# 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 $\lambda$ 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 $\xi_a^0$ 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 $\varphi_{T+P}^\tau$ 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 $\tau$ 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 $\xi_a^K$, $a \in B^K$, 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 $\xi_a^K$ 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 $t_n = n\tau \leq T$, and we have the estimate

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

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 $\ell^1$ 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 $\ell^1$. In these notes, $\ell^1$-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 $\beta$ is a cut-off function such that $\beta(x) = x$ for $|x| \leq \mathsf{cfl}$ and $\beta(x) = 0$ for $|x| > \mathsf{cfl}$ where the constant $\mathsf{cfl}$ 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 $\tau$ 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 $\xi_a^n$ 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 $\varphi_T^{\tau}$ 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

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

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

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

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 $\beta(x) = \frac{2}{\tau}\arctan(x/2)$. 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 $\alpha$, $c$, $x_0$ and $\theta_0$ 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

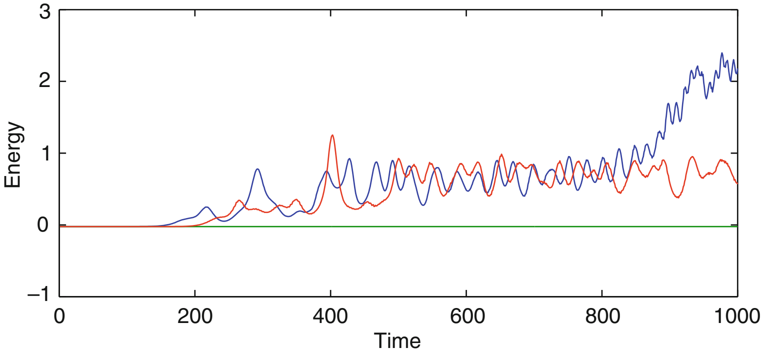

Figure I.1. Evolution of energy for the Strang splitting with $\mathsf{cfl} = 19.8$, 9.9 and 1.9.

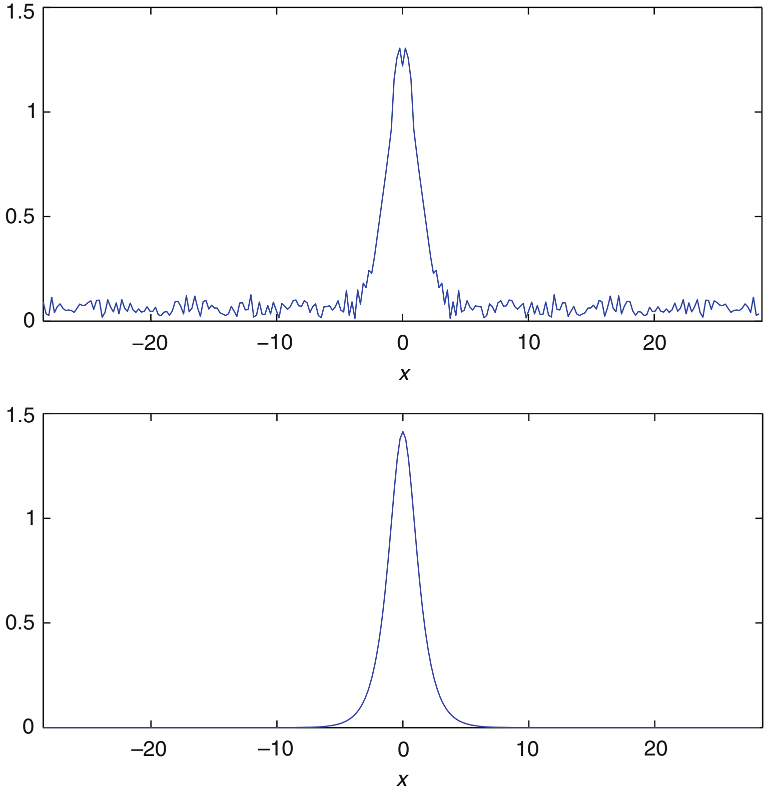

Figure I.2. $|u^n(x)|$ for the Lie splitting with cfl = 19.8 at time $t = 300$ (top) and cfl = 1.9 at time $t = 10000$ (bottom).

