \dots, N$, do not depend on K by the result of Proposition III.18. ■

As in the linear case, the specificity of the fully discrete case with CFL is that the Hamiltonian

$$\sum_{a \in B^K} \frac{1}{\tau} \beta(\tau \omega_a) \xi_a \eta_a$$

controls the H^1 norm of $z^K \in \mathcal{A}^K$. In particular, we get the following result:

Proposition VI.9. *Under the hypothesis of the previous theorem, there exists a constant c depending on N, M_0 and τ_0 such that the following holds: Let $z^{K,n} = (\xi^{K,n}, \bar{\xi}^{K,n})$ be the sequence defined by*

$$z^{K,n+1} = \varphi^{\tau}_{P^K} \circ \varphi^{1}_{A^K_0} \left(z^{K,n} \right)$$

with initial value $z^{K,0} \in \mathcal{A}^K$, and where $\tau \leq \tau_0$ and K satisfy the CFL condition (VI.29). Assume that for all K and all n, we have $z^{K,n} \in B_M$ with $M \leq \min(M_0, 1)$. Then for all $n\tau \leq c\tau^{-N}$, we have

$$\left| \sum_{a \in B^K} \frac{1}{\tau} \beta \left(\tau |a|^2 \right) |(\xi^{K,n})_a|^2 - \sum_{a \in B^K} \frac{1}{\tau} \beta(\tau |a|^2) |(\xi^{K,0})_a|^2 \right| \leq C M^4, \qquad \text{(VI.33)}$$

for some constant C independent of K.

Proof. The proof is the same as the proof of Corollary VI.6: we first can prove (compare (VI.26)) that for all n,

$$\left| H^K_\tau (z^n) - H^K_\tau \left(z^0 \right) \right| \leq (n\tau) c^{-1} M^{2N+1} \tau^N,$$

where H^K_τ is defined in Theorem VI.8, and for some constant c independent of M, τ and n. But this shows that

$$\left| H^K_\tau (z^n) - H^K_\tau (z^0) \right| \leq M^{2N+1},$$

for $n\tau \leq c\tau^{-N}$. Using estimates on the homogeneous polynomials Z_j of degree $2j + 2$, we obtain that (VI.33) holds with an error of the form

$$\mathcal{O} \left(M^4 + M^6 + \cdots + M^{2N+1} \right) = \mathcal{O}(M^4)$$

provided that $M \leq 1$ (and $N \geq 2$). ■

In particular, we get the following corollary, which is the fully discrete version of the global existence result of Proposition III.14:

Corollary VI.10. *Under the assumption of Theorem VI.8, there exist ε_0 such that the following holds: For all $\varepsilon \le \varepsilon_0$, $K \ge 1$ and $\tau \le \tau_0$ satisfying (VI.29), let $U^{K,0}(x) = \sum_{a \in B^K} \xi_a^{K,0} e^{ia\,x}$ and assume that*

$$\left\| U^{K,0} \right\|_{H^1} = \left\| \xi_a^{K,0} \right\|_{\ell_1^2} = \varepsilon.$$

Then if $U^{K,n}(x) = \sum_{a \in B^K} \xi_a^{K,n} e^{ia\,x}$ is the function obtained after n iterations of the splitting method $\varphi^1_{A_0^K} \circ \varphi^\tau_{P^K}$, we have

$$\forall n\tau \le c\tau^{-N}, \qquad \left\| U^{K,n} \right\|_{H^1} \le A\varepsilon$$

for some constants c and A independent of K.

Proof. We set for all $n \ge 0$, $z^{K,n} = (\xi^{K,n}, \bar{\xi}^{K,n}) \in \mathcal{A}^K$. Recall that there exists a constant C such that for all $z \in \mathcal{A}^K$, we have (see (III.12))

$$\|z\|_{\ell^1} \le C \, \|z\|_{\ell_1^2}. \tag{VI.34}$$

Hence we have that $z^{K,0} \in B_{C\varepsilon}$. Let us set $M = A\varepsilon$, where the constant A will be defined later. We can always assume that $M \le 1 \le M_0$ in Theorem VI.8, and that $A \ge C$.

As long as $z^{K,n} \in B_M$ and $n\tau \le c\tau^{-N}$ the preservation relation (VI.33) implies that

$$\begin{aligned}\sum_{a \in B^K} \frac{1}{\tau} \beta\left(\tau |a|^2\right) \left| \left(\xi^{K,n}\right)_a \right|^2 &\le C_1 M^4 + \sum_{a \in B^K} \frac{1}{\tau} \beta\left(\tau |a|^2\right) \left| \left(\xi^{K,0}\right)_a \right|^2 \\ &\le C_1 A^4 \varepsilon^4 + \varepsilon^2,\end{aligned} \tag{VI.35}$$

for some constants C_1 independent of K. Here, we used the fact that $\beta(x) \le x$ for all $x \ge 0$, and that $\left\| \xi_a^{K,0} \right\|_{\ell_1^2} = \varepsilon$.

Now for all $a \in B^K$, using equation (VI.29), we have $\tau a^2 \le 4\beta^{-1}(2\pi/N + 1)$, and hence there exists $\beta_0 \le 1$ such that, using the fact that for all $a \in B^K$, $\beta(\tau a^2) \ge \beta_0 \tau a^2$. Using (VI.35), we get the following: as long as $z^{K,n} \in B_M$ and $n\tau \le c\tau^{-N}$, we have

$$\left\| U^{K,n} \right\|_{H^1}^2 \le \frac{1}{\beta_0} \sum_{a \in B^K} \frac{1}{\tau} \beta\left(\tau |a|^2\right) \left| \left(\xi^{K,n}\right)_a \right|^2 \le \frac{1}{\beta_0} C_1 A^4 \varepsilon^4 + \frac{1}{\beta_0} \varepsilon^2$$

and hence using (VI.34),

$$\left\| z^{K,n} \right\|_{\ell^1}^2 \le \frac{C^2}{\beta_0} C_1 A^4 \varepsilon^4 + \frac{C^2}{\beta_0} \varepsilon^2.$$

Taking A such that $A^2 = \frac{2C^2}{\beta_0} \geq C^2$, there exists ε_0 such that for $\varepsilon \leq \varepsilon_0$, the previous relation yields

$$\left\| z^{K,n} \right\|_{\ell^1}^2 \leq A^2 \varepsilon^2 = M^2.$$

This implies that for all $n\tau \leq C_N \tau^{-N}$, $\left\| z^{K,n} \right\|_{\ell^1} \leq M$, and the relation (VI.35) yields (as we have $C \geq 1$ in (VI.34))

$$\left\| U^{K,n} \right\|_{H^1}^2 \leq A^2 \varepsilon^2$$

and this finishes the proof. ■

VII Introduction to long time analysis

In this chapter, we still consider the cubic nonlinear Schrödinger equation

$$i\partial_t u = -\Delta u + \lambda |u|^2 u, \tag{VII.1}$$

on the torus $\mathbb{T}^d$ with $d = 1$ or $d = 2$, and we will assume that the initial condition is small, i.e. that $u(0,x)$ is of order $\delta > 0$ in ℓ^1, with $\delta \to 0$ a small parameter. We are interested in the behavior of the solution with respect to δ. By making the change of unknown $u \mapsto u/\delta$, and by setting $\varepsilon = \delta^2$, we see that it is equivalent to consider the family of equations

$$i\partial_t u = -\Delta u + \varepsilon\lambda |u|^2 u, \quad u(0,x) = u^0(x) \simeq 1, \tag{VII.2}$$

where $\varepsilon \to 0$ is a small parameter and u^0 is fixed and independent of ε. Here we will assume that $\lambda \in \{\pm 1\}$.

As we shall see below, this equation is *resonant* because all the frequencies of the linear operator $-\Delta$ are integers. In contrast, when the linear operator is slightly perturbed, and of the form $u \mapsto -\Delta + V \star u$ for some potential V, then for a large class of potential V, the frequencies become *non resonant*, and many results exist concerning the long time behavior of the corresponding solution $u(t,x)$. Note that the fact that V acts as a convolution and not as a multiplication allows us to calculate explicitly the spectrum of the operator $-\Delta + V\star$ which turns out to be diagonal in the Fourier basis. Such a model has become popular to analyze resonances effects and the long time behavior of (VI.1), see for instance [6], [11], [16], [22] and the references therein. Typically, for a *generic* potential V making the frequencies non resonant, it can be shown that the preservation of the action – see (I.19) – holds for a very long time or order $c_r\varepsilon^{-r}$ for all r (with a constant c_r depending on r) or even exponentially large when the solution is analytic.