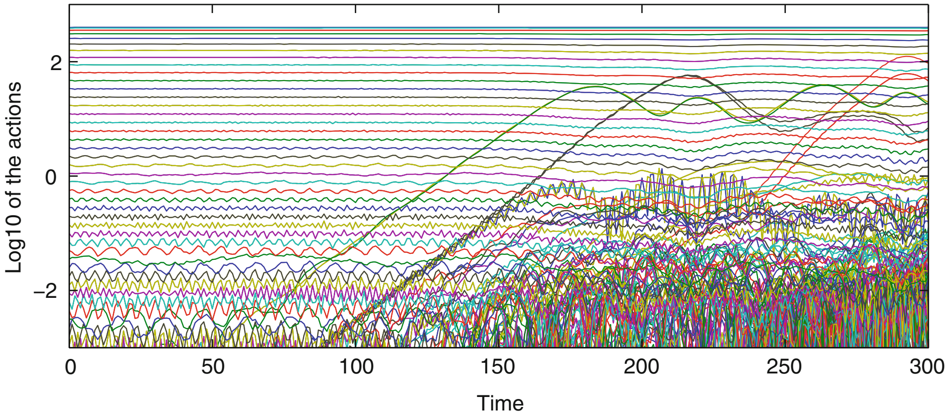

Figure I.3. Evolution of the actions for the Lie splitting with cfl = 19.8.

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.

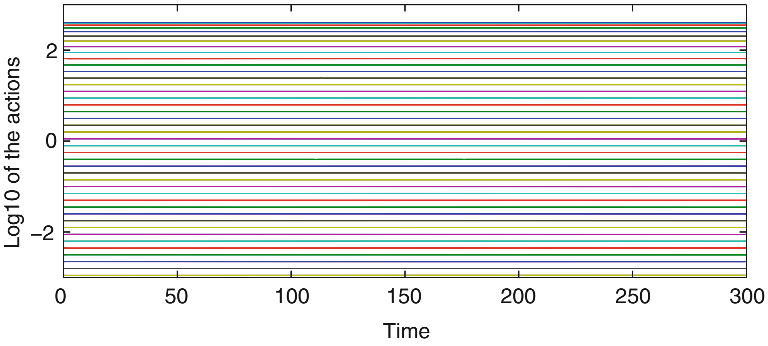

Figure I.4. Evolution of the actions for the Lie splitting with cfl = 1.9.