Here, we will only consider the case where $V = 0$, and show that the situation differs significantly between the dimension 1 or 2. In the sequel, we will not give results for *very* long times (of order ε^{-r} for any r), but only for times of order ε^{-1}. We will then discuss the behaviors of fully discrete numerical solutions over this time scale.

In a first step, we will consider the case of the dimension 1, and we will prove that no significant energy exchanges between the modes ξ_a, $a \in \mathbb{Z}$ can be observed: the *actions* $|\xi_a|^2$ of the solution are almost preserved, see (I.19), over a long time $t \leq \varepsilon^{-1}$. The situation is very different in dimension 2, where it can be proved that there is an *energy cascade* transferring energy from the low modes to arbitrary high modes (see Figure I.12 and [7]). In these two situations, we will then analyze the long time behavior of the fully discrete solutions obtained by the splitting schemes studied in the previous chapters. The main tool will be the modified energy constructed above.

1 Resonant system

The goal of this Section is to show that the understanding of the qualitative behavior of the solution of (VII.2) relies on the analysis of a *resonant* system driving the energy exchanges between the modes, at least in the time scale $t \leq \varepsilon^{-1}$.

1.1 An approximation result. With the notation of Chapter III, equation (VII.2) can be written

$$\begin{aligned} \dot\xi_a &= -i|a|^2\xi_a - i\varepsilon\lambda \sum_{a=b-c+d} \xi_b\eta_c\xi_d, \\ \dot\eta_a &= i|a|^2\eta_a + i\varepsilon\lambda \sum_{a=b-c+d} \eta_b\xi_c\eta_d. \end{aligned} \tag{VII.3}$$

The next result shows that over the time scale $\mathcal{O}(\varepsilon^{-1})$ the solution of the previous system is well represented by the superposition of the solution of the linear flow and the solution of a resonant system. This is a relatively standard result. We refer to [7] for a discussion on more general approaches and references.

Proposition VII.1. *Let $z^0 = (\xi^0, \bar\xi^0) \in \ell^1$. There exist constants ε_0, C and $T > 0$ such that for all $\varepsilon < \varepsilon_0$ and for $t \in [0, T/\varepsilon]$, there exists $z(t) = (\xi(t), \bar\xi(t))$ a solution to* (VII.3) *with $z(0) = z^0$. Moreover, we have*

$$\sum_{a\in\mathbb{Z}^d} \left| \xi_a(t) - e^{-it|a|^2} y_a(\varepsilon t) \right| \leq C\varepsilon, \tag{VII.4}$$

where $y(t) = (y_a(t))_{a\in\mathbb{Z}^d}$ is the solution of the system

$$\dot y_a = -i\lambda \sum_{\substack{a=b-c+d \\ |a|^2=|b|^2-|c|^2+|d|^2}} y_b\bar y_c y_d, \tag{VII.5}$$

with initial value $y(0) = \xi^0$, and for $t \in [0, T]$.

Proof. We define $Y(t)$ by

$$Y_a(t) = e^{it|a|^2}\xi_a(t).$$

We calculate that $Y(t)$ satisfies the equation

$$\dot Y_a = -i\varepsilon\lambda \sum_{a=b-c+d} Y_b\bar Y_c Y_d e^{-it\Omega_{acbd}}$$

where $\Omega_{acbd} = |a|^2 + |c|^2 - |b|^2 - |d|^2$, or equivalently

$$Y_a(t) = \xi_a^0 - i\varepsilon\lambda \sum_{a=b-c+d} \int_0^t Y_b\bar Y_c Y_d e^{-is\Omega_{acbd}} \mathrm{d}s.$$

By making a change of time $\sigma = \varepsilon t$, we define $\mathrm{y}(\sigma) = Y(\sigma/\varepsilon) = Y(t)$. This function satisfies

$$\begin{aligned}\mathrm{y}_a(\sigma) &= \xi^0 - i\lambda \sum_{a=b-c+d} \int_0^\sigma \mathrm{y}_b(s)\bar{\mathrm{y}}_c(s)\mathrm{y}_d(s)e^{-i\frac{s}{\varepsilon}\Omega_{acbd}}\mathrm{d}s \\ &= \xi^0 - i\lambda \sum_{\substack{a=b-c+d\\|a|^2=|b|^2-|c|^2+|d|^2}} \int_0^\sigma \mathrm{y}_b(s)\bar{\mathrm{y}}_c(s)\mathrm{y}_d(s)\mathrm{d}s + R_a(\mathrm{y},\varepsilon)\end{aligned}$$

where

$$R_a(\mathrm{y},\varepsilon) = -i\lambda \sum_{\substack{a=b-c+d\\ \Omega_{acbd}\neq 0}} \int_0^\sigma \mathrm{y}_b(s)\bar{\mathrm{y}}_c(s)\mathrm{y}_d(s)e^{-i\frac{s}{\varepsilon}\Omega_{acbd}}\mathrm{d}s.$$

Assume that $\|\mathrm{y}(\sigma)\|_{\ell^1} \leq M$ for $\sigma \in [0,T]$. Then we have from the previous equation that $\left\|\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathrm{y}(\sigma)\right\|_{\ell^1} \leq TM^3$ for $\sigma \in [0,T]$. Using an integration by parts, we thus see that

$$\begin{aligned}R_a(\mathrm{y},\varepsilon) = \varepsilon\lambda \sum_{\substack{a=b-c+d\\ \Omega_{acbd}\neq 0}} \Bigg(&\int_0^\sigma \frac{e^{-i\frac{s}{\varepsilon}\Omega_{acbd}}}{\Omega_{acbd}} \frac{\mathrm{d}}{\mathrm{d}s}\left(\mathrm{y}_b(s)\bar{\mathrm{y}}_c(s)\mathrm{y}_d(s)\right)\mathrm{d}s \\ &- \left[\frac{e^{-i\frac{s}{\varepsilon}\Omega_{acbd}}}{\Omega_{acbd}}\mathrm{y}_b(s)\bar{\mathrm{y}}_c(s)\mathrm{y}_d(s)\right]_0^\sigma\Bigg)\end{aligned}$$

and hence as $\Omega_{acbd} \neq 0$ implies $|\Omega_{acbd}| \geq 1$, we see that

$$\|R(\mathrm{y},\varepsilon)\|_{\ell^1} \leq \varepsilon C(M) \tag{VII.6}$$

where the constant C depends on M.

Now let T be such that there exists a solution $y(\sigma)$ to the equation (VII.5) on the time interval $\sigma \in [0,T]$. Such a time is given by Proposition III.6. In particular, there exists M such that $\|y(\sigma)\|_{\ell^1} \leq M/2$ for $\sigma \in [0,T]$.

We define the vector field F by the relation

$$F_a(\xi) = -i\lambda \sum_{\substack{a=b-c+d\\|a|^2=|b|^2-|c|^2+|d|^2}} \xi_b\bar{\xi}_c\xi_d. \tag{VII.7}$$

We have by definition

$$y_a(\sigma) = \xi^0 + \int_0^\sigma F_a(y(s))\,\mathrm{d}s$$

for $\sigma \in [0,T]$, and moreover

$$\mathrm{y}_a(\sigma) = \xi^0 + \int_0^\sigma F_a(\mathrm{y}(s))\,\mathrm{d}s + R(\mathrm{y},\varepsilon).$$

Using the fact that F is Lipschitz on the bounded sets of ℓ^1, we see that there exists a constant L such that as long as $\|\mathrm{y}(\sigma)\|_{\ell^1} \le M$, we can write using (VII.6)

$$\|y(\sigma) - \mathrm{y}(\sigma)\|_{\ell^1} \le L \int_0^\sigma \|y(s) - \mathrm{y}(s)\|_{\ell^1}\,\mathrm{d}s + \varepsilon\, C(M).$$

Using the Gronwall Lemma, we first see that if $\varepsilon < \varepsilon_0$ is small enough, we have $\|\mathrm{y}(\sigma)\|_{\ell^1} \le M$ for all times $\sigma \in [0,T]$ and moreover,

$$\|y(\sigma) - \mathrm{y}(\sigma)\|_{\ell^1} \le C\varepsilon$$

where C depends on T and M. Going back to the original time $t = \sigma/\varepsilon$ then yields the result. ■

1.2 The resonance modulus. We consider the resonant system (VII.5) associated with the vector field F_a defined in (VII.7). First, we note that this equation is again a Hamiltonian equation associated with the Hamiltonian

$$Z(\xi,\eta) = \frac{\lambda}{2} \sum_{\substack{a+b=c+d \\ |a|^2+|b|^2=|c|^2+|d|^2}} \xi_a \xi_b \eta_c \eta_d. \tag{VII.8}$$