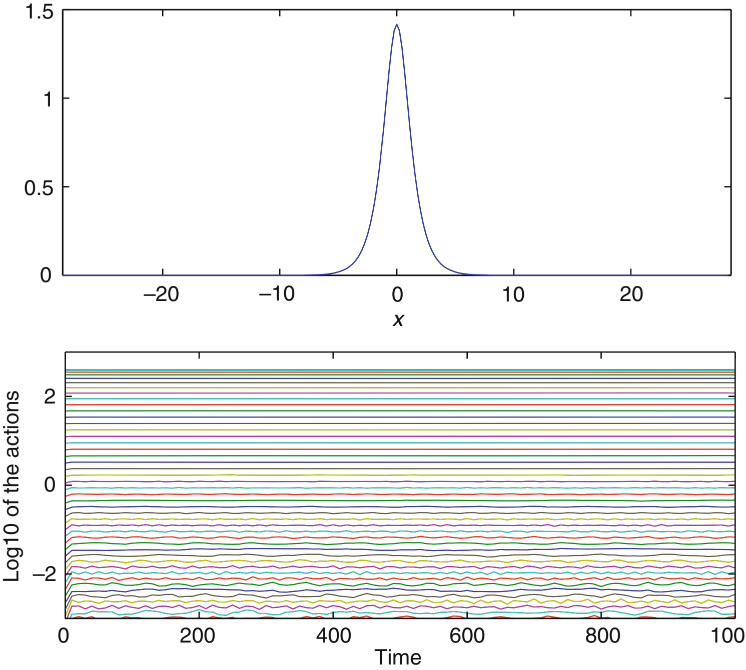

Figure I.5. Implicit-explicit integrator with cfl $= 19.8$. Profile at $t = 1000$ (top) and evolution of the actions (bottom).

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 $\tau$, we define a numerical solution $u_\tau^n$ 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 $\tau$ 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 $\tau^{-1}$), 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 $\tau$ 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,

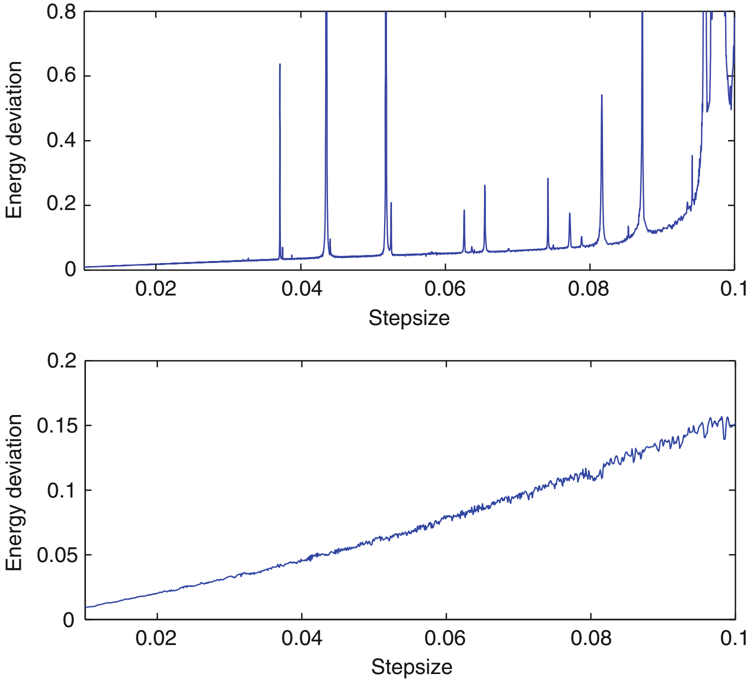

Figure I.6. Energy deviation as function of a time step for a Lie splitting (top) and the implicit-explicit scheme (bottom).

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:

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

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

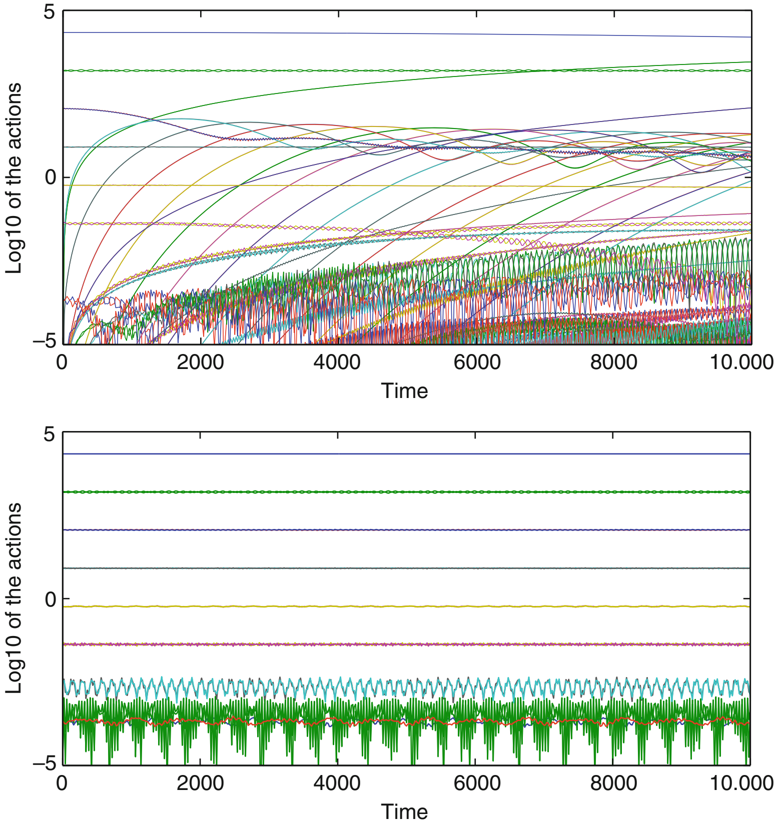

Figure I.7. Evolution of the actions (linear case) for a Lie splitting with resonant step size (top) and non resonant step size (bottom).

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 $\tau$! 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 $\delta$ 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

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

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

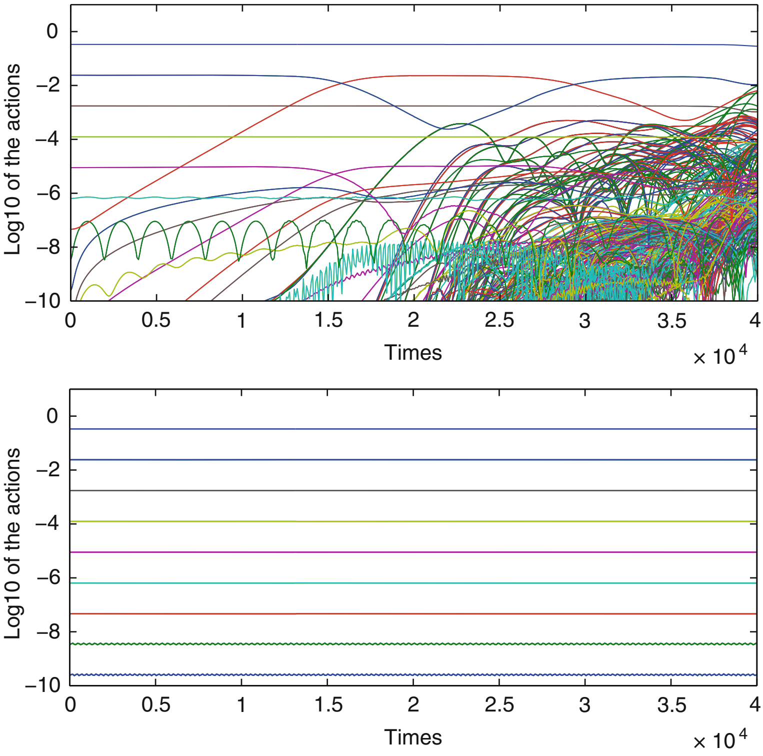

Figure I.8. Evolution of the actions in dimension 1 for resonant and non resonant step sizes.