To understand the qualitative behavior of $y(t)$, we are led to study the *resonant set*

$$\mathcal{K} = \{a,b,c,d \in \mathbb{Z}^d \mid a+b-c-d = 0, \text{ and } |a|^2+|b|^2-|c|^2-|d|^2 = 0\}, \tag{VII.9}$$

which drives the energy exchanges between the frequencies in the dynamical system (VII.5). The following lemma describes the geometric structure of this set of frequencies:

Lemma VII.2. *A quadruplet $(a,b,c,d) \in \mathbb{Z}^d$ is in $\mathcal{K}$ precisely when the endpoints of the vectors a,b,c,d form four corners of a non-degenerate rectangle with a and b opposing each other, or when this quadruplet corresponds to one of the two following degenerate cases: $(a=c, b=d)$, or $(a=d, b=c)$.*

Proof. Using the first relation, we obtain $|a+b|^2 = |c+d|^2$, and hence $a\cdot b = c\cdot d$. Then we calculate that

$$\begin{aligned}(a-d)\cdot(b-d) &= a\cdot b - d\cdot(a+b-d)\\ &= -d\cdot(a+b-c-d) = 0,\end{aligned}$$

which implies the statement. ■

Note that in dimension 1, only the second part of this lemma is applicable. We will first study this situation for the continuous, semi-discrete and fully discrete solutions.

2 The one-dimensional case

2.1 Long time preservation of the actions. As seen from Lemma VII.2, the resonant Hamiltonian Z defined by (VII.8) does not contain many terms in dimension 1: only monomials $\xi_a \xi_b \eta_c \eta_d$ such that $(a = c, b = d)$, or $(a = d, b = c)$. We thus have explicitly

$$Z = \frac{\lambda}{2}\left(\sum_{a \in \mathbb{Z}} I_a^2 + 2 \sum_{\substack{a,b \in \mathbb{Z} \\ a \neq b}} I_a I_b \right) \tag{VII.10}$$

which means that it depends *only on the actions*

$$I_a(z) := \xi_a \eta_a, \quad a \in \mathbb{Z}^d. \tag{VII.11}$$

Note that for real $z = (\xi, \bar{\xi})$ and in any dimension d, we have that $I_a(z) = |\xi_a|^2$. Thus this term represents the energy of the mode corresponding to $a \in \mathbb{Z}^d$ in the function $u(x) = \sum_{a \in \mathbb{Z}^d} \xi_a e^{i a \cdot x}$.

In view of (VII.10), we have for $d = 1$,

$$\forall a \in \mathbb{Z}, \quad \{I_a, Z\} = 0,$$

which means that for all $a \in \mathbb{Z}$ the actions $I_a(y(t))$ are constant along the solution $y_a(t)$ of (VII.5). As a consequence, we get the following result, which is valid only in dimension 1.

Theorem VII.3. *Let $z^0 = (\xi^0, \bar{\xi}^0) \in \ell^1$. There exist constants ε_0, C and $T > 0$ such that for all $\varepsilon < \varepsilon_0$ and $t \in [0, T/\varepsilon]$, there exists $z(t) = (\xi(t), \bar{\xi}(t))$ a solution to* (VII.3) *with $z(0) = z^0$. Moreover, we have for all $t \leq T/\varepsilon$,*

$$\sum_{a \in \mathbb{Z}} |I_a(z(t)) - I_a(z(0))| \leq C\varepsilon. \tag{VII.12}$$

Proof. Using (VII.4) and (VII.5), there exists M such that for all $\varepsilon \leq \varepsilon_0$ and $t \leq T/\varepsilon$, we have $\|z(t)\|_{\ell^1} \leq M$. For $a \in \mathbb{Z}$, we can write

$$\begin{aligned} &|I_a(z(t)) - I_a(z(0))| \\ &\quad \leq |I_a(z(t)) - I_a(y(t))| + |I_a(y(t)) - I_a(y(0))| + |I_a(y(0)) - I_a(z(0))|. \end{aligned}$$

In the right-hand side of this equation, we have used the previous remarks that $I_a(y(t)) = I_a(y(0))$ for all $t \leq T/\varepsilon$, so the second term cancels. Moreover, the third term vanishes, as $y(0) = z(0)$. For the first term, we first note that for all $a \in \mathbb{Z}$, $I_a(y(t)) = I_a(\tilde{y}(t))$ where $\tilde{y}(t)$ is defined by $\tilde{y}_a(t) = e^{-it|a|^2} y_a(t)$, $a \in \mathbb{Z}$. Hence as $z(t)$ and $\tilde{y}(t)$ remain bounded by M in ℓ^1, we easily obtain

$$\sum_{a\in\mathbb{Z}} |I_a(z(t)) - I_a(y(t))| \le \sum_{a\in\mathbb{Z}} 2M|y_a(t) - \tilde{y}_a(t)| \le 2MC\varepsilon$$

using (VII.4). This yields the result up to a slight modification of the constant C. ■

The previous result shows the *almost preservation of the actions* over a time of order ε^{-1}. Note that this time is larger than the one given by standard *a priori* estimates applied to (VII.2).

This result holds in the very specific case of the resonant cubic nonlinear Schrödinger equation in dimension 1. We will see later that this result is no longer true in dimension 2. Moreover, it is important to mention that the preservation result (VII.12) holds true for much longer times using the *integrability* of NLS in dimension 1 (see [37]) and the existence of a *global* change of variable putting the system into an integrable form preserving some modified actions (see [23]).

We will not give more details, but will rather discuss the ability of numerical schemes to reproduce the qualitative behavior described by the preservation result (VII.12).

2.2 Aliasing and numerical resonances in dimension one. As we have seen in Chapter III, the semi-discrete Hamiltonian obtained by Fourier pseudo-spectral method applied to (VII.2) is given by (see (III.40))

$$H^K(\xi,\eta) = \sum_{a\in B^K} |a|^2\xi_a\eta_a + \frac{\lambda}{2}\sum_{\substack{a+b-c-d=mK\\ a_i\in B^K,\,|m|\le 1}} \xi_a\xi_b\eta_c\eta_d.$$

where B^K is the discrete set of frequencies defined by (III.35). The solution $z^K(t) = (\xi^K(t),\eta^K(t))$ to the corresponding Hamiltonian system satisfies the equation (compare (VII.3)), for all $a\in B^K$,

$$\begin{aligned}\dot{\xi}_a^K &= -i|a|^2\xi_a^K - i\varepsilon\lambda \sum_{\substack{a=b-c+d+mK\\ m=-1,0,1}} \xi_b^K\eta_c^K\xi_d^K,\\ \dot{\eta}_a^K &= i|a|^2\eta_a^K + i\varepsilon\lambda \sum_{\substack{a=b-c+d+mK\\ m=-1,0,1}} \eta_b^K\xi_c^K\eta_d^K.\end{aligned} \tag{VII.13}$$

We see that we can perform the same analysis as before, using the method of Proposition VII.1. We get the following semi-discrete approximation result:

Proposition VII.4. *Let $C_0 > 0$ be a constant. There exist constants ε_0, C and $T > 0$ such that for all $\varepsilon < \varepsilon_0$ all $K \ge 1$ and all $z^{K,0} = (\xi^{K,0},\bar{\xi}^{K,0})$ such that $\left\|z^{K,0}\right\|_{\ell^1} \le$*

C_0, there exists for $t \in [0, T/\varepsilon]$ a solution $z^K(t) = (\xi^K(t), \bar{\xi}^K(t))$ to (VII.13) with $z(0) = z^{K,0}$. Moreover, we have

$$\sum_{a \in B^K} \left| \xi_a^K(t) - e^{-it|a|^2} y_a^K(\varepsilon t) \right| \leq C\varepsilon, \tag{VII.14}$$

where $y^K(t) = (y_a^K(t))_{a \in \mathbb{Z}^d}$ is the solution of the semi-discrete resonant system

$$\dot{y}_a^K = -i\lambda \sum_{\substack{a=b-c+d+mK \\ m=-1,0,1 \\ |a|^2=|b|^2-|c|^2+|d|^2}} y_b^K \bar{y}_c^K y_d^K, \tag{VII.15}$$

with initial value $y^K(0) = \xi^{K,0}$, and for $t \in [0, T]$.

The proof of this proposition is actually exactly the same as the proof of Proposition VII.1, using the fact that ℓ^1 estimates are the same with or without the aliasing relation between the frequencies.

We see that the main difference with the continuous problem studied in the previous section is that we have to consider the *discrete resonant set* defined by

$$\begin{aligned}\mathcal{K}^K = \big\{(a,b,c,d) \in (B^K)^4 \big| |a|^2 + |b|^2 - |c|^2 - |d|^2 = 0 \quad \text{and} \\ a + b - c - d = mK, \quad m \in \{0, \pm 1\}\big\}.\end{aligned}$$

As we will see in the lemma below, this discrete resonant set is more complicated than for the continuous case, and depends on the arithmetic nature of the integer K.

Lemma VII.5. *The following holds:*

(i) Assume that $K \geq 3$ is a prime number. Then $\mathcal{K}^K$ contains only terms such that $a = d$ and $b = c$, or $b = d$ and $a = c$.

(ii) Assume that $K/2 \geq 3$ is a prime number. Then $\mathcal{K}^K$ contains only terms such that $a = d$ and $b = c$, or $b = d$ and $a = c$, and the terms such that $a = -d$ and $b = -c$, or $a = -c$ and $b = -d$, under the constraint $a + b = mK/2$, $m = \pm 1$.

Proof. Calculating modulo K, we see from the first relation that

$$|a + b|^2 = |c + d|^2 \quad \text{modulo } K,$$

and hence $2ab = 2cd$ modulo K. Using the same calculation as in the proof of Lemma VII.2, we see that we must have

$$2(a - d)(b - d) = 0 \quad \text{modulo } K. \tag{VII.16}$$

(i) Assume that K is prime, then the previous relation implies that $a = d$ or $b = d$ modulo K. But as the indices are in $B^K = \{-P, \ldots, P\}$ where $K = 2P + 1$, we must have $a = d$ or $b = d$. This shows the first part of the lemma.

(ii) To prove the second, let us write $K = 2P$ with P prime. In this situation, we have $B^K = \{-P, \dots, P-1\}$. The relation (VII.16) implies now that $a = d$ or $b = d$ modulo P. Hence we can have for instance $a = d + \delta P$ with $\delta \in \{0, \pm 1\}$. In this situation, we have $a+b-c-d = 2mP = b-c+\delta P$, which implies $b = c + (2m-\delta)P$. If $\delta = 0$, we have $a = d$ and $b = c + 2mP$ which is possible only for $m = 0$ (as b and c are in B^K).

When $\delta = \pm 1$, we must have for the same reason $m = \delta$, and $b = c + mP$ and $a = d + mP$.
This implies that $a^2 = d^2 + 2ma\ P - P^2$ and $b^2 = c^2 + 2mcP + P^2$ and hence

$$a^2 + b^2 - c^2 - d^2 = 2mP(a+c) = 0.$$

But this shows that $a + c = 0$, and hence $a = -c$. From this relation, we easily deduce that $a + b = mP$, and $b = -a + mP = -d$. The other cases are treated similarly.
Conversely, we easily verify that all the points $a = -c$ and $b = -d$ or $a = -d$ and $b = -c$ under the constraint $a + b = mP$ are all in $\mathcal{K}^K$. ■

The following result constitutes a semi-discrete version of Theorem VII.3. It is a consequence of the previous lemma and of the approximation result given by Proposition VII.4:

Theorem VII.6. *Let $C_0 > 0$ be a constant. There exist constants ε_0, C and $T > 0$ such that for all $\varepsilon < \varepsilon_0$, all $K \geq 1$ and all $z^{K,0} = (\xi^{K,0}, \bar{\xi}^{K,0})$ such that $\left\|z^{K,0}\right\|_{\ell^1} \leq C_0$, there exists $z^K(t) = (\xi^K(t), \bar{\xi}^K(t))$ a solution to (VII.13) for $t \in [0, T/\varepsilon]$ with $z^K(0) = z^{K,0}$. Moreover, we have the following preservation properties: for all $t \leq T/\varepsilon$,*

(i) If K is a prime number, then

$$\sum_{a \in B^K} \left| I_a\left(z^K(t)\right) - I_a\left(z^K(0)\right)\right| \leq C\varepsilon. \tag{VII.17}$$

(ii) If $K = 2P$ and P is a prime number, then

$$\sum_{a=0}^{P} \left| J_a\left(z^K(t)\right) - J_a\left(z^K(0)\right)\right| \leq C\varepsilon, \tag{VII.18}$$

where for all $a = 0, \dots, P$, $J_a(z) = I_a(z) + I_{-a}(z)$.

Proof. The first part is a consequence of Lemma VII.5 and Proposition VII.4, using the same technique as in the proof of Proposition VII.3.
To prove the second part, we note that the Hamiltonian associated with the semi-discrete resonant system (VII.15) in the case $K = 2P$ and P is prime is written,

using the previous lemma,

$$Z^K(\xi,\eta) = \frac{\lambda}{2}\left(\sum_{a\in\mathbb{Z}} I_a^2 + 2\sum_{a\neq b} I_a\, I_b\right) + \lambda \sum_{\substack{a,b\in B^K,\, m\in\{\pm 1\}\\ a+b=mK/2}} \xi_a\eta_{-a}\xi_b\eta_{-b}. \tag{VII.19}$$

Hence, we see that the semi-discrete resonant Hamiltonian does not commute with the actions. However, it commutes with the *super actions* $J_a = I_a + I_{-a}$, for $a \in \{0,\dots,K/2\}$. To see this, we calculate

$$\begin{aligned}\{J_a, \xi_a\eta_{-a}\} &= \{I_a, \xi_a\eta_{-a}\} + \{I_{-a}, \xi_a\eta_{-a}\}\\ &= -\xi_a\eta_{-a} + \eta_{-a}\xi_a = 0.\end{aligned}$$

The equation (VII.18) is then easily shown using the fact that J_a is constant along the flow of (VII.15). ■

Note that the preservation of the super actions J_a yields the preservation of the actions in the case where the initial data $u^0(x)$ is odd or even, i.e. satisfy $u^0(-x) = u^0(x)$ or $u^0(-x) = -u^0(x)$. Actually such a property is carried to the exact solution $u(t,x)$ for all time, as well as for the semi-discrete solution, as can be easily verified. But in such a situation, we have $|u_a(x)|^2 = I_a = |u_{-a}(x)|^2 = I_{-a}$.

In more general situations, nonlinear instabilities can be observed. The example given in the introduction (see equation (I.21)) is constructed by noticing that when $K = 30 = 2\times 3\times 5$, we have $(-5, 14, -10, -11)$ belonging to $\mathcal{K}^K$ (for $m = 1$).

2.3 Fully discrete schemes. We now consider the case of fully discrete solutions obtained by splitting schemes. Actually, the results of the previous chapter show that the discrete dynamics obtained by a fully discrete splitting method can be interpreted as the exact flow of the modified Hamiltonian H_τ^K given by Theorem VI.8, up to an error of order $\mathcal{O}(\tau^{N+1})$ where N depends on the CFL number.

In the following, we will only consider the case of the "standard" splitting, for which the filter function β is the identity. In this case, the modified Hamiltonian can be written (see equation (VI.32) with $\Lambda_{abcd} = \tau(a^2+b^2-c^2-d^2)$)

$$\begin{aligned}H_\tau^K(\xi,\eta) = \sum_{a\in B^K} |a|^2\xi_a\eta_a + \frac{\varepsilon\lambda}{2}\\ \times \sum_{\substack{a+b-c-d=mK\\ (a,b,c,d)\in(B^K)^4,\,|m|\le 1}} \frac{i\tau\Omega_{abcd}}{\exp(i\tau\Omega_{abcd})-1}\xi_a\xi_b\eta_c\eta_d + \mathcal{O}\left(\varepsilon^2 z^6\right)\end{aligned} \tag{VII.20}$$

where we recall that $\Omega_{abcd} = a^2+b^2-c^2-d^2$. Note that in the construction of the modified energy, as the polynomial P is of order $\|P\| \simeq \varepsilon$, we easily see that for all

n, $\|Z_n\| \simeq \varepsilon^n$. Hence, the estimate (VI.30) is written here as

$$\left\|\varphi^{\tau}_{P^K} \circ \varphi^{1}_{A_0^K}\left(z^K\right) - \varphi^{\tau}_{H_\tau^K}\left(z^K\right)\right\|_{\ell^1} \leq C_N M^{2N+1} \varepsilon^{N+1} \tau^{N+1} \tag{VII.21}$$

if $\left\|z^K\right\|_{\ell^1} \leq M$ and ε is smaller than some fixed ε_0. Using the expression of the Hamiltonian (VII.20), we see that we can perform a similar analysis as before for the exact flow $\varphi^{\tau}_{H_\tau^K}$. In particular, as the difference between the first two terms $H_\tau^{K,1}$ (see (VI.31)) of H_τ^K and the full Hamiltonian H_τ^K is of order ε^2 for bounded z, a result similar to Proposition VII.4 and Theorem VII.6 can be derived. However, the proof is more complicated than in the semi-discrete case, and requires some preliminary results.