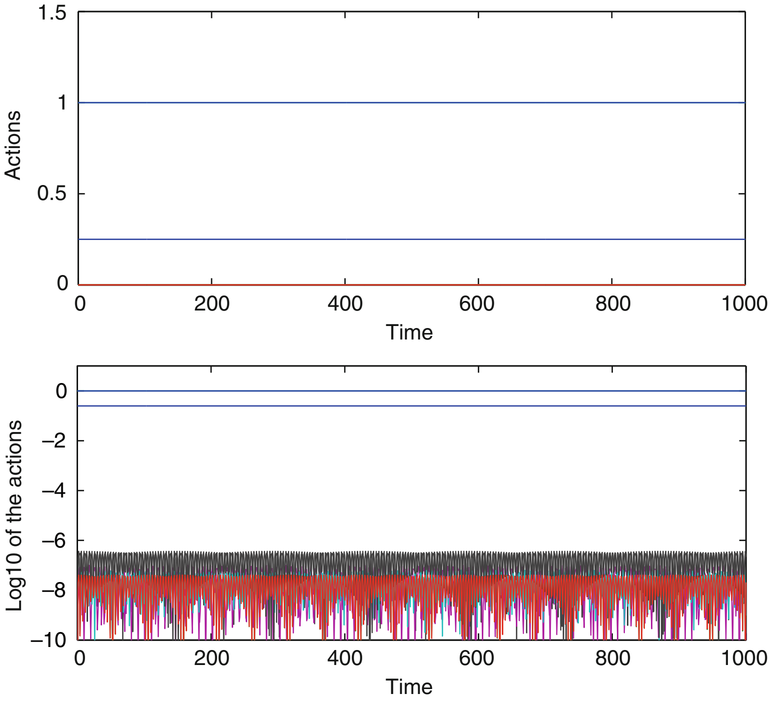

Figure I.9. Evolution of the actions in dimension 1 for $K = 31$ grid points.

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 $K = 34 = 2 \times 17$).

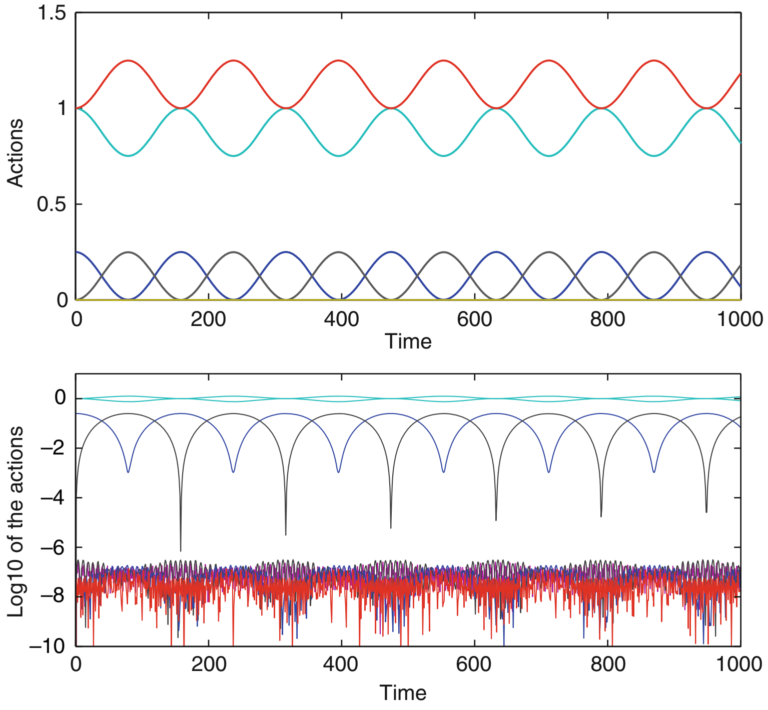

Figure I.10. Evolution of the actions in dimension 1 for $K = 34$ grid points.

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:

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

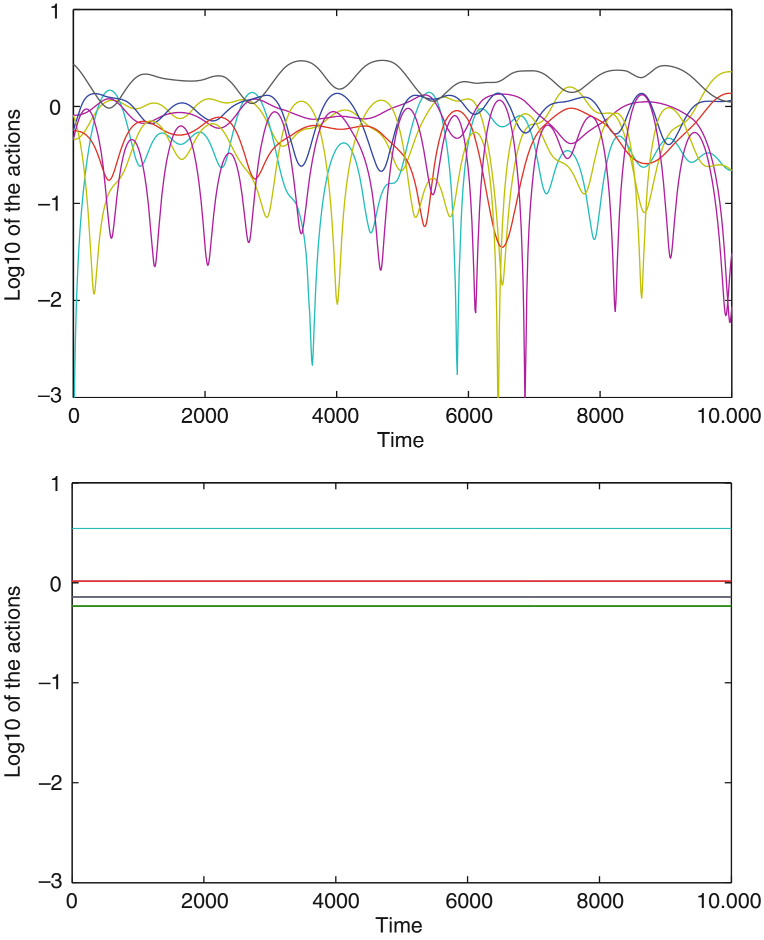

Figure I.11. Evolution of the actions for $K = 30$ (top) and $K = 31$ (bottom).

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).

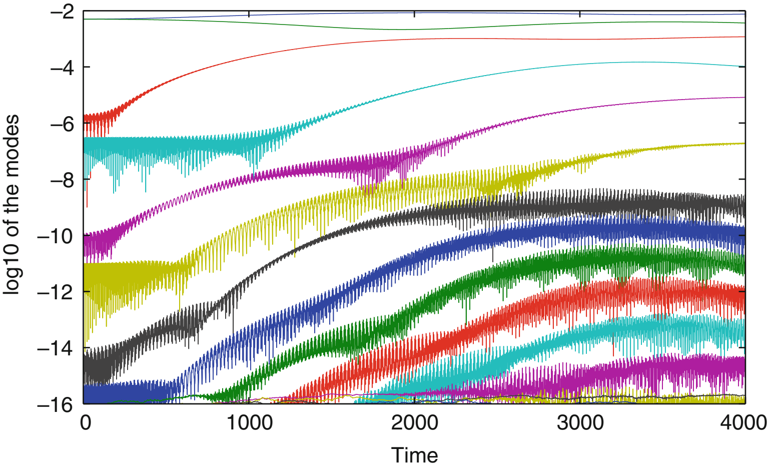

Figure I.12. Energy cascade.

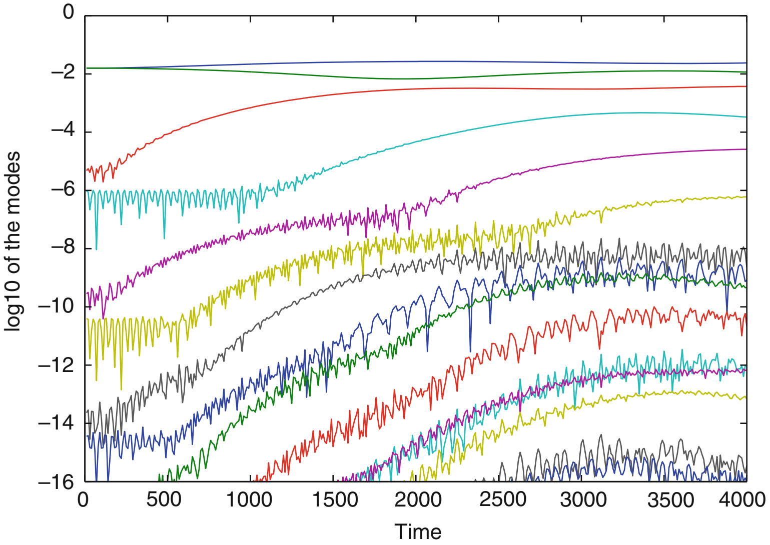

Figure I.13. Explicit scheme, $\tau = 0.1$, grid $128 \times 128$.