Lemma VII.7. *Let M be a fixed number. Then there exists a constant C depending on M such that for all ε the following holds: Assume that K is prime, and for $a \in B^K$, let us define*

$$G_a(\xi, \eta) := -\frac{\lambda}{2} \operatorname{Re}\left[\sum_{\substack{a=-b+c+d+mK \\ |m| \leq 1 \\ \Omega_{abcd} \neq 0}} \frac{i\tau}{\exp(i\tau\Omega_{abcd}) - 1} \xi_a \xi_b \eta_c \eta_d\right], \tag{VII.22}$$

where $\Omega_{abcd} = a^2 + b^2 - c^2 - d^2$, and let $z^{K,0} = (\xi^{K,0}, \eta^{K,0})$ be such that $\left\|z^{K,0}\right\|_{\ell^1} \leq M$. Let us define $z^{K,1} = \varphi^{\tau}_{H_\tau^K}(z^{K,0})$. Then we have

$$\sum_{a \in B^K} \left| I_a^{\varepsilon}\left(z^{K,1}\right) - I_a^{\varepsilon}\left(z^{K,0}\right)\right| \leq C\varepsilon^2\tau$$

where for all ε and all $a \in B^K$,

$$I_a^{\varepsilon}(z) = I_a(z) + \varepsilon G_a(z). \tag{VII.23}$$

Proof. For $t \in (0, \tau)$, let $z^K(t) = \varphi^{t}_{H_\tau^K}(z^{K,0}) = (\xi^K(t), \bar{\xi}^K(t))$ be the solution of the modified Hamiltonian system (VII.20). We can write for all $a \in B^K$,

$$\begin{aligned} \dot{\xi}_a^K &= -i|a|^2 \xi_a^K - i\varepsilon\lambda \\ &\times \sum_{\substack{a=-b+c+d+mK \\ |m| \leq 1}} \frac{i\tau\Omega_{abcd}}{\exp(i\tau\Omega_{abcd}) - 1} \bar{\xi}_b^K \xi_c^K \xi_d^K + \varepsilon^2 X_a^K\left(\xi^K, \bar{\xi}^K\right) \end{aligned}$$

where $X_a^K(z)$ is a Hamiltonian vector field that is bounded for bounded $z \in \ell^1$. With the notation of the previous chapter, the Hamiltonian function associated with the vector field X^K is $(Z_2 + \cdots + Z_N)/\varepsilon^2$.

Following the proof of Proposition VII.1, we define $Y^K(t)$ by

$$Y_a^K(t) = e^{it|a|^2}\xi_a^K(t), \quad a \in B^K.$$

Using the same method as in the proof of Proposition VII.4, we calculate that $I_a^K(t) := I_a(Y^K(t))$ satisfies the equation

$$\dot{I}_a^K(t) = -\operatorname{Re}\left[i\varepsilon\lambda \sum_{\substack{a=-b+c+d+mK \\ |m|\le 1}} \frac{i\tau\Omega_{abcd}}{\exp(i\tau\Omega_{abcd})-1} \bar{Y}_a^K \bar{Y}_b^K Y_c^K Y_d^Y e^{-it\Omega_{abcd}} \right] \\ + \varepsilon^2 \tilde{X}_a^K\left(Y^K\right)$$

where $\tilde{X}^K$ is a polynomial vector field in Y^K, bounded in ℓ^1 if Y^K is in ℓ^1. Hence, if we assume that $\left\|Y_a^K(t)\right\|_{\ell^1} \le M$, then there exists a constant $C(M)$ such that for $t \in [0, \tau]$,

$$I_a^K(t) = I_a^K(0) - \operatorname{Re}\left[i\varepsilon\lambda \sum_{\substack{a=-b+c+d=mK \\ |m|\le 1}} \frac{i\tau\Omega_{abcd}}{\exp\left(i\tau\Omega_{abcd}\right)-1} \right. \\ \left. \times \int_0^t \bar{Y}_a^K \bar{Y}_b^K Y_c^K Y_d^K e^{-is\Omega_{abcd}} \mathrm{d}s \right] + R_a^K\left(t, Y^K\right)$$

where

$$\left\|R_a^K(t, Y^K)\right\|_{\ell^1} \le C(M)\varepsilon^2\tau. \tag{VII.24}$$

Now let us define

$$G_a^K(t) = G_a\left(\xi^K(t)\right) = -\operatorname{Re}\left[\lambda \sum_{\substack{a=-b+c+d=mK \\ |m|\le 1 \\ \Omega_{abcd}\neq 0}} \frac{i\tau}{\exp\left(i\tau\Omega_{abcd}\right)-1} \bar{\xi}_a^K \bar{\xi}_b^K \xi_c^K \xi_d^K \right].$$

The previous estimate combined with an integration by parts, and the fact that K is prime (see Lemma VII.5) shows that

$$I_a^K(\tau) + \varepsilon G_a^K(\tau) = I_a^K(0) + \varepsilon G_a^K(0) + \tilde{R}_a^K\left(\tau, Y^K\right)$$

where $\left\|\tilde{R}_a(t, Y^K)\right\|_{\ell^1} \le C\varepsilon^2\tau$ if $\xi^K(0) \in B_M$. This shows the result. ■

Theorem VII.8. *Let N and $M > 0$ be fixed. There exists a constant C, T, τ_0 and ε_0 such that for all $\varepsilon \leq \varepsilon_0$, the following holds: For all prime integer K and all τ such that the CFL condition*

$$\tau K^2 < \frac{8\pi}{N+1}$$

holds (compare (VI.29)), let $z^{K,0} = (\xi^{K,0}, \bar{\xi}^{K,0})$ be such that $\left\| z^{K,0} \right\|_{\ell^1} \leq M/4$, and for all $n \in \mathbb{N}$, let

$$z^{K,n+1} := \varphi^{\tau}_{P^K} \circ \varphi^{1}_{A_0^K} \left(z^{K,n} \right)$$

be the fully discrete numerical solution obtained by the splitting methods applied to the semi-discretized Hamiltonian

$$P^K(\xi,\eta) = \frac{\varepsilon\lambda}{2} \sum_{\substack{a+b-c-d=mK \\ |m|\leq 1}} \xi_a \xi_b \eta_c \eta_d \quad \text{and} \quad A^K(\xi,\eta) = \sum_{a\in B^K} \tau \xi_a \eta_a .$$

Then we have

$$\sum_{a\in B^K} \left| I_a\left(z^{K,n}\right) - I_a\left(z^{K,0}\right) \right| \leq C\varepsilon, \quad \text{for} \quad n\tau \leq \frac{T}{\varepsilon}.$$

If $K = 2P$ with P prime, the same results hold true for the super actions $J_a = I_a + I_{-a}$.

Proof. Using the previous result, combined with (VII.21), we easily see that as long as $z^{K,n}$ is in B_M, we have the estimate

$$\begin{aligned}
&\sum_{a\in B^K} \left| \left(I_a(z^{K,n+1}) + \varepsilon G_a(z^{K,n+1})\right) - \left(I_a\left(z^{K,n}\right) + \varepsilon G_a\left(z^{K,n}\right)\right) \right| \\
&\leq C\varepsilon^2\tau + \varepsilon \sum_{a\in B^K} \left| G_a\left(z^{K,n+1}\right) - G_a\left(\varphi^{\tau}_{H^K_\tau}\left(z^{K,n}\right)\right) \right| \\
&\leq C\left(\varepsilon^2\tau + C\varepsilon^{N+2}\tau^{N+1}\right),
\end{aligned}$$

where the constant C depends on M. Hence if we assume that $\left\| z^{K,0} \right\|_{\ell^1} \leq M/4$, we obtain that $\sum_{a\in\mathbb{Z}} I_a^{\varepsilon}(z^{K,0}) \leq M/2$ if $\varepsilon \leq \varepsilon_0$ sufficiently small (see (VII.23)). Now using the previous estimate, we have that as long as $\left\| z^{K,n} \right\|_{\ell^1} \leq M$,

$$\sum_{a\in B^K} \left| I_a^{\varepsilon}(z^{K,n}) - I_a^{\varepsilon}(z^{K,0}) \right| \leq (n\tau)C\left(\varepsilon^2 + \varepsilon^{N+2}\tau^{N}\right).$$

But this shows that for

$$n\tau \leq T_{\varepsilon,\tau} := \frac{1}{C}\min\left(\varepsilon^{-1}, \varepsilon^{-N-1}\tau^{-N}\right),$$

and $\left\| z^{K,n} \right\|_{\ell^1} \leq M$, we have

$$\sum_{a \in B^K} \left| I_a^\varepsilon \left(z^{K,n} \right) - I_a^\varepsilon \left(z^{K,0} \right) \right| \leq \varepsilon.$$

But this shows that there exists a constant C depending on M such that as long as $z^{K,n}$ is in B_M and $n\tau \leq T_{\varepsilon,\tau}$,

$$\sum_{a \in B^K} \left| I_a \left(z^{K,n} \right) - I_a \left(z^{K,0} \right) \right| \leq C\varepsilon.$$

But for ε sufficiently small, this proves that $z^{K,n}$ is in B_M for all n such that $n\tau \leq T_{\varepsilon,\tau}$. This proves the result, as $N \geq 1$ and $\tau \leq \tau_0$, so that $T_{\varepsilon,\tau} = C^{-1}\varepsilon^{-1}$ for ε_0 sufficiently small. ■

Note that this result explains the behavior observed in Figures I.9, I.10 and I.11 in the introduction.

3 The case of dimension two

3.1 Energy cascades. We consider now the case of the Schrödinger equation (VII.1) set on a two-dimensional torus $\mathbb{T}^2$. In this situation, Proposition VII.1 applies, and the analysis of the long time qualitative behavior of the solution $u(t)$ of (VII.1) can be made through analysis of the resonant system (VII.5). The main difference with the one-dimensional case is that the frequencies can now interact in this resonant system, provided they are geometrically distributed on corners of rectangles (see Lemma VII.2). Using this, we can prove the following result (see [7]):

Theorem VII.9. *Let $d \geqslant 2$, and $u^0 \in C^\infty(\mathbb{T}^2)$ given by*

$$u^0(x) = 1 + 2\cos x_1 + 2\cos x_2.$$

For $\lambda \in \{\pm 1\}$, the following holds. There exist $\varepsilon_0, T, C_0, C > 0$ and a family $(c_a)_{a \in \mathcal{N}_*}$, with $c_a \neq 0$ for all a, such that for $0 < \varepsilon \leqslant \varepsilon_0$, (VI.1) has a unique solution $u(t,x) = \sum_{a \in \mathbb{Z}^d} \xi_a(t) e^{ia\cdot x} \in C([0, T/\varepsilon]; \ell^1)$, and:

$$\forall a \in \mathcal{N}_*,\ \forall t \in [0, T/\varepsilon], \quad \left| \xi_a(t) - c_a (\varepsilon t)^{|a|^2 - 1} \right| \leqslant (C_0 \varepsilon t)^{|a|^2} + C\varepsilon,$$

where the set $\mathcal{N}_*$ is given by

$$\mathcal{N}_* = \{(0, \pm 2^p), (\pm 2^p, 0), (\pm 2^p, \pm 2^p), (\mp 2^p, \pm 2^p),\ p \in \mathbb{N}\}. \tag{VII.25}$$

Arbitrarily high modes appear with equal intensity along a cascade of time layers:

$$\forall \gamma \in]0,1[,\ \forall \theta < \frac{1}{4},\ \forall \alpha > 0, \quad \exists \varepsilon_1 \in]0,\varepsilon_0], \quad \forall \varepsilon \in]0,\varepsilon_1],$$

$$\forall a \in \mathcal{N}_*,\ |a| < \alpha \left(\log \frac{1}{\varepsilon}\right)^{\theta}, \quad \left|\xi_a\left(\frac{2}{\varepsilon^{1-\gamma/(|a|^2-1)}}\right)\right| \geqslant \frac{\varepsilon^{\gamma}}{4}.$$

This theorem shows that for some high modes $a \in \mathcal{N}_*$ satisfying $|a| < \alpha \left(\log \frac{1}{\varepsilon}\right)^{\theta}$, the action $I_a(t)$ of the solution of (VII.1) with initial data $u^0(x_1,x_2) = 1+2\cos x_1 + 2\cos x_2$ satisfies

$$I_a(t_a) \geq c\varepsilon^{2\gamma},$$

for some time t_a increasing with $|a|$. This is very different from (VII.12), and shows the possibility of energy transfer from low to high modes

In the following, we will not give a complete prove of this theorem, and refer to [7] for the details. However, we would like to give a hint and explain the reason making this initial data produce an energy cascade.

To do this, we now turn to the analysis of the resonant system (VII.5). The main remark for the forthcoming analysis is that new modes can be generated by nonlinear interaction: we may have $y_a \neq 0$ even though $\xi_a^0 = 0$ in the system (VII.5).

Recall that nonlinear interactions in the resonant system are created from frequencies lying on rectangles, according to VII.2. Let us introduce the set of initial modes:

$$J_0 = \left\{a \in \mathbb{Z}^2 \mid \xi_a^0 \neq 0\right\}.$$

In view of (VII.5), modes which appear after one iteration of Lemma VII.2 are given by:

$$J_1 = \left\{a \in \mathbb{Z}^2 \setminus J_0 \mid \dot{y}_a(0) \neq 0\right\}.$$

One may also think of J_1 in terms of Picard iteration. Plugging the initial modes (from J_0) into the nonlinear Duhamel's term and passing to the limit $\varepsilon \to 0$, J_1 corresponds to the new modes resulting from this manipulation. More generally, modes appearing after k iterations exactly are characterized by:

$$J_k = \left\{a \in \mathbb{Z}^2 \setminus \bigcup_{\ell=0}^{k-1} J_\ell \mid \frac{\mathrm{d}^k}{\mathrm{d}t^k} y_a(0) \neq 0\right\}.$$

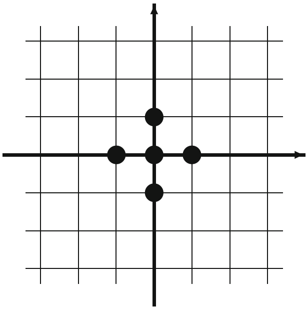
We now consider the initial datum

$$u^0(x) = 1 + 2\cos x_1 + 2\cos x_2 = 1 + e^{ix_1} + e^{-ix_1} + e^{ix_2} + e^{-ix_2}. \qquad \text{(VII.26)}$$

The corresponding set of initial modes is given by

$$J_0 = \{(0,0), (1,0), (-1,0), (0,1), (0,-1)\}.$$

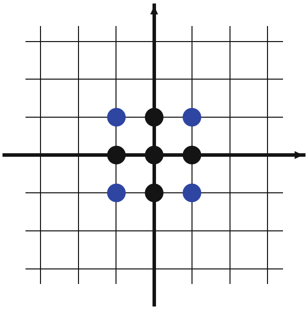
It is represented on the following figure:



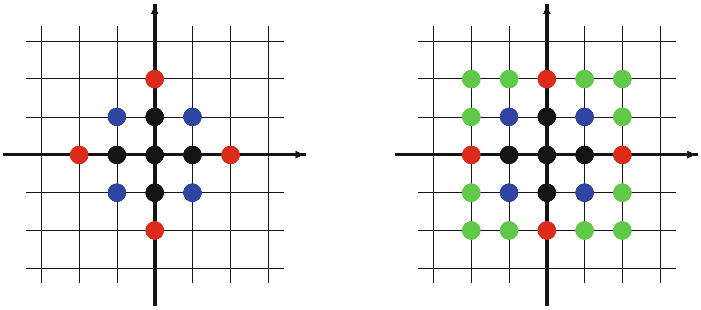
After one iteration of Lemma VII.2, four points appear:

$$J_1 = \{(1,1),(1,-1),(-1,-1),(-1,1)\},$$

as plotted below.