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 $128 \times 128$ 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.

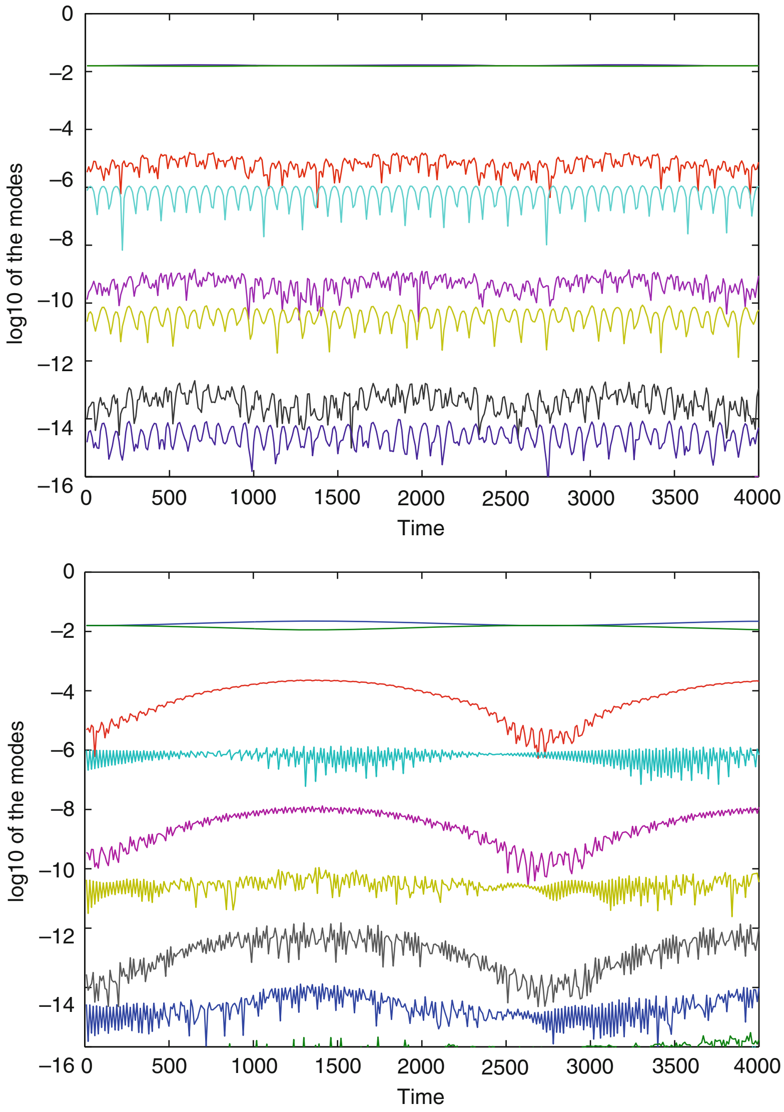

Figure I.14. Implicit-explicit integrator, $\tau = 0.1$ and $\tau = 0.05$.

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 $H_\tau$ 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 $\ell^1$, and where $C_N$ depends on the size of the function $u$ in $\ell^1$. 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 $\ell^1$ norm in estimate (I.24) can be replaced by the Sobolev norm $H^1$, and that the modified Hamiltonian $H_\tau$ controls the $H^1$ norm of (small) fully discrete solutions.

With this modified energy $H_\tau$ 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.