The next two steps are described geometrically:



As suggested by these illustrations, we can prove by induction:

Lemma VII.10. *Let $p \in \mathbb{N}$.*

- *The set of relevant modes after $2p$ iterations is the square of length 2^p whose diagonals are parallel to the axes:*

$$\mathcal{N}^{(2p)} := \bigcup_{\ell=0}^{2p} J_\ell = \{(a_1,a_2) \mid |a_1| + |a_2| \leqslant 2^p\}\,.$$

- *The set of relevant modes after* $2p+1$ *iterations is the square of length* 2^{p+1} *whose sides are parallel to the axes:*

$$\mathcal{N}^{(2p+1)} := \bigcup_{\ell=0}^{2p+1} J_\ell = \{(a_1,a_2) \mid \max(|a_1|,|a_2|) \leqslant 2^p\}\,.$$

After an infinite number of iterations, the whole lattice $\mathbb{Z}^2$ is generated:

$$\bigcup_{k\geqslant 0} \mathcal{N}^{(k)} = \mathbb{Z}^2.$$

Among these sets, our interest will focus on *extremal modes*: for $p \in \mathbb{N}$,

$$\mathcal{N}_*^{(2p)} := \{(a_1,a_2) \in \{(0,\pm 2^p),(\pm 2^p,0)\}\}\,,$$
$$\mathcal{N}_*^{(2p+1)} := \{(a_1,a_2) \in \{(\pm 2^p,\pm 2^p),(\mp 2^p,\pm 2^p)\}\}\,.$$

These sets correspond to the edges of the squares obtained successively by iteration of Lemma VII.2 on J_0. The set $\mathcal{N}_*$ defined in Theorem VII.9 corresponds to

$$\mathcal{N}_* = \bigcup_{k\geqslant 0} \mathcal{N}_*^{(k)}.$$

The important property associated to these extremal points is that they are generated in a unique fashion:

Lemma VII.11. *Let* $n \geqslant 1$, *and* $a \in \mathcal{N}_*^{(n)}$. *There exists a unique pair* $(b,c) \in \mathcal{N}^{(n-1)} \times \mathcal{N}^{(n-1)}$ *such that* a *is generated by the interaction of the modes* 0, b *and* c, *up to the permutation of* b *and* c. *More precisely,* b *and* c *are extremal points generated at the previous step:* $b,c \in \mathcal{N}_*^{(n-1)}$.

Note however that points in $\mathcal{N}_*^{(n)}$ are generated in a non-unique fashion by the interaction of modes in $\mathbb{Z}^d$. For instance, $(1,1) \in J_1$ is generated after one step only by the interaction of $(0,0)$, $(1,0)$ and $(0,1)$. On the other hand, we see that after two iterations, $(1,1)$ is fed also by the interaction of the other three points in $\mathcal{N}_*^{(1)}$, $(-1,1)$, $(-1,-1)$ and $(1,-1)$. After three iterations, there are even more three-wave interactions affecting $(1,1)$.

The method to prove Theorem VII.9 is to show that we can compute the first non-zero term in the Taylor expansion of solution $y_m(t)$ of (VII.5) at $t=0$, for $m \in \mathcal{N}_*$.

Let $n \geqslant 1$ and $a \in \mathcal{N}_*^{(n)}$. Note that since we have considered initial coefficients which are all equal to 1 – see (VII.26) – and because of the symmetry in (VII.5), the coefficients $y_a(t)$ do not depend on $a \in \mathcal{N}_*^{(n)}$ but only on n.

Hence we have

$$y_a(t) = \frac{t^{\alpha(n)}}{\alpha(n)!}\frac{\mathrm{d}^{\alpha(n)}y_a}{\mathrm{d}t^{\alpha(n)}}(0) + \frac{t^{\alpha(n)+1}}{\alpha(n)!}\int_0^1 (1-\theta)^{\alpha(n)}\frac{\mathrm{d}^{\alpha(n)+1}y_a}{\mathrm{d}t^{\alpha(n)+1}}(\theta t)\mathrm{d}\theta,$$

for some $\alpha(n) \in \mathbb{N}$ still to be determined.

The second term can be controlled, and is order $t^{\alpha(n)+1}$. In fact, by analyticity of the function $t \mapsto y_a(t)$, we can show that there exists $C_0 > 0$ *independent of a and n* such that

$$r_a(t) = \frac{t^{\alpha(n)+1}}{\alpha(n)!} \int_0^1 (1-\theta)^{\alpha(n)} \frac{\mathrm{d}^{\alpha(n)+1} y_a}{\mathrm{d}t^{\alpha(n)+1}}(\theta t)\mathrm{d}\theta$$

satisfies

$$|r_a(t)| \leqslant (C_0 t)^{\alpha(n)+1}. \tag{VII.27}$$

We refer to [7] for the details. Next, we write

$$y_a(t) = c(n)t^{\alpha(n)} + r_a(t), \tag{VII.28}$$

and we determine $c(n)$ and $\alpha(n)$ thanks to the iterative approach analyzed in the previous paragraph. In view of Lemma VII.11, we have

$$i\dot{y}_a = 2\lambda c(n-1)^2 t^{2\alpha(n-1)} + \mathcal{O}\left(t^{2\alpha(n-1)+1}\right),$$

where the factor 2 accounts for the fact that the vectors b and c can be exchanged in Lemma VII.11. We infer the relations:

$$\alpha(n) = 2\alpha(n-1) + 1 \quad ; \quad \alpha(0) = 0.$$
$$c(n) = -2i\lambda \frac{c(n-1)^2}{2\alpha(n-1)+1} \quad ; \quad c(0) = 1.$$

We first derive

$$\alpha(n) = 2^n - 1.$$

We can then compute, $c(1) = -2i\lambda$, and for $n \geqslant 1$:

$$c(n+1) = i\frac{(2\lambda)^{\sum_{k=0}^n 2^k}}{\prod_{k=1}^{n+1}\left(2^k-1\right)^{2^{n+1-k}}} = i\frac{(2\lambda)^{2^{n+1}-1}}{\prod_{k=1}^{n+1}\left(2^k-1\right)^{2^{n+1-k}}}.$$

We can then infer the first estimate of Theorem VII.9: by Proposition VII.1, there exists C independent of a and ε such that for $0 < \varepsilon \leqslant \varepsilon_0$,

$$|\xi_a(t) - y_a(\varepsilon t)| \leqslant C\varepsilon, \quad 0 \leqslant t \leqslant \frac{T}{\varepsilon}.$$

We notice that since for $a \in \mathcal{N}_*^{(n)}$, $|a| = 2^{n/2}$, regardless of the parity of n, we have $\alpha(n) = |j|^2 - 1$. For $j \in \mathcal{N}_*$, we then use (VII.28) and (VII.27), and the estimate follows, with $c_j = c(n)$.

To prove the last estimate of Theorem VII.9, we must examine more closely the behavior of $c(n)$. In [7], it is proved that for all $n \geq 1$, we have

$$|c(n)| \geqslant 2^{-2^n}.$$

We can now gather all the estimates together:

$$\begin{aligned} |\xi_a(t)| &\geqslant \left|c(n)\,(\varepsilon t)^{\alpha(n)}\right| - (C_0\varepsilon t)^{\alpha(n)+1} - C\varepsilon \\ &\geqslant \frac{1}{2}\left(\frac{\varepsilon t}{2}\right)^{2^n-1} - (C_0\varepsilon t)^{2^n} - C\varepsilon \\ &\geqslant \frac{1}{2}\left(\frac{\varepsilon t}{2}\right)^{2^n-1}\left(1-(2C_0)^{2^n}\varepsilon t\right) - C\varepsilon. \end{aligned} \tag{VII.29}$$

To conclude, we simply consider t such that

$$\left(\frac{\varepsilon t}{2}\right)^{2^n-1} = \varepsilon^\gamma, \text{ that is } t = \frac{2}{\varepsilon^{1-\gamma/\alpha(n)}}. \tag{VII.30}$$

Hence for the time t given in (VII.30), since $\alpha(n) = |a|^2 - 1$, we have

$$\begin{aligned} (2C_0)^{2^n}\varepsilon t &= (2C_0)^{|a|^2}\,\varepsilon^{\gamma/(|a|^2-1)} \\ &= \exp\left(|a|^2\log(2C_0) - \frac{\gamma}{|a|^2-1}\log\left(\frac{1}{\varepsilon}\right)\right). \end{aligned}$$

Assuming the spectral localization

$$|a| \leqslant \alpha\left(\log\frac{1}{\varepsilon}\right)^\theta,$$

we get for ε small enough

$$(2C_0)^{2^n}\varepsilon t \leqslant \exp\left(\alpha^2\left(\log\frac{1}{\varepsilon}\right)^{2\theta}\log(2C_0) - \frac{\gamma}{\alpha^2}\left(\log\frac{1}{\varepsilon}\right)^{1-2\theta}\right).$$

The argument of the exponential goes to $-\infty$ as $\varepsilon \to 0$ provided that

$$\gamma > 0 \quad\text{and}\quad \theta < \frac{1}{4},$$

in which case we have $1 - (2C_0)^{2^n}\varepsilon t > 3/4$ for ε sufficiently small. Inequality (VII.29) then yields the result, owing to the fact that $C\varepsilon^2$ is negligible compared to ε^γ when $0 \leqslant \gamma < 1$.

Finally, we note that the choice (VII.30) is consistent with $\varepsilon t \in [0, T]$ for $\varepsilon \leqslant \varepsilon_0$, for some $\varepsilon_0 > 0$ uniform in a satisfying the above spectral localization, since

$$\varepsilon^{\gamma/\alpha(n)} = e^{-\frac{\gamma}{\alpha(n)}\log\frac{1}{\varepsilon}} \leqslant \exp\left(-\frac{\gamma}{\alpha^2}\left(\log\frac{1}{\varepsilon}\right)^{1-2\theta}\right) \underset{\varepsilon\to 0}{\longrightarrow} 0.$$

3.2 Simulating energy cascades. The energy cascade described above can be seen in Figure I.12 where the step-size τ is very small, and the number of modes K sufficiently large. However, as seen in Figure I.13 and I.14, the correct reproduction of the energy exchanges is not guaranteed in the case of implicit schemes.

Actually, the results of Chapter VI remain valid in the two-dimensional case. Considering the numerical fully discrete solution obtained from a scheme of the form $\varphi^{\tau}_{P^K} \circ \varphi^{1}_{A^K}$ where

$$A^K(\xi,\eta) = \sum_{a \in B^K} \lambda_a \xi_a \eta_a$$

where B^K is a (two-dimensional) finite subset of indices, and where $\lambda_a = \beta(\tau|a|^2)$, it can be shown that a result similar to Theorem VI.8 can be proven, with a modified Hamiltonian of the form

$$\begin{aligned} H^K_\tau(\xi,\eta) = {} & \sum_{a \in B^K} \frac{1}{\tau}\beta(\tau|a|^2)\xi_a\eta_a + \frac{\varepsilon\lambda}{2} \\ & \times \sum_{\substack{a+b-c-d=mK \\ m \in \mathbb{Z}^2,\, |m| \le 1}} \frac{i\Lambda_{abcd}}{e^{i\Lambda_{abcd}} - 1} \xi_a\xi_b\eta_c\eta_d + \mathcal{O}(\varepsilon^2 z) \end{aligned}$$

where $\Lambda_{abcd} = \lambda_a + \lambda_b - \lambda_c - \lambda_d$.

Hence we see that we can perform a similar analysis as in Proposition VII.1 for instance, but the corresponding resonance modulus will be

$$\begin{aligned} \big\{(a,b,c,d) \in \mathbb{Z}^2 \,|\, a+b-c-d = mK, \quad m \in \mathbb{Z}^2 \\ \text{and} \quad \lambda_a + \lambda_b - \lambda_c - \lambda_d = 0\big\}. \end{aligned}$$

We thus see that except the case where $\beta(x) = x$, this resonance modulus does not satisfy Lemma VII.2 in general, and the numerical solution will be unable to reproduce correctly the energy exchanges. This is particularly the case for implicit-explicit schemes based on the filter function $\beta(x) = 2\arctan(x/2)$. Note however that if τ is sufficiently small, then we have $\lambda_a = \tau|a|^2 + \mathcal{O}(\tau^2)$ at least for the low frequencies, and we thus expect that, at least for them, the energy exchanges will be correctly reproduced.

By using similar technics as developed in the previous section, it is possible to prove that Theorem VII.9 can be extended to the situation where $\beta(x) = x$. We do not give the details here. Note that as the phenomena is a propagation of energy to high frequencies, there is no aliasing problem until the frequencies of order $K/2$ are reached by the cascade.

Bibliography

[1] U.M. Ascher and S. Reich, *The midpoint scheme and variants for Hamiltonian systems: advantages and pitfalls,* SIAM J. Sci. Comput. 21 (1999), 1045–1065.

[2] H.F. Baker, *Alternants and continuous groups*, Proc. of London Math. Soc. 3 (1905), 24–47.

[3] W. Bao and J. Shen, *A fourth-order time-splitting Laguerre–Hermite pseudospectral method for Bose–Einstein condensates*, SIAM J. Sci. Comput. 26 (2005), 2010–2028.

[4] W. Bao and J. Shen, *A generalised-Laguerre–Hermite pseudospectral method for computing symmetric and central vortex states in Bose–Einstein condensates*, J. Comput. Phys. 227 (2008), 9778–9793.

[5] G. Benettin and A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys. 74 (1994), 1117–1143.

[6] D. Bambusi and B. Grébert, *Birkhoff normal form for PDE's with tame modulus*. Duke Math. J. 135 no. 3 (2006), 507–567.

[7] R. Carles and E. Faou, *Energy cascades for NLS on the torus*. http://arxiv.org/abs/1010.5173

[8] T. Cazenave *Semilinear Schrödinger equations*. Courant Lecture Notes in Mathematics, 10. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2003.

[9] D. Cohen, E. Hairer, and C. Lubich, *Conservation of energy, momentum and actions in numerical discretizations of nonlinear wave equations*, Numer. Math. 110 (2008), 113–143.

[10] R. Courant, K. Friedrichs, and H. Lewy, *Über die partiellen Differenzengleichungen der mathematischen Physik*, Math. Ann. 100 (1928), 32–74.

[11] H.L. Eliasson and S.B. Kuksin, *KAM for non-linear Schroedinger equation*, Ann. Math. 172 (2010), 371–435.

[12] A. Debussche and E. Faou, *Modified energy for split-step methods applied to the linear Schrödinger equation*, SIAM J. Numer. Anal. 47 (2009), 3705–3719.

[13] G. Dujardin and E. Faou, *Normal form and long time analysis of splitting schemes for the linear Schrödinger equation with small potential*, Numer. Math. 106, 2 (2007), 223–262.

[14] E. Faou, V. Gradinaru, and C. Lubich, *Computing semi-classical quantum dynamics with Hagedorn wavepackets*, SIAM J. Sci. Comp. 31 (2009), 3027–3041.

[15] E. Faou and B. Grébert, *Hamiltonian interpolation of splitting approximation for Hamiltonian PDEs*. Found. Comput. Math. 11 (2011), 381–415.

[16] E. Faou and B. Grébert, *A Nekhoroshev type theorem for the nonlinear Schrödinger equation on the torus*. http://arxiv.org/abs/1003.4845

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

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

[19] L. Gauckler, *Convergence of a split-step Hermite method for the Gross–Pitaevskii equation*, IMA J. Numer. Anal. 31(2) (2011), 396–415.

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

[21] L. Gauckler and C. Lubich, *Nonlinear Schrödinger equations and their spectral discretizations over long times*, Found. Comput. Math. 10 (2010), 141–169.

[22] B. Grébert, *Birkhoff normal form and Hamiltonian PDEs.* Séminaires et Congrès 15 (2007), 1–46.

[23] B. Grébert, T. Kappeler, and J. Pöschel, *Normal form theory for the NLS equation*. http://arxiv.org/abs/0907.3938

[24] E. Hairer, *Backward analysis of numerical integrators and symplectic methods.* Ann. Numer. Math. 1 (1994), 107–132.

[25] E. Hairer and C. Lubich, *The life-span of backward error analysis for numerical integrators*, Numer. Math. 76 (1997), 441–462.

[26] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Second Edition. Springer 2006.

[27] E. Hansen and A. Ostermann, *Exponential splitting for unbounded operators.* Math. Comp. 78 (2009), 1485–1496.

[28] F. Hausdorff, *Die symbolische Exponentialformel in der Gruppentheorie.* Ber. Sächs. Akad. Wiss. 58 (1906), 19–48.

[29] T. Jahnke and C. Lubich, *Error bounds for exponential operator splittings*, BIT 40 (2000), 735–744.

[30] C. Lubich, *From quantum to classical molecular dynamics: reduced models and numerical analysis*. European Math. Soc., 2008.

[31] C. Lubich, *On splitting methods for Schrödinger–Poisson and cubic nonlinear Schrödinger equations*, Math. Comp. 77 (2008), 2141–2153.

[32] J. Moser, *Lectures on Hamiltonian systems*, Mem. Am. Math. Soc. 81 (1968), 1–60.

[33] S. Reich, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal. 36 (1999), 1549–1570.

[34] B. Leimkuhler and S. Reich, *Simulating Hamiltonian dynamics*. Cambridge Monographs on Applied and Computational Mathematics, 14. Cambridge University Press, Cambridge, 2004.

[35] A. Stern and E. Grinspun, *Implicit-explicit variational integration of highly oscillatory problems*, preprint (2008).

[36] C. Sulem and P.-L. Sulem, *The nonlinear Schrödinger equation. Self focusing and wave collapse*. Appl. Math. Sci., 139. Springer, New-York, 1999.

[37] V.E. Zakharov and A.B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*. Sov. Phys. JETP 34(1) (1972), 62–69.

Index